APPLICATIONS OF HYBRID AND DIGITAL COMPUTATION METHODS IN AEROSPACE-RELATED SCIENCES AND ENGINEERING

FINAL REPORT

Submitted to

National Aeronautics and Space Administration

Washington, D.C.

Grant No. NGL-44-005-084

(NASA-CR-157364) APPLICATIONS OF HYBRID AND DIGITAL COMPUTATION METHODS IN AEROSPACE-RELATED SCIENCES AND ENGINEERING Final Report (Houston Univ.) 486 p HC A21/MF A01 CSCL 12A G3/64

N78-29820

Unclas 28492

by

College of Engineering

University of Houston

REPRODUCED BY
NATIONAL TECHNICAL INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

521

TABLE OF CONTENTS

SUMMARY 1

INDIVIDUAL PROJECT ABSTRACTS AND PUBLICATIONS 4

(84 PROJECTS)

APPENDIX

A. Name of Participating Departments and Faculty 461

B. EGR 630 Hybrid Computation, Typical Laboratory/ Homework Assignments 464

C. Description of Engineering Systems Simulation Laboratory, College of Engineering 469

APPLICATIONS OF HYBRID AND DIGITAL COMPUTATION METHODS IN AEROSPACE-RELATED SCIENCES AND ENGINEERING

FINAL REPORT

Submitted to

National Aeronautics and Space Administration

Washington, D.C.

Grant No. NGL-44-005-084

by

College of Engineering

University of Houston

C. J. Huang
Co-Principal Investigator

R. L. Motard
Co-Principal Investigator

APPLICATIONS OF HYBRID AND DIGITAL COMPUTATION METHODS IN AEROSPACE-RELATED SCIENCE AND ENGINEERING

Final Report

NASA Grant NGL-44-005-084

SUMMARY

The Cullen College of Engineering at the University of Houston has been actively developing its educational and research capability in applications of digital/hybrid computation in solving engineering and scientific problems.

The Grant NGL-44-005-084 from National Aeronautics and Space Administration provided the funds for computer time and personnel required for research projects applying new methods in hybrid and digital computation.

(Continued)

The accomplishments of this program may be summarized as follows:

1. The grant supported application of hybrid/digital computation of 84 research projects in the following aerospace related disciplines.

   Departments

   Chemical Engineering

   Civil Engineering

   Electrical Engineering

   Industrial Engineering

   Mechanical Engineering

   Psychology

   Interdisciplinary Programs

   Acoustics

   Aerospace Engineering

   Energy Sciences & Engineering

   Environmental Engineering

   Systems Engineering

   There are altogether 21 faculty members and 68 graduate students participated in the programs. The latter include 36 Ph.D. students and 32 M.S. students. The departments, names of faculty members and number of graduate students of the above 84 projects are given in Appendix A. The project abstracts and publications of these 84 projects are given in the following chapters.

2. Under the sponsorship of this grant, a new course in hibrid computation (EGR 630) has been offered to graduate students in engineering. Typical laboratory/homework assignments of EGR 630 are included in Appendix B.

3. The Engineering Systems Simulation Laboratory of the College of Engineering is the focal point providing assistance to the above educational and research programs. The description of the Engineering Systems Simulation Laboratory is given in Appendix C.

The College of Engineering of the University of Houston gratefully acknowledges the NASA support which made the above accomplishments possible. Building upon the progresses made under this grant, the College of Engineering at the University of Houston will continue to make a stride in developing new techniques in hybrid and digital computation for aero-space related research.

PROJECT ABSTRACTS AND

PUBLICATION REPRINTS,

NAMES OF DEPARTMENT,

FACULTY ADVISOR, AND

GRADUATE STUDENTS

84 PROJECTS

Project

(A) Project Title: Continued Fraction Inversion by Routh's Algorithm

(B) Project Abstract:

The operation of converting a continued fraction into a rational transfer function of two polynomials is tedious. By the use of state-space techniques and Routh's algorithm, a new method is established for performing the continued fraction inversion.

(C) Publication: IEEE Transactions on Circuit Theory, CT-16, No. 2, 197 (1969)

(D) Year: 1969

(E) Department: Electrical Engineering

(F) Student Name: L. S. Shieh

(G) Faculty Advisor: Professor C. F. Chen

Continued Fraction Inversion by Routh's Algorithm

CHIH-FAN CHEN, SENIOR MEMBER, IEEE, AND LEANG-SAN SHIEH, STUDENT MEMBER, IEEE

Abstract—The operation of converting a continued fraction into a rational transfer function of two polynomials is tedious. By the use of state-space techniques and Routh's algorithm, a new method is established for performing the continued fraction inversion.

Introduction

EXPANDING a rational transfer function into a continued fraction and inverting a continued fraction to a transfer function are two fundamentally important operations in network synthesis, control system analysis, etc. Theoretically the two operations are trivial. One involves many divisions and the other is related to many multiplications. Practically speaking, however, when the order is high, the heavy labor of doing the multiplications and divisions is unavoidable. Facing the tedious work, we naturally think of an algorithmic approach to the problem in order that we can use the digital computer to free us from drudgery.

Routh's algorithm and continued fractions were first associated by Wall [1] in 1945. Frank [2] extended and modified his work further in 1946. However, they applied Routh's algorithm only to the expansion aspect, not to the inversion problem. It is known that the latter is much more difficult and tedious than the former.

This paper attempts to develop an algorithmic method for solving the inversion problem. In other words, how do we convert a continued fraction into a rational fraction of two polynomials in the easiest way.

Three Forms of Continued Fractions

Consider the following rational function:

$$\frac{g_2(s)}{g_1(s)} = \frac{A_{2,n}s^{n-1} + \cdots + A_{24}s^3 + A_{23}s^2 + A_{22}s + A_{21}}{A_{1,(n+1)}s^n + \cdots + A_{14}s^3 + A_{13}s^2 + A_{12}s + A_{11}} \tag{1}$$

where A_{ij} are constants.

We can expand (1) into several continued fraction forms. There are, however, three most important ones in engineering applications.

1) *The Stieltjes Form* [3]:

$$\frac{g_2(s)}{g_1(s)} = \cfrac{1}{a_1 s + b_1 + \cfrac{1}{a_2 s + b_2 + \cfrac{1}{a_3 s + b_3 + \cfrac{1}{\ddots}}}} \tag{2}$$

Manuscript received May 15, 1968; revised August 15, 1968, and September 11, 1968. This work was supported in part by the National Aeronautics and Space Administration under Grant NGL-44-005-084.

The authors are with the Department of Electrical Engineering, University of Houston, Houston, Tex.

Wall developed a technique to expand (1) into (2) by using Routh's algorithm. Frank extended this work to complex coefficient systems. Their main application is to stability theory and the proof of Routh's criterion. Dudnikov [4] also used this form for identifying system coefficients.

2) *The Cauer First Form* [3]:

$$\frac{g_2(s)}{g_1(s)} = \cfrac{1}{a_1 s + \cfrac{1}{a_2 + \cfrac{1}{a_3 s + \cfrac{1}{\ddots}}}} \tag{3}$$

It is well known that this form is used to synthesize ladder network driving-point impedances. We are, however, particularly interested in the third form.

3) *The Cauer Second Form:*

$$\frac{g_2(s)}{g_1(s)} = \cfrac{1}{h_1 + \cfrac{1}{\frac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\frac{h_4}{s} + \cfrac{1}{\ddots}}}}} \tag{4}$$

This form can be obtained from (1) by first arranging the two polynomials into the ascending order:

$$\frac{g_2(s)}{g_1(s)} = \frac{A_{21} + A_{22}s + A_{23}s^2 + \cdots + A_{2,n}s^{n-1}}{A_{11} + A_{12}s + A_{13}s^2 + \cdots + A_{1,(n+1)}s^n} \tag{1a}$$

and then expanding it into (4).

The second Cauer form on which we will concentrate is not only important in RC network synthesis but also plays a significant role in control systems analysis [5].

Expansion by Routh's Algorithm

It is known that Routh's array can be expressed by the following double subscript notation [6]:

$$\begin{array}{llll} A_{11} & A_{12} & A_{13} & \cdots \\ A_{21} & A_{22} & A_{23} & \cdots \\ A_{31} & A_{32} & \cdots & \\ A_{41} & & & \\ \cdots & & & \end{array} \tag{5}$$

and the elements of the third, fourth, and subsequent rows can be evaluated from the following relation [6]:

$$A_{j,k} = A_{j-2,k+1} - \frac{A_{j-2,1}A_{j-1,k+1}}{A_{j-1,1}}$$

$$j = 3, 4, \cdots, n+1; \quad k = 1, 2, \cdots. \tag{6}$$

β.6

Equation (6) is also known as the Routh algorithm.

Now, performing long division on (1a), we have

$$\frac{g_2(s)}{g_1(s)} = \cfrac{1}{\cfrac{A_{11}}{A_{21}} + \cfrac{\left(\frac{A_{21}A_{12} - A_{11}A_{22}}{A_{21}}\right)s + \left(\frac{A_{21}A_{13} - A_{11}A_{23}}{A_{21}}\right)s^2 + \cdots}{A_{21} + A_{22}s + A_{23}s^2 + \cdots + A_{2,n}s^{n-1}}} \tag{7}$$

in which

$$\left(\frac{A_{21}A_{12} - A_{11}A_{22}}{A_{21}}\right)$$

and

$$\left(\frac{A_{21}A_{13} - A_{11}A_{23}}{A_{21}}\right)$$

can be written as A_{31} and A_{32}, respectively, where A_{31} and A_{32} are defined in (6).

Therefore, we have

$$\frac{g_2(s)}{g_1(s)} = \cfrac{1}{\cfrac{A_{11}}{A_{21}} + \cfrac{A_{31}s + A_{32}s^2 + \cdots}{A_{21} + A_{22}s + A_{23}s^2 + \cdots}}.$$

Dividing again, we obtain

$$\cfrac{1}{\cfrac{A_{11}}{A_{21}} + \cfrac{s}{\cfrac{A_{21}}{A_{31}} + \cfrac{\left(\frac{A_{22}A_{31} - A_{32}A_{21}}{A_{31}}\right)s + \cdots}{A_{31} + A_{32}s + \cdots}}}$$

or

$$\cfrac{1}{\cfrac{A_{11}}{A_{21}} + \cfrac{s}{\cfrac{A_{21}}{A_{31}} + \cfrac{A_{41}s + A_{42}s^2 + \cdots}{A_{31} + A_{32}s + \cdots}}}. \tag{8}$$

Finally, we have the expansion

$$\cfrac{1}{\cfrac{A_{11}}{A_{21}} + \cfrac{s}{\cfrac{A_{21}}{A_{31}} + \cfrac{s}{\cfrac{A_{31}}{A_{41}} + \cfrac{s}{\ddots}}}} \tag{9}$$

This can be written in the form:

$$\cfrac{1}{h_1 + \cfrac{1}{\cfrac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\cfrac{h_4}{s} + \cfrac{1}{\ddots}}}}} \tag{10}$$

where

$$h_p = \frac{A_{p,1}}{A_{(p+1),1}}, \qquad p = 1, 2, \cdots, 2n, \quad h_p \neq 0. \tag{11}$$

Clearly, the elements of continued fraction (10), h_p, can be obtained by the quotients of members of the first column in Routh's array [7].

As an illustration, consider the following transfer function.

$$F(s) = \frac{360 + 171s + 10s^2}{720 + 702s + 71s^2 + s^3}. \tag{12}$$

A continued fraction expansion like (10) is desired. Write the first and second rows of the Routh's array by copying the coefficients of the denominator and the numerator, respectively, and then use (6) to generate the lower rows:

720	702	71	1
360	171	10	
360	51	1	
120	9		
24	1		
4			
1			

(13)

Values h_p are then found from (13) by taking the ratio of the neighboring terms of the first column:

$$h_1 = \frac{720}{360} = 2 \left\langle \begin{matrix} 720 & 702 & 71 & 1 \end{matrix} \right.$$

$$h_2 = \frac{360}{360} = 1 \left\langle \begin{matrix} 360 & 171 & 10 \end{matrix} \right.$$

$$h_3 = \frac{360}{120} = 3 \left\langle \begin{matrix} 360 & 51 & 1 \end{matrix} \right.$$

$$h_4 = \frac{120}{24} = 5 \left\langle \begin{matrix} 120 & 9 \end{matrix} \right.$$

$$h_5 = \frac{24}{4} = 6 \left\langle \begin{matrix} 24 & 1 \end{matrix} \right.$$

$$h_6 = \frac{4}{1} = 4 \left\langle \begin{matrix} 4 \\ 1 \end{matrix} \right.$$

p. 7

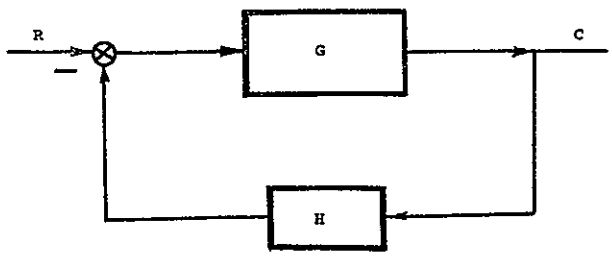


Fig. 1. A typical feedback system.

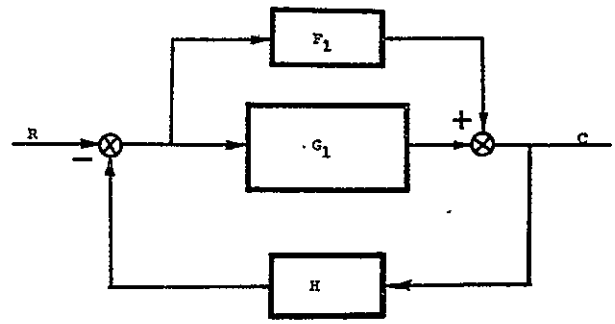


Fig. 2. Feedback and feedforward controls.

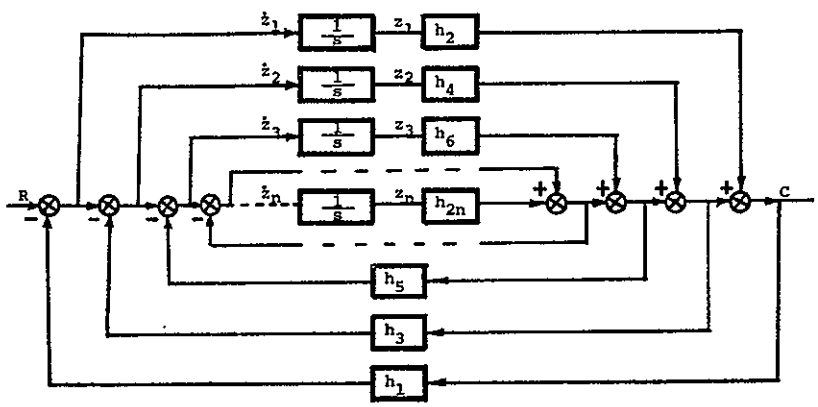


Fig. 3. Block diagram corresponds to continued fraction expansion.

from which the continued fraction is written immediately:

$$F(s) = \cfrac{1}{2 + \cfrac{1}{\frac{1}{s} + \cfrac{1}{3 + \cfrac{1}{\frac{5}{s} + \cfrac{1}{6 + \cfrac{1}{\frac{4}{s}}}}}}} \tag{14}$$

Block Diagram Representation

It is known that (10) in general or (14) in particular can be interpreted as the driving-point impedance function of an RC ladder network. However, we rather interpret it as the transfer function of a general feedback system.

Consider a typical feedback system shown in Fig. 1. The closed loop or overall transfer function is

$$\frac{C}{R} = \frac{G}{1 + GH}. \tag{15}$$

Dividing the numerator and denominator by G, we obtain

$$\frac{C}{R} = \frac{1}{H + \frac{1}{G}}. \tag{16}$$

Equation (16) can be considered as the simplest continued fraction. If we have a feedback system with a minor feedforward loop, as shown in Fig. 2, the corresponding overall transfer function is then

$$\frac{C}{R} = \frac{G_1 + F_1}{1 + (G_1 + F_1)H}. \tag{17}$$

This can be rewritten as a continued fraction:

$$\frac{C}{R} = \frac{1}{H + \frac{1}{F_1 + G_1}}. \tag{18}$$

If the subsystem G_1 is expanded again, we finally get the following general form:

$$\frac{C}{R} = \cfrac{1}{h_1 + \cfrac{1}{\frac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\ddots}}}}, \tag{19}$$

which is exactly (10). Therefore, continued fraction form (10) can always be interpreted as the block diagram shown in Fig. 3.

State-Space Formulation

In Fig. 3, after each integrator, if we assign a name as a state variable, the state matrix equation and the output equation can be easily written as

p. 8

$$[\dot{z}] = [H][z] + [D]r \tag{20}$$

$$c = [Q][z] \tag{21}$$

where

$$[H] = -\begin{bmatrix} h_2h_1 & h_4h_1 & h_6h_1 & \cdots & h_{2n}h_1 \\ h_2h_1 & h_4(h_1+h_3) & h_6(h_1+h_3) & \cdots & h_{2n}(h_1+h_3) \\ h_2h_1 & h_4(h_1+h_3) & h_6(h_1+h_3+h_5) & \cdots & h_{2n}(h_1+h_3+h_5) \\ \vdots & \vdots & \vdots & & \vdots \\ h_2h_1 & h_4(h_1+h_3) & h_6(h_1+h_3+h_5) & \cdots & h_{2n}(h_1+h_3+\cdots+h_{2n-1}) \end{bmatrix} \tag{22}$$

$$[D] = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{23}$$

and

$$[Q] = [h_2, h_4, h_6 \cdots h_{2n}]. \tag{24}$$

It is seen that the elements in the state matrix (22) are simple combinations of the quotients obtained from the continued fraction expression.

Next we would like to find the relationship between this state formulation and the phase variable form.

$$[\dot{x}] = [A][x] + [B]r \tag{25}$$

$$c = [C][x] \tag{26}$$

where $[x]$ is the phase variable vector, and

$$[A] = \begin{bmatrix} 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 1 & \cdot & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdot \\ & & & & 1 \\ -A_{11} & -A_{12} & -A_{13} & \cdot & -A_{1n} \end{bmatrix} \tag{27}$$

$$[B] = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \tag{28}$$

$$[C] = [A_{21}, A_{22}, \cdots, A_{2n}] \tag{29}$$

if the original transfer function is normalized as follows.

$$F(s) = \frac{A_{2,n}s^{n-1} + A_{2,n-1}s^{n-2} + \cdots + A_{22}s + A_{21}}{s^n + A_{1,n}s^{n-1} + A_{1,n-1}s^{n-2} + \cdots + A_{12}s + A_{11}} = \frac{A_{21} + A_{22}s + \cdots + A_{2,n-1}s^{n-2} + A_{2,n}s^{n-1}}{A_{11} + A_{12}s + \cdots + A_{1,n-1}s^{n-2} + A_{1,n}s^{n-1} + s^n}. \tag{30}$$

It can easily be proven that the two forms are related by the following linear transformation.

$$[z] = [P]\,[x] \tag{31}$$

where

$$[P] = \begin{bmatrix} A_{31} & A_{32} & A_{33} & \cdot & \cdot & A_{3,n-1} & A_{3,n} \\ 0 & A_{51} & A_{52} & \cdot & \cdot & A_{5,n-2} & A_{5,n-1} \\ 0 & 0 & A_{71} & A_{72} & \cdot & \cdot & A_{7,n-2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & A_{2n-1,1} & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & A_{2n+1,1} \end{bmatrix}. \tag{32}$$

It is believed that the construction of (32) is new. The proof of the similarity transformation (31) can be done by using the Krylov transform matrix [8] with an input constraint.

The matrix $[P]$ is an $n \times n$ upper triangular matrix. The elements in the triangle are copied directly from the elements of the Routh array. The elements of the third row in Routh's array is that of the first row of the $[P]$ matrix. In general, the $(2n+1)$th row of Routh's array is the nth row of the $[P]$ matrix.

Example for Obtaining the $[H]$ Matrix

Consider (12) again:

$$\frac{C(s)}{R(s)} = \frac{360 + 171s + 10s^2}{720 + 702s + 71s^2 + s^3}. \tag{12}$$

The second Cauer state form is required. The phase variable form of (12) is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -720 & -702 & -71 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} r$$

$$C = [360, 171, 10] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

From these state and output equations, we formulate the Routh array:

1.9

$$
\begin{array}{cccc}
720 & 702 & 71 & 1 \\
360 & 171 & 10 & \\
\hline
\multicolumn{1}{|c}{360} & \multicolumn{1}{c}{51} & \multicolumn{1}{c|}{1} & \\
\hline
120 & 9 & & \\
\hline
\multicolumn{1}{|c}{24} & \multicolumn{1}{c|}{1} & & \\
\hline
4 & & & \\
\hline
\multicolumn{1}{|c|}{1} & & &
\end{array}
$$

and then use the third, fifth, and seventh row to form the $[P]$ matrix.

$$[P] = \begin{bmatrix} 360 & 51 & 1 \\ 0 & 24 & 1 \\ 0 & 0 & 1 \end{bmatrix}. \tag{33}$$

Substituting this $[P]$ matrix into (31) and then into (25) and (26), we obtain

$$[\dot{z}] = [P][A][P]^{-1}[z] + [P][B]r \tag{34}$$

$$= [H][z] + [D]r \tag{34a}$$

$$c = [C][P]^{-1}[z] \tag{35}$$

$$= [Q][z]. \tag{35a}$$

For this problem, numerically we have

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = -\begin{bmatrix} 2 & 10 & 8 \\ 2 & 25 & 20 \\ 2 & 25 & 44 \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} r \tag{34b}$$

$$c = [1, 5, 4]\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \tag{35b}$$

where

$$[H] = \begin{bmatrix} -2 & -10 & -8 \\ -2 & -25 & -20 \\ -2 & -25 & -44 \end{bmatrix}.$$

The Inversion Process

We restate the inversion problem here. If the elements of a continued fraction are given, or h_p are known, what is the corresponding rational function.

Based on the block diagram (Fig. 3), we can write the state equations immediately.

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \vdots \\ \dot{z}_n \end{bmatrix} = \begin{bmatrix} -h_2h_1 & -h_4h_1 & \cdots & -h_{2n}h_1 \\ -h_2h_1 & -h_4(h_1+h_3) & \cdots & -h_{2n}(h_1+h_3) \\ \vdots & \vdots & & \vdots \\ -h_2h_1 & -h_4(h_1+h_3) & \cdots & -h_{2n}(h_1+h_3+\cdots+h_{2n-1}) \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} r. \tag{36}$$

Laplace transforming gives

$$s[Z(s)] - [Z(0)] = [H]\,[Z(s)] + [D]\,R(s). \tag{37}$$

The determinant of the system matrix will give the characteristic equation [9]

$$|s[I] - [H]| = 0. \tag{38}$$

The coefficients of this equation are the elements of the first row of the required Routh's array.

The next step is to let h_1 and h_2 equal to zero, and we have a reduced system whose state matrix is

$$[H]_3 = \begin{bmatrix} -h_4h_3 & \cdots & -h_{2n}h_3 \\ -h_4h_3 & \cdots & -h_{2n}(h_3+h_5) \\ \vdots & \vdots & \vdots \\ -h_4h_3 & \cdots & -h_{2n}(h_3+\cdots+h_{2n-1}) \end{bmatrix}. \tag{39}$$

The corresponding characteristic equation of this system is found by [9]

$$|s[I] - [H]_3| = 0, \tag{40}$$

which gives the elements of the third row of the Routh's array.

Following a similar reasoning, we find the fifth row, seventh row, or $(2n-1)$th row by evaluating [9]

$$|s[I] - [H]_5| = 0 \tag{41}$$

$$|s[I] - [H]_7| = 0, \tag{42}$$

respectively, where $[H]_5$, $[H]_7$, etc. are defined by $[H]_5 \mid h_1, h_2, h_3, h_4 \to 0$, $[H]_7 \mid h_1, h_2, h_3, h_4, h_5, h_6 \to 0$.

Once the values of the elements of the odd rows have been found, the $[P]$ matrix is determined.

The values of the even rows can be evaluated from the output equation (35).

Example for Inversion

A continued fraction is given as follows:

$$F(s) = \cfrac{1}{1 + \cfrac{1}{\frac{4}{s} + \cfrac{1}{2 + \cfrac{1}{\frac{3}{s} + \cfrac{1}{1 + \cfrac{1}{\frac{5}{s}}}}}}}. \tag{43}$$

Find the corresponding rational function.

First we form the $[H]$ matrix using (22):

p. 10

$$[H] = \begin{bmatrix} -4 & -3 & -5 \\ -4 & -9 & -15 \\ -4 & -9 & -20 \end{bmatrix} \tag{44}$$

from which the characteristic equation is evaluated by substituting (44) into (38):

$$|s[I] - [H]| = 120 + 129s + 33s^2 + s^3. \tag{45}$$

The third row is found by

$$[H]_3 = -\begin{bmatrix} h_4h_3 & h_6h_3 \\ h_4h_3 & h_6(h_3 + h_5) \end{bmatrix} = \begin{bmatrix} -6 & -10 \\ -6 & -15 \end{bmatrix}$$

and this gives

$$|s[I] - [H]_3| = 30 + 21s + s^2 = g_3(s). \tag{46}$$

Similarly, the fifth row is found as

$$[H]_5 = -[h_6h_5] = [-5] \tag{47}$$

and

$$|s[I] - [H]_5| = 5 + s = g_5(s).$$

The linear transformation matrix $[P]$ is then formed:

$$[P] = \begin{bmatrix} 30 & 21 & 1 \\ 0 & 5 & 1 \\ 0 & 0 & 1 \end{bmatrix}. \tag{48}$$

The elements of the second row, $g_2(s)$, are found from the output equation (21) and (31), or

$$[h_2, h_4, h_6][P] = [4, 3, 5]\begin{bmatrix} 30 & 21 & 1 \\ 0 & 5 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$
$$= [120, 99, 12]. \tag{49}$$

Therefore, the required rational function is obtained from (45) and (49) as

$$F(s) = \frac{120 + 99s + 12s^2}{120 + 129s + 33s^2 + s^3}.$$

Conclusions

Based on the state-space formulation, an algorithmic method for inverting a continued fraction to a rational fraction of two polynomials is established. This completes Wall's, Frank's, and Fryer's work on the application of Routh's algorithm to continued fractions. It is believed that the results are very useful in circuit theory and system analysis.

Two digital computer programs (one for expansion and one for inversion) have been written and can be obtained from the authors.

Acknowledgment

The authors wish to acknowledge the helpful and constructive comments of the reviewers.

References

[1] H. S. Wall, "Polynomials whose zeros have negative real parts," *Am. Math. Monthly*, vol. 52, pp. 308–322, 1945.
[2] E. Frank, "On the zeros of polynomial with complex coefficients," *Am. Math. Monthly*, vol. 52, pp. 144–157, 1946.
[3] H. S. Wall, *Analytic Theory of Continued Fraction*. Princeton, N. J.: Van Nostrand, 1948.
[4] E. E. Dudnikov, "Determination of transfer function coefficients of linear system from the initial portion of an experimentally obtained amplitude phase characteristic," *Automatic and Remote Control*, vol. 20, May 1959.
[5] C. F. Chen and L. S. Shieh, "A novel approach to linear model simplification," *1968 Joint Automatic Control Conf.* (Michigan), pp. 454–461.
[6] C. F. Chen and C. Hsu, "The determination of root loci using Routh's algorithm," *J. Franklin Inst.*, vol. 281, pp. 114–121, February 1966.
[7] W. D. Fryer, "Applications of Routh's algorithm to network-theory problems," *IRE Trans. Circuit Theory*, vol. CT-6, pp. 144–149, June 1959.
[8] F. R. Gantmacher, *The Theory of Matrices*, vol. 1. New York: Chelsea, 1959, pp. 202–214.
[9] R. H. Pennington, *Introductory Computer Methods and Numerical Analysis*. New York: MacMillan, 1965, pp. 355–359.
[10] C. F. Chen and I. J. Haas, *Elements of Control Systems Analysis: Classical and Modern Approaches*. Englewood Cliffs, N.J.: Prentice-Hall, 1968, pp. 303–305.

p. 11

Project

(A) Project Title: Analysis of Irrational Transfer Functions for Distributed-Parameter Systems

(B) Project Abstract:

This project is intended as an exposition of some approximation techniques on irrational transfer functions which are frequently encountered in control systems. Two algorithms for continued fraction expansion and inversion are established.

(C) Publication: IEEE Transactions on Aerospace and Electronics Systems, AES-5, No. 6, p. 967-973, 1969.

(D) Year: 1969

(E) Department: Electrical Engineering

(F) Student Name: L. S. Shieh

(G) Faculty Advisor: Professor C. F. Chen

Analysis of Irrational Transfer Functions for Distributed-Parameter Systems

LEANG-SAN SHIEH, Student Member, IEEE
CHIH-FAN CHEN, Senior Member, IEEE
University of Houston
Houston, Tex.

Abstract

This paper is intended as an exposition of some approximation techniques on irrational transfer functions which are frequently encountered in control systems. Two algorithms for continued fraction expansion and inversion are established.

Manuscript received April 1, 1968.
This work was supported in part by the National Aeronautics and Space Administration under Grant NGL-44-005-084.

C. F. Chen is now with the University of Cambridge, Cambridge, England.

I. Introduction

Control engineers often encounter a large number of control systems which involve distributed parameters. Thermal processes, hole diffusion of transistors and electromagnetic devices are typical examples [1]–[3]. Correspondingly, the mathematical descriptions in the Laplace transform domain for these elements usually contain the operator "S" under the radical sign. To analyze or synthesize an isolated irrational transfer function is not particularly difficult; however, when the element in question is represented by an irrational function in a closed loop system, the problem becomes very complicated.

Historically, the first irrational transfer function in engineering was noticed by Heaviside [4]. He observed that the impedance of an infinite RC cable is $1/\sqrt{S}$. Subsequently, many kinds of irrational functions have been derived from mathematical models. Some typical ones are listed as follows [5]:

1)	$\frac{1}{\sqrt{S}}$	$\frac{1}{\sqrt{\Pi t}}$
2)	$\frac{1}{S\sqrt{S+1}}$	$\operatorname{erf}(\sqrt{t})$
3)	$\frac{1}{\sqrt{S^2+1}}$	$J_0(t)$

If any one of these functions is contained in a closed loop system, the analysis is indeed quite tedious. For overcoming this difficulty, several methods have been developed. The historical developments will be reviewed first.

The Method Based on the Logarithmic Potential

Lerner [6] used a method for constructing the broadband impedance which is similar to potential analog approximation methods. He found that an infinite array of alternating poles and zeros placed along the negative real axis in the complex frequency plane produces a good approximation.

The Method Based on a Regular Newton's Process

The main contributors are Carlson and Halijak [7]. Their approximation is to predistort the algebraic expression $f(S)=S^n-a=0$. The resulting approximation in real variables has the unique property of preserving upper and lower approximations to the nth root of the real number "a" By using this regular Newton's process, they generate rational functions for the approximation.

The Method Based on Substitution

Kilomeitseva and Netushil [8] conceived a novel approach to this problem. They substituted $\sqrt{S}$ by p. An irrational transfer function of S then becomes a rational one of p. The regular Heaviside expansion is performed

4.13

on the new transfer function, and the inverse process is taken by using certain typical irrational function tables.

In reviewing the methods mentioned above, one finds that the first is for driving point impedance approximation only, while the second method is limited to fractional capacitors. The third method required predetermination of the roots and many necessary graphs for computing the behavior of the system Therefore, the applications are limited to very special cases and procedures involved are rather complicated

II. The Continued Fraction Approach

This paper attempts to use the continued fraction expansion to approximate the irrational function first, then truncate the unimportant parts under a reasonable error tolerance, and finally perform the inverse process to change the truncated continued fraction into a ratio of two polynomials

There are several problems involved in this approach.

1) Why should we use the continued fraction expansion? There are many forms of continued fraction expansions. Which one do we have to use?
2) Either the expansion or the inversion is tedious, if more terms are desired. What technique should we develop, such that the methods become practical. In other words, the method must be computer oriented.

According to Lerner's theory, we can answer the first question immediately. If the poles and zeros are alternately distributed, a good approximation can be achieved. The following continued fraction usually gives a satisfactory solution.

$$f(S) = b_1 + \cfrac{S}{b_2 + \cfrac{S}{b_3 + \cfrac{S}{b_4 + \cfrac{S}{\vdots}}}} \tag{1}$$

The accuracy of an approximation depends on where we truncate the function, of course. In general, if more terms are taken, a better result can be obtained. In view of the availability of high speed, large capacity digital computers, one can take as many elements as desired. However, two new problems arise:

1) expanding an irrational function becomes increasingly tedious;
2) the inversion process becomes very laborious.

These two fundamental problems must be solved. This paper will develop a new approach to solving these problems, one by one.

III. Continued Fraction Expansion

In control system studies, some particular irrational functions are usually of interest. The following examples are well known:

$$\begin{aligned} &1)\ \sqrt{S} \\ &2)\ \sqrt{S+1} \\ &3)\ \sqrt{S}+1 \\ &4)\ \sqrt{S^2+1}. \end{aligned} \tag{2}$$

These functions can be considered as special forms of $\sqrt{S+a}$. If we want to approximate this function by a continued fraction, a simpler technique can be applied.

Let

$$\sqrt{S+a} = B \tag{3}$$

or

$$S + a = B^2.$$

Adding B to both sides, we have

$$S + a + B = B(1 + B).$$

Rewriting gives

$$\begin{aligned} B &= \frac{S + a + B}{1 + B} \\ &= \frac{1 + B + S + (a-1)}{1 + B} \end{aligned} \tag{4}$$

or

$$\sqrt{S+a} = 1 + \frac{S + (a-1)}{1 + B}. \tag{5}$$

Continually substituting (4) into (5), we obtain

$$\begin{aligned} \sqrt{S+a} &= 1 + \cfrac{S + (a-1)}{1 + 1 + \cfrac{S + (a-1)}{1 + B}} \\ &= 1 + \cfrac{S + (a-1)}{2 + \cfrac{S + (a-1)}{2 + \cfrac{S + (a-1)}{\vdots}}} \end{aligned} \tag{6}$$

If $a=0$, (6) becomes

$$\sqrt{S} = 1 + \cfrac{S-1}{2 + \cfrac{S-1}{2 + \cfrac{S-1}{2 + \cfrac{S-1}{\vdots}}}} \tag{7}$$

If $a=1$, (6) is reduced to

$$\sqrt{S+1} = 1 + \cfrac{S}{2 + \cfrac{S}{2 + \cfrac{S}{\vdots}}} \tag{8}$$

p. 14

The third function of (2) is easily seen to be

$$\sqrt{S}+1 = 2 + \cfrac{S-1}{2+\cfrac{S-1}{2+\cfrac{S-1}{\ddots}}} \tag{9}$$

Replacing S by S^2 in (6) and letting $a=1$, we have

$$\sqrt{S^2+1} = 1 + \cfrac{S^2}{2+\cfrac{S^2}{2+\cfrac{S^2}{\ddots}}} \tag{10}$$

It is seen that to expand an irrational function into a standard continued fraction form (1) is relatively easy and the technique is simpler than the existing methods, for example, Auslander's differentiation method [9].

Comparing (7), (8), and (10) with (1), we see that $b_1=1$, and $b_2=b_3=b_4 \cdots =2$. This regularity helps in developing an algorithm for the inversion process.

IV. Continued Fraction Inversion

Inverting the continued fraction (1) into a ratio of two polynomials, we would like to take the advantage of the fact that $b_1=1$ and $b_2=b_3=b_4 \cdots =2$. We modify (1) into the following form by adding a unit to each side and taking the reciprocals:

$$\frac{1}{f(S)+1} = \cfrac{1}{b+\cfrac{S}{b+\cfrac{S}{b+\cfrac{S}{\ddots}}}} \tag{11}$$

$b=2$ in this case.

If we truncate (11) and only keep two b's, we have

$$\frac{1}{f(S)+1} = \frac{1}{b+(S/b)} \tag{12}$$

$$= \frac{b}{b^2+S}. \tag{12a}$$

Keeping three b's gives

$$\frac{1}{f(S)+1} \cong \cfrac{1}{b+\cfrac{S}{b+(S/b)}} \tag{13}$$

$$= \frac{b^2+S}{b^3+2bS}. \tag{13a}$$

In general, the function $1/[f(S)+1]$ can be approximated by a ratio of two polynomials

$$\frac{1}{f(S)+1} = \frac{P(b,S)}{Q(b,S)}$$

where P and Q are polynomials of S, and the coefficients are in terms of b.

If the order of the continued fraction is high, particular consideration should be given in finding the corresponding polynomials, which means solving the inverse problem.

For performing the inversion process, Table I is established from which we can read the coefficients of the two polynomials directly. The table is constructed in the following way.

1) the elements in the "0" row are $(S+a-1)^1$, $(S+a-1)^2$, $(S+a-1)^3$, $\cdots$ as indices, or $B(0, k) = (S+a-1)^k$, $k=1,2, \cdots n$.
2) the elements in the "0" column are b^1, b^2, $\cdots$ as indices, or $B(j, 0)=b^j$, $j=1, 2, \cdots 2n$.
3) the element $B(j, k)=\partial[B(j-1, k-1)]/(k)\partial b$, $j=2, 3, \cdots 2n$, $k=1,2, \cdots n$, $j>k$.
4) the element $B(j, k)=0$, $j=1, 2, \cdots 2n$, $k=1,2, \cdots n$, $j\leq k$ where $B(j, k)$ is an element at j row and K column.

The result obtained after the differentiations have been performed is shown in Table II.

The table is ready to be used as an aid in the inversion process. We take the following example for an illustration: find several rational transfer function approximations for the irrational function $\sqrt{S+a}$.

1) Expand $\sqrt{S+a}$ into continued fraction (6) or

$$\sqrt{S+a} = 1 + \cfrac{S+(a-1)}{2+\cfrac{S+(a-1)}{2+\cfrac{S+(a-1)}{\ddots}}} \tag{14}$$

2) Because the first quotient of (14) is 1, instead of 2, modify (14) into the standard form in order to use the table:

$$\frac{1}{\sqrt{S+a}+1} = \cfrac{1}{2+\cfrac{S+(a-1)}{2+\cfrac{S+(a-1)}{2+\cfrac{S+(a-1)}{\ddots}}}} \tag{15}$$

3) Truncate (15) by keeping two quotients,

$$\frac{1}{\sqrt{S+a}+1} \cong \frac{1}{2+[S+(a-1)/2]} \tag{16a}$$

$$= \frac{2}{4+[S+(a-1)]}. \tag{16b}$$

1.15

TABLE I

$j \backslash k$	$(s+a-1)^1$	$(S+a-1)^2$	$(s+a-1)^3$	$(S+a-1)^4$	$\cdots$
b^1	0	0	0	0	$\cdots$
b^2	$\frac{\partial(b^1)}{\partial b}$	0	0	0	$\cdots$
b^3	$\frac{\partial(b^2)}{\partial b}$	$\frac{\partial^2(b^1)}{2!\partial b^2}$	0	0	$\cdots$
b^4	$\frac{\partial(b^3)}{\partial b}$	$\frac{\partial^2(b^2)}{2!b^2}$	$\frac{\partial^3(b^1)}{3!\partial b^3}$	0	$\cdots$
b^5	$\frac{\partial(b^4)}{\partial b}$	$\frac{\partial^2(b^3)}{2!\partial b^2}$	$\frac{\partial^3(b^2)}{3!\partial b^3}$	$\frac{\partial^4(b^1)}{4!\partial b^4}$	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

TABLE II

$j \backslash k$	$(S+a-1)^1$	$(S+a-1)^2$	$(S+a-1)^3$	$(S+a-1)^4$	$\cdots$
b^1	0	0	0	0	$\cdots$
b^2	1	0	0	0	$\cdots$
b^3	$2b$	0	0	0	$\cdots$
b^4	$3b^2$	1	0	0	$\cdots$
b^5	$4b^3$	$3b$	0	0	$\cdots$
b^6	$5b^4$	$6b^2$	1	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$

When we use the table, the coefficients of (16b) can be read directly from row 1 and 2, or

$j \backslash k$	$(S+a-1)^1$	
b^1	0	$\leftarrow$ 1st row
b^2	1	$\leftarrow$ 2nd row

Therefore, where $b=2$ we have

$$\frac{b}{b^2+[S+(a-1)]}=\frac{2}{4+[S+(a-1)]}.$$

4) If we truncate (15) by keeping six quotients,

$$\frac{1}{\sqrt{S+a}+1}$$

ORIGINAL PAGE IS OF POOR QUALITY

$$\cong\cfrac{1}{2+\cfrac{S+(a-1)}{2+\cfrac{S+(a-1)}{2+\cfrac{S+(a-1)}{2+\cfrac{S+(a-1)}{2+\cfrac{S+(a-1)}{2}}}}}} \tag{17}$$

Inversing (17) gives

p.16

$$\frac{1}{\sqrt{S+a}+1} \cong \frac{b^5+4b^3(S+a-1)^1+3b(S+a-1)^2}{b^6+5b^4(S+a-1)^1+6b^2(S+a-1)^2+(s+a-1)^3}. \tag{17a}$$

By substituting $b=2$ into (17a) we have

$$\frac{1}{\sqrt{S+a}+1} \cong \frac{32+32(S+a-1)+6(S+a-1)^2}{64+80(S+a-1)+24(S+a-1)^2+(s+a-1)^3}. \tag{17b}$$

Then $\sqrt{S+a}$ can be easily obtained by using (17b):

$$\sqrt{S+a} \cong \frac{32+48(S+a-1)+18(S+a-1)^2+(S+a-1)^3}{32+32(S+a-1)+6(S+a-1)^2}. \tag{18}$$

In general, if j quotients are taken, then

$$\sqrt{S+a} \cong \frac{\text{polynomial at } j\text{th row} - \text{polynomial at } (j-1)\text{th row}}{\text{polynomial at } (j-1)\text{th row}} \quad = 2, 3, \cdots 2n. \tag{19}$$

V. Applications

We consider the error function $1/(S\sqrt{S+1})$ as the first application. From (8) we have the expression

$$\frac{1}{S\sqrt{S+1}} = \frac{1}{1+\cfrac{S}{2+\cfrac{S}{2+\cfrac{S}{\ddots}}}} \cdot \frac{1}{S} \tag{20}$$

If Table II is used, we simply substitute $a=1$, $b=2$ into Table II and obtain

$$\frac{1}{S\sqrt{S+1}} \cong \frac{\text{polynomial at } (j-1)\text{th row}}{\text{polynomial at } (j)\text{th row} - \text{polynomial at } (j-1)\text{th row}} \cdot \frac{1}{S}.$$

when j quotients are taken.

Let $j=2$ as a special case in the approximation; then

$$\frac{1}{S\sqrt{S+1}} \cong \frac{2}{S+2} \cdot \frac{1}{S} = \frac{2}{S^2+2s}. \tag{21}$$

The time domain [10] curve of (21) is shown in Fig 1 (the curve marked by $n=2$). We see that even though $j=2$, it is a very good approximation to the original curve.

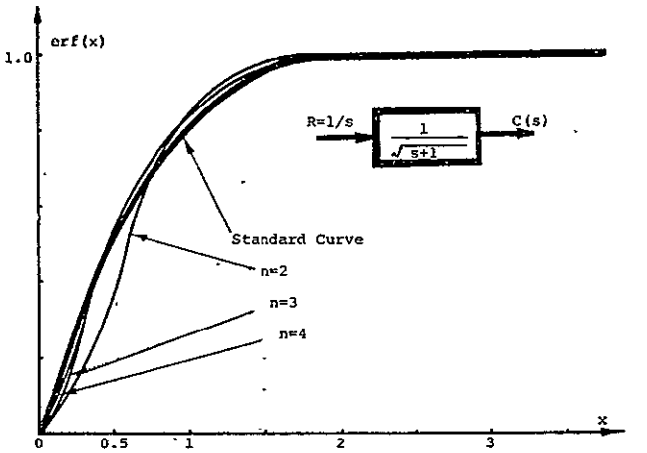


Fig. 1. Response curve of $1/(s\sqrt{s+1})$ (even number quotients being taken).

If $j=4$, we have

$$\frac{1}{S\sqrt{S+1}} \cong \frac{4S+8}{S^2+8S+8} \cdot \frac{1}{S} \cong \frac{4S+8}{S^3+8S^2+8S}.$$

If $j=6$, the result is

$$\frac{1}{S\sqrt{S+1}} \cong \frac{6S^2+32S+32}{S^3+18S^2+48S+32} \cdot \frac{1}{S} \cong \frac{6S^2+32S+32}{S^4+18S^3+48S^2+32S}.$$

The corresponding time curves are indicated by $n=3$ and $n=4$, respectively. Of course, we obtain a better result if higher quotients are taken. The comparison of data is shown in Table III.

Fig. 2 shows the different approximation when an odd number of quotients are taken. In other words, if $i=3, 5, 7, 9, 11$, the corresponding time curves are indicated by $n=2, 3, 4, 5, 6 \cdots$.

For an automatic control system which, in addition to components with lumped parameters, contains one or more elements with distributed parameters, the transfer function is written in the following form:

$$W_3 = \frac{W_1 W_2}{1+W_1 W_2} \tag{22}$$

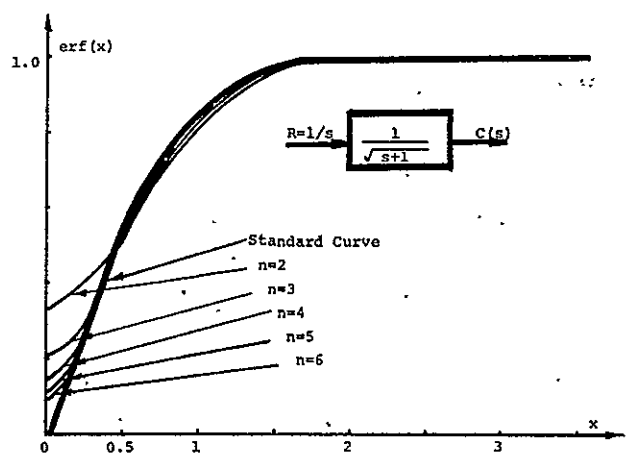
where W_1 is a transfer function containing distributed

p. 17

TABLE III

x	$t=x^2$	Exact erf $(\sqrt{t})$	Approximations $n=2$	$n=3$	$n=4$	$n=5$	$n=6$
0	0	0	0	0	0	0	0
0.02	0.0004	0.0226	0.0008	0.0016	0.0024	0.0032	0.0039
0.04	0.0016	0.0451	0.0032	0.0064	0.0095	0.0126	0.0156
0.06	0.0036	0.0676	0.0072	0.0142	0.0211	0.0277	0.0339
0.08	0.0064	0.0901	0.0127	0.0251	0.0369	0.0478	0.0576
0.1	0.01	0.1125	0.0198	0.0388	0.0564	0.0719	0.0851
0.2	0.04	0.2227	0.0769	0.1424	0.1895	0.2172	0.2295
0.3	0.09	0.3286	0.1647	0.2796	0.3319	0.3426	0.3377
0.4	0.16	0.4284	0.2739	0.4178	0.4466	0.4374	0.4292
0.5	0.25	0.5205	0.3935	0.5363	0.5349	0.5215	0.5185
0.6	0.36	0.6039	0.5132	0.6293	0.6096	0.6016	0.6024
0.7	0.49	0.6778	0.6247	0.7007	0.6775	0.6756	0.6775
0.8	0.64	0.7421	0.7218	0.7575	0.7394	0.7409	0.7423
0.9	0.81	0.7969	0.8021	0.8045	0.7941	0.7967	0.7972
1.0	1.0	0.8427	0.8647	0.8445	0.8408	0.8429	0.8428
1.2	1.44	0.9103	0.9439	0.9074	0.9101	0.9105	0.9103
1.3	1.69	0.9340	0.9659	0.9310	0.9341	0.9341	0.9339
1.4	1.96	0.9523	0.9802	0.9497	0.9526	0.9523	0.9523
1.5	2.25	0.9661	0.9889	0.9642	0.9664	0.9661	0.9661
2	4	0.9953	0.9996	0.9954	0.9953	0.9953	0.9953
3	9	0.9999	1.0000	0.9999	0.9999	0.9999	0.9999

Fig. 2. Response curve of $1/(s\sqrt{s+1})$ (odd number quotients being taken).



parameters and W_2 is the element involving lumped parameters.

Now we study the given feedback system of (22). The transfer functions are defined as

$$W_1 = \frac{100}{1+0.63\sqrt{S}};$$

$$W_2 = \frac{1}{1+S}.$$

Then

$$W_3 = \frac{\dfrac{100}{(1+S)\left[1+0.63\dfrac{g_2(S)}{g_1(S)}\right]}}{1+\dfrac{100}{(1+S)\left[1+0.63\dfrac{g_2(S)}{g_1(S)}\right]}}$$

$$= \frac{100g_1(S)}{(101+S)g_1(S)+0.63(1+S)g_2(S)}.$$

and when a unit step input is applied, we easily obtain the output

$$C(S) = \frac{100S^2+1000S+500}{4.15S^4+120.45S^3+1021.93S^2+505.63S}.$$

If

$$\sqrt{S} = \frac{g_2(S)}{g_1(S)} \cong \frac{5S^2+10S+1}{S^2+10S+5}$$

The corresponding time curve is shown in Fig. 3 by $n=4$ If a better approximation is used, or letting

$$\sqrt{S} \cong \frac{g_2(S)}{g_1(S)} \cong \frac{9S^4+84S^3+126S^2+36S+1}{S^4+36S^3+126S^2+84S+9},$$

the corresponding output will be

$$C(S) \cong \frac{100S^4+3600S^3+12600S^2+8400S+900}{6.67S^6+195.59S^5+3894.3S^4+12912.06S^3+8516.31S^2+909.63S}.$$

Assume that we can approximate $\sqrt{S}$ by

$$\sqrt{S} = \frac{g_2(S)}{g_1(S)}.$$

VI. Conclusion

A method for approximating irrational transfer functions by rational ones through continued fraction expansions and inversions is established. Compared with the

p.18

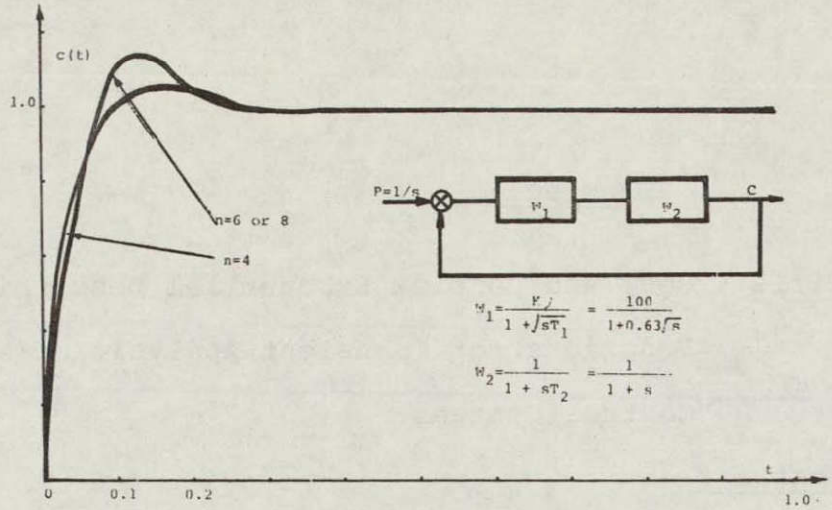


Fig. 3. Response curves of a closed loop system.

ORIGINAL PAGE IS OF POOR QUALITY

existing techniques, the new approach is algorithmic in nature and digital computer oriented. Two examples have been included for illustrating the accuracy and power of the method.

References

[1] J. G. Truxal, *Automatic Feedback Control System Synthesis.* New York: McGraw-Hill, 1955, pp. 546–558.

[2] D. P. Campbell, *Process Dynamics.* New York: Wiley, 1958, pp. 104–112.

[3] E. V. Bohn, *The Transform Analysis of Linear Systems.* Reading, Mass.: Addison-Wesley, 1963, pp. 281–303.

[4] O. Heaviside, *Electromagnetic Theory.* New York: Dover, 1950, pp. 128–129.

[5] J. A. Aseltine, *Transform Methods in Linear System Analysis.* New York: McGraw-Hill, 1958, p. 289.

[6] R. M. Lerner, "The design of a constant-angle or power-law magnitude impedance," *IEEE Trans. Circuit Theory*, vol. CT-10, pp. 98–107, March 1963.

[7] G. E. Carlson and C. A. Halijak, "Approximation of fractional capacitors $(1/S)^{1/n}$ by a regular Newton process," *IEEE Trans. Circuit Theory*, vol. CT-11, pp. 210–213, June 1964.

[8] M. B. Kilomeitseva and A. V. Netushil, "Transients in automatic control systems with irrational transfer functions," *Automation and Remote Control* (English transl. of *Automat. i Telemekh.*), vol. 26, pp. 359–364, February 1965.

[9] G. Auslender, "On the expansion of a function in series and continued fractions," *Zh. Vych. Mat.*, vol. 3, no. 3, pp. 565–568, 1963.

[10] C. F. Chen and L. S. Shieh, "Generalization and computerization of the Heaviside expansion for performing the inverse Laplace transform of high order systems," *IEEE Region 3 Conv. Rec.*, 1967, pp. 285–294.

Leang-San Shieh (S'68) was born in Taiwan, China, on January 10, 1934. He received the B.S.E.E. degree from the National Taiwan University in 1958, and the M.S.E.E. degree from the University of Houston, Houston, Tex., in 1968, where he was a Graduate Assistant. He is presently a Doctoral Fellow, working toward the Ph.D. degree at the University of Houston.

From 1958 to 1960 he served in the Chinese Navy. From 1960 to 1961 he was a Design Engineer at the Taiwan Fluorescent Lamp Company, and from 1961 to 1965 he was a System Engineer at the Taiwan Power Company.

Chih-Fan Chen (SM'61) was born in China. He received the B.S.E.E. degree from Peiyang University in 1948, the M.S.E.E. degree from the University of Pennsylvania, Philadelphia, in 1957, and the Honorary L.L.D. degree from Lewis College, Lockport, Ill., in 1965.

From 1966 to 1969 he was a Professor of Electrical Engineering at the University of Houston, Houston, Tex. He is now with the University of Cambridge, England. He is the author of more than 40 technical papers on automatic control theory and computing sciences, and is co-author of the book, *Elements of Control Systems Analysis: Classical and Modern Approaches* (Prentice-Hall, 1968).

p. 19

Project

(A) Project Title: Real and Complex Exponential Describing Functions for Transient Analysis of Nonlinear Control Systems

(B) Project Abstract:

Describing functions as an analysis tool for studying the transient response of a large class of nonlinear feedback systems are developed. The two classes of describing functions developed are identified as real exponential and complex exponential describing functions and are intended for analysis of higher order systems. The real exponential describing functions are employed when a system is overdamped and the complex exponential describing functions are associated with the analysis of underdamped systems.

Signals in $L_2(-\infty,t]$, a subspace of the space of square integrable signals defined on $(-\infty,t]$, are approximated by the sum of n signals in $L_2^{1,m}(-\infty,t]$, one-dimensional subspaces of $L_2(-\infty,t]$ spanned by the mth function from a set of time reversed orthogonalized real or complex exponential functions, where $m = 1,\ldots,n$. A system mapping $L_2(-\infty,5]$ into itself is associated with a system mapping $L_2^{1,m}(-\infty,t]$ into itself; the latter system is characterized by a gain - real or complex exponential describing function. The approximate response is found by adding n approximation components resulting from multiple one-dimensional mappings where n

is the order of the nonlinear differential equation describing the system's behavior.

The contraction-mapping fixed-point theorem is also used to determine conditions for the existence of a solution prior to the use of the exponential describing functions for obtaining an approximate response.

(C) Publication: Ph.D. Dissertation in Electrical Engineering

(D) Year: 1969

(E) Department: Electrical Engineering

(F) Student Name: Aristides D. Charalampous

(G) Faculty Advisor: Professor C. F. Chen

Project

(A) Project Title: A New Approach to Matrix Heaviside Expansion

(B) Project Abstract:

This paper presents a new procedure for the general matrix Heaviside expansion. The transfer matrix to be expanded can have many eigenvalues, each of which can have any multiplicity. The derivations of the formulas are based on Krylov's matrix and Vandermonde's transformation, and take advantage of using the particular nature of the inverse Jordan matrix. The results are extremely simple.

(C) Publication: International Journal of Control, 11, 431 (1970)

(D) Year: 1970

(E) Department: Electrical Engineering

(F) Student Name: R. E. Yates

(G) Student Advisor: Professor C. F. Chen

INT. J. CONTROL, 1970, VOL. 11, NO. 3, 431–448

A new approach to matrix Heaviside expansion†

C. F. CHEN‡

Engineering Department, University of Cambridge,
Cambridge, England

R. E. YATES

U.S. Army Missile Command, Redstone Arsenal, Alabama, and
University of Houston, Houston, Texas

[Received 14 January 1969]

This paper presents a new procedure for the general *matrix* Heaviside expansion. The transfer matrix to be expanded can have many eigenvalues, each of which can have any multiplicity. The derivations of the formulas are based on Krylov's matrix and Vandermonde's transformation, and take advantage of using the particular nature of the inverse Jordan matrix. The results are extremely simple.

1. Introduction

The general formulation and solution of state equations in dynamical systems analysis is based on the following two postulates:

(1) The method of analysis and form for an n-degree system of any complexity is the same as for a 1-degree system of the same general type, provided each quantity is replaced by an appropriate matrix.

(2) A matrix equation true in one reference frame remains invariant in form upon transformation to a new reference frame of the same type.

These two generalization postulates (Kron 1939, Bewley 1961) in fact, were introduced by Kron for the tensor method more than 30 years ago.

Heaviside's expansion (Van Valkenburg 1964) is a fundamental operation in transfer function analysis. It should be extended and applied to a transfer matrix without any theoretical difficulty. With a proper interpretation, Sylvester's expansion formula (Gantmacher 1959), can be considered as the matrix Heaviside expansion for the distinct eigenvalue case. For the general case, Chen and Parker (1966) have generalized the Heaviside expansion by using complex algebra, which can be considered as an extension of Goldstone's technique (Kuo 1966) from the scalar function to the matrix function. However, it is well known that Goldstone's approach is not very suitable for digital computation. Recently, Rao and Ahmed (1968) developed a recursive formula for solving the multiple eigenvalue problem. Unfortunately, their recursive formula is only good for the transfer matrix with *one* high-order eigenvalue. Therefore, there is a lack of a computer method for solving the matrix Heaviside expansion of a transfer matrix with several multiple eigenvalues. In other

† Communicated by Professor Chen.

‡ Permanent address: Electrical Engineering Department, University of Houston, Houston, Texas.

1.23

words, the matrix Heaviside expansion problem in general has not yet been completely solved.

This paper, based on Kron's two generalization postulates, develops a procedure for the general matrix Heaviside expansion. The transfer matrix to be expanded can have many eigenvalues, each of which can have any multiplicity. The derivations of the useful formulas are straightforward and the results are particularly simple and especially suitable for digital computation.

2. Derivation of transfer matrix

Transfer matrices come from many formulations in systems analysis. In the state space approach, usually, a transfer matrix comes from the Laplace transform of a transition matrix.

Consider a set of state equations in the matrix form:

$$[\dot{x}] = [A][x], \tag{1}$$

where $[A]$ is an $n \times n$ constant matrix. Laplace transform (1) and then solve for $[X(s)]$:

$$\begin{aligned}[X(s)] &= [sI - A]^{-1}[x(0)] \\ &= [\Phi_x(s)][x(0)],\end{aligned} \tag{2}$$

where $[\Phi_x(s)]$ is an $n \times n$ matrix, each element of which is a transfer function the ratio of two polynomials of s. The inverse Laplace transform of $[\Phi_x(s)]$ or $[\phi_x(t)]$ is usually called the transition matrix of (1).

To take the inverse Laplace transform of this $n \times n$ transfer function is very laborious. We would like to develop a new approach to solve the matrix Heaviside problem. The inverse transform problem then can be solved readily. We will use $[\Phi_x(s)]$ as a vehicle to develop the method of expansion; of course, the matrix Heaviside expansion technique can be applied to other transfer matrices as well. It is readily seen that

$$\begin{aligned}[\Phi_x(s)] &= [sI - A]^{-1} \\ &= \frac{\operatorname{Adj}(sI - A)}{\det(sI - A)},\end{aligned} \tag{3}$$

where Adj means the adjoint matrix, and

$$\det(sI - A) = s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_{n-1} s + a_n. \tag{4}$$

If the characteristic equation of $[\Phi_x(s)]$ involves distinct roots, only, the regular Heaviside expansion technique can be applied:

$$[\Phi_x(s)] = \frac{\operatorname{Adj}(sI - A)}{\det(sI - A)} = \sum_{j=1}^{n} \frac{[k_j]}{(s - \lambda_j)}, \tag{5}$$

where

$$[k_j] = \left[\frac{\operatorname{Adj}(sI - A)}{\det(sI - A)}(s - \lambda_j)\right]\Bigg|_{s \to \lambda_j}. \tag{6}$$

Hence

$$[\phi_x(t)] = \sum_{j=1}^{n} [k_j] \exp(\lambda_j t).$$

This extension is clearly explained in Chen and Parker's (1966) paper.

When multiple roots are involved in the characteristic equation, the regular Heaviside function differentiation technique can be extended; however, it is very cumbersome. We will use two similarity transformations to find the equivalent primitive systems in the Kron sense first and then evaluate the transfer matrix by expansion.

Our problem, therefore, is to expand the transfer matrix, $[\Phi_x(s)]$, into partial fraction matrix; or let

$$[\Phi_x(s)] = \sum_1^k \frac{[K_k]}{(s-\lambda_1)^k} + \sum_1^q \frac{[Q_q]}{(s-\lambda_2)^q} + \dots$$
$$= \frac{\text{Adj}\,[sI-A]}{(s-\lambda_1)^k(s-\lambda_2)^q\dots}, \tag{7}$$

where $[K_k]$ and $[Q_q]$ are $n \times n$ constant matrices which are to be determined.

3. Krylov's transformation

The first similaity transformation we want to perform on (1) is Krylov's transformation.

Krylov's transformation is a particular matrix which can transform a general matrix $[A]$ into a standard form. Kron's second postulate justifies this transformation.

Krylov's transformation is as follows:

$$[H][x] = [y], \tag{8}$$

where $[H]$ is formed by a set of chain vectors:

$$[H] = \begin{bmatrix} l[I] \\ l[A] \\ l[A]^2 \\ \vdots \\ l[A]^{n-1} \end{bmatrix} \tag{9}$$

in which l is any row vector such that the determinant of $[H]$ is not equal to zero.

Substituting (8) into (1), we obtain:

$$[\dot{y}] = [H][A][H]^{-1}[y] = [\alpha][y], \tag{10}$$

where $[\alpha]$ is a standard form or the companion matrix. Then (10) becomes:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{y}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 \\ -a_n & -a_{n-1} & \dots & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \tag{11}$$

in which the elements of the last row of $[\alpha]$ correspond to the characteristic equation coefficients of $[A]$ respectively.

The $[H]$ matrix was first proposed by Krylov (1931); Gantmacher (1959) modified it and used it to find the characteristic equation of a general matrix.

p. 25

The same matrix with an input constraint then was rediscovered and extended by Wonham and Johnson (1963, 1964).

Krylov's transformation not only simplifies the computation of the coefficients of the characteristic equation but also offers a link connecting a general system and a primitive system in the Kron sense.

It is very easy to prove that the $[H]$ matrix does indeed transform a general matrix into its companion form in the following manner.

Start with the Cayley–Hamilton theorem which is that any matrix $[A]$ satisfies its characteristic equation:

$$s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n = 0, \tag{12}$$

i.e.

$$[A]^n + a_1[A]^{n-1} + a_2[A]^{n-2} + \ldots + a_{n-1}[A] + a_n[I] = 0. \tag{13}$$

Both sides of (13) are multiplied by a row vector $[l]$:

$$l[A]^n + a_1 l[A]^{n-1} + a_2 l[A]^{n-2} + \ldots + a_{n-1} l[A] + a_n l[I] = 0. \tag{14}$$

Rearranging:

$$l[A]^n = -a_n l[I] - a_{n-1} l[A] - \ldots - a_1 l[A]^{n-1}. \tag{15}$$

We also write some trivial identities:

$$\left.\begin{aligned} l[I][A] &= l[A], \\ l[A][A] &= l[A]^2, \\ l[A]^2[A] &= l[A]^3, \\ &\vdots \\ l[A]^{n-2}[A] &= l[A]^{n-1}. \end{aligned}\right\} \tag{16}$$

Writing (16) and (15) together into a matrix form, we have:

$$\begin{bmatrix} l[I] \\ l[A] \\ \cdot \\ \cdot \\ l[A]^{n-1} \end{bmatrix} [A] = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 0 & 0 & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -a_n & -a_{n-1} & \cdot & \cdot & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} l[I] \\ l[A] \\ \cdot \\ \cdot \\ l[A]^{n-1} \end{bmatrix}, \tag{17}$$

which is:

$$[H][A][H]^{-1} = [\alpha]. \tag{18}$$

Therefore, we have proven the Krylov transformation.

4. Vandermonde's transformation

Once the system described by a general matrix differential eqn. (1) has been changed to the companion matrix form (10) by using Krylov's transformation matrix, we can diagonalize (10) immediately by the well-known Vandermonde transformation.

Let

$$[y] = [V][z], \tag{19}$$

p. 26

where

$$[V]=\begin{bmatrix} 1 & 1 & . & . & 1 \\ \lambda_1 & \lambda_2 & . & . & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & . & . & \lambda_n^2 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \lambda_1^{n-1} & \lambda_2^{n-1} & . & . & \lambda_n^{n-1} \end{bmatrix}, \tag{20}$$

if the system has only distinct roots.

Or, let

$$[y]=[W][z], \tag{21}$$

where

$$[W]=\begin{bmatrix} 1 & 0 & 0 & \vdots & 0 & 1 \\ \lambda_1 & \frac{\partial}{\partial\lambda}(\lambda)\Big|_{\lambda_1} & 0 & \vdots & 0 & \lambda\Big|_{\lambda_2} \\ \lambda_1^2 & \frac{\partial}{\partial\lambda}(\lambda^2)\Big|_{\lambda_1} & \frac{1}{2!}\frac{\partial^2}{\partial\lambda^2}(\lambda^2)\Big|_{\lambda_1} & \vdots & 0 & \lambda^2\Big|_{\lambda_2} \\ . & & . & \vdots & . & . \\ . & & . & \vdots & . & . \\ . & & . & \vdots & . & . \\ \lambda_1^{n-1} & \frac{\partial}{\partial\lambda}(\lambda^{n-1})\Big|_{\lambda_1} & \frac{1}{2!}\frac{\partial^{m-1}}{\partial\lambda^{m-1}}(\lambda^{n-1})\Big|_{\lambda_1} & \vdots & \frac{1}{(m-1)!}\frac{\partial^{(m-1)}}{\partial\lambda^{(m-1)}}(\lambda^{n-1})\Big|_{\lambda_1} & \lambda^{n-1}\Big|_{\lambda_2} \end{bmatrix} \tag{22}$$

if the characteristic equation in question has a multiple root λ_1 of multiplicity m and a root λ_2 of multiplicity 1. This is a typical example from which there is no loss of generality in writing $[W]$.

Then the system is described in z coordinate as follows:

$$[\dot{z}]=[V]^{-1}[\alpha][V][z]=[\Lambda][z] \tag{23}$$

for the distinct case, and

$$[\dot{z}]=[W]^{-1}[\alpha][W][z]=[J][z] \tag{24}$$

for the multiple case, where $[\Lambda]$ and $[J]$ are:

$$[\Lambda]=\begin{bmatrix} \lambda_1 & & & & & \\ & \lambda_2 & & & & \\ & & \lambda_3 & & & \\ & & & . & & \\ & & & & . & \\ & & & & & . \\ & & & & & & \lambda_n \end{bmatrix} \tag{23a}$$

p.27

and

$$[J]=\begin{bmatrix} \lambda_1 & 1 & & & & & \\ & \lambda_1 & 1 & & & & \\ & & \lambda_1 & \cdot & & & \\ & & & \cdot & \cdot & & \\ & & & & \cdot & \cdot & \\ & & & & & \cdot & 1 \\ & & & & & & \lambda_1 \\ & & & & & & & \lambda_2 \end{bmatrix} \tag{24a}$$

respectively.

In Kron's terminology, eqn. (23) is called the corresponding primitive system. Putting a general system into this form not only helps us solve the equation easily, but also gives us much insight.

It is very interesting to note that the Vandermonde matrix $[V]$ and the modified Vandermonde matrix $[W]$ can be derived from Krylov's matrix. Here we only derive the $[V]$ matrix from the $[H]$ matrix.

Consider z coordinates as a special case of x coordinates. From (23) we have:

$$[V][\Lambda][V]^{-1}=[\alpha]. \tag{25}$$

Comparing (10) and (25), we see that the $[V]$ matrix is only a special case of $[H]$.

Let us assume in (25) that

$$[\Lambda]=\begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & & \lambda_n \end{bmatrix}.$$

Then

$$[\Lambda]^2=\begin{bmatrix} \lambda_1^2 & & & & \\ & \lambda_2^2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & & \lambda_n^2 \end{bmatrix}$$

and

$$[\Lambda]^{n-1}=\begin{bmatrix} \lambda_1^{n-1} & & & & \\ & \lambda_2^{n-1} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & & \lambda_n^{n-1} \end{bmatrix}.$$

f. 28

The $[H]$ matrix is formed accordingly:

$$[H]=\begin{bmatrix} l \\ l[\Lambda] \\ l[\Lambda]^2 \\ \cdot \\ \cdot\cdot \\ \cdot \\ l[\Lambda]^{n-1} \end{bmatrix}, \tag{9}$$

and let

$$l=[1,1,1,\ldots 1],$$

we have arrived at:

$$[H]=\begin{bmatrix} 1 & 1 & 1 & . & 1 \\ \lambda_1 & \lambda_2 & \lambda_3 & . & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & . & \lambda_n^2 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \lambda_3^{n-1} & . & \lambda_n^{n-1} \end{bmatrix},$$

which is exactly the Vandermonde matrix.

If a similar reasoning is followed, the $[W]$ matrix can also be derived from the $[H]$ matrix.

5. General Heaviside expansion

Because the Krylov matrix and the Vandermonde matrix are so easy to form and to use, we would take the advantages to perform the similarity transformations on a general system in order that we will naturally obtain a simple procedure for the general Heaviside expansion.

Consider the general system again:

$$[\dot{x}]=[A][x], \tag{1}$$

whose transfer matrix is:

$$[\Phi_x(s)]=[sI-A]^{-1}, \tag{2a}$$

as we derived before.

Now, we use the following two similarity transformations:

$$[x]=[H]^{-1}[y],$$

$$[y]=[W][z],$$

to change (1) into a primitive system finally:

$$[\dot{z}]=[W]^{-1}[H][A][H]^{-1}[W][z] \tag{26}$$

or

$$[\dot{z}]=[J][z], \tag{26a}$$

p.29

ORIGINAL PAGE IS
OF POOR QUALITY

where

$$[J]=\left[\begin{array}{cccccc:cccc}
\lambda_1 & 1 & & & & & & & & \\
 & \lambda_1 & 1 & & & & & & & \\
 & & \lambda_1 & 1 & & & & & & \\
 & & & \lambda_1 & 1 & & & & & \\
 & & & & \lambda_1 & & & & & \\
 & & & & & \ddots & & & & \\
 & & & & & & \ddots & & & \\
\hdashline
 & & & & & & & \lambda_2 & 1 & \\
 & & & & & & & & \lambda_2 & 1 \\
 & & & & & & & & & \lambda_2 \\
 & & & & & & & & & \ddots \\
 & & & & & & & & & \ddots
\end{array}\right]
\begin{array}{l}
\left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\\1\\1\end{array}}\right\} k\times k \\
\left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\end{array}}\right\} q\times q
\end{array} \tag{27}$$

Laplace transforming (26*a*) and solving for the transfer matrix defined in z coordinates:

$$[\Phi_z(s)]=[sI-J]^{-1}. \tag{28}$$

The inverse of (28) can be directly written:

$$[\Phi_z(s)]=\left[\begin{array}{cccccc:cccc}
\frac{1}{(s-\lambda_1)} & \frac{1}{(s-\lambda_1)^2} & \frac{1}{(s-\lambda_1)^3} & \frac{1}{(s-\lambda_1)^4} & \cdots & & & & & \\
 & \frac{1}{(s-\lambda_1)} & \frac{1}{(s-\lambda_1)^2} & \frac{1}{(s-\lambda_1)^3} & \cdots & & & & & \\
 & & \frac{1}{(s-\lambda_1)} & \frac{1}{(s-\lambda_1)^2} & \cdots & & & & & \\
 & & & \frac{1}{(s-\lambda_1)} & \cdots & & & & & \\
 & & & & \ddots & & & & & \\
\hdashline
 & & & & & & \frac{1}{(s-\lambda_2)} & \frac{1}{(s-\lambda_2)^2} & \cdots & \\
 & & & & & & & \frac{1}{(s-\lambda_2)} & \cdots & \\
 & & & & & & & & \ddots & \\
 & & & & & & & & & \ddots
\end{array}\right]
\begin{array}{l}
\left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\\1\\1\end{array}}\right\} k\times k \\
\left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\end{array}}\right\} q\times q
\end{array} \tag{29}$$

The simplicity of (29) helps us solve the general Heaviside expansion problem.

When changing $[\Phi_z(s)]$ back to $[\Phi_x(s)]$ by use of (8) and (21), we easily find that they are related by:

$$[\Phi_x(s)] = [H]^{-1}[W][\Phi_z(s)][W]^{-1}[H]. \tag{30}$$

Equation (30) is our basic formula for the expansion.

Substituting (7) and (29) into (30) we obtain:

ORIGINAL PAGE IS
OF POOR QUALITY

$$\sum_1^k \frac{[K_k]}{(s-\lambda_1)^k} + \sum_1^q \frac{[Q_q]}{(s-\lambda_2)^q} + \dots$$

$$= [H]^{-1}[W] \left[\begin{array}{cccc|c} \frac{1}{s-\lambda_1} & \frac{1}{(s-\lambda_1)^2} & \frac{1}{(s-\lambda_1)^3} & \cdots & \\ & \frac{1}{(s-\lambda_1)} & \frac{1}{(s-\lambda_1)^2} & \cdots & k\times k \\ & & \frac{1}{(s-\lambda_1)} & \cdots & \\ \hline & & & \ddots & q\times q \end{array}\right] [W]^{-1}[H]. \tag{31}$$

It is noted that both $[H]$ and $[W]$ and their inverses are all constant matrices while $[K_k]$ and $[Q_q]$ are to be determined.

We multiply both sides of (31) by a scalar $(s-\lambda_1)^k$, then letting $s=\lambda_1$, $[K_k]$ is determined:

$$[K_k] = [H]^{-1}[W] \left[\begin{array}{ccccccc|ccc} 0 & 0 & 0 & . & . & 0 & 1 & & & \\ 0 & 0 & 0 & .. & . & 0 & 0 & & & \\ 0 & 0 & 0 & . & . & 0 & 0 & & k\times k & \\ . & . & . & . & . & . & . & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & \\ \hline & & & & & & & 0 & & \\ & & & & & & & & \ddots & \\ & & & & & & & & & 0 \\ \hline & & & & & & & & & 0 \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 0 \end{array}\right] \times [W]^{-1}[H]. \tag{32}$$

p. 31

Only the element at the first row and kth column has a value of unity; all other elements are equal to zero. Similarly:

$$[K_1]=[H]^{-1}[W]\left[\begin{array}{ccccccc|c} 1 & 0 & 0 & 0 & . & . & 0 & \\ 0 & 1 & 0 & 0 & . & . & 0 & \\ 0 & 0 & 1 & 0 & . & . & 0 & k\times k \\ . & . & . & . & . & . & . & \\ . & . & . & . & . & . & . & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \\ \hline & & & & & & & \\ & & & & & & & \end{array}\right][W]^{-1}[H]. \tag{33}$$

In eqn. (33) only the elements on the diagonal of the first Jordan block have values of unity; all other elements are equal to zero.

Next, we find $[K_2]$:

$$[K_2]=[H]^{-1}[W]\left[\begin{array}{cccccc|c} 0 & 1 & 0 & 0 & . & . & \\ 0 & 0 & 1 & 0 & 0 & . & \\ 0 & 0 & 0 & 1 & 0 & . & k\times k \\ . & . & . & . & . & . & \\ . & . & . & . & 0 & . & \\ \hline & & & & & & \\ & & & & & & \end{array}\right][W]^{-1}[H]. \tag{34}$$

Only the elements just above the main diagonal of the first block are unity; all other elements are equal to zero.

The matrix coefficients from $[K_1]$ to $[K_k]$ are determined by the above simple procedure. It is interesting to note how suitable this procedure is for digital computer programming.

Following the same process, we find:

$$[Q_q]=[H]^{-1}[W]\left[\begin{array}{c|ccccc|c} & & & & & & \\ & & & & & & k\times k \\ \hline & 0 & 0 & . & . & 1 & \\ & 0 & 0 & 0 & 0 & 0 & \\ & 0 & 0 & 0 & . & 0 & q\times q \\ & . & . & . & . & . & \\ & 0 & 0 & 0 & 0 & . & \\ \hline & & & & & & \end{array}\right][W]^{-1}[H]. \tag{35}$$

Only the element at the right upper corner of the second Jordan block equals unity. All other elements equal zero.

p. 32

Similarly:

ORIGINAL PAGE IS
OF POOR QUALITY

$$[Q_1]=[H]^{-1}[W]\left[\begin{array}{c|cccccc} & & & & & & \\ \hline & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & \cdot & & \\ & & & & & \cdot & \\ & & & & & & 1 \end{array}\right]\begin{array}{l} k\times k \\ \\ \\ \\ \\ \\ \\ q\times q \end{array}\Bigg][W]^{-1}[H], \quad (36)$$

etc.

Therefore, all $[K_k][Q_q]\ldots$ have been evaluated by simply changing appropriate elements into unity or zero. These simple results are believed to be new.

If we degenerate the problem into the distinct eigenvalue case, the formula is simpler. Expanding (30) for this case, we have:

$$\frac{[R_1]}{s-\lambda_1}+\frac{[R_2]}{s-\lambda_2}+\frac{[R_3]}{s-\lambda_3}+\ldots$$

$$=[H]^{-1}[V]\begin{bmatrix} \frac{1}{s-\lambda_1} & & & & & \\ & \frac{1}{s-\lambda_2} & & & & \\ & & \frac{1}{s-\lambda_3} & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & \frac{1}{s-\lambda_n} \end{bmatrix}[V]^{-1}[H]. \quad (37)$$

The unknown constant matrices are:

$$[R_1]=[H]^{-1}[V]\begin{bmatrix} 1 & 0 & 0 & . & . & . & 0 \\ 0 & 0 & 0 & . & . & . & 0 \\ 0 & 0 & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & . & 0 \end{bmatrix}[V]^{-1}[H], \quad (38)$$

$$[R_2]=[H]^{-1}[V]\begin{bmatrix} 0 & 0 & . & . & . & . & 0 \\ 0 & 1 & 0 & . & . & . & 0 \\ 0 & 0 & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & 0 \end{bmatrix}[V]^{-1}[H], \quad (39)$$

etc.

It should be noted that we have two by-products in the approach; first, in the use of the $[H]$ matrix for performing the similarity transformation we automatically find the characteristic equation. This way is simpler than Newton's formula (Chen and Haas 1968) or Faddeev's method (Chen and Haas 1968). Secondly, it is easily shown that the product of $[H]^{-1}[V]$ is the modal matrix, each column of which is an eigenvector. Therefore, this is an easy way to find eigenvectors. The arbitrary constants in the regular approach to the eigenvalue problem become an arbitrary vector $[l]$.

6. Conclusions

A procedure for the general matrix Heaviside expansion is established. It is particularly suitable for digital computation in state variable analysis.

Historically, researchers in the control engineering field have long been investigating the relationship between the inverse Vandermonde matrix and the residues of the regular Heaviside expansion. Tou (1964), Brule (1964) and Reis (1967) have presented their results. However, they always restrict themselves to the distinct eigenvalue case. No general or multiple eigenvalue case has been given. On the other hand, in the areas from the companion matrix formulation to the controllability and observability tests (Kalman *et al.* 1963) we have long been interested in Krylov's matrix; however, we have never obtained a united picture. This paper presents the general view and points out the essential applications and their relations.

```
C
C
C   N = DIMENSION OF TRANSFER FUNCTION MATRIX
C   A = N X N MATRIX OF GENERAL FORM  (XDOT) = (A)(X)
C   VECTRL = ANY ROW VECTOR SUCH THAT THE INVERSE OF THE KRYLOV
C            TRANSFORMATION MATRIX (H) EXISTS
C   PHIZ = SIMPLIFIED PHIZ(S) MATRIX ALL ELEMENTS UNITY OR ZERO
C   W = GENERAL VANDERMONDE MATRIX
C
C
C              SUBROUTINES
C
C   MATX(N,A,B,C) RETURNS (C) = (A)(B)   ALL N X N
C   MINV = LIBRARY DOUBLE PRECISION MATRIX INVERSE SUBROUTINE
C
C
C
C
      DOUBLE PRECISION U,A,W,H,P2,T,P1,HIV,VIH,D,PHIZ
      DIMENSION P1(4,4),LL(4,4),M(4,4)
      DIMENSION VECTRL(4),A(4,4),W(4,4),P2(4,4),T(4,4),HIV(4,4),VIH(4,4)
     1  ,H(4,4),PHIZ(4,4)
   10 FORMAT(4F10.0)
   11 FORMAT(4E16.7)
   13 FORMAT (I3)
      READ (5,13) N
      READ (5,10)(VECTRL(I),I=1,N)
      READ (5,10) (( A(I,J),J=1,N),I=1,N)
      READ (5,10) ((W(I,J),J=1,N),I=1,N)
      WRITE (6,205)
  205 FORMAT(///,'GENERAL STATE EQN (XDOT) = (A)(X)',//)
      WRITE (6,11) (( A(I,J),J=1,N),I=1,N)
      DO 27 I = 1,N
      DO 27 J =1,N
   27 P2(I,J) =  A(I,J)
      DO 24 I = 1,N
   24 H(1,I) =VECTRL(I)
      DO 28 I = 2,N
      DO 25 J =1,N
      H(I,J) = 0.
      DO 25 K =1,N
   25 H(I,J) = VECTRL(K)*P2(K,J) + H(I,J)
      CALL MATX(N,P2,A,T)
      DO 28 III = 1,N
      DO 28 JJJ =1,N
   28 P2(III,JJJ) = T(III,JJJ)
      WRITE (6,206)
  206 FORMAT(///,'H MATRIX TO TRANSFORM (A) TO (ALFA)',//)
      WRITE (6,11) ((H(I,J),J=1,N),I=1,N)
      DO 29 I = 1,N
      DO 29 J =1,N
   29 P1(I,J) = H(I,J)
      CALL MINV(P1,N,D,LL,M)
      CALL MATX(N,P1,W,HIV)
      CALL MATX(N,A,P1,P2)
      CALL MATX(N,H,P2,T)
      WRITE (6,208)
  208 FORMAT(///,'(YDOT) = (ALFA)(Y) WHERE (ALFA) = COMPANION FORM',//)
      WRITE(6,11) (( T(I,J),J=1,N),I=1,N)
      DO 33 I = 1,N
      DO 33 J =1,N
   33 P1(I,J) = W(I,J)
```

p.35

ORIGINAL PAGE IS
OF POOR QUALITY

```
      WRITE (6,212)
  212 FORMAT(///,'GENERAL VANDERMONDE MATRIX',//)
      WRITE (6,11)(( W(I,J),J=1,N),I=1,N)
      CALL MINV(P1,N,D,LL,M)
      CALL MATX(N,P1,H,VIH)
      CALL MATX(N,P1,T,A)
      CALL MATX(N,A,W,P1)
      WRITE (6,210)
  210 FORMAT (///,'JORDAN OR DIAGONAL FORM',//)
      WRITE (6,11) ((P1(I,J),J=1,N),I=1,N)
      DO 800 KK =1,N
      READ (5,10) ((PHIZ(I,J),J=1,N),I=1,N)
      CALL MATX(N,HIV,PHIZ,A)
      CALL MATX(N,A,VIH,P2)
  802 CONTINUE
      WRITE (6,803) KK,KK
  803 FORMAT(///,'K(',I3,') MATRIX FOR ROOT(',I3,')',//)
  800 WRITE (6,11) ((P2(I,J) ,J=1,N),I=1,N)
      STOP
      END


      SUBROUTINE MATX(N,A,B,C)
      DOUBLE PRECISION A,B,C
      DIMENSION A(N,N),B(N,N),C(N,N)
      DO 10 I =1,N
      DO 10 J =1,N
      C(I,J) = 0.
      DO 10 K =1,N
   10 C(I,J) = A(I,K)*B(K,J) + C(I,J)
      RETURN
      END


PROGRAM END
```

p. 36

Appendix

The method is best illustrated by the following example.

Consider a given system:

$$[\dot{x}] = [A][x], \tag{A 1}$$

where

$$[A] = \begin{bmatrix} -5 & 2 & 0 & 0 \\ 0 & -4 & 0 & 0 \\ -3 & 2 & -4 & -1 \\ -3 & 2 & 0 & -4 \end{bmatrix}. \tag{A 2}$$

We use the similarity transformation:

$$[y] = [H][x], \tag{A 3}$$

where

$$[H] = \begin{bmatrix} l[I] \\ l[A] \\ l[A]^2 \\ l[A]^3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -11 & 2 & -4 & -5 \\ 82 & -48 & 16 & 24 \\ -530 & 436 & -64 & -112 \end{bmatrix} \tag{A 4}$$

to transform (A 1) into y coordinates:

$$[\dot{y}] = [H][A][H]^{-1}[y] = [\alpha][y] \tag{A 5}$$

in which

$$[\alpha] = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -320 & -304 & -108 & -17 \end{bmatrix}. \tag{A6}$$

Performing another similarity transformation, or letting

$$[W][z] = [y], \tag{A 7}$$

where $[W]$ was defined in (22), we obtain:

$$\begin{aligned} [\dot{z}] &= [W]^{-1}[H][A][H]^{-1}[W][z] \\ &= [J][z], \end{aligned} \tag{A 8}$$

in which $[J]$ is the Jordan matrix. For this example which has a multiple root, -4, with multiplicity 3 and a root, -5, with multiplicity 1. The corresponding Jordan matrix is:

$$[J] = \begin{bmatrix} -4 & 1 & 0 & 0 \\ 0 & -4 & 1 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & -5 \end{bmatrix}. \tag{A9}$$

p.37

The Laplace transform of the transition matrix in z coordinates is readily seen:

$$[\Phi_z(s)]=[sI-J]^{-1}=\begin{bmatrix} s+4 & -1 & 0 & 0 \\ 0 & s+4 & -1 & 0 \\ 0 & 0 & s+4 & 0 \\ 0 & 0 & 0 & s+5 \end{bmatrix}^{-1}$$

$$=\begin{bmatrix} \frac{1}{(s+4)} & \frac{1}{(s+4)^2} & \frac{1}{(s+4)^3} & 0 \\ 0 & \frac{1}{(s+4)} & \frac{1}{(s+4)^2} & 0 \\ 0 & 0 & \frac{1}{(s+4)} & 0 \\ 0 & 0 & 0 & \frac{1}{(s+5)} \end{bmatrix}. \tag{A 10}$$

We return to the x coordinates and have:

$$[\Phi_x(s)]=[H]^{-1}[W][\Phi_z(s)][W]^{-1}[H]$$

$$=[H]^{-1}[W]\begin{bmatrix} \frac{1}{(s+4)} & \frac{1}{(s+4)^2} & \frac{1}{(s+4)^3} & 0 \\ 0 & \frac{1}{(s+4)} & \frac{1}{(s+4)^2} & 0 \\ 0 & 0 & \frac{1}{(s+4)} & 0 \\ 0 & 0 & 0 & \frac{1}{(s+5)} \end{bmatrix}[W]^{-1}[H]. \tag{A 11}$$

Assuming constant matrices $[K_1]$, $[K_2]$, $[K_3]$ and $[Q_1]$, we obtain:

$$[\Phi_x(s)]=\frac{[K_1]}{(s+4)}+\frac{[K_2]}{(s+4)^2}+\frac{[K_3]}{(s+4)^3}\frac{[Q_1]}{(s+5)}. \tag{A 12}$$

Equating (A 11) and (A 12) gives:

$$\frac{[K_1]}{(s+4)}+\frac{[K_2]}{(s+4)^2}+\frac{[K_3]}{(s+4)^3}+\frac{[Q_1]}{(s+5)}$$

$$=[H]^{-1}[W]\begin{bmatrix} \frac{1}{(s+4)} & \frac{1}{(s+4)^2} & \frac{1}{(s+4)^3} & 0 \\ 0 & \frac{1}{(s+4)} & \frac{1}{(s+4)^2} & 0 \\ 0 & 0 & \frac{1}{(s+4)} & 0 \\ 0 & 0 & 0 & \frac{1}{(s+5)} \end{bmatrix}[W]^{-1}[H]. \tag{A 13}$$

p. 38

Now, to evaluate the constant matrices of (A 13), multiply both sides by $(s+5)$ and then let $s=-5$. We have $[Q_1]$:

$$[Q_1]=[H]^{-1}[W]\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1\end{bmatrix}\quad [W]^{-1}[H]=\begin{bmatrix}1&-2&0&0\\0&0&0&0\\6&-12&0&0\\3&-6&0&0\end{bmatrix}.$$

Similarly, we multiply both sides of (A 13) by $(s+4)^3$, and letting $s=-4$:

$$[K_3]=[H]^{-1}[W]\begin{bmatrix}0&0&1&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\quad [W]^{-1}[H]=\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&4&0&0\\0&0&0&0\end{bmatrix}.$$

Similarly, $[K_2]$ and $[K_1]$ can be obtained by inspection. Thus, we have:

$$[K_2]=[H]^{-1}[W]\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&0\\0&0&0&0\end{bmatrix}[W]^{-1}[H]=\begin{bmatrix}0&0&0&0\\0&0&0&0\\3&-10&0&-1\\0&-4&0&0\end{bmatrix}.$$

By similar reasoning, we obtain:

$$[K_1]=[H]^{-1}[W]\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix}[W]^{-1}[H]=\begin{bmatrix}0&2&0&0\\0&1&0&0\\-6&12&1&0\\-3&6&0&1\end{bmatrix}.$$

Once the matrix coefficients have been determined, the transition matrix can be directly written as follows:

$$[\phi_x(t)]=[K_1]\exp(-4t)+[K_2]t\exp(-4t)$$
$$+[K_3](t^2/2)\exp(-4t)+[Q_1]\exp(-5t.)$$

A digital computer programme for the general Heaviside expansion is included.

Acknowledgments

The authors are grateful to Professor J. F. Coales of Cambridge University and Professor C. J. Huang of the University of Houston for constant encouragement. This work was supported in part by the National Aeronautics and Space Administration grant NGL-44-005-084.

References

Bewley, L. V., 1961, *Tensor Analysis of Electric Circuits and Machines* (New York : The Ronald Press), p. 11.

Brule, J. D., 1964, *I.E.E.E. Trans. autom. Control*, 9, 314.

Chen, C. F., and Haas, I. J., 1968, *Elements of Control Systems Analysis: Classical and Modern Approaches* (Prentice-Hall), p. 339.

p. 39

CHEN, C. F., and PARKER, R. R., 1966, *I.E.E.E. Trans. Education*, **9**, 209.
CHEN, C. F., and YATES, R. E., 1967, *J. bas. Engng, A.M.S.E. Trans.*, **89**, 269.
GANTMACHER, F. R., 1959, *The Theory of Matrices* (New York : Chelsea Publishing Co.), p. 202.
KALMAN, R. E., HO, Y. C., and NARENDRA, K. S., 1963, *Controllability of Linear Dynamical Systems* (New York : John Wiley & Sons).
KRON, G., 1939, *Tensor Analysis of Networks* (New York : John Wiley & Sons).
KRYLOV, A. N., 1931, *Izv. Akad. Nauk SSSR*, Ser. Fiz.-Mat., **4**, 491.
KUO, F. F., 1966, *Network Analysis and Synthesis*, 2nd edition (John Wiley & Sons), p. 153.
RAO, K. R., and AHMED, N., 1968, *Proc. I.E.E.E.*, **56**, 884.
REIS, G. C., 1967, *I.E.E.E. Trans. autom. Control*, **12**, 793.
TOU, J. T., 1964, *I.E.E.E. Trans. autom. Control*, **9**, 314.
WONHAM, W. M., and JOHNSON, C. D., 1963, *Proceedings of the 4th Joint Automatic Control Conference*, p. 101; 1964, *J. bas. Engng, A.S.M.E. Trans.*, **86**, 107.
VAN VALKENBURG, M. E., 1964, *Network Analysis*, 2nd edition, p. 170.

p.40

Project

(A) Project Title: A Normalized Multidimensional Newton-Raphson Method

(B) Project Abstract:

The only difficulty involved in the Newton-Raphson method is how to make the initial guess. This project presents a two-transformation technique which enables us to make the initial guess unnecessary. Several characteristic equations are tested.

(C) Publication: International Journal of Control, 12, 273 (1970)

(D) Year: 1970

(E) Department: Electrical Engineering

(F) Student Name: N. R. Strader

(G) Faculty Advisor: Professor C. F. Chen

INT. J. CONTROL, 1970, VOL. 12, NO. 2, 273–279

A normalized multidimensional Newton-Raphson method†

C. F. CHEN‡
Engineering Department, University of Cambridge

and N. R. STRADER
Electrical Engineering Department, University of Houston

[Received 10 June 1969, Revised 28 February 1970]

The only difficulty involved in the Newton–Raphson method is how to make the initial guess. This paper presents a two-transformation technique which enables us to make the initial guess unnecessary. Several characteristic equations are tested.

1. Introduction

The scalar Newton–Raphson method for solving a high-order algebraic equation:

$$f(s) = s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_n = 0 \tag{1}$$

is written as:

$$s_{i+1} = s_i - f'(s_i)^{-1} f(s_i). \tag{2}$$

If we make the right guess for s_0, after the ith iteration, we will obtain one of the roots, s_{i+1}. However, if the guess is not good, we may never obtain a solution. There is no systematic method to use as a guide line to make the initial guess.

2. The multidimensional Newton–Raphson method

If there is a set of simultaneous algebraic equations:

$$f(\mathbf{s}) = 0 \tag{3}$$

or

$$\left.\begin{aligned} f_1(s_1, s_2, \ldots, s_n) &= 0, \\ f_2(s_1, s_2, \ldots, s_n) &= 0, \\ &\vdots \\ f_n(s_1, s_2, \ldots, s_n) &= 0. \end{aligned}\right\} \tag{4}$$

and it is desired to find $s_1, s_2, \ldots$ and s_n, we can use the general multidimensional Newton–Raphson method (Bellman and Kalaha 1965 or Childs 1967):

$$\mathbf{s}_{i+1} = \mathbf{s}_i + -\left[\frac{\partial \mathbf{f}}{\partial \mathbf{s}}\right]^{-1} \mathbf{f}_i \tag{5}$$

† Communicated by Professor Chen.
‡ Permanent address: University of Houston, Houston, Texas.

p. 42

where **s** is a vector and

$$\frac{\partial \mathbf{f}}{\partial \mathbf{s}} = \begin{bmatrix} \frac{\partial f_1}{\partial s_1} & \frac{\partial f_1}{\partial s_2} & \frac{\partial f_1}{\partial s_3} & \cdots \\ \frac{\partial f_2}{\partial s_1} & \frac{\partial f_2}{\partial s_2} & \frac{\partial f_2}{\partial s_3} & \cdots \\ \frac{\partial f_3}{\partial s_1} & \frac{\partial f_3}{\partial s_2} & \frac{\partial f_3}{\partial f_3} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}, \tag{6}$$

which is usually called the Jacobian.

3. Equivalent equations

Instead of solving the characteristic eqn (1) by using (2), we would rather look at the problem from a different angle.

First, the equation:

$$f(s) = s^n + a_1 s^{n-1} + a_2 s^{n-2} + \ldots + a_n = 0$$

can always be decomposed into a set of equivalent equations:

$$\left.\begin{aligned} \alpha+\beta+\gamma+\ldots &= a_1, \\ \alpha\beta+\beta\gamma+\ldots &= a_2, \\ &\vdots \\ \alpha\beta\gamma\ldots &= a_n, \end{aligned}\right\} \tag{7}$$

where $-\alpha, -\beta, -\gamma, \ldots$ are the required roots. Therefore, finding the solution of (7) is equivalent to finding the roots of (1). We can use (5), or the multi-dimensional Newton–Raphson formula to obtain the values of $\alpha, \beta, \gamma, \ldots$, etc. This new viewpoint is similar to that of changing a high-order differential equation into a set of first-order state equations. Because of this change, new light is shed on the problem.

4. First transformation

The roots of the characteristic eqn. (1) are distributed in the s-plane. If we consider that each root is a unit mass, the s-plane must have a centre of gravity. From Evans' root-locus method (Chen and Haas 1968), the centre is determined by the arithmetic mean of the roots, or

$$\text{c.g.} = \frac{\alpha+\beta+\gamma+\ldots}{n} = k. \tag{8}$$

If we move the origin of the s-plane to the centre of gravity, we would have a more balanced picture.

It is well known that the centre of gravity can also be determined by the second coefficient of the characteristic equation. Therefore, we perform the first transformation by letting:

$$y = s - k. \tag{9}$$

Substituting (9) into (1) yields:

$$y^n + b_2 y^{n-2} + b_3 y^{n-3} + \ldots + b_n = 0. \tag{10}$$

p. 43

It is noted that the second term is missing from the y equation.

Let us use an example for illustration:

$$s^3+13s^2+26s+12 = 0. \tag{11}$$

The value of k is found by:

$$k = \tfrac{13}{3} = 4{\cdot}33. \tag{12}$$

Letting:

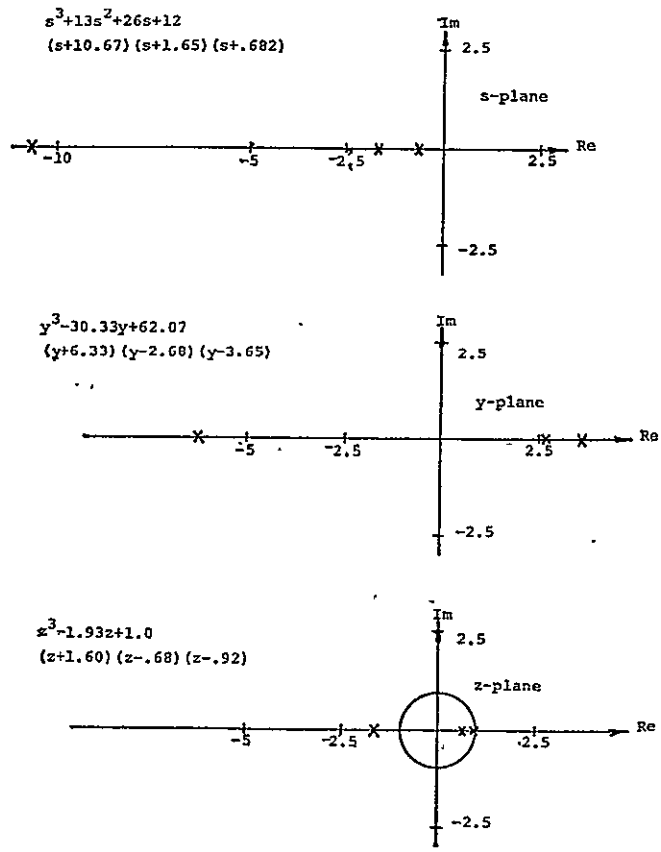
$$y = s-4{\cdot}33 \tag{13}$$

and substituting into (11) gives:

$$y^3-30{\cdot}33y+62{\cdot}07 = 0. \tag{14}$$

The root distributions in the s plane and in the y plane are shown in fig. 1.

Fig. 1



5. Second transformation

The first transformation is a general practice for solving equations. It coincides with the technique used in the cubic equation formula and in the quadratic equation formula.

The geometrical interpretation of the first transformation, of course, is to relocate the origin at the centre of gravity.

However, after the first transformation, the root distribution is still not very uniform. Some roots in the y plane are too far away from the origin while others are too close. We can change this by using another transformation and letting

$$y = \sqrt[n]{(b_n)}\, z. \tag{15}$$

In our example we use:

$$y = \sqrt[3]{(62{\cdot}07)}\, z. \tag{16}$$

Equation (15) can be called the geometrical mean transformation. The characteristic equation of our example becomes:

$$(\sqrt[3]{(62{\cdot}07)}\, z)^3 - 30{\cdot}3\,.\,(\sqrt[3]{(62{\cdot}07)})\,.\,z + 62{\cdot}07 = 0$$

or

$$z^3 - 1{\cdot}93z + 1 = 0 \tag{17}$$

and, in general,

$$z^n + \ldots + 1 = 0. \tag{18}$$

For eqn. (18), we can always make the initial guess by simply omitting the middle terms. In other words, the first guesses for the roots are:

$$z = \sqrt[n]{(-1)}, \quad = -1\angle\frac{360^\circ}{n}. \tag{19}$$

Because we have used the geometrical mean to normalize the last term and make the roots redistributed more uniformly around the unit circle, we call our approach a normalized multidimensional Newton–Raphson method.

For this example problem, fig. 1 shows the root distributions in the s plane, y plane and z plane.

6. A fifth-order example

For the equation:

$$f(s) = s^5 + 16{\cdot}2s^4 + 97{\cdot}25s^3 + 359{\cdot}6s^2 + 72{\cdot}7s + 17 \tag{20}$$

we use the new method to find the solution.

First, take the following linear transformation:

$$y = s - \frac{16{\cdot}2}{5}$$

Then we have:

$$y^5 - 7{\cdot}73y^3 + 94{\cdot}57y^2 - 847{\cdot}83y + 1676{\cdot}88 = 0. \tag{21}$$

Performing the second transformation by letting:

$$y = \sqrt[5]{(1676{\cdot}88)}\, z$$

we have:

$$z^5 - 0{\cdot}396z^3 + 1{\cdot}099z^2 - 2{\cdot}23z + 1 = 0. \tag{22}$$

Now we can solve for the roots of (22). The initial guess is the following vector:

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix} = -1 \angle \frac{360^\circ i}{5} \quad \text{for } i = 0, 1, 2, 3, 4,$$

or

$$z_1 = -0{\cdot}30901 + j0{\cdot}95106, \quad z_2 = -1{\cdot}0000, \quad z_3 = -0{\cdot}30902 - j0{\cdot}95105,$$

$$z_4 = 0{\cdot}80901 - j0{\cdot}58779 \quad \text{and} \quad z_5 = 0{\cdot}80902 + j0{\cdot}58778.$$

Fig. 2

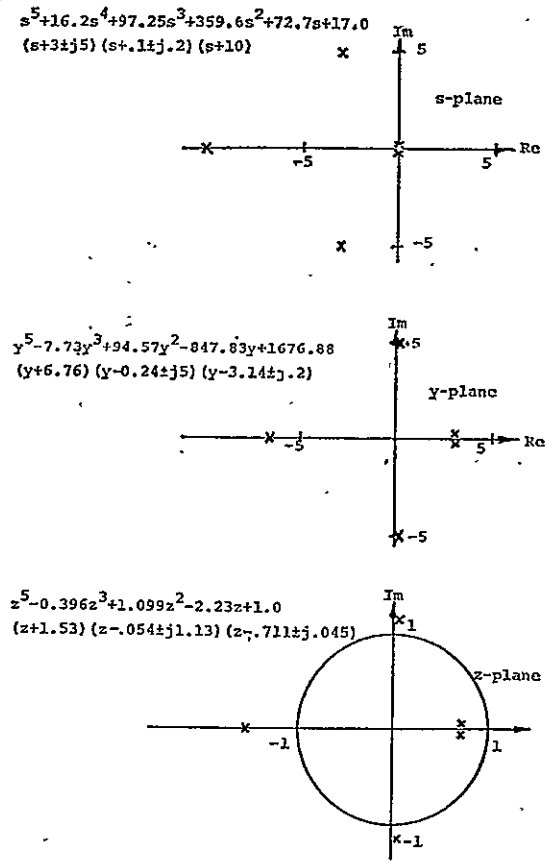
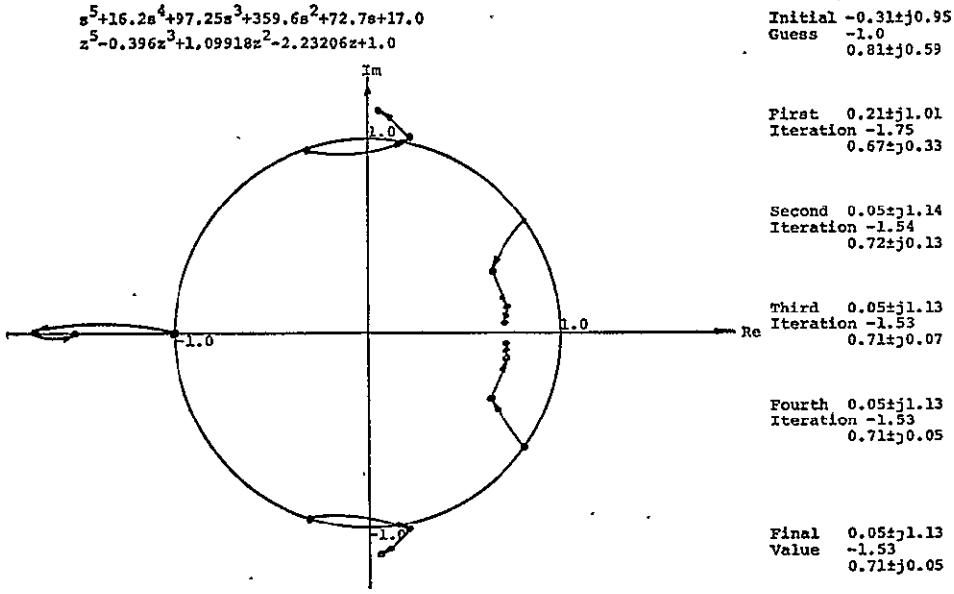


Figure 2 shows the root distributions in the s plane, y plane, and z plane. It is evident that our five values of the first guess are uniformly distributed on the unit circle in the z plane. They approach the actual roots by several iterations as shown in fig. 3.

p.46

Fig. 3



7. Computer program

The normalized multi-dimensional Newton–Raphson method was conceived by the first author and the computer experiments were performed by the second author on the SDS Sigma 7 computer. The programme was written using double-precision complex variables to allow accurate processing of complex roots.

The necessary inputs to the programme include the order of the equation, the tolerance for solution and the coefficients of all terms of the equation. The coefficients are read in as real numbers and divided by a_0 to normalize the highest-order term.

The first transformation is performed by using eqn. (9) and then solving for the new coefficients in the y plane. The second transformation is performed by using eqn. (15) and then solving for the new coefficients in the z plane. The initial estimate for the roots is made by using eqn. (19). This leaves the first guess for the roots uniformly distributed around the unit circle.

The iteration technique of eqn. (5) is used to converge to the actual roots of the equation in the z plane. The process is stopped when each of the equations of (7), after subtracting a_j from the ith equation, is close enough to zero to be within the tolerance specified for solution.

8. Other examples

The method was tested for a tenth-order equation:

$$s^{10}+12s^9+68{\cdot}75s^8+249{\cdot}5s^7+637s^6+1187{\cdot}5s^5 \\ +1613{\cdot}75s^4+1553s^3+994{\cdot}5s^2+373s+60=0$$

with roots at:

$$s_{1,2}=-1\pm j1,\quad s_{3,4}=0{\cdot}5\pm j\sqrt{3{\cdot}75},\quad s_{5,6}=-2\pm j1,$$
$$s_7=-0{\cdot}5,\quad s_8=-1,\quad s_9=-1{\cdot}5,\quad s_{10}=-2.$$

p.47

The following roots were calculated with our method:

$$s_{1,2} = -0{\cdot}99997 \pm j0{\cdot}99998, \quad s_{3,4} = -0{\cdot}49999 \pm j1{\cdot}93649,$$
$$s_{5,6} = -199977 \pm j1{\cdot}00027, \quad s_7 = -0{\cdot}50000;$$
$$s_8 = -1{\cdot}00005, \quad s_9 = -1{\cdot}49904, \quad s_{10} = -2{\cdot}00145.$$

An equation with repeated roots:

$$s^5 + 5{\cdot}4s^4 + 11{\cdot}64s^3 + 12{\cdot}52s^2 + 6{\cdot}72s + 1{\cdot}44 = 0$$

was also tested. The actual roots were:

$$s_{1,2,3} = -1, \quad s_{4,5} = -1{\cdot}2.$$

The following roots were calculated with the normalized method:

$$s_{1,2} = -1{\cdot}00633 \pm j0{\cdot}01398, \quad s_3 = -0{\cdot}98671,$$
$$s_{4,5} = -1{\cdot}20032 \pm j0{\cdot}00809.$$

It is interesting to note that the method is so good for a repeated root case.

9. Conclusions

A normalized multi-dimensional Newton–Raphson method is established. After two transformations, we solve the normalized equation and then perform the inverse transforms to obtain the solution. The initial guess which is the most difficult part of the original Newton–Raphson method becomes unnecessary.

Acknowledgments

The first author was supported in part by NASA Grant NGL–44–005–084; the second author was supported by NSF Grant GF–973. The constant encouragement and helpful discussions of Professor C. J. Huang of the University of Houston, Professor J. F. Coales and Doctor A. T. Fuller of the University of Cambridge are acknowledged.

References

Bellman, R. E., and Kalaba, R. E., 1965, *Quasilinearization and Nonlinear Boundary-value Problems* (New York: American Elsevier Publishing Company), p. 7.

Chen, C. F., and Haas, I. J., 1968, *Elements of Control Systems Analysis* (New Jersey: Prentice Hall), p. 117.

Childs, B., 1967, "Identification", unpublished lecture notes. University of Houston.

p. 48

Project

(A) Project Title: An Algebraic Approach to System Identification and Compensator Design

(B) Project Abstract:

In modern design of Control Systems, the synthesis techniques originated by Letov, Kalman, Bass and Tyler start with a certain functional--the quadratic performance index. However, the quadratic performance idex is not very suitable for using the industrial specifications. In other words, a link between the classical and modern methods is missing. We have to seek the missing link in the algebraic domain. This is mainly due to the digital computer consideration.

This dissertation research attempts to consider the design of control systems into two problems:

1. the Identification Problem,
2. the Compensation Problem.

Both problems are investigated from the algebraic viewpoint.

As far as the identification problem is concerned, three methods are developed. If the specifications are given in the time domain completely, a z transform technique is developed which is an extension of the application of the powerful sensitivity matrix. If the specifications are given in the frequency domain completely, Chen-Phillip's and Chen, Knox and Shieh's methods are further studied. If the

specifications are given in a hybrid form which means some index in the time domain, others in the frequency domain or in the complex domain, an original synthesis technique is established. The technique is by using the multidimensional Newton method to synthesize the transfer function from hybrid information.

The compensation problem is investigated by establishing a new form which is similar to the Cauer second form in circuit theory. The judgment of the approximation and the error estimation are based on the Minimum Integral Square criterion.

The general design philosophy is outlined as follows: To synthesize a desirable transfer function based on the hybrid specification. After finding the closed loop transfer function with an assigned compensator and simplifying the conpensated overall transfer function, we equate it with the model we synthesized before and use the Newton multidimensional method to obtain the paramteres.

(C) Publication: Ph.D. Dissertation in Electrical Engineering

(D) Year: 1970

(E) Department: Electrical Engineering

(F) Student Name: Leang-San Shieh

(G) Faculty Advisor: Professor C. F. Chen

Project

(A) Project Title: Simple Methods for Identifying Linear Systems from Frequency or Time Response Data

(B) Project Abstract:

A decomposing method is derived for identifying linear system transfer functions if response data in either the frequency domain or time domain are known. The method is based on the application of the second Cauer form or continued fraction expansion. The dominant factors of an unknown system can be systematically identified.

(C) Publication: International Journal of Control, 13, 1027 (1971)

(D) Year: 1971

(E) Department: Electrical Engineering

(F) Student Name: L. S. Shieh

(G) Faculty Advisor: Professors C. F. Chen and C. J. Huang

INT. J. CONTROL, 1971, VOL. 13, NO. 6, 1027–1039

Simple methods for identifying linear systems from frequency or time response data†

L. S. SHIEH
Electrical Engineering Department, University of Houston, Houston, Texas 77004

C. F. CHEN‡
Engineering Department, University of Cambridge, England

C. J. HUANG
Chemical Engineering Department, University of Houston, Houston, Texas 77004

A decomposing method is derived for identifying linear system transfer functions if response data in either the frequency domain or time domain are known. The method is based on the application of the second Cauer form or continued fraction expansion. The dominant factors of an unknown system can be systematically identified.

1. Introduction

In 1965 Chen and Philip, suggested a method for transfer function fitting from the frequency response data and in 1968 Chen and Knox extended the idea to the time domain. Their methods are based mainly on Bush and Caldwell's (1945) transfer function decomposition.

This paper based on the second Cauer form continued fraction expansion proposes two methods; one in the frequency domain and the other in the time domain, for constant coefficient linear system identification.

Consider a system transfer function of the form:

$$\frac{C(s)}{R(s)}=\frac{A_{2n}s^{n-1}+\ldots+A_{23}s^2+A_{22}s+A_{21}}{A_{1,(n+1)}s^n+\ldots+A_{13}s^2+A_{12}s+A_{11}}, \tag{1}$$

where $R(S)$ is an input, $C(s)$ is an output and A_{ij} are constants. Rearrange the numerator and denominator polynomials of eqn. (1) into ascending order:

$$\frac{C(s)}{R(s)}=\frac{A_{21}+A_{22}s+A_{23}s^2+\ldots+A_{2,n}s^{n-1}}{A_{11}+A_{12}s+A_{13}s^2+\ldots+A_{1,(n+1)}s^n}, \tag{1 a}$$

then using synthetic division on eqn. (1 a) we have:

$$\frac{C(s)}{R(s)}=\frac{1}{\dfrac{A_{11}}{A_{21}}+\dfrac{s\left[\left(\dfrac{A_{21}A_{12}-A_{11}A_{22}}{A_{21}}\right)+\left(\dfrac{A_{21}A_{13}-A_{11}A_{23}}{A_{21}}\right)s+\ldots\right]}{A_{21}+A_{22}s+A_{23}s^2+\ldots+A_{2,n}s^{n-1}}}$$

$$=\frac{1}{\dfrac{A_{11}}{A_{21}}+\dfrac{s(A_{31}+A_{32}s+\ldots)}{A_{21}+A_{22}s+A_{23}s^2+\ldots+A_{2,n}s^{n-1}}}, \tag{1 b}$$

† Communicated by Associate Dean of Faculties, Dr. C. J. Huang.
‡ Permanent address: Electrical Engineering Department, University of Houston, Houston, Texas 77004.

p.52

in which:

$$A_{31} = \frac{A_{21}A_{12} - A_{11}A_{22}}{A_{21}},$$

$$A_{32} = \frac{A_{21}A_{13} - A_{11}A_{23}}{A_{21}}.$$

Dividing repeatedly, we have:

$$\frac{C(s)}{R(s)} = \cfrac{1}{\cfrac{A_{11}}{A_{21}} + \cfrac{s}{\cfrac{A_{21}}{A_{31}} + \cfrac{s}{\cfrac{A_{31}}{A_{41}} + \cfrac{s}{\ddots}}}}, \tag{2}$$

where

$$A_{j,k} = A_{j-2,k+1} - \frac{A_{j-2,1}A_{j-1,k+1}}{A_{j-1,1}}, \quad j = 3, 4, \ldots, \quad k = 1, 2, \ldots. \tag{2 a}$$

An alternate form of eqn. (2) can be written in the form:

$$\frac{C(s)}{R(s)} = \cfrac{1}{h_1 + \cfrac{1}{\cfrac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\cfrac{h_4}{s} + \cfrac{1}{\ddots \, \cfrac{h_{2n}}{s}}}}}}, \tag{3}$$

where

$$h_p = \frac{A_{p,1}}{A_{(p+1),1}}, \quad p = 1, 2, \ldots, 2n, \quad h_p \neq 0.$$

Equation (3) is the second Cauer form.

Based on this form, Chen and Shieh (1968) constructed a linear model simplification technique. This revealed the fact that a transfer function is dominated by the first several quotients of h_p. Indeed the high-order transfer function can be reduced to a low-order model by simply taking the first several quotients of the continued fraction expansion. Later Towill and Mehdi (1970) compared this method (Chen and Shieh 1968) with several commonly used low-order models and investigated their sensitivity problems.

2. Identification based on frequency response data

For illustration and no loss of generality we consider a second-order system with its transfer function as follows:

$$\frac{C(s)}{R(s)} = \frac{b_1 s + b_2}{s^2 + a_1 s + a_2}, \tag{4}$$

where a_i, b_i are constants.

1.53

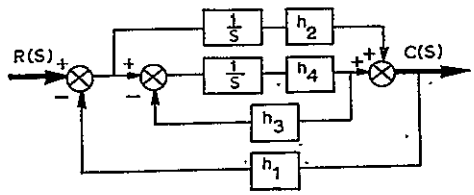
The continued fraction expansion of the second Cauer form for eqn. (4) is:

$$\frac{C(s)}{R(s)} = \cfrac{1}{h_1 + \cfrac{1}{\cfrac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\cfrac{h_4}{s}}}}}. \tag{5}$$

The block diagram representation for eqn. (5) is shown in fig. 1. The rational function for the continued fraction inversion of eqn. (5) is:

$$\frac{C(s)}{R(s)} = \frac{(h_2+h_4)\,s + h_2 h_3 h_4}{s^2 + (h_1 h_2 + h_1 h_4 + h_3 h_4)\,s + h_1 h_2 h_3 h_4}. \tag{5 a}$$

Fig. 1



Synthesis of transfer function.

Comparing eqns. (5 *a*) and (4) we have:

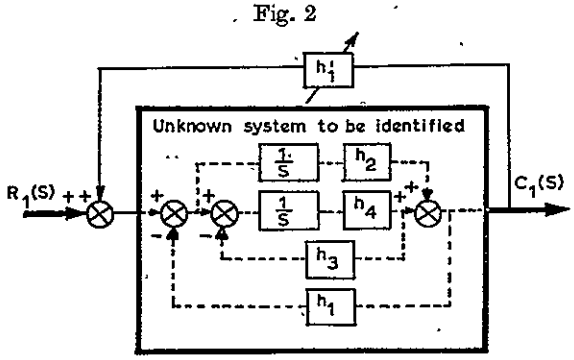
$$\left.\begin{aligned} a_1 &= h_1 h_2 + h_1 h_4 + h_3 h_4, \\ a_2 &= h_1 h_2 h_3 h_4, \\ b_1 &= h_2 + h_4, \\ b_2 &= h_2 h_3 h_4. \end{aligned}\right\} \tag{5 b}$$

Our goal is to identify the unknown quotients h_1, h_2, h_3 and h_4. The procedures are shown by the following steps.

(1) *Identifying* h_1

Suppose an unknown system which can be decomposed into a continued fraction of the form of eqn. (5) is put in the solid line block as shown in fig. 2.

Fig. 2



Identifying h_1.

A.54

The unknown quotients h_1, h_2, h_3 and h_4 are to be identified. We add a positive feedback gain h_1' to the unknown system; the block diagram of the modified unknown system is shown in fig. 2. The corresponding mathematic equation can be written as:

$$\frac{C_1(s)}{R_1(s)} = \cfrac{1}{(h_1-h_1')+\cfrac{1}{\cfrac{h_2}{s}+\cfrac{1}{h_3+\cfrac{1}{\cfrac{h_4}{s}}}}}. \tag{6}$$

The rational transfer function of eqn. (6) is:

$$\frac{C_1(s)}{R_1(s)} = \frac{(h_2+h_4)s+h_2h_3h_4}{s^2+[(h_1-h_1')h_2+(h_1-h_1')h_4+h_3h_4]s+(h_1-h_1')h_2h_3h_4}. \tag{7}$$

Equation (7) is a particular system in which its denominator has all coefficients if $h_1-h'\neq 0$. It is similar to the type '0' system in the feedback system terminology. We simply call eqn. (7) type '0' system. The frequency response data in the low-frequency region on the Bode plot shows the 0-slope response; however, if $h_1-h_1'=0$, eqn. (7) can be simplified as:

$$\frac{C_1(s)}{R_1(s)} = \frac{(h_2+h_4)s+h_2h_3h_4}{s(s+h_3h_4)}. \tag{8}$$

Equation (8) is a type 1 system. The frequency response in the low-frequency region on the Bode plot shows the -1 slope character. Due to the fact that the factor $|1/s|$ is more affected in the lower frequency region than the other factors in eqn. 8, the change of a system from a type 0 to a type 1 will show the eminent difference on the Bode plot. This peculiarity easily lets us judge the accurate value of h_1. In other words, in the lower-frequency region, we adjust h_1' until the frequency-response data of the modified system changes its slope from 0 to -1. Then we have the accurate h_1 value. Therefore, the h_1 value is unique.

(2) *Identifying* h_2

From step (1), if $h_1-h_1'=0$, the corresponding transfer function of the modified system can be written as:

$$\frac{C_1(s)}{R_1(s)} = \frac{h_2}{s}+\cfrac{1}{h_3+\cfrac{1}{\cfrac{h_4}{s}}}. \tag{9}$$

We add an integrator on the feed-forward link with negative gain h_2' to the modified system of fig. 2. The block diagram of the new modified system is shown in fig. 3. The transfer function is:

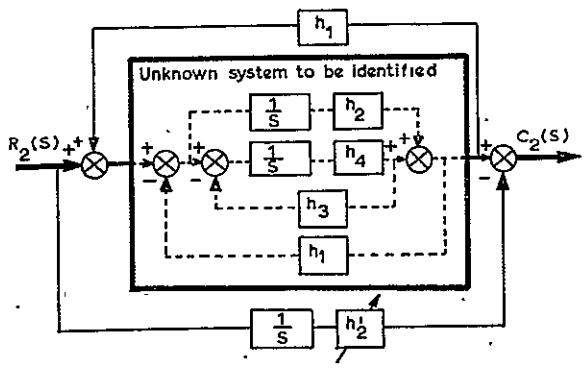
$$\frac{C_2(s)}{R_2(s)} = \frac{h_2-h_2'}{s}+\cfrac{1}{h_3+\cfrac{1}{\cfrac{h_4}{s}}}. \tag{10}$$

p. 55

The rational transfer function becomes:

$$\frac{C_2(s)}{R_2(s)} = \frac{[(h_2-h_2')+h_4]s+(h_2-h_2')h_3h_4}{s(s+h_3h_4)}. \tag{11}$$

Fig. 3



Identifying h_2.

Equation (11) is a type 1 system if $h_2-h_2' \neq 0$. The frequency response in the low-frequency region of the Bode plot has the -1 slope feature; however, when we adjust h_2' such that $h_2-h_2' = 0$, then eqn. (11) changes to:

$$\frac{C_2(s)}{R_2(s)} = \frac{h_4}{s+h_3h_4}. \tag{12}$$

Equation (12) is a type 0 system again. Of course the frequency response in the low-frequency region of the Bode plot shows 0 slope peculiarity. Therefore by adjusting h_2' we transfer the modified type 1 system of eqn. (9) to a type 0 system of eqn. (12). When the 0 slope appears in the low-frequency region on a Bode plot, we have the correct h_2 value.

(3) *Identifying* h_3

From step 2 we obtain a correct h_2 value, and the transfer function of fig. 3 becomes:

$$\frac{C_2(s)}{R_2(s)} = \frac{1}{h_3+\dfrac{1}{\dfrac{h_4}{s}}} = \frac{h_4}{s+h_3h_4}. \tag{13}$$

Equation (13) is a type 0 system again. Following step 1 we add a positive feedback gain to the system of fig. 3. The block diagram of the new modified system is shown in fig. 4. The transfer function is:

$$\frac{C_3(s)}{R_3(s)} = \frac{h_4}{s+(h_3-h_3')h_4}. \tag{14}$$

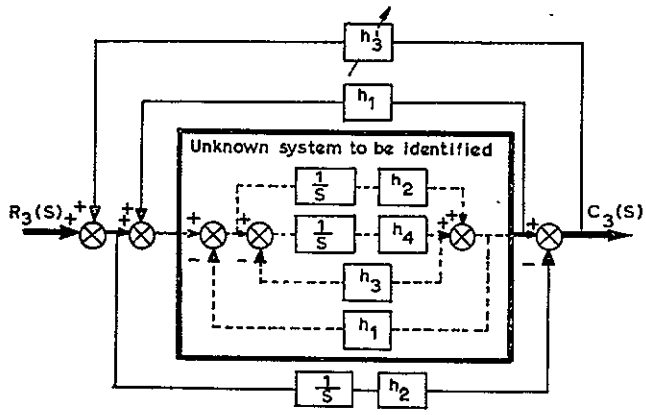
ORIGINAL PAGE IS
OF POOR QUALITY

Using the same procedure as step 1, we can identify the h_3 value correctly. The final system can be written as:

$$\frac{C_3(s)}{R_3(s)} = \frac{h_4}{s}. \tag{15}$$

ORIGINAL PAGE IS
OF POOR QUALITY

Fig. 4



Identifying h_3.

(4) *Identifying* h_4

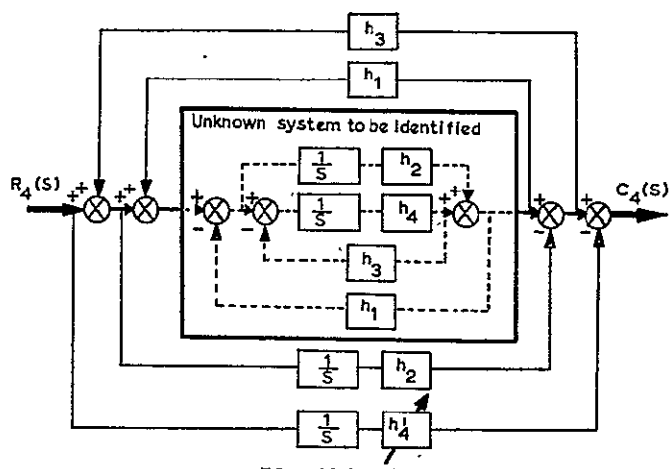
Equation (15) is a type 1 system. Again, we add a feed-forward link with a negative gain h_4' to the system of fig. 4. The block diagram of the new modified system is shown in fig. 5. Then the corresponding transfer function becomes:

$$\frac{C_4(s)}{R_4(s)} = \frac{h_4 - h_4'}{s}. \tag{16}$$

In the low-frequency region, if we choose the correct h_4' then the 0-slope feature of frequency response will appear on the Bode plot.

Since for any value of $h_4 - h_4' \neq 0$ we have a straight line with slope -1 which passes through the crossover frequency at $h_4 - h_4'$. For the case

Fig. 5



Identifying h_4.

J.57

$h_4 - h_4' = 0$ the amplitude of the frequency response on the Bode plot is 0 dB. Then the process of identification is completed.

Finally, we have all quotients h_1, h_2, h_3 and h_4. The required transfer function is:

$$\frac{C(s)}{R(s)} = \frac{(h_2+h_4)\,s + h_2 h_3 h_4 h_5}{s^2 + (h_1 h_2 + h_1 h_4 + h_3 h_4)\,s + h_1 h_2 h_3 h_4}. \tag{17}$$

Example 1

Consider the following transfer function which has a pair of complex poles to be identified:

$$\frac{C(s)}{R(s)} = \frac{-2s+6}{s^2+4s+6}. \tag{18}$$

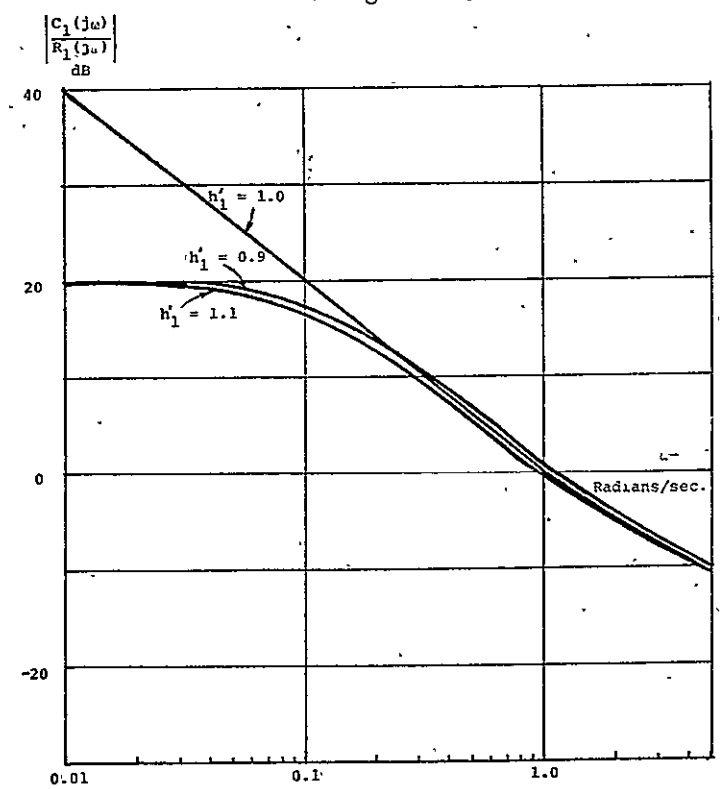
Assume eqn. (18) can be expanded into the following continued fraction expansion:

$$\frac{C(s)}{R(s)} = \cfrac{1}{h_1 + \cfrac{1}{\cfrac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\cfrac{h_4}{s}}}}}. \tag{18 a}$$

where h_j are unknown quotients to be identified.

Following step 1 and comparing the frequency response data of fig. 6 we

Fig. 6



Frequency response obtained from step 1.

p.58

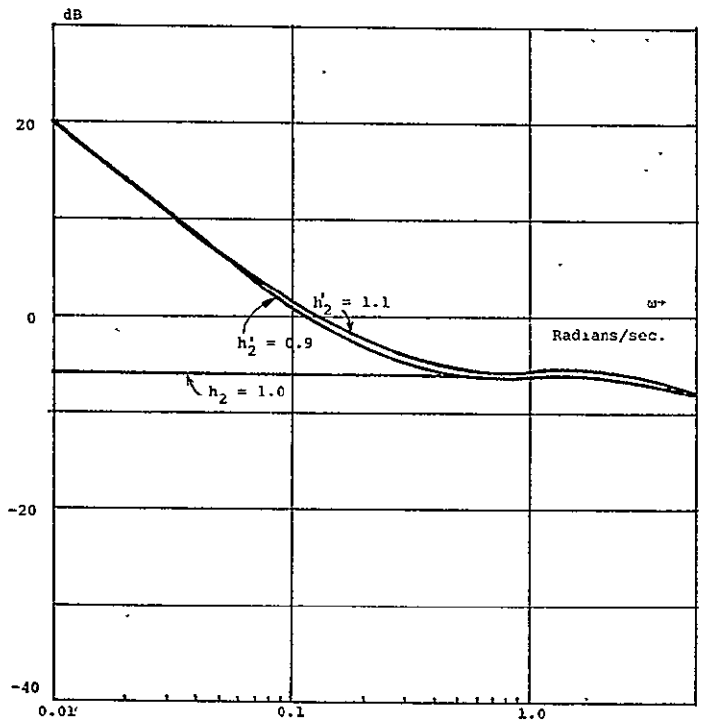
find that with a slight change of the exact h_1 value we have a very clear eminent difference of the amplitude on a Bode plot. In other words, on the Bode plot if h_1' is 0·9 or 1·1 we have a type 0 system which shows the 0 slope character. But if $h_1' = 1$, then we have a frequency response data which shows -1 slope feature on the Bode plot. Of course the system has been changed to a type 1 system, therefore $h_1 = 1$ is the required value.

Following step 2 and comparing the frequency response data of Fig. 7 we can easily obtain h_2. Since in low-frequency region if h_2' is 0·9 or 1·1 the two curves in fig. 7 almost coincide; however, if h_2' is 1 then we obtain a type 0 system with its 0 slope character on a Bode plot. Therefore the required h_2 is 1.

Using the same procedures as step 3 and comparing the frequency response curves as shown in fig. 8 we can easily figure out $h_3 = -2$.

Fig. 7

ORIGINAL PAGE IS
OF POOR QUALITY

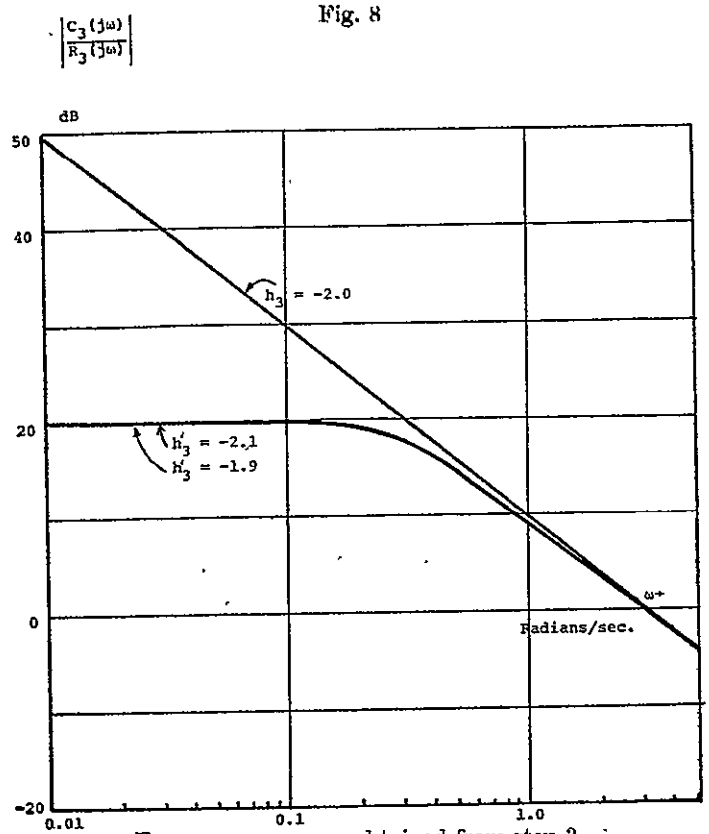


Frequency response obtained from step 2.

Following step 4 and comparing the frequency response curves of fig. 9, we obtain the required $h_4 = -3$. Based on fig. 9 we observe that when $h_4 = -3$ the magnitude in dB on a Bode plot is zero but if $h_4 - h_4' \neq 0$ a straight line with slope -1 passes through the crossover frequency at $h_1 - h_4'$. Then we have completed the identifying process.

p.59

Fig. 8



Frequency response obtained from step 3.

The required unknown system is:

$$\frac{C(s)}{R(s)} = \cfrac{1}{1+\cfrac{1}{\cfrac{1}{s}+\cfrac{1}{-2+\cfrac{1}{\cfrac{-3}{s}}}}}$$

$$= \frac{-2s+6}{s^2+4s+6}. \tag{19}$$

3. Identification based on time-response data

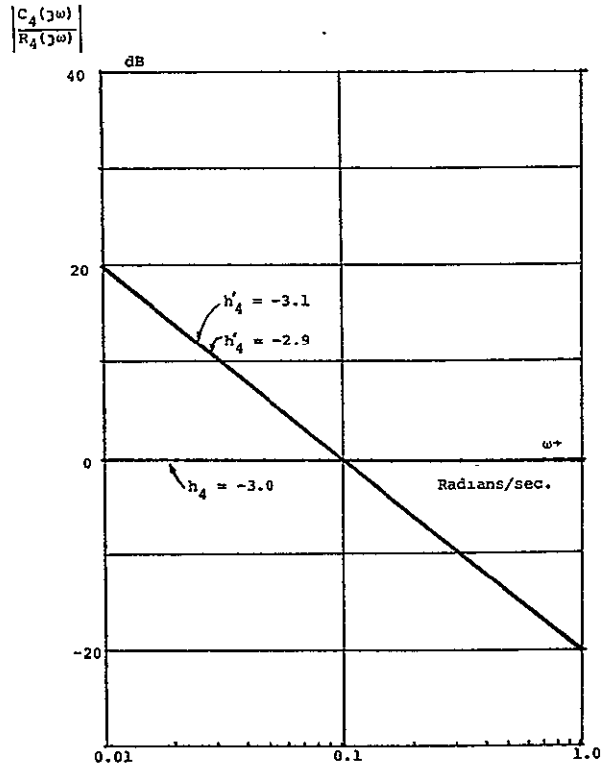
Suppose we have an unknown system which can be decomposed into continued fraction expansion as eqn. (5) and the block diagram is shown in fig. 1. The rational function corresponding to eqn. (5) is

$$\frac{C(s)}{R(s)} = \frac{(h_2+h_4)\,s+h_2h_3h_4}{s^2+(h_1h_2+h_1h_4+h_3h_4)\,s+h_1h_2h_3h_4}. \tag{20}$$

Our problem is to identify all these h's in the time domain. We have the following steps:

ORIGINAL PAGE IS
OF POOR QUALITY

Fig. 9



Frequency response obtained from step 4.

(1) *Identifying* h_1

A unit step function is applied to the unknown system or eqn. (20). The output equation can be written as:

$$C(s) = \frac{1}{s} \cdot \frac{(h_2+h_4)\,s + h_2 h_3 h_4}{s^2 + (h_1 h_2 + h_1 h_4 + h_3 h_4)\,s + h_1 h_2 h_3 h_4}. \tag{20 a}$$

Apply the final value theorem to eqn. (20 a) to find the steady-state value of $C(t)_{s.s}$, or:

$$\begin{aligned} C(t)_{s.s} &= \lim_{s\to 0} s \,.\, C(s) \\ &= \lim_{s\to 0} s \,.\, \frac{(h_2+h_4)\,s + h_2 h_3 h_4}{s[s^2 + (h_1 h_2 + h_1 h_4 + h_3 h_4)\,s + h_1 h_2 h_3 h_4]} \\ &= \frac{1}{h_1}. \end{aligned} \tag{21}$$

From eqn. (21) we find that the first feedback gain h_1 can be obtained by taking the reciprocal value of the final values of the unit step response, and adding a positive feedback gain h_1 to the unknown system. The block diagram

p. 61

is shown in fig. 2. The corresponding modified transfer function is:

$$\frac{C_1(s)}{R_1(s)} = \frac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\cfrac{h_4}{s}}} \tag{22}$$

$$= \frac{h_2}{s} + \frac{h_4}{s + h_3 h_4}.$$

(2) *Identifying* h_2

Taking the unit impulse function as an input and applying it to this modified system which is shown in fig. 2 the resulting transfer function becomes:

$$C_1(s) = \frac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\cfrac{h_4}{s}}}$$

or:

$$= \frac{h_2}{s} + \frac{h_4}{s + h_3 h_4}, \tag{23}$$

and the final value is measured. It is evident that the second term in eqn. (23) cannot contribute to the final value of the impulse response of eqn. (23). Only the first term decides the final value of the response. From the final value, we obtain the h_2 value. We then add an integrator with negative gain h_2 on the feed forward link as shown in fig. 3. The resulting transfer function becomes:

$$\frac{C_2(s)}{R_2(s)} = \cfrac{1}{h_3 + \cfrac{1}{\cfrac{h_4}{s}}}$$

or:

$$= \frac{h_4}{s + h_3 h_4}. \tag{24}$$

(3) *Identifying* h_3

A unit step input is applied to the modified system which is shown in fig. 3 or the corresponding eqn. (24) and the final value is measured. We then have:

$$C_2(t)s \,.\, s = \frac{1}{h_3}.$$

From the final value, we calculate its reciprocal value and get h_3. Again, following step 2, a positive feedback gain h_3 is added to the system of fig. 3. The corresponding new system is shown in fig. 4 and the resulting transfer function is:

$$\frac{C_3(s)}{R_3(s)} = \frac{h_4}{s}. \tag{25}$$

p.62

(4) *Identifying* h_4

Following step 2, applying unit impulse function to the modified system of fig. 4, the final value is the required h_4. Again, adding an integrator with negative gain h_4 to the system we will, at last, obtain a horizontal line only.

All h_i are determined. After substituting these values into eqn. (20) we get the desired transfer function of the unknown system.

Example 2

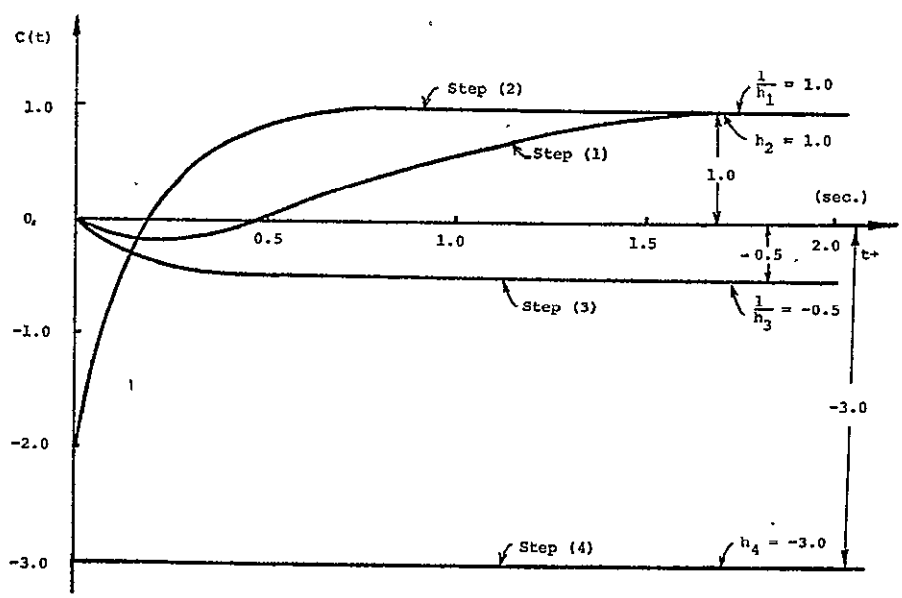
Consider the same example as shown in example 1. Rewrite eqns. (18):

$$\frac{C(s)}{R(s)} = \frac{-2s+6}{s^2+4s+6} \tag{26}$$

$$= \cfrac{1}{h_1 + \cfrac{1}{\cfrac{h_2}{s} + \cfrac{1}{h_3 + \cfrac{1}{\cfrac{h_4}{s}}}}}. \tag{26 a}$$

Assume that eqn. (26) is the unknown system to be identified. Following steps 1, 2, 3 and 4 we have the time response data shown in fig. 10. From the final value we can easily obtain the required unknown quotients of $h_1 = 1$, $h_2 = 1$, $h_3 = -2$, $h_4 = -3$.

Fig. 10



Time responses obtained from steps 1–4.

Substituting all of these h values into eqn. (26 b) we have:

$$\frac{C(s)}{R(s)} = \cfrac{1}{1+\cfrac{1}{\cfrac{1}{s}+\cfrac{1}{-2+\cfrac{1}{\cfrac{-3}{s}}}}} \tag{27}$$

or:

$$= \frac{-2s+6}{s^2+4s+6}. \tag{28}$$

4. Discussion

So far the examples we have discussed are only the second-order system with four unknown quotients; however, we can apply the same procedures for high-order systems. Theoretically, we can find all the h_i values of an unknown system. After all h_i factors have been found a method (Chen and Shieh 1969) of continued fraction inversion can be used to convert the continued fraction into a rational function.

To illustrate the above processes in a convenient way the response data, in the frequency and in the time domain of the above two examples, are obtained by calculations from the digital computer. Several assumptions were made to get the exact solution; for instance, in example 2, we have assumed that we can set up an ideal unit step function and unit impulse function as inputs and also that we should have prior knowledge that the linear system is constant coefficient and stable.

5. Conclusion

Based on the second Cauer form continued fraction expansion, two methods for identifying a transfer function are discussed. No prior knowledge of the order of the numerator and denominator is necessary. No complex pole problem is involved. It evaluates the parameters of the system from the frequency response data as well as that of time. Integrators are mainly used instead of differentiators. Due to the fact that the most important quotients can be identified first, these methods give us a satisfactory accuracy.

Acknowledgment

The authors wish to thank the National Aeronautics and Space Administration for support of this work under Grant NGL 44–005–084.

References

Bush, V., and Caldwell, S. H., 1945, *J. Franklin Inst.*, **240**, 255.

Chen, C. F., and Knox, A. E., 1968, *Int. J. Electronics*, **24**, 521.

Chen, C. F., and Philip, B. L., 1965, *I.E.E.E. Trans. autom. Control*, **10**, 356.

Chen, C. F., and Shieh, L. S., 1968, *Int. J. Control*, **8**, 561; 1969, *I.E.E.E. Trans. Circuit Theory*, **16**, 197.

Towill, D. R., and Mehdi, Z., 1970, *Measurement Contr.*, **3**, T1.

Project

(A) Project Title: Analysis and Design of Multivariable Systems via the Krylov Transformation

(B) Project Abstract:

The analysis and design of automatic control systems may be divided into the classical and the modern approach. The modern or state variable approach has been the result of early work by Kalman, Bucy, and others. This modern approach has received much attention in recent literature and in academic circles. However, the classical methods of Bode, Black, and others are more often used to approach practical problems.

This project bridges this gap via the use of the Krylov transformation. A new insight into Laplace transform inversion is developed. The method of the Heaviside expansion is extended to the modern case. The relationship between Liapunov functions in various reference frames is given. Finally, the power and flexibility of the Krylov transformation is used to design a system in the modern sense to classical design criteria.

Several example problems are worked to illustrate the principles involved.

(C) Publication: Ph.D. Dissertation in Electrical Engineering

(D) Year: 1972

(E) Department: Electrical Engineering

(F) Student Name: Robert E. Yates

(G) Faculty Advisor: Professor C. F. Chen

Project

(A) Project Title: On Identifying Transfer Functions and State Equations for Linear Systems

(B) Project Abstract:

Two methods are established for identifying constant-coefficient C^{2n}-type noise-free linear systems if the time response data of the input-output or of all states are known. 2n response data are required to identify an nth-order transfer function or state equation for an unknown linear system. The order of the unknown system can be identified by checking a sequence of determinants. The Z transform and its inversion are mainly used.

(C) Publication: IEEE Transactions on Aerospace and Electronics Systems, AES 8, No. 6, p. 811 (1972)

(D) Year: 1972

(E) Department: Electrical Engineering

(F) Student Name: L. S. Shieh

(G) Faculty Advisor: Professors C. F. Chen and C. J. Huang

On Identifying Transfer Functions and State Equations for Linear Systems

L. S. SHIEH, Member, IEEE
C. F. CHEN, Senior Member, IEEE
C. J. HUANG
University of Houston
Houston, Tex. 77004

Abstract

Two methods are established for identifying constant-coefficient C^{2n}-type noise-free linear systems if the time response data of the input-output or of all states are known. $2n$ response data are required to identify an nth-order transfer function or state equation for an unknown linear system. The order of the unknown system can be identified by checking a sequence of determinants. The Z transform and its inversion are mainly used.

Manuscript received July 28, 1970; Released for publication June 15, 1972

The work presented in this paper was supported by NASA under Grant NGL 44-005-084.

The determination of transfer function coefficients of a linear system from the system impulse response was shown by Kekre [1]. Bellman, Kagiwada and Kalaba [2] applied a method of Legendre-Gauss quadrature approximation to identify linear systems from observed samples the input and output. Later, Cook, Denman, and Carr [3] proposed an alternate method by using Laguerre-Gauss quadrature approximation for the same problem. These methods lead one to the idea that the Laplace transform is applicable not only to certain classes of explicit time functions, but also to sampled time response data which are obtained from Laplace transformable implicit time functions. This paper proposes two methods: one for identifying transfer functions in "s" from the input-output time response data, the other for identifying state equations in the time domain from the zero input time response data.

II. Identifying Transfer Functions from Time Response Data

Consider a constant-coefficient, C^{2n}-type noise-free linear system whose transfer function is

$$X(s) = \frac{b_1 s^{n-1} + b_2 s^{n-2} + \cdots + b_n}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \cdots + a_n}. \tag{1}$$

An alternate way to represent (1) is

$$x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_n x = 0 \tag{2}$$

with a set of initial conditions

$$\begin{bmatrix} x(0) \\ x'(0) \\ \vdots \\ x^{n-1}(0) \end{bmatrix}, \tag{3}$$

where $a_i, b_i, i = 1, 2, \cdots, n$ are constants.

Differentiate (2) $n - 1$ times and the resulting equations become

$$x^{n+1} + a_1 x^n + a_2 x^{n-1} + \cdots + a_n x' = 0$$

$$x^{n+2} + a_1 x^{n+1} + a_2 x^n + \cdots + a_n x'' = 0$$

ORIGINAL PAGE IS OF POOR QUALITY

$$x^{2n-1} + a_1 x^{2n-2} + a_2 x^{2n-3} + \cdots + a_n x^{n-1} = 0. \tag{4}$$

J. 68

ORIGINAL PAGE IS
OF POOR QUALITY

Rearrange (2) and (4) into a matrix form:

$$[\mathbf{x}] = (-1)\,[X]\,[\mathbf{a}] \tag{5}$$

where

$$[\mathbf{x}] = \begin{bmatrix} x^n \\ x^{n+1} \\ \cdot \\ \cdot \\ \cdot \\ x^{2n-1} \end{bmatrix}, \qquad [\mathbf{a}] = \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{bmatrix}$$

and

$$[X] = \begin{bmatrix} x^{n-1} & x^{n-2} & \cdots & x'x \\ x^n & x^{n-1} & \cdots & x''x' \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ x^{2n-2} & x^{2n-3} & \cdots & x^n x^{n-1} \end{bmatrix}. \tag{6}$$

Both sides of (5) are premultiplied by the inverse matrix $[X]$ of (6); then the $n \times 1$ constant vector $[\mathbf{a}]$ can be evaluated as follows:

$$\begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{bmatrix} = (-1) \begin{bmatrix} x^{n-1} & x^{n-2} & \cdots & x'x \\ x^n & x^{n-1} & \cdots & x''x' \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ x^{2n-2} & x^{2n-3} & \cdots & x^n x^{n-1} \end{bmatrix}^{-1} \begin{bmatrix} x^n \\ x^{n+1} \\ \cdot \\ \cdot \\ \cdot \\ x^{2n-1} \end{bmatrix}. \tag{7}$$

If all the elements of $[X]$ and $[\mathbf{x}]$ in (7) at specific time are known, then the denominator coefficient of a transfer function can be immediately obtained.

The order of the unknown system can be determined by checking the determinants of partitioned matrices from $[X]$ in (6); in other words, $j \times j$, $j = 2, 3, \cdots, n$ square matrices can be sequently partitioned from $[X]$ by taking from the upper right-hand corner of $[X]$ in (6). If one of these $j \times j$ square matrices is the largest matrix such that its determinant is nonzero, then the order of this unknown system is j. This is because a jth-order differential equation with j terms' unknown coefficients requires only j linear independent equations.

Stanley [4] linked the relationship between both coefficients of the denominator and numerator of a transfer function with a set of initial conditions. The relation is

$$\begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ a_1 & 1 & \cdots & 0 \\ a_2 & a_1 & 1 \cdots & 0 \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ a^{n-1} & \cdots & a_1 & 1 \end{bmatrix} \begin{bmatrix} x(0) \\ x'(0) \\ \cdot \\ \cdot \\ \cdot \\ x^{n-1}(0) \end{bmatrix}. \tag{8}$$

We expand (8) and rearrange it into the following form:

$$\begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix} = \begin{bmatrix} x(0) & 0 & \cdots & 0 \\ x'(0) & x(0) & \cdots & 0 \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ x^{n-1}(0) & \cdots & x'(0) & x(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \cdot \\ \cdot \\ \cdot \\ a_{n-1} \end{bmatrix}. \tag{9}$$

Substituting (7) into (9), we obtain the complete transfer function.

In the above description an assumption has been made that we can differentiate the given response curve $2n - 1$ times such that $2n$ values at a specific time can be obtained. By using these values we can formulate (7) and (9); then the required transfer function can be constructed. However, any numerical differentiation will generate errors and the result will be inaccurate. In other words, if we can obtain all the information from this given response curve without any numerical modification, we will get the correct solutions. The Z transform and its inversion are used to achieve this goal.

p.69

Consider a response curve $x(t)$ is given; a sample function with sampling period T is added to the response curve. The sampled function $x(t)$ can be written as $x^*(t)$. Then,

$$x^*(t) = x(0)\delta(t) + x(T)\delta(t - T) + x(2T)\delta(t - 2T) + \cdots. \quad (10)$$

Performing the Laplace transform on (10), we have

$$X^*(s) = x(0)e^{o} + x(T)e^{-Ts} + x(2T)e^{-2Ts} + \cdots. \quad (10a)$$

Following the definition of the Z transform by letting

$$s = \frac{1}{T} lnz \quad \text{or } e^{Ts} = z, \quad (10b)$$

and substituting (10b) into (10a), we have

$$X^*\left(\frac{1}{T} lnz\right) = X(z) = x(0)z^{o} + x(T)z^{-1} + x(2T)z^{-2} + \cdots + x(\overline{n-1}T)z^{-(n-1)} + \cdots. \quad (11)$$

Equation (11) can be considered as the result of the long division of two polynomials, or

$$X(z) = \frac{z(e_1 z^{n-1} + e_2 z^{n-2} + \cdots + e_n)}{z^n + d_1 z^{n-1} + d_2 z^{n-2} + \cdots + d_n} \quad (12)$$

where d_i and e_i, $i = 1, 2, \cdots, n$ are unknown constants to be determined. Equation (12) can be represented by a difference equation as follows:

$$x(t + nT) + d_1 x(t + \overline{n-1}T) + d_2 x(t + \overline{n-2}T) + \cdots + d_n x(t) = 0 \quad (13)$$

with a set of discrete points

$$\begin{bmatrix} x(0) \\ x(T) \\ \cdot \\ \cdot \\ \cdot \\ x(\overline{n-1}T) \end{bmatrix}. \quad (14)$$

Applying the E operator $n - 1$ times to (14), we have

$$x(t + \overline{n+1}T) + d_1 x(t + nT) + d_2 x(t + \overline{n-1}T) + \cdots + d_n x(t + T) = 0$$

$$x(t + \overline{n+2}T) + d_1 x(t + \overline{n+1}T) + d_2 x(t + nT) + \cdots + d_n x(t + 2T) = 0$$

$$\vdots$$

$$x(t + \overline{2n-1}T) + d_1 x(t + \overline{2n-2}T) + d_2 x(t + \overline{2n-3}T) + \cdots + d_n x(t + \overline{n-1}T) = 0. \quad (15)$$

We then rearrange (13) and (15) into a matrix form, and set $t = 0$, which yields

$$\begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ d_n \end{bmatrix} = (-1) \begin{bmatrix} x(\overline{n-1}T) & x(\overline{n-2}T) & \cdots & x(T) & x(0) \\ x(nT) & x(\overline{n-1}T) & \cdots & x(2T) & x(T) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x(\overline{2n-2}T) & x(\overline{2n-3}T) & \cdots & x(nT) & x(\overline{n-1}T) \end{bmatrix}^{-1} \begin{bmatrix} x(nT) \\ x(\overline{n+1}T) \\ \cdot \\ \cdot \\ x(\overline{2n-1}T) \end{bmatrix}. \quad (16)$$

The compact form will read

$$[\mathbf{d}] = (-1)\,[X_z]^{-1}\,[\mathbf{x}_z] \tag{17}$$

where

$[X_z] =$

ORIGINAL PAGE IS OF POOR QUALITY

$$\left[\begin{array}{ccc:cc} x(\overline{n-1}T) & \cdots & x(2T) & x(T) & x(0) \\ x(nT) & \cdots & x(3T) & x(2T) & x(T) \\ \vdots & & & & \\ \vdots & \cdots & x(4T) & x(3T) & x(2T) \\ \vdots & & & & \vdots \\ x(\overline{2n-2}T) & \cdots & x(\overline{n+1}T) & x(nT) & x(\overline{n-1}T) \end{array}\right] \tag{18}$$

The coefficients of the numerator of (12) can be obtained by the Chen and Shieh [5] algorithm, or

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} x(0) & 0 & \cdots & 0 \\ x(T) & x(0) & \cdots & \vdots \\ x(2T) & x(T) & \cdots & \vdots \\ \vdots & & & \vdots \\ x(\overline{n-1}T) & \cdots & x(T) & x(0) \end{bmatrix} \begin{bmatrix} 1 \\ d_1 \\ \vdots \\ d_{n-1} \end{bmatrix} \tag{19}$$

Again, the order of this unknown system can be obtained by checking the determinants of the square matrices which are taken from the upper right-hand corner of $[X_z]$ of (18). Finally, we have the required pulse transfer function

$$X(z) = \frac{z(e_1 z^{n-1} + e_2 z^{n-2} + \cdots + e_n)}{z^n + d_1 z^{n-1} + d_2 z^{n-2} + \cdots + d_n}. \tag{20}$$

Since we have prior knowledge that the system is a C^{2n}-type noise-free system, the required transfer function in "s" can be obtained by the Z transform and its inverse transform [6]. One of the most commonly used transform pairs is

$$\frac{C_l z}{z - \lambda_l} \rightarrow \frac{C_l}{s - \frac{1}{T} ln\lambda_l}. \tag{21}$$

Taking the partial fraction expansion of (20) yields

$$X(z) = \frac{C_1 z}{z - \lambda_1} + \frac{C_2 z}{z - \lambda_2} + \cdots + \frac{C_n z}{z - \lambda_n} \tag{22}$$

where the λ_i, $i = 1, 2, \cdots, n$ are the eigenvalues of the characteristic equation of (20), and C_i are the residues corresponding to poles λ_i. Apply the transform pair of (21) to (22); the required transfer function then is

$$X(s) = \frac{C_1}{s - \frac{1}{T} ln\lambda_1} + \frac{C_2}{s - \frac{1}{T} ln\lambda_2} + \cdots + \frac{C_n}{s - \frac{1}{T} ln\lambda_n}$$

$$= \frac{b_1 s^{n-1} + b_2 s^{n-2} + \cdots + b_n}{s^n + a_1 s^{n-1} + a_2 s^{n-2} + \cdots + a_n}. \tag{23}$$

Example 1

A $\sin \omega t$ function with $\omega = 1$ is applied as an input to an unknown system. The time response data with sampling period $T = \pi/2$ are recorded as follows:

$$x^*(t) = 1\delta(t) + 0\delta\left(t - \frac{\pi}{2}\right) + 1\delta(t - \pi) + 2\delta\left(t - \frac{3\pi}{2}\right) + 1\delta(t - 2\pi) + 0\delta\left(t - \frac{5\pi}{2}\right) + 1\delta(t - 3\pi) + \cdots. \tag{24}$$

The transfer function in "s" of this unknown system is required.

The following steps are followed.

1) Write the discrete equation for (24),

$$X(z) = x(0)z^0 + x(T)z^{-1} + x(2T)z^{-2} + x(3T)z^{-3} + \cdots$$

$$= 1z^0 + 0z^{-1} + 1z^{-2} + 2z^{-3} + 1z^{-4} + 0z^{-5} + \cdots \tag{25}$$

where the sampling period $T = \pi/2$.

2) Construct the $[X_z]$ matrix and check the order of this transfer function. The $[X_z]$ matrix is

$$[X_z] = \left[\begin{array}{c:c:cc} \cdot & x(3T) & x(2T) & x(T) & x(0) \\ \cdot & x(4T) & x(3T) & x(2T) & x(T) \\ \cdot & x(5T) & x(4T) & x(3T) & x(2T) \\ \cdot & x(6T) & x(5T) & x(4T) & x(3T) \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right]. \tag{26}$$

7.71

Checking the determinants of $j \times j, j = 2, 3, \cdots, n$ square matrices which are taken from the upper right-hand corner of $[X_z]$ of (26), if $j = 2$,

$$\det \begin{bmatrix} x(T) & x(0) \\ x(2T) & x(T) \end{bmatrix} = \det \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \neq 0,$$

and if $j = 3$,

$$\det \begin{bmatrix} x(2T) & x(T) & x(0) \\ x(3T) & x(2T) & x(T) \\ x(4T) & x(3T) & x(2T) \end{bmatrix} = \det \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \neq 0.$$

When j is higher than 3, and all the determinants obtained from the upper right-hand corner of $[X_z]$ in (26) are zero, we conclude that the order of this transfer function is 3 or $n = 3$.

3) Substitute $2n$ discrete values of $x^*(t)$ into (16) and (19), giving

$$\begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} = (-1) \begin{bmatrix} x(2T) & x(T) & x(0) \\ x(3T) & x(2T) & x(T) \\ x(4T) & x(3T) & x(2T) \end{bmatrix}^{-1} \cdot \begin{bmatrix} x(3T) \\ x(4T) \\ x(5T) \end{bmatrix}$$

$$= (-1) \begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}$$

and

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} x(0) & 0 & 0 \\ x(T) & x(0) & 0 \\ x(2T) & x(T) & x(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}.$$

The pulse-output function is

$$X(z) = \frac{z(e_1 z^2 + e_2 z + e_3)}{z^3 + d_1 z^2 + d_2 z + d_3} = \frac{z(z^2 - z + 2)}{z^3 - z^2 + z - 1}. \tag{27}$$

4) Transform (27) from the z plane to the s plane by performing (21). The required output function is

$$X(s) = \frac{s^2 - s + 1}{s(s^2 + 1)}. \tag{28}$$

5) Since the input function $R(s)$ of this unknown system is $(1/s^2 + 1)$, the required transfer function for this unknown system is

$$\frac{X(s)}{R(s)} = \frac{s^2 - s + 1}{s}. \tag{29}$$

III. Identifying Zero-Input State Equations from Time Response Data

The same methods mentioned above can be extended to a system whose zero input response data, at each state, is known. By observing the time response data, the state equation in the time domain can be identified.

Consider a system whose state equation is

$$[\dot{\mathbf{x}}] = [A]\,[\mathbf{x}] \tag{30}$$

where

$$[\dot{\mathbf{x}}] = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \dot{x}_n \end{bmatrix}, \quad [\mathbf{x}] = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

and

$$[A] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}. \tag{31}$$

Differentiate (30) $n - 1$ times and rearrange into a matrix form. We then have

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 & \dot{x}_1 & \cdots & x_1^{n-1} \\ x_2 & \dot{x}_2 & \cdots & x_2^{n-1} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ x_n & \dot{x}_n & \cdots & x_n^{n-1} \end{bmatrix}$$

$$= \begin{bmatrix} \dot{x}_1 & x_1'' & \cdots & x_1^n \\ \dot{x}_2 & x_2'' & \cdots & x_2^n \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \dot{x}_n & x_n'' & \cdots & x_n^n \end{bmatrix}. \tag{32}$$

The unknown coefficient matrix $[A]$ can be obtained by rearranging (32), giving

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} \dot{x}_1 & x_1'' & \cdots & x_1^n \\ \dot{x}_2 & x_2'' & \cdots & x_2^n \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \dot{x}_n & x_n'' & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} x_1 & \dot{x}_1 & \cdots & x_1^{n-1} \\ x_2 & \dot{x}_2 & \cdots & x_2^{n-1} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ x_n & \dot{x}_n & \cdots & x_n^{n-1} \end{bmatrix}^{-1}. \tag{33}$$

If all x_i^j values $i = 1, \cdots, n, j = 0, 1, \cdots, n$ can be evaluated, then the unknown constant coefficient matrix $[A]$ can be obtained. But we are interested in using the given information $x_i(t), i = 1, \cdots, n$ only. Again, the Z transform and its inverse transform [6] can be applied.

Based on the given $x_i(t), i = 1, \cdots, n$ curves we can express this system by the following discrete state equation:

$$[\mathbf{x}(\overline{k+1}T)] = [D][\mathbf{x}(kT)] \tag{34}$$

or

$$\begin{bmatrix} x_1(\overline{k+1}T) \\ x_2(\overline{k+1}T) \\ \cdot \\ \cdot \\ \cdot \\ x_n(\overline{k+1}T) \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \begin{bmatrix} x_1(kT) \\ x_2(kT) \\ \cdot \\ \cdot \\ \cdot \\ x_n(kT) \end{bmatrix}. \tag{35}$$

Apply the E operator on (35) $n-1$ times; then

$$\begin{bmatrix} x_1(\overline{k+2}T) \\ x_2(\overline{k+2}T) \\ \cdot \\ \cdot \\ \cdot \\ x_n(\overline{k+2}T) \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \begin{bmatrix} x_1(\overline{k+1}T) \\ x_2(\overline{k+1}T) \\ \cdot \\ \cdot \\ \cdot \\ x_n(\overline{k+1}T) \end{bmatrix}$$

$$\begin{bmatrix} x_1(\overline{k+n}T) \\ x_2(\overline{k+n}T) \\ \cdot \\ \cdot \\ \cdot \\ x_n(\overline{k+n}T) \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

p. 73

$$\begin{bmatrix} x_1(\overline{k+n-1}T) \\ x_2(\overline{k+n-1}T) \\ \cdot \\ \cdot \\ \cdot \\ x_n(\overline{k+n-1}T) \end{bmatrix}. \quad (36)$$

Rearranging (35) and (36), we have

$$\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} x_1(\overline{k+1}T) & x_1(\overline{k+2}T) & \cdots & x_1(\overline{k+n}T) \\ x_2(\overline{k+1}T) & x_2(\overline{k+2}T) & \cdots & x_2(\overline{k+n}T) \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ x_n(\overline{k+1}T) & x_n(\overline{k+2}T) & \cdots & x_n(\overline{k+n}T) \end{bmatrix}$$

$$\begin{bmatrix} x_1(kT) & x_1(\overline{k+1}T) & \cdots & x_1(\overline{k+n-1}T) \\ x_2(kT) & x_2(\overline{k+1}T) & \cdots & x_2(\overline{k+n-1}T) \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ x_n(kT) & x_n(\overline{k+1}T) & \cdots & x_n(\overline{k+n-1}T) \end{bmatrix}^{-1}. \quad (37)$$

$[D]$ can be obtained by substituting the sampled values of $X_i^*(t), i = 1, \cdots, n$ into (37).

Rewrite (34):

$$[\mathbf{x}(\overline{k+1}T)] = [D]\,[\mathbf{x}(kT)]. \quad (38)$$

This time we take the Z transform on (38), which yields

$$z[\mathbf{X}(z) - \mathbf{x}(0)] = [D]\,[\mathbf{X}(z)]$$

or

$$[\mathbf{X}(z)] = z[zI - D]^{-1}\,[\mathbf{x}(0)]. \quad (39)$$

Evaluating $z[zI - D]^{-1}$, the resulting matrix can be shown as follows:

$$z[zI - D]^{-1}$$

$$= z\begin{bmatrix} \sum_{j=1}^{n} \frac{b_{11j}}{z-\lambda_j} & \sum_{j=1}^{n} \frac{b_{12j}}{z-\lambda_j} & \cdots & \sum_{j=1}^{n} \frac{b_{1nj}}{z-\lambda_j} \\ \sum_{j=1}^{n} \frac{b_{21j}}{z-\lambda_j} & \sum_{j=1}^{n} \frac{b_{22j}}{z-\lambda_j} & \cdots & \sum_{j=1}^{n} \frac{b_{2nj}}{z-\lambda_j} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \sum_{j=1}^{n} \frac{b_{n1j}}{z-\lambda_j} & \cdot & \cdots & \sum_{j=1}^{n} \frac{b_{nnj}}{z-\lambda_j} \end{bmatrix} \quad (40)$$

where λ_j is an eigenvalue of $|zI - D|$, which can be distinct or repeated. b_{ilj} located at the ith row and lth column of the $[zI - D]^{-1}$ matrix is a residue corresponding to pole λ_j. Again, using (21), the resulting matrix yields

$$\begin{bmatrix} X_1(s) \\ X_2(s) \\ \vdots \\ X_n(s) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^{n} \frac{b_{11j}}{s - \frac{1}{T}\ln\lambda_j} & \sum_{j=1}^{n} \frac{b_{12j}}{s - \frac{1}{T}\ln\lambda_j} & \cdots & \sum_{j=1}^{n} \frac{b_{1nj}}{s - \frac{1}{T}\ln\lambda_j} \\ \sum_{j=1}^{n} \frac{b_{21j}}{s - \frac{1}{T}\ln\lambda_j} & \cdot & \cdots & \sum_{j=1}^{n} \frac{b_{2nj}}{s - \frac{1}{T}\ln\lambda_j} \\ \vdots & & & \\ \sum_{j=1}^{n} \frac{b_{n1j}}{s - \frac{1}{T}\ln\lambda_j} & \cdot & \cdots & \sum_{j=1}^{n} \frac{b_{nnj}}{s - \frac{1}{T}\ln\lambda_j} \end{bmatrix}$$

$$\begin{bmatrix} x_1(0) \\ x_2(0) \\ \vdots \\ x_n(0) \end{bmatrix}. \quad (41)$$

The compact form of (41) is

$$[X(s)] = [sI - A]^{-1}[x(0)] \tag{42}$$

where $[A]$ is the required system constant matrix such that

$$[\dot{x}] = [A]\,[x]. \tag{43}$$

Comparing (41) with (42), the transition matrix $[\Phi(t)]$ of (43) can be written as

$$[\Phi(t)] = L^{-1}[sI - A]^{-1}$$

$$= \begin{bmatrix} \sum_{j=1}^{n} b_{11j}e^{(t/T)\ln\lambda_j} & \sum_{j=1}^{n} b_{12j}e^{(t/T)\ln\lambda_j} & \cdots & \sum_{j=1}^{n} b_{1nj}e^{(t/T)\ln\lambda_j} \\ \sum_{j=1}^{n} b_{21j}e^{(t/T)\ln\lambda_j} & \cdot & \cdots & \sum_{j=1}^{n} b_{2nj}e^{(t/T)\ln\lambda_j} \\ \vdots & & & \\ \sum_{j=1}^{n} b_{n1j}e^{(t/T)\ln\lambda_j} & \cdot & \cdots & \sum_{j=1}^{n} b_{nnj}e^{(t/T)\ln\lambda_j} \end{bmatrix}. \tag{44}$$

The solution of (43) is

$$[x] = [\Phi(t)]\,[x(0)]. \tag{45}$$

From (43) and (45), we can easily find that

$$[A] = [\dot{\Phi}(t)]\,[\Phi(t)]^{-1}. \tag{46}$$

At this point, we can make the following observation: The $[A]$ matrix can be directly formulated from equation (40), by using the Z transform and its inverse transform table [6], or by performing the transform pair

$$\frac{bz}{z - \lambda_j} \rightarrow \frac{b}{T}\ln\lambda_j. \tag{47}$$

The required $[A]$ matrix

$$[A] = \begin{bmatrix} \sum_{j=1}^{n} \frac{b_{11j}\ln\lambda_j}{T} & \sum_{j=1}^{n} \frac{b_{12j}\ln\lambda_j}{T} & \cdots & \sum_{j=1}^{n} \frac{b_{1nj}\ln\lambda_j}{T} \\ \sum_{j=1}^{n} \frac{b_{21j}\ln\lambda_j}{T} & \cdot & \cdots & \sum_{j=1}^{n} \frac{b_{2nj}\ln\lambda_j}{T} \\ \vdots & & & \\ \sum_{j=1}^{n} \frac{b_{n1j}\ln\lambda_j}{T} & \cdot & \cdots & \sum_{j=1}^{n} \frac{b_{nnj}\ln\lambda_j}{T} \end{bmatrix}. \tag{48}$$

Recall that T is a sampling period.

The required state equation is

$$[\dot{x}] = [A]\,[x]. \tag{49}$$

Example 2

ORIGINAL PAGE IS OF POOR QUALITY

A set of initial conditions

$$\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

is added to an unknown system to generate the zero input response. The response data with sampling period T is recorded as follows:

$$\begin{bmatrix} x_1(T) \\ x_2(T) \end{bmatrix} = \begin{bmatrix} -3e^{-T} + 4e^{-2T} \\ 3e^{-T} - 2e^{-2T} \end{bmatrix},$$

$$\begin{bmatrix} x_1(2T) \\ x_2(2T) \end{bmatrix} = \begin{bmatrix} -3e^{-2T} + 4e^{-4T} \\ 3e^{-2T} - 2e^{-4T} \end{bmatrix}. \tag{50}$$

The constant coefficient matrix $[A]$ and the transition matrix $[\Phi(t)]$ in (49) are required.

First, follow (37) to obtain the $[D]$ matrix,

$$[D] = \begin{bmatrix} -3e^{-T} + 4e^{-2T} & -3e^{-2T} + 4e^{-4T} \\ 3e^{-T} - 2e^{-2T} & 3e^{-2T} - 2e^{-4T} \end{bmatrix} \cdot \begin{bmatrix} 1 & -3e^{-T} + 4e^{-2T} \\ 1 & 3e^{-T} - 2e^{-2T} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 2e^{-2T} - e^{-T} & 2e^{-2T} - 2e^{-T} \\ -e^{-2T} + e^{-T} & -e^{-2T} + 2e^{-T} \end{bmatrix}. \tag{51}$$

Then, following (40),

$$z[zI - D]^{-1}$$

$$= z\begin{bmatrix} z - (2e^{-2T} - e^{T}) & -2(e^{-2T} - e^{-T}) \\ e^{-2T} - e^{T} & z + (e^{-2T} - 2e^{-T}) \end{bmatrix}^{-1}$$

$$= z\begin{bmatrix} \frac{2}{z - e^{-2T}} + \frac{-1}{z - e^{-T}} & \frac{2}{z - e^{-2T}} + \frac{-2}{z - e^{-T}} \\ \frac{-1}{z - e^{-2T}} + \frac{1}{z - e^{-T}} & \frac{-1}{z - e^{-2T}} + \frac{2}{z - e^{-T}} \end{bmatrix}. \tag{52}$$

p. 75

The required $[A]$ matrix can be obtained by using (46) or (48). The result is

$$[A] = \begin{bmatrix} 2\frac{1}{T}ln(e^{-2T}) - \frac{1}{T}ln(e^{-T}) & 2\frac{1}{T}ln(e^{-2T}) - \frac{2}{T}ln(e^{-T}) \\ -\frac{1}{T}ln(e^{-2T}) + \frac{1}{T}ln(e^{-T}) & -\frac{1}{T}ln(e^{-2T}) + \frac{2}{T}ln(e^{-T}) \end{bmatrix}$$

$$= \begin{bmatrix} -3 & -2 \\ 1 & 0 \end{bmatrix}.$$

The state equation is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -3 & -2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{53}$$

By using (44) and (52), the transition matrix is

$$[\Phi(t)] = \begin{bmatrix} 2e^{-2t} - e^{-t} & 2e^{-2t} - 2e^{-t} \\ -e^{-2t} + e^{-t} & -e^{-2t} + 2e^{-t} \end{bmatrix}.$$

IV. Discussion

So far the methods we have discussed involve a multiple-valued logarithmic function. Unless the branch cut of this function can be determined, many solutions can be obtained. How to choose a suitable sampling period T so that the principal values of this multiple-valued function can be used is a very complicated problem. In sampled data systems many people attacked this problem; for example, Mitchell and McDaniel's [7] adaptive sampling technique and Taits [8] sampling criteria, etc.

For the following numerical example, $T = 0.02$ is used. Given:

$$\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 66.7 \end{bmatrix}. \tag{54}$$

$$\begin{bmatrix} x_1(T) \\ x_2(T) \end{bmatrix} = \begin{bmatrix} 1.24 \\ 57.075 \end{bmatrix}$$

$$\begin{bmatrix} x_1(2T) \\ x_2(2T) \end{bmatrix} = \begin{bmatrix} 2.2759 \\ 46.383 \end{bmatrix}$$

and

$$\begin{bmatrix} x_1(3T) \\ x_2(3T) \end{bmatrix} = \begin{bmatrix} 3.092 \\ 35.183 \end{bmatrix}. \tag{54}$$

Following the procedures mentioned above, we obtain the required state equation as follows:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -0.000087 & 0.9999769 \\ -106.53339 & -6.669595 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{55}$$

The original generating equation is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -106.5 & -6.67 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{56}$$

Compared with (56), the answer is quite satisfactory.

V. Conclusion

Two methods, one for fitting transfer functions, the other for estimating state equations, are discussed through the use of the Z transform and its inversion.

References

[1] H. B. Kekre, "Further extension to a time-to-frequency domain matrix formulation," *Proc. IEEE*, vol. 57, pp. 1192-1195, June 1969.

[2] R. Bellman, H. Kagiwada, and R. Kalaba, "Identification of linear systems via numerical inversion of Laplace transforms," *IEEE Trans. Automatic Control*, vol. AC-10, pp. 111-112, January 1965.

[3] G. E. Cook, E. D. Denman, and H. M. Carr, "Numerical inversion of Laplace transforms by the Laguerre-Gauss quadrature approximation," *IEEE Trans. Automatic Control*, vol. AC-12, pp. 623-625, October 1967.

[4] W. D. Stanley, "A time-to-frequency domain matrix formulation," *Proc. IEEE*, vol. 52, pp. 874-875, July 1964.

[5] C. F. Chen and L. S. Shieh, "A single matrix formula for the closed form solution of the inverse Z transform," *Internatl. J. Electron.*, vol. 27, pp. 535-548, 1969.

[6] J. T. Tou, *Digital and Sampled Data Control Systems*. New York: McGraw Hill, 1959, pp. 588-592.

[7] J. R. Mitchell, and W. L. McDaniel, Jr., "Adaptive sampling technique," *IEEE Trans. Automatic Control*, vol. AC-14, pp. 200-201, April 1969.

[8] K. E. Tait, "Further sampling interval critera and their evaluation in discrete-continuous feedback control systems," *Internatl. J. Control*, vol. 4, no. 4, pp. 297-324, 1966.

ORIGINAL PAGE IS
OF POOR QUALITY

Leang-San Shieh (S'68—M'70) was born in Taiwan, China, on January 10, 1934. He received the B.S. degree in electrical engineering from the National Taiwan University in 1958, and the M.S. and Ph.D. degrees in electrical engineering from the University of Houston, Houston, Tex., in 1968 and 1970, respectively.

From 1958 to 1960 he served in the Chinese Navy ROTC. From 1960 to 1961 he was a Design Engineer at the Taiwan Flourescent Lamp Company, and from 1961 to 1965 he was a Systems Engineer at the Taiwan Electric Power Company. At present, he is an Assistant Professor of Electrical Engineering at the University of Houston.

Dr. Shieh is a member of Sigma Xi.

C. F. Chen (SM'61) was born in China. He received the B.S.E.E. degree from Peiyang University, the M.S. degree from the University of Pennsylvania, Philadelphia, and the Ph.D. degree from Cambridge University, England.

He has taught electrical engineering at Christian Brothers College, Memphis, Tenn. from 1957-64. Since 1966 he has been a Professor of Electrical Engineering at the University of Houston, Houston, Tex. His current research interests involve large-scale systems, model reductions, and missile guidance. He has also been a consultant to Battelle Laboratories for the Army Missile Command. He is the author of more than 40 technical papers on automatic control theory and computing sciences, and is co-author of the book, *Elements of Control Systems Analysis: Classical and Modern Approaches* (Prentice-Hall, 1968).

C. J. Huang received the B.Sc. degree in chemical engineering from National Taiwan University, Taiwan, China, in 1949, and the Ph.D. degree in chemical engineering from the University of Toronto, Canada, in 1955.

He has been with the University of Houston, Houston, Tex., since 1955, teaching and conducting research in the fields of mass transfer operation and computer process control. He has served as Chairman of the Department of Chemical Engineering and Associate Dean of Engineering. He is currently Assistant Vice President, Associate Dean of Faculties, and Professor of Chemical Engineering. He has published extensively in the above fields, and three of his publications were chosen for Best Publication Award of the Year by the South Texas Section of the American Institute of Chemical Engineers.

Project

(A) Project Title: State Space Approach to Mixed Boundary Value Problems

(B) Project Abstract:

A state space procedure for the formulation and solution of the mixed boundary value problems is established. It is a natural extension of the method used in the initial value problems; however, certain special theorems and rules should be developed. The scope of the applications of the approach includes the beam, arch, axisymmetrical shell problems in structural analysis, boundary layer problems in fluid mechanics and eigenvalue problems for deformable bodies, etc. Many classical methods in these fields developed by, for example, Holzer, Prohl, Myklestad, Thomson, Love-Meissner, etc. can be either simplified or unified under new light shed by the state variable approach. A beam problem is included as an illustration.

(C) Publication: International Journal of Control, 17, 863 (1973)

(D) Year: 1973

(E) Department: Electrical Engineering

(F) Student Name: M. M. Chen

(G) Faculty Advisor: Professor C. F. Chen

INT. J. CONTROL, 1973, VOL. 17, NO. 4, 863–868

State space approach to mixed boundary value problems†

C. F. CHEN
Department of Electrical Engineering, University of Houston, Houston, Texas

M. M. CHEN
Department of Aerospace Engineering, Boston University, Boston, Massachusetts

ORIGINAL PAGE IS
OF POOR QUALITY

[Received 24 April 1972]

A state space procedure for the formulation and solution of the mixed boundary value problems is established. It is a natural extension of the method used in the initial value problems; however, certain special theorems and rules should be developed. The scope of the applications of the approach includes the beam, arch, axisymmetrical shell problems in structural analysis, boundary layer problems in fluid mechanics and eigenvalue problems for deformable bodies, etc. Many classical methods in these fields developed by, for example, Holzer, Prohl, Myklestad, Thomson, Love–Meissner, etc. can be either simplied or unified under new light shed by the state variable approach. A beam problem is included as an illustration.

1. Introduction

The state space formulation for a one-dimensional linear system can be expressed as follows (Chen and Haas 1968):

$$[\dot{z}] = [A(x)][z] + [B(x)], \tag{1}$$

$$[z(b)] = [k], \tag{2}$$

where

$[z]$ = state vector,
$[A(x)]$ = property matrix of the system,
$[B(x)]$ = input vector,
$[k]$ = state vector evaluated at the boundary.

The solution of eqns. (1) and (2), when $[B(x)] = 0$, represents zero input response, and it is given as follows:

$$\begin{bmatrix} z_1(x) \\ z_2(x) \\ \vdots \\ z_n(x) \end{bmatrix} = \begin{bmatrix} \phi_{11}(x) & \dots & \phi_{1n}(x) \\ \vdots & & \vdots \\ \vdots & & \vdots \\ \phi_{n1}(x) & \dots & \phi_{nn}(x) \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ \vdots \\ c_n \end{bmatrix} \tag{3}$$

or simply

$$[z] = [\phi(x)][c], \tag{3 a}$$

where $[\phi(x)]$ is called the transition matrix and $[c]$ is a vector denoting the n arbitrary constants.

† Communicated by the Authors.

p. 79

For $[B(x)] \neq [0]$, the response appears to be

$$[z(x)] = [\phi(x)][c] + \int_0^x [\phi(x)][\phi(\tau)]^{-1}[B(\tau)]\, d\tau \tag{4}$$

in which the last term is the matrix form of a convolution integral.

For

$$[A(x)] = [A], \tag{5}$$

where $[A]$ is a constant property matrix, we have the following relationship for $[\phi(\tau)]$:

$$[\phi(\tau)]^{-1} = [\phi(-\tau)]. \tag{6}$$

Then (4) is rewritten as follows:

$$[z(x)] = [\phi(x)][c] + [u(x)], \tag{7}$$

where

$$[u(x)] = [\phi(x)] \int_0^x [\phi(\tau)]^{-1}[B(\tau)]\, d\tau. \tag{8}$$

2. Partition

By partitioning eqn. (7), we obtain

$$\begin{bmatrix} z_1 \\ \vdots \\ z_r \\ \hline z_{r+1} \\ \vdots \\ z_n \end{bmatrix} = \left[\begin{array}{ccc|ccc} \phi_{11} & \cdots & \phi_{1r} & \phi_{1,r+1} & \cdots & \phi_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ \phi_{r1} & \cdots & \phi_{rr} & \phi_{r,r+1} & \cdots & \phi_{rr} \\ \hline \phi_{r+1,1} & \cdots & \phi_{r+1,r} & \phi_{r+1,r+1} & \cdots & \phi_{r+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ \phi_{n1} & \cdots & \phi_{nr} & \phi_{n,n+1} & \cdots & \phi_{nn} \end{array}\right] \begin{bmatrix} c_1 \\ \vdots \\ c_r \\ \hline c_{r+1} \\ \vdots \\ c_n \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_r \\ \hline u_{r+1} \\ \vdots \\ u_n \end{bmatrix} \tag{9}$$

or briefly, we can write eqn. (9) as

$$\begin{bmatrix} Z_1 \\ \hline Z_2 \end{bmatrix} \left[\begin{array}{c|c} \Phi_{11} & \Phi_{12} \\ \hline \Phi_{21} & \Phi_{22} \end{array}\right] \begin{bmatrix} C_1 \\ \hline C_2 \end{bmatrix} + \begin{bmatrix} U_1 \\ \hline U_2 \end{bmatrix}. \tag{10}$$

At the boundaries, eqn. (10) can be written as, for $x = \alpha$

$$\begin{bmatrix} Z_1^\alpha \\ \hline Z_2^\alpha \end{bmatrix} = \left[\begin{array}{c|c} \Phi_{11}^\alpha & \Phi_{12}^\alpha \\ \hline \Phi_{21}^\alpha & \Phi_{22}^\alpha \end{array}\right] \begin{bmatrix} C_1 \\ \hline C_2 \end{bmatrix} + \begin{bmatrix} U_1^\alpha \\ \hline U_2^\alpha \end{bmatrix}, \tag{11}$$

and for $x = \beta$

$$\begin{bmatrix} Z_1^\beta \\ \hline Z_2^\beta \end{bmatrix} = \left[\begin{array}{c|c} \Phi_{11}^\beta & \Phi_{12}^\beta \\ \hline \Phi_{21}^\beta & \Phi_{22}^\beta \end{array}\right] \begin{bmatrix} C_1 \\ \hline C_2 \end{bmatrix} + \begin{bmatrix} U_1^\beta \\ \hline U_2^\beta \end{bmatrix}, \tag{12}$$

where $(\cdot)^\alpha$ and $(\cdot)^\beta$ designate the value of the appropriate matrix or vector $(\cdot)$ at $x = \alpha$ and $x = \beta$ respectively.

p.80

Eliminating $\begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$ from eqns. (11) and (12), we obtain either

ORIGINAL PAGE IS
OF POOR QUALITY

$$\left[\begin{array}{c} Z_1^\beta - U_1^\beta \\ \hline Z_2^\beta - U_2^\beta \end{array}\right] = \left[\begin{array}{c|c} \Phi_{11}^\beta & \Phi_{12}^\beta \\ \hline \Phi_{21}^\beta & \Phi_{22}^\beta \end{array}\right] \left[\begin{array}{c|c} \Phi_{11}^\alpha & \Phi_{12}^\alpha \\ \hline \Phi_{21}^\alpha & \Phi_{22}^\alpha \end{array}\right]^{-1} + \left[\begin{array}{c} Z_1^\alpha - U_1^\alpha \\ \hline Z_2^\alpha - U_2^\alpha \end{array}\right] \tag{13}$$

or

$$\left[\begin{array}{c} Z_1^\alpha - U_1^\alpha \\ \hline Z_2^\alpha - U_2^\alpha \end{array}\right] = \left[\begin{array}{c|c} \Phi_{11}^\alpha & \Phi_{12}^\alpha \\ \hline \Phi_{21}^\alpha & \Phi_{22}^\alpha \end{array}\right] \left[\begin{array}{c|c} \Phi_{11}^\beta & \Phi_{12}^\beta \\ \hline \Phi_{21}^\beta & \Phi_{22}^\beta \end{array}\right]^{-1} + \left[\begin{array}{c} Z_1^\beta - U_1^\beta \\ \hline Z_2^\beta - U_2^\beta \end{array}\right] \tag{14}$$

in which two of the matrices $[Z_1^\alpha]$, $[Z_2^\alpha]$, $[Z_1^\beta]$ and $[Z_2^\beta]$ are prescribed.

3. Classification and solution

Three cases in general can be classified.

Case I. $[Z_1^\alpha]$ and $[Z_2^\alpha]$ are known or $[Z_1^\beta]$ and $[Z_2^\beta]$ are given, the problem is degenerated to an initial value problem.

Case II. $[Z_1^\alpha]$ and $[Z_2^\beta]$ are known; a cantilever beam is a typical example. Then $[Z_1^\alpha]$ is interpretated as the displacement and slope of the beam at fixed end α and $[Z_2^\beta]$ is the bending moment and shear at free end β.

Case III. $[Z_1^\alpha]$ and $[Z_1^\beta]$ are known. A simply supported beam is a typical example. We can consider that $[Z_1^\alpha]$ and $[Z_1^\beta]$ are the displacement and bending moment of the beam at α and β respectively. Both Case II and Case III are classified as the mixed boundary value problems.

Because Case I is the initial value problem, the solution is readily seen as

$$\left[\begin{array}{c} Z_1 \\ \hline Z_2 \end{array}\right] = \left[\begin{array}{c|c} \Phi_{11} & \Phi_{12} \\ \hline \Phi_{21} & \Phi_{22} \end{array}\right] \left[\begin{array}{c|c} \Phi_{11}^\alpha & \Phi_{12}^\alpha \\ \hline \Phi_{21}^\alpha & \Phi_{22}^\alpha \end{array}\right]^{-1} \left[\begin{array}{c} K_1 - U_1^\alpha \\ \hline K_2 - U_2^\alpha \end{array}\right] + \left[\begin{array}{c} U_1 \\ \hline U_2 \end{array}\right], \tag{15}$$

where $\begin{bmatrix} K_1 \\ \hline K_2 \end{bmatrix} = [k]$, the given boundary conditions.

Then we consider Case II.

When $[Z_1^\alpha]$ and $[Z_2^\beta]$ are given and equal to $[K_1]$ and $[K_2]$ respectively, the corresponding solution can be derived from (10)–(13). The result is

$$\left[\begin{array}{c} Z_1 \\ \hline Z_2 \end{array}\right] = \left[\begin{array}{c|c} \Phi_{11} & \Phi_{12} \\ \hline \Phi_{21} & \Phi_{22} \end{array}\right] \left(\left[\begin{array}{c|c} \Psi_{11}^\alpha & \Psi_{12}^\alpha \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} K_1 - U_1^\alpha \\ \hline K_2 - U_2^\alpha \end{array}\right] + \left[\begin{array}{c|c} 0 & 0 \\ \hline \Psi_{21}^\beta & \Psi_{22}^\beta \end{array}\right] \left[\begin{array}{c} K_1 - U_1^\beta \\ \hline K_2 - U_2^\beta \end{array}\right] \right) + \left[\begin{array}{c} U_1 \\ \hline U_2 \end{array}\right], \tag{16}$$

where the $[\Psi]$ matrix is defined as follows:

$$\left(\left[\begin{array}{c|c} \Phi_{11} & \Phi_{12} \\ \hline \Phi_{21} & \Phi_{22} \end{array}\right]_\alpha \right)^{-1} = \left[\begin{array}{c|c} \Psi_{11}^\alpha & \Psi_{12}^\alpha \\ \hline \Psi_{21}^\alpha & \Psi_{22}^\alpha \end{array}\right] \tag{17}$$

p. 81

Case III is with the following conditions.

$$[Z_1^\alpha]=[K_1] \quad \text{and} \quad [Z_1^\beta]=[K_2] \quad \text{which are known.}$$

Matrices $[Z_2^\alpha]$ and $[Z_2^\beta]$ should be evaluated first. This can be done by the combination of (11), (12) and $[K_1]$ and $[K_2]$. We have

$$\begin{bmatrix} Z_2 \\ \hline Z_2 \end{bmatrix} = \left[\begin{array}{c|c} \Psi_{12}^\alpha & -\Psi_{12}^\beta \\ \hline \Psi_{22} & -\Psi_{22}^\beta \end{array}\right]^{-1} \left[\begin{array}{c|c|c|c} -\Psi_{11}^\alpha & \Psi_{11}^\beta & -\Psi_{12}^\alpha & \Psi_{12}^\beta \\ \hline -\Psi_{21}^\alpha & \Psi_{21}^\beta & -\Psi_{22}^\alpha & \Psi_{22}^\beta \end{array}\right] \begin{bmatrix} K_1-U_1^\alpha \\ \hline K_2-U_2^\beta \\ \hline 0-U_2^\alpha \\ \hline 0-U_2^\beta \end{bmatrix} \tag{18}$$

Substituting (18) into (11) and (12) we find C_1 and C_2 and then use (10) to obtain the solution.

4. Illustrative example

The state equation, eqn. (1), for a cantilever beam subjected to a lateral pressure loading, $p(x)$, is as follows :

$$[\dot{z}]=[A][z]+[b], \tag{1}$$

where

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} W \\ \theta \\ M \\ V \end{bmatrix}, \quad [A]=\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{EI} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad [b]=\begin{bmatrix} 0 \\ 0 \\ 0 \\ p(x) \end{bmatrix} \tag{19}$$

and EI denotes the bending stiffness. The state variable W, θ, M, V are displacement, slope, bending moment and shear respectively (Timoshenko and MacCullough 1940).

Taking the Laplace transform, we obtain

$$[Z(s)]=\Big(s[I]-[A]\Big)^{-1}([C]-[B(s)]) \tag{20}$$

or

$$[Z(s)]=\frac{1}{s^4}\begin{bmatrix} s^3 & s^2 & \frac{s}{EI} & 1 \\ 0 & s^3 & \frac{s^2}{EI} & \frac{s}{EI} \\ 0 & 0 & s^3 & s^2 \\ 0 & 0 & 0 & s^3 \end{bmatrix}([C]+[B(s)]). \tag{21}$$

The inverse Laplace transform of eqn. (21) yields the following :

$$[z(x)]=[\phi(x)][C]+[u(x)], \tag{22}$$

where

$$[u(x)]=[\phi(x)]\int_0^x [\phi(\tau)]^{-1}[b(\tau)]\,d\tau \tag{23}$$

p. 82

and

$$[\phi(x)]=\begin{bmatrix} 1 & x & \dfrac{x^2}{2EI} & \dfrac{x^3}{6EI} \\ 0 & 1 & \dfrac{x}{EI} & \dfrac{x^2}{2EI} \\ 0 & 0 & 1 & x \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{24}$$

For constant $[A]$ and uniform pressure p_0, eqn. (23) yields

ORIGINAL PAGE IS
OF POOR QUALITY

$$[u(x)]\triangleq\begin{bmatrix} U_1 \\ U_2 \end{bmatrix}=[\phi(x)]\int_0^x[\phi(-\tau)][b(\tau)]\,d\tau=p_0\begin{bmatrix} \dfrac{x^4}{24EI} \\ \dfrac{x^3}{6EI} \\ \dfrac{x^2}{2} \\ x \end{bmatrix}. \tag{25}$$

The boundary conditions for a cantilever beam can be expressed as

$$\left.\begin{aligned} \begin{bmatrix} z_1(0) \\ z_2(0) \end{bmatrix}&\triangleq[K_1]=[0], \\ \begin{bmatrix} z_3(l) \\ z_4(l) \end{bmatrix}&\triangleq[K_2]=[0]. \end{aligned}\right\} \tag{26}$$

Thus

$$[U_1{}^\alpha]=[U_2{}^\alpha]=[0],\ [U_2{}^\beta]=p_0\begin{bmatrix} \dfrac{l^2}{2} \\ l \end{bmatrix}. \tag{27}$$

The submatrices of $[\phi(x)]^{-1}$ are as follows :

$$[\Psi_{22}]=[\Psi_{11}]=\begin{bmatrix} 1 & -x \\ 0 & 1 \end{bmatrix},\quad [\Psi_{12}]=\begin{bmatrix} \dfrac{x^2}{2EI} & \dfrac{-x^3}{6EI} \\ \dfrac{-x}{EI} & \dfrac{x^2}{2EI} \end{bmatrix},\quad [\Psi_{21}]=[0], \tag{28}$$

By eqns. (26), (27) and (28), the solution (eqn. (16)) yields

$$\begin{bmatrix} Z_1 \\ \hline Z_2 \end{bmatrix}=\begin{bmatrix} -\Phi_{12} & \Psi_{22}{}^\beta & U_2{}^\beta \\ \hline -\Phi_{22} & \Psi_{22}{}^\beta & U_2{}^\beta \end{bmatrix}+\begin{bmatrix} U_1 \\ \hline U_2 \end{bmatrix}, \tag{29}$$

where

$$\left.\begin{aligned} &[\Phi_{12}]=\begin{bmatrix} \dfrac{x^2}{2EI} & \dfrac{x^3}{6EI} \\ \dfrac{x}{EI} & \dfrac{x^2}{2EI} \end{bmatrix},\quad [\Phi_{22}]=\begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix}, \\ &[U_2{}^\beta]=p_0\begin{bmatrix} \dfrac{l^2}{2} \\ l \end{bmatrix}\qquad [\Psi_{22}{}^\beta]=\begin{bmatrix} 1 & -l \\ 0 & 1 \end{bmatrix}. \end{aligned}\right\} \tag{30}$$

p. 83

Upon substitution we obtain

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = -p_0 \begin{bmatrix} \dfrac{-l^2x^2}{4EI} + \dfrac{lx^3}{6EI} \\ \dfrac{-l^2x}{2EI} + \dfrac{lx^2}{2EI} \\ \dfrac{-l^2}{2} + lx \\ l \end{bmatrix} + p_0 \begin{bmatrix} \dfrac{x^4}{24EI} \\ \dfrac{x^3}{6EI} \\ \dfrac{x^2}{2} \\ x \end{bmatrix},$$

which is the solution of a cantilever beam.

5. Conclusions

In general, the mixed boundary value problems are less amenable to numerical computation. It is seen that our systematic procedure established as well as simple formulae derived make the modern powerful tool of engineering —digital computers—more suitable for analysing this class of problems.

Acknowledgments

This work is supported in part by the National Aeronautics and Space Administration Grant NGL–44–005–084.

References

Chen, C. F., and Haas, I. J., 1968, *Elements of Control Systems Analysis* (Englewood Cliffs, New Jersey : Prentice-Hall Co.).

Timoshenko, S., and MacCullough, G. H., 1940, *Elements of Strength of Materials*, 2nd edition (D. van Nostrand Co. Inc.).

p.84

Project

(A) Project Title: A New Formulation of the Hermite Criterion

(B) Project Abstract:

The Routh algorithm is used to generate the parameters of the Hermite criterion. The new formulation is simpler than the original. The relationship among the Hermite, the symmetrical Hurwitz, the Kalman-Bertram, and the Routh criteria is naturally established.

(C) Publication: International Journal of Control, 19, 457 (1974)

(D) Year: 1974

(E) Department: Electrical Engineering

(F) Author: Professor C. F. Chen

INT. J. CONTROL, 1974, VOL. 19, NO. 4, 757–764

A new formulation of the Hermite criterion

C. F. CHEN†

The Routh algorithm is used to generate the parameters of the Hermite criterion. The new formulation is simpler than the original. The relationship among the Hermite, the symmetrical Hurwitz, the Kalman–Bertram, and the Routh criteria is naturally established.

1. Introduction

The Hermite matrix (1854) is formulated in terms of the coefficients of a characteristic polynomial. The parameters of the form are determined by an algorithm which is not very convenient, and the matrix itself appears very complicated.

Since Parks (1962) rediscovered Hermite's matrix and Ralston (1962) independently established the symmetrical Hurwitz criterion, the Hermite criterion has played an increasingly important role in system analysis. Recently, Jury and Anderson (1972) discussed the simplified stability criteria and Anderson (1972) also developed the reduced Hermite form. They, however, have not overcome the basic difficulty : the elements generated by the algorithm are too complicated for general use.

This paper will give a new form and a new derivation of the Hermite criterion. The form is simple and the derivation is unified.

2. Canonical forms and canonical transformations

It is known that a general linear system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \tag{1}$$

can be changed into a companion form by the use of Krylov's transformation (Gantmacher 1959)

$$\mathbf{K}\mathbf{x} = \mathbf{y} \tag{2}$$

where

$$\mathbf{K} = \begin{bmatrix} \mathbf{q}' \\ \mathbf{q}'\mathbf{A} \\ \vdots \\ \mathbf{q}'\mathbf{A}^{n-1} \end{bmatrix} \tag{3}$$

in which $\mathbf{q}'$ is any row vector such that $\mathbf{K}$ is not singular. The prime means transposition. We then have

$$\dot{\mathbf{y}} = \mathbf{K}\mathbf{A}\mathbf{K}^{-1}\mathbf{y} \triangleq \alpha\mathbf{y} \tag{4}$$

where

$$\alpha = \begin{bmatrix} 0 & 1 & 0 & . & . & . & 0 \\ 0 & 0 & 1 & . & . & . & 0 \\ 0 & 0 & 0 & 1 & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & 1 \\ -a_n & -a_{n-1} & . & . & . & -a_2 & -a_1 \end{bmatrix} \tag{5}$$

Received 30 April 1973.

† Department of Electrical Engineering, University of Houston, Texas 77004.

p. 86

ORIGINAL PAGE IS
OF POOR QUALITY

The characteristic equation can be directly obtained from (5) :

$$\lambda^n + a_1\lambda^{n-1} + \ldots + a_{n-1}\lambda + a_n = 0 \tag{6}$$

Based on (6) we now write the Routh array :

$$\begin{array}{c|ccccc|ccc}
1 & a_2 & a_4 & \cdot & \cdot & \cdot & c_{11} & c_{12} & c_{13} \\
a_1 & a_3 & a_5 & \cdot & \cdot & & c_{21} & c_{22} & c_{23} \\
\left(\dfrac{a_1a_2-a_3}{a_1}\right) & \cdot & \cdot & & & & c_{31} & \cdot & \cdot \\
\cdot & \cdot & \cdot & & & & & & \\
\cdot & & & & & & \cdot & \cdot & \\
\cdot & & & & & & \cdot & & \\
 & & & & & & c_{n,1} & & \\
a_n & & & & & & c_{n+1,1} & &
\end{array} \tag{7}$$

The right-hand side of the verticle line of (7) is the corresponding double script notation of the Routh array. By the use of the convenient notation the Routh algorithm (Chen and Haas 1968) is simply as follows :

$$c_{j,k} = c_{j-2,k+1} - \frac{c_{j-2,1}c_{j-1,k+1}}{c_{j-1,1}} \tag{7 a}$$

Then we perform another linear transform by letting

$$\mathbf{L}\mathbf{y} = \mathbf{z} \tag{8}$$

where

$$\mathbf{L} = \begin{bmatrix}
1 & 0 & \cdots & & & & & 0 & 0 \\
0 & 1 & & & & & & & \vdots \\
\dfrac{c_{n-1,2}}{c_{n-1,1}} & 0 & \ddots & & & & & & \vdots \\
0 & \ddots & \ddots & \ddots & & & & & \vdots \\
\dfrac{c_{n-3,3}}{c_{n-3,1}} & \ddots & \ddots & \ddots & 1 & 0 & 0 & 0 \\
0 & \ddots & \ddots & \ddots & 0 & 1 & 0 & 0 \\
\vdots & \ddots & \ddots & 0 & \dfrac{c_{32}}{c_{31}} & 0 & 1 & 0 \\
0 & \cdots & \dfrac{c_{23}}{c_{21}} & 0 & \dfrac{c_{22}}{c_{21}} & 0 & 1
\end{bmatrix} \tag{9}$$

This transformation was found by Chen and Chu (1966). For convenience, we call (9) the Chen–Chu transformation. Then the system becomes

$$\begin{aligned}\dot{\mathbf{z}} &= \mathbf{L}\alpha\mathbf{L}^{-1}\mathbf{z} \\ &\triangleq \beta\mathbf{z}\end{aligned} \tag{10}$$

where
$$\boldsymbol{\beta}=\begin{bmatrix} -\frac{c_{n+1,1}}{c_{n-1,1}} & & & & & \\ & \ddots & & & \mathbf{O} & \\ & & 0 & & & \\ & & -\frac{c_{51}}{c_{31}} & 0 & 1 & \\ & & & -\frac{c_{41}}{c_{21}} & 0 & 1 \\ & \mathbf{O} & & & -\frac{c_{31}}{c_{11}} & -\frac{c_{21}}{c_{11}} \end{bmatrix} \tag{11}$$

Form (10) is the Schwarz form (1956). The fact that its elements in (11) are in terms of the Routh elements is due to Chen and Chu (1966). Schwarz originally used computer techniques to search this matrix. Bellman (1970), Kalman (1960) and others followed him.

We perform one more transformation, letting

$$\mathbf{Mz}=\mathbf{w} \tag{12}$$

where
$$\mathbf{M}=\begin{bmatrix} & & & & \sqrt{(c_{11}c_{21})} \\ & & & -\sqrt{(c_{21}c_{31})} & \\ & & \sqrt{(c_{31}c_{41})} & & \\ & \iddots & & & \\ (-1)^{\rho+1}\sqrt{(c_{n,1}c_{n+1,1})} & & & & \end{bmatrix} \tag{13}$$

where ρ is the row number index of $\mathbf{M}$.

Then we have

$$\begin{aligned}\dot{\mathbf{w}}&=\mathbf{M}\boldsymbol{\beta}\mathbf{M}^{-1}\mathbf{w}\\ &\triangleq\boldsymbol{\gamma}\mathbf{w}\end{aligned} \tag{14}$$

where
$$\boldsymbol{\gamma}=\begin{bmatrix} -\frac{c_{21}}{c_{11}} & \sqrt{\frac{c_{31}}{c_{11}}} & & & & \\ -\sqrt{\frac{c_{31}}{c_{11}}} & 0 & \sqrt{\frac{c_{41}}{c_{21}}} & & \mathbf{O} & \\ & -\sqrt{\frac{c_{41}}{c_{21}}} & 0 & \sqrt{\frac{c_{51}}{c_{31}}} & & \\ & & -\sqrt{\frac{c_{51}}{c_{31}}} & \ddots & \ddots & \\ & \mathbf{O} & & \ddots & \ddots & \sqrt{\frac{c_{n+1,1}}{c_{n-1,1}}} \\ & & & & \sqrt{\frac{c_{n+1,1}}{c_{n-1,1}}} & 0 \end{bmatrix} \tag{15}$$

p. 88

Form (15) was first found by Puri and Weygandt (1963) and **M** was discovered by Power (1969). The fact that all the elements are in terms of the Routh array is due to the author.

Let us summarize the transformations as follows:

ORIGINAL PAGE IS
OF POOR QUALITY

$$\left.\begin{array}{lll} \dot{\mathbf{w}}=\gamma\mathbf{w} & \text{(Puri–Weygandt coordinates)} \\ \dot{\mathbf{z}}=\mathbf{M}^{-1}\gamma\mathbf{M}\mathbf{z} & \text{(Schwarz coordinates)} \\ \dot{\mathbf{y}}=\mathbf{L}^{-1}\mathbf{K}^{-1}\gamma\mathbf{M}\mathbf{L}\mathbf{y} & \text{(phase variable coordinates)} \\ \dot{\mathbf{x}}=\mathbf{K}^{-1}\mathbf{L}^{-1}\mathbf{M}^{-1}\gamma\mathbf{M}\mathbf{L}\mathbf{K}\mathbf{x} & \text{(general coordinates)} \\ \text{where} & \\ \quad \mathbf{K} \text{ is the Kyrlov transformation} & \\ \quad \mathbf{L} \text{ is the Chen–Chu transformation} & \\ \quad \mathbf{M} \text{ is the Power transformation} & \end{array}\right\} \tag{16}$$

It was very unfortunate that Puri and Weygandt formulated $(ML)^{-1}$ as one matrix and that matrix is in terms of the characteristic coefficients; therefore they made the formulation very complicated: Butchart (1965), on the other hand, formed (L^{-1}) first, and the elements are expressed by Hurwitz determinants; he, therefore, made the transformation unnecessarily complex. If we form M and L first and write their terms by the Routh array parameters as we did, the whole process is much simplified.

3. Liapunov functions

In the Puri–Weygandt coordinates, we synthesize a Liapunov function

$$v_w=\mathbf{w}'\mathbf{I}\mathbf{w} \tag{17}$$

where **I** is the identity matrix.

In the Schwarz coordinates, the function is

$$\begin{aligned} v_z &= \mathbf{z}'\mathbf{M}'\mathbf{I}\mathbf{M}\mathbf{z} \\ &\triangleq \mathbf{z}'\mathbf{G}\mathbf{z} \end{aligned} \tag{18}$$

where

$$\mathbf{G}=\mathbf{M}'\mathbf{I}\mathbf{M}=\begin{bmatrix} c_{n,1}c_{n+1,1} & & & & \\ & \ddots & & & \\ & & c_{31}c_{41} & & \\ & & & c_{21}c_{31} & \\ & & & & c_{11}c_{21} \end{bmatrix} \tag{19}$$

(19) was discovered by Kalman and Bertram†; Chen and Chu (1966) express the elements by the Routh array parameters while Kalman and Bertram (1960) used Schwarz's elements.

In the phase variable coordinates, the same function becomes

$$\begin{aligned} v_y &= \mathbf{y}'\mathbf{L}'\mathbf{M}'\mathbf{I}\mathbf{M}\mathbf{L}\mathbf{y} \\ &= \mathbf{y}'\mathbf{H}\mathbf{y} \end{aligned} \tag{20}$$

where

$$\mathbf{H}=\mathbf{L}'\mathbf{M}'\mathbf{I}\mathbf{M}\mathbf{L} \tag{21}$$

† Kalman and Bertram (1960) used Schwarz parameters to form (19); their matrix is somewhat complicated than the present form. For details, see Chen and Chu (1966).

The elements of the matrix **H** are in terms of the Routh parameters; when we convert them into characteristic coefficients, the matrix **H** is identical with the Hermite criterion. Therefore, Hermite's criterion can be considered as a Liapunov function in the phase variable coordinate. Parks recognized this fact and Anderson used this fact to prove the reduced Hermite criterion. However, each of them failed to express it by the Routh parameters; and therefore they kept the Hermite criterion as complicated as it was. Now, not only do we have a simpler form but also we see the links among Routh's, Kalman and Bertram's and Hermite's criteria.

4. A new form of Hermite criterion

For illustration, let us consider a fourth-order system :

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \dot{y}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -a_4 & -a_3 & -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \tag{22}$$

Performing the Chen–Chu transformation

$$\mathbf{L}\mathbf{y} = \mathbf{z}$$

where

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \dfrac{c_{32}}{c_{31}} & 0 & 1 & 0 \\ 0 & \dfrac{c_{22}}{c_{21}} & 0 & 1 \end{bmatrix} \tag{23}$$

in which c_{ij} are evaluated from the Routh array, or

$$\begin{array}{ccc|ccc} 1 & a_2 & a_4 & c_{11} & c_{12} & c_{13} \\ a_1 & a_3 & & c_{21} & c_{22} & \\ \dfrac{a_1a_2-a_3}{a_1} & a_4 & & c_{31} & c_{32} & \\ \dfrac{\left(\dfrac{a_1a_2-a_3}{a_1}a_3-a_1a_4\right)}{\dfrac{a_1a_2-a_3}{a_1}} & & & c_{41} & & \\ a_4 & & & c_{51} & & \end{array} \tag{24}$$

7.90

The Liapunov function in the Schwarz coordinates is

$$
\begin{aligned}
v_y &= \mathbf{y}'\mathbf{L}'\mathbf{M}'\mathbf{I}\mathbf{M}\mathbf{L}\mathbf{y} \\
&= \mathbf{y}'\mathbf{L}'\mathbf{G}\mathbf{L}\mathbf{y} \\
&= \mathbf{y}' \begin{bmatrix} 1 & 0 & \frac{c_{32}}{c_{31}} & 0 \\ 0 & 1 & 0 & \frac{c_{22}}{c_{21}} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_{41}c_{51} & & & \\ & c_{31}c_{41} & & \\ & & c_{21}c_{31} & \\ & & & c_{11}c_{21} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{c_{32}}{c_{31}} & 0 & 1 & 0 \\ 0 & \frac{c_{22}}{c_{21}} & 0 & 1 \end{bmatrix} \mathbf{y} \qquad (25) \\
&= \mathbf{y}' \begin{bmatrix} c_{41}c_{51} + \frac{c_{32}{}^2 c_{21}}{c_{31}} & 0 & c_{32}c_{21} & 0 \\ 0 & c_{31}c_{41} + \frac{c_{22}{}^2 c_{11}}{c_{21}} & 0 & c_{22}c_{11} \\ c_{21}c_{32} & 0 & c_{21}c_{31} & 0 \\ 0 & c_{11}c_{22} & 0 & c_{11}c_{21} \end{bmatrix} \mathbf{y} \qquad (25\,a)
\end{aligned}
$$

The matrix (25 a) is the new Hermite form for the fourth-order system which is neater than the original Hermite formulation. Instead of using the coefficients a_i of the characteristic equation, we describe all the elements in terms of the Routh parameters. However, the expansion form of (25 a) that is (25) is even simpler ; the core matrix appears as a combination of three matrices. We, therefore, claim that the product of three matrices in (25) is a new form of Hermite's criterion. From this special form, we generalize to the Hermite criterion for the nth-order system as follows :

$$
v_y = \mathbf{y}' \begin{bmatrix} 1 & 0 & \frac{c_{n-1,\,2}}{c_{n-1,\,1}} & & & \\ & 1 & 0 & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & \frac{c_{22}}{c_{21}} & \\ & & 1 & \ddots & 0 & \\ & & & & \ddots & 1 \end{bmatrix} \begin{bmatrix} c_{n1}c_{n+1,\,1} & & & \\ & \ddots & & \\ & & c_{21}c_{31} & \\ & & & c_{11}c_{21} \end{bmatrix} \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ \frac{c_{n-1,\,2}}{c_{n-1,\,1}} & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{c_{22}}{c_{21}} & 0 & 1 \end{bmatrix} \mathbf{y} \qquad (26)
$$

f.91

Examining (26) we see that it is a product of three simple matrices : namely, the transpose of the Chen–Chu matrix ; the modified Kalman–Bertram matrix ; and the Chen–Chu matrix.

5. Derivatives of Liapunov functions

Assume that for the system

$$\dot{\mathbf{w}} = \gamma \mathbf{w} \tag{14}$$

we have a Liapunov function

$$v_w = \mathbf{w}'\mathbf{I}\mathbf{w} \tag{17}$$

Substituting (15) into the derivative function of (17) gives

$$\dot{v}_w = \mathbf{w}' \begin{bmatrix} -2\dfrac{c_{21}}{c_{11}} & 0 & 0 & . & . & 0 \\ 0 & 0 & 0 & . & . & 0 \\ 0 & 0 & . & . & . & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & 0 \end{bmatrix} \mathbf{w} \tag{27}$$

This semidefinite function can be described in the Schwarz coordinates

$$\begin{aligned} \dot{v}_z &= \mathbf{z}'\mathbf{M}'(\gamma + \gamma^{\mathrm{T}})\mathbf{M}\mathbf{z} \\ &= \mathbf{z}' \begin{bmatrix} 0 & 0 & 0 & . & . & 0 \\ 0 & 0 & 0 & . & . & 0 \\ 0 & 0 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & -2c_{21}^2 \end{bmatrix} \mathbf{z} \end{aligned} \tag{28}$$

Function (27) was used by Puri and Weygandt, while function (28) was frequently applied by Kalman and Bertram and Parks. Of course, they used Schwarz's elements to express them.

In the phase variable coordinate, the derivative of the Liapunov function is easily found :

$$\begin{aligned} \dot{v}_y &= \mathbf{y}'\mathbf{L}' \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2c_{21}^2 \end{bmatrix} \mathbf{L}\mathbf{y} \\ &= \mathbf{y}' \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -2c_{22}^2 & 0 & -2c_{22}c_{21} \\ 0 & 0 & 0 & 0 \\ 0 & -2c_{22}c_{21} & 0 & -2c_{21}^2 \end{bmatrix} \mathbf{y} \end{aligned} \tag{29}$$

Inspecting either (27), (28) or (29) we see that they are always negative semidefined if c_{21} is not zero. The definition of c_{21} is a_1 which is indeed a positive

p.92

real number. Therefore, when we use Hermite's criterion, as a Liapunov function in (29), we never worry about the definiteness of its derivative. It is also evident that the central matrix of (26) which is Kalman and Bertram's matrix is positive definite, if and only if c_{i1} are all positive. Obviously, this statement is simply the Routh criterion.

6. Conclusions

For Hermite's criterion of stability, we derived a new form (25) which is less complicated than its original formulation. The derivation is simple and new. The relationship among Hermite's criterion, Kalman and Bertram's criterion and the Routh criterion is naturally shown by the use of canonical forms of companion, Schwarz, and Puri and Weygandt through the applications of canonical transformations developed by Krylov, Chen and Chu, and Power. It should be emphasized that this new development is based on the repeated applications of Routh's elements instead of using characteristic coefficients or Hurwitz determinants.

ACKNOWLEDGMENT

This work is supported in part by the National Aeronautics and Space Administration Grant NGL–44–005–084.

REFERENCES

ANDERSON, B D. O., 1972, *I.E.E.E. Trans autom. Control*, **17**, 669
BELLMAN, R., 1970, *Introduction to Matrix Analysis*, second edition (McGraw-Hill Book Company), p. 256.
BUTCHART, R. L., 1965, *Int. J. Control*, **1**, 201.
CHEN, C. F., and CHU, H., 1966, *I.E.E.E Trans. autom. Control*, **11**, 303.
CHEN, C. F., and HAAS, I. J., 1968, *Elements of Control Systems* (Prentice-Hall).
GANTMACHER, F. R., 1959, *The Theory of Matrices*, Vol. I (New York · Chelsea Publishing Co.).
HERMITE, C., 1854, *J. Reine angew. Math.*, **52**, 39 and works 1, p. 397.
JURY, E. I., and ANDERSON, B. D. O., 1972, *I.E.E.E. Trans. autom Control*, **17**, 371.
KALMAN, R E., and BERTRAM, J. E., 1960, *Trans. Am Soc. mech. Engrs* D, 371.
PARKS, P. C., 1962, *Proc Camb phil. Soc.*, **58**, 694.
POWER, H. M., 1969, *Proc. Instn elect Engrs*, **116**, **7**, 1245.
PURI, N N., and WEYGANDT, C. N., 1963, *J. Franklin Inst.*, **276**, 365.
RALSTON, A., 1962, *I.R.E Trans. autom. Control*, **7**, 50.
SCHWARZ, H. R., 1956, *Z. angew. Math Phys.*, **7**, 473

p.93

Project

(A) Project Title: Evaluation of Irrational and Transcendental Transfer Functions via the Fast Fourier Transform

(B) Project Abstract:

The analytic methods for evaluating the time response of irrational transfer functions are incomplete. The graphical approaches for solving the problem are inaccurate. This paper attempts to attack the general inverse Laplace transform is developed. Several typical but difficult cases are studied and the results are extraordinarily satisfactory.

(C) Publication: International Jountal of Control, 35, No. 2, p. 267 (1973).

(D) Year: 1973

(E) Department: Electrical Engineering

(F) Student Name: R. F. Chiu

(G) Faculty Advisor: Professor C. F. Chen

INT. J. ELECTRONICS, 1973, VOL. 35, NO. 2, 267–276

Evaluation of irrational and transcendental transfer functions via the fast Fourier transform †

C. F. CHEN and R. F. CHIU ‡
Electrical Engineering Department,
University of Houston, Houston, Texas

[Received 5 May 1972]

The analytic methods for evaluating the time response of irrational transfer functions are incomplete. The graphical approaches for solving the problem are inaccurate. This paper attempts to attack the general inverse Laplace transform problem under the new light shed by the Fast Fourier transform. An algorithm for the inverse Laplace transform is developed. Several typical but difficult cases are studied and the results are extraordinarily satisfactory.

1. Introduction

A large number of circuits, processes or systems have distributed parameters and/or delay elements. (1) Thermal processes, (2) hole diffusion of transistors, (3) electromagnetic devices, and (4) transmission lines are typical examples (Truxal 1955, Campbell 1958, and Bohn 1963). The mathematical models in the Laplace transform domain for these elements contain either the operators s under the radical sign or other irrational or transcendental functions. To find the inverse Laplace transforms of these functions is an extremely important topic in analysis and a very difficult problem in general.

For solving the problem, the methods so far developed can be summarized into three schools:

(1) Approximation methods

Let a rational transfer function approximate an irrational or transcendental transfer function. The main contributions include Pade's approximation developed by Stewart (1960), Carlson and Haljak's approach (Carlson and Haljak 1964) to use a regular Newton's process generating rational functions, Lerner's work (Lerner 1965) on potential analogue approximation, and Chen and Shieh's exhaustively calculating the approximate high order rational transfer functions (Chen and Shieh 1967) for typical irrational and transcendental functions by using a digital computer approach.

(2) Analytic methods

Among the available analytic methods, the most notable one is developed by Netushil and Kilomeitseva (1965). They use special functions to make the transform table serving the particular purpose.

(3) Graphical methods

Convert the irrational transfer function in the Laplace domain into the frequency domain, then use Laguerre polynomials or Chebyshev polynomials

† Communicated by Professor C. F. Chen.

‡ Dr. R. F. Chiu is now with the International Communications Corporation, Miami, Florida.

7.95

to calculate the inverse function numerically. Chen's Inverse Laplace transform formula in terms of Laguerre polynomials is a typical method (Chen 1966) in this area.

The approximation approaches in the first school are lacking accuracy, while the analytic method is not complete. (For example, Kilomeitseva and Netushil cannot treat the repeated root case.) The third school is a practical method. However, if we convert it into a computer-aided approach, we will face either speed or storage problems.

This paper solves the problem by using the Fast Fourier transform techniques.

2. Difficulties in analytic approach

Kilomeitseva and Netushil's (1965) theory can be summarized below in order to examine the difficulties involved.

An irrational transfer function is usually given as follows:

$$W(s)=\frac{a_0 s^m + a_1 s^{m-1} + \ldots + a_m}{b_0\sqrt{s^{2n+1}} + b_1\sqrt{s^{2n}} + \ldots b_{2n+1}}. \tag{1}$$

Let $\sqrt{s}=z$; eqn. (1) becomes a rational transfer function of z; or

$$W(z)=\frac{a_0 z^{2m} + a_1 z^{2(m-1)} + \ldots + a_m}{b_0 z^{2n+1} + b_1 z^{2n} + \ldots + b_{2n+1}}, \tag{2}$$

or simplify

$$=\frac{\sum_{i=0}^{m} a_i z^{2(m-i)}}{\sum_{i=0}^{2n+1} b_i z^{2n+1-i}}. \tag{3}$$

This rational fraction may be expanded into partial fractions

$$W(z)=\sum_{i=1}^{2n+1}\frac{A_i}{z-z_i}. \tag{4}$$

If there is no repeated root in the characteristic equation of (2), the inverse Laplace transform of the typical term (4) with a unit step as input can be derived by using the well known pair:

$f(t)$	$F(s)$
$\exp(a^2t)[b-a\,\mathrm{erf}\,(a\sqrt{t})]-b\exp(b^2t)\,\mathrm{erfc}\,(b\sqrt{t})$	$\dfrac{b^2-a^2}{(s-a^2)(b+\sqrt{s})}$

and letting $a=0$ to get the reduced form

$$b-b\exp(b^2t)\,\mathrm{erfc}\,(b\sqrt{t})\,\Big|\,\frac{b^2}{s(\sqrt{s}+b)}. \tag{5}$$

This inverse formula finally can be written as follows:

$$L^{-1}\left[\frac{A}{s(\sqrt{s}+a)}\right]=\frac{A}{a}[1-\exp(a^2t)\,\mathrm{erfc}\,(a\sqrt{t})]. \tag{6}$$

7.96

ORIGINAL PAGE IS
OF POOR QUALITY

If complex roots are involved, pair (5) cannot be directly applied. Kilomeitseva and Netushil's approach is to reform the pair into a new function.

ORIGINAL PAGE IS
OF POOR QUALITY

For example, assume the complex conjugate roots are z_k and z_{k+1}.

$$W_1(z)=\frac{A_k}{z-z_k}+\frac{A_{k+1}}{z-z_{k+1}}=\frac{E}{(z-z_k)(z-z_{k+1})}+\frac{Dz}{(z-z_k)(z-z_{k+1})}. \tag{7}$$

The inverse Laplace transform will be

$$L^{-1}\frac{W_1(z)}{s}=E\left[\frac{1}{\rho}m_0(\rho\sqrt{t},\,\theta)\right]+D\left[\frac{1}{\rho^2}m_1(\rho\sqrt{t},\,\theta)\right] \tag{8}$$

where ρ and θ are defined by $z_k=\rho\exp(j\theta)$, and m_0 and m_1 are given by very special graphical forms in terms of very complicated special functions and their approximations.

The difficulties involved in their approach are as follows:

(1) If there is a repeated root appearing in the transfer function, no corresponding inverse formula is available.

(2) If there is a pair of complex roots involved, the inverse formula cannot be evaluated without a particular table or curves.

(3) Even all roots are real; because the inverse formula is not in terms of elementary functions, it does not reveal too much information without plotting the corresponding curves again.

3. Difficulties in graphical methods

We start with a general transfer function which could involve irrational elements and/or transcendental elements.

$$Y(s,\,\sqrt{s},\,\exp(-\tau s)). \tag{9}$$

The stability of the system is examined first. It can be easily done by one of the existing methods; for example, Brim's approach. After knowing the system is asymptotically stable, we can substitute 's' by 'jw'. (We note that when the system is not asymptotically stable, we can make it to be asymptotically stable by choosing a suitable positive number c and substituting s with $s+c$, and time the final result by $\exp(ct)$.)

Then we have

$$Y[jw,\,\sqrt{jw},\,\exp(-\tau jw)]. \tag{10}$$

The magnitude and phase versus frequency curves can be drawn. Based on those curves, one can obtain the time curve numerically. The well-known method along this line are Floyd's method (Brown and Campbell 1948), Guilleman's method (Truxal 1955) and Chen's method (Chen 1966). The basic formula they used is

$$y(t)=\frac{2}{\pi}\int_0^\infty R(w)\cos tw\,dw$$

where $R(w)$ is the real part of $Y(w)$.

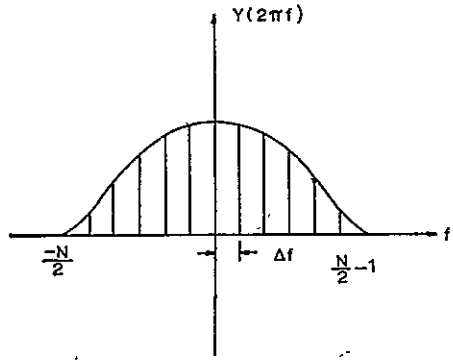
Because those three methods are basically graphical, it is difficult to obtain accurate answers.

p.97

4. Principle of the new approach

We rather look at this inverse problem from a different angle. Consider the given frequency response as in fig. 1. First, we take N points in the frequency axis as shown in fig. 1. It is noted that in the negative frequency

Fig. 1



part, we truncate at $(N/2)\Delta f$, while in the positive frequency part, we truncate at $[(N/2)-1]\Delta f$. We divide the intervals in this way, because N has to be even in the Fast Fourier transform. When N is large, this kind of dividing is justified.

Then

$$y(t)=\sum_{n=(-N/2)}^{(N/2)-1} Y(2\pi n\Delta f)\Delta f \exp(j2n\pi\Delta ft), \quad 0\leqslant t<\frac{1}{\Delta f}. \tag{11}$$

It can be written into two parts;

$$y(t)=\sum_{n=(-N/2)}^{-1} Y(2\pi n\Delta f)\Delta f \exp(j2n\pi\Delta ft) + \sum_{n=0}^{(N/2)-1} Y(2\pi n\Delta f)\Delta f \exp(j2\pi n\Delta ft). \tag{12}$$

Let $n'=N+n$

$$n=n'-N.$$

The first term of (12) becomes

$$=\sum_{n'=(N/2)}^{N-1} Y[2(n'-N)\pi\Delta f]\Delta f \exp[j2\pi(n'-N)\,\Delta ft].$$

We change the dummy variable n' back to n

$$=\sum_{n=(N/2)}^{N-1} Y[2(n-N)\pi\Delta f]\Delta f \exp[j2\pi(n-N)\,\Delta ft]$$

or

$$y(t)=\sum_{n=0}^{(N/2)-1} Y(2\pi n\Delta f)\Delta f \exp(j2n\pi\Delta ft) + \sum_{n=(N/2)}^{N-1} Y[2\pi(n-N)\Delta f]\Delta f \exp[j2(n-N)\pi\Delta ft], \tag{13}$$

P. 98

when

$$t=\frac{k}{N\Delta f}, \quad k=1, 2, \ldots, N-1$$

$$y\left(\frac{k}{N\Delta f}\right)=\sum_{n=0}^{(N/2)-1} Y(2\pi n\Delta f)\Delta f \exp\left(j2\pi n\Delta f\frac{k}{N\Delta f}\right)$$

ORIGINAL PAGE IS
OF POOR QUALITY

$$+\sum_{n=(N/2)}^{N-1} Y[2\pi(n-N)\Delta f]\Delta f \exp\left[j2(n-N)\pi\Delta f\frac{k}{N\Delta f}\right]$$

$$=\sum_{n=0}^{(N/2)-1} Y(2\pi n\Delta f)\Delta f \exp[j(2nk\pi/N)]$$

$$+\sum_{n=(N/2)}^{N-1} Y[2\pi(n-N)\Delta f]\Delta f \exp[j(2nk\pi/N)]. \tag{14}$$

Let

$$S_n=Y(2\pi n\Delta f)\Delta f, \qquad 0\leqslant n<\frac{N}{2}$$

$$=Y[2\pi(n-N)\Delta f]\Delta f, \quad \frac{N}{2}\leqslant n<N$$

and

$$A_k=y\left(\frac{k}{N\Delta f}\right).$$

Then we finally have

$$A_k=\sum_{n=0}^{N-1} S_n \exp(j2nk\pi/N), \quad k=1, 2, \ldots, N-1. \tag{15}$$

Equation (15) is the standard form on which the Fast Fourier transform is based.

5. Proof of the fundamental formula

The fundamental formula of our approach is (11). Its proof is shown as follows:

Let the inverse Laplace transform of the transfer functions $Y(s, \sqrt{s}, \exp(-\tau s))$ be $y(t)$. If it is asymptotically stable, we can write

$$y(t)=\int_{-\infty}^{+\infty} Y(2\pi f)\exp(j2\pi ft)\,df.$$

Define

$$Y^*(2\pi f)=\sum_{n=-\infty}^{+\infty} Y(2\pi n\Delta f)\Delta f\delta(f-n\Delta f)$$

and

$$y^*(t)=\int_{-\infty}^{+\infty} Y^*(2\pi f)\exp(j2\pi ft)\,df,$$

p. 99

then

$$Y^*(2\pi f)=\sum_{n=-\infty}^{n=\infty} Y(2\pi n\Delta f)\Delta f\delta(f-n\Delta f)$$

$$=Y(2\pi f)\Delta f\sum_{n=-\infty}^{\infty}\delta(f-n\Delta f)$$

$$=Y(2\pi f)\Delta f\sum_{n=-\infty}^{\infty}\frac{1}{\Delta f}\exp(j2n\pi f/\Delta f)$$

$$=\sum_{n=-\infty}^{\infty} Y(2\pi f)\exp(j2n\pi f/\Delta f)$$

and

$$y^*(t)=\int_{-\infty}^{+\infty}\left[\sum_{n=-\infty}^{n=\infty} Y(2\pi f)\exp(j2n\pi f/\Delta f)\right]\exp(j2\pi ft)\,df$$

$$=\sum_{n=-\infty}^{\infty}\int_{-\infty}^{+\infty} Y(2\pi f)\exp([j2\pi f(t+(n/\Delta f))]\,df$$

$$=\sum_{n=-\infty}^{n=\infty} y\left(t+\frac{n}{\Delta f}\right).$$

If $y(t)$ is 0 or negligible when $t<0$ and $t\geqslant(1/\Delta f)$, then $y^*(t)=y(t)$ when $0\leqslant t<(1/\Delta f)$.

Because most control systems are low-pass filters, $Y(2\pi f)$ can be neglected when $|f|$ is very large. Let us truncate $Y(2\pi f)$ at $f=[(N/2)-1]\Delta f$ on the left hand and at $f=(-N/2)\Delta f$ on the right hand side, then $Y^*(2\pi f)$ becomes

$$Y^*(2\pi f)=\sum_{n=(-N/2)}^{n=(N/2)-1} Y(2\pi n\Delta f)\Delta f\delta(f-n\Delta f)$$

and

$$y^*(t)=\int_{-\infty}^{+\infty}\sum_{n=(-N/2)}^{n=(N/2)-1} Y(2\pi n\Delta f)\Delta f\delta(f-n\Delta f)\exp(j2\pi ft)\,df,$$

or

$$y(t)=\sum_{n=(-N/2)}^{n=(N/2)-1} Y(2\pi n\Delta f)\Delta f\exp[j2\pi(n\Delta f)t],$$

when $0\leqslant t<(1/\Delta f)$.

Therefore, (11) is proved.

6. Irrational transfer function evaluation

Based on (15), a computer programme for evaluating general transfer functions is written as shown in the appendix.

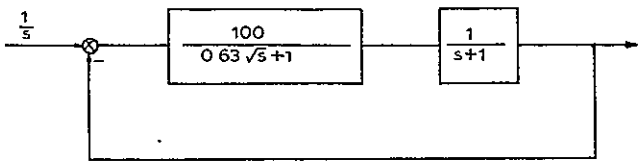
The example shown in fig. 2 has been tested.

In the example, there is an irrational function element in the open loop transfer function.

$$G(s)=\frac{100}{(s+1)(0{\cdot}63\sqrt{s}+1)}.$$

p. 100

Fig. 2



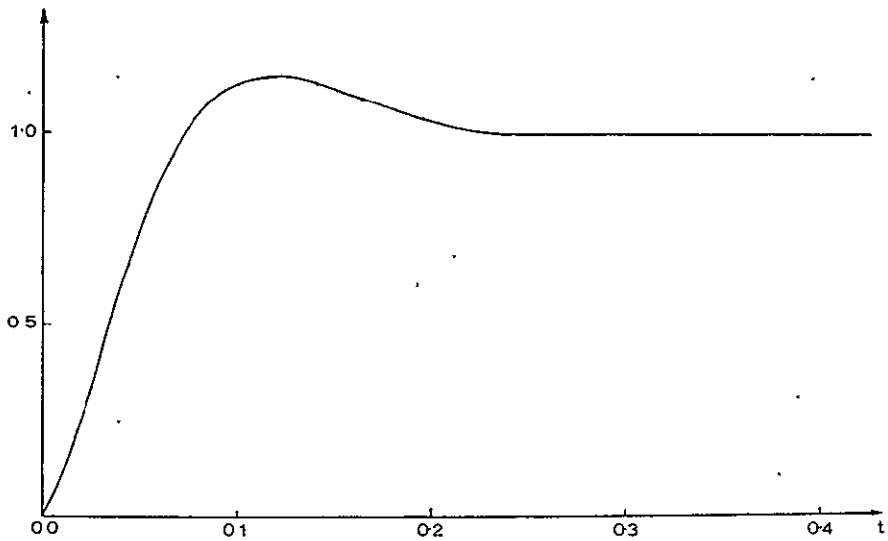
ORIGINAL PAGE IS
OF POOR QUALITY

The corresponding closed loop transfer function is

$$\frac{C}{R}(s)=\frac{100}{(s+1)(0{\cdot}63\sqrt{s}+1)+100}.$$

The computer programme evaluates the impulse response first by taking $2^{12}=4096$ points from the frequency response as the first step and then calculating A_k by (15). If the unit step response is desired, we simply perform numerical integration on the unit impulse response. The resultant curve is shown in fig. 3. This example was taken from Kilomeitseva and Netushil, and we found that their answer had a slight error.

Fig. 3



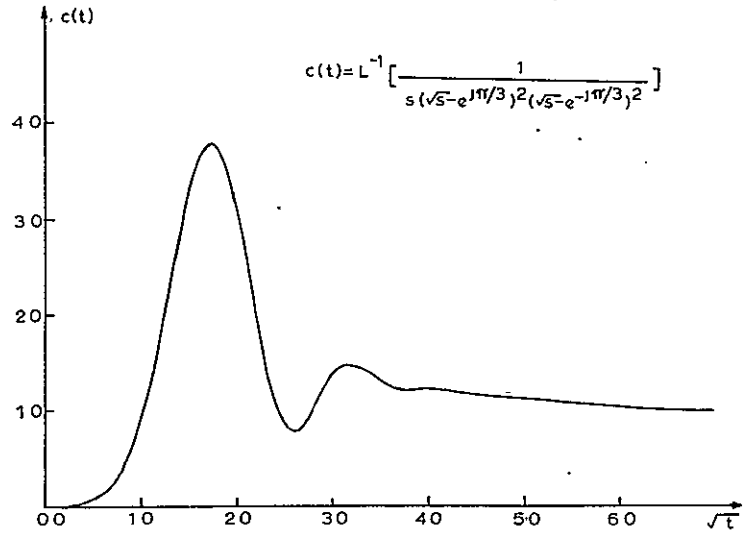
7. Repeat root case

Theoretically, there is no analytic formula available for evaluating a system which has repeated root, for example, the transfer function

$$\frac{1}{s(\sqrt{s}-\exp[j(\pi/3)])^2(\sqrt{s}-\exp[-j(\pi/3)])^2}$$

is a difficult case for Kilomeitseva and Netushil's method. However, if we use our new approach, we don't face any particular difficulties at all. We take $2^{13}=8192$ points from the frequency response and then obtain the answer as shown in fig. 4.

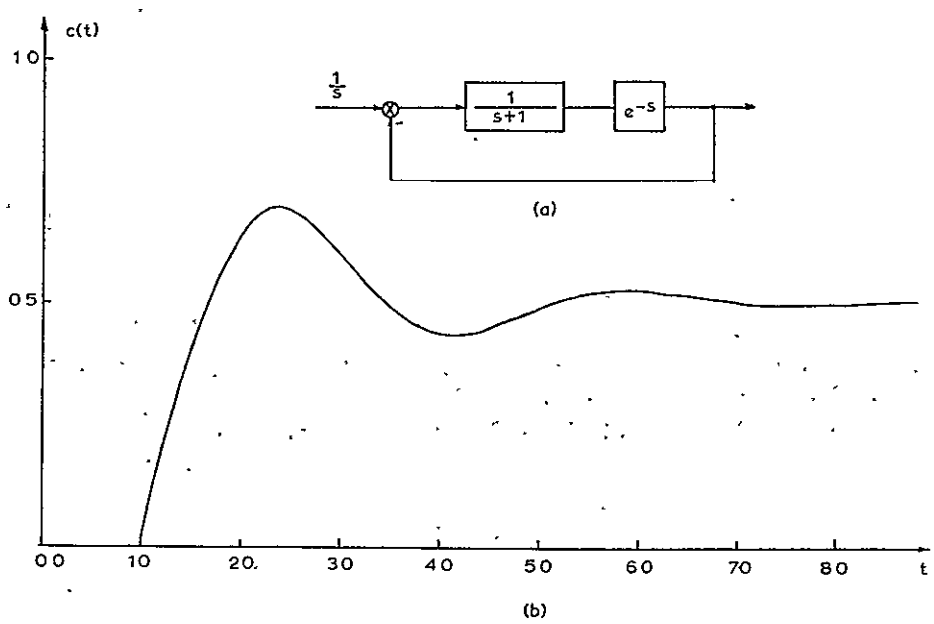
Fig. 4



8. Transcendental transfer function

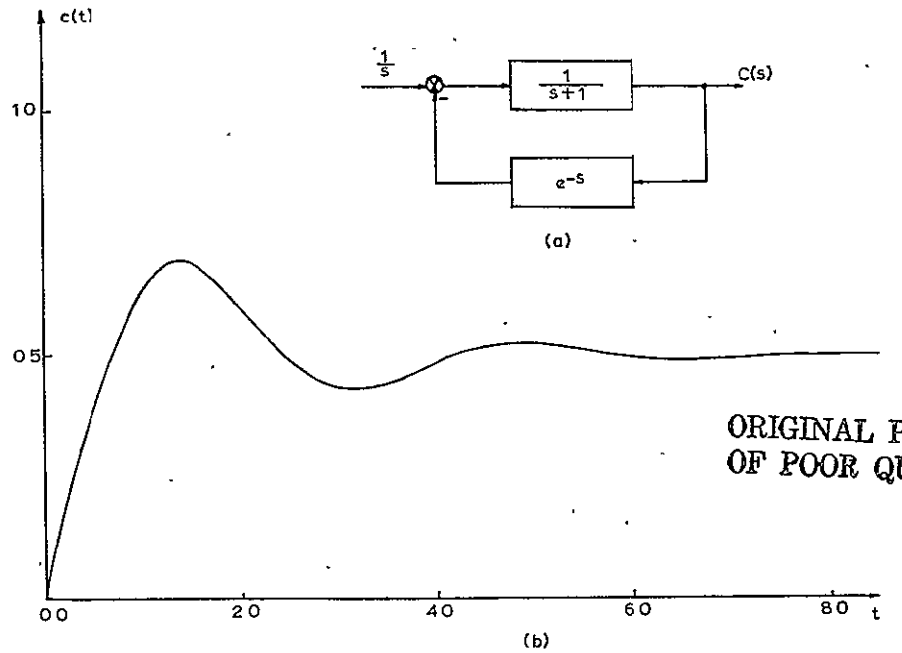
If a delay element is either in the feedforward loop (fig. 5 *a*), or in the feedback loop (fig. 6 *a*), and their responses are desired, the problem is very difficult. However, with our approach to solve, it is still routine. We take $2^{13}=8192$ points, the results are shown in fig. 5 *b* and 6 *b* respectively.

Fig. 5



p.102

Fig. 6



ORIGINAL PAGE IS
OF POOR QUALITY

9. Conclusion

The Fast Fourier transform technique was originated by mathematicians for evaluating Fourier coefficients and has been extended by communication engineers for application in spectra analysis and design. This paper establishes a technique for performing the inverse Laplace transform of irrational and transcendental transfer functions via the Fast Fourier transform.

References

Bohn, E. T., 1963, *The Transform Analysis of Linear Systems* (Reading, Mass.: Addison Wesley Publishing Co.), p. 281.

Brown, G. S., and Campbell, D. P., 1948, *Principles of Servomechanism* (New York: John Wiley and Sons), p. 334.

Campbell, D. P., 1958, *Process Dynamics* (New York: John Wiley and Sons), p. 104.

Carlson, G. E., and Haljak, C. A., 1964, *I.E.E.E. Trans. Circuit Theory*, **11**, 210.

Chen, C. F., 1966, *I.E.E.E. Int. Conv. Rec.*, Part 7, p. 281.

Chen, C. F., and Shieh, L. S., 1967, Analysis of Irrational Transfer Functions, I.E.E.E. Region III Record.

Chen, C. F., and Hass, I. J., 1968, *Elements of Control Systems Analysis* (Prentice Hall Publishing Company).

Cockran, W. T., , 1967, *I.E.E.E. Proc.*, **55**, 1667.

Cooley, J. W., and Tukey, J. W., 1965, *An Algorithm for the Machine Calculation of Complex Fourier Series of Computer* Vol. 19 (Association for Computing Machinery), p. 297.

Dubner, H., and Ahate, J., 1968, *Numerical Inversion of Laplace Transforms by Relating Them to the Finite Fourier Cosine Transform*, Vol. 15 (Association for Computing Machinery), p. 115.

p.103

KILOMEITSEVA, M. B., and NETUSHIL, A. V., 1965, English Translation of *Automatika i Telemekhanika*, **26**, 359.
LERNER, R. M., 1965, *I.E.E.E. Trans. Circuit Theory*, **10**, 98.
POPOULIS, A., 1964, *The Fourier Integral and its Applications* (New York : McGraw-Hill Book Company).
STEWART, J. L., 1960, *Inst. Radio Engrs*, **00**, 2003.
TRUXAL, J. G., 1955, *Automatic Feedback Control System Synthesis* (New York : McGraw-Hill Book Co.), p. 546.

Appendix

```
C     A COMPUTER PROGRAM FOR COMPUTING THE INVERSE LAPLASE TRANSFORM
      COMPLEX A,T1,DES,CFUN,EA,EB,S
      DIMENSION A(8200)
      READ(5,1) M,T
    1 FORMAT(I10,F15.5)
      WRITE(6,2)
    2 FORMAT(10X,'TIME',10X,'SQRT(T)',10X,'X(T)')
      N=2**M
      DET=T/N
      DES=(0.,1.)*2.*3.1416/T
      NH=N/2
      DO 100 I=1,N
      NA=I-1
      NN=0
      DO  200 J=1,M
      NN=NN*2
      NB=NA-NA/2*2
      NA=NA/2
  200 NN=NN+NB
      IF(I-NH) 250,250,260
  250 A(NN+1)=CFUN((I-1)*DES)
      GO TO 100
  260 A(NN+1)=CFUN((-N+I-1)*DES)
  100 CONTINUE
      DO 300 I=1,M
      IA=2**(M-I)
      L=2**(I-1)
      DO 400 J=1,IA
      DO 400 K=1,L
      T1=A(2*(J-1)*L+L+K)*CEXP((0., 1.)*2*3.1416*(K-1)/(2**I))
      A(2*(J-1)*L+L+K)=A(2*(J-1)*L+K)-T1
  400 A(2*(J-1)*L+K)=A(2*(J-1)*L+K)+T1
  300 CONTINUE
      NHH=(N-1)/2
      AL=0
      DO 500 I=1,NHH
      TM=2.*I*DET
      TR=SQRT(TM)
      AL=AL+(REAL(A(2*I-1))+4.*REAL(A(2*I))+REAL(A(2*I+1)))/(3.*N)
  500 WRITE(6,501) TM,TR,AL
  501 FORMAT(10X,3F16.6)
      STOP
      END

      COMPLEX FUNCTION CFUN(S)
      COMPLEX S,EA,EB
      EA=(0.,1.)*2*3.1416/360.*70.
      EB=EA
      CFUN=CSQRT(S)/((CSQRT(S)-CEXP(EA))*(CSQRT(S)-CEXP(EB)))
      RETURN
      END
```

p.104

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: New Theorems of Association of Variables in Multiple Dimension Laplace Transform

(B) Project Abstract:

In non-linear systems analysis, multiple dimension Laplace transform is often applied in solving the volterra model. The special technique for the inverse Laplace transform solution is called the association of variables. Three new theorems are developed for the theory of association. Compared with the inspection method of Brilliant, the pair listing method of Lubbock, the new approach is much simpler and easier and more systematic. Several illustrative examples are included.

(C) Publication: International Journal of Systems Science, 4, 647 (1974)

(D) Year: 1974

(E) Department: Electrical Engineering

(F) Student Name: R. F. Chiu

(G) Faculty Advisor: Professor C. F. Chen

INT. J. SYSTEMS SCI., 1973, VOL. 4, NO. 4, 647–664

New theorems of association of variables in multiple dimensional Laplace transform

C. F. CHEN

Electrical Engineering Department,
University of Houston, Houston, Texas, U.S.A.

and R. F. CHIU

International Communication Company,
Miami, Florida, U.S.A.

[Received 8 September 1972]

In non-linear systems analysis, multiple dimensional Laplace transform is often applied in solving the volterra model. The special technique for the inverse Laplace transform solution is called the association of variables. Three new theorems are developed for the theory of association. Compared with the inspection method of Brilliant, the pair listing method of Lubbock, the new approach is much simpler and easier and more systematic. Several illustrative examples are included.

1. Introduction

Suppose we have a function $F(s_1, s_2, \ldots, s_n)$; its n-dimensional inverse Laplace transform can be found by

$$f(t_1, t_2, \ldots, t_n) = \frac{1}{(2\pi j)^n} \int_{\alpha_1 - j\infty}^{\alpha_1 + j\infty} \cdots \int_{\alpha_n - j\infty}^{\alpha_n + j\infty} \exp\left(\sum_{i=1}^{n} s_i t_i\right) \cdot F(s_1, s_2, \ldots, s_n) \prod_{i=1}^{n} ds_i. \tag{1}$$

Formula (1) is a general one. In certain types of analysis, particularly in Volterra series applications (Volterra 1930, Wiener 1942) on non-linear systems (Brilliant 1958, Barrett 1963) we are only interested in the special case: $t_1 = t_2 = \ldots = t_n = t$. We denote this time function of the special case by $g(t)$, or

$$g(t) \triangleq f(t_1, t_2, \ldots, t_n)|_{t_1 = t_2 = \ldots = t_n = t}. \tag{2}$$

In single-dimensional Laplace transform, there must be a corresponding Laplace transform of $g(t)$, $G(s)$. Then we have the correspondence

$$G(s) = L[g(t)]. \tag{3}$$

Our problems in the special case can be restated as follows:

For given $F(s_1, s_2, \ldots, s_n)$, find $f(t, t, t, \ldots, t)$. There are two ways to accomplish this purpose:

1. Using (1) to find $f(t_1, t_2, \ldots, t_n)$ first and then substitute $t_1, t_2, \ldots, t_n$ by t.
2. From $F(s_1, s_2, \ldots, s_n)$ to find $G(s)$ first and then evaluate the single dimensional inverse Laplace transform $g(t)$ which is the answer.

The second approach is called association of variables. The function $G(s)$ is called the associated transform.

The following table can easily illustrate the relationship.

$$\begin{array}{ccc} F(s_1, s_2, \ldots, s_n) & \xrightarrow{\text{multiple } L^{-1}} & f(t_1, t_2, \ldots, t_n) \\ \Big\downarrow A_n & & \Big\downarrow t_1=t_2 \ldots = t \\ G(s) & \xrightarrow{\text{single } L^{-1}} & g(t) \end{array}$$

ORIGINAL PAGE IS
OF POOR QUALITY

where A_n means the association process for finding $G(s)$ form $F(s_1, s_2, \ldots, s_n)$. Before developing the associate variable theory, let us familiarize ourselves with the inversion process by using the direct inversion formula (1).

Example 1

Given

$$\left.\begin{aligned} F(s_1, s_2) &= \frac{1}{(s_1+1)(s_2^2+3s_2+2)} \\ \text{find} \quad & \\ f(t_1, t_2)|_{t_1=t_2=t} \text{ or } & g(t). \end{aligned}\right\} \tag{4}$$

By (1)

$$f(t_1, t_2)|_{t_1=t_2=t} = \frac{1}{(2\pi j)^2} \int_{\alpha_1-j\infty}^{\alpha_1+j\infty} \int_{\alpha_2-j\infty}^{\alpha_2+j\infty} \frac{\exp(s_1t+s_2t)\, ds_1\, ds_2}{(s_1+1)(s_2^2+3s_2+2)},$$

$$= \left(\frac{1}{2\pi j} \int_{\alpha_1-j\infty}^{\alpha_1+j\infty} \frac{\exp(s_1t)\, ds_1}{s_1+1}\right)\left(\frac{1}{2\pi j} \int_{\alpha_2-j\infty}^{\alpha_2+j\infty} \frac{\exp(s_2t)\, ds_2}{s_2^2+3s_2+2}\right). \tag{5}$$

Each one is a single-dimensional inverse Laplace transform. Therefore the inversions can be evaluated individually and we obtain

$$= [\exp(-t)] \,.\, [\exp(-t) - \exp(-2t)] = \exp(-2t) - \exp(-3t) = g(t). \tag{6}$$

For this simple example, we did not encounter difficulties, however, for more complicated problems, the direct inversion method becomes laborious.

2. Theorems of association of variables

To find $G(s)$ from $F(s_1, s_2, \ldots, s_n)$ directly is by no means easy as we have seen in the previous example. Originally, George (1958, 1959) used the method of inspection which is not a systematic method at all. Flake (1962) followed George and still used the inspection method to do the work. Recently, Lubbock and Bansal (1969) developed a set of pairs to show the association correspondence. However, Lubbock and Bansal's method is by no means general. Particularly, in all their association pairs, there is no one with numerator dynamics. In other words, they cannot solve the problems involving initial conditions.

p.107

In this paper, we shall develop the basic theorems for the associate variables. Once the fundamental theorems have been established, we can produce as many associate pairs as we want, and use them flexibly.

3. Complex translation theorem

If $F(s_1, s_2, \ldots, s_n)$ can be written in the following factorial form

$$F(s_1, s_2, \ldots, s_n) = \frac{1}{s_k + a} F_1(s_1, s_2, \ldots, s_{k-1}, s_{k+1}, s_n), \tag{7}$$

and if

$$F_1(s_1, s_2, \ldots, s_{k-1}, s_{k+1}, s_n) \xrightarrow{A_{n-1}} G_1(s), \tag{8}$$

we have the theorem as follows:

$$F(s_1, s_2, \ldots, s_n) \xrightarrow{A_n} G(s) = G_1(s+a) \tag{9}$$

Proof

By eqns. (1) and (2), we have

$$\begin{aligned} g(t) &= f(t_1, t_2, \ldots, t_n)|_{t_1=t_2=\ldots=t_n=t} \\ &= \frac{1}{(2\pi j)^n} \int_{\alpha_1 - j\infty}^{\alpha_1 + j\infty} \cdots \int_{\alpha_n - j\infty}^{\alpha_n + j\infty} F(s_1, s_2, \ldots, s_n) \exp\left(\sum_{i=1}^{n} s_i t\right) \prod_{i=1}^{n} ds_1 \\ &= \frac{1}{(2\pi j)} \int_{\alpha_k - j\infty}^{\alpha_k + j\infty} \frac{1}{s_k + a} \exp(s_k t)\, ds_k \\ &\quad \cdot \frac{1}{(2\pi j)^{n-1}} \int_{\alpha_1 - j\infty}^{\alpha_1 + j\infty} \cdots \int_{\alpha_l - j\infty}^{\alpha_l + j\infty} \cdots \int_{\alpha_n + j\infty}^{\alpha_n - j\infty} F(s_1, \ldots, s_{k-1}, s_{k+1}, s_n) \\ &\qquad \cdot \exp\left(\sum_{\substack{i=1 \\ i \neq k}}^{n} s_i t\right) \prod_{\substack{i=1 \\ i \neq k}}^{n} ds_i = \exp(-at) g_1(t). \end{aligned} \tag{10}$$

Laplace transforming both sides of (10) yields

$$G(s) = L[\exp(-at) g_1(t)] = G_1(s+a).$$

Theorem is proven.

Example 2

Let us consider the previous example again, but use the association variables to solve it.

Given

$$F(s_1, s_2) = \frac{1}{(s_1+1)(s_2^2 + 3s_2 + 2)}. \tag{11}$$

find

$$f(t, t) = g(t).$$

ORIGINAL PAGE IS
OF POOR QUALITY

Let

$$F_1(s_2) = \frac{1}{s_2^2 + 3s_2 + 2}. \tag{12}$$

Association transforming yields

$$F_1(s_2) \xrightarrow{A_1} \frac{1}{s^2 + 3s + 2}, \tag{13}$$

which is identical to (12), because there is only one variable involved.

We then rewrite (11) as follows:

$$F(s_1, s_2) = \frac{1}{s_1 + 1} \cdot \frac{1}{(s_2^2 + 2s_2 + 2)}$$

$$= \frac{1}{s_1 + 1} F_1(s_2). \tag{14}$$

This is the standard form for applying the complex translation theorem. Performing the association process gives:

$$\xrightarrow{A_2} = \frac{1}{(s+1)^2 + 3(s+1) + 2}.$$

Simplifying

$$= \frac{1}{s^2 + 5s + 6}. \tag{15}$$

This is a single inverse Laplace transform problem, the answer is easily obtained as follows:

$$g(t) = \exp(-2t) - \exp(-3t),$$

which is also the answer of our problem.

Example 3

Find the inverse Laplace transform of

$$F(s_1, s_2, s_3) = \frac{1}{(s_1 + 1)(s_2^2 + 3s_2 + 2)(s_3 + 2)}, \tag{16}$$

if

$$t_1 = t_2 = t_3 = t.$$

Again, we use the association of variables by writing (16) in two parts:

$$F(s_1, s_2, s_3) = \frac{1}{(s_1 + 1)(s_2^2 + 3s_2 + 2)} \cdot \frac{1}{s_3 + 2} \tag{17}$$

p.109

or

$$= F_1(s_1, s_2) \cdot \frac{1}{s_3+2}.$$

Based on the results obtained in the previous example, we know

$$F_1(s_1, s_2) \xrightarrow{A_2} \frac{1}{s^2+5s+6},$$

then

$$F(s_1, s_2, s_3) = \frac{1}{s_3+2} F_1(s_1, s_2),$$

$$\xrightarrow{A_2} \frac{1}{(s+2)^2+5(s+2)+6}$$

$$= \frac{1}{s^2+9s+20}. \tag{18}$$

The single-dimensional inverse Laplace transform of (18) can be easily obtained even by inspection.

4. Complex convolution theorem

Suppose the function $F(s_1, s_2, \ldots, s_n)$ can be factored out into the following form,

$$F(s_1, s_2, \ldots, s_n) = F_1(s_1, s_2, \ldots, s_m) \cdot F_2(s_{m+1}, s_{m+2}, \ldots, s_n). \tag{19}$$

The complex convolution theorem states that

$$G(s) = G_1(s) \circledast G_2(s), \tag{20}$$

where $G_1(s)$ and $G_2(s)$ are defined as

$$F_1(s_1, s_2, \ldots, s_m) \xrightarrow{A_m} G_1(s)\,; \quad F_2(s_{m+1}, \ldots, s_n) \xrightarrow{A_{n-m}} G_2(s), \tag{21}$$

and $\circledast$ means complex convolution

$$G_1(s) \circledast G_2(s) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} G_1(s-w) G_2(w)\, dw,$$

with a suitable c.

Proof

$$g(t) = f(t_1, t_2, \ldots, t_n)|_{t_1=t_2=\ldots=t_n=t}.$$

By definition

$$= \frac{1}{(2\pi j)^n} \int_{\alpha_1-j\infty}^{\alpha_1+j\infty} \cdots \int_{\alpha_n-j\infty}^{\alpha_n+j\infty} F(s_1, s_2, \ldots, s_n) \exp\left(\sum_{i=1}^{n} s_i t\right) \prod_{i=1}^{n} ds_i$$

Substituting the F function by its product form, we get

ORIGINAL PAGE IS OF POOR QUALITY

$$=\left[\frac{1}{(2\pi j)^m}\int_{\alpha_1-j\infty}^{\alpha_1+j\infty}\cdots\int_{\alpha_m-j\infty}^{\alpha_m+j\infty}F_1(s_1,\ldots,s_m)\cdot\exp\left(\sum_{i=1}^{m}s_it\right)\prod_{i=1}^{m}ds_i\right]$$

$$\cdot\left[\frac{1}{(2\pi j)^{n-m}}\int_{\alpha_{m+1}-j\infty}^{\alpha_{m+1}+j\infty}\cdots\int_{\alpha_n-j\infty}^{\alpha_n+j\infty}F_2(s_{m+1},\ldots,s_n)\right.$$

$$\left.\times\exp\left(\sum_{i=m+1}^{n}s_it_i\right)\prod_{i=m+1}^{n}ds_i\right]=g_1(t)\cdot g_2(t). \qquad (22)$$

Laplace transforming both sides of (22), we obtain the result

$$G(s)=G_1(s)\circledast G_2(s).$$

The theorem is proven.

Example 2

Use the Complex Convolution Theorem to take the inverse Laplace transform of the previous example

$$F(s_1, s_2)=\frac{1}{(s_1+1)(s_2{}^2+3s_2+2)}$$

under the special condition, $t_1=t_2=t$.

$$F(s_1, s_2)=\frac{1}{(s_1+1)(s_2{}^2+3s_2+2)}.$$

Writing into the product form

$$=\frac{1}{(s_1+1)}\cdot\frac{1}{(s_2{}^2+3s_2+2)}.$$

Defining

$$=F_1(s_1)\cdot F_2(s_2), \qquad (23)$$

where

$$F_1(s_1)=\frac{1}{s_1+1}\xrightarrow{A_1}\frac{1}{s+1}=G_1(s), \qquad (24)$$

and

$$F_2(s)=\frac{1}{(s_2+1)(s_2+2)}\xrightarrow{A_1}\frac{1}{(s+1)(s+2)}=G_2(\qquad 25)$$

p. 111

Applying the complex convolution theorem

$$G(s)=\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} G_1(s-w)G_2(w)\,dw$$

$$=\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty} \frac{1}{s-w+1}\cdot\frac{1}{(w+1)(w+2)}\,dw.$$

Using the standard residue method yields

$$=\frac{1}{2\pi j}\,2\pi j\left(\frac{1}{s+1+1}+\frac{-1}{s+1+2}\right), \tag{26}$$

$$G(s)=\frac{1}{(s^2+5s+6)}, \tag{27}$$

$$g(t)=\exp(-2t)-\exp(-3t).$$

Example 5

Use the complex convolution theorem to find the associated transform of the following:

$$F(s_1, s_2, s_3)=\frac{1}{(s_1+1)(s_2{}^2+3s_2+2)(s_3+2)}.$$

We define

$$F_1(s_1, s_2)=\frac{1}{(s_1+1)(s_2{}^2+3s_2+2)}\xrightarrow{A_1} G_1(s), \tag{28}$$

$$F_2(s_3)=\frac{1}{s_3+2}\xrightarrow{A_1}\frac{1}{s+2}=G_2(s). \tag{29}$$

Applying the theorem, we have

$$G(s)=G_1(s)\circledast G_2(s).$$

From the previous example, we found that

$$G_1(s)=\frac{1}{(s+2)(s+3)}. \tag{30}$$

Convoluting $G_2(s)$ with eqn. (30), we obtain

$$F(s)=\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty}\frac{1}{s-w+2}\cdot\frac{1}{(w+2)(w+3)}\,dw.$$

Using the residue method again, we finally have

$$F(s)=\frac{1}{2\pi j}\cdot 2\pi j\left(\frac{1}{s+2+2}+\frac{-1}{s+3+2}\right)=\frac{1}{(s+4)(s+5)}. \tag{31}$$

p. 112

ORIGINAL PAGE IS
OF POOR QUALITY

5. Real convolution theorem

If a given function $F(s_1, s_2, \ldots, s_n)$ has the following particular structure:

$$F(s_1, s_2, \ldots, s_n) = H(s_1+s_2+\ldots+s_n)F_1(s_1, s_2, \ldots, s_n), \tag{32}$$

and

$$F_1(s_1, s_2, \ldots, s_n) \xrightarrow{A_n} G_1(s), \tag{33}$$

the real convolution theorem states that

$$F(s_1, s_2, \ldots, s_n) \xrightarrow{A_n} G(s) = H(s)G_1(s). \tag{34}$$

Proof

In multiple Laplace transformation theory it can be shown that (Chiu 1971)

$$H(s_1+s_2+\ldots+s_n)F_1(s_1, s_2, \ldots, s_n)$$
$$= \int_0^\infty \ldots \int_0^\infty \exp\left(-\sum_{i=1}^n s_i t_i\right) f(t_1, t_2, \ldots, t_n) \prod_{i=1}^n dt_i, \tag{35}$$

where

$$f(t_1, t_2, \ldots, t_n) = \int_0^{\min(t_1, t_2, \ldots, t_n)} h(\tau) f_1(t_1-\tau, t_2-\tau, \ldots, t_n-\tau)\, d\tau, \tag{36}$$

and $h(t)$, $f_1(t_1, t_2, \ldots, t_n)$ are the inverse Laplace transforms of $H(s)$ and $F_1(s_1, s_2, \ldots, s_n)$, respectively.

Based on the inversion formula (1), we have

$$f(t_1, t_2, \ldots, t_n) = \frac{1}{(2\pi j)^n} \int_{\alpha_1-j\infty}^{\alpha_1+j\infty} \ldots \int_{\alpha_n-j\infty}^{\alpha_n+j\infty} \exp\left(\sum_{i=1}^n s_i t_i\right)$$
$$\cdot F(s_1, s_2, \ldots, s_n) \prod_{i=1}^n ds_i$$
$$= \int_0^{\min(t_1, t_2, \ldots, t_n)} h(\tau) f_1(t_1-\tau, t_2-\tau, \ldots, t_n-\tau)\, d\tau. \tag{37}$$

By definition

$$g(t) = f(t_1, t_2, \ldots, t_n)|_{t_1=t_2=\ldots=t_n=t}$$
$$= f(t, t, \ldots, t) = \int_0^t h(\tau) f_1(t-\tau, t-\tau, \ldots, t-\tau)\, d\tau$$
$$= \int_0^t h(\tau) g_1(t-\tau)\, d\tau. \tag{38}$$

By the theory of convolution, we have

$$G(s) = L[g(t)] = L[h(t)] \,.\, L[g_1(t)] = H(s)G_1(s). \tag{39}$$

Example 6

Find the associated transform of

$$F(s_1, s_2) = \frac{1}{(s_1+s_2)^2+2(s_1+s_2)+1} \cdot \frac{1}{(s_1+1)(s_2^2+3s_2+2)}$$
$$= H(s_1+s_2)F_1(s_1, s_2), \tag{40}$$

where

$$F_1(s_1, s_2) = \frac{1}{(s_1+1)(s_2^2+3s_2+2)},$$

and

$$H(s_1+s_2) = \frac{1}{(s_1+s_2)^2+2(s_1+s_2)+1}.$$

We have found $G_1(s)$ in Example 4 as

$$F(s_1, s_2) \xrightarrow{A_s} G_1(s) = \frac{1}{(s+2)(s+3)}. \tag{41}$$

Applying the real convolution theorem, we have

$$F(s_1, s_2) \xrightarrow{A_s} \frac{1}{s^2+2s+1} G_1(s) = \frac{1}{(s+1)^2(s+2)(s+3)}. \tag{42}$$

Example 7

Find the associated transform of

$$F(s_1, s_2, s_3) = \frac{1}{(s_1+s_2+s_3)^2+4(s_1+s_2+s_3)+3}$$
$$\cdot \frac{1}{(s_1+1)(s_2^2+3s_2+2)(s_3+2)}$$
$$= H(s_1+s_2+s_3)F_1(s_1, s_2, s_3), \tag{43}$$

where

$$F_1(s_1, s_2, s_3) = \frac{1}{(s_1+1)(s_2^2+3s_2+2)(s_3+2)},$$

$$H(s_1+s_2+s_3) = \frac{1}{(s_1+s_2+s_3)^2+4(s_1+s_2+s_3)+3}.$$

We found in Example 5 that

$$F_1(s_1, s_2, s_3) \xrightarrow{A_s} G_1(s) = \frac{1}{(s+4)(s+5)}. \tag{44}$$

p.114

ORIGINAL PAGE IS
OF POOR QUALITY

By applying the real convolution theorem, we have

$$F(s_1, s_2, s_3) \xrightarrow{A_3} \frac{1}{s^2+4s+3} G_1(s) = \frac{1}{(s+1)(s+3)} \cdot \frac{1}{(s+4)(s+5)}$$

$$= \frac{1}{(s+1)(s+3)(s+4)(s+5)}. \tag{45}$$

6. Conclusions

The three theorems established in the previous sections are rigorous and powerful in performing the inverse Laplace transform under the special conditions. Because they consist of a systematic approach to the problem, the guess work, inspection work and tedious tables can be eliminated. For the user's convenience, we still derive some pairs as an appendix which are mainly derived by using the three theorems.

To apply the technique is not a difficult matter, once we grasp the idea and flexibly use the theorems. Several simple as well as complicated examples are included.

p.115

Appendix

	Multiple dimensional Laplace transform	Associated transform
(1)	$\dfrac{K}{(s_1+a)(s_2+b)}$	$\dfrac{K}{(s+a+b)}$
(2)	$\dfrac{K}{a(s_1+s_2)^2+b(s_1+s_2)+c}$	$\dfrac{K}{as^2+bs+c}$
(3)	$\dfrac{K}{[c(s_1+s_2)^2+d(s_1+s_2)+e](s_1+a)(s_2+b)}$	$\dfrac{K}{(cs^2+ds+e)(s+a+b)}$
(4)	$\dfrac{K}{(s_1+a)(s_1+b)(s_2+a)(s_2+b)}$	$\dfrac{2K}{(s+a+b)(s+2a)(s+2b)}$
(5)	$\dfrac{K}{[c(s_1+s_2)^2+d(s_1+s_2)+e]}\times\dfrac{1}{(s_1+a)(s_1+b)(s_2+a)(s_2+b)}$	$\dfrac{2K}{(cs^2+ds+e)(s+a+b)(s+2a)(s+2b)}$
(6)	$\dfrac{K(s_1+c)(s_2+c)}{(s_1+a)(s_1+b)(s_2+a)(s_2+b)}$	$\dfrac{K}{(b-a)^2}\left[\dfrac{(c-a)^2}{s+2a}-\dfrac{2(c-a)(c-b)}{s+a+b}+\dfrac{(c-b)^2}{s+2b}\right]$
(7)	$\dfrac{K(s_1+c)(s_2+c)}{d(s_1+s_2)^2+e(s_1+s_2)+f}\times\dfrac{1}{(s_1+a)(s_1+b)(s_2+a)(s_2+b)}$	$\dfrac{K}{(b-a)^2}\cdot\dfrac{1}{ds^2+es+f}\cdot\left[\dfrac{(c-a)^2}{s+2a}-\dfrac{2(c-a)(c-b)}{s+a+b}+\dfrac{(c-b)^2}{s+2b}\right]$
(8)	$\dfrac{K}{a(s_1+s_2+s_3)^2+b(s_1+s_2+s_3)+c}$	$\dfrac{K}{as^2+bs+c}$
(9)	$\dfrac{K}{(s_1+a)(s_2+b)(s_3+c)}$	$\dfrac{K}{s+a+b+c}$

p.116

	Multiple dimensional Laplace transform	Associated transform
(10)	$\dfrac{K}{[d(s_1+s_2)^2+e(s_1+s_2)+f]}\times\dfrac{1}{(s_1+a)(s_2+b)(s_3+c)}$	$\dfrac{K}{[d(s+c)^2+e(s+c)+f](s+a+b+c)}$
(11)	$\dfrac{K}{[h(s_1+s_2+s_3)^2+p(s_1+s_2+s_3)+q]}\times\dfrac{1}{[d(s_1+s_2)^2+e(s_1+s_2)+f](s_1+a)(s_2+b)(s_3+c)}$	$\dfrac{K}{(hs^2+ps+q)[d(s+c)^2+e(s+c)+f](s+a+b+c)}$
(12)	$\dfrac{K}{\prod_{i=1}^{3}(s_i+a)(s_i+b)}$	$\dfrac{K}{(a-b)^3}\left(\dfrac{1}{s+3b}-\dfrac{3}{s+2a+b}+\dfrac{3}{s+a+2b}-\dfrac{1}{s+3a}\right)$
(13)	$K\prod_{i=1}^{3}\dfrac{s_i+c}{(s_i+a)(s_i+b)}$	$\dfrac{K}{(a-b)^3}\left[\dfrac{(a-c)^3}{s+3a}-\dfrac{3(a-c)^2(b-c)}{s+2a+b}+\dfrac{3(a-c)(b-c)^2}{s+a+2b}-\dfrac{(c-b)^3}{s+3b}\right]$
(14)	$\dfrac{K}{[c(s_1+s_2+s_3)^2+d(s_1+s_2+s_3)+e]}\times\dfrac{1}{\prod_{i=1}^{3}s_i+a)(s_i+b)}$	$\dfrac{K}{(a-b)^3}\cdot\dfrac{1}{cs^2+ds+e}\left(\dfrac{1}{s+3b}-\dfrac{3}{s+2a+b}+\dfrac{3}{s+a+2b}-\dfrac{1}{s+3a}\right)$
(15)	$\dfrac{K}{[d(s_1+s_2+s_3)^2+e(s_1+s_2+s_3)+f]}\times\prod_{i=1}^{3}\dfrac{s_i+c}{(s_i+a)(s_i+b)}$	$\dfrac{K}{(a-b)^3}\cdot\dfrac{1}{ds^2+es+f}\left[\dfrac{(a-c)^3}{s+3a}-\dfrac{3(a-c)^2(b-c)}{s+2a+b}+\dfrac{3(a-c)(b-c)^2}{s+a+2b}-\dfrac{(c-b)^3}{s+3b}\right]$
(16)	$\dfrac{K}{(s_1+a)^2(s_2+a)^2(s_3+a)^2}$	$\dfrac{6K}{(s+a)^4}$

B-117

ORIGINAL PAGE IS
OF POOR QUALITY

Multiple dimensional Laplace transform	Associated transform
(17) $\dfrac{K(s_1+b)(s_2+b)(s_3+b)}{(s_1+a)^2(s_2+a)^2(s_3+a)^2}$	$K\left[\dfrac{1}{s+3a}-\dfrac{3(a-b)}{(s+3a)^2}+\dfrac{6(a-b)^2}{(s+3a)^3}-\dfrac{6(a-b)^3}{(s+3a)^4}\right]$
(18) $\dfrac{K}{[b(s_1+s_2+s_3)^2+c(s_1+s_2+s_3)+d]}\times\dfrac{1}{\prod_{i=1}^{3}(s_i+a)^2}$	$\dfrac{6K}{(bs^2+cs+e)(s+a)^4}$
(19) $\dfrac{K\prod_{i=1}^{3}(s_i+b)}{[c(s_1+s_2+s_3)^2+d(s_1+s_2+s_3)+e]}\times\dfrac{1}{\prod_{i=1}^{3}(s_i+a)^2}$	$\dfrac{K}{cs^2+ds+e}\left[\dfrac{1}{s+3a}-\dfrac{3(a-b)}{(s+3a)^2}+\dfrac{6(a-b)^2}{(s+3a)^3}-\dfrac{6(a-b)^3}{(s+3a)^4}\right]$
(20) $\dfrac{K}{[c(s_1+s_2)+d(s_1+s_2)+e]\prod_{i=1}^{3}(s_i+a)(s_i+b)}$	$\dfrac{2K}{a-b}\left[\dfrac{1}{[c(s+b)^2+d(s+b)+e](s+2a+b)(s+a+2b)(s+3b)}-\dfrac{1}{[c(s+a)^2+d(s+a)+e](s+3a)(s+2a+b)(s+a+2b)}\right]$
(21) $\dfrac{K}{[f(s_1+s_2+s_3)^2+g(s_1+s_2+s_3)+h]}\times\dfrac{1}{[c(s_1+s_2)^2+d(s_1+s_2)+e]\prod_{i=1}^{3}(s_i+a)(s_i+b)}$	$\dfrac{\frac{2K}{a-b}}{fs^2+gs+h}\left[\dfrac{1}{[c(s+b)^2+d(s+b)+e]}\times\dfrac{1}{(s+2a+b)(s+a+2b)(s+3b)}-\dfrac{1}{[c(s+a)^2+d(s+a)+e](s+3a)(s+2a+b)(s+a+2b)}\right]$

A.11.8

References

Barrett, J. F., 1963, *J. Electron. Control*, **15**, 567.

Brilliant, M. B., 1958, *Theory of the Analysis of Non-linear Systems*, Report 345 (Research Laboratory of Electronics, M.I.T.).

Chiu, R. F., 1971, University of Houston Ph. D. Dissertation.

Flake, R. . H., 1962, *A.I.E.E. Trans.*, 1962, 330.

George, D. A., 1958, *An Algebra of Continuous Systems*, Quarterly Progress Report (Research Laboratory of Electronics, M.I.T.) ; 1959, *Continuous Non-linear Systems*, Report 355 (Research Laboratory of Electronics, M.I.T.).

Lubbock, J. K., and Bansal, V. S., 1969, *Proc. I.E.E.*, **116**, 2075.

Wiener, N., 1942, *Response of a Non-linear Device to Noise*, Report 129 (Radiation Laboratory, M.I.T.).

Volterra, V., 1930, *Theory of Functionals and of Integral and Integro-differential Equations* (London : Blackie & Sons).

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: A General Frequency Stability Criterion for Multi-Input-Output, Lumped and Distributed-Parameter Feedback Systems

(B) Project Abstract:

The original Nyquist criterion is based on the comparison of the encirclement of the frequency plot of the return ratio function with the number of poles and the number of zeros of the same function to determine the closed-loop stability of a feedback system. The extensions of the return ratio idea to the stability study of multi-variable feedback systems have used the same terminology and followed a similar course. For the multi-input-output case, the use of the Nyquish criterion or its extension is by no means a simple matter. This paper establishes a new frequency stability criterion which converts the Nyquist criterion from a return ratio oriented approach to a return difference oriented one. Instead of examining the encirclement of the return ratio function to a critical point, we examine the phase change of the positive frequency of the return difference function, and the number of zeros of the positive frequency of the return difference function. This result simplifies the stability study of multi-input-output lumped systems tremendously, and covers multi-input-output distributed-parameters systems naturally. For illustration,

several typical examples - single-input-output feedback systems with minimum phase or non-minimum phase open-loop transfer functions, multi-input-output feedback systems with stable or unstable open-loop transfer matrices, multi-input-output feedback systems with irrational or transendental type distributed-parameter open-loop transfer matrices - are included.

(C) Publication: International Journal of Control, 23, No. 3, 341 (1976)

(D) Year: Completed in 1974

(E) Department: Electrical Engineering

(F) Student Name: Y. T. Tsay

(G) Faculty Advisor: Professor C. F. Chen

INT. J. CONTROL, 1976, VOL. 23, NO. 3, 341–359

A general frequency stability criterion for multi-input–output, lumped and distributed-parameter feedback systems

C. F. CHEN† and Y. T. TSAY†

The original Nyquist criterion is based on the comparison of the encirclement of the frequency plot of the return ratio function with the number of poles and the number of zeros of the same function to determine the closed-loop stability of a feedback system. The extensions of the return ratio idea to the stability study of multivariable feedback systems have used the same terminology and followed a similar course. For the multi-input–output case, the use of the Nyquist criterion or its extension is by no means a simple matter. This paper establishes a new frequency stability criterion which converts the Nyquist criterion from a return ratio oriented approach to a return difference oriented one. Instead of examining the encirclement of the return ratio function to a critical point, we examine the phase change of the positive frequency of the return difference function, and the number of zeros of the positive frequency of the return difference function. This result simplifies the stability study of multi-input–output lumped systems tremendously, and covers multi-input–output distributed-parameters systems naturally. For illustration, several typical examples—single-input–output feedback systems with minimum phase or non-minimum phase open-loop transfer functions, multi-input–output feedback systems with stable or unstable open-loop transfer matrices, multi-input–output feedback systems with irrational or transcendental type distributed-parameter open-loop transfer matrices—are included.

1. Introduction

Since Chen (1968) and Hsu and Chen (1968) generalized the scalar return difference formula to a matrix form, much light has been shed on the analysis of multivariable feedback systems. Indeed, the return difference idea is fundamental for all feedback theories, serving as a quantitative measure of the various consequences of the use of feedback (Truxal 1955).

Historically, Bode (1945) originated the return difference idea. Kalman (1964) showed that the return difference gives the simplest presentation of the optimal condition. Chen and Tsang (1965, 1967) established a stability criterion for a single-input–output feedback system based on the return difference. It is known that Nyquist's original work is based on the return ratio, and not on the return difference. For the single-input–output case, the differentiation between using the return ratio or using the return difference is trivial and insignificant. Researchers and engineers, therefore, have followed Nyquist's derivation, terminology and formulation closely and faithfully. For several decades, the return ratio has always been used and the return difference has only been mentioned. For the study of multivariable feedback systems, it has been the same tradition. Rosenbrock's (1969) investigation of the Nyquist method, which starts with the return difference but ends with the return ratio, is a typical example. MacFarlane's (1970) paper on the return difference, which is entitled the return difference but essentially uses

Received 27 June 1975.

† Electrical Engineering Department, University of Houston, Houston, Texas 77004.

p.122

ORIGINAL PAGE IS
OF POOR QUALITY

return ratio, is another. However, to extend the return difference from the single-input–output case to the multi-input–output case is natural but to generalize the return ratio is artificial. Although Chen's establishment of the matrix return difference formula gives a fresh starting point, there is no proper development for investigating the stability of multivariable feedback systems. Because of this, Pontryagin's stability criterion for time-delay feedback systems cannot be generalized to multivariable systems. Desoer's stability theorem (1975) has limited applications, Brin's (1962) stability criterion is applicable only to a very special class. In the meantime, practical engineers badly need a unified stability criterion for a general multi-input–output feedback system.

This paper establishes a frequency stability criterion which uses the comparison of the number of poles of the open-loop transfer matrix with the phase change of the return difference. The criterion is applicable to (1) single classical feedback systems, (2) multi-input–output feedback systems, (3) feedback systems with non-minimum open-loop transfer matrices or with minimum ones, (4) feedback systems with irrational transfer elements and/or transcendental transfer elements. In all, it is a general frequency stability criterion for lumped and distributed-parameter multi-input–output feedback systems.

2. Stability criterion based on return difference

Consider the typical feedback system shown in Fig. 1 in which $g(s)$ is the feedforward transfer function; $h(s)$ the feedback transfer function. The product

$$g(s)h(s) \triangleq q(s) \tag{1}$$

is called the return ratio and

$$p(s) = 1 + q(s) \tag{2}$$

is called the return difference.

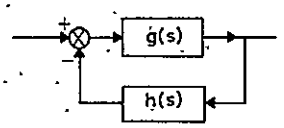


Figure 1. A typical single-input–output feedback system.

The differences between (1) and (2) are trivial indeed for the single-input–output case.

Assume that

$$q(s) = \frac{a(s)}{b(s)} \tag{3}$$

where $a(s)$ and $b(s)$ are relative prime polynomials.

Then

$$p(s) = \frac{a(s) + b(s)}{b(s)} \tag{4}$$

p. 123

The closed-loop transfer function is given by

$$f(s) = \frac{g(s)}{1 + g(s)h(s)} \tag{5}$$

$$= \frac{g(s)b(s)}{a(s) + b(s)} \tag{6}$$

If there is no common factor between the denominator $g(s)$ and the numerator of $h(s)$, then $g(s)b(s)$ is a polynomial. Therefore $b(s)$ is the open-loop characteristic polynomial, or the open-loop characteristic function in general; and $a(s)+b(s)$ is the closed-loop polynomial, or the closed-loop characteristic function in general. The return difference is then

$$p(s) = \frac{a(s)+b(s)}{b(s)} \quad \frac{\text{closed-loop characteristic function}}{\text{open-loop characteristic function}} \tag{7}$$

The multi-input–output case is shown in Fig. 2.

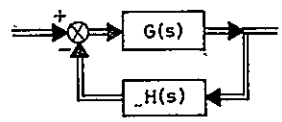


Figure 2. A typical multi-input–output feedback system.

Chen (1968) published and then proved (Hsu and Chen 1968) the classical formula which states that the return difference of a multi-input–output system is

$$p(s) = \det\{1 + GH\} = \frac{\text{closed-loop characteristic function}}{\text{open-loop characteristic function}} \tag{8}$$

where $G(s)$ and $H(s)$ are forward and feedback transfer matrices respectively.

If the return ratio of the multi-input–output system is defined from an extension of the simple-input–output case, we have

$$q(s) = \det[GH] \tag{9}$$

We see that the relationship between the return difference and the return ratio for the multi-input–output system is not as trivial as that for the single-input–output system.

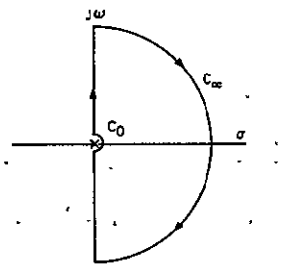
Therefore, when we analyse a multi-input–output system by using the Nyquist criterion, it will make a great difference whether we use the return difference or the return ratio. Unfortunately, Rosenbrock and MacFarlane although extending the return difference formula to design, they really used a return ratio approach. Desoer established a theorem to use the return difference in the stability study to a rather larger class; however, he still maps the whole imaginary axis and semicircles as Nyquist did and also limits the scope of applications. Apparently all these deviations are caused by the long-time influence of Nyquist's way of thinking.

p.124

3. Modified Nyquist stability criterion

We would like to modify the original Nyquist criterion first: the fundamental change is that instead of using the return ratio, we use the return difference; instead of mapping the whole imaginary axis and semicircles, we map the positive frequency segments.

Chen and Tsang (1965, 1967) derived a frequency stability criterion for a single-input–output system based on the return difference. Here we rigorously generalize the criterion for both single-input–output and multi-input–output cases. The modified Nyquist criterion is mapping the straight line along the $j\omega$ axis for $\omega = 0^+$ to $\omega = \infty$ and then count the total phase change of the corresponding mapping as shown in Fig. 4. Because the original Nyquist criterion is known to map the whole imaginary axis and the right-half plane circle as shown in Fig. 3, the modified criterion is much simpler, even from the very beginning.



ORIGINAL PAGE IS
OF POOR QUALITY

Figure 3. Nyquist criterion uses the whole imaginary axis and right-half circle.

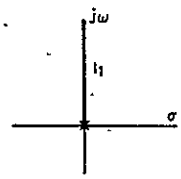


Figure 4. New criterion uses the positive imaginary axis only.

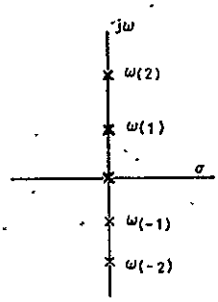


Figure 5. Definition of open segment.

p.125

Definition 1

The open segment on the $j\omega$ axis from $\omega=\omega_i$ to $\omega=\omega_{i+1}$ is the open interval $(j\omega_i, j\omega_{i+1})$ on the $j\omega$ axis (see Fig. 5). If there is a β_0-ple zero of $b(s)$ at the origin and β_i-ple zero at ω_i for $i=\pm 1, \pm 2, \ldots, \pm \rho$, then the mapping segments are $(0, j\omega_1), (j\omega_1, j\omega_2), (j\omega_2, j\omega_3) \ldots (j\omega_\rho, j\omega_\infty)$.

Definition 2

The phase angle change of the mapping of an open segment $l(i)=(j\omega_i, j\omega_{i+1})$ on the $j\omega$ axis by a certain complex function $\eta(s)$ is defined by

$$\Delta\theta_{l(i)}=\theta[j(\omega_{i+1}^-)]-\theta[j(\omega_i^+)] \tag{10}$$

where

$$\theta(j\omega_{i-1})=\lim_{\epsilon\to 0}\theta[j(\omega_{i+1}-\epsilon)],\ \theta(j\omega_i^+)=\lim_{\epsilon\to 0}\theta[j\omega_i+\epsilon)]$$

and $\theta(j\omega)$ is the phase angle of the complex function $\eta(s)$ for $s=j\omega$.

Definition 3

If the open segments are $(0, j\omega_1), (j\omega_1, j\omega_2) \ldots (j\omega_i, j\omega_{i+1}) \ldots (j\omega_\rho, j_\infty)$, the total phase change of the mapping of all these open segments on the positive $j\omega$ axis by a certain complex function of (s) is defined by

$$\Delta\theta=\sum_{i=1}^{\rho+1}\Delta\theta_{l(i)} \tag{11}$$

where $\Delta\theta_{l(i)}$ is the phase angle change due to the ith open segment mapping. Let the return difference of the system have the following properties :

[1] $p(j\omega)$ is finite and non-zero.

[2] There are β poles of $p(s)$ on the $j\omega$ axis and γ poles of $p(s)$ in the right-half plane, where β and γ are finite integers, and $\beta=\sum_{i=-p}^{p}\beta_i$.

[3] $p^*(s)=p(s^*)$ where * means complex conjugate.

For deriving the modified Nyquist criterion, let us start with the corresponding Nyquist contour of Fig. 5 as shown in Fig. 6.

From the principle of argument, it is known that the total phase change of $P(s)$ along the contour Γ is given by

$$\Delta\theta_\Gamma=-2\pi(M-\gamma) \tag{12}$$

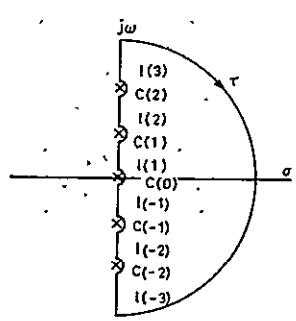


Figure 6. Nyquist's original contour.

p.126

where M is the number of closed-loop characteristic roots with positive real parts, γ is the number of open-loop characteristic roots with positive real parts. In detail, the total phase change of $P(s)$ along the contour Γ is evaluated by the summation of individual phase changes along $l_{(i)}$, $i = \pm 1, \pm 2, \ldots$, semicircles $c_{(i)}$, $i = 0, \pm 1, \pm 2, \ldots, \pm \rho$ and c_∞.

In other words,

$$\Delta\theta_\Gamma = \sum_{i=1}^{l+1} [\Delta\theta_{l(i)} + \Delta\theta_{l(-i)}] + \sum_{i=-\rho}^{\rho} \Delta\theta_{c(i)} + \Delta\theta_{c(\infty)} \tag{13}$$

since $p(j\omega)$ is non-zero finite, therefore $\theta_{c(\infty)} = 0$.

Assume that

$$P(j\omega) = U(\omega) + jV(\omega) \tag{14}$$

where $U(\omega)$ and $V(\omega)$ are real functions of ω.

ORIGINAL PAGE IS
OF POOR QUALITY

From property [3], we have

$$P^*(j\omega) = U(\omega) - jV(\omega) \tag{15}$$

$$= P(-j\omega)$$

$$= U(-\omega) + jV(-\omega) \tag{16}$$

The phase angle of $P(j\omega)$ is given by

$$\theta(j\omega) = \tan^{-1} \frac{V(\omega)}{U(\omega)} \tag{17}$$

and that of $P(-j\omega)$ is given by

$$\theta(j\omega) = \tan^{-1} \frac{V(-\omega)}{U(-\omega)} \tag{18}$$

However, from eqn. (15) and (16), we have

$$U(\omega) = U(-\omega) \tag{19}$$

and

$$V(\omega) = -V(-\omega) \tag{20}$$

Therefore,

$$\theta(j\omega) = -\theta(-j\omega) \tag{21}$$

Now we consider the phase angle change due to $l_{(i)}$:

$$\Delta\theta_{l(i)} = \theta(j\omega_{i-1}) - \theta(j\omega_i^+) \tag{22}$$

and that along $l_{(-i)}$:

$$\Delta\theta_{l(-i)} = \theta(-j\omega_i^+) - \theta(-j\omega_{i+1}^-) \tag{23}$$

Combining with (21), we can rewrite (23) as follows:

$$\Delta\theta_{l(-i)} = \theta(j\omega_{i+1}^-) - \theta(j\omega_i^+) \tag{24}$$

then we have

$$\Delta\theta_{l(i)} = \Delta\theta_{l(-i)} \tag{25}$$

Substituting eqn. (25) and the relation $\Delta\theta_{c(\infty)} = 0$ into eqn. (13) yields

$$\Delta\theta_\Gamma = 2\sum_{i=1}^{\rho+1} \Delta\theta_{l(i)} + \sum_{i=p}^{\rho} \Delta\theta_{c(i)} \tag{26}$$

Next, we consider the phase angle change due to semicircles $c_{(i)}$. Assume that there is β_i poles of $p(s)$ at $\omega = \omega_i$. Then

$$\Delta\theta_{c(i)} = -\beta_i\pi \tag{27}$$

Then

$$\sum_{i=-\rho}^{\rho} \Delta\theta_{c(i)} = -\pi \sum_{i=-\rho}^{\rho} \beta_i \tag{28}$$

From property (2), eqn. (28) becomes

$$\sum_{i=-\rho}^{\rho} \Delta\theta_{c(i)} = -\beta\pi \tag{29}$$

Then combining eqns. (29), (12) and (26) we have

$$\sum_{i=1}^{\rho+1} \Delta\theta_{l(i)} = \frac{\pi}{2}[-2(M-\gamma)+\beta] \tag{30}$$

$$= \frac{\pi}{2}[\beta+2(\gamma-M)] \tag{31}$$

Now we obtain the modified Nyquist stability criterion.

Criterion

If the return difference of a general feedback system, single-input–output or multi-input–output has the properties [1] to [3], the system is stable if and only if the total phase angle change along open segments $(j0, j\omega_1)$, $(j\omega_1, j\omega_2) \ldots (j\omega_\rho, j\infty)$ is equal to $(\pi/2)(\beta+2\gamma)$ where β and γ are the number of open-loop poles on the imaginary axis and in the right-half plane respectively.

It is easy to prove the criterion by substituting $M=0$ into (31), because the system is stable if and only if the closed-loop characteristic polynomial has no roots in the right-half plane.

Therefore, we have

$$\Delta\theta = \frac{\pi}{2}(\beta+2\gamma) \tag{32}$$

4. General case

Let us then consider the general case, which means that

$$P(j\infty)=0, \quad P(-j\infty)=0 \quad \text{or} \quad P(j\infty)=\infty \quad \text{and} \quad P(-j\infty)=\infty$$

In this case $\theta_{c(\infty)} \neq 0$. Assume that

$$\lim_{s\to\infty} P(s) = \lim_{s\to\infty} as^{-k}$$

where a is non-zero and k is real. This means that the phase angle of $P(j\omega)$ as $\omega\to\infty$ is equal to $[-k\pi/2]$. Then the previous case in which $p(j\infty)$ and $P(-j\infty)$ is non-zero finite becomes a special case: $k=0$.

For this general case the phase angle change along the semicircle of infinite radius $c(\infty)$ is

$$\begin{aligned}\Delta\theta_{c(\infty)} &= \theta(-j\infty)-\theta(j\infty)\\ &= k\frac{\pi}{2}-\left(-k\frac{\pi}{2}\right)\\ &= k\pi\end{aligned} \tag{34}$$

p. 128

Equation (13) should be rewritten for this general case as

$$\Delta\theta_\Gamma = 2\sum_{i=1}^{\rho+1} \Delta\theta_{l(i)} - \beta\pi + k\pi \tag{35}$$

Therefore, from eqns. (12) and (35), we have

$$\sum_{i=1}^{\rho+1} \Delta\theta_{l(i)} = \frac{\pi}{2}[\beta - k + 2(\gamma - M)] \tag{36}$$

Thus we state the modified Nyquist criterion for the general case as follows.

If the return difference $p(s)$ of a single-input–output or multi-input–output system has the following properties :

ORIGINAL PAGE IS OF POOR QUALITY

[1] the numbers of poles of $P(s)$ on the $j\omega$ axis and in the right-half plane are β and γ respectively ;
[2] $P^*(s) = P(s^*)$;
[3] $P(j\omega) \neq 0$ for finite ω.

The system is stable if and only if the total phase angle change along the segments $(0, j\omega_1)$, $(j\omega_1, j\omega_2)$ ... $(j\omega_\rho, j\infty)$ is equal to $(\pi/2)(\beta - k + 2\gamma)$ where

$$k = \frac{-2}{\pi}[\theta(j\infty)] \tag{37}$$

It is easy to prove this criteria : simply substituting $M = 0$ into (36) yields

$$\Delta\theta = \sum_{i=1}^{\rho+1} \Delta\theta_{l(i)} = \frac{\pi}{2}(\beta - k + 2\gamma) \tag{38}$$

The most important feature of this criterion is that we never bother with the number of roots of the numerator of the open-loop transfer function while the original Nyquist criterion has to examine it all the time. This feature makes the stability determination of a class of infinite dimensional multi-input–output feedback systems possible.

5. Application to lumped-parameter feedback systems

5.1. *Single-input–output system with non-minimum open-loop transfer function*

Consider a feedback system with the following open-loop transfer function :

$$g(s)h(s) = \frac{-(s-1)(s+4)}{(s+2)^2} \tag{39}$$

which is a non-minimum phase. Investigate the stability of the system.

If we use the original Nyquist criterion to study this problem, we have to start with the return ratio (39) and count the right-half plane roots of the numerator.

If the new stability criterion is used, we start with the return difference function :

$$P(s) = 1 + g(s)h(s) = \frac{s+8}{(s+2)^2} \tag{40}$$

p.129

Then plot the phase angle of the return difference

$$\angle P(j\omega) = \tan^{-1}\frac{j\omega}{8} - 2\tan^{-1}\frac{j\omega}{2} \tag{41}$$

as shown in Fig. 7.

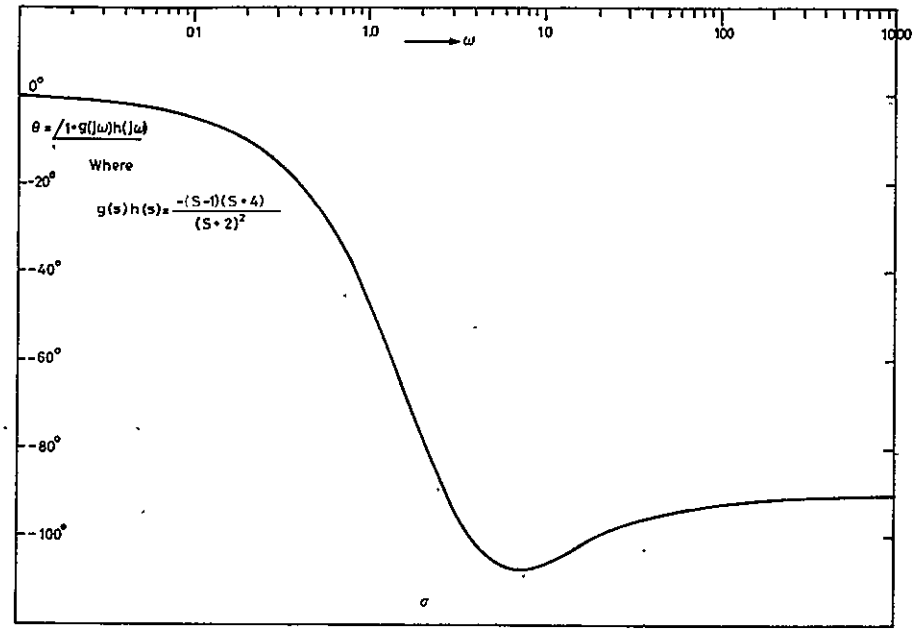


Figure 7. Phase change of the return difference function of a single-input–output system with non-minimum open-loop transfer function.

Examining the open-loop polynomial or the denominator of (39), we find that there are no roots on the imaginary axis or in the right-half plane. Then we have

$$\beta = 0 \quad \text{and} \quad \gamma = 0$$

Also,

$$\lim_{s\to\infty} P(s) = \lim_{s\to\infty} \frac{1}{s}$$

Therefore $k = 1$.

The criterion for stability is

$$\Delta\theta = \frac{\pi}{2}(\beta + 2\gamma - k) = -\frac{\pi}{2}$$

From Fig. 7 we see that at $\omega = 0^+$ the phase angle is zero degree while at $\omega = \infty$ the phase angle is negative 90°. The phase angle change is therefore

$$\Delta\theta = -\frac{\pi}{2}$$

we then conclude that the system is stable.

p.130

It is interesting to note that we never count the number of right-half plane roots of the numerator of the return ratio function. This is a great advantage of the new criterion. This problem cannot be solved by DeSoer's criteria since $\inf|P(s)|=0$ for $\operatorname{Re} s \geqslant 0$.

5.2. *Single-input–output systems with oscillating open-loop transfer function*

If the open-loop transfer function of a feedback system is as follows:

$$g(s)h(s)=\frac{2(s+1)^2(s+2)}{s^2(s^2+1)} \tag{42}$$

we are then interested in the stability of the system. The return ratio function has two roots at the origin and two roots on the imaginary axis respectively.

For applying the new criterion, we find the return difference function first:

$$P(s)=1+g(s)h(s)=\frac{s^4+2s^3+9s^2+10s+4}{s^2(s^2+1)} \tag{43}$$

Inspecting (43), we have $\beta=4$, $\gamma=0$, and $k=0$. Therefore the stability criterion is $\Delta\theta=2\pi$.

We then plot the phase curve of the return difference function as shown in Fig. 8. We have the following observation:

ω	θ
0^+	$-180°$
1^-	$-63{\cdot}4°$
1^+	$-243{\cdot}4°$
∞	$0°$

The phase change is found as follows:

$$\Delta\theta=\sum_{i=1}^{2}\theta_{l(i)}=[-63{\cdot}4°-(-180°)]+[0°+243{\cdot}4°]$$
$$=360°$$

Based on our new criterion, we conclude that the system is stable.

The interesting part of this example is the discontinuity of the phase curve at $\omega=1$. Even with this discontinuity, the total phase change still meets the stability criterion.

This case has been excluded by Chen and Tsang when they established their criterion. In this sense, the new criterion has a larger scope of application, although the criterion for the special case is the same criterion derived by Chen and Tsang in 1965.

p.131

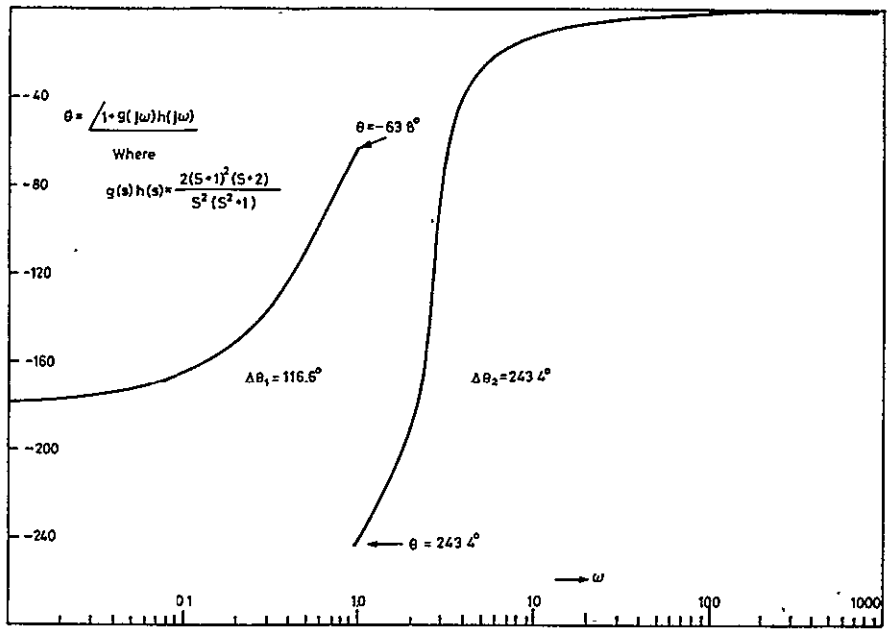


Figure 8. Phase change of the return difference function of a single-input–output system with oscillating open-loop transfer function.

5.3. *Multi-input–output feedback system with peculiar return difference function*

Consider the following multi-input–output system with

$$G(s)=\begin{bmatrix} \dfrac{-s}{s-1} & \dfrac{s}{s+1} \\ 1 & \dfrac{-2}{s+1} \end{bmatrix} \tag{44}$$

and

$$H(s)=1$$

Let us examine the return difference first:

$$\begin{aligned} P(s) &= \det\{1+G(s)H(s)\} \\ &= -1 \end{aligned} \tag{45}$$

There is no phase angle change at all or the phase change is equal to zero.

According to the stability criterion, however, the system is stable if and only if $\Delta\theta=(\pi/2)(\beta+2\gamma)=\pi$, because $\beta=0$, $\gamma=1$. We therefore conclude that the system is unstable.

The problem is taken from C. T. Chen's book (1968) where he warned that the return difference function sometimes does not offer stability information and his remedy was to pose more restrictions or conditions on this peculiar situation.

Our new criterion, on the other hand, works fine and need not pose any new restrictions.

p.132

5.4. *Multi-input–output feedback system with unstable open-loop transfer matrix*

Consider the following multi-input–output feedback system :

$$G(s)=\begin{bmatrix} \dfrac{-s^2+s+1}{(s+1)(s-1)} & \dfrac{1}{(s-1)} \\ \dfrac{1}{(s+1)(s-1)} & \dfrac{1}{(s-1)} \end{bmatrix}. \tag{46}$$

and

$$H(s)=1. \tag{47}$$

First, we find the return difference function :

$$\begin{aligned} P(s) &= \det\,[1+G(s)] \\ &= \frac{(s+1)}{(s-1)(s+1)} = \frac{1}{(s-1)} \end{aligned} \tag{48}$$

ORIGINAL PAGE IS
OF POOR QUALITY

from which we have $\beta=0$, $\gamma=1$ and

$$\theta(j\infty) = -\lim_{\omega\to\infty} \tan^{-1}\frac{\omega}{-1} = -\frac{\pi}{2} \tag{49}$$

Therefore $k=1$.

Substituting these data into (38) we obtain

$$\begin{aligned} \Delta\theta &= \frac{\pi}{2}(2\gamma-k+\beta) \\ &= \frac{\pi}{2}(2-1+0) = \frac{\pi}{2} \end{aligned} \tag{50}$$

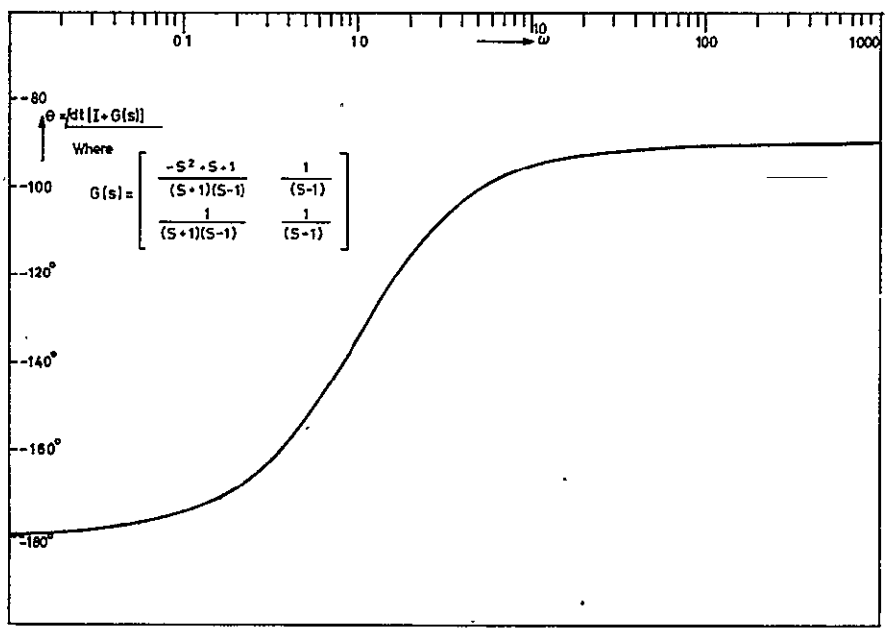


Figure 9. Phase change of the return difference function of a multi-input–output system with unstable open-loop transfer matrix.

p.133

Then we draw the phase plot of the return difference function (48). The phase angle change is $\pi/2$. Therefore we claim that the system is stable. This example is also posed by Chen in his text; he warned that (48) should be carefully examined, otherwise we could easily arrive at a wrong conclusion. To use our new criterion, however, this example is simply a routine exercise—there are no particular difficulties.

It is known that DeSoer's recent work is also based on the return difference. But for this simple example, his criterion is not applicable, because of the fact: $\inf_{\mathrm{Re}\, s \geq 0} |P(s)| = 0$.

6. Application to distributed-parameter systems

The stability criterion given in § 3 and § 4 can be extended to the study of the stability of feedback systems with distributed parameters as far as β and γ of the system is finite; there is no restriction on the return difference being a rational function of s in the proof of the criterion. Of course, the mapping of the return difference along the $j\omega$ axis should be taken in the principle sheet of Riemann surfaces.

6.1. *Multi-input–output feedback systems involving* $\exp(-\tau s)$

Consider the multi-input–output system with

$$G(s) = \begin{bmatrix} \dfrac{\exp(-s)}{s-1} & \frac{1}{2} \\ 1 & \dfrac{2\exp(-2s)}{s+2} \end{bmatrix} \tag{51}$$

and

$$H(s) = 1 \tag{52}$$

First of all, we find the return difference function:

$$P(s) = \det[1 + G(s)]$$

$$= \frac{\exp(-s)(s+2) + 2(s-1)\exp(-2s) + 2\exp(-3s) + 0{\cdot}5(s-1)(s+2)}{(s-1)(s+2)} \tag{53}$$

The open-loop characteristic function is the denominator of (53), therefore, we have $\beta = 0$, $\gamma = 1$.

Then we find

$$\lim_{s \to \infty} P(s) = 0{\cdot}5 \neq 0$$

Therefore $k = 0$. The total phase change should be

$$\Delta\theta = \frac{\pi}{2}(\beta - k + 2\gamma)$$

$$= \frac{\pi}{2}(0 - 0 + 2) = \pi \tag{54 a}$$

if we expect its stability.

p. 134

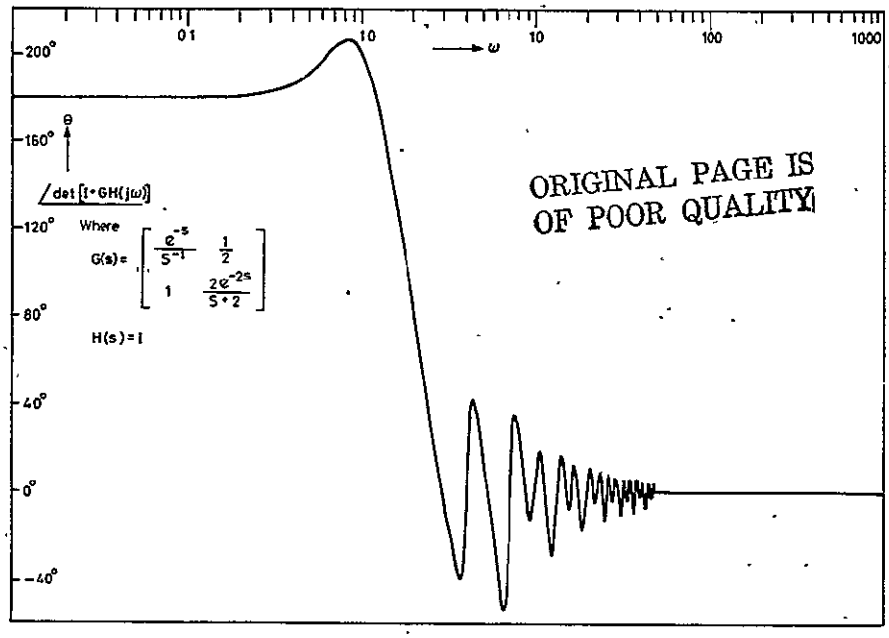


Figure 10. Phase change of the return difference function of a multi-input–output feedback system with exp $(-\tau s)$.

When we plot the phase angle of the return difference function, we have the graph shown in Fig. 10. It is very interesting to note that when $\omega = 0^+$, the phase is 180°; when $\omega = \infty$, the phase is 0°. There are many oscillatory types of various starts with $\omega = 2$ and beyond. Based on our new criterion, we are only interested in the phase angle change between the two ends. That is

$$\Delta\theta = 0° - 180° = -180° \tag{54 b}$$

From a comparison of (54 *a*) and (54 *b*), we conclude that the system is unstable.

While Pontryagin's technique for a time-delay system has not been extended to multi-input–output systems and Desoer's criterion for this type of time-delay system does not work (because there is an unstable pole in the open-loop characteristic function), by using our new criterion, the problem is easily solved.

6.2. *An equivalent system*

There is a very important theorem on block diagram equivalence for negative feedback systems. The theorem states (Chen 1974): as far as the input–output properties are concerned, the block diagrams shown in Fig. 11 are equivalent.

It is easy to prove the equivalency as follows.

The closed-loop transfer matrices of the left system is

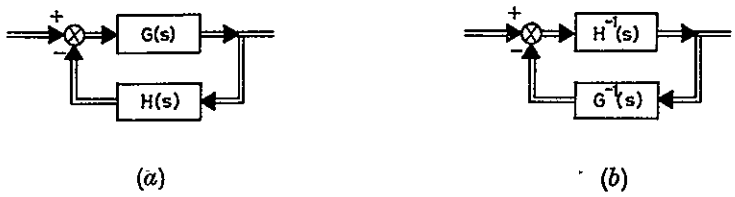
$$F(s) = [1 + G(s)H(s)]^{-1}G(s) \tag{55}$$

If the inverses of $G(s)$ and $H(s)$ exist, we have

$$\begin{aligned} F(s) &= [G(s)H(s)(H^{-1}(s)G^{-1}(s)+1)]^{-1}G(s) \\ &= [1+H^{-1}(s)G^{-1}(s)]^{-1}[G(s)H(s)]^{-1}G(s) \\ &= [1+H^{-1}(s)G^{-1}(s)]^{-1}H^{-1}(s) \end{aligned} \tag{56}$$

Comparing (55) with (56), we have the proof of the equivalence.

As far as the stability of Fig. 11 (*a*) is concerned, we can test the stability of (*b*).



(*a*) (*b*)

Figure 11. Equivalent systems.

This equivalent block diagram makes the following fact obvious. The inverse Nyquist plot is an unnecessary complication and so is Rosenbrock's inverse Nyquist arrays. We use the equivalent block diagrams of Fig. 11 to deal with a class of problems in the next section.

6.3. *Feedback systems involving* $\sqrt{s}$

Consider the following system :

$$g(s) = \frac{100}{(0{\cdot}63\sqrt{s}+1)(s+1)} \tag{57}$$

$$h(s) = 1 \tag{58}$$

This problem has been studied by Chen and Chiu (1973) by the fast Fourier transfer form, and we know that it is stable. We would like to use the new criterion to reach that answer.

Since there is $\sqrt{s}$ in the denominator of the open-loop transfer function, it is not easy to count the number of open-loop poles.

Instead of dealing with this problem directly, we can determine the stability of an equivalent system with the following :

$$g_1(s) = 1 \tag{59}$$

$$h_1(s) = \frac{(0{\cdot}63\sqrt{s}+1)(s+1)}{100} \tag{60}$$

Then we examine the return difference of the equivalent system :

$$P(s) = 1 + \frac{1}{100}(0{\cdot}63\sqrt{s}+1)(s+1) \tag{61}$$

We find that

$$\beta=0 \quad \text{and} \quad \gamma=0$$

because there is only a constant in the denominator of (60) and

$$\lim_{s\to\infty} P(s)=\lim_{s\to\infty} 0{\cdot}0063s^{3/2}$$

from which we have

$$k=-\tfrac{3}{2} \tag{62}$$

Therefore, the stability criterion is given by

$$\Delta\theta=\frac{\pi}{2}(\beta-k+2\gamma)$$

$$=\frac{\pi}{2}(+\tfrac{3}{2})=135^\circ \tag{63}$$

When we plot the phase angle function of (61), we see that the phase change is 135° as shown in Fig. 12. Therefore the equivalent feedback system is stable; so is the original system.

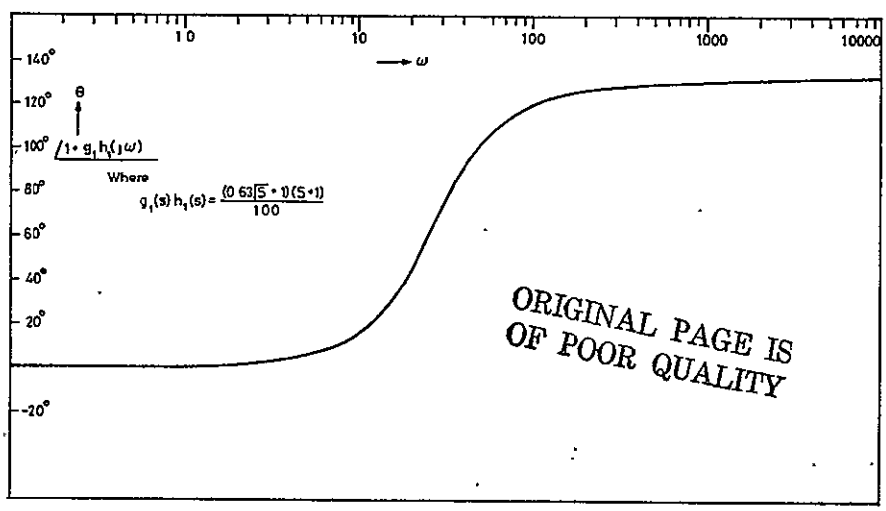


Figure 12. Phase change of the return difference function of a single-input–output feedback system with $\sqrt{s}$.

6.4. *Feedback systems involving* exp $(-\sqrt{s})$

Consider the feedback system with the following element:

$$g(s)=\frac{5(s+0{\cdot}3)\exp(-\sqrt{s})}{s^2+40(s+0{\cdot}3)\exp(-\sqrt{s})} \tag{64}$$

and

$$h(s)=\frac{100(s+0{\cdot}1)(s+0{\cdot}2)}{(s+30)(s+20)} \tag{65}$$

p.137

When we use our new criterion, we need to count the number of poles of the open-loop characteristic function in the right-half plane and on the imaginary axis. In other words, we have first to know the root numbers of the denominator of (64) in the right-half plane. There is no direct way to know the root situation of the equation :

$$s^2+40(s+0{\cdot}3)\exp(-\sqrt{s})=0 \tag{66}$$

However, we can set up an auxilliary system with a return difference function as follows :

$$1+\frac{40(s+0{\cdot}3)\exp(-\sqrt{s})}{s^2} \tag{67}$$

That means the open-loop transfer function is

$$\frac{40(s+0{\cdot}3)\exp(-\sqrt{s})}{s^2}$$

Examining the auxilliary system we find

$$\beta=2, \quad \gamma=0 \quad \text{and} \quad k=0$$

Substituting these values into (36), we have

$$\Delta\theta=\frac{\pi}{2}(2-0+2(0-M))$$

$$M=-\frac{\Delta\theta}{\pi}+1 \tag{68}$$

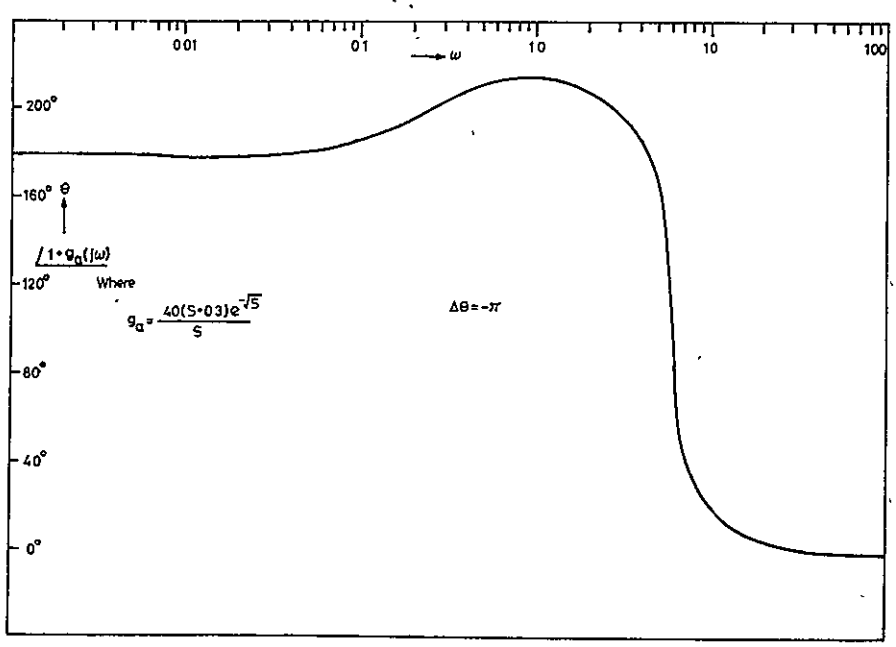


Figure 13. Phase change of the return difference function of an auxiliary feedback system for problem shown in § 6.4.

p.138

The phase curve of (67) is then plotted as shown in Fig. 13. By inspection, the phase change is $\Delta\theta = -180°$; substituting this $\Delta\theta$ into (68), we have $M = 2$. This means there are two roots of the original $g(s)$ in the right-half plane.

Now we deal with the original system shown in (64) and (65). The return difference function is

$$1 + g(s)h(s) = 1 + \frac{500(s+0{\cdot}1)(s+0{\cdot}2)(s+0{\cdot}3)\exp(-\sqrt{s})}{(s+20)(s+30)[s^2+40(s+0{\cdot}3)\exp(-\sqrt{s})]} \tag{69}$$

From (68) we know there are two roots of the open-loop characteristic function, or the denominator of (49), in the right-half plane. The phase curve of (69) is then plotted as shown in Fig. 14. The phase change is 2π. The stability criterion is obtained by using $\beta = 0$, $k = 0$, and $\gamma = 2$ and we get

$$\Delta\theta = \frac{\pi}{2}(0 + 0 + 2 \times 2) = 2\pi \tag{70}$$

Therefore, the system is stable.

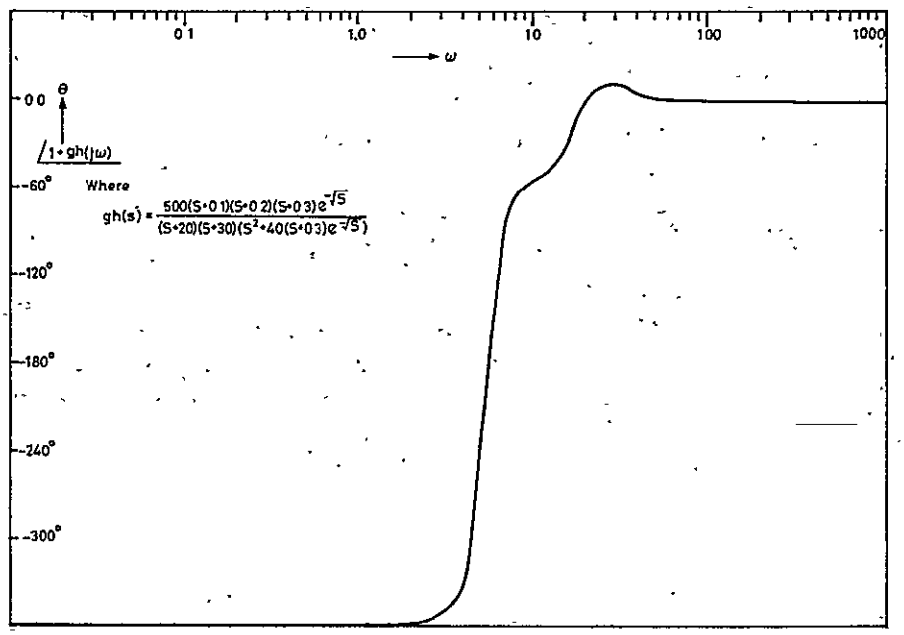


Figure 14. Phase change of problem § 6.4.

7. Conclusions

A new frequency stability criterion for feedback systems, single-input–output or multi-input–output, is established. The criterion is applicable to lumped-parameter systems as well as to distributed-parameter systems. The procedure is to compare the number of non-left-half plane open-loop poles with the phase change of the return difference phase plot.

Compared with the original Nyquist criterion, the new criterion uses the return difference instead of using the return ratio of Nyquist. This allows an extension to multi-input–output system as a routine exercise.

Compared with the Pontryagin stability criterion, the new criterion is not only applicable to systems with time delays but also to systems with irrational functions.

Compared with the Brin stability criterion, the new criterion is for a general system while Brin's criterion is for a very special narrow class. Because Brin's criterion is an extension of Mikhailov's criterion, therefore the new criterion is naturally better than Mikhailov's.

Compared with the Desoer stability criterion, the scope of application is much larger. Examples in §§ (5.1), (5.4), (6.1) and (6.4) cannot be solved by Desoer's criterion.

The only restriction to our criterion is that the unstable open-loop poles must be finite, or reducible to finite.

References

Bode, H. W., 1945, *Network Analysis and Feedback Amplifier Design* (New York: Van Nostrand).

Brin, I. A., 1962, *Autom. Remote Control*, **23**, 798.

Chen, C. F., 1974, Comments on DeSoer's Presentation at the Eighth Asilomar Conference.

Chen, C. T., 1968, *Proc. I.E.E.E.*, **56**, 821.

Chen, C. F., and Chiu, R. F., 1973, *Int. J. Control*, **15**, 267.

Chen, C. F., and Haas, I. J., 1968, *Elements of Control Systems Analysis* (Prentice-Hall).

Chen, C. F., and Tsang, N. F., 1965, *Proceedings of the Pan American Congress of Mechanical and Electrical Engineering and Related Branches*, Mexico, October; 1967, *I.E.E.E. Trans. Educ.*, **10**, 180.

DeSoer, C. A., and Vodyasagar, M., 1975, *Feedback Systems: Input–Output Properties* (New York: Academic Press), p. 92.

Hsu, C. H., and Chen, C. T., 1968, *Proc. I.E.E.E.*, **56**, 2061.

Kalman, R. E., 1964, *Bas. Engng*, D, **86**, 51.

MacFarlane, A. G. J., 1970, *Proc. I.E.E.E.*, **117**, 2037.

Pontryagin, L. S., 1942, *Izv. Akad. Nauk. SSSR*, **6**, 115.

Porter, Brian, 1942 *Stability Criteria for Linear Dynamical Systems* (London: Oliver & Boyd), p. 146.

Rosenbrock, H. H. 1969, *Proc. I.E.E.E.*, **116**, 1929.

Truxal, J. G., 1955, *Automatic Feedback Control System Synthesis* (McGraw-Hill).

p. 140

Project

(A) Project Title: Walsh Series Analysis in Optimal Control

(B) Project Abstract:

This paper is concerned with the determination of sub-optimal feedback laws for the linear systems with quadratic performance criteria. The time-varying gains are approximated by the piecewise constant gains which are naturally determined by using Walsh functions. An increase of the number of intervals of Walsh functions enables us to approximate the true optimal control more closely; and a decrease of the number of intervals makes the implementation easier. Therefore the proposed method is simple in theory and flexible in practice. The beginning part of the paper, being tutorial in nature, is on Walsh functions, the middle part develops an operational method for solving state equations and the final part is concentrated on the Walsh functions approach to the solution of piecewise constant gains of optimal control.

(C) Publication: International Journal of Control, 21, 881 (1975)

(D) Year: Completed in 1974

(E) Department: Electrical Engineering

(F) Student Name: C. H. Hsiao

(G) Faculty Advisor: Professor C. F. Chen

INT. J. CONTROL, 1975, VOL. 21, NO. 6, 881–897

Walsh series analysis in optimal control

C. F. CHEN† and C. H. HSIAO†

This paper is concerned with the determination of suboptimal feedback laws for the linear systems with quadratic performance criteria. The time-varying gains are approximated by the piecewise constant gains which are naturally determined by using Walsh functions. An increase of the number of intervals of Walsh functions enables us to approximate the true optimal control more closely; and a decrease of the number of intervals makes the implementation easier. Therefore the proposed method is simple in theory and flexible in practice. The beginning part of the paper, being tutorial in nature, is on Walsh functions, the middle part developes an operational method for solving state equations and the final part is concentrated on the Walsh functions approach to the solution of piecewise constant gains of optimal control.

1. Introduction

The control of a linear system with respect to quadratic performance criteria often involves time-varying gain design. In theory, the methods are well established and documented. In implementation, however, it is still a difficult task. Kleinman *et al.* (1968), Kleinman and Athans (1968) and Fortmann (1967) proposed a very elegant approach to the problem by taking practical engineering constraints into consideration: It was done by prespecifying the structural form of time-varying feedback gains, while leaving various free parameters to be chosen in an optimal fashion. Regarding the theoretical investigation, their thinking showed certain similarities with the direct method of Ritz in the calculus of variation.

This paper presents a new approach to the optimal problem by using the Walsh functions (Proceedings 1970–1972, Harmuth 1969, Corrington 1973). The approach should be classified as a direct method but is more powerful than the direct methods of Ritz and Euler (Elsgolc 1961) or Gelerkin (Schechter 1967) on the one hand and is much simpler than the procedure proposed by Kleinman and Athans (1968) on the other. The piecewise constant gains so obtained by the new approach are naturally formed, equally distributed and therefore can be easily implemented.

We shall start with the introduction of Walsh functions.

2. Fourier series and Walsh series

In the direct methods of the calculus of variation, we often use Fourier series, power series, etc. with coefficients to be determined as the initial step. Our new approach starts with Walsh series.

It is well known that a function which is periodic may be expanded into Fourier series. Analogously speaking, a function, $f(t)$, which is absolutely integrable in [0, 1) may be expanded into Walsh series

$$f(t) = c_0\phi_0(t) + c_1\phi_1(t) + c_2\phi_2(t) + \ldots + c_n\phi_n(t) + \ldots \qquad (1)$$

Received 18 December 1974.

† Electrical Engineering Department, University of Houston, Houston, Texas 77004.

p.142

ORIGINAL PAGE IS
OF POOR QUALITY

where

$$c_n = \int_0^1 \phi_n(t) f(t)\, dt \tag{2}$$

are determined in the following sense

$$\lim_{n\to\infty} \int_0^1 \left[f(t) - \sum_{n=0}^{\infty} c_n \phi_n(t) \right]^2 dt = 0 \tag{3}$$

The functions $\phi_0(t), \phi_1(t), \ldots, \phi_n(t)$ are a set of square waves which are orthonormal. ϕ_0 to ϕ_{15} are shown in Fig. 1. We call $\phi_i(t)$, $i = 0, 1, \ldots, n$, the set of Walsh functions in the dyadic order.

However, the regularity of the set of Walsh functions in the dyadic order is not very easily seen. It is usual to decompose each Walsh function into more elementary square waves, or Rademacher (1922) functions.

Rademacher function, $r_k(t)$, are a set of square waves of unit height with periods equal to $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, 2^{(1-k)}$. Alternatively, we can state that the number of cycles of the square waves of $r_k(t)$ is 2^{k-1}. The first five waves of Rademacher are shown in Fig. 2. It is seen that

$$\left.\begin{aligned}
\phi_0(t) &= r_0(t) \\
\phi_1(t) &= r_1(t) \\
\phi_2(t) &= (r_2(t))^1 (r_1(t))^0 \\
\phi_3(t) &= (r_2(t))^1 (r_1(t))^1 \\
\phi_4(t) &= (r_3(t))^1 (r_2(t))^0 (r_1(t))^0 \\
\phi_5(t) &= (r_3(t))^1 (r_2(t))^0 (r_1(t))^1 \\
\phi_6(t) &= (r_3(t))^1 (r_2(t))^1 (r_1(t))^0 \\
\phi_7(t) &= (r_3(t))^1 (r_2(t))^1 (r_1(t))^1 \\
&\vdots \\
\phi_i(t) &= (r_k(t))^{b_k} (r_{k-1}(t))^{b_{k-1}} (r_{k-2}(t))^{b_{k-2}} \ldots
\end{aligned}\right\} \tag{4}$$

where

$$k = [\log_2 i] + 1 \tag{5}$$

where $[\cdot]$ means taking the greatest integer of '$\cdot$', and $b_k, b_{k-1}, \ldots, b_1$ is the binary number expression of i.

To draw the wave form of any Walsh function is a trivial matter by using the above-mentioned decomposition techniques.

Let us then return to the Walsh coefficient evaluation of a Walsh series for a function.

Consider a given function $f(t) = t$. It is desired to expand it into Walsh series.

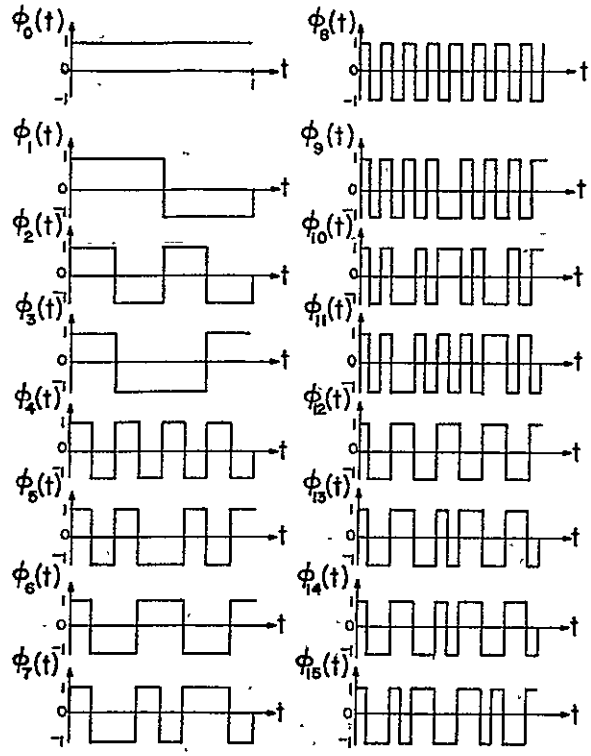


Figure 1.

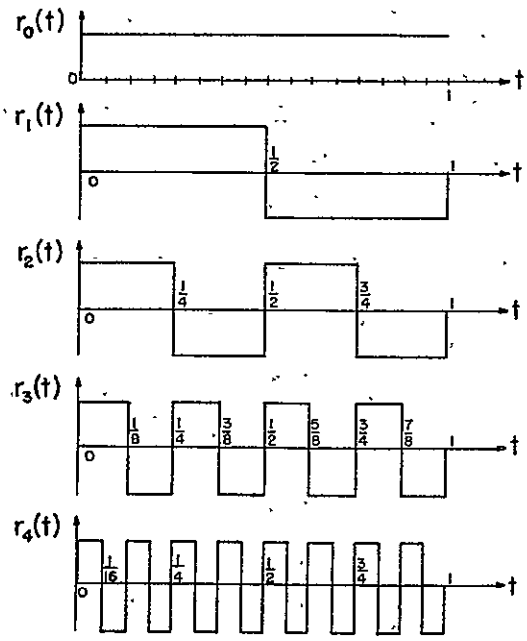


Figure 2.

4.144

Substituting $f(t)=t$ into (2), we have

ORIGINAL PAGE IS
OF POOR QUALITY

$$c_n = \int_0^1 t\phi_n(t)\,dt = \begin{cases} \frac{1}{2}, & n=0 \\ -\frac{1}{4}, & n=1 \\ -\frac{1}{8}, & n=2 \\ 0, & n=3 \\ -\frac{1}{16}, & n=4 \\ \cdots\cdots\cdots\cdots\cdots \end{cases} \tag{6 a}$$

or

$$f(t)=t \cong \tfrac{1}{2}\phi_0(t) - \tfrac{1}{4}\phi_1(t) - \tfrac{1}{8}\phi_2(t) - \tfrac{1}{16}\phi_4(t) - \tfrac{1}{32}\phi_8(t) - \tfrac{1}{64}\phi_{16}(t) \ldots \tag{6 b}$$

The original curve $f(t)=t$ and its Walsh series approximations are shown in Fig. 3. They are stairwise waves. The most crude representation is obtained by taking one term of the Walsh series, or $\frac{1}{2}\phi_0$; the second one consists of two terms $\frac{1}{2}\phi_0 - \frac{1}{4}\phi_1$. The figure shows up to the four-term approximation which is

$$f(t)=t \cong \tfrac{1}{2}\phi_0 - \tfrac{1}{4}\phi_1 - \tfrac{1}{8}\phi_2 - \tfrac{1}{16}\phi_4 \tag{6 c}$$

From the coefficient determination process, we see that the similarities between Fourier series and Walsh series are obvious.

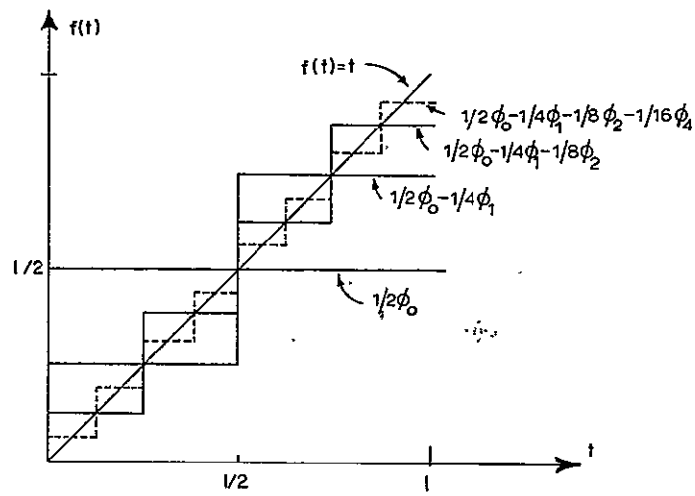


Figure 3.

3. Discrete formula

When we deal with the Fourier theory, if the given function is not in its analytic form but in tabulated data or in its graphical form and its Fourier series is desired, we would use a set of discrete formula. Similarly, we can

p.145

derive a corresponding set of discrete formula for Walsh series. They are

$$f_k = \sum_{n=0}^{m-1} \psi_{kn} c_n, \qquad k = 0, 1, 2, \ldots, (m-1) \tag{7 a}$$

$$c_n = \sum_{k=0}^{m-1} \phi_{nk} f_k \cdot \frac{1}{m}, \quad n = 0, 1, 2, \ldots, (m-1) \tag{7 b}$$

where f_k is the average value of the function in question in the kth sub-interval and ϕ_{nk} the value of the nth Walsh function in the kth sub-interval and m is the total number of sub-intervals from 0 to 1. It can be shown that $\psi_{nk} = \phi_{kn}$.

For illustration, consider the given data shown in Fig. 4 or

k	0	1	2	3
f_k	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{5}{8}$	$\frac{7}{8}$

Its Walsh series coefficients are required.

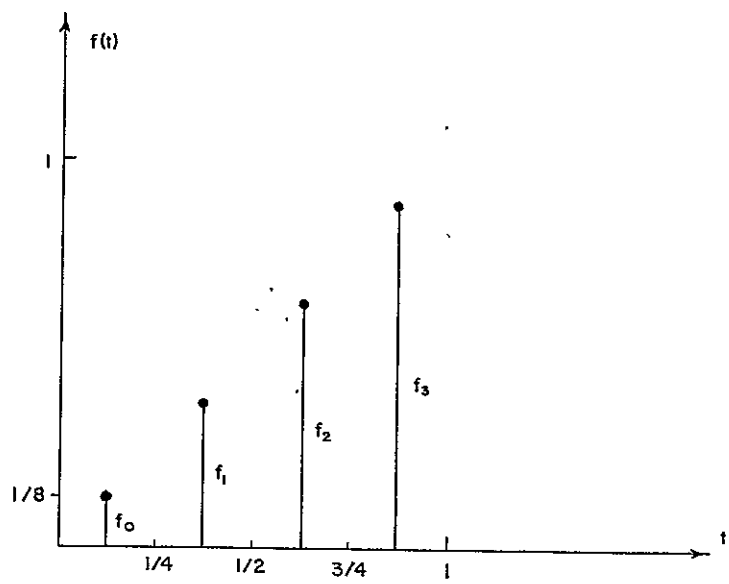


Figure 4.

Equation (7 *b*) can be written in a matrix form, once m is assigned. For example, for $m = 4$, (7 *b*) becomes

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} \cong \begin{bmatrix} \phi_{00} & \phi_{01} & \phi_{02} & \phi_{03} \\ \phi_{10} & \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{20} & \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{30} & \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{bmatrix} \cdot \frac{1}{4} \tag{8}$$

where f_k are the given data in Fig. 4; ϕ_{kn} can be defined clearly by using Fig. 2. For example, if we divide $\phi_0(t)$ from 0 to 1 into four sub-intervals, we can describe $\phi_0(t)$ by 1, 1, 1, 1. Similarly, $\phi_1(t)$ is described by 1, 1, −1, −1, etc.

The coefficients c_0, c_1, c_2 and c_3 can be evaluated by substitutions

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} \cong \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{8} \\ \frac{3}{8} \\ \frac{5}{8} \\ \frac{7}{8} \end{bmatrix} \cdot \frac{1}{4} = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{4} \\ -\frac{1}{8} \\ 0 \end{bmatrix} \tag{9}$$

This means

$$f(t) = t \cong \tfrac{1}{2}\phi_0 - \tfrac{1}{4}\phi_1 - \tfrac{1}{8}\phi_2$$

It is seen that the Walsh series coefficients evaluated from (2) analytically or from (9) numerically are the same.

Equation (7) can be written into matrix form

$$\mathbf{c} = \mathbf{\Phi f} \,.\, 1/m \tag{10}$$

where $\mathbf{\Phi}$ is called the Walsh matrix and can be easily derived from the Walsh wave configuration as shown in Fig. 5.

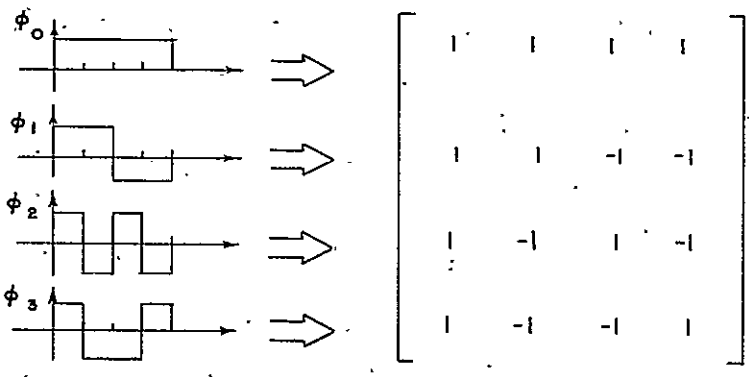


Figure 5.

Similarly we can write (7 *a*) into matrix form

$$\mathbf{f} = \mathbf{\Psi c} \tag{11}$$

Apparently

$$\mathbf{\Phi}^{-1} m = \mathbf{\Psi} \tag{12}$$

Once $\mathbf{\Phi}$ is defined from the basic definition of Walsh functions, $\mathbf{\Psi}$ follows. Equation (12) is one of the nice properties of Walsh functions. Formulae (10) and (11) are comparable to the discrete formula of Fourier series.

4. Integration and operational matrix

Through using (10), the evaluation Walsh coefficients can be obtained by multiplying a constant matrix by the discrete samples, and vice versa. In

p.147

this section we derive a method by which we can perform any integration by multiplying a constant matrix also.

Let us take $\phi_0, \phi_1, \ldots, \phi_7$ and integrate them, we then have various triangular waves as shown in Fig. 6. If we evaluate the Walsh coefficients for these triangular waves, we will easily arrive at the following :

$$\begin{bmatrix} \int \phi_0\,dt \\ \int \phi_1\,dt \\ \int \phi_2\,dt \\ \int \phi_3\,dt \\ \int \phi_4\,dt \\ \int \phi_5\,dt \\ \int \phi_6\,dt \\ \int \phi_7\,dt \end{bmatrix} = \left[\begin{array}{cc|cc|cccc} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{8} & 0 & -\frac{1}{16} & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & -\frac{1}{8} & 0 & -\frac{1}{16} & 0 & 0 \\ \hline \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{16} & 0 \\ 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{16} \\ \hline \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \\ \phi_7 \end{bmatrix} \tag{13 a}$$

or

$$\int \boldsymbol{\phi}_{(8)}\,dt = \mathbf{P}_{(8\times 8)}\boldsymbol{\phi}_{(8)} \tag{13 b}$$

The subscript means the dimension taken. It is preferable to take 2^{Ω}, where Ω is an integer, as a dimension number.

It is noted that

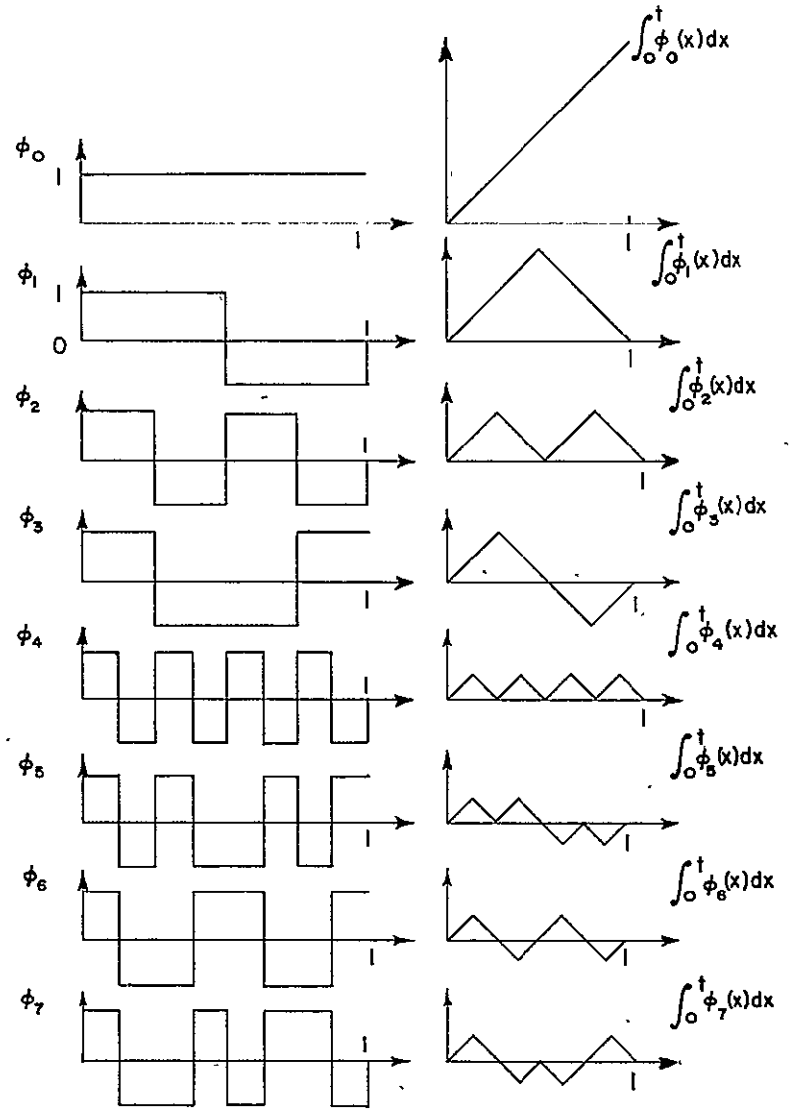
$$\int \phi_0\,dt = t$$

therefore the Walsh coefficients of $\int \phi_0\,dt$ are found in (6 *b*) which is simply the first row of (13 *a*) times the $\boldsymbol{\phi}(t)$ vector.

Equation (13) is for $m=8$; a general formula $\mathbf{P}_{(m\times m)}$ can be written as follows :

$$\mathbf{P}_{(m\times m)} = \left[\begin{array}{cc|c|c|c} \frac{1}{2} & & & & \\ & \diagdown & -\frac{1}{m/2}\mathbf{I}_{(m/8)} & & \\ \hline \multicolumn{2}{c|}{\frac{1}{m/2}\mathbf{I}_{(m/8)}} & \mathbf{0}_{(m/8)} & -\frac{1}{m}\mathbf{I}_{(m/4)} & \\ \hline \multicolumn{3}{c|}{\frac{1}{m}\mathbf{I}_{(m/4)}} & \mathbf{0}_{(m/4)} & -\frac{1}{2m}\mathbf{I}_{(m/2)} \\ \hline \multicolumn{4}{c|}{\frac{1}{2m}\mathbf{I}_{(m/2)}} & \mathbf{0}_{(m/2)} \end{array}\right] \tag{14}$$

It is interesting to note that if we partition (14) into four parts as shown, the left upper part of $\mathbf{P}_{(m\times m)}$ is identical to $\mathbf{P}_{(m/2\times m/2)}$, and the left upper corner of $\mathbf{P}_{(m/2\times m/2)}$ is $\mathbf{P}_{(m/4\times m/4)}$. Therefore, this regularity of the structure of the $\mathbf{P}$



ORIGINAL PAGE IS
OF POOR QUALITY

Figure 6.

matrix enables us to write the reduced Nth enlarged matrices to any dimension, as far as the dimension number is equal to 2^{Ω}, where Ω is an integer number.

5. State equation solution by the Kronecker product formula

Consider the following state equation :

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$
$$\mathbf{x}(0) = \mathbf{x}_0 \tag{15}$$

where $\mathbf{x}$ is a state vector of n components, $\mathbf{u}$ is an input vector of l components. $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ and $n \times l$ matrices respectively. We would like to establish a procedure to solve the state equations via the Walsh series.

First of all, we approximate the rate variable vector $\dot{\mathbf{x}}$ by a set of m-term Walsh series whose $n \times m$ coefficients are to be determined. Let

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \vdots \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} c_{10} & c_{11} & c_{12} & \cdots & c_{1(m-1)} \\ c_{20} & c_{21} & c_{22} & \cdots & \vdots \\ c_{30} & c_{31} & c_{32} & \cdots & \vdots \\ \vdots & & & & \vdots \\ c_{n0} & c_{n1} & c_{n2} & \cdots & c_{n(m-1)} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \vdots \\ \phi_n \end{bmatrix} \tag{16 a}$$

or simply written as

$$\dot{\mathbf{x}} = \begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \\ \vdots \\ \mathbf{c}_n' \end{bmatrix} \boldsymbol{\phi} \triangleq \mathbf{C}_{(n \times m)} \boldsymbol{\phi}_{(m)} \tag{16 b}$$

where ' ' means transposition.

It should be noted that this initial step is quite different from that of the regular procedure of a series solution of a differential equation. We assume the rate variable $\dot{\mathbf{x}}$ as an undetermined vector series, instead of assuming the state vector itself.

The state variable $\mathbf{x}$ may be obtained by integration,

$$\mathbf{x}(t) = \mathbf{C} \int_0^t \boldsymbol{\phi}(\lambda)\, d\lambda + \mathbf{x}_0 \tag{17}$$

The integration can be performed approximately by using the $\mathbf{P}$ matrix :

$$\int_0^t \boldsymbol{\phi}(\lambda)\, d\lambda \doteq \mathbf{P}\boldsymbol{\phi}(t) \tag{18}$$

The input vector can also be expressed by the Walsh series :

$$\mathbf{u} = \begin{bmatrix} h_{10} & h_{11} & h_{12} & \cdots & h_{1(m-1)} \\ h_{20} & h_{21} & h_{22} & \cdots & h_{2(m-1)} \\ \vdots & & & & \\ h_{l0} & h_{l1} & h_{l2} & \cdots & h_{l(m-1)} \end{bmatrix} \boldsymbol{\phi} \tag{19 a}$$

or

$$\mathbf{u} \triangleq \mathbf{H}\boldsymbol{\phi} \tag{19 b}$$

where $\mathbf{H}$ is a constant matrix, the determination of which can be achieved by the techniques illustrated in § 2.

Substituting (16 *b*), (17), (18) and (19 *b*) into (15), yields

$$\mathbf{C}\boldsymbol{\phi} = \mathbf{A}\{\mathbf{C}\mathbf{P}\boldsymbol{\phi} + \mathbf{x}_0\} + \mathbf{B}\mathbf{H}\boldsymbol{\phi} \tag{20}$$

$\mathbf{A}\mathbf{x}_0$ can be written as a form of vector, or

$$\mathbf{A}\mathbf{x}_0 = \mathbf{A}\mathbf{x}_0\phi_0 = [\mathbf{A}\mathbf{x}_0, \underbrace{0, 0, \ldots, 0}_{(m-1)\text{ column}}]\boldsymbol{\phi} = \mathbf{G}\boldsymbol{\phi} \tag{21}$$

ORIGINAL PAGE IS
OF POOR QUALITY

Finally we have

$$\mathbf{C}=\mathbf{ACP}+\mathbf{G}+\mathbf{BH}$$

The last two terms on the right-hand side are given and can be combined into one term, by letting

$$\mathbf{G}+\mathbf{BH}=\mathbf{K}$$

Then

$$\mathbf{C}=\mathbf{ACP}+\mathbf{K} \tag{22}$$

It can be shown that

$$\begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_n \end{bmatrix} = [\mathbf{P}'\otimes\mathbf{A}] \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_n \end{bmatrix} + \begin{bmatrix} \mathbf{k}_1 \\ \mathbf{k}_2 \\ \vdots \\ \mathbf{k}_n \end{bmatrix} \tag{23}$$

where $\mathbf{P}'\otimes\mathbf{A}$ is a Kronecker product defined as

$$\mathbf{P}'\otimes\mathbf{A} \triangleq \begin{bmatrix} a_{11}\mathbf{P}' & a_{12}\mathbf{P}' & \cdots & a_{1n}\mathbf{P}' \\ a_{21}\mathbf{P}' & a_{22}\mathbf{P}' & \cdots & a_{2n}\mathbf{P}' \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1}\mathbf{P}' & a_{n2}\mathbf{P}' & \cdots & a_{nm}\mathbf{P}' \end{bmatrix} \tag{24}$$

The solution of $\mathbf{C}$ comes from (23) directly,

$$\begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_n \end{bmatrix} = [\mathbf{I}-\mathbf{P}'\otimes\mathbf{A}] \begin{bmatrix} \mathbf{k}_1 \\ \mathbf{k}_2 \\ \vdots \\ \mathbf{k}_n \end{bmatrix} \tag{25}$$

Once $\mathbf{C}$ has been decided, the rate variable Walsh series representation is determined. The state variable vector is then found by substitution :

$$\mathbf{x}=\mathbf{CP}\boldsymbol{\phi}(t)+\mathbf{x}_0 \tag{26}$$

A high-order differential equation with constant coefficients can always be written as a state equation. Therefore eqns. (25) and (26) enable us to solve linear time-invariant systems elegantly and completely. Of course, the answer is in terms of Walsh functions.

6. State equation example

A gate function shown in Fig. 7 (*b*) is applied to a circuit shown in Fig. 7 (*a*). Find the response of the circuit.

The parameters of the circuit are

$$R=50\,\Omega, \quad C=5\,\mu\text{F}$$

and the initial condition is

$$v(0)=1 \text{ volt}$$

p.151

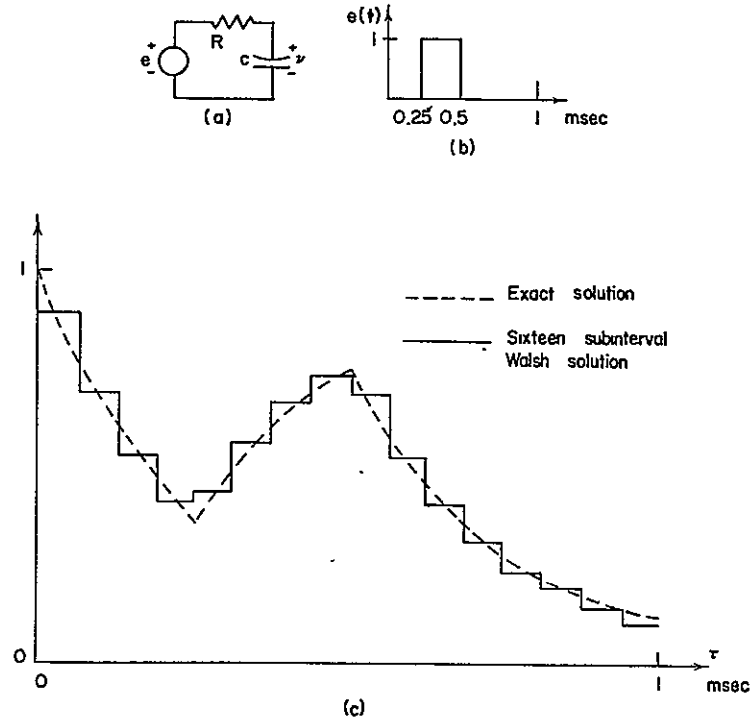


Figure 7.

The governing equation of the system is

$$dv/dt = -4\times 10^3 v + 4\times 10^3 e \tag{27 a}$$

First we would like to normalize the equation by using

$$\tau = 1000t$$

Then the governing equation is changed to

$$dv/d\tau = -4v + 4e \tag{27 b}$$

Let us assume that

$$\hat{v} = c_0\phi_0 + c_1\phi_1 + \ldots + c_{15}\phi_{15} = \mathbf{c}'\boldsymbol{\phi} \tag{28}$$

where both $\mathbf{c}$ and $\boldsymbol{\phi}$ are vectors with 16 components.

The input function $e(\tau)$ can be decomposed into a Walsh series

$$e(\tau) = [\tfrac{1}{4}, \tfrac{1}{4}, -\tfrac{1}{4}, -\tfrac{1}{4}, \underbrace{0, \ldots, 0}_{12 \text{ zeros}}]\boldsymbol{\phi} \tag{29}$$

$$= \mathbf{H}\boldsymbol{\phi}$$

The $\mathbf{G}$ matrix is written as

$$\mathbf{G} = [\mathbf{A}v(0), \underbrace{0, \ldots, 0}_{15 \text{ zeros}}]$$

$$= [-4, 0, \ldots, 0]$$

1.152

and the **K** matrix becomes

$$\begin{aligned}\mathbf{K} &= \mathbf{BH}+\mathbf{G}\\ &= [-3,\ 1,\ -1,\ -1,\ \underbrace{0,\ 0,\ \ldots,\ 0}_{12\ \text{zeros}}]\\ &= \mathbf{k}'\end{aligned}$$

ORIGINAL PAGE IS
OF POOR QUALITY

The coefficient vector **c** is determined by

$$\mathbf{c} = \mathbf{P}' \otimes \mathbf{A}\mathbf{c} + \mathbf{k} \tag{30}$$

where

$$\mathbf{P}' \otimes \mathbf{A} = \left[\begin{array}{cccccccc|c}
-2 & -2 & -\frac{1}{2} & 0 & -\frac{1}{4} & 0 & 0 & 0 & \\
1 & 0 & 0 & -\frac{1}{2} & 0 & -\frac{1}{4} & 0 & 0 & \\
\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & 0 & \\
0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{4} & -\frac{1}{8}\mathbf{I}_{(8)} \\
\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & \\
\hline
\frac{1}{8} & & & & & & & & \\
 & \frac{1}{8} & & & & & & & \\
 & & \frac{1}{8} & & & & & & \\
 & & & \frac{1}{8} & & & & & \mathbf{0}_{(8)} \\
 & & & & \frac{1}{8} & & & & \\
 & & & & & \frac{1}{8} & & & \\
 & & & & & & \frac{1}{8} & & \\
 & & & & & & & \frac{1}{8} &
\end{array}\right], \quad \mathbf{k} = \begin{bmatrix} -3 \\ +1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \hline 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The answer is obtained from solving (30) for **c** first.

$$\mathbf{c} = \begin{bmatrix} -0{\cdot}9022495 \\ -0{\cdot}4476435 \\ -1{\cdot}3779225 \\ -0{\cdot}7389225 \\ -0{\cdot}2220921 \\ 0{\cdot}1092045 \\ -0{\cdot}3391808 \\ -0{\cdot}1818889 \\ -0{\cdot}1127811 \\ 0{\cdot}0554554 \\ -0{\cdot}1722403 \\ -0{\cdot}0923653 \\ 0{\cdot}0136505 \\ -0{\cdot}0423976 \\ -0{\cdot}0227360 \end{bmatrix} \tag{31}$$

The required $v(\tau)$ is obtained by substituting $\mathbf{c}$ into (26)

$$\begin{aligned} v(\tau) &= \mathbf{c}'\mathbf{P}\boldsymbol{\phi}(\tau) \\ &= 0{\cdot}4701407\phi_0 + 0{\cdot}1417552\phi_1 + 0{\cdot}0861996\phi_2 - 0{\cdot}0697089\phi_3 \\ &\quad + 0{\cdot}0555230\phi_4 - 0{\cdot}0273011\phi_5 + 0{\cdot}0847951\phi_6 + 0{\cdot}0454721\phi_7 \\ &\quad + 0{\cdot}0281952\phi_8 - 0{\cdot}0138380\phi_9 + 0{\cdot}0430600\phi_{10} + 0{\cdot}0230903\phi_{11} \\ &\quad + 0{\cdot}0069403\phi_{12} - 0{\cdot}0034126\phi_{13} + 0{\cdot}0105994\phi_{14} + 0{\cdot}005680\phi_{15} \end{aligned}$$

The comparison between the Walsh series solution and the actual result is shown in Fig. 7 (*c*).

7. Optimal problem with constraint

The optimal control of a linear time-invariant system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{32}$$

with quadratic performance index

$$\delta = \tfrac{1}{2}\int_0^{t_f} (\mathbf{x}'\mathbf{Q}\mathbf{x} + \mathbf{u}'\mathbf{R}\mathbf{u})\, dt \tag{33}$$

is well known to be

$$\mathbf{u}^* = \mathbf{R}^{-1}\mathbf{B}'\mathbf{p}(t) \tag{34}$$

where $p(t)$ satisfies the following canonical equation :

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B}\mathbf{R}^{-1}\mathbf{B}' \\ \mathbf{Q} & -\mathbf{A}' \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{p} \end{bmatrix} \tag{35}$$

and

$$\mathbf{p}(t_f) = \mathbf{0} \tag{36}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \tag{37}$$

Equation (36) is the transversality condition.

It is more convenient to change the independent variable by defining

$$\tau = t_f - t \tag{38}$$

Then (34) becomes

$$\begin{bmatrix} \dot{\mathbf{x}}(\tau) \\ \dot{\mathbf{p}}(\tau) \end{bmatrix} = \begin{bmatrix} -\mathbf{A} & -\mathbf{B}\mathbf{R}^{-1}\mathbf{B}' \\ -\mathbf{Q} & \mathbf{A}' \end{bmatrix} \begin{bmatrix} \mathbf{x}(\tau) \\ \mathbf{p}(\tau) \end{bmatrix} \triangleq -\mathbf{M} \begin{bmatrix} \mathbf{x}(\tau) \\ \mathbf{p}(\tau) \end{bmatrix} \tag{39}$$

The transition matrix of (39) is

$$\exp(-\mathbf{M}t) = \begin{bmatrix} \eta_{11}(\tau) & \eta_{12}(\tau) \\ \eta_{21}(\tau) & \eta_{22}(\tau) \end{bmatrix} \tag{40}$$

Since $\mathbf{p}(\tau = 0) = 0$, the solution of (39) can be written as

$$\mathbf{x}(\tau) = \eta_{11}(\tau)\mathbf{x}(\tau = 0) \tag{41 a}$$

$$\mathbf{p}(\tau) = \eta_{21}(\tau)\mathbf{x}(\tau = 0) \tag{42}$$

p. 154

From (41), we have

$$\mathbf{x}(\tau=0)=\eta_{11}^{-1}(\tau)\mathbf{x}(\tau) \tag{41 b}$$

Substituting (41 a) into (42), yields

$$\mathbf{p}(\tau)=\eta_{21}(\tau)\eta_{11}^{-1}(\tau)\mathbf{x}(\tau)$$

Then

$$\begin{aligned}\mathbf{u}^*(t)&=\mathbf{R}^{-1}\mathbf{B}'\eta_{21}(t_f-t)\eta^{-1}(t_f-t)\mathbf{x}(t_f-t)\\&\triangleq-\mathbf{g}(t_f-t)\mathbf{x}(t_f-t)\end{aligned} \tag{42}$$

where $\mathbf{g}(\tau)$ is the optimal feedback gain matrix.

8. Walsh series solution to the problem

Because the Walsh series is defined from the 0 to 1 interval, we normalize the problem first by using

$$\lambda=\frac{\tau}{t_f} \tag{43}$$

Then (39) becomes

$$\begin{bmatrix}\dot{\mathbf{x}}(\lambda)\\ \dot{\mathbf{p}}(\lambda)\end{bmatrix}=-t_f\mathbf{M}\begin{bmatrix}\mathbf{x}(\lambda)\\ \mathbf{p}(\lambda)\end{bmatrix},\quad 0\leqslant\lambda<1 \tag{44}$$

Next, assume $\dot{\mathbf{x}}(\lambda)$ and $\dot{\mathbf{p}}(\lambda)$ to be expanded into a Walsh series whose coefficients are to be determined :

$$\begin{bmatrix}\dot{\mathbf{x}}(\lambda)\\ \dot{\mathbf{p}}(\lambda)\end{bmatrix}=\begin{bmatrix}\mathbf{c}_1'\\ \mathbf{c}_2'\\ \vdots\\ \mathbf{c}_m'\end{bmatrix}\boldsymbol{\phi} \tag{45}$$

where $\mathbf{c}_i$ and $\boldsymbol{\phi}$ are vectors with m components.

Then use (18)

$$\int_0^\lambda \boldsymbol{\phi}(\nu)\,d\nu\doteq\mathbf{P}\boldsymbol{\phi}(\lambda) \tag{18}$$

to perform integration on (45) :

$$\begin{bmatrix}\mathbf{x}(\lambda)\\ \mathbf{p}(\lambda)\end{bmatrix}\doteq\begin{bmatrix}\mathbf{c}_1'\\ \mathbf{c}_2'\\ \vdots\\ \mathbf{c}_{2n}'\end{bmatrix}\mathbf{P}\boldsymbol{\phi}(\lambda)+\begin{bmatrix}\mathbf{x}(\lambda=0)\\ 0\end{bmatrix} \tag{46}$$

Substituting (46) and (45) into (44) gives

$$\begin{bmatrix}\mathbf{c}_1'\\ \mathbf{c}_2'\\ \vdots\\ \mathbf{c}_{2n}'\end{bmatrix}\boldsymbol{\phi}(\lambda)=-t_f\mathbf{M}\left\{\begin{bmatrix}\mathbf{c}_1'\\ \mathbf{c}_2'\\ \vdots\\ \mathbf{c}_{2n}'\end{bmatrix}\mathbf{P}+\begin{bmatrix}\mathbf{x}(\lambda=0) & \mathbf{0}_{2n},\ \ldots,\ \mathbf{0}_{2n}\\ \mathbf{0}_n & \end{bmatrix}\right\}\boldsymbol{\phi}(\lambda) \tag{47 a}$$

p.155.

Defining

$$-t_f\mathbf{M}\begin{bmatrix}\mathbf{x}(=0) & \\ & \underbrace{\mathbf{0}_{2n},\ \ldots,\ \mathbf{0}_{2n}}_{(m-1)\text{ columns}} \\ \mathbf{0}_n & \end{bmatrix} \triangleq \begin{bmatrix}\boldsymbol{\rho}_1 \\ \boldsymbol{\rho}_2 \\ \vdots \\ \boldsymbol{\rho}_{2n}\end{bmatrix} \tag{48}$$

Then (47) is simplified into

$$\begin{bmatrix}\mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_{2n}\end{bmatrix} = [\mathbf{1}+t_f\mathbf{P}\otimes\mathbf{M}]^{-1}\begin{bmatrix}\boldsymbol{\rho}_1 \\ \boldsymbol{\rho}_2 \\ \vdots \\ \boldsymbol{\rho}_{2n}\end{bmatrix} \tag{47 b}$$

Solving (47 *b*) for $\mathbf{c}_i$, we obtain the Walsh coefficients of the rate variable $\dot{\mathbf{x}}(\lambda)$ and the rate co-stable variable $\dot{\mathbf{p}}(\lambda)$. Then substitute them into (46). The answers to $\mathbf{x}(\lambda)$ and $\mathbf{p}(\lambda)$ in terms of the Walsh functions are finally found.

9. Example of feedback gain determination

Consider the system

$$\dot{\mathbf{x}} = \begin{bmatrix}0 & 0 \\ 1 & 0\end{bmatrix}\mathbf{x} + \begin{bmatrix}1 \\ 0\end{bmatrix}u \tag{49}$$

with

$$\mathbf{x}(0) = \begin{bmatrix}0 \\ 10\end{bmatrix}$$

$$\mathbf{x}(t_f) \text{ unspecified} \tag{50}$$

The performance index is

$$J = \tfrac{1}{2}\int_0^{t_f} (\mathbf{x}'\mathbf{Q}\mathbf{x} + \mathbf{u}'\mathbf{R}\mathbf{u})\,dt \tag{51}$$

where

$$\mathbf{Q} = \begin{bmatrix}0 & 0 \\ 0 & 4\end{bmatrix}, \quad \mathbf{R} = 1$$

the terminal time t_f is $\pi/2$.

The canonical equation for the system is then

$$\begin{bmatrix}\dot{\mathbf{x}}(\lambda) \\ \dot{\mathbf{p}}(\lambda)\end{bmatrix} = \begin{bmatrix}0 & 0 & -\pi/2 & 0 \\ -\pi/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \pi/2 \\ 0 & -\pi/2 & 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{x}(\lambda) \\ \mathbf{p}(\lambda)\end{bmatrix} \tag{52}$$

p156

Express $\dot{\mathbf{x}}$ and $\dot{\mathbf{p}}$ with the Walsh series of 16 terms whose coefficients are to be determined:

$$\begin{bmatrix} \dot{\mathbf{x}}(\lambda) \\ \dot{\mathbf{p}}(\lambda) \end{bmatrix} = \begin{bmatrix} c_{10} & c_{11} & \cdots & c_{1,\,15} \\ c_{20} & c_{21} & \cdots & c_{2,\,15} \\ c_{30} & c_{31} & \cdots & c_{3,\,15} \\ c_{40} & c_{41} & \cdots & c_{4,\,15} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_{15} \end{bmatrix} \tag{53}$$

using (53), (18) to find $\mathbf{C}$ from (47 *b*). Then $\dot{\mathbf{x}}(\lambda)$ and $\dot{\mathbf{p}}(\lambda)$ are determined by (45); so are $\mathbf{x}(\lambda)$ and $\mathbf{p}(\lambda)$ by (46). Finally, we have the optimal feedback gains as follows:

λ	g_1	g_2
1/32	$0{\cdot}2011657 \times 10^{-6}$	$0{\cdot}9639859 \times 10^{-2}$
3/32	$0{\cdot}3784919 \times 10^{-2}$	0·04818776
5/32	0·0189553	0·1251763
7/32	0·05276887	0·2400212
9/32	0·1122418	0·3909498
11/32	0·2028542	0·5740402
13/32	0·3275995	0·7822832
15/32	0·4855795	1·005100
17/32	0·6710423	1·228877
19/32	0·8735645	1·438799
21/32	1·079055	1·621599
23/32	1·275569	1·768292
25/32	1·450156	1·875617
27/32	1·596604	1·945752
29/32	1·712762	1·984705
31/32	1·800200	2·000170

ORIGINAL PAGE IS
OF POOR QUALITY

The analytic solution of this problem is

$$g_1 = [\sinh(\pi - 2t) - \sin(\pi - 2t)] \Big/ \left[\cosh^2\left(\frac{\pi}{2} - t\right) + \cos^2\left(\frac{\pi}{2} - t\right)\right]$$

$$g_2 = [\cosh(\pi - 2t) - \cos(\pi - 2t)] \Big/ \left[\cosh^2\left(\frac{\pi}{2} - t\right) + \cos^2\left(\frac{\pi}{2} - t\right)\right]$$

The comparison of the analytical solution and the Walsh solution is shown in Fig. 8.

10. Conclusion

The Walsh function method for determining the optimal piecewise constant gains for a linear system is established. The basic formula is (47 *b*). Compared with the method of Kleinman *et al.* (1968), the proposed approach is much simpler in analysis and easier in implementation. It is believed that this is the first time in using the Walsh series to approach the most interesting and highly important problem in optimal control.

p.157

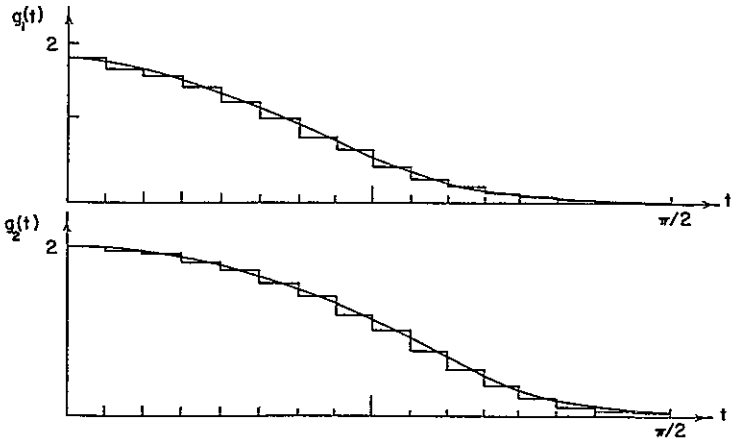


Figure 8.

References

Athans, M., 1967, *Inf. Control*, **10**, 335.

Corrington, M. S., 1973, *I.E.E.E. Trans. Circuit Theory*, **20**, 470.

Elsgolc, L. E., 1961, *Calculus of Variations* (New York: Pergamon Press).

Fortmann, T., 1967, MIT Electronics Systems Laboratory Report ESL–R–326, October.

Harmuth, N. F., 1969, *I.E.E.E. Spectrum*, **6**, 82.

Kleinman, D. L., and Athans, M., 1968, *I.E.E.E. Trans. autom. Control*, **13**, 150.

Kleinman, D. L., Fortmann, T., and Athans, M., 1968, *I.E.E.E. Trans. autom. Control*, **13**, 354.

Proceedings of Walsh Functions Symposium, 1970, 1971, 1972, Naval Research Laboratories, Washington, D.C.

Rademacher, H., 1922, *Math. Ann.*, **87**, 712.

Schechter, R. S., 1967, *The Variational Method in Engineering* (New York: McGraw-Hill Co.), p. 23.

Walsh, J. L., 1923, *Am. J. Math.*, **45**, 5.

Project

(A) Project Title: Design of Piecewise Constant Gains for Optimal Control via Walsh Functions

(B) Project Abstract:

This project developed a technique for determining time-varying feedback gains of linear systems with quadratic performance criteria. The gains are approximated by the piecewise constants which are naturally determined by Walsh functions. After introducing Walsh Functions in the beginning we develop an operational matrix for solving state equations. Then using the operational matrix we solve the piecewise constant gains problem.

(C) Publication: IEEE Transactions on Automatic Control, vol. AC-20, No. 5, P. 596 (1975)

(D) Year: 1975

(E) Department: Electrical Engineering

(F) Student Name: Chi-Huang Hsiao

(G) Faculty Advisor: Professor C. F. Chen

Design of Piecewise Constant Gains for Optimal Control via Walsh Functions

CHIH-FAN CHEN, SENIOR MEMBER, IEEE, AND CHI-HUANG HSIAO, STUDENT MEMBER, IEEE

Abstract—This paper presents a technique for determinating time-varying feedback gains of linear systems with quadratic performance criteria. The gains are approximated by the piecewise constants which are naturally determined by Walsh functions. After introducing Walsh functions in the beginning we develop an operational matrix for solving state equations. Then using the operational matrix we solve the piecewise constant gains problem.

I. Introduction

It is known that the control of a linear system with respect to quadratic performance criteria often involves time varying gain design. Kleinman, Fortmann, and Athans proposed a very elegant approach [1]–[3] to the problem by taking practical engineering constraints into consideration: it was done by prespecifying the structural form of time varying feedback gains, while leaving various free parameters to be chosen in an optimal fashion. Regarding the theoretical investigation, their thinking showed certain similarities between their approach and the direct method of Ritz in calculus of variation.

This paper presents a new approach to the optimal problem by using the Walsh functions [4]–[6]. The approach should be classified as a direct method, but is more powerful than the direct methods of Ritz [7], Euler [7], or Gelerkin [8] on the one hand and is simpler than the procedure proposed by Kleinman, Fortmann, and Athans on the other. The piecewise constant gains obtained by the approach are naturally formed, equally distributed and therefore can be easily implemented.

As a start, we will briefly introduce Walsh functions.

II. Fourier Series and Walsh Series

In direct methods of calculus of variation, we often use Fourier series, power series, etc., with coefficients to be determined as the initial step. Our new approach starts with Walsh series.

It is well known that a function which is periodic may be expanded into Fourier series. Analogously speaking, a function, $f(t)$, which is absolutely integrable in $(0,1]$ may be expanded into Walsh series.

Manuscript received August 30, 1974; revised January 29, 1975. Paper recommended by D. L. Kleinman, Chairman of the IEEE S-CS Optimal Systems Committee.

C.-F. Chen is with the Department of Electrical Engineering, University of Houston, Houston, Tex. 77004.

C.-H. Hsiao was with the Department of Electrical Engineering, University of Houston, Houston, Tex. 77004. He is now with Cheng Kung University, Taiwan, China.

$$f(t)=c_0\phi_0(t)+c_1\phi_1(t)+c_2\phi_2(t)+\cdots+c_n\phi_n(t)+\cdots \tag{1}$$

where

$$c_n=\int_0^1 \phi_n(t)f(t)\,dt \tag{2}$$

are determined such that the following integral square error is minimized:

$$\epsilon=\int_0^1\left[f(t)-\sum_{n=0}^{N} c_n\phi_n(t)\right]^2 dt. \tag{3}$$

The Walsh functions $\phi_0(t)$, $\phi_1(t),\cdots,\phi_n(t)$ are a set of square waves which are orthonormal. Fig. 1. shows the Walsh functions from ϕ_0 to ϕ_{11} in the dyadic order. However, the regularity of the set of Walsh functions in the dyadic order is not very easily seen. Each Walsh function can be decomposed into more elementary square waves, or Rademacher functions [10].

Rademacher functions, $r_k(t)$, are a set of square waves of unit height with periods equal to 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8},\cdots,2^{(1-k)}$. Alternatively, we can state that the number of cycles of the square waves of $r_k(t)$ is 2^{k-1}. The first six waves of Rademacher are shown in Fig. 2. It is seen that

$$\begin{aligned}\phi_0(t)&=r_0(t)\\ \phi_1(t)&=r_1(t)\\ &\vdots\\ \phi_i(t)&=(r_k(t))^{b_k}(r_{k-1}(t))^{b_{k-1}}(r_{k-2}(t))^{b_{k-2}}\cdots\end{aligned} \tag{4}$$

where

$$k=[\log_2 i]+1 \tag{5}$$

where $[\bullet]$ means taking the integer part of "$\bullet$" and $b_k,b_{k-1},\cdots,b_1$ is the binary number expression of i.

To draw the wave form of any Walsh function becomes a trivial matter if we use the above mentioned decomposition technique.

Let us then return to the Walsh coefficient evaluation for a function. Consider a given function $f(t)=t$. It is desired to expand it into Walsh series. Substituting $f(t)=t$

Copyright ©1975 by The Institute of Electrical and Electronics Engineers, Inc.
Printed in U.S.A. Annals No. 510AC002

1160

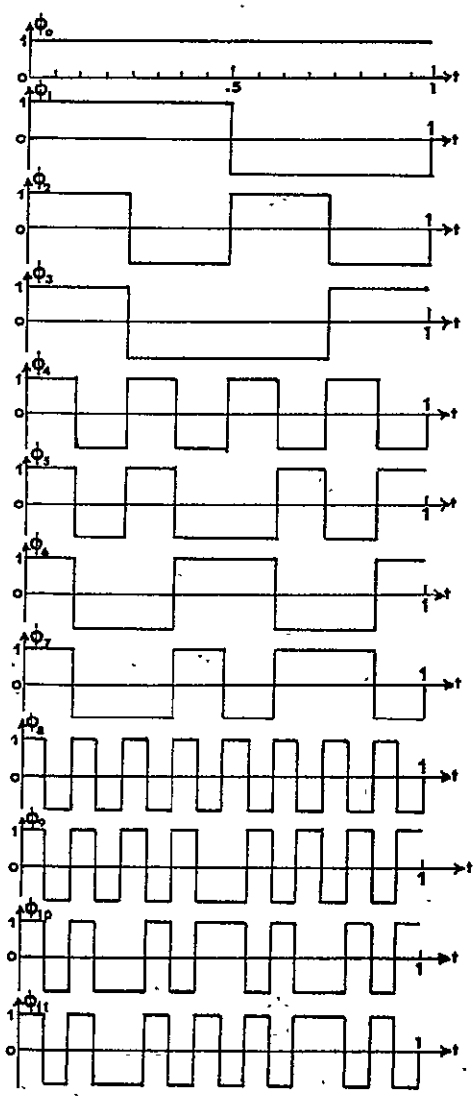


Fig. 1. Dyadic ordered Walsh functions.

into (2), we have

$$f(t)=t=\tfrac{1}{2}\phi_0(t)-\tfrac{1}{4}\phi_1(t)-\tfrac{1}{8}\phi_2(t)-\tfrac{1}{16}\phi_4(t)-\tfrac{1}{32}\phi_8(t)-\tfrac{1}{64}\phi_{16}(t)\cdots. \quad (6)$$

The original curve $f(t)=t$ and its Walsh series approximations are shown in Fig. 3. They are stairwise waves. The crudest representation is obtained by taking one term of the Walsh series, or $\frac{1}{2}\phi_0$; the second one consists of two terms $\frac{1}{2}\phi_0-\frac{1}{4}\phi_1$. The figure shows up to a four term approximation which is

$$f(t)=t\simeq\tfrac{1}{2}\phi_0-\tfrac{1}{4}\phi_1-\tfrac{1}{8}\phi_2-\tfrac{1}{16}\phi_4. \quad (6a)$$

From the coefficient evaluation process, we easily see that the similarities between Fourier series and Walsh series are obvious.

III. Integration and Operational Matrix

In this section we will derive a method by which we can perform any integration by multiplying a constant matrix.

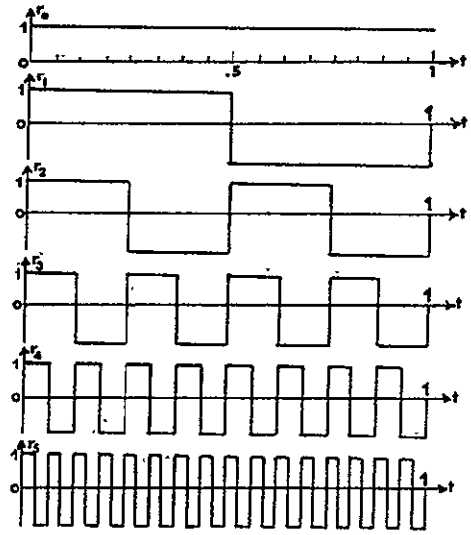


Fig. 2. Rademacher functions.

Let us take $\phi_0,\phi_1,\cdots,\phi_7$ and integrate them; we will have various triangular waves [6]. If we evaluate the Walsh coefficients for these triangular waves, we will easily arrive at the following formula for approximation:

ORIGINAL PAGE IS
OF POOR QUALITY

$$\begin{bmatrix}\int\phi_0dt\\ \int\phi_1dt\\ \int\phi_2dt\\ \int\phi_3dt\\ \int\phi_4dt\\ \int\phi_5dt\\ \int\phi_6dt\\ \int\phi_7dt\end{bmatrix}=\begin{bmatrix}\frac{1}{2} & \frac{-1}{4} & \frac{-1}{8} & 0 & \frac{-1}{16} & 0 & 0 & 0\\ \frac{1}{4} & 0 & 0 & \frac{-1}{8} & 0 & \frac{-1}{16} & 0 & 0\\ \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{-1}{16} & 0\\ 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & \frac{-1}{16}\\ \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\phi_0\\ \phi_1\\ \phi_2\\ \phi_3\\ \phi_4\\ \phi_5\\ \phi_6\\ \phi_7\end{bmatrix} \quad (7)$$

or

$$\int\Phi_{(8)}dt=P_{(8\times 8)}\Phi_{(8)}. \quad (7a)$$

The subscript means the dimension taken. It is preferable to take 2^{Ω}, where Ω is an integer, as a dimension number. Making this choice will enable us to obtain simpler results and to have a much easier calculation.

It is noted that

$$\int\phi_0dt=t;$$

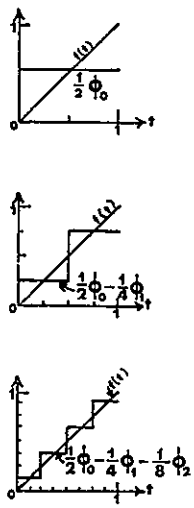


Fig. 3. Expanding a ramp function into Walsh functions.

therefore the Walsh coefficients of $\int \phi_0 dt$ is found in (6a) which is simply the first row of (7) times the $\phi(t)$ vector.

Equation (7) is for $m=8$, a general formula $P_{(m\times m)}$ can be written as follows:

$$P_{(m\times m)} = \left[\begin{array}{c|c|c|c} \frac{1}{2} & \frac{-2}{m} I_{(m/8)} & \frac{-1}{m} I_{(m/4)} & \\ \hline \frac{2}{m} I_{(m/8)} & 0_{(m/8)} & & \frac{-1}{2m} I_{(m/2)} \\ \hline & \frac{1}{m} I_{(m/4)} & 0_{(m/4)} & \\ \hline & \frac{1}{2m} I_{(m/2)} & & 0_{(m/2)} \end{array}\right]. \quad (8)$$

It is interesting to note that if (8) is partitioned into four parts as shown, the upper left part of $P_{(m\times m)}$ is identical to $P_{(\overline{m/2}\times\overline{m/2})}$; and the upper left corner of $P_{(\overline{m/2}\times\overline{m/2})}$ is $P_{(\overline{m/4}\times\overline{m/4})}$. Therefore, this regularity of the structure of the P matrix enables us to write the mth enlarged matrix to any dimension, if the dimension number is restricted to 2^{Ω} where Ω is an integer number.

IV. State Equation Solution by Kronecker Product Formula

Consider the following state equation:

$$\dot{x} = Ax + Bu, \qquad x(0) = x_0 \quad (9)$$

where x is a state vector of n components and u is an input vector of l components. A and B are $n\times n$ and $n\times l$ matrices, respectively. We would like to establish a procedure for solving the state equations via Walsh series.

First of all, we assume the rate variable vector $\dot{x}$ to be a set of m-term Walsh series whose $n\times m$ coefficients are to be determined. Let

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} c_{10} & c_{11} & c_{12} & \cdots & c_{1(m-1)} \\ c_{20} & c_{21} & c_{22} & \cdots & c_{2(m-1)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ c_{n0} & c_{n1} & c_{n2} & \cdots & c_{n(m-1)} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_{(m-1)} \end{bmatrix}. \quad (10)$$

We write each column as a vector and define the rectangular matrix as C. Then (10) becomes

$$\dot{x} = [c_0, c_1, \cdots, c_{m-1}]\phi \triangleq C\phi. \quad (10a)$$

It should be noted that this initial step is quite different than that of the regular procedure of the series solution of a differential equation. We assume the rate variable $\dot{x}$ as an undetermined vector series, instead of assuming the state variable itself.

The state variable x may be obtained by integration:

$$x(t) = C\int_0^t \phi(s)ds + x_0. \quad (11)$$

The integration can be performed approximately by using the P matrix.

$$\int_0^t \phi(s)ds = P\phi(t). \quad (7a)$$

Also the input vector can also be expressed by Walsh series:

$$u = H\phi \quad (12)$$

where H is a $l\times m$ matrix; the determination of which can be achieved by the techniques illustrated in Section II.

Substituting (10a), (11), (7a), and (12) into (9) yields

$$C\phi = A(CP\phi + x_0) + BH\phi. \quad (13)$$

Ax_0 can be written as the product of a matrix G and the vector ϕ.

$$Ax_0 = Ax_0\phi_0 = [Ax_0, \underbrace{0, 0, \cdots, 0}_{(m-1)\text{columns}}]\phi \triangleq G\phi. \quad (14)$$

Finally we have

$$C = ACP + G + BH. \quad (15)$$

The last two terms of (15) can be combined by letting

$$G + BH \triangleq K. \quad (16)$$

Then

$$C = ACP + K. \quad (15a)$$

p. 162

If we rearrange the $n \times m$ matrix C as an nm-vector c by changing its first column into the first n components of the vector; the second column, the second n components of the vector, etc.; and rearrange K in the same manner; we finally obtain an even simpler form in terms of a Kronecker product for (15a);

$$c = [A \otimes P']c + k \tag{17}$$

where $A \otimes P'$ is a Kronecker product defined as

$$A \otimes P' \triangleq \begin{bmatrix} p_{11}A & p_{21}A & \cdots & p_{m1}A \\ p_{12}A & p_{22}A & \cdots & p_{m2}A \\ \cdots & \cdots & \cdots & \cdots \\ p_{1m}A & p_{2m}A & \cdots & p_{mm}A \end{bmatrix} \tag{18}$$

and P' is the transposition of P.

The solution of c comes from (17) directly,

$$c = [I - A \otimes P']^{-1}k. \tag{19}$$

Once c has been decided, the Walsh series representation for the rate variable is determined. The state variable vector is then found by substitution.

$$x = CP\phi + x_0. \tag{20}$$

When m is large, the computation of (19) becomes a difficult problem, we take advantage of the special pattern of matrix P and develop an algorithm for calculation. The details are shown in the Appendix.

A high-order differential equation with constant coefficients can always be written as a state equation. Therefore (19) and (20) enable us to solve linear time invariant systems elegantly and completely. Of course, the answer is in terms of Walsh functions.

A subroutine WALDE has been written in Fortran language for solving an nth-order differential equation.

V. Optimal Problem

The optimal control of a linear time-invariant system

$$\dot{x} = Ax + Bu \tag{21}$$

with quadratic performance index

$$\delta = \tfrac{1}{2}\int_0^{t_f}(x'Qx + u'Ru)\,dt \tag{22}$$

is well known to be

$$u^* = R^{-1}B'p(t) \tag{23}$$

where $p(t)$ satisfies the following canonical equation:

$$\begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} A & BR^{-1}B' \\ Q & -A' \end{bmatrix}\begin{bmatrix} x \\ p \end{bmatrix} \tag{24}$$

and the boundary conditions are specified as

$$x(0) = x_0 \tag{25}$$

$$p(t_f) = 0. \tag{26}$$

ORIGINAL PAGE IS
OF POOR QUALITY

Equation (26) is the transversality condition.

It is more convenient to change the independent variable by defining

$$\tau = t_f - t; \tag{27}$$

then (24) becomes

$$\begin{bmatrix} \dot{x}(\tau) \\ \dot{p}(\tau) \end{bmatrix} = \begin{bmatrix} -A & -BR^{-1}B' \\ -Q & A' \end{bmatrix}\begin{bmatrix} x(\tau) \\ p(\tau) \end{bmatrix} \triangleq -M\begin{bmatrix} x(\tau) \\ p(\tau) \end{bmatrix}. \tag{28}$$

The transition matrix of (28) is

$$e^{-Mt} = \begin{bmatrix} \eta_{11}(\tau) & \eta_{12}(\tau) \\ \eta_{21}(\tau) & \eta_{22}(\tau) \end{bmatrix}. \tag{29}$$

Since $p(\tau = 0) = 0$, the solution of (28) can be written as

$$x(\tau) = \eta_{11}(\tau)x(\tau = 0) \tag{30}$$

$$p(\tau) = \eta_{21}(\tau)x(\tau = 0). \tag{31}$$

From (30), we have

$$x(\tau = 0) = \eta_{11}^{-1}(\tau)x(\tau). \tag{30a}$$

Substituting (30a) into (31) yields

$$p(\tau) = \eta_{21}(\tau)\eta_{11}^{-1}(\tau)x(\tau). \tag{31a}$$

Then optimal control is reduced to

$$\begin{aligned} u^*(t) &= R^{-1}B'\eta_{21}(t_f - t)\eta_{11}^{-1}(t_f - t)x(t_f - t) \\ &\triangleq -L(t_f - t)x(t_f - t) \end{aligned} \tag{32}$$

where $L(t_f - t)$ is the optimal feedback gain matrix.

VI. Walsh Series Solution to the Problem

Because Walsh series is defined in the 0 to 1 interval we normalize the problem first by using

$$\lambda = \tau / t_f; \tag{33}$$

then (28) becomes

$$\begin{bmatrix} \dot{x}(\lambda) \\ \dot{p}(\lambda) \end{bmatrix} = -t_f M\begin{bmatrix} x(\lambda) \\ p(\lambda) \end{bmatrix}, \qquad 0 < \lambda < 1. \tag{34}$$

Next, assume $\dot{x}(\lambda)$ and $\dot{p}(\lambda)$ to be expanded into Walsh series whose coefficients are to be determined.

$$\begin{bmatrix} \dot{x}(\lambda) \\ \dot{p}(\lambda) \end{bmatrix} = C\phi(\lambda) \tag{35}$$

where C is an $2n \times m$ matrix, and $\phi(\lambda)$, an m-vector.

p. 163

Then (7a) is applied to perform integration on (35):

$$\begin{bmatrix} x(\lambda) \\ p(\lambda) \end{bmatrix} = CP\phi(\lambda) + \begin{bmatrix} x(\lambda=0) \\ 0_n \end{bmatrix}. \tag{36}$$

Substituting (36) and (35) into (34) gives

$$C\phi(\lambda) = -t_f M \left\{ CP + \begin{bmatrix} x(\lambda=0) & 0_{2n}, \cdots, 0_{2n} \\ 0_n & \end{bmatrix} \right\} \phi(\lambda). \tag{37}$$

Defining k as

$$k = \begin{bmatrix} -t_f A x(\lambda=0) \\ -t_f Q x(\lambda=0) \\ 0_{2n} \\ \vdots \\ 0_{2n} \end{bmatrix}. \tag{38}$$

then (37) is simplified into

$$c = [I + t_f M \otimes P']^{-1} k. \tag{37a}$$

Solving (37a) for c, we obtain the Walsh coefficients of the rate variable $\dot{x}(\lambda)$ and the rate co-state variable $\dot{p}(\lambda)$. Then substitute them into (36). The answer of $x(\lambda)$ and $p(\lambda)$ in terms of Walsh function are finally found.

VII. Example of Feedback Gain Determination

Let us consider Kleinman's example [1], [2],

$$\dot{x}(t) = Ax(t) + Bu(t)$$
$$= \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & -2 \\ 0 & 2 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} u(t), \qquad x(0) = x_0. \tag{39}$$

The performance index is specified as

$$\delta = \frac{1}{2} \int_0^2 (x'Qx + u'Ru)\, dt$$
$$= \frac{1}{2} \int_0^2 \left\{ x' \begin{bmatrix} 2 & -2 & 0 \\ -2 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} x + 2u^2 \right\} dt. \tag{40}$$

Since $t_f = 2 \neq 1$, we need to normalize the time scale.

$$\lambda = (t_f - t)/t_f = 1 - 0.5t. \tag{41}$$

The canonical equation then becomes

$$\begin{bmatrix} \dot{x}(\lambda) \\ \dot{p}(\lambda) \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & -4 & -4 & 2 \\ 0 & 0 & 4 & -4 & -4 & 2 \\ 0 & -4 & 0 & 2 & 2 & -1 \\ -4 & 4 & 0 & -2 & 0 & 0 \\ 4 & -4 & 0 & 0 & 0 & -4 \\ 0 & 0 & 0 & 0 & 4 & 0 \end{bmatrix} \begin{bmatrix} x(\lambda) \\ p(\lambda) \end{bmatrix}. \tag{42}$$

Assuming that the Walsh expansion of $\dot{x}(\lambda)$ and $\dot{p}(\lambda)$ may be approximately expressed with 16 terms, we have

$$\begin{bmatrix} \dot{x}(\lambda) \\ \dot{p}(\lambda) \end{bmatrix} = C\phi(\lambda) \tag{43}$$

where C is a 6×16 matrix to be determined and $\phi(\lambda)$ a Walsh vector of 16 components. Letting $x_1(\lambda=0)=1$ and all other initial conditions $(\lambda=0)$ equal to zero, and applying (37a), we obtain the first column of $\eta_{11}(\lambda)$ and $\eta_{21}(\lambda)$. Then, letting $x_2(\lambda=0)=1$, we obtain the second column of $\eta_{11}(\lambda)$ and $\eta_{21}(\lambda)$; and letting $x_3(\lambda=0)=1$, we obtain the third column. The optimal gain is evaluated with (32), which is a multiplication of matrices. The waveform is shown at the lower right hand corner of Fig. 4.

Taking the average for each pair of consecutive values of the above obtained optimal gains with 16 subintervals, we obtain a new set of piecewise feedback gain for 8 subintervals. Continuing this procedure, we obtain constant feedback gain for 4, 2, and 1 subintervals. All of them are shown in Fig. 4.

The cost matrix associated with the gain matrix $L(t)$ is a unique solution of the following:

$$\dot{V}(t) = -V(t)[A - BL(t)] - [A - BL(t)]'V(t) - Q - L'(t)RL(t) \tag{44}$$

$$V(t_f) = 0 \tag{44a}$$

and μ is defined as the trace of $V(0)$, namely

$$\mu = \mathrm{tr}[V(0)]. \tag{45}$$

For a detailed explanation of $V(t)$ and μ, the reader is referred to [1].

We evaluate $V(0)$ and μ for each control law we have previously derived. These μ's are tabulated below.

Number of Subintervals	16	8	4	2	1
μ	1.7125	1.7115	1.7114	1.7471	1.9063

These values are slightly larger than those of Kleinman, but our method is much simpler.

VIII. Conclusion

A Walsh function method for determining the optimal piecewise constant gains for a linear system is established. The basic formula is (37a). Compared with Kleinman, Fortman and Athans method, the proposed approach is much simpler in analysis and easier in implementation.

Appendix

Algorithm for Solving C Via the Kronecker Product Formula

In this Appendix we derive a recursive algorithm to solve C from (19) instead of inversing $[I - A \otimes P']$ directly. Let us illustrate the procedures for $m = 2^3 = 8$.

p.164

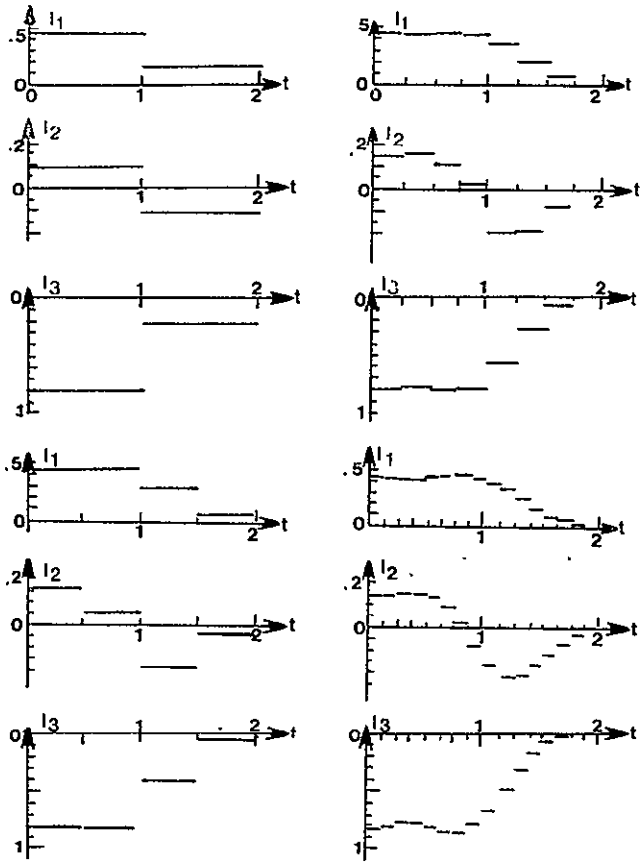


Fig. 4. Solving Kleinman's problem via the Walsh function method.

Step 1: Equation (15a) may be rewritten explicitly for $m=8$.

$$\begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \\ \mathbf{c}_4 \\ \mathbf{c}_5 \\ \mathbf{c}_6 \\ \mathbf{c}_7 \end{bmatrix} = \begin{bmatrix} (\frac{1}{2})A & (\frac{1}{4})A & (\frac{1}{8})A & 0 & (\frac{1}{16})A & 0 & 0 & 0 \\ (\frac{-1}{4})A & 0 & 0 & (\frac{1}{8})A & 0 & (\frac{1}{16})A & 0 & 0 \\ (\frac{-1}{8})A & 0 & 0 & 0 & 0 & 0 & (\frac{1}{16})A & 0 \\ 0 & (\frac{-1}{8})A & 0 & 0 & 0 & 0 & 0 & (\frac{1}{16})A \\ (\frac{-1}{16})A & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & (\frac{-1}{16})A & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & (\frac{-1}{16})A & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & (\frac{-1}{16})A & 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \\ \mathbf{c}_4 \\ \mathbf{c}_5 \\ \mathbf{c}_6 \\ \mathbf{c}_7 \end{bmatrix} + \begin{bmatrix} \mathbf{k}_0 \\ \mathbf{k}_1 \\ \mathbf{k}_2 \\ \mathbf{k}_3 \\ \mathbf{k}_4 \\ \mathbf{k}_5 \\ \mathbf{k}_6 \\ \mathbf{k}_7 \end{bmatrix}. \tag{A1}$$

It is obvious that $\mathbf{c}_4$, $\mathbf{c}_5$, $\mathbf{c}_6$, and $\mathbf{c}_7$ may be expressed in terms of $\mathbf{c}_0$, $\mathbf{c}_1$, $\mathbf{c}_2$, and $\mathbf{c}_3$.

$$\begin{bmatrix} \mathbf{c}_4 \\ \mathbf{c}_5 \\ \mathbf{c}_6 \\ \mathbf{c}_7 \end{bmatrix} = \begin{bmatrix} (\frac{-1}{16})A & 0 & 0 & 0 \\ 0 & (\frac{-1}{16})A & 0 & 0 \\ 0 & 0 & (\frac{-1}{16})A & 0 \\ 0 & 0 & 0 & (\frac{-1}{16})A \end{bmatrix} \cdot \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{k}_4 \\ \mathbf{k}_5 \\ \mathbf{k}_6 \\ \mathbf{k}_7 \end{bmatrix}. \tag{A2}$$

ORIGINAL PAGE IS
OF POOR QUALITY

In order to keep a consistent notation, we may define

$$G_3 \triangleq I \tag{A3}$$

$$R_3 \triangleq -(1/2^{3+1})G_3^{-1}A; \tag{A4}$$

then

$$\mathbf{c}_{i+4} = R_3\mathbf{c}_i + \mathbf{k}_{i+4}, \qquad i=0,1,2,3. \tag{A5}$$

Substituting (A2) into (A1), we have

$$\begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{bmatrix} = \begin{bmatrix} (\frac{1}{2})A + (\frac{-1}{256})A^2 & (+\frac{1}{4})A & (\frac{1}{8})A & 0 \\ (\frac{-1}{4})A & (\frac{-1}{256})A^2 & 0 & (\frac{1}{8})A \\ (\frac{-1}{8})A & 0 & (\frac{-1}{256})A^2 & 0 \\ 0 & (\frac{-1}{8})A & 0 & (\frac{-1}{256})A^2 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{k}_0 + (\frac{1}{16})A\mathbf{k}_4 \\ \mathbf{k}_1 + (\frac{1}{16})A\mathbf{k}_5 \\ \mathbf{k}_2 + (\frac{1}{16})A\mathbf{k}_6 \\ \mathbf{k}_3 + (\frac{1}{16})A\mathbf{k}_7 \end{bmatrix}. \tag{A6}$$

Now the diagonal elements are no longer null matrices. We may define the new diagonal matrix as F_3, and the new inputs as $\mathbf{k}_{i,\mathrm{III}}$.

$$F_3 \triangleq \tfrac{-1}{256}A^2 = \frac{1}{2^{3+1}}AR_3 \tag{A7}$$

$$\mathbf{k}_{i,\mathrm{III}} \triangleq \mathbf{k}_i + \frac{1}{2^{3+1}}AG_3^{-1}\mathbf{k}_{i+4}, \qquad i=0,1,2,3. \tag{A8}$$

Step 2: Eliminating $\mathbf{c}_2$, $\mathbf{c}_3$ from the lower half of (A6), we have

$$\begin{bmatrix} \mathbf{c}_2 \\ \mathbf{c}_3 \end{bmatrix} = \begin{bmatrix} (\frac{-1}{8})(I+\frac{1}{256}A^2)^{-1}A & 0 \\ 0 & (-\frac{1}{8})(I+\frac{1}{256}A^2)^{-1}A \end{bmatrix} \cdot \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{bmatrix} + (I+\tfrac{1}{256}A^2)^{-1}\begin{bmatrix} \mathbf{k}_{2,\mathrm{III}} \\ \mathbf{k}_{3,\mathrm{III}} \end{bmatrix}. \tag{A9}$$

This may be put into recursive form by defining

$$G_2 \triangleq I - F_3 = I + \tfrac{1}{256}A^2 \tag{A10}$$

$$R_2 \triangleq \frac{1}{2^{2+1}}G_2^{-1}A = (-\tfrac{1}{8})(I+\tfrac{1}{256}A^2)^{-1}A; \tag{A11}$$

then

$$\mathbf{c}_{i+2} = R_2\mathbf{c}_i + G_2^{-1}\mathbf{k}_{(i+2),\mathrm{III}}, \qquad i=0,1. \tag{A12}$$

Substituting (A9) and (A12) into (A6), we have

$$\begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{bmatrix} = \begin{bmatrix} (\frac{1}{2})A + F_3 \frac{1}{64} A\left(I + \frac{1}{256}A^2\right)^{-1} A & (\frac{1}{4})A \\ (\frac{-1}{4})A & F_3 \frac{-1}{64} A\left(I + \frac{1}{256}A^2\right)^{-1} A \end{bmatrix} \cdot \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{bmatrix} + \begin{bmatrix} \mathbf{k}_{0,\mathrm{III}} + \frac{1}{8} A G_2^{-1} k_{2,\mathrm{III}} \\ \mathbf{k}_{1,\mathrm{III}} + \frac{1}{8} A G_2^{-1} k_{3,\mathrm{III}} \end{bmatrix}. \quad \text{(A13)}$$

For recursive operation, we should define again

$$F_2 \triangleq F_3 + \frac{1}{2^{2+1}} A R_2 = -\tfrac{1}{256}A^2 - \tfrac{1}{64}A\left(I + \tfrac{1}{256}A^2\right)^{-1} A \quad \text{(A14)}$$

$$\mathbf{k}_{i,\mathrm{II}} \triangleq \mathbf{k}_{i,\mathrm{III}} + \frac{1}{2^{2+1}} A G_2^{-1} \mathbf{k}_{i+2,\mathrm{III}}, \qquad i = 0, 1. \quad \text{(A15)}$$

Step 3: Eliminating $\mathbf{c}_1$ from (A13) and (A14), we have

$$\mathbf{c}_1 = \left(\frac{-1}{4}\right)(I - F_2)^{-1} A \mathbf{c}_0 + (I - F_2)^{-1} \mathbf{k}_{i,\mathrm{II}} \triangleq R_1 \mathbf{c}_0 + G_1^{-1} \mathbf{k}_{i,\mathrm{II}} \quad \text{(A16)}$$

where G_1, R_1 are defined in the same form as G_2, R_2, G_3, R_3.

$$G_1 \triangleq I - F_2 \quad \text{(A17)}$$

$$R_1 \triangleq -\frac{1}{2^{i+1}} G_1^{-1} A. \quad \text{(A18)}$$

Substituting (A16) and (A14) into the upper half of (A13), we have

$$\mathbf{c}_0 = \left[(\tfrac{1}{2})A + F_2 + (\tfrac{1}{4})AR_1\right]\mathbf{c}_0 + \mathbf{k}_{0,\mathrm{II}} + (\tfrac{1}{4})AG_1^{-1}\mathbf{k}_{i,\mathrm{II}}.$$

Therefore

$$\mathbf{c}_0 = G_0^{-1} \mathbf{k}_{0,\mathrm{I}} \quad \text{(A19)}$$

where

$$G_0 \triangleq I - (\tfrac{1}{2})A - F_1 \quad \text{(A20)}$$

$$F_1 \triangleq F_2 + \frac{1}{2^{1+1}} A R_1 \quad \text{(A21)}$$

$$\mathbf{k}_{0,\mathrm{I}} \triangleq \mathbf{k}_{0,\mathrm{II}} + \frac{1}{2^{1+1}} A G_1^{-1} \mathbf{k}_{1,\mathrm{II}}. \quad \text{(A22)}$$

Step 4: $\mathbf{c}_1$ is obtained by substituting $\mathbf{c}_0$ into (A16). Similarly, $\mathbf{c}_2$, $\mathbf{c}_3$, and $\mathbf{c}_4$, $\mathbf{c}_5$, $\mathbf{c}_6$, $\mathbf{c}_7$ are obtained by applying (A12) and (A5), respectively. This completes the derivation of the algorithm for solving c from (28), with $m = 8$.

In general, for $m = 2^\alpha$, α is any positive integer, we start with

$$G_\alpha = I, \quad R_\alpha = -\frac{1}{2^{\alpha+1}} A, \quad F_\alpha = \frac{1}{2^{\alpha+1}} A R_\alpha$$

$$F_{\alpha+1} = 0, \quad k_{i,\alpha+1} = k_i. \quad \text{(A23)}$$

Then, we can calculate R_β and $\mathbf{k}_{i,\beta}$, $\beta = \alpha, \alpha-1, \cdots, 1$, $i = 0, 1, \cdots, 2^{\beta-1}$, from the following recursive formulas:

$$G_\beta = I - F_{\beta+1}$$

$$R_\beta = -2^{-\beta-1} G_\beta^{-1} A$$

$$F_\beta = F_{\beta+1} + 2^{-\beta-1} A R_\beta$$

$$\mathbf{k}_{i,\beta} = \mathbf{k}_{i,\beta+1} + 2^{-\beta-1} A G_\beta^{-1} \mathbf{k}_{(i+2^{\beta-1}),\beta+1}. \quad \text{(A24)}$$

Then we obtain $\mathbf{c}_0$:

$$\mathbf{c}_0 = G_0^{-1} \mathbf{k}_{0,\mathrm{I}}, \; G_0 = \mathbf{1} - \tfrac{1}{2}A - F_1. \quad \text{(A25)}$$

All the other vectors $\mathbf{c}_j, j = 0, 1, \cdots, (m-1)$ are found by substituting them with

$$\mathbf{c}_{i+2\beta-1} = R_\beta \mathbf{c}_i + G_\beta^{-1} \mathbf{k}_{(i+2^{\beta-1}),\beta+1},$$

$$\beta = 1, 2, \cdots, \alpha \quad i = 0, 1, \cdots, (2^{\beta-1} - 1). \quad \text{(A26)}$$

In the above algorithm, we always work with matrices (G, F, R) of $n \times n$. Therefore there is no need to operate with larger matrices. For instance, if we wish to apply 128 Walsh functions to solve a 6th-order differential equation, we only work with matrices 6×6. The calculation of matrix inverse of $[I - A \otimes P']$, which is of 768×768, is avoided. Therefore, we have saved computing time and storage. In addition, we have reduced round-off errors significantly.

References

[1] D. L. Kleinman, T. Fortmann, and M. Athans, "On the design of linear systems with piecewise-constant feedback gains," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 354–361, Aug. 1968.

[2] D. L. Kleinman, and M. Athans, "The design of suboptimal linear time-varying systems," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 150–159, Apr. 1968.

[3] T. Fortmann, "Optimal piece constant solutions of the linear regulator problems," MIT Electron. Syst. Lab. Rep. ESL-326, Oct. 1967.

[4] *Proc. Walsh Functions Symp.*, Nav. Res. Lab., Washington, D. C., 1970, 1971, 1972.

[5] N. F. Harmuth, "Application of Walsh functions in communication," *IEEE Spectrum*, pp. 82–91, Nov. 1969.

[6] M. S. Corrington, "Solution of differential and integral equations with Walsh functions," *IEEE Trans. Circuit Theory*, vol. CT-20, pp. 470–476., Sept. 1973.

[7] L. E. Elsgoe, *Calculus of Variations*. New York: Pergamon, 1961.

[8] R. S. Schechter, *The Variation Method in Engineering*. New York: McGraw-Hill, 1967, pp. 23–26.

[9] J. L. Walsh, "A closed set of orthogonal functions," *Amer. J. Math.*, vol. 45, pp. 5–24, 1923.

[10] H. Rademacher, "Einge Satze Uber Reihen Von Allgemen Orthogonal Function," *Ann. Math.*, vol. 87, pp. 712–738, 1922.

[11] M. Athans, "The matrix minimum principle," *Inform. Contr.*, Nov. 1967.

Chih-Fan Chen (SM'61) was born in Pa, Hopei, China on June 19, 1924. He received the B.S. degree from Pei-Yang University, China, the M.S. degree from the University of Pennsylvania, Philadelphia, and the Ph.D. degree from Cambridge University, Cambridge, England.

During the 1950's he was a Research Associate at the University of Pennsylvania and Princeton University, Princeton, N. J. In the early 1960's he was a Professor at the Christian Brothers College, Memphis, Tenn. Since 1966 he has been Professor of Electrical Engineering, University of Houston, Houston, Tex. In 1962 he

p.166

was a Technical Assistant at the International Atomic Energy Agency. During the summers of 1963, 1964, and 1967 he was a Visiting Scholar at the Argonne National Laboratory. During the summers of 1972, 1974, and 1975 he was a Consultant for the U.S. Army Missile Command. He was a Visiting Professor at the National Taiwan University, the National Tsing Hwa University, and the National Cheng Kung University during the summers of 1962, 1971, and 1973, respectively. He is author and co-author of more than 50 papers on identification, model reduction and stability, and co-author of the book, *Elements of Control Systems Analysis* (Prentice-Hall, 1968). He was awarded the Science Faculty Fellowship twice by the National Science Foundation and was elected to Fellow at St. Edmond House, Cambridge University, England.

Dr. Chen is a member of the American Society for Engineering Education, Sigma Xi, and Phi Kappa Phi.

Chi-Huang Hsiao (S'72) was born in Hong Kong, China on February 25, 1932. He received the B.S. degree from the Chung Cheng Institute of Technology, China, the M.S. degree from the National Chiao Tung University, Taiwan, China, the M.S.E.E. degree from Columbia University, New York, N.Y., and the Ph.D. degree from Cheng Kung University, Tawain, China.

Since 1969 he has been with the Chung Shan Institute of Science and Technology, China. During the academic year of 1973–1974 he was with the Department of Electrical Engineering, University of Houston, Houston, Tex. His research interests are in the fields of stochastic processes, digital simulation, and time domain synthesis.

ORIGINAL PAGE IS
OF POOR QUALITY

p.167

Project

(A) Project Title: A State-Space Approach to Walsh Series Solution of Linear Systems

(B) Project Abstract:

A state-space procedure for solving linear dynamic systems by the Walsh series is developed. A new operational matrix plays the main role and a new Kronecker product formula is established. The laborious use of Corrington's table is eliminated. Several examplex illustrate the process and demonstrate the power of the approach.

(C) Publication: International Journal of Systems Science, 6, 833 (1975)

(D) Year: 1975

(E) Department: Electrical Engineering

(F) Student Name: C. H. Hsiao

(G) Faculty Advisor: Professor C. F. Chen

INT. J. SYSTEMS SCI., 1975, VOL. 6, NO. 9, 833–858

A state-space approach to Walsh series solution of linear systems

C. F. CHEN† and C. H. HSIAO‡

A state-space procedure for solving linear dynamic systems by the Walsh series is developed. A new operational matrix plays the main role and a new Kronecker product formula is established. The laborious use of Corrington's tables is eliminated. Several examples illustrate the process and demonstrate the power of the approach.

1. Introduction

Walsh's series has been widely used in the analysis of communication theory (Harmuth 1969, 1972, Lee 1970, Jones 1972, Dust 1972), optical engineering (Gibbs and Gebbie 1969, Pratt *et al* 1972, Kennedy 1971), bioscience (Milne *et al.* 1972, Thomas and Welch 1972), data processing (Andrews 1972, Kuhn *et al.* 1963, Shanks 1969, Ahmed 1972, Chen 1972), electromagnetic radiation (Harmuth 1970, 1972, Pearlman 1970), pattern recognition (Ito 1970, Andrews 1971, Carl and Kabrisky 1971, Clark *et al.* 1972), and control systems (Dinh *et al.* 1972, Picher 1970 a, b, Gibbs and Millard 1969). In a recent paper (Corrington 1973), Corrington applied it to a more fundamental problem : the solution of linear or non-linear differential and integral equations. His technique is (1) to assume a Walsh series whose coefficients are to be determined, (2) to perform integrations by using n tables for an nth order differential equation, and (3) to iterate to a convergent answer.

This paper presents a state space formulation, in which only an operational matrix is involved. For the linear case, iteration steps are eliminated. The new approach is much simpler in theory and more suitable for digital computation.

1.1. *Rademacher functions*

Notation of the Walsh function approach has not been unified. A brief introduction is summarized as follows.

In 1922, Rademacher (1922) developed a set of functions shown in Fig. 1 which is a set of square waves of unit height with periods equal to 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, ..., $2^{(1-k)}$, respectively. In general the number of cycles of the square wave of $r_k(t)$ is 2^{k-1} (Fig. 1). Obviously, it is a set of odd functions about $t=\frac{1}{2}$ and an orthonormal system. For example

$$\int_0^1 r_1(t)r_2(t)\,dt=0 \tag{1 a}$$

$$\int_0^1 r_2^2(t)\,dt=1 \tag{1 b}$$

Received 30 October 1974.

† Department of Electrical Engineering, University of Houston, Houston, Texas 77004, U.S.A.

‡ Chung Shan Institute of Science, Lung Tan, Taiwan, Republic of China.

J.169

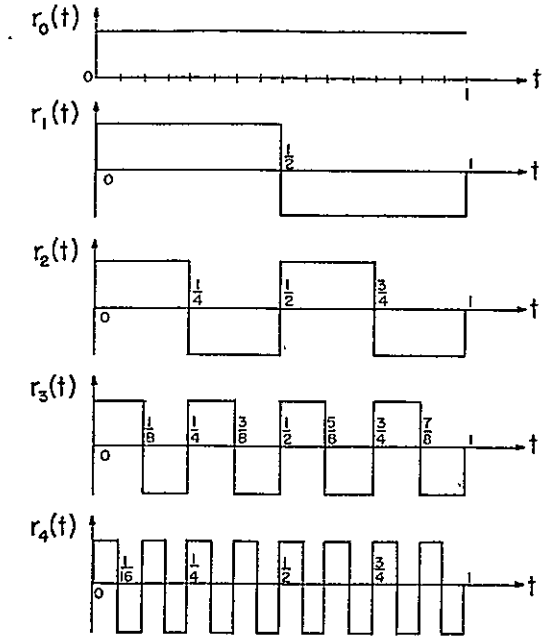


Figure 1. Rademacher functions.

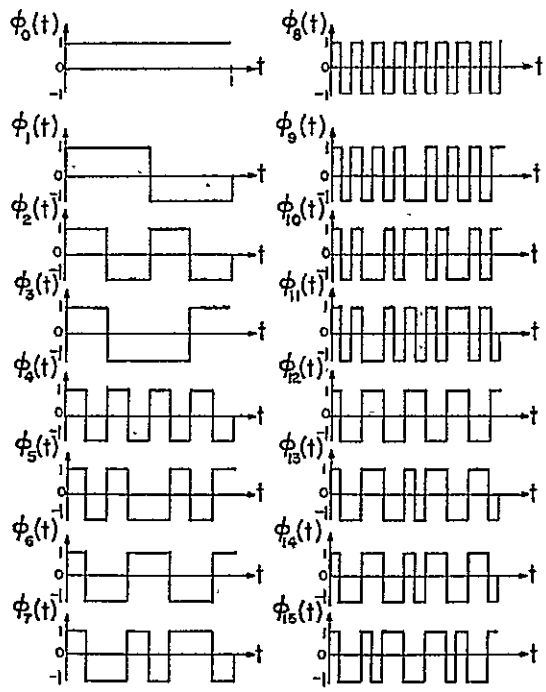


Figure 2. Walsh functions.

p. 170

ORIGINAL PAGE IS
OF POOR QUALITY

Or, in general

$$\int_0^1 r_m(t) r_n(t)\, dt = \begin{cases} 0 & \text{if } m \neq n \\ 1 & \text{if } m = n \end{cases} \tag{1}$$

Because the Rademacher set consists only of odd functions, it is incomplete. In 1923 Walsh (1923) independently developed a complete set which is known as Walsh functions (Fig. 2).

1.2. *Derivation of Walsh functions*

The set of Walsh functions is similar to and derivable from Rademacher functions (Paley 1932); and it is a complete set of orthonormal systems. The relationship between Rademacher functions and Walsh functions is as follows:

$$\begin{aligned}
\phi_0(t) &= r_0(t) \\
\phi_1(t) &= r_1(t) \\
\phi_2(t) &= (r_2(t))^1 (r_1(t))^0 \\
\phi_3(t) &= (r_2(t))^1 (r_1(t))^1 \\
\phi_4(t) &= (r_3(t))^1 (r_2(t))^0 (r_1(t))^0 \\
\phi_5(t) &= (r_3(t))^1 (r_2(t))^0 (r_1(t))^1 \\
\phi_6(t) &= (r_3(t))^1 (r_2(t))^1 (r_1(t))^0 \\
\phi_7(t) &= (r_3(t))(r_2(t))^1 (r_1(t))^1 \\
&\vdots \\
\phi_n(t) &= (r_q(t))^\alpha (r_{q-1}(t))^\beta (r_{q-2}(t))^\gamma \ldots
\end{aligned} \tag{2}$$

ORIGINAL PAGE IS
OF POOR QUALITY

where

$$q = [\log_2 n] + 1 \tag{3}$$

in which $[\cdot]$ means taking the greatest integer of '$\cdot$'. And,

$$\alpha \cdot 2^{q-1} + \beta \cdot 2^{q-2} + r \cdot 2^{q-3} + \ldots = n \tag{4}$$

Or $\alpha\beta\gamma \ldots$ is the binary expansion of n.

1.3. *Illustrative example*

Express Walsh function $\phi_9(t)$ by Rademacher functions.

$$\left.\begin{aligned}
n &= 9 \\
q &= [\log_2 n] + 1 \\
&= 4 \\
9 &= \underset{\alpha}{1} \cdot 2^3 + \underset{\beta}{0} \cdot 2^2 + \underset{\gamma}{0} \cdot 2^1 + \underset{\delta}{1} \cdot 2^0
\end{aligned}\right\} \tag{4 a}$$

Substituting the values of α, β, γ and δ of (4 a) into the general expression of (2), we obtain

$$\phi_9(t) = (\gamma_4(t))^1 (\gamma_3(t))^0 (\gamma_2(t))^0 (\gamma_1(t))^1$$

which means that Walsh function $\phi_9(t)$ is decomposed into Rademacher functions as shown in Fig. 3.

p. 171

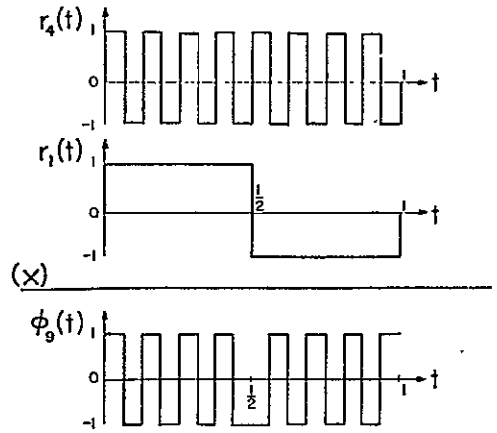


Figure 3. Walsh function $\phi_9(t)$ is obtained from the product Rademacher functions.

By inspection, we have the following corresponding relationship.

$$\begin{aligned}
\phi_0(t) &= r_0(t) \\
\phi_2(t) &= r_2(t) \\
\phi_4(t) &= r_3(t) \\
\phi_8(t) &= r_4(t) \\
&\vdots \\
\phi_n(t) &= r_q(t), \quad n = 2^{q-1}
\end{aligned} \tag{2 a}$$

It is seen that the Rademacher functions are contained in the Walsh functions; while the latter is a complete orthonormal set and the former is not.

2. Walsh coefficient evaluation

A function $f(t)$ that is absolutely integrable in [0, 1) may be expanded as a Walsh series.

$$\begin{aligned}
f(t) &= c_0\phi_0(t) + c_1\phi_1(t) + c_2\phi_2(t) + \ldots c_n\phi_n(t) + \ldots \\
&= \sum_{n=0}^{\infty} c_n\phi_n(t)
\end{aligned} \tag{5}$$

where $\{c_n\}$ are called the coefficients of Walsh series.

Our problem is to determine the coefficient c_n so that the integral square error satisfies,

$$\lim_{N\to\infty} \int_0^1 [f(t) - \sum_{n=0}^{N} c_n\phi_n(t)]^2\, dt = 0 \tag{6}$$

Multiplying $\phi_n(t)$ on both sides of (5) and then integrating from 0 to 1, we have

$$\int_0^1 f(t)\phi_n(t)\, dt = \int_0^1 c_0\phi_0(t)\phi_n(t)\, dt + \int_0^1 c_1\phi_1(t)\phi_n(t)\, dt + \ldots \times \int_0^1 c_n\phi_n^2(t)\, dt + \ldots \tag{7}$$

p. 172

Every term on the right-hand side of (7) is equal to zero, due to the orthogonal property, except the square of $\phi_n(t)$ term ; therefore

$$c_n = \int_0^1 \phi_n(t) f(t)\, dt \tag{8}$$

2.1. *Gate function example*

Let us expand the gate function shown in Fig. 4.

$$f(t) = \begin{cases} 1 & \frac{1}{4} \leqslant t \leqslant \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

as a Walsh series.

Assume

$$f(t) = \sum_{n=0}^{\infty} c_n \phi_n(t)$$

ORIGINAL PAGE IS
OF POOR QUALITY

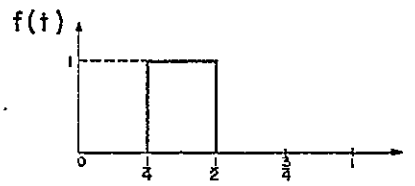


Figure 4. A gate function.

The Walsh coefficients can be obtained by substituting into (8)

$$c_0 = \int_0^1 f(t) \cdot \phi_0(t)\, dt = \int_{1/4}^{1/2} 1 \cdot 1\, dt = \tfrac{1}{4}$$

$$c_1 = \int_0^1 f(t) \cdot \phi_1(t)\, dt = \int_{1/4}^{1/2} 1 \cdot 1\, dt = \tfrac{1}{4}$$

$$c_2 = \int_0^1 f(t) \cdot \phi_2(t)\, dt = \int_{1/4}^{1/2} 1 \cdot (-1)\, dt = -\tfrac{1}{4}$$

$$c_3 = \int_0^1 f(t) \cdot \phi_3(t)\, dt = \int_{1/4}^{1/2} 1 \cdot (-1)\, dt = -\tfrac{1}{4}$$

$$c_4 = c_5 = c_6 = \ldots = 0$$

Therefore

$$f(t) = \tfrac{1}{4}\phi_0(t) + \tfrac{1}{4}\phi_1(t) - \tfrac{1}{4}\phi_2(t) - \tfrac{1}{4}\phi_3(t)$$

2.2. *Sine function example*

Express a sinusoidal wave form

$$f(t) = \sin(\pi t), \quad 0 < t < 1$$

into Walsh series (say, only taking the first eight terms).

p. 173

Assume

$$f(t)=\sin(\pi t)\cong\sum_{n=0}^{7}c_n\phi_n(t)\triangleq f^*(t)$$

Substituting into (10), we obtain

$$c_0=\int_0^1\sin(\pi t)\,dt=\frac{1}{\pi}[-\cos(\pi t)]_0^1=\frac{2}{\pi}=0{\cdot}637$$

$$c_1=\int_0^{1/2}\sin(\pi t)\,dt+\int_{1/2}^{1}-\sin(\pi t)\,dt=0$$

$$c_2=\int_0^{1/4}\sin(\pi t)\cdot dt-\int_{1/4}^{1/2}\sin(\pi t)+\int_{1/2}^{3/4}\sin(\pi t)\,dt-\int_{3/4}^{1}\sin(\pi t)\,dt=0$$

$$c_3=2\left\{\int_0^{1/4}(\pi t)\,dt-\int_{1/4}^{1/2}\sin(\pi t)\,dt\right\}=\frac{2}{\pi}\left\{-\frac{\sqrt{2}}{2}-1)-\frac{\sqrt{2}}{2}\right\}$$

$$=-0{\cdot}263$$

$$c_4=0$$

$$c_5=2\left\{\int_0^{1/8}\sin(\pi t)\,dt-\int_{1/8}^{1/4}\sin(\pi t)\,dt+\int_{1/4}^{3/8}\sin(\pi t)-\int_{3/8}^{1/2}\sin(\pi t)\,dt\right\}$$

$$=-0{\cdot}126$$

$$c_6=2\left\{\int_0^{1/8}\sin(\pi t)\,dt-\int_{1/8}^{3/8}\sin(\pi t)\,dt+\int_{1/8}^{1/2}\sin(\pi t)\,dt\right\}$$

$$=-0{\cdot}0573$$

$$c_7=0$$

The resultant of the eight components is tabulated below and the comparison of the original curve and the partial sum of the Walsh series is shown in Fig. 5.

t	1/16	3/16	5/16	7/16	9/16	11/16	13/16	15/16
$f^*(t)$	0·196	0·55	0·825	0·976	0·976	0·825	0·55	0·196

(T 1)

2.3. *Ramp function example*

Let us expand a unit ramp function t which is shown in Fig. 6 into Walsh series ; i.e.

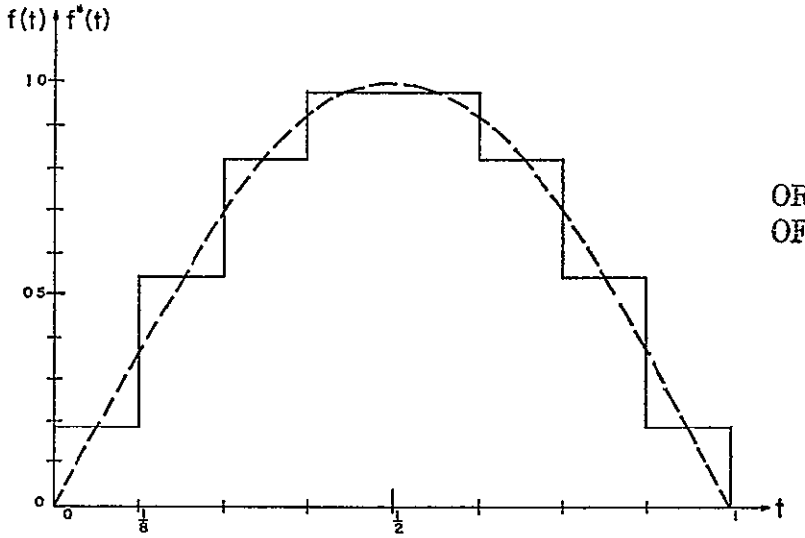
$$f(t)=t$$

Substituting into (8) yields

$$\left.\begin{aligned}c_0&=\int_0^1\phi_0(t)\cdot t\cdot dt=\tfrac{1}{2}\\ c_1&=\int_0^1\phi_1(t)\cdot t\cdot dt=-\tfrac{1}{4}\\ c_2&=\int_0^1\phi_2(t)\cdot t\cdot dt=-\tfrac{1}{8}\\ c_3&=\int_0^1\phi_3(t)\cdot t\cdot dt=0\\ &\vdots\end{aligned}\right\}\qquad(9)$$

Therefore

$$t\cong\tfrac{1}{2}\phi_0(t)-\tfrac{1}{4}\phi_1(t)-\tfrac{1}{8}\phi_2(t)+\ \ldots$$



ORIGINAL PAGE IS
OF POOR QUALITY

Figure 5. Sine function and its Walsh series approximation.

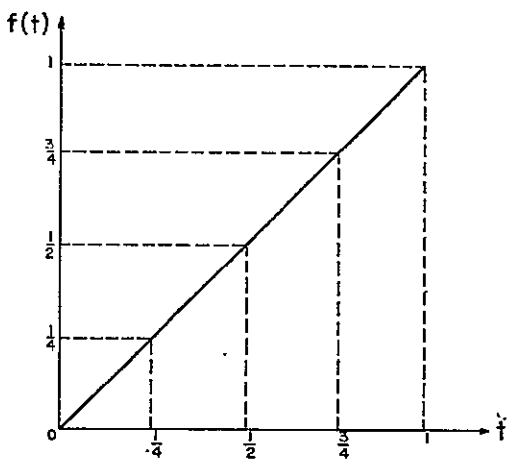


Figure 6. Unit ramp function.

3. Discrete formula

If the given function is not described in a closed form and its Walsh series is desired, we can easily modify (5) and (8) into their discrete forms :

$$\bar{f}_k = \sum_{n=0}^{m-1} \phi_{kn} c_n, \qquad k=0, 1, 2, \ldots, (m-1) \tag{5 a}$$

$$c_n = \sum_{k=0}^{m-1} \phi_{nk} \bar{f}_k \cdot \frac{1}{m}, \quad n=0, 1, 2, \ldots, (m-1) \tag{8 a}$$

where $\bar{f}_k$ is the average value of the function in question in the kth subinterval, ϕ_{nk} is the value of the nth Walsh function in the kth subinterval, and m is the total number of subdivisions.

p. 175

3.1. *Ramp function example again*

For illustrating the use of discrete formulae, let us evaluate the Walsh series for the ramp function again. The given set of data $\{f_k\}$ is as follows :

k	f_k
0	0
1	1/4
2	2/4
3	3/4
4	4/4

(T 2)

We convert the table into a discrete form by taking the average value over each interval and obtain the following corresponding table.

k	$\bar{f}_k$
0	$\frac{1}{2}(0+\frac{1}{4})=\frac{1}{8}$
1	$\frac{1}{2}(\frac{1}{4}+\frac{2}{4})=\frac{3}{8}$
2	$\frac{1}{2}(\frac{2}{4}+\frac{3}{4})=\frac{5}{8}$
3	$\frac{1}{2}(\frac{3}{4}+\frac{4}{4})=\frac{7}{8}$

(T 3)

Substituting (T 3) into (8 a) and writing it in a matrix form, we have

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \bar{f}_1 \\ \bar{f}_2 \\ \bar{f}_3 \\ \bar{f}_4 \end{bmatrix} \cdot \frac{1}{4} \tag{10 a}$$

In a compact form

$$\mathbf{c} = \mathbf{W}\bar{\mathbf{f}} \cdot \frac{1}{m} \tag{10}$$

where $\mathbf{W}$ is called the Walsh matrix and is obtained from the definition of the Walsh functions. Figure 7 shows the correspondence.

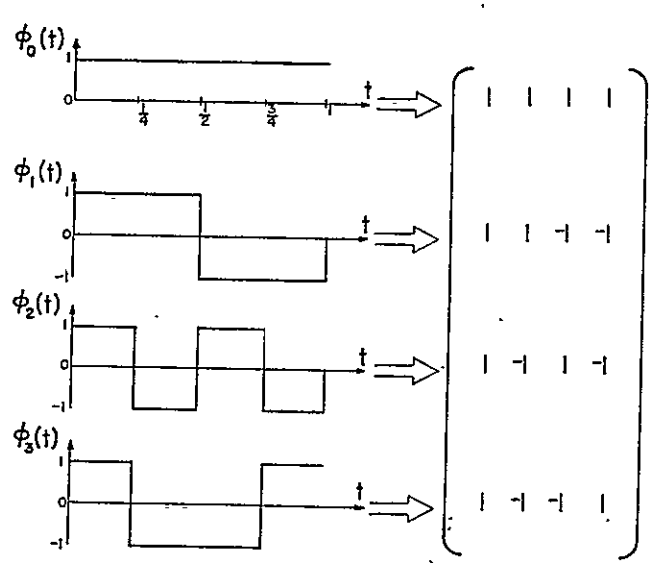


Figure 7. A set of Walsh functions and its Walsh matrix correspondence.

p.176

Substituting $\bar{f}_k$ of (T 3) into (10 *a*) and evaluating c_n, we obtain the same results as shown in (9) of the previous section.

3.2. *Triangular function example*

Let us consider a triangular function shown in Fig. 8. The given data are in numerical form, or

k	f_k
0	0
1	$\frac{1}{4}$
2	$\frac{1}{2}$
3	$\frac{1}{4}$
4	0

(T 4)

The Walsh coefficients of this function are desired.

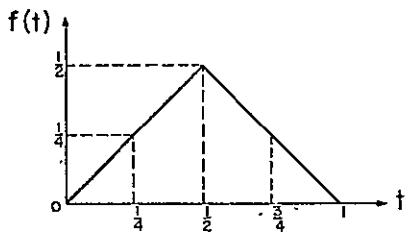


Figure 8. A triangular function.

ORIGINAL PAGE IS
OF POOR QUALITY

First of all, we take the average values of what in each interval and have the following,

k	$\bar{f}_k$
0	$\frac{1}{2}(0+\frac{1}{4})=\frac{1}{8}$
1	$\frac{1}{2}(\frac{1}{4}+\frac{1}{2})=\frac{3}{8}$
2	$\frac{1}{2}(\frac{1}{2}+\frac{1}{4})=\frac{3}{8}$
3	$\frac{1}{2}(\frac{1}{4}+0)=\frac{1}{8}$

(T 5)

Substituting (T 5) into (8 *a*) yields,

$$\mathbf{c} \triangleq \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \mathbf{W}\bar{\mathbf{f}} \cdot \frac{1}{m} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{8} \\ \frac{3}{8} \\ \frac{3}{8} \\ \frac{1}{8} \end{bmatrix} \cdot \frac{1}{4} = \begin{bmatrix} \frac{1}{4} \\ 0 \\ 0 \\ -\frac{1}{8} \end{bmatrix}$$

3.3. *Double triangular function example*

The double triangular functions shown in Fig. 9 are of our particular interest. Their Walsh coefficients can be evaluated routinely by first taking the average numerical values and then substituting into (8 *a*).

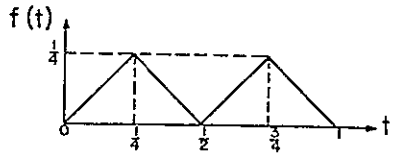


Figure 9. A double triangular function.

k	f_k	$\bar{f}_k$
0	0	
1	$\frac{1}{4}$	$\frac{1}{2}(0+\frac{1}{4})=\frac{1}{8}$
2	0	$\frac{1}{2}(\frac{1}{4}+0)=\frac{1}{8}$
3	$\frac{1}{4}$	$\frac{1}{2}(0+\frac{1}{4})=\frac{1}{8}$
4	0	$\frac{1}{2}(\frac{1}{4}+0)=\frac{1}{8}$

(T 6)

$$\mathbf{c} \triangleq \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \mathbf{W}\bar{\mathbf{f}} \cdot \frac{1}{m} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{8} \\ \frac{1}{8} \\ \frac{1}{8} \\ \frac{1}{8} \end{bmatrix} \cdot \frac{1}{4} = \begin{bmatrix} \frac{1}{8} \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

3.4. *Alternating triangular function example*

The alternating triangular function shown in Fig. 10 is nearly the same as that of Fig. 9 except with the second triangle inverted. The sampled values of the alternating triangular function and its average value in each interval are tabulated below :

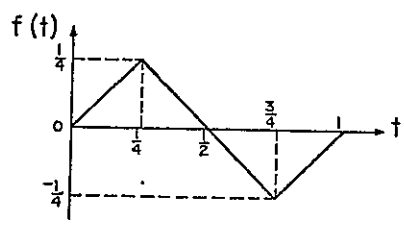


Figure 10. An alternating triangular function.

k	f_k	$\bar{f}_k$
0	0	
1	$\frac{1}{4}$	$\frac{1}{2}(0+\frac{1}{4}(=\frac{1}{8}$
2	0	$\frac{1}{2}(\frac{1}{4}+0)=\frac{1}{8}$
3	$-\frac{1}{4}$	$\frac{1}{2}(0-\frac{1}{4})=-\frac{1}{8}$
4	0	$\frac{1}{2}(-\frac{1}{4}+0)=-\frac{1}{8}$

(T 7)

Substituting (T 7) into (8 *a*) again, we have

$$\mathbf{c} \triangleq \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \mathbf{W}\bar{\mathbf{f}} \cdot \frac{1}{m} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{8} \\ \frac{1}{8} \\ -\frac{1}{8} \\ -\frac{1}{8} \end{bmatrix} \cdot \frac{1}{4} = \begin{bmatrix} 0 \\ \frac{1}{8} \\ 0 \\ 0 \end{bmatrix}$$

f. 178

ORIGINAL PAGE IS
OF POOR QUALITY

4. Derivation of operational matrix

We recall that Walsh functions are a set of rectangular waves. Their integrals are various triangular waves. We summarize the facts as shown in Fig. 11.

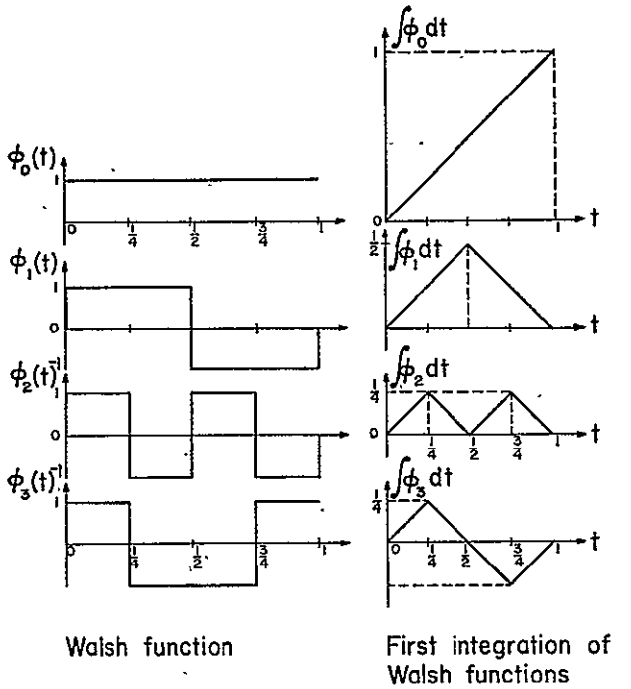


Figure 11. Four-interval Walsh functions and their first integrals.

The first integration of Walsh functions are expressible by Walsh functions and each one has been evaluated in the previous section. Therefore, we can write the relationship between Walsh functions and their integrals in the matrix form

$$\begin{bmatrix} \int \phi_0\, dt \\ \int \phi_1\, dt \\ \int \phi_2\, dt \\ \int \phi_3\, dt \end{bmatrix} \cong \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{8} & 0 \\ \frac{1}{4} & 0 & 0 & -\frac{1}{8} \\ \frac{1}{8} & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix} \tag{11 a}$$

or in compact form,

$$\int \boldsymbol{\Phi}_{(4)}\, dt = \mathbf{P}_{(4\times 4)} \boldsymbol{\Phi}_{(4)} \tag{11}$$

P is called the operational matrix which relates the Walsh functions and their integrals. It is chosen as a square matrix for convenient computation and its dimension depends on the number of components chosen.

4.1. 8×8 *operational matrix*

Equation (11) is an approximate formula, its accuracy depends on the dimension of $\boldsymbol{\Phi}$ or **P**. We can follow a similar reasoning stated in § 3 to derive a larger **P** matrix. If we have 8 subdivisions between 0 and 1, we have eight components of the Walsh series. The Walsh components and their integrals are shown in Fig. 12.

p.179

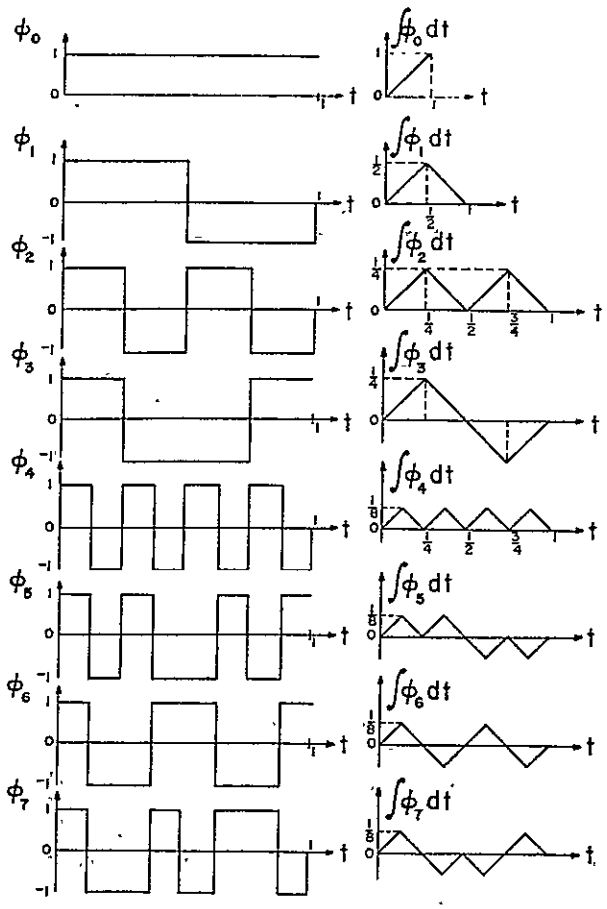


Figure 12. Eight-interval Walsh functions and their first integrals.

Evaluating analytically or numerically of the triangular functions, we obtain the following matrix

$$\begin{bmatrix} \int \phi_0\, dt \\ \int \phi_1\, dt \\ \int \phi_2\, dt \\ \int \phi_3\, dt \\ \int \phi_4\, dt \\ \int \phi_5\, dt \\ \int \phi_6\, dt \\ \int \phi_7\, dt \end{bmatrix} = \left[\begin{array}{cccc|cccc} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{8} & 0 & -\frac{1}{16} & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & -\frac{1}{8} & 0 & -\frac{1}{16} & 0 & 0 \\ \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{16} & 0 \\ 0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{16} \\ \hline \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \\ \phi_7 \end{bmatrix} \tag{12 a}$$

p. 180

Or,

$$\int \boldsymbol{\Phi}_{(8)}\, dt = \mathbf{P}_{(8\times 8)}\boldsymbol{\Phi}_{(8)} \tag{12}$$

It is interesting to note that the upper left corner of $\mathbf{P}_{(8\times 8)}$ is exactly $\mathbf{P}_{(4\times 4)}$ as shown in (12). The upper right corner matrix and the lower left down corner matrix are diagonal matrices, and the lower right corner matrix is simply a null matrix.

ORIGINAL PAGE IS
OF POOR QUALITY

4.2. 16×16 *operational matrix*

Following a similar reasoning line, we can easily establish the 16×16 operational matrix as shown in eqn. (13)

$$\mathbf{P}_{(16\times 16)} = \left[\begin{array}{cccc|cccc|cccc|cccc}
\frac{1}{2} & -\frac{1}{4} & -\frac{1}{8} & 0 & -\frac{1}{16} & 0 & 0 & 0 & -\frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & 0 & -\frac{1}{8} & 0 & -\frac{1}{16} & 0 & 0 & 0 & -\frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
\frac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{16} & 0 & 0 & 0 & -\frac{1}{32} & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\frac{1}{16} & 0 & 0 & 0 & -\frac{1}{32} & 0 & 0 & 0 & 0 \\
\hline
\frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{32} & 0 & 0 & 0 \\
0 & \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{32} & 0 & 0 \\
0 & 0 & -\frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{32} & 0 \\
0 & 0 & 0 & -\frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{32} \\
\hline
\frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{32} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right] \tag{13}$$

This looks like table (2) of Corrington. However, he did not present his table (2) as a matrix nor used it as a matrix.

p.181

If we partition 16×16 operational matrix, we observe that the upper left corner is the same as $\mathbf{P}_{(8\times 8)}$, and then partition the submatrix $\mathbf{P}_{(8\times 8)}$, the upper left corner of which is $\mathbf{P}_{(4\times 4)}$, etc. Therefore we have

$$\mathbf{P}_{(16\times 16)} = \left[\begin{array}{cc|c|c|c} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{8}\mathbf{I}_{(2)} & & \\ \hline \frac{1}{4} & 0 & & -\frac{1}{16}\mathbf{I}_{(4)} & \\ \hline \multicolumn{2}{c|}{\frac{1}{8}\mathbf{I}_{(2)}} & \mathbf{0}_{(2)} & & \\ \hline \multicolumn{3}{c|}{\frac{1}{16}\mathbf{I}_{(4)}} & \mathbf{0}_{(4)} & -\frac{1}{32}\mathbf{I}_{(8)} \\ \hline \multicolumn{4}{c|}{\frac{1}{32}\mathbf{I}_{(8)}} & \mathbf{0}_{(8)} \end{array}\right] \tag{14}$$

This regular pattern enables us to construct an operational matrix with any large dimension. Because Corrington used the arrangement of (13) as a table, he did not recognize the regularity of the pattern shown in (14).

4.3. *General operational matrix*

When we assume the undetermined coefficients series matrix in the first step of solution, we chose m terms ; the value of m is assumed to be

$$m = 2^{\alpha} \tag{15}$$

where α is a positive integer. This assumption which is only for the convenience of computation, of course, is not absolutely necessary.

If we do choose m so that (15) holds, the general operational matrix can be written as follows :

$$\mathbf{P}_{(m\times m)} = \left[\begin{array}{c|c|c|c} \frac{1}{2} \ddots & -\frac{1}{m/2}\mathbf{I}_{(m/8)} & -\frac{1}{m}\mathbf{I}_{(m/4)} & \\ \hline \frac{1}{m/2}\mathbf{I}_{(m/8)} & \mathbf{0}_{(m/8)} & & -\frac{1}{2m}\mathbf{I}_{(m/2)} \\ \hline \multicolumn{2}{c|}{\frac{1}{m}\mathbf{I}_{(m/4)}} & \mathbf{0}_{(m/4)} & \\ \hline \multicolumn{3}{c|}{\frac{1}{2m}\mathbf{I}_{(m/2)}} & \mathbf{0}_{(m/2)} \end{array}\right] \tag{16}$$

5. Solution of state equations

Corrington derived n different tables for solving an nth-order differential equation. It is tedious and unnecessary. From a state space viewpoint, any nth order differential equation can be converted into a set of n state equations, and one table suffices for the solution. The new approach is to take advantage of the state space formulation and derive a unique Kronecker product formula for the solution.

J.182

5.1. *Derivation of solution formula*

Given a set of state equations

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{17}$$

where $\mathbf{x}$ is a state vector of n components, $\mathbf{u}$ is an input vector of l components. $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ and $n \times l$ matrices, respectively. For solving this problem by the Walsh series approach, we assume the rate vector $\dot{\mathbf{x}}$ instead of state vector $\mathbf{x}$ as a set of Walsh series. Let

ORIGINAL PAGE IS OF POOR QUALITY

$$\left.\begin{aligned} \dot{x}_1 &= c_{10}\phi_0 + c_{11}\phi_1 + c_{12}\phi_2 + \cdots \\ \dot{x}_2 &= c_{20}\phi_0 + c_{21}\phi_1 + c_{22}\phi_2 + \cdots \\ &\vdots \\ x_n &= c_{n0}\phi_0 + c_{n1}\phi_1 + c_{n2}\phi_2 + \cdots \end{aligned}\right\} \tag{18}$$

where c_{ij} are constants to be determined. Once we know the solution $\dot{\mathbf{x}}$ we can obtain the solution $\mathbf{x}$ in a straightforward manner.

Because Walsh series is not only orthonormal but also convergent fast, we can use a finite number of terms, say m terms, to approximate the actual solution. In other words, it is justified to assume that

$$\dot{\mathbf{x}} \cong \begin{bmatrix} c_{10} & c_{11} \cdots & c_{1(m-1)} \\ c_{20} & c_{21} \cdots & c_{2(m-1)} \\ \cdots & \cdots \cdots & \cdots \\ \cdots & \cdots \cdots & \cdots \\ c_{n0} & c_{n1} \cdots & c_{n(m-1)} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{(m-1)} \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \\ \vdots \\ \mathbf{c}_n' \end{bmatrix} \boldsymbol{\phi} \tag{19 a}$$

or

$$\dot{\mathbf{x}} \cong \mathbf{C}_{(n\times m)}\boldsymbol{\phi}_{(m)} \tag{19}$$

where $\mathbf{C}_{(n\times m)}$ is an $n \times m$ rectangular matrix and $\boldsymbol{\phi}_{(m)}$ is a vector with m components, or

$$\mathbf{C}_{(n\times m)} = \begin{bmatrix} c_{10} & c_{11} & c_{12} \cdots & c_{1(m-1)} \\ c_{20} & c_{21} & c_{22} \cdots & c_{2(m-1)} \\ \cdots & \cdots & \cdots \cdots & \cdots \\ \cdots & \cdots & \cdots \cdots & \cdots \\ c_{n0} & c_{n1} & c_{n2} \cdots & c_{n(m-1)} \end{bmatrix} \tag{20}$$

$$\triangleq [\mathbf{c}_0 \quad \mathbf{c}_1 \quad \mathbf{c}_2 \cdots c_{(m-1)}]$$

$$\boldsymbol{\phi}_{(m)}' = [\phi_0, \phi_1, \phi_2, \ldots, \phi_{m-1}] \tag{21}$$

Prime means transpose.

The state variable $\mathbf{x}$ may be obtained by integration

$$\mathbf{x}(t) = C \int_0^t \boldsymbol{\phi}(\lambda)\, d\lambda + \mathbf{x}_0 \tag{22}$$

However, the integral can be evaluated approximately via the operational matrix $\mathbf{P}$ as we mentioned in the previous section. We then have,

$$\int_0 \boldsymbol{\phi}(\lambda)\, d\lambda = \mathbf{P}\boldsymbol{\phi}(t) \tag{23}$$

p.183

The input function $\mathbf{u}(t)$ can be also expressed as a Walsh series

$$\mathbf{u}(t) \triangleq \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ \vdots \\ u(t) \end{bmatrix} \cong \begin{bmatrix} h_{10} & h_{11} & h_{12} \dots h_{1(m-1)} \\ h_{20} & h_{21} & h_{22} \dots h_{2(m-1)} \\ \dots & \dots & \dots \dots \dots \\ \dots & \dots & \dots \dots \dots \\ h_{l0} & h_{l1} & h_{l2} \dots h_{l(m-1)} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \vdots \\ \phi_{m-1} \end{bmatrix}$$

$$\cong \mathbf{H}\boldsymbol{\phi} \tag{24}$$

$\mathbf{u}(t)$ is a known vector, i.e. all the elements of $\mathbf{H}$ matrix are known.

Substituting (20), (23), and (24) into (17), we have

$$\mathbf{C}\boldsymbol{\phi} = \mathbf{ACP}\boldsymbol{\phi} + \mathbf{Ax}_0 + \mathbf{BH}\boldsymbol{\phi} \tag{25}$$

However, $\mathbf{Ax}_0$ is a constant vector and is expressible as $\mathbf{Ax}_0\boldsymbol{\phi}_0$, or

$$\mathbf{Ax}_0 = \mathbf{Ax}_0\boldsymbol{\phi}_0 = [\mathbf{Ax}_0, \underbrace{\mathbf{0}, \dots, \mathbf{0}}_{(m-1)\text{ columns}}]\boldsymbol{\phi} \triangleq \mathbf{G}\boldsymbol{\phi} \tag{26}$$

Then substituting (24) into (23) yields

$$\mathbf{C} = \mathbf{ACP} + \mathbf{G} + \mathbf{BH} \triangleq \mathbf{ACP} + \mathbf{K} \tag{27}$$

where $\mathbf{K} \triangleq \mathbf{G} + \mathbf{BH}$ is an $n \times m$ matrix. The first column of $\mathbf{K}$ may be defined as $\mathbf{k}_0$; the second column, $\mathbf{k}_1$, etc., as we defined for $\mathbf{C}$ in (20), then eqn. (27) is expressible in terms of these vectors,

$$[\mathbf{c}_0 \ \mathbf{c}_1 \dots \mathbf{c}_{(m-1)}] = \mathbf{A}[\mathbf{c}_0 \ \mathbf{c}_1 \dots \mathbf{c}_{(m-1)}]\mathbf{P} + [\mathbf{k}_0 \ \mathbf{k}_1 \dots \mathbf{k}_{(m-1)}] \tag{27 a}$$

If we re-arrange $\mathbf{C}$ as a vector with nm elements by changing its first column into the first n components of the vector and then the second column, etc.; and re-arrange $\mathbf{K}$ in the same manner, finally we obtain an even simpler form in terms of a Kronecker product for (27)

$$\begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_{(m-1)} \end{bmatrix} = [\mathbf{A} \otimes \mathbf{P}'] \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_{(m-1)} \end{bmatrix} + \begin{bmatrix} \mathbf{k}_0 \\ \mathbf{k}_1 \\ \vdots \\ \mathbf{K}_{(m-1)} \end{bmatrix} \triangleq [\mathbf{A} \otimes \mathbf{P}']\mathbf{c} + \mathbf{k} \tag{28}$$

where $\mathbf{A} \otimes \mathbf{P}'$ is the Kronecker product defined as

$$\mathbf{A} \otimes \mathbf{P}' = \begin{bmatrix} p_{11}\mathbf{A} & p_{21}\mathbf{A} \dots p_{m1}\mathbf{A} \\ p_{12}\mathbf{A} & p_{22}\mathbf{A} \dots p_{m2}\mathbf{A} \\ \dots\dots\dots & \dots\dots\dots \\ \dots\dots\dots & \dots\dots\dots \\ p_{1m}\mathbf{A} & p_{2m}\mathbf{A} \dots p_{mm}\mathbf{A} \end{bmatrix} \tag{29}$$

The solution of $\mathbf{c}$ comes from (28) directly

$$\mathbf{c} = [\mathbf{I} - \mathbf{A} \otimes \mathbf{P}']^{-1}\mathbf{k} \tag{30}$$

After $\mathbf{C}$ is determined the solution $\mathbf{x}$ is obtained. The solution $\mathbf{x}$ is easily found by substituting $\mathbf{C}$ into (22), namely

$$\mathbf{x}(t) = \mathbf{CP}\boldsymbol{\phi}(t) + \mathbf{x}_0 \tag{31}$$

p.184

5.2. *Free system example*

For the free system for which

$$\dot{x} = -4x, \quad x(0) = 1$$

we should like to get the solution via the Walsh series approach.

First of all, we assume $\dot{x}$ has a Walsh expansion with $m = 4$ undetermined coefficients,

$$\dot{x}(t) = \sum_{i=0}^{\infty} c_i \phi_i(t)$$

$$= [c_0, c_1, c_2, c_3] \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix} = \mathbf{c}'\boldsymbol{\phi}(t)$$

ORIGINAL PAGE IS
OF POOR QUALITY

Next, the expression of $\mathbf{x}$ in terms of $\boldsymbol{\phi}$ is obtained through integration

$$x(t) = \mathbf{c}' \int_0^t \boldsymbol{\phi}(\lambda)\, d\lambda + x_0 = \mathbf{c}'\mathbf{P}\boldsymbol{\phi} + x_0$$

$$= \mathbf{c}' \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{8} & 0 \\ \frac{1}{4} & 0 & 0 & -\frac{1}{8} \\ \frac{1}{8} & 0 & 0 & 0 \\ 0 & \frac{1}{8} & 0 & 0 \end{bmatrix} \boldsymbol{\phi} + [1, 0, 0, 0]\boldsymbol{\phi}$$

Combining $x(t)$ and $\dot{x}(t)$ with the differential equation, we have

$$\mathbf{c}'\boldsymbol{\phi} = -4\mathbf{c}'\mathbf{P}\boldsymbol{\phi} + [-4, 0, 0, 0]\boldsymbol{\phi}$$

$$\mathbf{c} = [\mathbf{I} + 4\mathbf{P}']^{-1} \begin{bmatrix} -4 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 & 1 & \frac{1}{2} & 0 \\ -1 & 1 & 0 & \frac{1}{2} \\ -\frac{1}{2} & 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 0 & 1 \end{bmatrix} \begin{bmatrix} -4 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{80}{81} \\ -\frac{64}{81} \\ -\frac{40}{81} \\ -\frac{32}{81} \end{bmatrix}$$

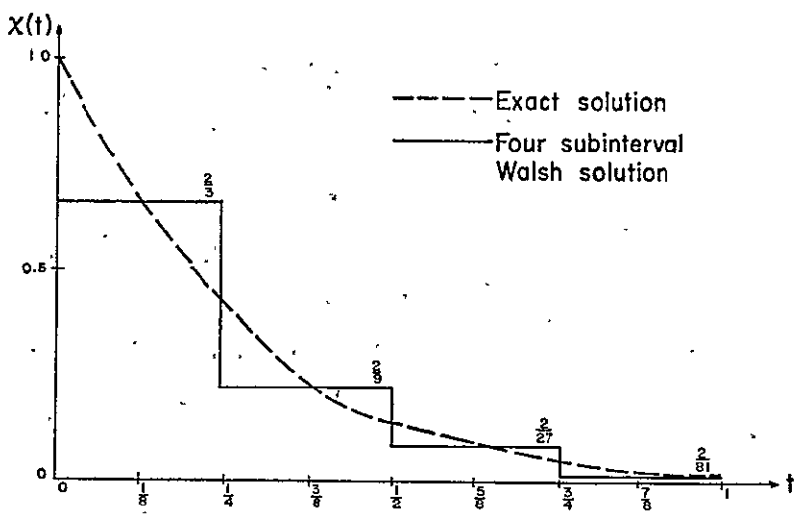


Figure 13. Walsh series solution of a free system.

p. 185

Therefore,

$$\dot{x}(t) \cong \mathbf{c}'\boldsymbol{\phi}(t) = -\tfrac{80}{81}\phi_0(t) - \tfrac{64}{81}\phi_1(t) - \tfrac{40}{81}\phi_2(t) - \tfrac{32}{81}\phi_3(t)$$

$$x(t) \cong \mathbf{c}'\mathbf{P}\boldsymbol{\phi}(t) + \mathbf{x}_0 = \tfrac{20}{81}\phi_0(t) + \tfrac{16}{81}\phi_1(t) + \tfrac{10}{81}\phi_2(t) + \tfrac{8}{81}\phi_3(t)$$

Figure 13 shows the result of the solution and the following table is the comparison between the Walsh approach and the exact solution.

t	1/8	3/8	5/8	7/5
$\mathbf{c}'\mathbf{P}\boldsymbol{\phi}(t)+x_0$	0·667	0·222	0·0741	0·0247
$\exp(-4t)$	0·693	0·223	0·105	0·03

(T 8)

5.3. *Forced system example*

Consider a set of differential equations to be given as

$$\begin{cases} \dot{x}_1 = -x_1 - 1{\cdot}8x_2 + 1{\cdot}8u, & x_1(0) = 0 \\ \dot{x}_2 = 5x_1 - x_2 \quad , & x_2(0) = 0 \end{cases}$$

$$u(t) = \text{unit step input}$$

It is required to find solutions for $x_1(t)$ and $x_2(t)$.

The **A** and **B** matrices of this system are as follows

$$\mathbf{A} = \begin{bmatrix} -1 & -1{\cdot}8 \\ 5 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1{\cdot}8 \\ 0 \end{bmatrix}$$

The rate vector $\dot{\mathbf{x}}(t)$ may be assumed as the Walsh series form.

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} c_{10}\phi_0 + c_{11}\phi_1 + c_{12}\phi_2 + \ldots \\ c_{20}\phi_0 + c_{21}\phi_1 + c_{22}\phi_2 + \ldots \end{bmatrix}$$

For easily showing the procedure let us take only 4 terms of Walsh series for each state variable, namely $m = 4$.

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} c_{10} & c_{11} & c_{12} & c_{13} \\ c_{20} & c_{21} & c_{22} & c_{23} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix} \triangleq \mathbf{C}\boldsymbol{\phi}$$

For unit step function, **H** may be written as

$$\mathbf{H} = [1, 0, 0, 0]$$

The **G** matrix is a null matrix for zero initial condition ; then **K** becomes

$$\mathbf{K} = \mathbf{B}\mathbf{H} + \mathbf{G} = \begin{bmatrix} 1{\cdot}8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = [\mathbf{k}_0, \mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3]$$

Substituting **A**, **P**′ and **K** into (28), we have

$$\begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{bmatrix} = \begin{bmatrix} \tfrac{1}{2}\mathbf{A} & \tfrac{1}{4}\mathbf{A} & \tfrac{1}{8}\mathbf{A} & \mathbf{0} \\ -\tfrac{1}{4}\mathbf{A} & \mathbf{0} & \mathbf{0} & \tfrac{1}{8}\mathbf{A} \\ -\tfrac{1}{8}\mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\tfrac{1}{8}\mathbf{A} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{k}_0 \\ \mathbf{k}_1 \\ \mathbf{k}_2 \\ \mathbf{k}_3 \end{bmatrix}$$

p.186

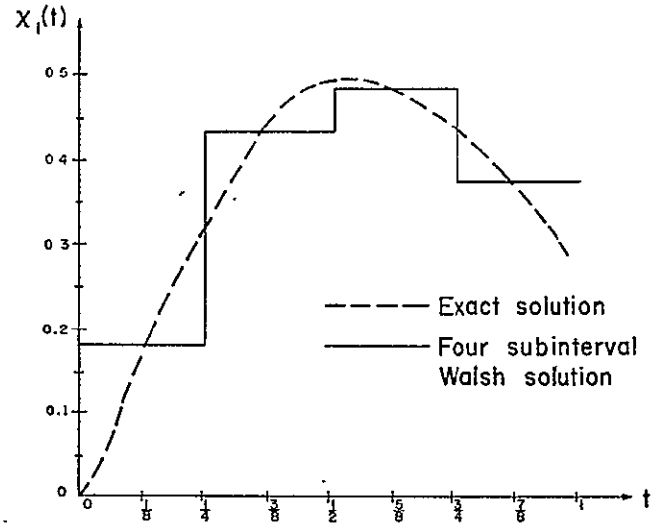


Figure 14. Walsh series solution of a forced system.

ORIGINAL PAGE IS OF POOR QUALITY

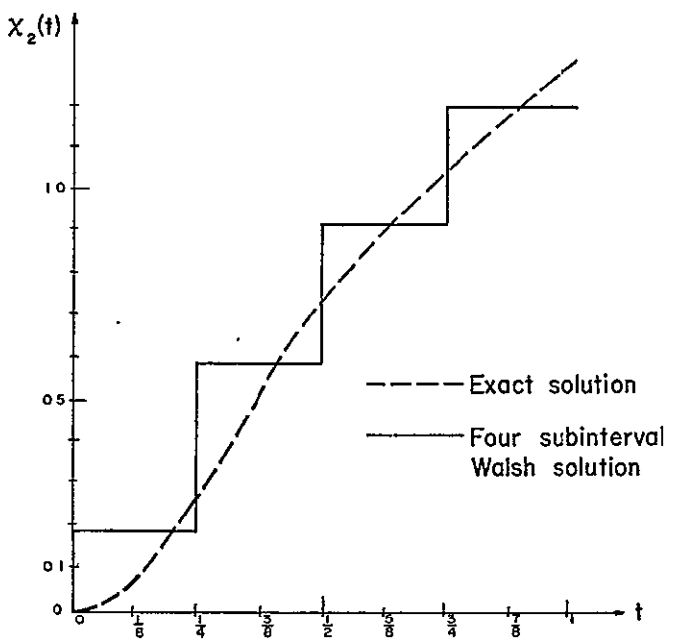


Figure 15. Walsh series solution for $X_2(t)$.

p. 187

Solving for **c** and multiplying **C** with **P**, we have

$$\mathbf{CP}=\begin{bmatrix} 0{\cdot}3717155 & -0{\cdot}0637155 & -0{\cdot}0381155 & -0{\cdot}0011377 \\ 0{\cdot}6240888 & -0{\cdot}3640888 & -0{\cdot}1543111 & -0{\cdot}0056888 \end{bmatrix}$$

Finally, the solution of x_1 and x_2 are :

$$x_1=0{\cdot}3717155\phi_0-0{\cdot}0637155\phi_1-0{\cdot}0381155\phi_2-0{\cdot}0011377\phi_3$$

$$x_2=0{\cdot}6240888\phi_0-0{\cdot}3640888\phi_1-0{\cdot}1543111\phi_2-0{\cdot}0056888\phi_3$$

The conventional solution and the Walsh series solution of x_1 and x_2 are drawn in Figs. 14 and Fig. 15, respectively.

5·4. *A circuit example* (Huelsman 1972)

Let us consider the circuit shown in Fig. 16. The initial conditions are zero and unit step function is applied. The governing differential equation is :

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}=\begin{bmatrix} -G_1/C_1 & 0 & 0 & -1/C_1 & 0 & -1/C_1 \\ 0 & -G_2/C_2 & 0 & 0 & -1/C_2 & 1/C_2 \\ 0 & 0 & 0 & 0 & 0 & 1/C_3 \\ 1/L_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/L_2 & 0 & 0 & 0 & 0 \\ 1/L_1 & -1/L_3 & -1/L_3 & 0 & 0 & -R_3/L_3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix}+\begin{bmatrix} 1/C_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}u$$

$$=\mathbf{A}_1\mathbf{x}+\mathbf{B}_1u$$

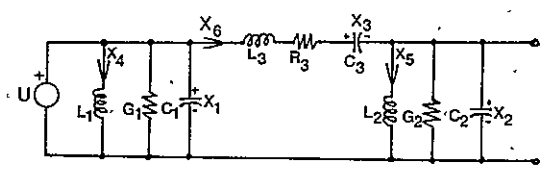


Figure 16. A circuit of sixth order.

If each parameter of the circuit is assumed to have a numerical value of 1, $\mathbf{A}_1$ and $\mathbf{B}_1$ become

$$\mathbf{A}_1=\begin{bmatrix} -1 & 0 & 0 & -1 & 0 & -1 \\ 0 & -1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & -1 \end{bmatrix},\quad \mathbf{B}_1=\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Suppose we are interested in applying the Walsh functions to obtain the solution for each state variable in the interval $0\leqslant t\leqslant 12{\cdot}8$ sec. The first step is to normalize the time scale by letting,

$$\tau=t/12{\cdot}8\triangleq t/t_f,\quad dt=t_f\cdot d\tau$$

p. 188.

Then

$$\frac{dx}{d\tau}=\frac{dx}{dt}\cdot\frac{dt}{d\tau}=t_f\cdot\mathbf{A}_1\mathbf{x}+t_f\mathbf{B}_1u\triangleq\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{u}$$

Next, expanding $u(t)$ into a Walsh series with m terms and applying (19), (24), and (27), we have

$$\dot{\mathbf{x}}=\mathbf{C}\boldsymbol{\phi}$$

$$\mathbf{u}=\mathbf{H}\boldsymbol{\phi}$$

$$\mathbf{C}=\mathbf{ACP}+\mathbf{K}$$

ORIGINAL PAGE IS OF POOR QUALITY

where $\mathbf{K}$ is related to $\mathbf{H}$ through (27).

A question arises here. 'How many terms should we use?' If we wish to obtain a quick answer and to sacrifice accuracy, we can use small number for m, say, let $m\leqslant 8$. On the other hand, if we want accurate answer and do not care about computation time, we can use very large value for m, say, $m\geqslant 128$.

Let us investigate both cases. Try $m=128$ first. The matrix $\mathbf{C}$ will contain 768 elements for $m=128$, $n=6$. If we use (30) directly, some difficulties might occur in obtaining the inverse of a square matrix of 768×768. But, recall that the $\mathbf{P}$ matrix has the special form as shown in (16). We can take advantage of the properties of $\mathbf{P}$ and use Gauss' method to eliminate $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{m-1}$, then calculate $\mathbf{c}_0$, finally calculate $\mathbf{c}_1$ through $\mathbf{c}_{m-1}$ via substitution. For detailed explanation, the reader is referred to Appendix.

The evaluation of $\mathbf{x}(t)$ via (31) should cause little trouble. Shank's (1969) algorithm may be applied here to speed the computation process. Using UNIVAC 1108 computer, we obtained 128 points for each of these state variables. The execution time including normalization of $\mathbf{A}_1$ and $\mathbf{B}_1$, transformation of $u(t)$ into Walsh series, and inverse transformation of Walsh coefficients for $\mathbf{x}$ into time function is 617 millisec. The waveform of x_2 is drawn in Fig. 17 which is checked with Huelsman's Fig. 6–7.7; his result is obtained by using the Runge–Kutta method.

Next, let us try $m=8$. $\mathbf{C}$ contains 48 elements. We may use (30) to get the inverse of a square matrix 48×48. Or, we may apply the algorithm in

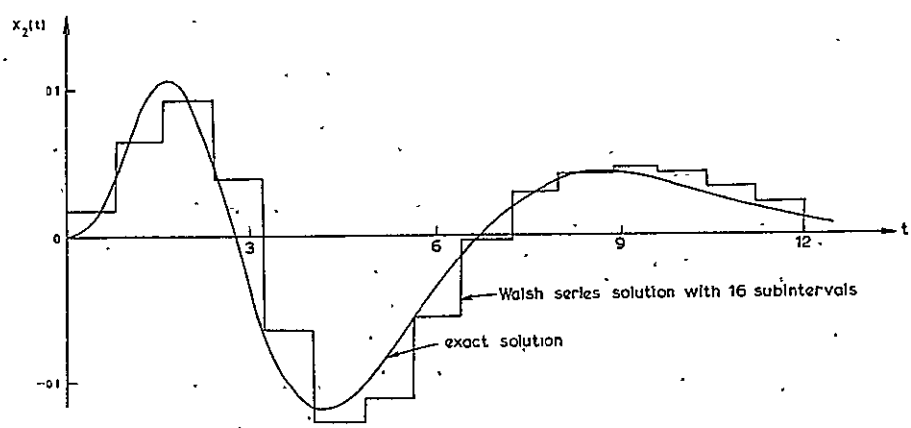


Figure 17. Walsh solution compared with the Runge–Kutta solution.

p. 189

Appendix. The following values are obtained with the Gauss elimination algorithm.

j	1	2	3	4
$C_{2,j}$	0·00941	−0·24960	0·48335	−0·04008
j	5	6	7	8
$C_{2,j}$	−0·00091	−0·21712	0·44308	0·04254

After multiplying **C** with **P** and taking the inverse Walsh transform, we have

t	0·8	2·4	4·0	5·6
$x_2(t)$	0·02942	0·05480	−0·03593	−0·12136
t	7·2	8·8	10·4	12·0
$x_2(t)$	−0·03264	0·06517	0·04748	0·01442

The solution $x_2(t)$ is drawn in Fig. 17 as stairs for comparison. The other values of $x_1(t)$, $x_3(t)$, $x_4(t)$, $x_5(t)$, $x_6(t)$ have been obtained also, but they are omitted here.

For comparison, we have tried to solve this problem with Runge–Kutta's method by letting subinterval of integration equal to 1·6 sec. Then, we obtain,

t	0	1·6	3·2	4·8	6·4
$x_2(t)$	0	−0·13653	−1·0352	−1·5209	2·5396
t	8·0	9·6	11·2	12·8	
$x_2(t)$	10·499	12·778	19·543	107·79	

In this case, Runge–Kutta's method fails completely as shown in Fig. 18 because of numerical instability.

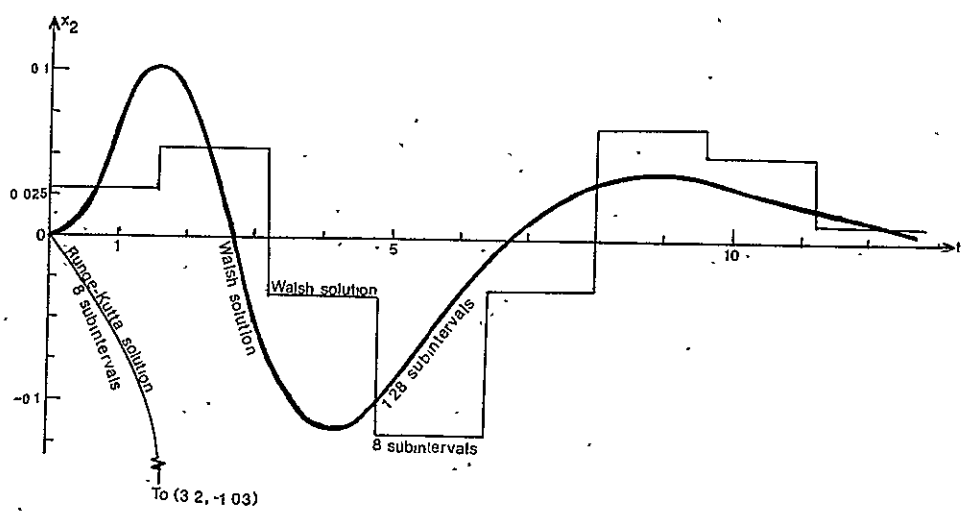


Figure 18. Runge–Kutta's method fails.

p.190

6. Conclusion

A simple procedure for solving state equations of linear systems via Walsh series is formulated. It involves (1) to assume the rate variables as Walsh series whose coefficients are to be determined; (2) to use an operational matrix to perform the integration; (3) to develop a new Kronecker product formula for the rate variable and (4) to determine the state variable from the rate variable so obtained and the given set of initial conditions.

Compared with Corrington's procedure, the new technique has several advantages: (1) while we only use an operational matrix he must use several tables. The tables so far he derived are only for solving low order differential equations. In other words, if the order is relatively higher, additional tables are not available. (2) We derive an exact formula which involves the Kronecker product only.

The disadvantage of our approach is that our answer is slightly less accurate than Corrington's. However, this shortcoming can be easily overcome by using more subdivisions and digital computation.

Appendix

Algorithm for solving **C** *via the Kronecker product formula*

In this appendix we derive a recursive algorithm to solve **C** from (28) instead of inversing $[\mathbf{I}-\mathbf{A}\otimes\mathbf{P}']$ directly. Let us illustrate the procedures for $m=2^3=8$.

Step 1

Equation (28) may be rewritten explicitly for $m=8$.

$$\begin{bmatrix}\mathbf{c}_0\\ \mathbf{c}_1\\ \mathbf{c}_2\\ \mathbf{c}_3\\ \mathbf{c}_4\\ \mathbf{c}_5\\ \mathbf{c}_6\\ \mathbf{c}_7\end{bmatrix}=\begin{bmatrix}\frac{1}{2}\mathbf{A} & \frac{1}{4}\mathbf{A} & \frac{1}{8}\mathbf{A} & 0 & \frac{1}{16}\mathbf{A} & 0 & 0 & 0\\ -\frac{1}{4}\mathbf{A} & 0 & 0 & \frac{1}{8}\mathbf{A} & 0 & \frac{1}{16}\mathbf{A} & 0 & 0\\ -\frac{1}{8}\mathbf{A} & 0 & 0 & 0 & 0 & 0 & \frac{1}{16}\mathbf{A} & 0\\ 0 & -\frac{1}{8}\mathbf{A} & 0 & 0 & 0 & 0 & 0 & \frac{1}{16}\mathbf{A}\\ -\frac{1}{16}\mathbf{A} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & -\frac{1}{16}\mathbf{A} & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & -\frac{1}{16}\mathbf{A} & 0 & 0 & 0 & 0 & 0\\ 0 & & 0 & -\frac{1}{16}\mathbf{A} & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{c}_0\\ \mathbf{c}_1\\ \mathbf{c}_2\\ \mathbf{c}_3\\ \mathbf{c}_4\\ \mathbf{c}_5\\ \mathbf{c}_6\\ \mathbf{c}_7\end{bmatrix}+\begin{bmatrix}\mathbf{k}_0\\ \mathbf{k}_1\\ \mathbf{k}_2\\ \mathbf{k}_3\\ \mathbf{k}_4\\ \mathbf{k}_5\\ \mathbf{k}_6\\ \mathbf{k}_7\end{bmatrix} \tag{A 1}$$

It is obvious that $\mathbf{c}_4$, $\mathbf{c}_5$, $\mathbf{c}_6$ and $\mathbf{c}_7$ may be expressed in terms of $\mathbf{c}_0$, $\mathbf{c}_1$, $\mathbf{c}_2$ and $\mathbf{c}_3$.

$$\begin{bmatrix}\mathbf{c}_4\\ \mathbf{c}_5\\ \mathbf{c}_6\\ \mathbf{c}_7\end{bmatrix}=\begin{bmatrix}-\frac{1}{16}\mathbf{A} & 0 & 0 & 0\\ 0 & -\frac{1}{16}\mathbf{A} & 0 & 0\\ 0 & 0 & -\frac{1}{16}\mathbf{A} & 0\\ 0 & 0 & 0 & -\frac{1}{16}\mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{c}_0\\ \mathbf{c}_1\\ \mathbf{c}_2\\ \mathbf{c}_3\end{bmatrix}+\begin{bmatrix}k_0\\ k_1\\ k_2\\ k_3\end{bmatrix}. \tag{A 2}$$

In order to keep consistent notation, we may define

$$\mathbf{G}_3 \triangleq \mathbf{I} \tag{A 3}$$

$$\mathbf{R}_3 \triangleq -(1/2^{3+1})\mathbf{G}_3^{-1}\mathbf{A} \tag{A 4}$$

then

$$\mathbf{c}_{i+4}=\mathbf{R}_3\mathbf{c}_i+k_{i+4}, \quad i=0, 1, 2, 3 \tag{A 5}$$

p. 191

Substituting (A 2) into (A 1), we have

$$\begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\mathbf{A} & -\frac{1}{256}\mathbf{A}^2 & \frac{1}{4}\mathbf{A} & \frac{1}{8}\mathbf{A} & 0 \\ & -\frac{1}{4}\mathbf{A} & -\frac{1}{256}\mathbf{A}^2 & 0 & \frac{1}{8}\mathbf{A} \\ & -\frac{1}{8}\mathbf{A} & 0 & -\frac{1}{256}\mathbf{A}^2 & 0 \\ & 0 & -\frac{1}{8}\mathbf{A} & 0 & -\frac{1}{256}\mathbf{A}^2 \end{bmatrix} \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{k}_0 + \frac{1}{16}\mathbf{A}\mathbf{k}_4 \\ \mathbf{k}_1 + \frac{1}{16}\mathbf{A}\mathbf{k}_5 \\ \mathbf{k}_2 + \frac{1}{16}\mathbf{A}\mathbf{k}_6 \\ \mathbf{k}_3 + \frac{1}{16}\mathbf{A}\mathbf{k}_7 \end{bmatrix} \tag{A 6}$$

Now we may define the new diagonal matrix as $\mathbf{F}_3$ and the new inputs as $\mathbf{k}_{i,\,\mathrm{III}}$.

$$\mathbf{F}_3 \triangleq -\frac{1}{256}\mathbf{A}^2 = \frac{1}{2^{3+1}}\,\mathbf{A}\mathbf{R}_3 \tag{A 7}$$

$$\mathbf{k}_{i,\,\mathrm{III}} \triangleq \mathbf{k}_i + \frac{1}{2^{3+1}}\,\mathbf{A}\mathbf{G}_3^{-1}\mathbf{k}_{i+4}, \quad i = 0, 1, 2, 3 \tag{A 8}$$

Step 2

Eliminating $\mathbf{c}_2$, $\mathbf{c}_3$ from the lower half of (A 6), we have

$$\begin{bmatrix} \mathbf{c}_2 \\ \mathbf{c}_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{8}(\mathbf{I}+\frac{1}{256}\mathbf{A}^2)^{-1}\mathbf{A} & 0 \\ 0 & -\frac{1}{8}(\mathbf{I}+\frac{1}{256}\mathbf{A}^2)^{-1}\mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{bmatrix} + (\mathbf{I}+\tfrac{1}{256}\mathbf{A}^2)^{-1} \begin{bmatrix} \mathbf{k}_{2,\,\mathrm{III}} \\ \mathbf{k}_{3,\,\mathrm{III}} \end{bmatrix} \tag{A 9}$$

This may be put into recursive form by defining

$$\mathbf{G}_2 \triangleq \mathbf{I} - \mathbf{F}_3 = \mathbf{I} + \tfrac{1}{256}\mathbf{A}^2 \tag{A 10}$$

$$\mathbf{R}_2 \triangleq -\frac{1}{2^{2+1}}\,\mathbf{G}_2^{-1}\mathbf{A} = -\tfrac{1}{8}(\mathbf{I}+\tfrac{1}{256}\mathbf{A}^2)^{-1}\mathbf{A} \tag{A 11}$$

then,

$$\mathbf{c}_{i+2} = \mathbf{R}_2\mathbf{c}_i + \mathbf{G}_2^{-1}\mathbf{k}_{(i+2),\,\mathrm{III}}, \quad i = 0, 1 \tag{A 12}$$

Substituting (A 9), (A 12) into (A 6), we have

$$\begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\mathbf{A} + \mathbf{F}_3 - \frac{1}{64}\mathbf{A}(\mathbf{I}+\frac{1}{256}\mathbf{A}^2)^{-1}\mathbf{A} & \frac{1}{4}\mathbf{A} \\ -\frac{1}{4}\mathbf{A} & \mathbf{F}_3 - \frac{1}{64}\mathbf{A}(\mathbf{I}+\frac{1}{256}\mathbf{A}^2)^{-1}\mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{bmatrix} + \begin{bmatrix} \mathbf{k}_{0,\,\mathrm{III}} + \frac{1}{8}\mathbf{A}\mathbf{G}_2^{-1}\mathbf{k}_{2,\,\mathrm{IV}} \\ \mathbf{k}_{1,\,\mathrm{III}} + \frac{1}{8}\mathbf{A}\mathbf{G}_2^{-1}\mathbf{k}_{3,\,\mathrm{IV}} \end{bmatrix} \tag{A 13}$$

For recursive operation, we should define again

$$\mathbf{F}_2 \triangleq \mathbf{F}_3 + \frac{1}{2^{2+1}}\,\mathbf{A}\mathbf{R}_2 = -\tfrac{1}{256}\mathbf{A}^2 - \tfrac{1}{64}\mathbf{A}(\mathbf{I}+\tfrac{1}{256}\mathbf{A}^2)^{-1}\mathbf{A} \tag{A 14}$$

$$\mathbf{k}_{i,\,\mathrm{II}} \triangleq \mathbf{k}_{i,\,\mathrm{III}} + \frac{1}{2^{2+1}}\,\mathbf{A}\mathbf{G}_2^{-1}\mathbf{k}_{i+2,\,\mathrm{III}}, \quad i = 0, 1 \tag{A 15}$$

p. 192

C-3

Step 3

Eliminating $\mathbf{c}_1$ from (A 13), (A 14), we have

$$\mathbf{c}_1 = -\tfrac{1}{4}(\mathbf{I}-\mathbf{F}_2)^{-1}\mathbf{A}\mathbf{c}_0 + (\mathbf{I}-\mathbf{F}_2)^{-1}\mathbf{k}_{i,\,\mathrm{II}} \triangleq \mathbf{R}_1\mathbf{c}_0 + \mathbf{G}_1^{-1}\mathbf{k}_{i,\,\mathrm{II}} \tag{A 16}$$

where $\mathbf{G}_1$, $\mathbf{R}_1$ are defined in the same form as $\mathbf{G}_2$, $\mathbf{R}_2$, $\mathbf{G}_3$, $\mathbf{R}_3$.

$$\mathbf{G}_1 \triangleq \mathbf{I}-\mathbf{F}_2 \tag{A 17}$$

$$\mathbf{R}_1 \triangleq -\frac{1}{2^{2+1}}\mathbf{G}_1^{-1}\mathbf{A} \tag{A 18}$$

Substituting (A 16), (A 14) into the upper half of (A 13), we have

$$\mathbf{c}_0 = [\tfrac{1}{2}\mathbf{A}+\mathbf{F}_2+\tfrac{1}{4}\mathbf{A}\mathbf{R}_1]\mathbf{c}_0 + \mathbf{k}_{0,\,\mathrm{II}} + \tfrac{1}{4}\mathbf{A}\mathbf{G}_1^{-1}\mathbf{k}_{i,\,\mathrm{II}}$$

Therefore

$$\mathbf{c}_0 = \mathbf{G}_0^{-1}\mathbf{k}_{0,\,\mathrm{I}} \tag{A 19}$$

where

$$\mathbf{G}_0 \triangleq \mathbf{I}-\tfrac{1}{2}\mathbf{A}-\mathbf{F}_1 \tag{A 20}$$

$$\mathbf{F}_1 \triangleq \mathbf{F}_2 + \frac{1}{2^{1+1}}\mathbf{A}\mathbf{R}_1 \tag{A 21}$$

$$\mathbf{k}_{0,\,\mathrm{I}} \triangleq \mathbf{k}_{0,\,\mathrm{II}} + \frac{1}{2^{1+1}}\mathbf{A}\mathbf{G}_1^{-1}\mathbf{k}_{1,\,\mathrm{II}} \tag{A 22}$$

Step 4

$\mathbf{c}_1$ is obtained by substituting $\mathbf{c}_0$ into (A 16). Similarly, $\mathbf{c}_2$, $\mathbf{c}_3$ and $\mathbf{c}_4$, $\mathbf{c}_5$, $\mathbf{c}_6$, $\mathbf{c}_7$ are obtained by applying (A 12) and (A 5), respectively. This completes the derivation of the algorithm for solving $\mathbf{C}$ from (28), with $m=8$.

In general, for $m=2^\alpha$, α is any positive integer, we start with

$$\mathbf{G}_\alpha = \mathbf{I}, \quad \mathbf{R}_\alpha = -\frac{1}{2^{\alpha+1}}\mathbf{A}, \quad \mathbf{F}_\alpha = \frac{1}{2^{\alpha+1}}\mathbf{A}\mathbf{R}_\alpha \tag{A 23}$$

Then, we can calculate $\mathbf{R}_\beta$ and $\mathbf{k}_{i,\,\beta}$, $i=0, 1, \ldots, 2^{\beta-1}$, $0<\beta\leqslant\alpha-1$ from the following recursive formulae.

$$\left.\begin{aligned} \mathbf{G}_\beta &= \mathbf{I}-\mathbf{F}_{\beta+1} \\ \mathbf{R}_\beta &= -2^{-\beta-1}\mathbf{G}_\beta^{-1}\mathbf{A} \\ \mathbf{F}_\beta &= \mathbf{F}_{\beta+1} + 2^{-\beta-1}\mathbf{A}\mathbf{R}_\beta \\ \mathbf{k}_{i,\,\beta} &= \mathbf{k}_{i,\,\beta+1} + 2^{-\beta-1}\mathbf{A}\mathbf{G}_\beta^{-1}\mathbf{k}_{i+2,\,\beta+1} \end{aligned}\right\} \tag{A 24}$$

Then we obtain $\mathbf{c}_0$

$$\mathbf{c}_0 = \mathbf{G}_0^{-1}\mathbf{k}_{0,\,\mathrm{I}}, \quad \mathbf{G}_0 = \mathbf{I}-\tfrac{1}{2}\mathbf{A}-\mathbf{F}_1 \tag{A 25}$$

All the other vectors $\mathbf{c}_i$, $i=0, 1, \ldots, (m-1)$ are found by substituting with

$$\mathbf{c}_{i+2^{\beta-1}} = \mathbf{R}_\beta \mathbf{c}_i + \mathbf{G}_\beta^{-1}\mathbf{k}_{(i+2^{\beta-1}),\,\beta} \tag{A 26}$$

In the above algorithm, we always work with matrices $(\mathbf{G}, \mathbf{F}, \mathbf{R})$ of $n\times n$. Therefore there is no need to operate with larger matrices. For instance, if we wish to apply Walsh functions to solve a sixth-order differential equation, we need only work with matrices 6×6. The calculation of matrix inverse of $[\mathbf{I}-\mathbf{A}\otimes\mathbf{P}']$, which is of 768×768, is avoided. Therefore, we have saved computing time, storage, and have reduced round-off errors significantly.

References

Ahmed, N., 1972, *Proc. I.E.E.E. Conf. on Decision and Control*, pp. 495–498.

Andrews, H. C., 1971, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs.), pp. 26–32 ; 1972, *Proc. of I.E.E.E. Conf. on Decision and Control*, pp. 492–494.

Carl, J. W., and Kabrisky, M., 1971, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 203–209.

Chen, C., 1972, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 64–67.

Clark, M. T., Swanson, J. E., and Sanders, J. A., 1972, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 169–172.

Corrington, M. S., 1973, *I.E.E.E. Trans. Circuit Theory*, **20**, 470.

Dinh, C. T. L., Goulet, R., and Van Houtte, N., 1972, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 362–368.

Dust, D. I., 1972, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 82–88.

Gibbs, J. E., and Gebbie, H. A., 1969, *Nature, Lond.*, **224**, 1012.

Gibbs, J. E., and Millard, M. J., 1969, DES Report No. 2 (Teddington, Middlesex : National Phys. Lab.).

Harmuth, H. F., 1969, *I.E.E.E. Spectrum*, Vol. 6, No. 11, p. 82 ; 1970, *Proc. Walsh Function Symp.* (Washington, D.C. : Naval Res. Labs), pp. 248–259 ; 1972, *Transmission of Information by Orthogonal Functions* (New York : Springer-Verlag) ; 1972, *Proc. Walsh Function Symp.* (Washington, D.C. : Naval Res. Labs), pp. 23–29.

Huelsman, L. P., 1972, *Basic Circuit Theory* (New Jersey, Englewood : Prentice-Hall), pp. 363–365.

Ito, T., 1970, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 128–137.

Jones, H. E., 1972, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 110–113.

Kennedy, J. D., 1971, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs); pp. 7–10.

Kuhn, B. G., Morey, K. H., and Smith, W. B., 1963, *I.E.E.E. Trans. Space Electron. Telemetry*, **9**, 63.

Lee, J. D., 1970, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 26–35.

Milne, P. J., Ahmed, N., Gallagher, R. R., and Harris, S. G., 1972, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 149–153.

Paley, R. E. A. C., 1932, *Proc. Lond. Math. Soc.*, 2nd series, **34**, 241.

Pearlman, J., 1970, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 65–69.

Picher, F., 1970 a, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 175–182 ; 1970 b, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 17–22.

Pratt, W. K., Welch, L. R., and Chen, W., 1972, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 229–234.

Rademacher, 1922, *Math. Annln*, **87**, 712.

Shanks, J. L., 1969, *I.E.E.E. Trans. Comput.*, Vol. C-18, No. 5, p. 457.

Thomas, C. W., and Welch, A. J., 1972, *Proc. Walsh Function Symp.* (Washington, D.C.: Naval Res. Labs), pp. 154–158.

Walsh, J. L., 1923, *Am. J. Math.*, **45**, 5.

p. 194

Project

(A) Project Title: Time-Domain Synthesis via Walsh Functions

(B) Project Abstract:

The project deals with the application of Walsh functions to the time-domain-synthesis problem, i.e., the determination of a suitable internal structure for a system from its prescribed external (input/output) behavior. The method is based on repeated integration, and a new operational matrix, which relates Walsh functions and their integrations, is defined. Examples of transfer-function and state-equation synthesis are used to illustrate the technique.

(C) Publication: Proceedings of IEE; 122, No. 5, 565 (1975).

(D) Year: 1975

(E) Department: Electrical Engineering

(F) Student Name: C. H. Hsiao

(G) Faculty Advisor: Professor C. F. Chen

Time-domain synthesis via Walsh functions

Prof. C.F. Chen, M.S., Ph.D., and C.H. Hsiao, M.S.

Indexing terms: Identification, Systems theory, Time-domain synthesis, Walsh functions

Abstract

The paper deals with the application of Walsh functions to the time-domain-synthesis problem, i.e., the determination of a suitable internal structure for a system from its prescribed external (input/output) behaviour. The method is based on repeated integration, and a new operational matrix, which relates Walsh functions and their integrations, is defined. Examples of transfer-function and state-equation synthesis are used to illustrate the technique.

List of symbols

$\phi_n(t)$ = Walsh function
$r_q(t)$ = Rademacher function
$b_q b_{q-1}$ = binary expression of n
$f(t)$ = an arbitrary function
c_n = Walsh-series coefficient
f_k = discrete value of $f(t)$ at k
W = Walsh matrix
$P(\sigma \times \sigma)$ = operational matrix with $\sigma \times \sigma$ dimensions
Φ_σ = truncated Walsh vector with σ elements
m = order of operational matrix
$y(t)$ = output variable
$u(t)$ = input variable
h_i = input variable Walsh-series coefficients
c = output-variable Walsh-coefficient vector
h = input-variable Walsh-coefficient vector
Φ' = rectangular matrix $2n \times m$
a = coefficient vector to be determined
b = coefficient vector to be determined
$Y(s)$ = Laplace transform of $y(t)$
$U(s)$ = Laplace transform of $u(t)$
C = $[c_1 c_2, \ldots, c_n]'$
$\bar{x}_1$ = average value of x_1
$\bar{x}_2$ = average value of x_2

ORIGINAL PAGE IS
OF POOR QUALITY

1 Introduction

Finding the simplest system that will realise a prescribed input/output behaviour has been a fundamental problem in systems theory since Guillemin's time.[1] Only a few techniques are generally known, and they, including the one developed by Guillemin himself, are based on repeated differentiation and contain all the inherent disadvantages involved therein.[2,3]

In searching for a new direction, the question naturally arises: why not use a principle based on integration instead? Walsh-function theory,[4] when used in the solving of differential equations, is such a principle. Corrington[5] used this approach on the solution problem. This paper proceeds in the inverse direction, attacking the time-domain-synthesis problem via Walsh functions.

Let us first briefly review Walsh functions.

2 Rademacher and Walsh functions

In recent years, Walsh-function theory has been innovated and applied to various fields in engineering and science.[6,7] The original papers were published in 1922 and 1923 by Rademacher[8] and Walsh,[9] respectively.

Rademacher's function is a set of square waves of unit height with periods equal to $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, 2^{(1-k)}$, respectively. The first four square waves are shown in Fig. 1. It is noted that the set involves only odd functions, and therefore it is not complete. In 1923, Walsh independently developed a complete set known as Walsh functions. The set of Walsh functions and the set of Rademacher functions have the following relations:

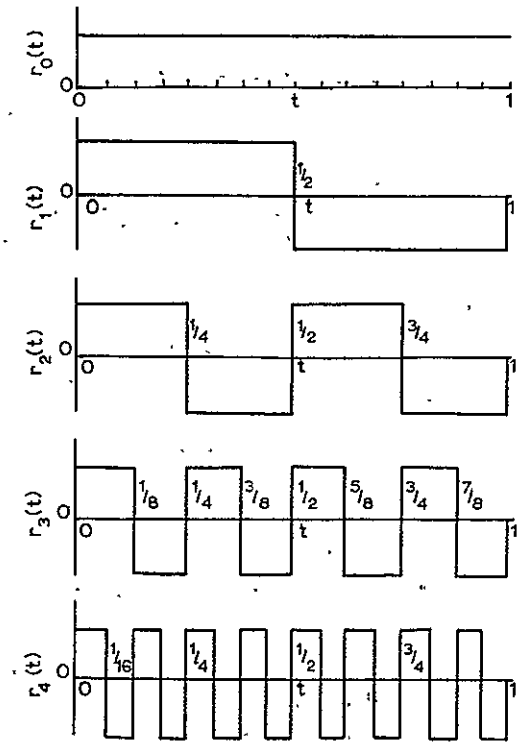


Fig. 1
Rademacher functions

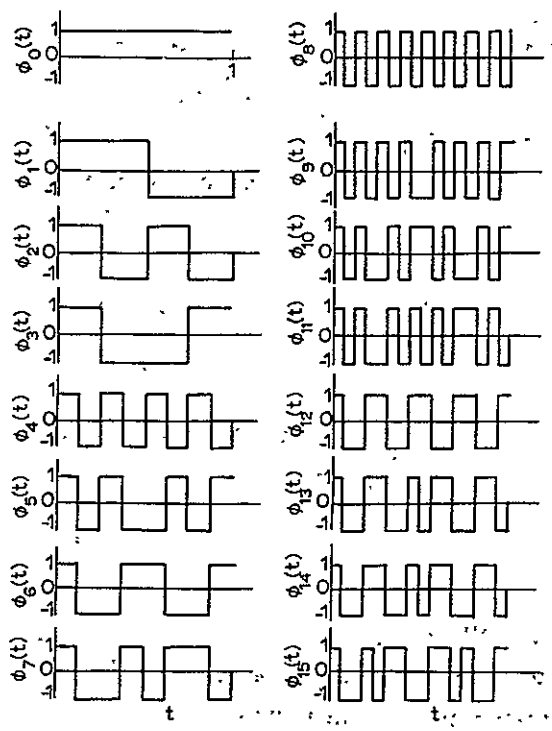


Fig. 2
Walsh functions

Paper 7389 C, first received 27th September and in revised form 17th December 1974

Prof. Chen is, and Mr Hsiao was formerly, with the Department of Electrical Engineering, University of Houston, Houston, Texas 77004, USA. Mr. Hsiao is now with Cheng Kung University and Chung Shan Institute, Taiwan

p.196

$$\left.\begin{aligned}
\phi_0(t) &= r_0(t)\\
\phi_1(t) &= r_1(t)\\
\phi_2(t) &= \{r_2(t)\}^1\,\{r_1(t)\}^0\\
\phi_3(t) &= \{r_2(t)\}^1\,\{r_1(t)\}^1\\
\phi_4(t) &= \{r_3(t)\}^1\,\{r_2(t)\}^0\,\{r_1(t)\}^0\\
\phi_5(t) &= \{r_3(t)\}^1\,\{r_2(t)\}^0\,\{r_1(t)\}^1\\
\phi_6(t) &= \{r_3(t)\}^1\,\{r_2(t)\}^1\,\{r_1(t)\}^0\\
\phi_7(t) &= \{r_3(t)\}^1\,\{r_2(t)\}^1\,\{r_1(t)\}^1\\
&\vdots
\end{aligned}\right\} \tag{1}$$

$$\phi_n(t) = \{r_q(t)\}^{b_q}\,\{r_{q-1}(t)\}^{b_{q-1}}\,\{r_{q-2}(t)\}^{b_{q-2}}\ldots \tag{2}$$

where

$$q = [\log_2 n] + 1 \tag{3}$$

in which $[\cdot]$ means 'taking the greatest integer of'. Therefore,

$$n = b_q 2^{q-1} + b_{q-1} 2^{q-2} + \ldots b_1 2^0 \tag{4}$$

where $b_q b_{q-1} \ldots b_1$ is the binary expression of n.

Therefore, if a particular Walsh function $\phi_n(t)$ is given and its Rademacher function components are required, we simply change n into binary form and then substitute in eqn. 2. For example, to find the Rademacher-function components of Walsh function $\phi_{10}(t)$,

$$\begin{aligned}
q &= [\log_2 10] + 1\\
&= 4\\
10 &= \underset{\substack{\uparrow\\ b_4}}{1\times 2^3} + \underset{\substack{\uparrow\\ b_3}}{0\times 2^2} + \underset{\substack{\uparrow\\ b_2}}{1\times 2^1} + \underset{\substack{\uparrow\\ b_1}}{0\times 2^0}
\end{aligned}$$

and therefore

$$\phi_{10}(t) = \{r_4(t)\}^1\,\{r_3(t)\}^0\,\{r_2(t)\}^1\,\{r_1(t)\}^0$$

The Walsh functions, like the Rademacher functions, are easy to draw.

3 Walsh-coefficient determination

A function $f(t)$ that is absolutely integrable in $[0, 1)$ may be expanded into a Walsh series:

$$f(t) = c_0\phi_0(t) + c_1\phi_1(t) + \ldots + c_n\phi_n(t) + \ldots \tag{5}$$

where c_n are coefficients of the Walsh series of $f(t)$.

It is desirable to determine the coefficients c_n such that the integral-square error satisfies the following relation:

$$\lim_{N\to\infty} \int_0^1 |f(t) - \sum_{n=0}^{N} c_n\phi_n(t)|^2 dt = 0 \tag{6}$$

Multiplying by $\phi_n(t)$ on both sides of eqn. 5 and then integrating each term from 0 to 1, we obtain

$$c_n = \int_0^1 \phi_n(t) f(t)\, dt \tag{7}$$

This is because of the orthonormal property of Walsh functions.

Let us illustrate the Walsh-series expansion by the following simple ramp-function example:

$$f(t) = t$$

Substituting $f(t)$ into eqn. 7 and taking only four terms, we obtain

$$c_0 = \int_0^1 \phi_0(t)\, t\, dt = \tfrac{1}{2}$$

$$c_1 = \int_0^1 \phi_1(t)\, t\, dt = -\tfrac{1}{4}$$

$$c_2 = \int_0^1 \phi_2(t)\, t\, dt = -\tfrac{1}{8}$$

$$c_3 = \int_0^1 \phi_3(t)\, t\, dt = 0$$

After substituting these coefficients into eqn. 5, we have

$$t = \tfrac{1}{2}\phi_0(t) - \tfrac{1}{4}\phi_1(t) - \tfrac{1}{8}\phi_2(t) + 0\,\phi_3(t)$$

Which is the four-term Walsh-series expansion of the ramp function.

3.1 Discrete formula

If the given function is not in its analytic form but in tabulated-data or graphical form, and if its Walsh-series expansion is desired, we would modify eqns. 5 and 7 into discrete forms:

$$f_k = \sum_{n=0}^{m-1} \phi_{kn} c_n \qquad k = 0, 1, 2, \ldots, m-1 \tag{5a}$$

$$c_n = \sum_{k=0}^{m-1} \phi_{nk} f_k \frac{1}{m} \qquad n = 0, 1, 2, \ldots, m-1 \tag{7a}$$

where f_k is the average value of the function in question in the kth subinterval, $\phi_{n,k}$ is the value of the nth Walsh function in the kth subinterval, and m is the total number of subintervals.

To illustrate the use of a discrete formula, let us evaluate the Walsh series again for the ramp function in its tabulated form. Given

k	0	1	2	3
f_k	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{5}{8}$	$\frac{7}{8}$

The corresponding graphical form is the ramp function.

Eqn. 7a in its expansion form for $m = 4$ is as follows:

$$\begin{bmatrix} c_0\\ c_1\\ c_2\\ c_3 \end{bmatrix} = \begin{bmatrix} \phi_{00} & \phi_{01} & \phi_{02} & \phi_{03}\\ \phi_{10} & \phi_{11} & \phi_{12} & \phi_{13}\\ \phi_{20} & \phi_{21} & \phi_{22} & \phi_{23}\\ \phi_{30} & \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix} \begin{bmatrix} f_0\\ f_1\\ f_2\\ f_3 \end{bmatrix} \times \tfrac{1}{4} \tag{8}$$

Substituting the tabulated data of the ramp function into eqn. 8, we have

$$\begin{bmatrix} c_0\\ c_1\\ c_2\\ c_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1\\ 1 & 1 & -1 & -1\\ 1 & -1 & 1 & -1\\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{8}\\ \frac{3}{8}\\ \frac{5}{8}\\ \frac{7}{8} \end{bmatrix} \times \tfrac{1}{4} \tag{9a}$$

$$= \begin{bmatrix} \frac{1}{2}\\ -\frac{1}{4}\\ -\frac{1}{8}\\ 0 \end{bmatrix} \tag{9b}$$

ORIGINAL PAGE IS
OF POOR QUALITY

The square matrix defined in eqn. 8 and the numerical values shown in eqn. 9a are easily recognised from the definition of Walsh functions.

It is seen that the Walsh series of a unit ramp function obtained from a discrete formula, or from eqns. 9, and that obtained from the analytic formula are, of course, the same.

Eqn. 8 can be written into a general compact form:

$$c = Wf\frac{1}{m} \tag{10}$$

where W is called the Walsh matrix.

3.2 Operational matrix

In the preceding Section, we showed that the ramp function can be expressed by a Walsh series, or

$$t \cong \tfrac{1}{2}\phi_0(t) - \tfrac{1}{4}\phi_1(t) - \tfrac{1}{8}\phi_2(t) \tag{11}$$

However, a ramp function can be considered as the first integral of a unit step function, or $\phi_0(t)$. Therefore, we write the following

$$\int_0^t \phi_0(x)\, dx \cong [\tfrac{1}{2}, -\tfrac{1}{4}, -\tfrac{1}{8}, 0] \begin{bmatrix} \phi_0\\ \phi_1\\ \phi_2\\ \phi_3 \end{bmatrix} \tag{12}$$

The first integral of $\phi_1(t)$ is a triangular function, and if we expand the triangular function into a Walsh series by using discrete formula with $m = 4$, we have

$$\int_0^t \phi_1(t)\, dx \cong [\tfrac{1}{4}, 0, 0, -\tfrac{1}{8}] \begin{bmatrix} \phi_0\\ \phi_1\\ \phi_2\\ \phi_3 \end{bmatrix} \tag{13}$$

Similarly, we can evaluate the Walsh-series coefficients of the first integration of $\phi_2(t)$ and $\phi_3(t)$ and easily obtain

$$\int_0^t \phi_2(x)\,dx \cong [\tfrac{1}{8},\ 0,\ 0,\ 0]\begin{bmatrix}\phi_0\\ \phi_1\\ \phi_2\\ \phi_3\end{bmatrix}. \tag{14}$$

and

$$\int_0^t \phi_3(x)\,dx \cong [0,\ \tfrac{1}{8},\ 0,\ 0]\begin{bmatrix}\phi_0\\ \phi_1\\ \phi_2\\ \phi_3\end{bmatrix}. \tag{15}$$

Combining eqns. 12–15, we have

$$\begin{bmatrix}\int_0^t \phi_0(x)\,dx\\ \int_0^t \phi_1(x)\,dx\\ \int_0^t \phi_2(x)\,dx\\ \int_0^t \phi_3(x)\,dx\end{bmatrix} \cong \begin{bmatrix}\tfrac{1}{2} & -\tfrac{1}{4} & -\tfrac{1}{8} & 0\\ \tfrac{1}{4} & 0 & 0 & -\tfrac{1}{8}\\ \tfrac{1}{8} & 0 & 0 & 0\\ 0 & \tfrac{1}{8} & 0 & 0\end{bmatrix}\begin{bmatrix}\phi_0(t)\\ \phi_1(t)\\ \phi_2(t)\\ \phi_3(t)\end{bmatrix}. \tag{16}$$

or in compact form

$$\int_0^t \varphi_{(4)}(x)\,dx \cong P_{(4\times 4)}\varphi_{(4)}(t) \tag{17}$$

$P_{(4\times 4)}$ is called the operational matrix of dimension 4 which relates Walsh functions and their integrals. It is chosen as a square matrix for the reason of convenient calculation.

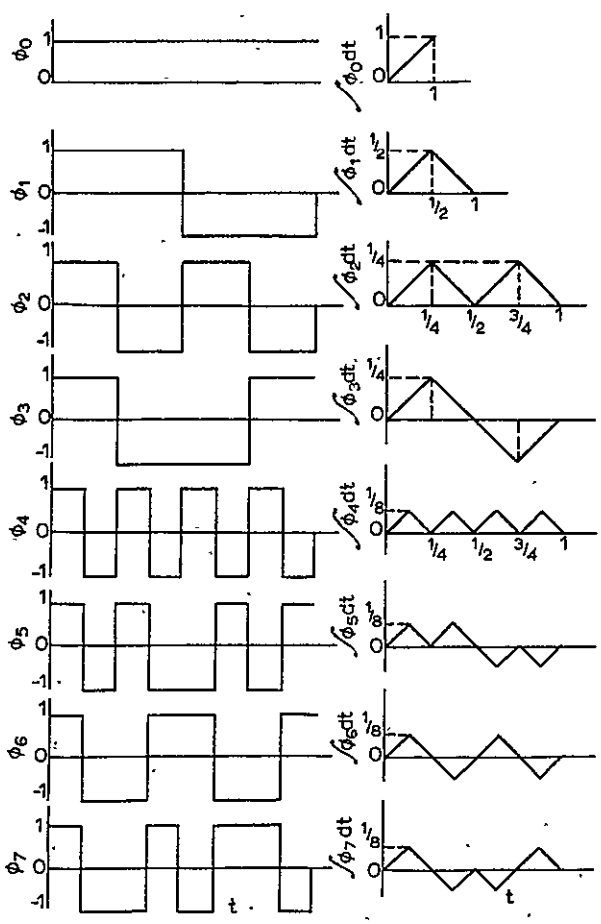


Fig. 3
Walsh functions and their first integrals

By the use of eqn. 17, integration becomes multiplication, therefore, we consider P as an operational matrix.

If we divided the unit [0, 1) into eight subintervals instead of four, and evaluated $\int \phi_0\,dt, \int \phi_1\,dt, \ldots, \int \phi_7\,dt$ by either an analytic method or a discrete formula, we would obtain a group of triangular waves as shown in Fig. 3. Then we could expand the triangular waves into Walsh functions, arriving at the following formula

$$\begin{bmatrix}\int \phi_0\,dt\\ \int \phi_1\,dt\\ \int \phi_2\,dt\\ \int \phi_3\,dt\\ \hline \int \phi_4\,dt\\ \int \phi_5\,dt\\ \int \phi_6\,dt\\ \int \phi_7\,dt\end{bmatrix} \cong \left[\begin{array}{cccc|cccc}\tfrac{1}{2} & -\tfrac{1}{4} & -\tfrac{1}{8} & 0 & -\tfrac{1}{16} & 0 & 0 & 0\\ \tfrac{1}{4} & 0 & 0 & -\tfrac{1}{8} & 0 & -\tfrac{1}{16} & 0 & 0\\ \tfrac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\tfrac{1}{16} & 0\\ 0 & \tfrac{1}{8} & 0 & 0 & 0 & 0 & 0 & -\tfrac{1}{16}\\ \hline \tfrac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & \tfrac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & \tfrac{1}{16} & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & \tfrac{1}{16} & 0 & 0 & 0 & 0\end{array}\right] \begin{bmatrix}\phi_0\\ \phi_1\\ \phi_2\\ \phi_3\\ \hline \phi_4\\ \phi_5\\ \phi_6\\ \phi_7\end{bmatrix} \tag{18}$$

which is

$$\int_0^t \varphi_{(8)}(x)\,dx \cong P_{(8\times 8)}\varphi_{(8)}(t) \tag{18a}$$

It is interesting to note that the upper left corner of $P_{(8\times 8)}$ is exactly $P_{(4\times 4)}$ in eqn. 17, the upper right corner and the lower corner are unit matrices multiplied by $-\frac{1}{16}$ and $\frac{1}{16}$, respectively, and the lower right corner is simply a null matrix.

Following a similar reasoning line, we can write a general expression for the operational matrix P of order m (which is a positive integer power of 2) as follows.

$$\left[\begin{array}{c|c|c|c}\frac{1}{2} & -\frac{2}{m}I_{\left(\frac{m}{8}\right)} & -\frac{1}{m}I_{\left(\frac{m}{4}\right)} & \\ \hline \frac{2}{m}I_{\left(\frac{m}{8}\right)} & O_{\left(\frac{m}{8}\right)} & & \\ \hline & \frac{1}{m}I_{\left(\frac{m}{4}\right)} & O_{\left(\frac{m}{4}\right)} & -\frac{1}{2m}I_{\left(\frac{m}{2}\right)}\\ \hline & & \frac{1}{2m}I_{\left(\frac{m}{2}\right)} & O_{\left(\frac{m}{2}\right)}\end{array}\right] \tag{19}$$

This operational matrix will play an important role in the time-domain-synthesis problems.

4 Principle of transfer-function synthesis

Consider the following differential equation:

$$y^{(n)} + a_1 y^{(n-1)} + \ldots a_{n-1}y^1 + a_n y = b_1 u^{(n-1)} + \ldots + b_{n-1}u + b_n u \tag{20}$$

where $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$ are unknowns while input u and output y are given analytically or numerically. Also, assume all initial conditions are equal to zero.

Integrating both sides of eqn. 20 n times, we have

$$\begin{aligned} y(t) + a_1 \int_0^t y(t)\,dt + \ldots \underbrace{a_n \int \ldots \int y(t)\,dt}_{n} \\ = b_1 \int_0^t u(t)\,dt + \ldots + \underbrace{b_n \int \ldots \int u(t)\,dt}_{n} \end{aligned} \tag{21}$$

Both given $y(t)$ and $u(t)$ may be expanded into Walsh series. We write

$$y(t) = c'\varphi(t) = c_0\phi_0 + c_1\phi_1 + c_2\phi_2 + \ldots \tag{22}$$

and

$$u(t) = h'\varphi(t) = h_0\phi_0 + h_1\phi_1 + h_2\phi_2 + \ldots \tag{23}$$

As explained in Section 3, the first integration of a Walsh function may be expressed approximately by

$$\int_0^t \phi(x)\,dx \cong P\varphi(t) \tag{24}$$

Substituting eqns. 22, 23 and 24 into eqn. 21, we have

$$\begin{aligned} c'[I + a_1P + a_2P^2 + \ldots + a_nP^n]\,\varphi(t) \\ = h'[b_1P + b_2P^2 + \ldots + b_nP^n]\,\varphi(t) \end{aligned} \tag{25}$$

Eqn. 25 must be satisfied for any value of t. Let us take, say, $2n$ samples, at $t_1, t_2, \ldots, t_{2n}$. We have

P-198

$$c'[I + a_1P + a_2P^2 + \dots a_nP^n]\varphi(t_1) = h'[b_1P + b_2P^2 + \dots + b_nP^n]\varphi(t_1)$$
$$c'[I + a_1P + a_2P^2 + \dots a_nP^n]\varphi(t_2) = h'[b_1P + b_2P^2 + \dots + b_nP^n]\varphi(t_2)$$
$$\vdots$$
$$c'[I + a_1P + a_2P^2 + \dots a_nP^n]\varphi(t_{2n}) = h'[b_1P + b_2P^2 + \dots + b_nP^n]\varphi(t_{2n}) \quad (26)$$

We can solve the $2n$ equations for the $2n$ unknowns a and b. Then let us define a matrix Φ of $2n \times m$ as

$$\Phi' \triangleq \begin{bmatrix} \varphi'(t_1) \\ \varphi'(t_2) \\ \vdots \\ \varphi'(t_{2n}) \end{bmatrix} \quad (27)$$

ORIGINAL PAGE IS
OF POOR QUALITY

Eqn. 26 is then simplified to

$$\{\Phi'[-P'c, -P'^2c, \dots, -P'^nc, P'h \dots P'^nh]\}\begin{bmatrix} a \\ b \end{bmatrix} = \Phi'c \quad (26a)$$

Performing inversions on the $2n \times 2n$ matrix in the $\{\cdot\}$, we obtain a and b:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \{\Phi'[-P'c \dots -P'^nc, P'h \dots P'^nh]\}^{-1}\Phi c \quad (28)$$

Eqn. 28 is the basic formula to fit eqn. 20; if input/output behaviour is given.

4.1 Illustrative example

Suppose the input to a system is a unit step and the output data are as follows:

t	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
$\bar{y}(t)$	0	0·632	0·865	0·95	0·982

It is required to find the differential equation or the transfer function of the system.

Let us average $y(t)$ of each subinterval first.

k	0	1	2	3
y_k	0·316	0·7485	0·9075	0·966

Applying the discrete formula shown in eqns 9 with $n = 4$, we obtain

$$c = W\bar{y}\frac{1}{m}$$
$$= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 0{\cdot}316 \\ 0{\cdot}7485 \\ 0{\cdot}9075 \\ 0{\cdot}966 \end{bmatrix} \times \frac{1}{4} = \begin{bmatrix} 0{\cdot}7345 \\ -0{\cdot}20225 \\ -0{\cdot}12275 \\ -0{\cdot}0935 \end{bmatrix}$$

Suppose the linear system is defined as

$$y + ay = bu$$
$$u = [1, 0, 0, 0]\,\varphi(t)$$

and choose $t_1 = \frac{1}{8}$ and $t_2 = \frac{3}{8}$. Then

$$\Phi' = \begin{bmatrix} \varphi'(\frac{1}{8}) \\ \varphi'(\frac{3}{8}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix}$$

Applying eqn. 28 with P defined in eqn. 16 yields

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 3{\cdot}48 \\ 3{\cdot}63 \end{bmatrix}$$

The required differential equation is therefore

$$\dot{y} + 3{\cdot}48y = 3{\cdot}63u \quad (29)$$

or the required transfer function is

$$\frac{Y(s)}{U(s)} = \frac{3{\cdot}63}{s + 3{\cdot}48} \quad (29a)$$

For the same problem, if we took more samples we would obtain a better result. To demonstrate this statement, let us take nine samples

t	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{4}{8}$	$\frac{5}{8}$	$\frac{6}{8}$	$\frac{7}{8}$	1
$y(t)$	0	0·394	0·632	0·789	0·865	0·918	0·950	0·970	0·982
$\bar{y}_k$		0·197	0·513	0·710	0·826	0·889	0·934	0·960	0·976

Using eqn. 9 again with $m = 8$, we have

$$c = Wy\tfrac{1}{8} = \tfrac{1}{8}[6{\cdot}063, -1{\cdot}511, -0{\cdot}941, -0{\cdot}711, -0{\cdot}495, -0{\cdot}369, -0{\cdot}231, -0{\cdot}169]$$

choosing $t_1 = 13/16$ and $t_2 = 15/16$, we find

$$\Phi = \begin{bmatrix} \varphi'(\frac{13}{16}) \\ \varphi'(\frac{15}{16}) \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

Applying eqn 28 with $P = P_{(8 \times 8)}$, we finally obtain

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 3{\cdot}9287759 \\ 3{\cdot}9243553 \end{bmatrix}$$

Therefore the differential equation of the system is

$$\dot{y} + 3{\cdot}9287759y = 3{\cdot}9243553u \quad (30)$$

When we took five samples, the differential equation fitted was eqn. 29; for nine samples, we obtain eqn. 30. The unit step response is actually taken from the following system:

$$\dot{y} + 4y = 4u$$

It is interesting to note that the result obtained by the five-point approach is acceptable, and that by the nine-point approach is quite satisfactory.

5 Principle of state-equation synthesis

Suppose we are interested in realising the following system:

$$\dot{x} = Ax \quad (31)$$

from a set of zero-input response data $x(t_i)$, $i = 0, 1, \dots, m$.

First of all, we expand $x(t)$ into a Walsh series with m terms

$$x(t) = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n' \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_{m-1} \end{bmatrix} = C\varphi(t) \quad (32)$$

where C is an $n \times m$ matrix; $\varphi(t)$, an $m \times 1$ vector.

Integrating both sides of eqn. 31 and using the expression of eqn. 32 and the operational matrix P, we have

$$\int_0^t \dot{x}(\lambda)d\lambda = A\int_0^t x(\lambda)d\lambda = AC\int_0^t \varphi(\lambda)d\lambda \quad (33)$$
$$x(t) - x(0) = ACP\varphi(t)$$
$$C\varphi(t) - [x(0), 0, \dots, 0]\,\varphi(t) = ACP\varphi(t) \quad (33a)$$

Eqn. 33a must hold for any value of t. Let us take n samples.

Define

$$\Phi \triangleq [\varphi(t_1), \varphi(t_2) \dots \varphi(t_n)] \quad (34)$$
$$\{C - [x(0), 0, 0, \dots, 0]\}\Phi = ACP\Phi \quad (35)$$

Then the system matrix A can be evaluated by

$$A = \{C - [x(0), 0, 0, \dots, 0]\}\Phi\{CP\Phi\}^{-1} \quad (36)$$

5.1 Illustrative example

A zero-input response of a system is tabulated as follows:

t	$x_1(t)$	$x_2(t)$	$\bar{x}_1$	$\bar{x}_2$
0	1	−1		
			0·8772	−0·9737
$\frac{1}{4}$	0·7545	−0·9494		
			0·6433	−0·8852
$\frac{1}{2}$	0·5321	−0·8231		
			0·4388	−0·7453
$\frac{3}{4}$	0·3454	−0·6676		
			0·2720	−0·5879
1	0·1985	−0·0508		

J.199

Using the discrete formula, we find

$$c_1 \cong W\bar{x}_1\frac{1}{m} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 0{\cdot}8772 \\ 0{\cdot}6433 \\ 0{\cdot}4388 \\ 0{\cdot}2720 \end{bmatrix} \times \tfrac{1}{4} = \begin{bmatrix} 2{\cdot}23153 \\ 0{\cdot}809797 \\ 0{\cdot}4007092 \\ 0{\cdot}0671214 \end{bmatrix} \times \tfrac{1}{4}$$

$$c_2 \cong W\bar{x}_2\frac{1}{m} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} -0{\cdot}8737 \\ -0{\cdot}8852 \\ -0{\cdot}7453 \\ -0{\cdot}5879 \end{bmatrix} \times \tfrac{1}{4} = \begin{bmatrix} -3{\cdot}192311 \\ -0{\cdot}525631 \\ -0{\cdot}2458685 \\ 0{\cdot}0689903 \end{bmatrix} \times \tfrac{1}{4}$$

$$\text{Let } \Phi = [\varphi(\tfrac{1}{8}), \varphi(\tfrac{3}{8})] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix}$$

The initial conditions are given as

$$x(0) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

We use $P_{(4\times 4)}$ formula, or

$$P_{(4\times 4)} = \left[\begin{array}{c|c|c|c} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{8} & 0 \\ \hline \frac{1}{4} & 0 & 0 & -\frac{1}{8} \\ \hline \frac{1}{8} & 0 & 0 & 0 \\ \hline 0 & \frac{1}{8} & 0 & 0 \end{array}\right]$$

Substituting $C, x(0), P$ and Φ into eqn. 36, we have

$$\{C - [x(0), 0, 0, 0]\}\,\Phi = \tfrac{1}{4}\begin{bmatrix} -0{\cdot}4908416 & -2{\cdot}911856 \\ 0{\cdot}1051796 & 1{\cdot}6481788 \end{bmatrix}$$

$$CP\Phi = \begin{bmatrix} 3{\cdot}5091584 & 16{\cdot}764101 \\ -3{\cdot}8948204 & -23{\cdot}186666 \end{bmatrix} \times \tfrac{1}{32}$$

Therefore

$$A = \tfrac{32}{4}\begin{bmatrix} -0{\cdot}4908416 & -2{\cdot}911856 \\ 0{\cdot}1051796 & 1{\cdot}6481788 \end{bmatrix} \begin{bmatrix} 3{\cdot}5091584 & 16{\cdot}764101 \\ -3{\cdot}8948204 & -23{\cdot}186666 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 0{\cdot}0198261 & 0{\cdot}9903525 \\ -1{\cdot}9962949 & -2{\cdot}0061511 \end{bmatrix} \tag{37}$$

The samples actually are taken from the zero-input response of a system

$$\dot{x} = Ax$$

where

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -2 \end{bmatrix} \tag{38}$$

Comparing eqn. 37 with eqn. 38, we see the power of the method.

6 Laboratory test

The common excitations usually used in a laboratory are neither unit-step-input nor initial conditions, but rather gate functions. When the input is a given gate function and the output is measured, it is desired to find the transfer function of the system. The classical approaches for the domain synthesis are not easily applied. The new Walsh-function method, however, can fit the transfer function as usual. The following example will demonstrate the procedure and the accurate results.

Suppose a gate function is applied to a linear system, and the output is recorded as follows.

t	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
$y(t)$	0	0·245	0·222	0·1867	0·1469
$\bar{y}_k$		0·1225	0·2335	0·20435	0·1668

The gate input is defined as

$$u(t) = \begin{cases} 1 & \text{or } 0 \leqslant t < \frac{1}{4} \\ 0 & \text{elsewhere} \end{cases}$$

Using a discrete formula, we can find the Walsh series for the output.

$$y(t) \cong c'\varphi = [c_0, c_1, c_2, c_3]\varphi$$

$$c' = W\bar{y}\frac{1}{m} = \tfrac{1}{4}[0{\cdot}72715, -0{\cdot}01515, -0{\cdot}07345, -0{\cdot}14855]$$

The Walsh-series expansion for the gate input can be easily shown as

$$u(t) = [\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}]\,\varphi = h'\varphi$$

$$h' = \tfrac{1}{4}[1, 1, 1, 1]$$

Using operational matrix P, we can find the first and second integrations of $y(t)$ and $u(t)$. For a second-order system, there are four unknowns, two as and two bs, and $m = 4$ for five discrete data. Therefore

$$\Phi = W = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

Substituting c, h, P and Φ we obtain

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1{\cdot}8173251 \\ 1{\cdot}8531946 \\ 1{\cdot}0107953 \\ 1{\cdot}7734002 \end{bmatrix}$$

That means the system may be approximately described as

$$y + 1{\cdot}8173251\,\dot{y} + 1{\cdot}8531946\,y = 1{\cdot}0107953\,\dot{u} + 1{\cdot}7734002\,u$$

To obtain better results, we may take more data, say 17 points as follows:

t	0	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{4}$	$\frac{5}{16}$	$\frac{3}{8}$	$\frac{7}{16}$	$\frac{1}{2}$
$y(t)$	0	0·0624	0·1244	0·1855	0·245	0·241	0·236	0·230	0·222

t	$\frac{9}{16}$	$\frac{5}{8}$	$\frac{11}{16}$	$\frac{3}{4}$	$\frac{13}{16}$	$\frac{7}{8}$	$\frac{15}{16}$	1
$y(t)$	0·214	0·205	0·1962	0·1866	0·1768	0·1669	0·1569	0·1469

And choose $t_1 = 25/32, t_2 = 27/32, t_3 = 29/32, t_4 = 31/32$, namely

$$\Phi' = \begin{bmatrix} \varphi'(25/32) \\ \varphi'(27/32) \\ \varphi'(29/32) \\ \varphi'(31/32) \end{bmatrix} =$$

$$= \begin{bmatrix} 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{bmatrix}$$

Then we obtain

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2{\cdot}050781 \\ 2{\cdot}039063 \\ 0{\cdot}9980469 \\ 2{\cdot}042969 \end{bmatrix}$$

$$y + 2{\cdot}050781\,\dot{y} + 2{\cdot}039063\,y = 0{\cdot}9980469\,\dot{u} + 2\,041969\,u$$

The actual system is

$$y + 2\dot{y} + 2y = \dot{u} + 2u.$$

7 Conclusions

A new time-domain synthesis method based on the Walsh functions has been established. With zero-state response given and the

transfer function of the system desired, by using the Walsh-function principle, we derived eqn. 28 to fit the transfer function. On the other hand, with zero-input response given and the state equation desired, we derived eqn. 36 to fit. Three examples are included for illustration.

If the order of the transfer function assumed is too high, the inverse of eqn. 28 does not exist. Then we can decrease the order and solve the problem again. Similarly, if the dimension of the state equation is assumed too high initially, the inversion of the matrix in eqn. 36 does not exist, and we should lower the dimension.

If what we fitted is a high-order transfer function and what we want is a lower order one, then we face a new problem, namely, the model reduction problem. The first author[10] has suggested continued-fraction methods to solve this problem effectively.

It is seen that the Walsh-function approach belongs to the school of least-squares identification. In general, however, least-squares identification is to use orthogonal polynomials or

$$C_i = \frac{\sum_{t=0}^{n} f(t)\,\psi_{ni}(t)}{\sum_{t=0}^{n} \psi_{ni}^2(t)} \qquad i = 1, \ldots, m$$

where $\psi(t)$ is an orthogonal polynomial, and C_i and $f(t)$ are coefficients and the given data, respectively, as defined before.

The advantages of the Walsh-function approach are apparent: (*a*) it is still least-squares identification, (*b*) the denominator becomes 1, because Walsh functions are not only orthogonal, but also orthonormal, (*c*) the fast Walsh algorithm[11, 12] is available.

8 References

1 GUILLEMIN, E.A.: 'Synthesis of passive networks' (Wiley, 1957)
2 KALMAN, R.E., and DECLARIS, N.: 'Aspects of network and system theory' (Holt, Rinehart & Winston, 1971), pp. 311–325
3 SHIEH, L.S., CHEN, C.F., and HUANG, C.J.: 'On identifying transfer functions and state equations for linear systems', *IEEE Trans.*, 1972, AES-8, pp. 811–820
4 PICHER, F.: 'Walsh-function and optimal linear systems'. Proceedings of the Walsh-function symposium, Washington DC, 1970, pp. 17–22
5 CORRINGTON, M.S.: 'Solution of differential and integral equations with Walsh functions', *IEEE Trans*, 1973, CT-20, pp. 470–475
6 LEE, J.D.: 'Review of recent work on applications of Walsh functions in communications', Proceedings of the Walsh-function symposium, Washington, DC, 1970, pp. 26–35
7 GIBBS, J.E., and GEBBIE, H A.: 'Application of Walsh function to transform spectroscopy', *Nature*, 1969, 224, pp. 1021–1013
8 RADEMACHER, H.: 'Einige Salze uber Reiben von allgemeinen Orthogonal functionen', *Math. Ann.*, 1922, 87, pp. 712–718
9 WALSH, J.L.: 'A closed set of orthogonal functions', *Am J. Math*, 1923, 45, pp. 5–24
10 CHEN, C.F.: 'Model reduction of multivariable systems via matrix continued fractions. Presented at the 5th International Federation of Automatic Control, Paris, France, 1972 and published in *Int J of Contr.*, 1974, 20, pp. 225–238
11 MANZ, J.H.. 'A sequency-ordered fast Walsh transform,' *IEEE Trans.*, 1972, AU-20, pp. 204–205
12 SHANKS, J.L.: 'Computation of the fast Walsh-Fourier transform,' *ibid.*, 1969, C-18, pp 457–459

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: A Technique for Expanding the Matrix Riccati Equation

(B) Project Abstract:

A new technique for expanding the matrix Riccati equation is established: From the given equation of the problem, groups of rules are formulated and based on these rules a computer program is written. It makes various well known numerical methods directly applicable to the problem.

(C) Publication: Computing and Electrical Engineering, 3, 193 (1976)

(D) Year: Completed in 1975

(E) Department: Electrical Engineering

(F) Authors: Professors C. F. Chen and L. S. Shieh

Comput. & Elect. Engng, Vol. 3, pp. 193–202. Pergamon Press 1976. Printed in Great Britain

A TECHNIQUE FOR EXPANDING THE MATRIX RICCATI EQUATION

C. F. CHEN and L. S. SHIEH
University of Houston, Houston, TX 77004 U.S.A.

ORIGINAL PAGE IS
OF POOR QUALITY

(Received 8 July 1975)

1. INTRODUCTION

It is well known[1–4] that optimizing the linear plant

$$\dot{\underset{\sim}{x}} = \underset{\approx}{A}\underset{\sim}{x} + \underset{\approx}{B}\underset{\sim}{u} \tag{1}$$

for a quadratic performance index

$$J = \frac{1}{2}\int_0^\infty (\underset{\sim}{x}^T \underset{\approx}{Q}\underset{\sim}{x} + \underset{\sim}{u}^T \underset{\approx}{R}\underset{\sim}{u})\mathrm{d}t \tag{2}$$

where $\underset{\approx}{Q}$ and $\underset{\approx}{R}$ are symmetric matrices, we obtain the following optimal control law:

$$u^*(t) = -\underset{\approx}{R}^{-1}\underset{\approx}{B}^T\underset{\sim}{\phi}(t) \tag{3}$$

where $\underset{\sim}{\phi}(t)$ is the vector related to $\underset{\sim}{x}(t)$ by

$$\begin{bmatrix} \dot{\underset{\sim}{x}} \\ \dot{\underset{\sim}{\phi}} \end{bmatrix} = \begin{bmatrix} \underset{\approx}{A} & -\underset{\approx}{C} \\ -\underset{\approx}{Q} & -\underset{\approx}{A}^T \end{bmatrix} \begin{bmatrix} \underset{\sim}{x} \\ \underset{\sim}{\phi} \end{bmatrix} \tag{4}$$

in which $\underset{\approx}{C} = \underset{\approx}{B}\underset{\approx}{R}^{-1}\underset{\approx}{B}^T$, subject to the boundary conditions:

$$\underset{\sim}{x}(0) = \underset{\sim}{x}_0 \tag{5}$$

$$\underset{\sim}{\phi}(\infty) = \underset{\sim}{0} \tag{6}$$

Solving (4) and then substituting the result into (3), the optimal control will be found.

An alternative consideration is to use the following linear transformation:

$$\underset{\sim}{\phi} = \underset{\approx}{P}\underset{\sim}{x}. \tag{7}$$

Substituting (7) into (4), we arrive at

$$\dot{\underset{\approx}{P}} = \underset{\approx}{P}\underset{\approx}{B}\underset{\approx}{R}^{-1}\underset{\approx}{B}^T\underset{\approx}{P} - (\underset{\approx}{P}\underset{\approx}{A} + \underset{\approx}{A}^T\underset{\approx}{P}) - \underset{\approx}{Q} \tag{8}$$

which is called the matrix Riccati Equation. Solving for $\underset{\approx}{P}$ and then substituting it into (3)–(7), we will obtain the same control law, of course.

2. DIFFICULTIES IN OBTAINING SOLUTIONS

In solving (8) if the dimension of the equation is high, to expand it into a set of simultaneous differential equations is by no means an easy task. Considering the difficulties involved in it, many pioneers use (4) directly. MacFarlane[5] and Potter[6] use the eigenvector solution, while Vaughan[7] offers a negative exponential solution. Fath[8] establishes a procedure which involves a particular decoupling matrix, etc.

The technique established in this paper is to obtain a set of simultaneous component equations of (8) by a digital computer with the original symbols preserved.

3. STATEMENT OF THE EXPANSION PROBLEM

In the matrix Riccati Equation (8), the matrices $\underset{\sim}{A}$, $\underset{\sim}{B}$, $\underset{\sim}{Q}$ and $\underset{\sim}{R}$ are constant matrices; dimensionally they are $n \times n$, $n \times m$, $n \times n$ and $m \times m$ matrices respectively; $\underset{\sim}{P}$ is an $n \times n$ matrix. Also, $\underset{\sim}{P}$, $\underset{\sim}{Q}$ and $\underset{\sim}{R}$ are symmetrical matrices.

Let us regroup eqn (8) first,

$$\dot{\underset{\sim}{P}} = (\underset{\sim}{P}\underset{\sim}{B}\underset{\sim}{R}^{-1}\underset{\sim}{B}^T\underset{\sim}{P}) - (\underset{\sim}{P}\underset{\sim}{A} + \underset{\sim}{A}^T\underset{\sim}{P}) - \underset{\sim}{Q}. \tag{8a}$$

Because $\underset{\sim}{B}$ and $\underset{\sim}{R}$ are given, we simplify (8a) again

$$\dot{\underset{\sim}{P}} = (\underset{\sim}{P}\underset{\sim}{C}\underset{\sim}{P}) - (\underset{\sim}{P}\underset{\sim}{A} + \underset{\sim}{A}^T\underset{\sim}{P}) - \underset{\sim}{Q} \tag{8b}$$

where

$$\underset{\sim}{C} = \underset{\sim}{B}\underset{\sim}{R}^{-1}\underset{\sim}{B}^T \tag{9}$$

as we defined before.

$\underset{\sim}{C}$ is still an $n \times n$ symmetrical matrix.

Then arrange the elements of $\underset{\sim}{P}$ into a vector $\vec{\underset{\sim}{p}}$ such that

$$\vec{\underset{\sim}{p}} = [P_{11}, P_{12}, P_{13} \ldots . P_{1n}, P_{22}, P_{23} \ldots . P_{n-1,n}, P_{nn}]^T$$

Also, organize another vector $\vec{\vec{\underset{\sim}{p}}}$ to express the quadratic elements

$$\vec{\vec{\underset{\sim}{p}}} = [P_{11}P_{11}, P_{11}P_{12}, \ldots P_{11}P_{1n} \ldots P_{11}P_{nn}, P_{12}P_{12},$$

$$P_{12}P_{13}, \ldots . P_{12}P_{1n}, \ldots P_{12}P_{nn}, \ldots . P_{(n-1)n}$$

$$P_{(n-1)n}, P_{(n-1)n}P_{nn}, P_{nn}P_{nn}]^T.$$

For matrix $\underset{\sim}{Q}$, we arrange it into a vector such that

$$\vec{\underset{\sim}{q}} = [Q_{11}, Q_{12}, Q_{13}, \ldots . Q_{1n}, Q_{22}, Q_{23}, \ldots Q_{(n-1)n}, Q_{nn}]^T.$$

It is desired to change the matrix Ricatti Equation (8b) into the following set of simultaneous component differential equations

$$\dot{\vec{\underset{\sim}{p}}} = \underset{\sim}{\Omega}\vec{\vec{\underset{\sim}{p}}} - \underset{\sim}{\alpha}\vec{\underset{\sim}{p}} - \vec{\underset{\sim}{q}}. \tag{10}$$

The problem is to find $\underset{\sim}{\Omega}$ and $\underset{\sim}{\alpha}$.

The dimension of the vector $\vec{\underset{\sim}{p}}$ or $\vec{\underset{\sim}{p}}$ is $n(n+1)/2$ and that of $\vec{\vec{\underset{\sim}{p}}}$ is $[n(n+1)(n^2+n+2)/8] \cdot \underset{\sim}{\alpha}$ is an $[n(n+1)/2] \times [n(n+1)/2]$ square matrix while $\underset{\sim}{\Omega}$ is a rectangular matrix with dimensions

$$\frac{n(n+1)}{2} \times \frac{n(n+1)(n^2+n+2)}{8}.$$

4. SIMPLE EXAMPLE FOR THE EXPANSION

To illustrate the notations in (10), we give a simple example as follows:

We rewrite (8)

$$\dot{\underset{\sim}{P}} = \underset{\sim}{P}\underset{\sim}{C}\underset{\sim}{P} - (\underset{\sim}{P}\underset{\sim}{A} + \underset{\sim}{A}^T\underset{\sim}{P}) - \underset{\sim}{Q}. \tag{8a}$$

p.204.

Consider that the order of the equation is two. The equation becomes

$$\begin{bmatrix} \dot{P}_{11} & \dot{P}_{12} \\ \dot{P}_{12} & \dot{P}_{22} \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ P_{12} & P_{22} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} \\ P_{12} & P_{22} \end{bmatrix}$$
$$-\left\{ \begin{bmatrix} P_{11} & P_{12} \\ P_{12} & P_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} \\ P_{12} & P_{22} \end{bmatrix} \right\} - \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12} & Q_{22} \end{bmatrix}. \tag{11}$$

Expanding (11), and arranging it into component equations, we have

$$\begin{bmatrix} \dot{P}_{11} \\ \dot{P}_{12} \\ \dot{P}_{22} \end{bmatrix} = \begin{bmatrix} C_{11} & 2C_{12} & 0 & C_{22} & 0 & 0 \\ 0 & C_{11} & C_{12} & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & C_{11} & 2C_{12} & C_{22} \end{bmatrix} \begin{bmatrix} P_{11}P_{11} \\ P_{11}P_{12} \\ P_{11}P_{22} \\ P_{12}P_{12} \\ P_{12}P_{22} \\ P_{22}P_{22} \end{bmatrix}$$

ORIGINAL PAGE IS
OF POOR QUALITY

$$-\begin{bmatrix} 2A_{11} & 2A_{21} & 0 \\ A_{12} & A_{11}+A_{22} & A_{21} \\ 0 & 2A_{12} & 2A_{22} \end{bmatrix}, \begin{bmatrix} P_{11} \\ P_{12} \\ P_{22} \end{bmatrix}, -\begin{bmatrix} Q_{11} \\ Q_{12} \\ Q_{22} \end{bmatrix}. \tag{12}$$

In this example

$$\underset{=}{\Omega} = \begin{bmatrix} C_{11} & 2C_{12} & 0 & C_{22} & 0 & 0 \\ 0 & C_{11} & C_{12} & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & C_{11} & 2C_{12} & C_{22} \end{bmatrix} \tag{13}$$

and

$$\underset{=}{\alpha} = \begin{bmatrix} 2A_{11} & 2A_{21} & 0 \\ A_{12} & A_{11}+A_{22} & A_{21} \\ 0 & 2A_{12} & 2A_{22} \end{bmatrix} \tag{14}$$

The example is the simplest problem of its kind because the dimension is two. Even so, we can see the complication of its $\underset{=}{\Omega}$ and its $\underset{=}{\alpha}$. When the dimension increases to a higher order, to find the $\underset{=}{\Omega}$ and $\underset{=}{\alpha}$ matrices is extremely difficult.

This paper attempts to establish two sets of rules for writing $\underset{=}{\Omega}$ and $\underset{=}{\alpha}$. Of course, they must constitute a computer oriented procedure.

5. RULES FOR FINDING $\underset{=}{\alpha}$

In eqn (10), we defined $\underset{=}{\alpha}$ as an $n(n+1)/2$ square matrix. This matrix coincides with the Liapunov function matrix. In a previous paper [10], an algorithm was developed for expanding the Liapunov matrix. After slight modification, that algorithm can be directly used for finding $\underset{=}{\alpha}$.

We use (K, L) and (I, J) as the row and column subscripted indices respectively. The following formula should be followed when forming the $\underset{=}{\alpha}$ matrix directly from the $\underset{=}{A}$ matrix.

(1)
$$\text{if } K = I,\ L \neq J, \rightarrow A(L, J)$$
$$\text{if } K \neq I,\ L = J, \rightarrow A(K, I)$$

(2)
$$\text{if } K \neq I,\ L \neq J,$$
$$\begin{cases} K = J,\ L \neq I \rightarrow A(L, I) \\ K \neq J,\ L = I \rightarrow A(K, J) \\ K \neq J,\ L \neq I \rightarrow 0 \end{cases}$$

(3)
$$\text{if } K = I, L = J$$
$$\begin{cases} K = J,\ L = I \rightarrow A(K, I) \\ K \neq K,\ L \neq I \rightarrow A(K, I) + A(L, J) \end{cases}$$

p. 205

(4) After the first three steps have been considered if $I=J$
all elements at row I should be multiplied by 2.

If the $\underset{=}{A}$ is a 3×3 matrix following the rules shown above, we will have $\underset{=}{\alpha}$ as follows:

		$K-L$					
		$1-1$	$1-2$	$1-3$	$2-2$	$2-3$	$3-3$
	$1-1$	$2A_{11}$	$2A_{21}$	$2A_{31}$	0	0	0
$I-J$	$1-2$	A_{12}	$A_{11}+A_{22}$	A_{32}	A_{21}	A_{31}	0
	$1-3$	A_{13}	A_{23}	$A_{11}+A_{33}$	0	A_{21}	A_{31}
	$2-2$	0	$2A_{12}$	0	$2A_{22}$	$2A_{32}$	0
	$2-3$	0	A_{13}	A_{12}	A_{23}	$A_{22}+A_{33}$	A_{32}
	$3-3$	0	0	$2A_{13}$	0	$2A_{23}$	$2A_{33}$

where A_{ij} are defined by

$$\underset{=}{A}=\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}.$$

Here the subscripted index $K-L$ and $I-J$ are the comparing indices. Each element in the table can be directly written by inspection. For Example, if we want to find the element Z

$$\frac{}{I-J}\left|\frac{K-L}{Z=?}\right.$$

we examine

(1) If $K=I$, $L\neq J$, then $Z=A(L,J)$

(2) if $K\neq I$, $L\neq J$, and $K\neq J$, $L\neq I$, then $Z=0$.

6. RULES FOR FINDING $\underset{=}{\Omega}$

Matrix $\underset{=}{\Omega}$ can be found in a similar way; however, because $\vec{\underset{=}{p}}$ is a much higher dimension vector, the rules for finding $\underset{=}{\Omega}$ are lengthier than those for finding $\underset{=}{\alpha}$.

Examine (10) again; we only consider the first term of the right hand side this time.

For convenience and clarity, we write the elements of $\vec{\underset{=}{p}}$ horizontally and still write $\vec{\underset{=}{p}}$ vertically. Namely, the arrangement is as follows

	$P_{11}P_{11}$	$P_{11}P_{12}$	$P_{11}P_{13}\ldots\ldots$
$\dot{P}_{11}$			
$\dot{P}_{12}$			
$\dot{P}_{13}$			

Then we write the subscripts symbolically

$$\frac{}{\dot{P}(I,J)}\left|\frac{P(K,L)P(M,N)}{Y=?}\right.$$

The value of Y is found by two steps.

(1) Organize a subindices matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

p. 206

first by following the rules indicated below:

(i)

if	$I=K$,	Then	$a=L$
	$I=L$,		$a=K$
	$I=M$		$b=N$
	$I=N$		$b=M$
if	$J=K$,	Then	$c=L$
	$J=L$		$c=K$
	$J=M$		$d=N$
	$J=N$		$d=M$

ORIGINAL PAGE IS
OF POOR QUALITY

(ii)

if	$I\neq K$,	$I\neq L$	Then	$a=0$
	$I\neq M$,	$I\neq L$		$b=0$
if	$J\neq K$,	$J\neq L$,		$c=0$
	$J\neq M$,	$J\neq N$,		$d=0$.

The element Y is found by

$$Y=C(a,d)+C(b,c)$$

subject to the following two conditions:

(i) if a, b, c or d equal to zero, the corresponding C term is equal to zero.

(ii) if $K=M$, $L=N$

Then

$$Y=\frac{C(a,d)+C(b,c)}{2}.$$

How to use these rules can be illustrated by the following 3×3 example.

$(K-L)(M-N)$

	$P_{11}P_{11}$	$P_{11}P_{12}$	$P_{11}P_{13}$	$P_{11}P_{22}$	$P_{11}P_{23}$	$P_{11}P_{33}$	$P_{12}P_{11}$	$P_{12}P_{12}$	$P_{12}P_{13}$	$P_{12}P_{23}$	$P_{12}P_{33}$	$P_{13}P_{11}$	$P_{13}P_{12}$	$P_{13}P_{13}$	$P_{13}P_{22}$	$P_{13}P_{23}$	$P_{13}P_{33}$	$P_{22}P_{22}$	$P_{22}P_{23}$	$P_{22}P_{33}$	$P_{23}P_{23}$	$P_{23}P_{33}$	$P_{33}P_{33}$
$\dot{P}_{11}$	C_{11}	$2C_{12}$	$2C_{13}$				C_{22}	$2C_{23}$				C_{33}											
$\dot{P}_{12}$		C_{11}		C_{12}	C_{13}		C_{12}	C_{13}	C_{22}	C_{23}				C_{23}	C_{33}								
$\dot{P}_{13}$ $I-J$			C_{13}		C_{13}	C_{13}		C_{12}		C_{22}	C_{23}	C_{13}			C_{23}	C_{33}							
$\dot{P}_{22}$							C_{11}		$2C_{12}$	$2C_{13}$							C_{22}	$2C_{23}$		C_{33}			
$\dot{P}_{23}$								C_{11}		C_{22}	C_{23}		C_{12}	C_{13}				C_{22}	C_{23}	C_{23}	C_{33}		
$\dot{P}_{33}$												C_{11}		$2C_{13}$	$2C_{13}$					C_{23}	$2C_{23}$	C_{33}	

7. COMPUTER PROGRAM

Based on the rules for finding the component equations, a computer program has been written.

The part of the expansion of the $\underset{\sim}{P}\underset{\sim}{A}+\underset{\sim}{A}^T\underset{\sim}{P}$ matrix is written as a subroutine called SUBROUTINE *PA*.

The part of the expansion of *PCP* is written as a subroutine called SUBROUTINE *PRP*.

The main program is to use the Runge Kutta four step formula to carry out of the evaluation of $\vec{\underset{\sim}{P}}$. Of course, the negative time increment is used. When the steady state values of the components of $\vec{\underset{\sim}{P}}$ vector have been reached, we use only those values to substitute into eqn (3) and (7). The optimal gains are obtained.

p. 207

The details of preparing the input cards can be summarized as follows

N—dimension of x
M—dimension of R
NM—points of solutions of Runge Kutta Program
TN—starting time
$DELT$—time increment (negative values should be used)

Then the following cards are the row elements of matrices $\underset{\approx}{A}$, $\underset{\approx}{B}$, $\underset{\approx}{Q}$, $\underset{\approx}{R}$. The last card is for PIN which means the initial values of the Runge Kutta Program, or the final values of $\underset{\approx}{\phi}(\infty)$.

8. ILLUSTRATIVE EXAMPLES

For the given plant

$$\dot{\underset{\approx}{x}} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \underset{\approx}{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

and for the quadratic performance index

$$J = \frac{1}{2} \int_0^\infty \left(\underset{\approx}{x}^T \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \underset{\approx}{x} + u^T \left(\frac{1}{2}\right) u \right) \mathrm{d}t$$

the input data cards read

$$N = 2, M = 1, NN = 100, TN = 0, DELT = -0.05.$$

The $\underset{\approx}{P}$ matrix is found from the computer output.

$$\underset{\approx}{P} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}.$$

When the order of the plant is higher, we can see the advantages of using the program. The following example will demonstrate the point.

A plant is given as follows

$$\underset{\approx}{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \underset{\approx}{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Following Fuller[11], we formulate the plant matrices $\underset{\approx}{A}$ and $\underset{\approx}{B}$ as shown. As far as the state feedback design problems are concerned, the example form is a very general one. Then we assigned the constants of $\underset{\approx}{Q}$ and R as follows:

$$\underset{\approx}{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0 & 0.001 & 0 \\ 0 & 0 & 0 & 0 & 0.0001 \end{bmatrix} \qquad \begin{aligned} &R = 1 \\ &PIN = \text{null matrix } 5 \times 5. \end{aligned}$$

The input card should be punched as follows

$$N = 5, M = 1, NN = 200, TN = 0, DELT = -0.05.$$

p.208

The result obtained is:

$$\underset{\approx}{P} = \begin{bmatrix} 3.219 & 5.152 & 5.128 & 3.185 & 0.996 \\ & 11.592 & 13.482 & 9.338 & 3.218 \\ & & 17.597 & 13.463 & 5.176 \\ & & & 11.497 & 5.168 \\ & \text{(Symmetry)} & & & 3.209 \end{bmatrix}$$

ORIGINAL PAGE IS
OF POOR QUALITY

CONCLUSIONS

A new technique for expanding the matrix Riccati equation is established: From the given equation of the problem, groups of rules are formulated and based on these rules a computer program is written. It makes various well known numerical methods directly applicable to the problem.

REFERENCES

1. K. E. Kalman, Contribution to the theory of optimal control. *Bol. Soc. Math. Mexicana* 5, 102–119 (1960).
2. M. Athans and P. L. Falb, *Optimal Control*, pp. 756–793. McGraw-Hill, New York (1966).
3. A. P. Sage, *Optimum Systems Control*, pp. 244–254. Prentice Hall, New Jersey (1968).
4. L. S. Pontryagin, V. G. Doltyanskii, R. V. Gankrelidze and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes.* Interscience, New York (1962).
5. A. G. J. MacFarlane, An eigenvector solution to the optimal linear regulator problem. *J. Electron. Control* 14 (6), 643–654 (June 1963).
6. J. E. Potter, Matrix quadratic solution, *SIAM J. Appl. Math.* 14 (3), 496–501 (May 1966).
7. D. R. Vaughan, A negative exponential solution for the linear optimal regulator problems. *Peprint of JACC*, 717–725 (June 1968).
8. A. F. Fath, Computational aspects of the linear optimal regulator problem. *Preprint of JACC*, 44–49 (Aug. 1969).
9. F. T. Man, The Davidon method of solution of the algebraic matrix Riccati Equation, *Int. J. Control* 10 (6), 713–719 (Dec. 1969).
10. C. F. Chen and L. S. Shieh, A note on expanding $PA + A^TP = -Q$, *IEEE Trans. Autom. Control* AC-13 (1), 122–123 (Feb. 1968).
11. A. T. Fuller, In-the-large stability of relay and saturating systems with linear controllers. *Int. J. Control* 10 (4) (Oct. 1969).
12. C. F. Chen and I. J. Haas, *Elements of Control Systems Analysis.* Prentice Hall, Englewood Cliffs, N.J. (1968).


```
C                SOLUTION  TO  RICCATI  EQUATIONS
C      P(DOT)=-P*A-A(TRANSPOSE)*P-C+P*B*R(INVERSE)*B(TRANSPOSE)*P
C      N*N  DIMENSION OF  A  OR  Q  OR  P  OR  PIN  MATRIX
C      N*M  DIMENSION OF  B  MATRIX,  M*M  DIMENSION OF  R  MATRIX
C      NN... NO. OF PLOT POINTS.   TN... STARTING TIME
C              MAIN PROGRAM BEGINS HERE
       DIMENSION PN(55),P(55),QQ(55,5),PDOT(55),A(10,10),B(10,10),
      1R(10,10),Q(10,10),PP(55),C(10,10),PP1(55),QQQ(55),RR(10,10),
      2PIN(10,10),PP2(55),D(10,10),E(10,10)
 1000 READ(5,500) N,M,NN,TN,DELT
  500 FORMAT(3I5,2F10.3)
      WRITE(6,600) N,M,NN,TN,DELT
  600 FORMAT(4HN...,I3,6H  M...,I3,7H  NN...,I3,7H  TN...,F16.6,
     19H  DELT...,F16.6)
      WRITE (6,610)
  610 FORMAT(//22HA...N*N  SYSTEM MATRIX/)
      DO 10 I=1,N
      READ(5,510) (A(I,J),J=1,N)
   10 WRITE(6,631) (A(I,J),J=1,N)
  510 FORMAT(4F20.6)
      WRITE(6,611)
  611 FORMAT(//21HB...N*M  INPUT MATRIX/)
      DO 11 I=1,N
      READ(5,510) (B(I,J),J=1,M)
   11 WRITE(6,631) (B(I,J),J=1,M)
      WRITE(6,612)
  612 FORMAT(//25HQ...N*N  SYMMETRIC MATRIX/)
      DO 12 I=1,N
      READ(5,510) (Q(I,J),J=1,N)
   12 WRITE(6,631) (Q(I,J),J=1,N)
      WRITE(6,613)
  613 FORMAT(//25HR...M*M  SYMMETRIC MATRIX/)
      DO 13 I=1,M
      READ(5,510) (R(I,J),J=1,M)
   13 WRITE(6,631) (R(I,J),J=1,M)
      WRITE(6,614)
  614 FORMAT(//35HPIN...N*N  INITIAL SYMMETRIC MATRIX/)
      DO 14 I=1,N
      READ(5,510) (PIN(I,J),J=1,N)
   14 WRITE(6,631) (PIN(I,J),J=1,N)
```

```
      NR=N*(N+1)/2
      LL=0
      DO 20 K=1,N
      DO 20 J=K,N
      LL=LL+1
      PN(LL)=PIN(K,J)
   20 QQQ(LL)=Q(K,J)
      IF (M.GT.1) GO TO 21
      R(1,1)=1./R(1,1)
      GO TO 22
   21 CALL INVER (R,M,RR,0,DET)
   22 CALL MALTP(N,M,B,M, R,C)
      DO 150 K=1,N
      DO 150 J=1,M
  150 E(J,K)=B(K,J)
      CALL MALTP(N,M,C,N,E,D)
      NNP=0
      DO 30 I=1,NR
   30 PDOT(I)=0.
      WRITE(6,630)
  630 FORMAT(/9X,1HT,12X,2HP1,11X,2HP2,11X,2HP3,11X,2HP4,11X,2HP5,11X,
     12HP6,11X,2HP7)
      WRITE(6,631) TN,(PN(I),I=1,NR)
  631 FORMAT(1X,9E13.5)
    1 L=1
      T=TN
      DO 40 K=1,NR
   40 P(K)=PN(K)
      GO TO 101
  100 DO 50 K=1,NR
   50 QQ(K,L)=DELT*PDOT(K)
      T=TN+DELT/2.
      DO 60 K=1,NR
   60 P(K)=PN(K)+QQ(K,L)/2.
      L=2
      GO TO 101
  200 DO 70 K=1,NR
   70 QQ(K,L)=DELT*PDOT(K)
      T=TN+DELT/2.
      DO 80 K=1,NR
   80 P(K)=PN(K)+QQ(K,L)/2.
      L=3
      GO TO 101
  300 DO 90 K=1,NR
   90 QC(K,L)=DELT*PDOT(K)
      T=TN+DELT
      DO 110 K=1,NR
  110 P(K)=PN(K)+QQ(K,L)
      L=4
      GO TO 101
  400 DO 120 K=1,NR
  120 QQ(K,L)=DELT*PDOT(K)
      GO TO 7
  101 CONTINUE
      CALL PA (N,A,NR,P,PP1)
      KL=0
      JJ=0
      MN=0
      DO 130 K7=1,N
      DO 130 L7=K7,N
      KL=KL+1
      LN=L7
      DO 131 M7=K7,N
      DO 131 N7=LN,N
      MN=MN+1
      JJ=JJ+1
      IF (N7.EQ.N) LN=M7+1
  131 PP(JJ)=P(KL)*P(MN)
  130 MN=KL
      CALL PRP(N,D,NR,PP,PP2)
      DO 160 K=1,NR
  160 PDOT(K)=PP2(K)-PP1(K)-QQQ(K)
      GO TO (100,200,300,400),L
    7 TN=TN+DELT
      DO 170 K=1,NR
  170 PN(K)=PN(K)+(1./6.)*(QQ(K,1)+2.*QQ(K,2)+2.*QQ(K,3)+QQ(K,4))
      WRITE(6,631) TN,(PN(K),K=1,NR)
      NNP=NNP+1
      IF (NNP.GT.NN) GO TO 1000
      GO TO 1
      END


      SUBROUTINE INVER  (A,N,B,M,DET)
      DIMENSION A(10,10),B(10,10),IPVOT(10),INDEX(10,2),PIVOT(10)
      EQUIVALENCE (IROW,JROW),(ICOL,JCOL)
   57 DET=1.
      DO 17 J=1,N
```

p.210

ORIGINAL PAGE IS
OF POOR QUALITY

```
 17 IPVOT(J)=0
    DO 135 I=1,N
    T=0.
    DO 9 J=1,N
    IF(IPVOT(J)-1) 13,9,13
 13 DO 23 K=1,N
    IF(IPVOT(K)-1) 43,23,81
 43 IF (ABS(T)-ABS(A(J,K))) 83,23,23
 83 IROW=J
    ICOL=K
    T=A(J,K)
 23 CONTINUE
  9 CONTINUE
        IPVOT(ICOL)=IPVOT(ICOL)+1
    IF(IROW-ICOL) 73,109,73
 73 DET=-DET
    DO 12 L=1,N
    T=A(IROW,L)
    A(IROW,L)=A(ICOL,L)
 12 A(ICOL,L)=T
    IF(M) 109,109,33
 33 DO 2 L=1,M
    T=B(IROW,L)
    B(IROW,L)=B(ICOL,L)
  2 B(ICOL,L)=T
109 INDEX(I,1)=IROW
    INDEX(I,2)=ICOL
    PIVOT(I)=A(ICOL,ICOL)
    DET=DET*PIVOT(I)
    A(ICOL,ICOL)=1.
    DO 205 L=1,N
205 A(ICOL,L)=A(ICOL,L)/PIVOT(I)
    IF(M) 347,347,66
 66 DO 52 L=1,M
 52 B(ICOL,L)=B(ICOL,L)/PIVOT(I)
347 DO 134 LI=1,N
    IF (LI-ICOL) 21,134,21
 21 T=A(LI,ICOL)
    A(LI,ICOL)=0.
    DO 89 L=1,N
 89 A(LI,L)=A(LI,L)-A(ICOL,L)*T
    IF(M) 134,134,18
 18 DO 68 L=1,M
 68 B(LI,L)=B(LI,L)-B(ICOL,L)*T
134 CONTINUE
135 CONTINUE
222 DO 3 I=1,N
    L=N-I+1
    IF(INDEX(L,1)-INDEX(L,2)) 19,3,19
 19 JROW=INDEX(L,1)
    JCOL=INDEX(L,2)
    DO 549 K=1,N
    T=A(K,JROW)
    A(K,JROW)=A(K,JCOL)
    A(K,JCOL)=T
549 CONTINUE
  3 CONTINUE
 81 RETURN
    END


    SUBROUTINE PA(N,A,NI,X,PP1)
    DIMENSION A(10,10),X(55),PP1(55),P(1 ,55)
    M=1
    KK=0
    DO 10 I=1,N
    DO 10 J=I,N
    KK=KK+1
    LM=0
    DO 20 K=1,N
    DO 20 L=K,N
    LM=LM+1
    IF (K.EQ.I.AND.L.NE.J) GO TO 21
    IF (K.NE.I.AND.L.EQ.J) GO TO 22
    IF (K.NE.I.AND.L.NE.J.AND.K.EQ.J.AND.L.NE.I) GO TO 23
    IF (K.NE.I.AND.L.NE.J.AND.K.NE.J.AND.L.EQ.I) GO TO 24
    IF (K.NE.I.AND.L.NE.J.AND.K.NE.J.AND.L.NE.I) GO TO 25
    IF (K.EQ.I.AND.L.EQ.J.AND.K.EQ.J.AND.L.EQ.I)  GO TO 26
    IF (K.EQ.I.AND.L.EQ.J.AND.K.NE.J.AND.L.NE.I) GO TO 27
 21 P(M,LM)=A(L,J)
    IF (I.EQ.J) P(M,LM)=2.*P(M,LM)
    GO TO 20
 22 P(M,LM)=A(K,I)
    IF (I.EQ.J) P(M,LM)=2.*P(M,LM)
    GO TO 20
 23 P(M,LM)=A(L,I)
    IF (I.EQ.J) P(M,LM)=2.*P(M,LM)
    GO TO 20
 24 P(M,LM)=A(K,J)
```

```
      IF (I.EQ.J) P(M,LM)=2.*P(M,LM)
      GO TO 20
   25 P(M,LM)=0.
      GO TO 20
   26 P(M,LM)=A(K,I)
      IF (I.EQ.J) P(M,LM)=2.*P(M,LM)
      GO TO 20
   27 P(M,LM)=A(K,I)+A(L,J)
      IF (I.EQ.J) P(M,LM)=2.*P(M,LM)
   20 CONTINUE
      S=0.
      DO 31 JJ=1,NI
   31 S=P(M,JJ)*X(JJ)+S
      PP1(KK)=S
   10 CONTINUE
      RETURN
      END


      SUBROUTINE PRP(NN,PBRBP,NR,PP,PP2)
      DIMENSION H(1 , 770),KN(2,2),PBRBP(10,10),PP(55),PP2(55)
      NC=NN*(NN+1)*(NN**2+NN+2)/8
      II=1
      KK=0
      JJ=0
      DO 20 I=1,NN
      DO 21 J=I,NN
      KK=KK+1
      DO 30 K=1,NN
      DO 31 L=K,NN
      LN=L
      DO 40 M=K,NN
      DO 41 N=LN,NN
      JJ=JJ+1
      IF (I.EQ.K) KN(1,1)=L
      IF (I.EQ.L) KN(1,1)=K
      IF (I.EQ.M) KN(1,2)=N
      IF (I.EQ.N) KN(1,2)=M
      IF (J.EQ.K) KN(2,1)=L
      IF (J.EQ.L) KN(2,1)=K
      IF (J.EQ.M) KN(2,2)=N
      IF (J.EQ.N) KN(2,2)=M
      IF (I.NE.K.AND.I.NE.L) KN(1,1)=0
      IF (I.NE.M.AND.I.NE.N) KN(1,2)=0
      IF (J.NE.K.AND.J.NE.L) KN(2,1)=0
      IF (J.NE.M.AND.J.NE.N) KN(2,2)=0
      IF((KN(1,1).EQ.0.OR.KN(2,2).EQ.0).AND.KN(1,2).NE.0.AND.KN(2,1).NE.
     10)   H(II,JJ)=PBRBP(KN(1,2),KN(2,1))
      IF((KN(1,2).EQ.0.OR.KN(2,1).EQ.0).AND.KN(1,1).NE.0.AND.KN(2,2).NE.
     10)   H(II,JJ)=PBRBP(KN(1,1),KN(2,2))
      IF((KN(1,1).EQ.0.OR.KN(2,2) .EQ.0).AND.(KN(1,2).EQ.0.OR.KN(2,1).
     1EQ.0)) H(II,JJ)=0.
      IF (KN(1,1).NE.0.AND.KN(2,2).NE.0.AND.KN(1,2).NE.0.AND.KN(2,1).
     1NE.0)  H(II,JJ)=PBRBP(KN(1,1),KN(2,2))+PBRBP (KN(1,2),KN(2,1))
      IF (K.EQ.M.AND.L.EQ.N) H(II,JJ)=H(II,JJ)/2.
      IF (N.EQ.NN) LN=M+1
   41 CONTINUE
   40 CONTINUE
   31 CONTINUE
   30 CONTINUE
      JJ=0
      S=0.
      DO 51 LL=1,NC
   51 S=H(II,LL)*PP(LL)+S
      PP2(KK)=S
   21 CONTINUE
   20 CONTINUE
      RETURN
      END


      SUBROUTINE MALTP(N,M,A,L,B,C)
      DIMENSION A(10,10),B(10,10),C(10,10)
      DO 10 I=1,N
      DO 10 J=1,L
      S=0.
      DO 10 K=1,M
      S=S+A(I,K)*B(K,J)
   10 C(I,J)=S
      RETURN
      END
```

PRECEDING PAGE BLANK NOT FILMED

(D) Year: 1973

(E) Department: Mechanical Engineering

(F) Student Name: Norman C. Martin

(G) Faculty Advisor: Professor B. D. Cook

Project

(A) Project Title: On the Stability of Poiseuille Pipe Flow

(B) Project Abstract:

The problem of the stability of Poiseuille pipe flow was studied numerically. The finite-difference equations which were solved are approximations to the nonlinear, axisymmetric, Navier-Stokes equations in cylindrical coordinates subject to a stream function perturbation. The disturbance to the stream function which was used is axisymmetric, oscillatory and fixed in space. The resulting solutions show the experimentally observed instability of the stream function and vorticity at Reynolds numbers of 10,000 and 100,000. The experimentally observed stability at a Reynolds number of 1,000 is also found.

(C) Publication: Ph.D. Dissertation in Mechanical Engineering

(D) Year: 1969

(E) Department: Mechanical Engineering

(F) Student Name: Henry Johnston Crowder

(G) Faculty Advisor: Prof. Charles Dalton

Project

(A) Project Title: The Fluid Resistance of Shrouded and Unshrounded Circular Cylinders in an Oscillatory Flow

(B) Project Abstract:

This investigation concerns itself with the measurement of a strain-gage signal responding to a sinusidal variation of tension and compression brought about by a circular cylinder oscillating with simple harmonic motion in a tank of water, otherwise at rest. The strain-gage is transformed into the fluid resistance acting on circular cylinders.

A series of plots of force coefficient versus Reynolds Number were developed to gain a better understanding of the phenomenon. The experimental apparatus is analogous to the distribution of wave forces present on a single fixed leg of an offshore structure.

The fundamental variables involved were the diameter of the cylinder (0.625-, 1.0, 1.5-inch) and amplitude and speed of oscillation, one- to six-inches and 10 to 0.60 rpm, respectively.

The effects on the fluid resistance acting on the two smaller cylinders enclosed in a concentric perforated shroud of the same outside diameter as the largest cylinder

were determined relative to the fluid resistance acting on the largest unshrouded cylinder. The effects of the shroud were investigated to ascertain the feasibility of reducing wave forces on the structural members of an offshore platform.

(C) Publication: M.S. Thesis in Mechanical Engineering

(D) Year: 1973

(E) Department: Mechanical Engineering

(F) Student Name: John P. Hunt

(G) Faculty Advisor: Prof. Charles Dalton

Project

(A) Project Title: Model Investigation: Effects of a Reef on Ocean Waves

(B) Project Abstract:

An obstruction in the path of a water wave will change the profile, energy and forces of the wave. In some case it is difficult to tell exactly what effect a particular geometry will have. A model was constructed for use in a wave tank to investigate a section of the flower gardens for a possible site of an offshore platform. A series of waves varying in height and period were used. It was found that the reef increased the wave forces due to increased mass transport and shoaling. The decision was made to choose a different location.

(C) Publication: M.S. Thesis in Mechanical Engineering

(D) Year: 1973

(E) Department: Mechanical Engineering

(F) Student Name: Paul G. Johnson

(G) Faculty Advisor: Prof. Charles Dalton

Project

(A) Project Title: A Four-Equation Model for Numerical Solution of the Turbulent Boundary Layer

(B) Project Abstract:

In this study, a four-equation model of turbulence is presented in order to solve the steady, incompressible two-dimensional turbulent boundary layer flow field. Those four equations are the continuity, momentum, turbulent kinetic energy, and the rate of dissipation equations. The solution is obtained in terms of the mean variables of the flow field.

Closure is obtained by assuming for each equation that all those terms containing fluctuating quantities are somehow related to the mean variables of the mean flow field.

A new two-layer eddy viscosity model is used. Near the wall, the eddy viscosity is assumed to be proportional to the second power of the vertical coordinate. For the outer layer, it is assumed that it is proportional to the ratio of the square of the turbulent kinetic energy to the rate of dissipation.

The model was solved numerically by a finite-difference technique, using a variables mesh system in order to have small increments near the wall. An implicit numerical procedure was used in order to speed the computation in the

downstream direction.

The model was applied to the computation of the incompressible turbulent boundary layer over a flat plate, and agreement with the Wieghardt data is excellent. Three cases with varying pressure gradients are also computed; they are based on the Ludwieg and Tilmann data. Those three cases were chosen because they are for boundary layers on flat surfaces with different pressure gradients, mild adverse, strong adverse, and favorable pressure gradient flow.

The calculation procedure requires that starting profiles be known for $\bar{u}$, $\bar{g}$, and $\bar{D}$. These starting profiles are obtained from Ludwieg and Tilmann. The eddy viscosity model at a given station in the outer region is based on the calculated values of $\bar{g}$ and $\bar{D}$ at the previous station.

(C) Publication: Ph.D. Dissertation in Mechanical Engineering.

(D) Year: 1974.

(E) Department: Mechanical Engineering

(F) Student Name: Hyppolito de Valle Pereira Filho

(G) Faculty Advisor: Prof. Charles Dalton

Project

(A) Project Title: A Nonlinear Diffraction Study of Inertia Forces on a Vertical Circular Cylinder

(B) Project Abstract:

In this study the inertia coefficient for a vertical circular cylinder subject to wave action in a finite depth of water is analyzed with respect to its independent variables. The primary purpose of the investigation is to develop criteria for choosing values of the inertia coefficient for use in wave force calculations. This is done keeping in mind that changes in wave height, wave period, water depth, elevation, phase, and cylinder diameter will effect the coefficient chosen.

The analysis begins by assuming an inviscid, irrotational wave of small amplitude which can be described by linear wave theory. The wave is allowed to diffract around a vertical circular cylinder. The pressure distribution around the cylinder is calculated by application of the complex potential for the incident and scattered wave systems. The force per unit length on the cylinder is determined by integration of the pressure around the cylinder. The expression for the inertia coefficient is then derived from the known water particle acceleration and wave force equations. Drag forces are not considered in this investi-

gation. The resulting equation for the inertia coefficient is applicable to cylinder and wave combinations with large diameter to wave length ratios. This provides an improvement to an existing diffraction study. The nonlinear velocity terms in the pressure expression are retained, thus providing still another modification to existing diffraction theory.

The results from this investigation indicate that the inertia coefficient is mainly a function of the ratio of cylinder diameter to wave length. Contributions made by the influence of depth and wave height are small compared to that made by the diameter to wave length ratio.

(C) Publication: M.S. Thesis in Mechanical Engineering

(D) Year: 1975

(E) Department: Mechanical Engineering

(F) Student Name: Jerry Lee Borrer

(G) Faculty Advisor: Prof. Charles Dalton

Project

(A) Project Title: Numerical Solutions for Recirculating Flow

(B) Project Abstract:

Numerical solutions have been obtained for the steady two-dimensional flow of a viscous incompressible fluid in rectangular cavities by solving various implicit finite-difference approximations of the Navier-Stokes equations. The sets of implicit difference equations were solved using a recently introduced iterative numerical procedure called the strongly implicit procedure (SIP). The strongly implicit procedure was found to be an effective and economical numerical procedure for obtaining iterative numerical solutions for sets of linear and/or nonlinear finite-difference equations. Various qualitative and quantitative comparisons have been made to determine the effects of the Reynolds number, the grid size, and the difference approximation on the numerical solutions and computational procedures. The results show that one difference scheme was particularly accurate for Reynolds numbers and grid sizes which satisfy the stability restriction. Streamline patterns and equivorticity plots for a range of Reynolds numbers are shown for various difference approximations and grid sizes. Changes in the principal features of the flow field have been discussed

11-6

and correlated with changes in the Reynolds number of the flow.

(C) Publication: M.S. Thesis in Mechanical Engineering

(D) Year: 1975

(E) Department: Mechanical Engineering

(F) Student Name: Jerry Lee Borrer

(G) Faculty Advisor: Prof. Charles Dalton

Project

(A) Project Title: The Effect of Inclination of a Conduit on Power Spectra of Wall Pressure Fluctuations in Two-Phase Flow

(B) Project Abstract:

The two-phase flow regime characterization by use of wall pressure fluctuation power spectra was proposed by Dr. M. G. Hubbard and Dr. A. E. Dukler to deal with horizontal flow. In this work, the power spectra of various flow regimes of two-phase flow with different inclinations were obtained and analyzed. The effect of inclination of a conduit on power spectra and flow mechanisms was investigated. Also a method of decomposion of power spectra was proposed. The results suggested that a two-phase flow process could be decomposed into several simple processes each of which was represented by an individual spectral component.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1971

(E) Department: Chemical Engineering

(F) Student Name: Jeng-Shong Liaw

(G) Faculty Advisor: Professor A. E. Dukler

p. 225

Project

(A) Project Title: Hybrid Computer Simulation of Turbulent Diffusion in the Atmosphere by Monte Carlo Methods

(B) Project Abstract:

Turbulent diffusion in the atmosphere was simulated by implementing a new Monte Carlo method on a hybrid computer. The new method involved the development of a stochastic Langevin equation which required the instantaneous wind velocity as input information to stimulate the diffusion process.

Although several established models were available for the mean wind profile, there were no models available for the fluctuating component of the velocity. Thus a model was developed by using empirical equations to describe the rms value as a function of position and by using independent Gaussian white noises of proper frequency range and power spectral densities.

The present method was evaluated by comparing the results to the theoretical dispersion in a homogeneous flow and to experimental concentration profiles in a boundary layer and in the atmosphere. All of the experimental flow fields were nonhomogeneous. Good agreement was found in all cases. The simulated concentration distributions were found to have a 95% statistical reliability by a chi-square

goodness-of-fit test.

A few of the major advantages of the present method are (1) since the current method simulates the diffusion process directly, it has great flexibility and the concept of eddy diffusion coefficients is not used, and (2) esentially all meteorological effects can be fully utilized. The present method is also applicable to multiple sources of almost any type.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1975

(E) Department: Chemical Engineering

(F) Student Name: Jerry A. Bullin

(G) Faculty Advisor: Professor A. E. Dukler

Project

(A) Project Title: Studies on Turbulent Diffusion

(B) Project Abstract:

Turbulent diffusion was studied both theoretically and experimentally. The purposes of this study were (1) to develop and test a new statistical model for turbulent diffusion which was reasonable in both theory and implementation, (2) to include the effect of shear stress on diffusion, (3) to extend the new model to a particulate system and (4) to obtain experimental data for particle dispersion to test the proposed model.

A Langevin equation model was developed by considering the fluctuating velocity as a stochastic process. The same model was also derived from a one dimensional lagrangian Navier-Stokew equation. This model was physically realistic.

The present model was implemented on a hybrid computer. The simulated results of turbulent diffusion were compared with the theoretical predictions for a homogeneous flow and with experimental concentration profiles in a boundary layer and in the atmosphere. Good agreement was found in all cases.

A technique to generate two random processes which are correlated with each other to any degree was developed. This method was used to investigate the shear effect on turbulent diffusion. In the presence of both mean velocity

gradient and shear, diffusion was shown to be independent of shear for long and short diffusion times but to be strongly dependent on shear for intermediate times.

The langevin equation model was extended to permit the modeling of particle dispersion. This was accomplished by deriving a relation between the power spectral density distributions of a particle and the background fluid from the equation of motion of a particle in a turbulent flow.

A series of experiments was designed and executed in a large wind tunnel using glass beads as diffusing particles. A solid feeder and a particle dispenser were built and a special sampling nozzle was designed for isokinetic sampling. Hot wire equipment was used to measure the fluid dynamical properties.

The conditions of the experiment were simulated using the developed model. Agreement with experiment was good.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1976

(E) Department: Department of Chemical Engineering

(F) Student Name: Naugab Lee

(G) Faculty Advisor: Professor A. E. Dukler

Project

(A) Project Title: On the Transmission of Sound Through Finite, Closed Shells: Statistical Energy Analysis, Modal Coupling, and Nonresonant Transmission

(B) Project Abstract:

An investigation of sound transmission into small enclosures considers the effects of acoustically induced coupling between shell modes. Using an integral equation approach, the transmission loss into a small rectangular box is computed and the level of cavity reactance examined. The noise reductions for a closed cylinder and a rectangular parallelopiped enclosure (with a single flexible panel) sitting in reverberant acoustic fields are computed and experimentally checked. The transmission by resonant and nonresonant shell modes is examined, especially in relation to the statistical energy analysis approach. The nature of the predominance of one type of transmission over the other is considered in relation to shell and cavity configurations and structural damping levels. A technique is given for estimating the predominance of resonant or nonresonant modes as the basis for computations of sound transmission by flat panels and cylindrical shells in reverberant acoustic fields.

(C) Publication: Ph.D. Dissertation in Mechanical Engineering

(D) Year: 1970

(E) Department: Mechanical Engineering

(F) Student Name: Larry Debbs Pope

(G) Faculty Advisor: Prof. R.D. Finch

Project

(A) Project Title: Feasibility Study of Flaw Detection in Railway Wheels Using Acoustic Signatures

(B) Project Abstract:

The research is a feasibility study on the use of acoustic signature for detection of flaws in railway wheels. The ultimate objective is the development of a wayside device capable of indicating defective wheels on moving cars or locomotives. A test stand was constructed so that wheels could be excited by random noise under simulated loads, and by impacting with various devices. Analytical and experimental determinations of the natural modes of vibrating wheels are reported. Differences in acoustic signature were found between good and flawed whells, including spectral changes and variations in the time decay of sound. Pattern recognition techniques were used for selecting good and bad wheels with a data processing scheme using a minicomputer.

(C) Publication: Ph.D. Dissertation in Mechanical Engineering

(D) Year: 1974

(E) Department: Mechanical Engineering

(F) Student Name: Kornel Nagy

(G) Faculty Advisor: Prof. R.D. Finch

Project

(A) Project Title: Respiratory System Dynamics

(B) Project Abstract:

The understanding of physiological mechanisms, involved in the dynamics of respiratory adjustments, and their quantitative representation is of major concern in physiology and medicine. A necessary prelude to any analysis of respiratory regulation is the construction of a respiratory plant model which accurately mimics changes in the various state variables following an external disturbance imposed on the system. Most investigators in the past have attempted to uncover the control hierarchies inherent in respiratory regulation without assessing the validity of their plant representations. As a result relatively little has been accomplished in the delineation of the actual transport processes within blooc and tissue fluids.

This report describes a first attempt at a detailed examination of the various aspects involved in the adjustment of O_2 and CO_2 stores as well as the maintenance of acid base balance in the body. Our approach is a pyramidal one focusing first on the different subsystems themselves, using experimental data wherever available to assess the representation, and then examining the over all plant model predictions.

To assess the general validity of the blood gas inter-

actions and relationships used in the study, we derive a theoretical CO_2 dissociation curve for blooc in vitro and compare its predictions with experimental data. The kinetics of blood gas transport and reaction processes are considered in the analysis of pulmonary gas exchange. Several different alternative transport hypothesis are implemented in our description of the muscle subsystem and appropriate conclusions drawn using open loop experimental data for muscle tissue. A multicompartmental description is presented for the brain which allows examination of the various transport hypotehsis presented in the physiological literature. Finally, by specifying the necessary elements for the entire closed loop flow path, simulation studies are presented for the overall respiratory plant under open control loop conditions.

Good agreement is obtained with the available experimental data for the individual elements as well as the overall plant. Needless to say, this study does not constitute the last word in analyses of the respiratory system but the more basic approach and evolutionary framework used here, hopefully represents an important step toward a meaningful description of respiratory system dynamics.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1975

(E) Department: Chemical Engineering

(F) Student Name: Akhil Bidani

(G) Faculty Advisor: Professor R. W. Flumerfelt

Project

(A) Project Title: A Transport and Thermodynamic Properties Package for Simulation and Design

(B) Project Abstract:

Transport and thermodynamic properties such as viscosity, thermal conductivity, enthalpy, etc., are frequently needed in chemical process computations. A computer program was written which calculates these and many other related properties rapidly and efficiently.

The system consists of an executor or main subroutine called ROADMP plus 36 subroutines. ROADMP decides what subroutines have to be used to calculate a given property from input data provided by the user.

All the subroutines, except ROADMP, consist of small programs that use physical property correlations taken from the literature. Extensive tests and error checks were made and are included, where possible, in the system.

The entire program consists of 3,200 FORTRAN statements and occupies 11K words of memory.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1969

(E) Department: Chemical Engineering

(F) Student Name: Roberto Mariano Beirute

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: Multicomponent Vapor-Liquid Equilibria Computation

(B) Project Abstract:

This thesis is concerned with the development of a comprehensive computer program for the estimation of multicomponent vapor-liquid equilibria in two phase systems at low to moderate pressures. The system which has been developed is referred to as "the KVALUE routines" and it fulfills its intended purpose in that it constitutes a versatile, convenient, and efficient vapor-liquid equilibria package which is suitable for use either alone or in conjunction with other computer programs. The KVALUE routines consist of approximately 2,000 FORTRAN instructions which require 21,454 words of storage when loaded into an IBM 360 Model 44 computer.

The KVALUE routines exhibit a number of very desirable features. The most desirable feature of the system lies in its versatility. Multiple routes leading to each of the major thermodynamic and physical properties allow the user a considerable degree of freedom with respect to the type and quantity of input data which he may provide. The system structure, and the data structure as well, are designed so as to make the package conveniently integrable with other computing systems. An auxiliary program, CURFIT,

which is included in the package, should serve as a valuable aid to the user in performing tasks related to data preparation and organization.

The techniques used are such that the computations are based upon a sound thermodynamic approach. The text contains a discussion of the theory which is incorporated in the estimation techniques that constitute the various algorithms. Included in this discussion are the following topics: the thermodynamic criteria for equilibrium, the computation of vapor phase fugacity coefficients using both the virial and the Redlich-Kwong equations of state, the correlation of activity coefficients to liquid phase composition via the equations of Wilson, Van Laar, and Hildebrand, the calculation of standard-state fugacities from thermodynamic considerations and from the Chao-Seader correlation, and techniques for evaluating the necessary physical properties and parameters. Also included in the text is a discussion of Marquardt's method and the manner in which this algorithmic process is implemented to obtain a solution to the system of nonlinear equations which must be solved when estimating multicomponent vapor-liquid equilibria.

The KVALUE routines have been thoroughly tested and have been shown to produce reliable results in most cases. Several examples are included in the text to illustrate the applicability of the KVALUE routines in typical situations and indicate the quality of the results obtainable.

The appendices include information pertaining to the use of the KVALUE routines as well as schematic diagrams, listings, and documentations for the various subroutines. Also included in the appendices are computer output listings for the example problems which are discussed in the text.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1969

(E) Department: Chemical Engineering

(F) Student Name: Raymond Alan Williams

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: Optimization by Signal Flow Graph Method

(B) Project Abstract:

Five related objectives were realized in this research:

1. A modified linear programming technique was developed. The procedure follows the simplex algorithm, but signal flow graph (SFG) methods rather than matrix manipulations are used.
2. It is shown that all the ordinary post-optimum analysis, including sensitivity analysis, may be performed by the SFG method using the final graph instead of the simplex final tableau.
3. The SFG methods to solve linear equations were used to obtain the gradient vectors of objective functions with equality constraints.
4. The techniques developed in 1 and 2 were incorporated into the method of "feasible direction" (MFD), one of the most powerful methods of constrained optimization.
5. Two large, nonlinear, heat exchanger networks were studied, and the total heat exchanger area was minimized using the MFD in conjunction with the LP system developed.

This research points to the possibility that signal flow graph methods can be a useful tool for solving linear and nonlinear constrained optimization problems.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1971

(E) Department: Chemical Engineering

(F) Student Name: Takashi Tonomura

(G) Faculty Advisor: Professor E. J. Henley

p. 240

Project

(A) Project Title: Application of a Generalized Urban Model to a Specific Region

(B) Project Abstract:

The purpose of the research was to take the generalized urban model of Professor J. W. Forrester and apply it to a specific area, namely Harris County, Texas. The only variables changed were those considered to be "region dependent." The model was initialized with data from the year 1950. Statistical data for 1960 and 1970 was used for checking the validity of the model. The statistical data used was obtained from the 1950, 1960, and 1970 census, along with data from various planning agencies in the county.

The results revealed that the model could undoubtedly be used as a planning tool for a specific region, as the final model appeared closely tuned to the major statistical variables. The lack of variables in the model which could be directly correlated with statistical data was thought to be the reason for any fluctuations in the subvariables. Sensitivity of the model to a large number of variables was a byproduct of the research.

(C) Publication: M.S. Thesis in Mechanical Engineering

(D) Year: 1971

(E) Department: Mechanical Engineering

(F) Student Name: Howell R. Porter, III

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: Systematic Flow Graph Analysis and Applications

(B) Project Abstract:

A generalized technique for the solution of engineering problems which can be represented in the form of flow graphs has been developed. The fundamental algorithm is one whereby the paths and loops in a flow graph can be systematically and selectively enumerated. This algorithm serves as the basis for a procedure whereby Mason's rule can be efficiently applied to generate characteristic equations, system determinants, transfer functions defining input-output relationships, sensitivity functions, and other important network functions related to signal flow graphs.

The theory developed provides the basis for a comprehensive computing system which is instrumental in solving many types of flow graph problems. The value of flow graph analysis in engineering science, and the diversified utility of the techniques developed herein are illustrated by six example problems: sensitivity analysis of a heat exchanger network, simulation and analysis of a chemical reactor control system, generation of closed-form expressions describing the steady-state performance of an absorption column, ordering of recycle calculations for a chemical process

simulation, computation of eigenvalues, and the solution of a typical transportation problem.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1971

(E) Department: Chemical Engineering

(F) Student Name: Raymond Alan Williams

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: Numerical Integration of Stiff, Sensitive and Multivalued Equations

(B) Project Abstract:

A technique for the numerical integration of stiff and multivalued ordinary differential equations has been developed. Any of the standard numerical integration methods (i.e., Runge Kutta, Adams-Moulton, etc.) may be employed. The technique utilizes a changeable independent variable of integration to conquer the numerical difficulties usually encountered in the integration of certain equations. Application of the method to several problems of interest to chemical engineers was made. In general, the technique works to increase accuracy and efficiency of solution.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1971

(E) Department: Chemical Engineering

(F) Student Name: Patrick Hugh Shannon

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: Simulation of Multicomponent Separation by an Adiabatic Cascade Technique

(B) Project Abstract:

Advances in the ability to estimate thermodynamic physical properties and the demand for solutions to difficult multistage separation problems have warranted further research to develop suitable computer algorithms.

Use of composition-dependent vapor-liquid equilibrium ratios and enthalpies preclude the utilization of all but the most recently developed algorithms and as yet no current technique is widely accepted as were the historic Thiele-Geddes and Lewis-Matheson algorithms. Additionally, no algorithm is able to solve all separation problems with a single technique. Thus, stability of convergence on classes of problems or specific examples, and inability to use composition dependent properties are major areas for improvement.

The proposed algorithm uses a complete Chao-Seader physical properties package as modified by Grayson-Streed. Liquid phase z-factors are computed by Yen-Woods coefficients.

Two major causes for instabilities in convergence are numerical round off error and structural ordering of equations for solution via successive approximation techniques. Both of these problems are largely eliminated in the

proposed algorithm.

Inherent instabilities of other algorithms due to structural arrangement are avoided by computing physical output quantities from physical inputs and round off is avoided by requiring the outputs to sum to the inputs.

A unique first order acceleration scheme was developed that is stable for absorption, reboiler absorption and distillation problems.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1973

(E) Department: Chemical Engineering

(F) Student Name: Larry J. McNiel

(G) Faculty Advisor: Professor E. J. Henley

p. 247

Project

(A) Project Title: Reliability Optimization of Process Systems Using Intermediate Storage Tanks

(B) Project Abstract:

The purpose of this study is to create a model that will predict process system reliability when storage tanks are used as back-up units.

The model is nondeterministic and only requires a knowledge of the probability of failure and repair of each unit. It assumes that the tanks will fail according to a step function; when empty the probability of failure is unity and when full it is zero.

Based on this model a computer program for optimizing the size of intermediate storage tanks in a process system with and without recycle was developed. The method of Paviani, et al. was used to find the optimum for a process system composed of three units and three tanks.

It was found that the benefits of having tanks decreased from the end to the beginning of the system and that the last tank in the system plays such an important role that the optimum will generally be obtained by sizing the product tanks as large as possible.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1973

(E) Department: Chemical Engineering

(F) Student Name: Erwin Rosen

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: The Optimization of Urban Systems Objective Functions

(B) Project Abstract:

A generalized computer program has been developed which enables the representation of the interaction of social segments of an urban structure. This program is presented in detail such that it can be adapted to any particular urban system of interest. However, the adaptation presented herein is to Harris County, Texas. The calibration has been accomplished utilizing data extracted from magnetic tapes produced by the Census Bureau, U.S. Department of Commerce (1970 census). This model is the Housing Allocation and Location Optimization (HALO) model.

The model functions in a dynamic manner over a selected global analysis period. The assumption is made that the global analysis period may be represented by a series of discretized analysis periods. During each of these periods, a certain portion of the households is assumed to enter the market seeking to relocate. The model must satisfy a set of quantitative constraints pertinent to the particular geographical region of interest to achieve equilibrium.

Several features of this model are unique and provide significant improvement over previous models:

The tremendous problem of evaluating preference

factors by regression techniques is completely eliminated by the inclusion of generalized weighting factors.

A multi-dimensional array is incorporated to control the supply and demand of the housing market. This feature reflects the various degrees of dependency of construction upon the economic situation experienced by different types of housing units.

The capability of selecting an optimum location within the geographical region is provided. A generalized user specified objective function is optimized in the global analysis period. This facilitates the selection of optimal sites for the location of various installations.

After the model is calibrated to any area of interest, it can be used to evaluate various policies.

Two different policies are implemented in this paper. A new school district is created from an existing school district and the affect on the housing distribution is presented. The housing pattern change caused by the completion of a new freeway is also presented. In each case, the impact caused by the policy implementation is evident.

(C) Publication: Ph.D. Dissertation in Electrical Engineering

(D) Year: 1973

(E) Department: Electrical Engineering

(F) Student Name: Joe W. Pyle

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: Reliability Analysis and Optimization of Complex Systems

(B) Project Abstract:

Most of the earlier literature on system reliability optimization consider only the simple series-parallel systems subject to one or two constraints. Practical systems have complex rather than the simple series parallel configurations. With a view towards solving these complex system reliability optimization problems, an efficient computer algorithm based on the path enumeration method has been developed. An important feature of the method is the module representation of the reliability graph which considerably simplifies the calculation of reliability and sensitivity functions of complex systems. A modified integer gradient method is used for system optimization. Although the method does not insure a global optimum, it does find various near-optimum solutions. From a practical consideration, this could provide for a wider choice during the design phase.

In this research an effort has also been made to apply basic reliability concepts to process plants. A new formulation of the optimal reliability design of process plants which takes into account the quantitative aspects of systems throughout is proposed. It is based on the k-out-of-n

configuration instead of the conventional parallel redundancy configuration. The problem is so formulated that determining the optimum configuration also determines the optimum capacity of units to be used at each stage of the system. A computer program based on a pseudo-Boolean algorithm is used to solve this non-linear integer programming problem.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1973

(E) Department: Chemical Engineering

(F) Student Name: Satish Loonkaran Gandhi

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: Synthesis of Process Flowsheets by a Theorem Proving Method

(B) Project Abstract:

Events occuring in chemical processing systems are described in terms of an axiomatic second-order theory. The axioms of the theory are mixing, splitting, reaction, change of pressure, change of enthalpy and equality axiom.

Theorem proving method based upon the resolution principle is used to test the hypothesis that a required slate of products is a logical consequence of the axioms and available raw materials. The proof, if it exists, yields the process plant flowsheet.

The computer implementation of the procedure is in LISP 1.5 programming language. An example of flowsheet synthesis is given.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1974

(E) Department: Chemical Engineering

(F) Student Name: Vladimir Mahalec

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: A Study of Integration Algorithms in Chemical and Physiological System Dynamics

(B) Project Abstract:

Solution of many chemical engineering problems requires the use of numerical integration techniques. The most popular integration technique for such problems is the classical fourth order Runge-Kutta that in some cases can be used only with very low efficiency.

In the last few years new methods have been developed which seem to be efficient for regular and stiff problems. Some of these new methods were compared in the solution of chemical and physiological systems. The program written by C. W. Gear, which includes two slightly different algorithms for stiff systems and a third algorithm for regular systems was found to be the most stable and efficient in all cases.

To increase accessibility of the Gear program for general engineering usage, a set of subroutines was written. These subroutines were designed with the following objectives:

a) Minimization of input requirements.

b) Minimization of interaction between user and integration program.

c) Giving the user the option of calculating the values of the dependent variables at certain

specific values of the independent variable.

d) Giving the user the option of dynamically selecting the algorithm to solve the problem in question efficiently.

The behavior of the different algorithms in the solution of selected problems is also discussed.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1974

(E) Department: Chemical Engineering

(F) Student Name: Francisco S. Castellanos

(G) Faculty Advisor: Professor E. J. Henley

Project

(A) Project Title: Water Vapor, CO_2 and Particulate Effects On The Atmospheric Temperature Profile

(B) Project Abstract:

A new approach is developed for the determination of the atmospheric temperature profile. Various concentrations of carbon dioxide, water vapor, and scattering particles are introduced to determine the perturbing effect on the temperature distribution in the atmosphere.

The solution is gained through the combination of the Edwards exponential wide band model equations and the Curtis-Godson transformation. The Curtis-Godson technique allows a transformation from the equations for a nonisothermal medium to those for the isothermal case while the Edwards band model equations provide a means of finding the absorption characteristics of the carbon dioxide and water vapor. The basic radiation equation of transfer is coupled with an energy balance equation through the use of a wide band model derivative approximation which reduces the complexity of the analysis considerably.

An approach is developed for the inclusion of particle scattering. The simple case of elastic scattering is considered and the equations are treated in a manner which allows the wide band model derivative approximation to be utilized again.

The doubling of the carbon dioxide concentration produced a 1.89^{o}K increase in the ground level temperature. Halving the concentration caused a decrease of 1.94^{o}K in the ground level temperature. These results agree reasonably well with those obtained by other investigators. In addition, since only a 25% increase in the concentration of carbon dioxide is expected from AD 1900 to AD 2000, there is no threat of a significant ground level temperature change due to the increase of the carbon dioxide in the atmosphere. In fact, the introduction of water vapor and scattering can negate the effects of the carbon dioxide completely.

(C) Publication: Ph.D. Dissertation in Mechanical Engineering and Proceedings of the 1972 Heat Transfer and Fluid Mechanics Institute, page 146-162.

(D) Year: 1971

(E) Department: Mechanical Engineering

(F) Student Name: Ross E. Ferland

(G) Faculty Advisor: Professor J. R. Howell

Project

(A) Project Title: Thermal Modelling of a Plate with Coupled Heat-Transfer Modes

(B) Project Abstract:

The thermal system considered here is a flat plate subjected to all three modes of heat exchange. The plate also has a constant rate of volumetric internal energy generation. According to the method for thermal modelling proposed here, the volumetric energy generation rate in the model is varied in a manner depending upon its geometric reduction and other parameters. The model then has the same dimensionless temperature profile as does the prototype. The experiments conducted give confirmation of theoretical results obtained earlier and also give encouraging results for the proposed modelling procedure.

(C) Publication: Ph.D. Dissertation in Mechanical Engineering

(D) Year: 1972

(E) Department: Mechanical Engineering

(F) Student Name: Manohar S. Sohal

(G) Faculty Advisor: Professor J.R. Howell

THERMAL MODELING OF A PLATE WITH COUPLED HEAT-TRANSFER MODES

Manohar S. Sohal* and John R. Howell†

University of Houston, Houston, Texas

Abstract

The thermal system considered here is a flat plate subjected to all three modes of heat exchange. The plate also has a constant rate of volumetric internal energy generation. According to the method for thermal modeling proposed here, the volumetric energy generation rate in the model is varied in a manner depending upon its geometric reduction and other parameters. The model then has the same dimensionless temperature profile as does the prototype. The experiments conducted give confirmation of theoretical results obtained earlier and also give encouraging results for the proposed modeling procedure.

Nomenclature

C_l = constant for a laminar flow, $0.623\ Pr^{-1/3}\ Re_L^{-1/2}$
C_t = constant for a turbulent flow, $3.323\ Pr^{-0.6}\ Re_L^{-0.8}$
I = current flowing in the plate
i = 1, 2, 3, ... n + 1; $i \leq j$
j = 1, 2, 3, ... n + 1
k_f = thermal conductivity of the fluid

Presented as paper 73-748 at AIAA 8th Thermophysics Conference, Palm Springs, Calif., July 16-18, 1973.

*Research Fellow, Dept. of Thermodynamics and Fluid Mechanics, University of Strathclyde, Glasgow, Scotland.

†Professor, Department of Mechanical Engineering.

k_p = thermal conductivity of the plate
N = nondimensional thermal conductivity parameter $k_p/\varepsilon\sigma T_\infty^3 L$
n = a positive integer
Pr = Prandtl number of the fluid
Q = total internal energy generation in the prototype
Q_m = total internal energy generation in the model
q = uniform volumetric internal energy generation rate in the prototype
$\dot{q}''$ = convective heat-transfer rate per unit area at the plate surface
q_{mx} = varying volumetric internal energy generation rate in the model
R = electric resistance of the plate
Re = Reynolds number of the fluid
S_{ij}^{ℓ} = series for laminar flow:

$$\frac{1}{j-1} + \frac{8}{21}\,\frac{i^{7/4}-(i-1)^{7/4}}{(j-1)^{7/4}} + \frac{2}{9}\,\frac{i^{5/2}-(i-1)^{5/2}}{(i-1)^{5/2}} + \cdots$$

ORIGINAL PAGE IS OF POOR QUALITY

S_{ij}^{t} = series for turbulent flow:

$$\frac{1}{j-1} + \frac{80}{171}\,\frac{i^{19/10}-(i-1)^{19/10}}{(j-1)^{19/10}} + \frac{160}{567}\,\frac{i^{14/5}-(i-1)^{14/5}}{(j-1)^{14/5}}$$

T_x = temperature of the plate surface
T_∞ = freestream temperature of the fluid
V = potential drop across the plate
W = width of the plate
X = nondimensional plate length parameter x/L, $0 \le X \le 1$
x = dimensional length of the plate
Y = nondimensional plate thickness parameter t/L
Z = nondimensional dummy variable for length ξ/L

Greek Letters

ε = total hemispherical emissivity of the plate
$\ddot{\theta}$ = the second derivative of nondimensional temperature
θ = nondimensional plate temperature T_x/T_∞
ρ = electrical resistivity of the plate
ξ = dimensional dummy variable for plate length

p.261

σ = Stefan-Boltzmann constant, 5.680×10^{12} w/cm^2-^{0}K^4
ϕ = nondimensional energy generation parameter, $qt/\varepsilon\sigma T_\infty^4$

I. Introduction

The satisfactory design of many systems requires thermal analysis and thermal testing of the hardware. Thermal scale modeling is a useful tool for the design of many systems. Scale modeling gains even more importance with the increasing size of space vehicles.

The two common approaches to modeling are dimensional analysis and thermal similitude. In the former method, certain dimensionless groups are equated which must contain all of the physical quantities and constants involved in the system. This approach has been used by Katz,[1] Vickers,[2] and MacGregor[3] to study the thermal behavior of the walls of the spacecraft. In the theory of thermal similitude, it is essential to know the explicit governing equations of the system. The model criteria are established from the governing equations. Then, if a similar but smaller-sized model is constructed, its thermal behavior would be given by the same governing equations at homologous locations in the model. This approach has been examplified in the works of Jones,[4] Rolling,[5] and Chao and Wedekind.[6] Jones[7] also has given a brief account of the theory of similitude applicable to spacecraft.

In this work, an attempt has been made to demonstrate, analytically and experimentally, a method of thermal modeling based upon the theory of thermal similitude. After solving numerically the governing equation for the temperature of the prototype, the dimensionless temperature profile of the model is forced to be the same as that of the prototype. From that, the local internal heat generation rate in the model is calculated for scaling purpose. The procedure of numerical solution and the scaling method are confirmed by testing geometrically similar models. The particular example chosen is a flat plate having internal thermal energy generation and subjected to all three modes of heat exchange.

II. Analysis

The thermal system considered here is a thin, semi-infinite flat plate placed in a stream of transparent air. The

p. 262

plate (see Fig. 1) has uniform volumetric internal energy generation and has one face insulated while the other rejects heat by convection and radiation. It is assumed that the properties of the plate and air are independent of any temperature variation, that the plate is a gray surface, and that the surroundings are black and isothermal at T_∞. The conditions on the surroundings can be relaxed by accounting for the emission and reflection of radiant energy from them. This can be done for a known geometry and known temperature of the surroundings. The nondimensional plate temperature in a laminar flow can be obtained by combining the governing energy equation for the plate with the equation describing the convective heat transfer from the plate. These are given by, respectively,

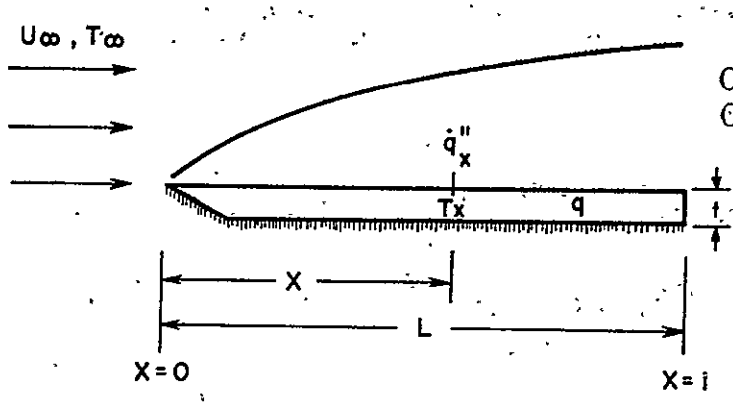
$$\dot{q}''_x = k_p t(d^2T_x/dx^2) - \varepsilon\sigma(T_x{}^4 - T_\infty{}^4) + qt \tag{1}$$

$$T_x - T_\infty = (0.623/k_f)\ Pr^{-1/3}\ Re_x{}^{-1/2} \int_0^x \left[1-\left(\frac{\xi}{x}\right)^{3/4}\right]^{-2/3} \dot{q}''_\xi\ d\xi \tag{2}$$

Combining Eqs. (1) and (2) and nondimensionalizing gives

$$\theta_X = 1 + C_1 \frac{k_p}{k_f} X^{0.5}\left[3.530\left(\frac{1+\phi}{N}\right) + \sum_{i=1}^{j}\left(Y\bar{\bar{\theta}}_i - \frac{\bar{\theta}_i{}^4}{N}\right)S_{ij}^{\ell}\right] \tag{3}$$

Details of the analysis are given in Ref. 8. The various nondimensional parameters for a plate of length L are defined as follows:



ORIGINAL PAGE IS
OF POOR QUALITY

Fig. 1 A flat-plate thermal system.

p. 263

$$X = x/L \qquad Z = \xi/L$$

$$\theta_X = T_x/T_\infty \qquad \theta_Z = T_\xi/T_\infty$$

$$Y = t/L \qquad C_1 = 0.623\ Pr^{-1/3}\ Re_L^{-1/2}$$

$$N = k_p/\bar{\varepsilon}\sigma T_\infty^3 L \qquad \phi = qt/\varepsilon\sigma T_\infty^4$$

$$S_{ij}^{\ell} = \frac{1}{j-1} + \frac{8}{21}\frac{i^{7/4}-(i-1)^{7/4}}{(j-1)^{7/4}} + \frac{2}{9}\frac{i^{5/2}-(i-1)^{5/2}}{(j-1)^{5/2}} + \frac{160}{1053}\frac{i^{13/4}-(i-1)^{13/4}}{(j-1)^{13/4}} + \cdots$$

If n is taken to be a positive integer, then $i = nZ + 1$ and $j = nX + 1$.

Similarly for a turbulent flow, the plate temperature is given by

$$\theta_X = 1 + C_t \frac{k_p}{k_f} X^{0.2} \left[9.827\left(\frac{1+\phi}{N}\right) + \sum_{i=1}^{j}\left(Y\bar{\bar{\theta}}_i - \frac{\bar{\theta}_i^{\,4}}{N}\right) S_{ij}^{t}\right] \tag{4}$$

where $C_t = 3.323\ Pr^{-0.6}\ Re_L^{-0.8}$, and

$$S_{ij}^{t} = \frac{1}{j-1} + \frac{80}{171}\frac{i^{19/10}-(i-1)^{19/10}}{(j-1)^{19/10}} + \frac{160}{567}\frac{i^{14/5}-(i-1)^{14/5}}{(j-1)^{14/5}} + \cdots$$

For thermal modeling, in order to avoid numerous changes, it is proposed here that only one dimension of the prototype be changed. Therefore, in order to have the same temperature at homologous locations of the prototype and the model, all physical properties are kept constant except for the reduced length of the model. By solving Eq. (3) for the volumetric energy generation q, we obtain

$$q = \frac{k_p T_\infty}{L\ t}\left\{\frac{1}{3.530}\left[\frac{(\theta_X - 1)}{C_1 (k_p/k_f) X^{0.5}} - \sum_{i=1}^{j}\left(Y\bar{\bar{\theta}}_i - \frac{\bar{\theta}_i^{\,4}}{N}\right) S_{ij}^{\ell}\right] - \frac{1}{N}\right\} \tag{5}$$

Next, we define r = prototype length/model length. If we apply Eq. (5) for a model, the varying volumetric energy generation along the length of the model q_{mx}, for the same temperature profile as in the prototype, would be given by

$$q_{mx} = \frac{k_p T_\infty}{L\ t}\left\{\frac{1}{3.530}\left[\frac{(\theta_X - 1) r^{0.5}}{C_1 (k_p/k_f) X^{0.5}} - \sum_{i=1}^{j}\left(r^2 Y\bar{\bar{\theta}}_i - \frac{\bar{\theta}_i^{\,4}}{N}\right) S_{ij}^{\ell}\right] - \frac{1}{N}\right\}$$

f.264 ...

ORIGINAL PAGE IS
OF POOR QUALITY

or

$$\frac{q_{mx}}{q} = 1 + \frac{N}{\phi}\frac{1}{3.530}\left[\frac{(\theta_X-1)(r^{0.5}-1)}{C_l(k_p/k_f)\ X^{0.5}} - \sum_{i=1}^{j}(r^2-1)Y\bar{\bar{\theta}}_i\ S_{ij}^{\ell}\right] \tag{6}$$

Similarly for a turbulent flow, we have

$$\frac{q_{mx}}{q} = 1 + \frac{N}{\phi}\frac{1}{9.827}\left[\frac{(\theta_X-1)(r^{0.2}-1)}{C_t(k_p/k_f)\ X^{0.2}} - \sum_{i=1}^{j}(r^2-1)Y\bar{\bar{\theta}}_i\ S_{ij}^{t}\right] \tag{7}$$

Note that, for no internal generation in the prototype (q=0), the equation must be modified by multiplying through by q, which then will cancel with part of ϕ. The method, however, will still apply. The q_{mx} necessary to produce a temperature profile in the model similar to that of the prototype with q=0 is then predicted by the modified Eq. (7).

III. Experimental Procedure

Before confirming the modeling procedure as just developed, an experiment was performed to see the validity of the numerical results of plate temperature as given by Eqs. (3) and (4). A 279.4 × 152.4 × 0.0508-mm 303 stainless-steel plate was tested in a 41.9 × 41.9 cm suction-type wind tunnel. The plate temperature was measured by installing 13 30-gauge chromel-alumel thermocouples on the lower side of the plate. The thermocouples were attached at the centerline of the plate width (279.4 mm) and along its length (152.4 mm), at equal intervals, except close to the leading edge, where the spacing was decreased. This centerline placement eliminated multidimensional conduction errors. The thermocouples were held in place by Borden Mystik Tape, no. 7366, which also insulated the thermocouple hot junctions from the electric current heating the plate. To insulate the lower face of the plate, it was attached by Dow Corning silicon adhesive to a 25.4-mm-thick Marinite-36 plate with a wedge-shaped leading edge machined at 10^0. This material manufactured by Johns-Manville, New York, has an average thermal conductivity of 0.967 Cal/hr-cm-^{0}C and specific heat of 0.27 Cal/g-^{0}C. Two copper electrodes were bolted to the base plate, with the test plate sandwiched between the electrodes and the base plate. This setup was then installed in the wind tunnel and connected to the various instruments needed to supply power and record the temperature and velocity.

PRECEDING PAGE BLANK NOT FILMED

The experiment was run for the maximum and minimum Reynolds numbers available from the tunnel; 15.6×10^4 and 4.92×10^4, based on the total length of the plate. Variation in Reynolds number was accomplished solely by changing the freestream velocity. For a current through the test plate of 100 amp, temperature data were recorded when the readings became approximately reproducible. Changes in electrode-plate resistance and plate surface emissivity due to high plate temperatures caused some initial fluctuations in the temperature readings. The interior surface of the test tunnel met the condition of being at T_∞ (room ambient temperature). The tunnel walls were probably near-black in the infra-red region, but measurement of the tunnel-wall absorptivity was not made. The temperatures recorded for laminar and turbulent flow cases are shown in Fig. 2, along with the theoretical results. The property values for 303 stainless steel, averaged over the experimental temperature range and obtained from Ref. 9, are total hemispherical emissivity $\varepsilon = 0.70$, electrical resistivity $\rho = 73.5$ μohm-cm, thermal conductivity $k_p = 142.0$ Cal/cm-hr-^{0}C.

The experimental temperature profiles were smoothed to pass through the experimental data results as indicated on the figures. The ratios q_{mx}/q, shown in Fig. 2, were calculated by using these profiles in Eq. (6) or (7) for a model of ½ full size, i.e., $r = 2$. The varying volumetric internal energy generation in the model q_{mx} was obtained by varying the resistance in the electrical path. Therefore, for an infinitesimal length ds of the plate (model) of electric resistance dR_x, resistivity ρ, and for a voltage drop of V across its width,

$$q_{mx} W_x t\, dx = V^2/dR_x = V^2/(\rho W_x/t\, dx)$$

or

$$W_x = V/\sqrt{q_{mx}/q}$$

ORIGINAL PAGE IS
OF POOR QUALITY

Let $W_x = W_o$ for $x = 0$, and, noting that $q_{mo}/q = 1$, we obtain

$$W_x = W_o/\sqrt{q_{mx}/q} \tag{8}$$

This equation gives the profile of the model plate for the corresponding values of q_{mx}/q for laminar and turbulent flows, respectively, assuming one-dimensional current flow

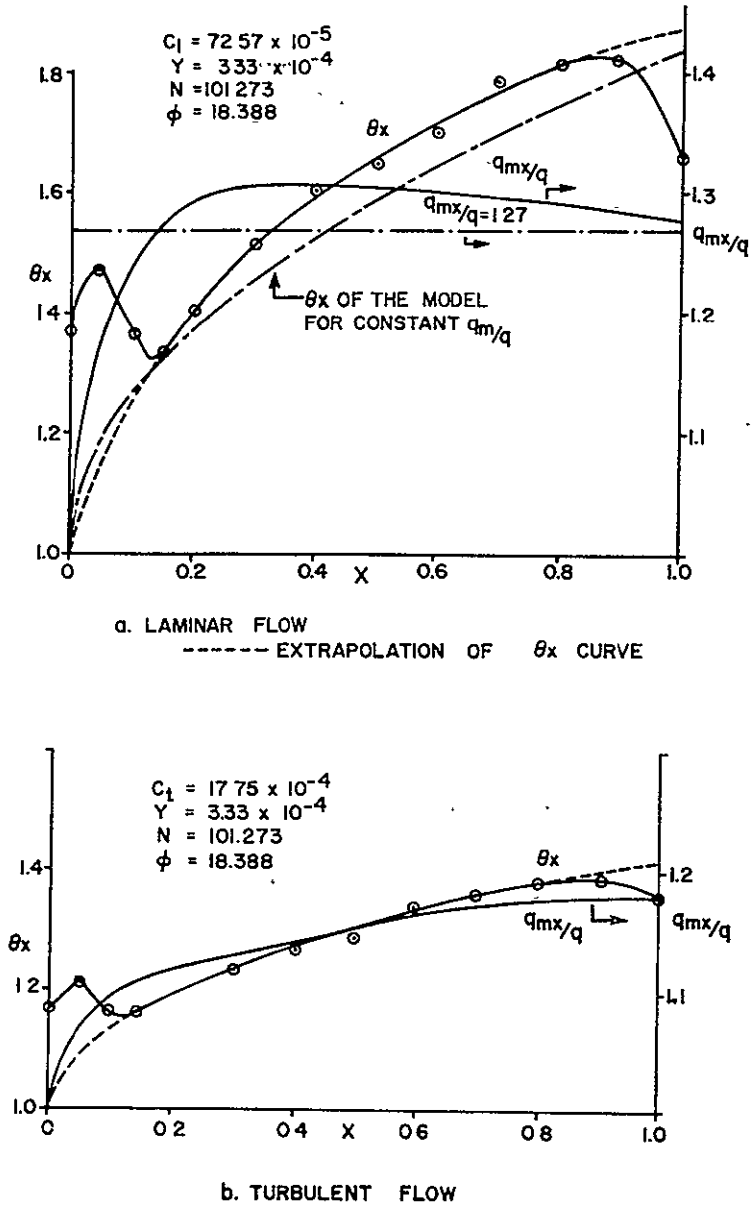


Fig. 2 Determination of q_{mx}/q from experimental temperature profile.

p. 268

across the test plate. In actual practice, the copper electrodes were shaped to give the desired width of the plate, as shown in Fig. 3 for W_o = 292.1 mm. The total energy required to produce the desired temperature profile in the model is then

ORIGINAL PAGE IS
OF POOR QUALITY

$$Q_m = q\ t\ W_o \int_0^L \sqrt{q_{mx}/q}\ dx \tag{9}$$

On evaluating the right-hand side of Eq. (9) graphically, the total energy required for laminar and turbulent flows is obtained as 157 and 148.2 w, respectively. Thus, the proposed method demands a specific energy input, as calculated previously, and then examination of the resulting temperature profile for the model as compared with that of the prototype. But, as no method was available to check the actual distribution of volumetric energy generation along the length of the model, an attempt was made to adjust the total power input to the model so that the temperature profile matched with that of the prototype.

IV. Discussion of the Results

Figure 4 presents the results of the experiments for the prototype, the model, and the theoretical results for the corresponding experimental data. Except for the leading edge and the trailing edge, comparison is reasonable within the experimental errors, the maximum deviation from the theoretical results being 7.0%. For the plate temperature at the leading edge to be equal to the freestream temperature, i.e., $\theta_o = 1$, the convective heat-transfer coefficient should approach an infinite value. This means that the boundary layer thickness at the leading edge and hence the leading edge itself should have zero thickness. This not being the case, the flow is likely to separate at the leading edge, giving rise to a separation laminar bubble, as reported by Tani[10] and Chang.[11] The boundary layer then begins at a point downstream of the leading edge, and the convective heat-transfer coefficient near the leading edge has some finite value, which in turn results in a higher temperature close to the leading edge. Similarly, the trailing edge separation, and formation of a vortex wake behind the trailing edge, could be the cause of plate temperature being below that predicted theoretically.

p.269

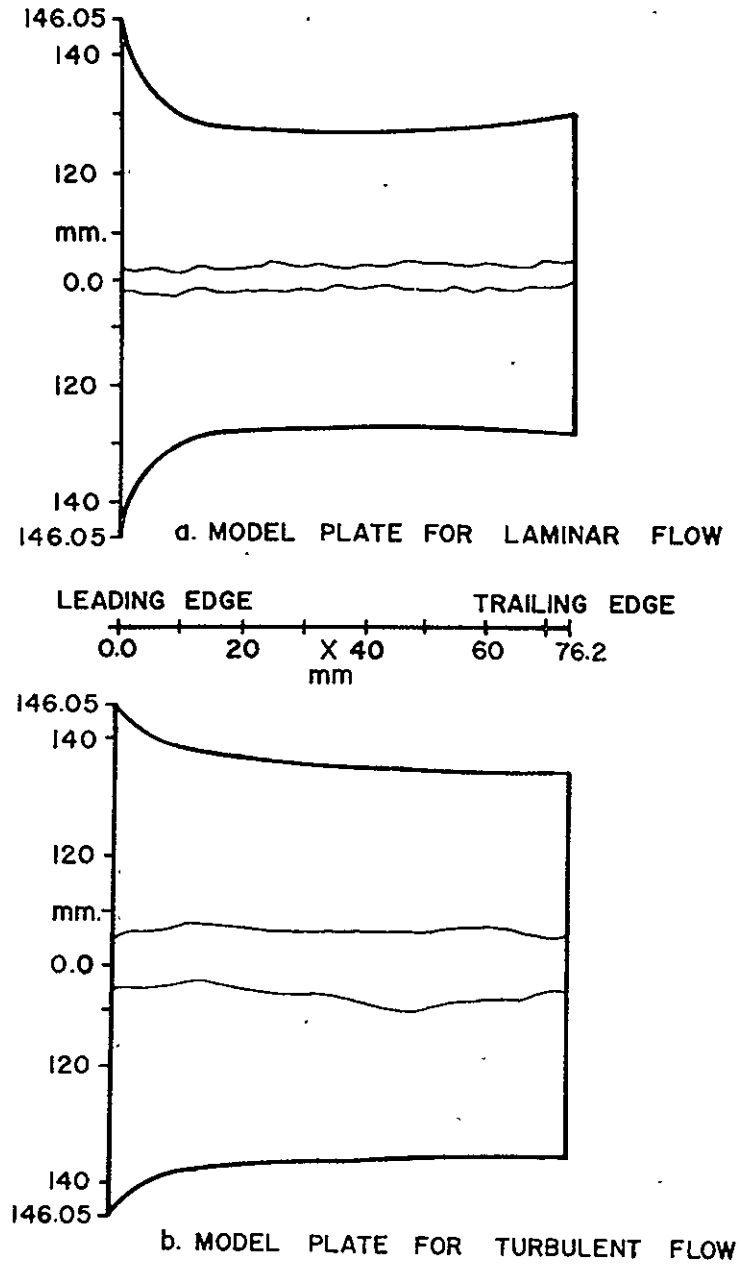
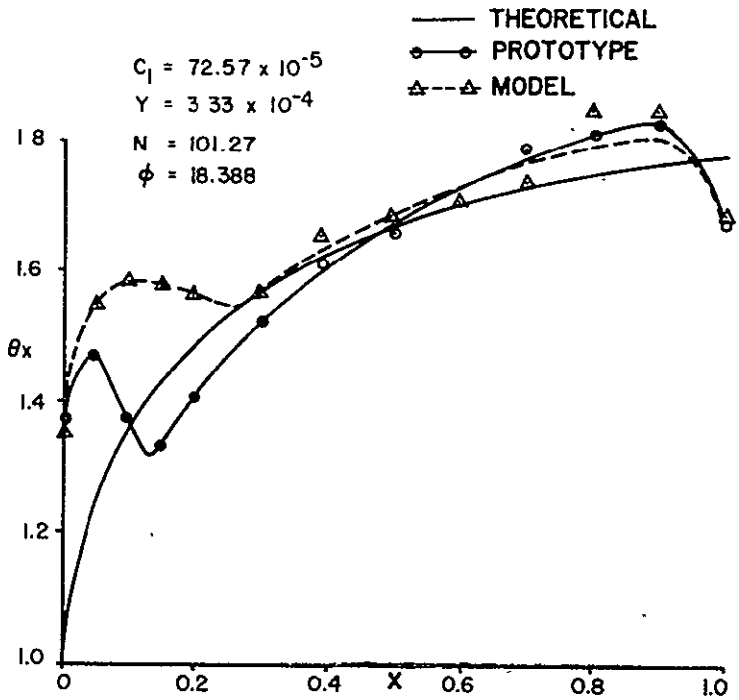
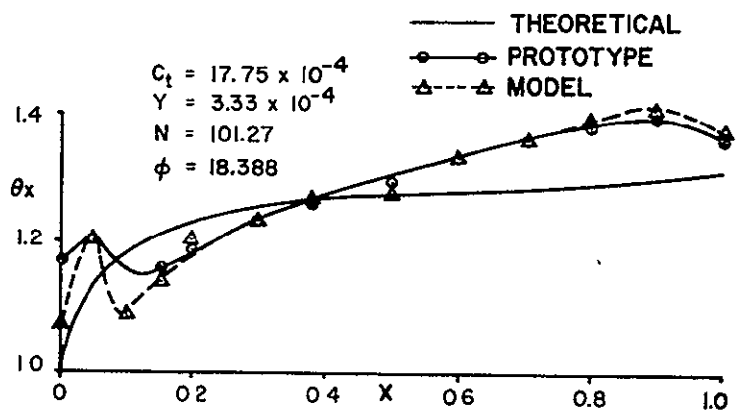


Fig. 3 Geometric profile of the plate (model) width.

1.270



a. LAMINAR FLOW



b TURBULENT FLOW

Fig. 4 Comparison of theoretical results and experimental values of temperature profile.

p. 271

For the plate (prototype), the total energy input for a current flow I can be obtained from $Q = I^2\rho W/Lt$.

It can be seen from Table 1 that the actual energy input as measured during the experiment is more than the corresponding value as calculated analytically. Some energy was lost by conduction through the insulating base plate, which had its lower face in the shape of a wedge with a reduced thickness near the leading edge. A thick leading edge would hamper the formation of boundary layer on the plate. The magnitude of the energy loss by conduction is of the order of 70 w. The electrical resistivity of the steel plate also increases with increase in temperature. This has not been accounted for in the theoretical analysis. The contact resistance between the copper electrodes and the steel plate also contributes somewhat to the energy unaccounted for. Therefore, either a much more accurate estimate of energy distribution is desired, or the present approach of model testing may be used. To correct for the energy loss, the actual energy input in the model is obtained by increasing the theoretically calculated values for the model by the same factor as in the case of the prototype. This is essentially the procedure used in the present case. Figure 4 shows encouraging experimental verification of the theoretical results of plate temperature and the modeling technique.

In an actual modeling of a prototype, this problem should be avoided by monitoring the true energy input into the model. Voltage taps that measure the actual drop across the model and avoid electrode contact losses are necessary when prototype data are not available.

An idea of the discrepancies caused in the plate (model) temperature because of its inaccurate curved boundaries can be had from Fig. 2. If the curved boundaries of the model test plate, as shown in Fig. 3, were replaced by straight ones, we would have a constant volumetric internal energy generation q_m in place of varying energy q_{mx}. Therefore, for a model in laminar flow and for a constant $q_{mx}/q = 1.27$, the temperature profile calculated from Eq. (6) is also shown in Fig. 2a. The difference between θ_x for constant and varying energy generation, although not severe, is noticeable.

V. Conclusions

The experimental results of plate temperature profile compare satisfactorily with those predicted earlier by the

p. 272

Table 1 Theoretically calculated and experimental values of energy required

		A	B	(A-B)/A
		Actual energy input, w	Energy required as per calculation, w	Energy unaccounted for, %
Prototype	Laminar & turbulent flows	100amp, 4.40V = 440.0	265.5	39.65
Model	Laminar flow	57 amp, 4.45V = 253.5	157.0	41.6
	Turbulent flow	62 amp, 5.04V = 312.5	148.2	52.6

theoretical method. For scaling purposes, the method of varying the internal energy generation does work. More sophisticated instrumentation to allow a complete energy loss analysis is necessary in order to use the method better.

The same method can be employed for different fluids and for different flow regimes. Study of two-dimensional and three-dimensional problems would be most desirable to make use of the method presented here. Further research is necessary in such cases to determine whether the method can be applied. It may be necessary to provide locally variable heat sources rather than Joulean heating controlled by boundary shaping. Still, the change of only one parameter, instead if changing many parameters as is the case in dimensional analysis modeling, lends itself as a valuable technique.

References

[1]Katz, A. J., "Thermal Testing," Space/Aeronautics (Aerospace Test Engineering - Part 2), Vol. 38, No. 10, Oct. 1962, pp. 30-34.

[2]Vickers, J. M. F., "Thermal Scale Modeling," Astronautics & Aeronautics, Vol. 3, No. 5, May 1965, pp. 34-39.

[3]MacGregor, R. K., "Spacecraft Thermal Design Verificaiion Through Modleing," AIAA Progress in Astronautics and Aeronautics: Fundamentals of Spacecraft Thermal Design, Vol. 29, edited by John W. Lucas, MIT Press, Cambridge, Mass., 1972,

p. 273

pp. 361-380.

[4]Jones, B. P., "Thermal Similitude Studies," Journal of Spacecraft and Rockets, Vol. 1, No. 4, July-Aug. 1964, pp. 364-369.

[5]Rolling, R. E., "Results of Transient Thermal Modeling in a Simulated Space Environment," AIAA Progress in Astronautics & Aeronautics: Thermophysics and Temperature Control of Spacecraft and Entry Vehicles, Vol. 18, edited by G. B. Heller, Academic Press, New York, 1966, pp. 627-659.

[6]Chao, B. T. and Wedekind, G. L., "Similarity Criteria for Thermal Modeling of Spacecraft," Journal of Spacecraft and Rockets, Vol. 2, No. 2, March-April 1965, pp. 146-152.

[7]Jones, B. P., "Theory of Thermal Similitude with Applications to Spacecraft - A Survey," Astronautica Acta, Vol. 12, No. 4, July-Aug. 1966, pp. 258-271.

[8]Sohal, M. S. and Howell, J. R., "Determination of Plate Temperature in Case of Combined Conduction, Convection and Radiation Heat Exchange," International Journal of Heat and Mass Transfer (to be published).

[9]Thermophysical Properties of Matter, TPRC Data series, Vols. 1-7, IFI/Plenum Press, N. Y., 1970.

[10]Tani, I., "Low Speed Flows Involving Bubble Separation," Progress in Aeronautical Sciences, edited by D. Kuchemann and L. H. G. Sterne, Vol. 5, Pergamon Press, Oxford, 1964, pp. 70-103.

[11]Chang, P. K., Separation of Flow, Pergamon Press, Oxford, 1970, p. 323.

p. 274

Project

(A) Project Title: Determination of Plate Temperature in Case of Conbined Conduction, Convection and Radiation Heat Exchange

(B) Project Abstract:

The singular, integro-differential equation for the temperature of a flat plate with internal energy generation and a fluid flowing over one of its faces, is solved numerically using the method of iteration. The present results compare well with those of Sparrow and Lin, except for the leading portions of the plate. It is also seen that the relations given by Cess for similar problems may not give converged solutions for all cases. The importance of conduction in a plate of high thermal conductivity and of radiation in cases of laminar flow has also been demonstrated.

(C) Publication: Int. Journal of Heat and Mass Transfer, 16, 2055 (1973).

(D) Year: 1973

(E) Department: Mechanical Engineering

(F) Student Name: Mahohar S. Sohal

(G) Faculty Advisor: Professor J. R. Howell

p. 275

Int J. Heat Mass Transfer. Vol 16, pp 2055–2066 Pergamon Press 1973 Printed in Great Britain

DETERMINATION OF PLATE TEMPERATURE IN CASE OF COMBINED CONDUCTION, CONVECTION AND RADIATION HEAT EXCHANGE

MANOHAR S. SOHAL* and JOHN R. HOWELL
Department of Mechanical Engineering, Cullen College of Engineering, University of Houston, Houston, Texas 77004, U.S.A.

(*Received* 20 *September* 1972 *and in revised form* 13 *April* 1973)

Abstract—The singular, integro-differential equation for the temperature of a flat plate with internal energy generation and a fluid flowing over one of its faces, is solved numerically using the method of iteration. The present results compare well with those of Sparrow and Lin, except for the leading portions of the plate. It is also seen that the relations given by Cess for similar problems may not give converged solutions for all cases. The importance of conduction in a plate of high thermal conductivity and of radiation in cases of laminar flow has also been demonstrated.

NOMENCLATURE

a, constant, $0 < a \leqslant 1$;
$a_1, a_2 \ldots$, constants;
b, constant, $0 < b < 1$;
$b_0, b_1, b_2 \ldots$, constants;
C_1, constant for a laminar flow, $0{\cdot}623\, Pr^{-\frac{1}{3}}\, Re_L^{-\frac{1}{2}}$;
C_t, constant for a turbulent flow, $3{\cdot}323\, Pr^{-0\,6}\, Re_L^{-0{\cdot}8}$;
F_z, $Y\dot{\theta}_z - \dfrac{\theta_z^4}{N}$;
h_x, convective heat-transfer coefficient;
H_z, $Y\dot{\theta}_z - \dfrac{\theta_z^4}{N} + \dfrac{1+\Phi}{N}$;
i, $i \leqslant j$, $1, 2, 3, \ldots, n+1$;
j, $1, 2, 3, \ldots, n+1$;
k_f, thermal conductivity of the fluid;
k_p, thermal conductivity of the plate;
L, length of plate;
n, a positive integer;
N, nondimensional thermal conductivity parameter, $\dfrac{k_p}{\epsilon\sigma T_\infty^3 L}$;
p, dummy variable for z;
Pr, Prandtl number of the fluid;
q, volumetric internal energy generation in the plate;
$\dot{q}_x''$, convective heat-transfer rate per unit area at the surface;
Re_L, Reynolds number based on length L, $(U_\infty L/\nu)$;
Re_x, Reynolds number based on length x, $(U_\infty x/\nu)$;
S_{ij}, series for a laminar flow,
$$\frac{1}{j-1} + \frac{8}{21}\,\frac{i^{\frac{7}{2}} - (i-1)^{\frac{7}{2}}}{(j-1)^{\frac{7}{2}}} + \frac{2}{9}\,\frac{i^{\frac{3}{2}} - (i-1)^{\frac{3}{2}}}{(j-1)^{\frac{3}{2}}} + \ldots;$$
S_{ij}^t, series for a turbulent flow,
$$\frac{1}{j-1} + \frac{80}{171}\,\frac{i^{\frac{19}{10}} - (i-1)^{\frac{19}{10}}}{(j-1)^{\frac{19}{10}}} + \frac{160}{567}\,\frac{i^{\frac{14}{5}} - (i-1)^{\frac{14}{5}}}{(j-1)^{\frac{14}{5}}} + \ldots;$$

* Now, Research Fellow, Heat Transfer Section, Technische Hogeschool, Eindhoven, The Netherlands.

p.276

t,	thickness of the plate;
T_x,	temperature of the plate surface;
T_∞,	freestream temperature;
U_∞,	freestream velocity of the fluid;
x,	length co-ordinate for the plate, $0 \leqslant x \leqslant L$;
X,	nondimensional plate length, $x/L, 0 \leqslant X \leqslant 1$;
Y,	nondimensional plate thickness, t/L;
Z,	nondimensional dummy variable for length, $\xi/L, 0 \leqslant Z \leqslant X$.

Greek letters

β,	$\beta(l, m)$, represents the Beta function of l and m;
Γ,	$\Gamma(l)$, represents the Gamma function of (l);
ΔX,	$1/n$;
ϵ,	hemispherical total emissivity of the plate;
$\dot{\theta}, -, \ddot{\theta}$,	1st, 2nd and 3rd derivatives of nondimensional temperature;
θ_x,	nondimensional temperature, T_x/T_∞;
θ_z,	T_ξ/T_∞;
ν,	kinematic viscosity of the fluid;
ξ,	dummy variable for the plate length, $0 \leqslant \xi \leqslant x$;
σ,	Stefan–Boltzmann constant, 5680×10^{12} W/cm^2 °K^4;
Φ,	nondimensional energy generation parameter, $qt/\epsilon\sigma T_\infty^4$;
Ψ,	nondimensional length parameter, (k_p/k_f) $[1/N(Re_L)]$ $a(X)^b$.

1. INTRODUCTION

WITH the increasing use of complex thermal systems, analysis of coupled problems becomes very important. In most such systems, heat exchange by only two of the three possible modes is considered. Some problems of combined conduction and radiation have been examined. Viskanta and Grosh [1] analyzed heat transfer by simultaneous conduction and radiation in a gas between two parallel plates. The nonlinear integro-differential equation was solved numerically by an iterative method after reducing it to a nonlinear Fredholm integral equation of the second kind. Howell [2] solved a combined conduction and radiation problem by a finite-difference technique, considering the radiant exchange terms involved in the equation to be independent of the conduction process. Doornink and Hering [3] gave numerical solutions to the transient simultaneous conductive and radiative transfer in a plane gray medium bounded by black walls. The singular nonlinear integro-partial differential equation was solved by representing its nonlinear function by a finite expansion in terms of elementary functions.

Other coupled problems of heat transfer have also received some attention. Oliver and McFadden [4] solved the problem of simultaneous convection and radiation in a laminar boundary layer on an isothermal flat plate by reducing the governing equations to the familiar equation of Blasius. Sparrow and Lin [5] carried out an analysis to determine the distribution of surface temperature on a flat plate undergoing heat exchange with the environment by both convection and radiation and having an internal heat source or sink. The nonlinear integral equation was solved numerically by changing the integral into a series summation and using a predictor–corrector numerical technique. Cess [6, 7] presented an analysis to determine the influence of radiation heat transfer upon the forced convection Nusselt number. Though the solutions presented by Cess do not converge for all the values of plate length, the results were used to find under what conditions radiation may be neglected.

The present study is aimed at determining the temperature profile of a thermal system involving internal energy generation and conduction. Externally, heat is rejected to a flowing trans-

p.27

parent gas by convection and to constant temperature black surroundings by radiation. Similar problems are encountered in the design of aircraft and missiles, hot wire anemometry, cooling of electronic instruments and other areas.

2. ANALYSIS OF THE PLATE TEMPERATURE DISTRIBUTION

Derivation of the governing equations

Consider a very thin flat plate of finite length and infinite width with uniform internal generation of thermal energy. Let there be a flow of transparent gas over one face of the plate and let the other face be insulated (see Fig. 1). It is assumed, here, that the thermal conductivity of

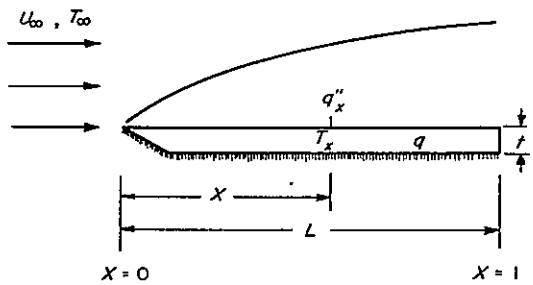


FIG. 1. Thermal model, a flat plate.

the plate and other properties of the plate and air remain constant along the length of the plate, i.e. the properties are independent of any temperature variation. The plate is taken to be a gray surface and the surroundings are considered to be black. The governing energy equation in the plate is obtained by considering an infinitesimal element dx of the plate at a distance x from its leading edge. This analysis yields

$$\dot{q}''_x = h_x(T_x - T_\infty) = k_p t \frac{d^2 T_x}{dx^2} - \epsilon\sigma(T_x^4 - T_\infty^4) + qt, \quad (1)$$

where $\dot{q}''_x$ is the convective heat-transfer rate per unit area of the surface, h_v is the convective heat-transfer coefficient, T_x is the plate temperature, T_∞ is the freestream temperature, k_p is the thermal conductivity of the plate, t is the plate thickness and q is the volumetric internal energy generation in the plate.

For a constant freestream flow along a semi-infinite plate, a solution can be obtained for the wall temperature for the case of arbitrary specified surface (convective) heat flux from Kays' text [8]. Considering a laminar flow

$$T_x - T_\infty = \frac{0{\cdot}623}{k_f} Pr^{-\frac{1}{3}} Re_x^{-\frac{1}{2}} \int_0^x \left[1 - \left(\frac{\xi}{x}\right)^{\frac{3}{4}}\right]^{-\frac{1}{3}} \dot{q}''_\xi \, d\xi, \quad (2)$$

where k_f, Pr and Re are the thermal conductivity, Prandtl number and Reynolds number respectively and ξ is the dummy variable.

Combining the equations (1) and (2) and further simplifying gives

$$T_x - T_\infty = \frac{0{\cdot}623}{k_f} Pr^{-\frac{1}{3}} Re_x^{-\frac{1}{2}} \int_0^x \left[1 - \left(\frac{\xi}{x}\right)^{\frac{3}{4}}\right]^{-\frac{1}{3}} \cdot \left[k_p t \frac{d^2 T_\xi}{d\xi^2} - \epsilon\sigma(T_\xi^4 - T_\infty^4) + qt\right] d\xi.$$

Define the various dimensionless numbers as follows:

$$\frac{x}{L} = X \qquad \frac{\xi}{L} = Z$$

$$\frac{T_x}{T_\infty} = \theta_x \qquad \frac{T_\xi}{T_\infty} = \theta_z$$

$$\frac{t}{L} = Y \qquad 0{\cdot}623\, Pr^{-\frac{1}{3}} Re_L^{-\frac{1}{2}} = C_1$$

$$\frac{k_p}{\epsilon\sigma T_\infty^3 L} = N \qquad \frac{qt}{\epsilon\sigma T_\infty^4} = \Phi.$$

Thus the nondimensional governing equation for the plate temperature in a laminar flow is

$$\theta_x = 1 + C_1 \frac{k_p}{k_f} X^{-\frac{1}{2}} \int_0^X \left[1 - \left(\frac{Z}{X}\right)^{\frac{3}{4}}\right]^{-\frac{1}{3}} \times \left[Y\ddot{\theta}_z - \frac{\theta_z^4}{N} + \frac{1 + \Phi}{N}\right] dZ. \quad (3)$$

p. 278

This is a singular, integro-differential equation with a nonlinearity in temperature. The kernel

$$H_z\left[1-\left(\frac{Z}{X}\right)^{\frac{3}{2}}\right]^{-\frac{2}{3}},$$

where

$$H_z=\left(Y\dot{\theta}_z-\frac{\theta_z^4}{N}+\frac{1+\Phi}{N}\right),$$

makes the equation singular as it has a weak singularity at $Z=X$. But the integral exists and converges to a finite value. By removing the singularity, the kernel can be transformed into a continuous function.

In a similar manner the nondimensional governing equation for turbulent flow can be obtained, using a relation given in reference [8].

$$\theta_x=1+C_t\frac{k_p}{k_t}X^{-0\cdot 8}\int_0^X\left[1-\left(\frac{Z}{X}\right)^{\frac{9}{10}}\right]^{-\frac{8}{9}}\cdot\left[Y\theta_z-\frac{\theta_z^4}{N}+\frac{1+\Phi}{N}\right]\mathrm{d}Z. \tag{4}$$

where

$$C_t=3\cdot 323\,Pr^{-0\cdot 6}\,Re_L^{-0\cdot 8}.$$

Closed form analytical solutions to equations of the type (3) and (4) are not known. Therefore they are solved by employing two numerical methods.

Simplification obtained on integrating by parts

The equation (3), for laminar flow, can be rewritten by integrating the right hand side by parts. Thus,

$$\theta_x=1+C_1\frac{k_p}{k_f}X^{-\frac{1}{2}}\times\left\{(H_z)_0^X\int_0^X\left[1-\left(\frac{Z}{X}\right)^{\frac{3}{2}}\right]^{-\frac{2}{3}}\mathrm{d}Z-\int_0\frac{\mathrm{d}H_z}{\mathrm{d}Z}\int\left[1-\left(\frac{Z}{X}\right)^{\frac{3}{2}}\right]^{-\frac{2}{3}}\mathrm{d}Z\,\mathrm{d}Z\right\}.$$

It can be easily seen that the term

$$\int[1-(Z/X)^{\frac{3}{2}}]^{-\frac{2}{3}}\,\mathrm{d}Z$$

is equivalent to

$$\int_0^Z[1-(p/X)^{\frac{3}{2}}]^{-\frac{2}{3}}\,\mathrm{d}p,$$

where p is a dummy variable of integration. Therefore, it is obtained,

$$\theta_x=1+C_1\frac{k_p}{k_f}X^{-\frac{1}{2}}\left\{\left[\left(Y\ddot{\theta}_x-\frac{\theta_x^4}{N}+\frac{1+\Phi}{N}\right)-\left(Y\dot{\theta}-\frac{\theta_0^4}{N}+\frac{1+\Phi}{N}\right)\right]\int_0^X\left[1-\left(\frac{Z}{X}\right)^{\frac{3}{2}}\right]^{-\frac{2}{3}}\times\mathrm{d}Z-\int_0^X\left(Y\dot{\theta}_z-\frac{4\theta_z^3\dot{\theta}_z}{N}\right)\times\int_0^Z\left[1-\left(\frac{p}{X}\right)^{\frac{3}{2}}\right]^{-\frac{2}{3}}\mathrm{d}p\,\mathrm{d}Z\right\}. \tag{5}$$

Applying equation (1), for the limiting case of $x\to 0$,

$$[T_x-T_\infty]_{x\to 0}=\frac{[k_p t\,\mathrm{d}^2T/\mathrm{d}x^2-\epsilon\sigma(T_x^4-T_\infty^4)+qt]x\to 0}{[h_x]_{x\to 0}}.$$

For a flat plate, the local convective heat-transfer coefficient varies as $1/(x)^{\frac{1}{2}}$, so as

$$x\to 0,\qquad h_x\to\infty,$$

or

$$\theta_0\to 1.$$

Again considering the energy equation (1) for $x\to 0$,

$$h_x(T_x-T_\infty)|_{x\to 0}=k_p t\left.\frac{\mathrm{d}^2T_x}{\mathrm{d}x^2}\right|_{x\to 0}-\epsilon\sigma(T_x^4-T_\infty^4)|_{x\to 0}+qt|_{x\to 0},$$

p.279

or

$$Y\dot{\theta}_0 = -\frac{\Phi}{N}.$$

Therefore, equation (5) reduces to

$$\theta_x = 1 + C_1 \frac{k_p}{k_f} X^{-\frac{1}{2}} \left\{ \left(Y\ddot{\theta}_x - \frac{\theta_x^4}{N} + \frac{1+\Phi}{N} \right) \times \int_0^X \left[1 - \left(\frac{Z}{X} \right)^{\frac{1}{2}} \right]^{-\frac{1}{2}} \mathrm{d}Z - \int_0^X \left(Y\dddot{\theta}_z - \frac{4\theta_z^3 \dot{\theta}_z}{N} \right) \int_0^Z \left[1 - \left(\frac{p}{X} \right)^{\frac{1}{2}} \right]^{-\frac{1}{2}} \mathrm{d}p\, \mathrm{d}Z \right\}. \quad (6)$$

The finite value of a singular integral of the type

$$\int_0^X [1 - (Z/X)^a]^{-b}\, \mathrm{d}Z$$

(where $0 < a \leqslant 1$ and $0 < b < 1$), can be obtained by making use of a Beta function. Thus,

$$\int_0^X \left[1 - \left(\frac{Z}{X} \right)^a \right]^{-b} \mathrm{d}Z = \frac{X}{a} \beta \left(1 - b, \frac{1}{a} \right).$$

[$\beta(l, m)$ represents the Beta function of l and m]

Thus when the integral

$$\int_0^Z [1 - (p/X)^{\frac{1}{2}}]^{-\frac{1}{2}}\, \mathrm{d}p$$

of equation (6) becomes singular, it can also be simplified by using a beta function for $Z = X$.

Thus the simplified governing equation for the plate temperature in a laminar flow becomes,

$$\theta_x = 1 + C_1 \frac{k_p}{k_f} X^{-0.5} \left\{ 3.530\, X \left(Y\ddot{\theta} - \frac{\theta_z^4}{N} + \frac{1+\Phi}{N} \right) - \int_0^X \left(Y\dddot{\theta}_z - \frac{4\theta_z^3 \dot{\theta}_z}{N} \right) \times \int_0^Z \left[1 - \left(\frac{p}{X} \right)^{\frac{1}{2}} \right]^{-\frac{1}{2}} \mathrm{d}p\, \mathrm{d}Z \right\}. \quad (7)$$

If the flow is turbulent, the governing equation (4) can be simplified by following a procedure similar to the laminar flow case, to obtain,

$$\theta_x = 1 + C_t \frac{k_p}{k_f} X^{-0.8} \left\{ 9.827\, X \left(Y\dddot{\theta}_z - \frac{\theta_z^4}{N} + \frac{1+\Phi}{N} \right) - \int_0^X \left(Y\ddddot{\theta}_z - \frac{4\theta_z^3 \dot{\theta}_z}{N} \right) \times \int_0^Z \left[1 - \left(\frac{p}{X} \right)^{\frac{9}{10}} \right]^{-\frac{8}{9}} \mathrm{d}p\, \mathrm{d}Z \right\}. \quad (8)$$

Simplification is obtained by expanding in a series.

The integral in the governing equation (3) can be simplified by using another method as illustrated by Sparrow and Lin [5]. The treatment presented by Sparrow and Lin is only for the case where the delay factor in the kernel is of the type $[1 - (Z/X)]^{-b}$ and not of the type $[1 - (Z/X)^a]^{-b}$ as used in the present analysis. The simplification in the form of delay factor used by Sparrow and Lin was obtained, as proposed by Hanna and Mayers [9], by using a superposition of step-changes in surface heat flux instead of a superposition of step-changes in surface temperature.

Divide the region between $X = 0$ and $X = 1$ into n equal parts, such that $\Delta X = 1/n$. Any value of X in the region $0 \leqslant X \leqslant 1$ can be denoted by $X = (j-1)\Delta X$ and $Z = (i-1)\Delta X$. Also, in equation (3) for $|Z/X| < 1$, the kernel can be expanded in a binomial series. Thus

$$\int_0^X F_z \left[1 - \left(\frac{Z}{X} \right)^{\frac{1}{2}} \right]^{-\frac{1}{2}} \mathrm{d}Z = \sum_{i=1}^{j} \int_{(i-1)\Delta X}^{i\Delta X} F_z \left[1 + \frac{2}{3} \left(\frac{Z}{X} \right)^{\frac{1}{2}} + \frac{5}{9} \left(\frac{Z}{X} \right)^{\frac{3}{2}} + \frac{40}{81} \left(\frac{Z}{X} \right)^{\frac{5}{2}} + \ldots \right] \mathrm{d}Z = \sum_{i=1}^{j} F_z (j-1)\, \Delta X \left[\frac{1}{j-1} + \frac{8}{21} \frac{i^{\frac{7}{2}} - (i-1)^{\frac{7}{2}}}{(j-1)^{\frac{7}{2}}} + \frac{2}{9} \frac{i^{\frac{9}{2}} - (i-1)^{\frac{9}{2}}}{(j-1)^{\frac{9}{2}}} + \ldots \right],$$

p.280

where

$$\bar{F}_z = \tfrac{1}{2}[F_{(i-1)\Delta X} + F_{i\Delta X}],$$
$$= \left[Y\bar{\theta}_i - \frac{\bar{\theta}_i^4}{N}\right].$$

Substitute this in equation (3) and rearrange some terms, to obtain

$$\theta_x = 1 + C_1 \frac{k_p}{k_f} X^{0\cdot5}\left[3\cdot530\left(\frac{1+\Phi}{N}\right) + \sum_{i=1}^{j}\left(Y\bar{\theta}_i - \frac{\bar{\theta}_i^4}{N}\right) S_{ij}^l\right], \quad (9)$$

where

$$S_{ij}^l = \frac{1}{j-1} + \frac{8}{21}\frac{i^{\frac{7}{4}} - (i-1)^{\frac{7}{4}}}{(j-1)^{\frac{7}{4}}} + \frac{2}{9} \times \frac{i^{\frac{3}{4}} - (i-1)^{\frac{3}{4}}}{(j-1)} + \frac{160}{1053}\frac{i^{\frac{13}{4}} - (i-1)^{\frac{13}{4}}}{(j-1)^{\frac{13}{4}}} + \dots$$

The equation (9) can be solved numerically for all values of $Z/X < 1$. However, for $Z/X = 1$ the equation in the form given is indeterminate. Once again use is made of beta function and in place of the summation, the integral is given by,

$$\int_0^X F_z\left[1 - \left(\frac{Z}{X}\right)^{\frac{3}{4}}\right]^{-\frac{4}{3}} dZ$$

$$Z = X \int_0^{X-\Delta X} F_z\left[1 - \left(\frac{Z}{X}\right)^{\frac{3}{4}}\right]^{-\frac{4}{3}} dZ + \bar{F}_z\left\{3\cdot530\,X - \int_0^{X-\Delta X}\left[1 - \left(\frac{Z}{X}\right)^{\frac{3}{4}}\right]^{-\frac{4}{3}} dZ\right\}.$$

In similar manner, the simplified equation for the plate temperature in turbulent flow is obtained, which is,

$$\theta_x = 1 + C_t \frac{k_p}{k_f} X^{0\cdot2}\left[9\cdot827\left(\frac{1+\Phi}{N}\right) + \sum_{i=1}^{j}\left(Y\bar{\theta}_i - \frac{\bar{\theta}_i^4}{N}\right) S_{ij}^t\right], \quad (10)$$

where

$$S_{ij}^t = \frac{1}{j-1} + \frac{80}{171}\frac{i^{\frac{17}{10}} - (i-1)^{\frac{17}{10}}}{(j-1)^{\frac{17}{10}}} + \frac{160}{567}\frac{i^{\frac{14}{5}} - (i-1)^{\frac{14}{5}}}{(j-1)^{\frac{14}{5}}} + \dots$$

For $Z = X$, again the beta function is used to solve the integral.

3. DISCUSSION OF THE RESULTS

Comparison of the present results with those available in the literature

The general procedure to solve the above simplified governing equations is the method of iterations. To find a root of the equation $f(x) = 0$, the method of iterations is concerned with the finding of numbers $x_0, x_1, x_2, \dots, S$, which converge to limit S such that the equation $f(x) = 0$ is satisfied by $x = S$. Therefore an initial guess for the temperature profile of the plate is made. An accurate curve-fitting method needs to be used, because the temperature derivatives at the various locations of the plate length are obtained by differentiating the polynomial representing the assumed temperature profile. The initial guess for the temperature profile is then corrected to approach the temperature profile output obtained by the above numerical procedure. This process is continued till the input and output temperature profiles match within a prescribed accuracy to give a solution to the governing equation. The detailed numerical program is given in [10].

The results for the plate temperature in laminar and turbulent flows are presented in Figs. 2 and 3 respectively. For comparison, plots from the results of Sparrow and Lin [5] and Cess [6] are also shown. The ratio h_{RAD}/h_{UHF} in the analysis of Sparrow and Lin can be modified for comparison with the present results. The ratio h_{RAD}/h_{UHF} is a measure of the relative strengths of radiative and convective heat transfer. Therefore,

p. 281

$$\frac{h_{\mathrm{RAD}}}{h_{\mathrm{UHF}}} = \frac{3{\cdot}511\, C_1}{N}\frac{k_p}{k_f} X^{0{\cdot}5}, \text{ for laminar flow,}$$

$$= \frac{10{\cdot}166\, C_t}{N}\frac{k_p}{k_f} X^{0{\cdot}2}, \text{ for turbulent flow.}$$

Also the parameter, $e/\epsilon\sigma T_\infty^4$ of Sparrow and Lin, is equivalent to $(1 + \Phi)$.

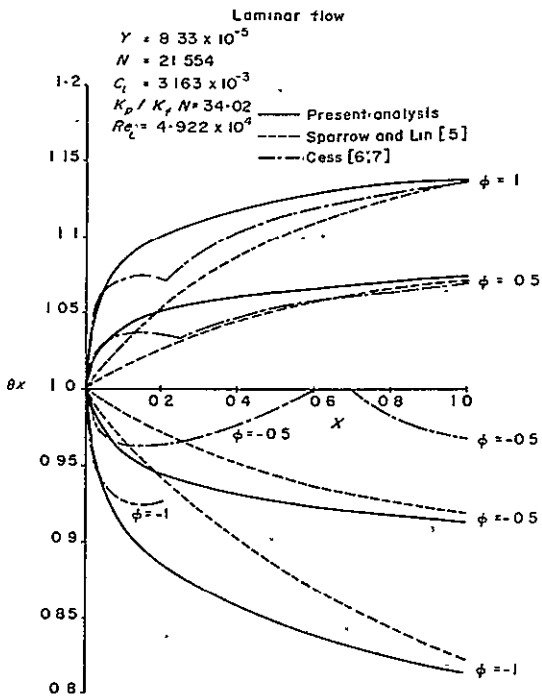


FIG. 2. Comparison of the plate temperature results in laminar flow.

From Figs. 2 and 3 it is seen that the results of Sparrow and Lin give a lower value of the plate temperature than the present results. The maximum deviation of about 5 per cent occurs for turbulent flow near $X = 0{\cdot}2$. The deviation subsequently decreases continuously. Both of the present methods (equations (7)–(10)) of solving the governing equation for plate temperature give identical results for all similar cases. Integration is essentially a summation process, which explains the identical nature of the results given by the present two methods.

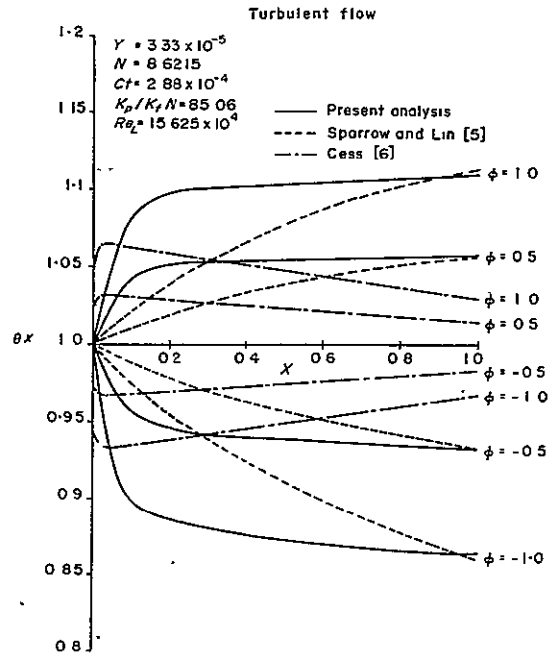


FIG. 3. Comparison of the plate temperature results in turbulent flow.

Therefore, the two methods have not been differentiated on the figures. The governing equations for the plate temperature in laminar and turbulent flows, equations (25) and (20) respectively given by Sparrow and Lin [5], are modified to correspond to the present analysis. For laminar flow, the analysis of Sparrow and Lin gives

$$\theta_x = 1 + \frac{h_{\mathrm{RAD}}}{h_{\mathrm{UHF}}}\left(\frac{e}{\epsilon\sigma T_\infty^4}\right) - \frac{1}{3X}\frac{h_{\mathrm{RAD}}}{h_{\mathrm{UHF}}} \times \int_0^x \frac{\theta_z^4}{(1 - Z/X)^{\frac{1}{3}}}\,\mathrm{d}Z,$$

$$= 1 + 3{\cdot}511\, C_1 \frac{k_p}{k_f} X^{0{\cdot}5}\left\{\left(\frac{1+\Phi}{N}\right) + \sum_{i=1}^{j} \times \left(-\frac{\theta_i^4}{N}\right)\left[\left(1 - \frac{i-1}{j}\right)^{\frac{1}{3}} - \left(1 - \frac{i}{j}\right)^{\frac{1}{3}}\right]\right\}. \quad (11)$$

p. 282

ORIGINAL PAGE IS
OF POOR QUALITY

Including the conduction term also,

$$\theta_x = 1 + 3{\cdot}511\, C_1 \frac{k_p}{k_f} X^{0{\cdot}5} \left\{ \left(\frac{1+\Phi}{N} \right) + \sum_{i=1}^{j} \times \left(Y\bar{\theta}_i - \frac{\bar{\theta}_i^4}{N} \right) \left[\left(1 - \frac{i-1}{j} \right)^{\frac{1}{2}} - \left(1 - \frac{i}{j} \right)^{\frac{1}{2}} \right] \right\}. \tag{12}$$

Similarly, the two equations for turbulent flow without and with conduction heat transfer would be,

$$\theta_x = 1 + 10{\cdot}166\, C_t \frac{k_p}{k_f} X^{0{\cdot}2} \left\{ \left(\frac{1+\Phi}{N} \right) + \sum_{i=1}^{j} \times \left(-\frac{\bar{\theta}_i^4}{N} \right) \left[\left(1 - \frac{i-1}{j} \right)^{\frac{1}{5}} - \left(1 - \frac{i}{j} \right)^{\frac{1}{5}} \right] \right\} \tag{13}$$

and

$$\theta_x = 1 + 10{\cdot}166\, C_t \frac{k_p}{k_f} X^{0{\cdot}2} \left\{ \left(\frac{1+\Phi}{N} \right) + \sum_{i=1}^{j} \times \left(Y\bar{\theta}_i - \frac{\bar{\theta}_i^4}{N} \right) \left[\left(1 - \frac{i-1}{j} \right)^{\frac{1}{5}} - \left(1 - \frac{i}{j} \right)^{\frac{1}{5}} \right] \right\}. \tag{14}$$

The results of Sparrow and Lin in Figs. 2 and 3 are the solutions of equations (11) and (13) respectively by using a predictor–corrector numerical technique. But it is found that by using the method of iterations, the results of equations (11) and (13) are exactly the same as those for equations (7) or (9) and (8) or (10) respectively. This is not because of neglecting conduction terms in equations (11) and (13), since the conduction term is immaterial for the small value of thermal conductivity k_p used.

The results for $N = 0{\cdot}0$, $86{\cdot}2$ can hardly be distinguished in the graphs presented here. In the absence of conduction in the plate and because radiation heat exchange is negligible close to the leading edge, convection is the dominating mode of heat transfer near the leading edge. For convection, $T_x \propto 1/h_x$. Thus while h_x drops asymptotically from a very large to a finite value, the temperature would obviously show a sharp increase along the corresponding plate length for the case of small values of thermal conductivity. This effect is not obvious from the plots of Sparrow and Lin. As shown in the subsequent figures (i.e. Figs. 4, 7 and 8), the figures for lower thermal conductivity show a steeper profile as compared to the curves for higher thermal conductivity. Therefore even the results for the boundary condition $\dot{q}_0'' = qt$ (case of zero thermal conductivity) should show a sharp increase in temperature close to the leading edge. Nevertheless, it seems that the method of iterations with $\Delta X = 0{\cdot}01$ and the predictor–corrector method as employed by Sparrow and Lin, converge to different results.

Equation (5) of Cess [6] gives the plate temperature as

$$\theta_x = 1 + a_1 \left(\Psi + \frac{a_2}{a_1} \Psi^2 + \ldots \right),$$

where $a_1, a_2, \ldots$ are the constants and for laminar flow Ψ is given by

$$\Psi = \frac{\epsilon \sigma T_\infty^3}{k_f} \sqrt{\left(\frac{\nu x}{U_\infty} \right)}, = \frac{k_p}{k_f} \frac{1}{N\sqrt{(Re_L)}} \sqrt{X}.$$

Therefore, for laminar flow,

$$\theta_x = 1 + \frac{\Phi}{0{\cdot}4059} \left[\frac{k_p}{k_f} \frac{1}{N\sqrt{(Re_L)}} (\sqrt{X}) - 8{\cdot}328 \left(\frac{k_p}{k_f} \frac{1}{N\sqrt{(Re_L)}} \sqrt{X} \right)^2 + \ldots \right]. \tag{15}$$

and for turbulent flow,

$$\theta_x = 1 + \frac{\Phi}{0{\cdot}0273} \left[\frac{k_p}{k_f} \frac{1}{N(Re_L)^{0{\cdot}8}} X^{0{\cdot}2} - 142{\cdot}27 \left(\frac{k_p}{k_f} \frac{1}{N(Re_L)^{0{\cdot}8}} X^{0{\cdot}2} \right)^2 + \ldots \right]. \tag{16}$$

For the plate temperature distribution in a laminar flow for higher values of Ψ, Cess [7]

p. 283

has given an alternate approach,

$$\theta_x = 1 + b_0\left[1 + \frac{b_1}{b_0}\Psi^{-1} + \ldots\right], \quad (17)$$

where

$$b_0 = (1 + \Phi)^{\frac{1}{4}} - 1,$$

and

$$\frac{b_1}{b_0} = -\frac{0{\cdot}2927}{4(1 + \Phi)^{\frac{3}{4}}}.$$

With these substitutions, equation (17) becomes

$$\theta_x = (1 + \Phi)^{\frac{1}{4}} - \frac{0{\cdot}2927}{4}\frac{[(1 + \Phi)^{\frac{1}{4}} - 1]}{(1 + \Phi)^{\frac{3}{4}}} \times \frac{k_f}{k_p}\frac{N\sqrt{(Re_L)}}{\sqrt{X}} + \ldots. \quad (18)$$

Results from equations (15), (16) and (18) are also shown in Figs. 2 and 3. The results of equation (15) converge only up to a small value of X around 0·1–0·2. The plots of equation (18) produce results which are physically consistent for $\Phi = 0{\cdot}5$ and 1·0, for $X = 0{\cdot}2$–1·0, but fail to give satisfactory results for $\Phi = -0{\cdot}5$. For turbulent flow also the equation (16) gives results which differ appreciably from the present results and those given by Sparrow and Lin. These plots show a steeper increase in the plate temperature near the leading edge for the same relations given by Cess [6] than the plots presented by Sparrow and Lin.

For both laminar and turbulent flows, the present two methods and the relations given by Sparrow and Lin, solved by using the present numerical technique give identical results, maximum deviation being less than 1 per cent. Therefore, to reduce the computer time, for all the curves presented henceforth, the method of iterations is used to solve the simpler equations (12) and (14), for laminar and turbulent flows respectively

Effect of various parameters

In most of the previous analyses, the conduction heat flow in the plate has been neglected.

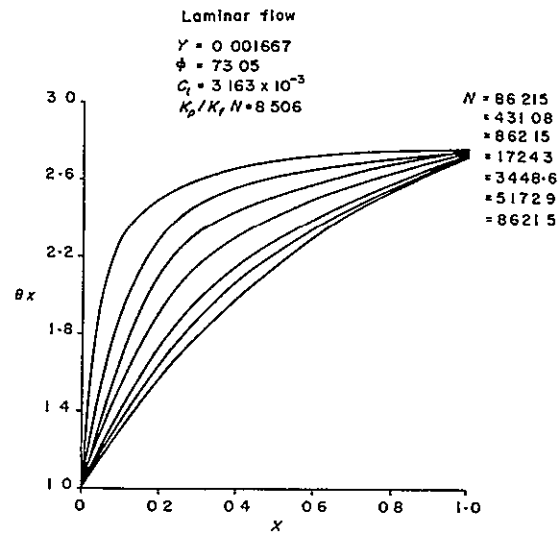


FIG. 4. Effect of thermal conductivity parameter N on the temperature of 0·254 mm thick plate in laminar flow.

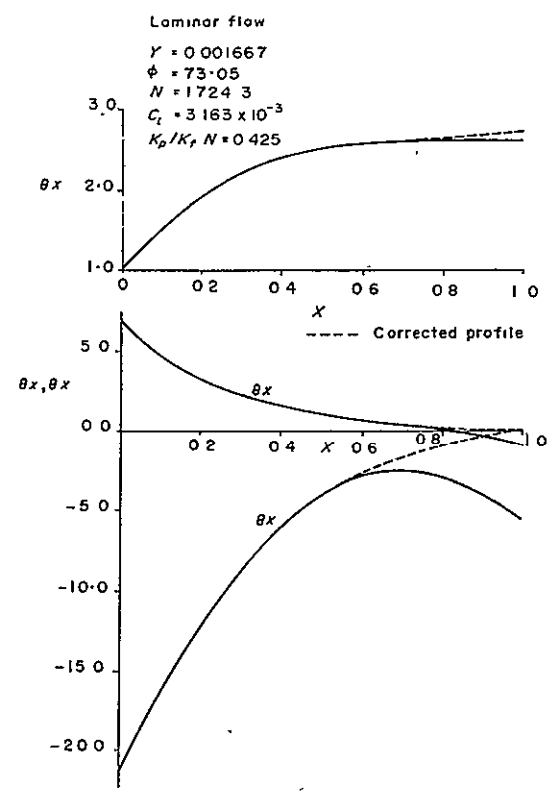


FIG. 5. Correction to be applied in the determination of θ_x.

p. 284

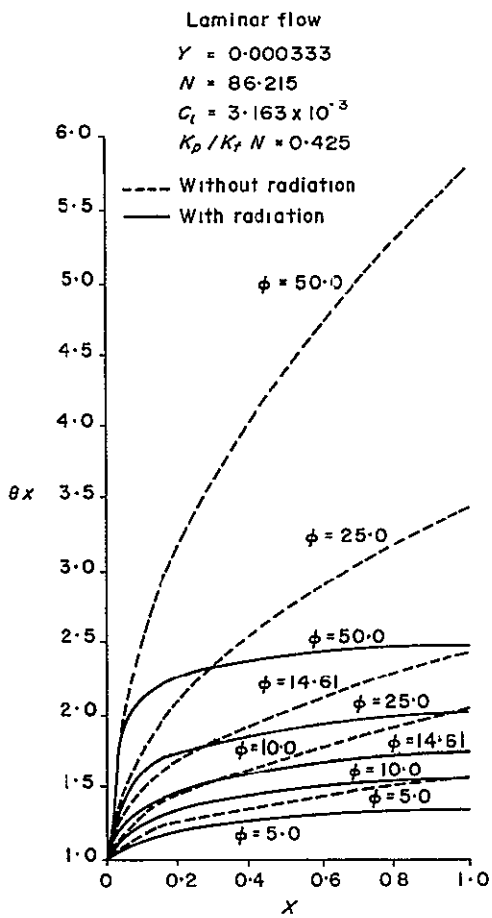


FIG. 6. Effect of the parameter Φ and radiation on the plate temperature in laminar flow.

Figure 4 shows the plate temperature for a wide range of thermal conductivity parameter N. It can be seen that the higher values of N bring down the plate temperature appreciably at smaller values of X, but with subsequent increase in X the effect of N continuously decreases. For a 0·254 mm ($Y = 0{\cdot}00166$) thick copper plate ($N \sim 1724{\cdot}3$) neglecting the thermal conductivity would cause an error of 62 per cent in the plate temperature at a location $X = 0{\cdot}05$ from the leading edge. For an infinitely large value of thermal conductivity, $\bar{\theta}_x$ would approach zero, i.e. $\bar{\theta}_x$ would tend to a constant value.

Thus, in order to study the effect of thermal conductivity, numerical determination of $\bar{\theta}_x$ becomes very critical. Because of the constant internal energy generation in the plate, its temperature has to increase with X. This means that $\bar{\theta}_x$ can never be negative, its value decreasing with x and becoming equal to zero in the extreme case. Therefore $\ddot{\theta}_x$ cannot have a point of inflection and its value has to approach close to zero at $X = 1$. The usual method of determining $\ddot{\theta}_x$ by differentiating the curve fit of the temperature profile is susceptible to some errors, which in turn may cause errors in the calculated temperature. This effect is shown in Fig. 5 by a solid curve in the range $X = 0{\cdot}7$–$1{\cdot}0$. Hence $\bar{\theta}_x$ is forced to follow a physically more realistic profile as shown by the dashed line in Fig. 5.

Figure 6 shows the effect of changing the parameter Φ and also the plate temperature with and without inclusion of the radiative transfer term. Changing Φ essentially means changing either

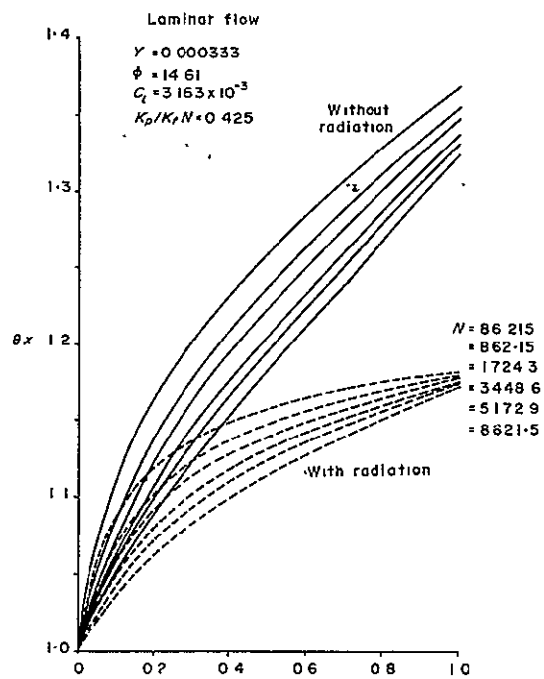


FIG. 7. Effection of the parameter N and radiation on the plate temperature in laminar flow.

J. 285

ORIGINAL PAGE IS
OF POOR QUALITY

internal energy generation or the plate thickness at low thermal conductivity. Thus, as would be expected, an increase in Φ increases the plate temperature as shown in Fig. 6. This figure also illustrates that radiation becomes more and more important with the increase in the plate temperature, which is caused by the higher volumetric energy input.

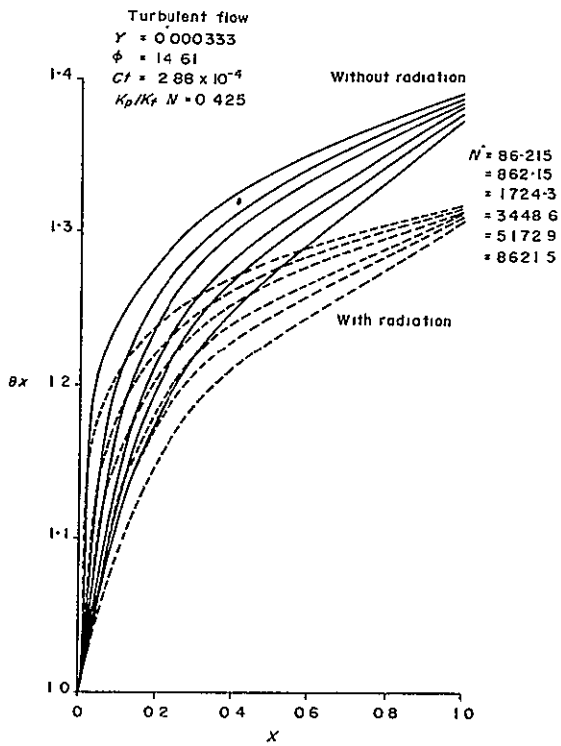


FIG. 8 ·Effect of the parameter N and radiation on the plate temperature in turbulent flow.

Figures 7 and 8 present the relative importance of conduction and radiation heat transfer for laminar and turbulent flows, respectively. The maximum error caused by neglecting the conduction heat transfer in a 0·0508 mm thick plate with $N = 1724{\cdot}3$ is about 10 per cent, as shown in Fig. 7. For the same volumetric energy generation, corresponding error for a 0·254 mm thick plate from Fig. 4, is 62 per cent. Absence of radiation could mean a maximum increase of about 43·5 per cent in the plate temperature in a laminar flow, the maximum increase occurring at the trailing edge of the plate when $N = 86{\cdot}215$. The corresponding increase in a turbulent flow flow is only about 5·55 per cent. This emphasizes the relative importance of radiative heat transfer in laminar and turbulent flows. Clearly, in turbulent flow convective heat transfer is much more dominant. Cess [6] has also reported that above a turbulent Reynolds number of $3{\cdot}5 \times 10^5$ the error caused by neglecting radiation effects would be less than 5 per cent.

4. CONCLUSIONS

The present work presents solutions for the plate temperature, where conduction heat transfer along the length of the plate becomes important. These are the cases of plates having a higher value of thermal conductivity and certain cases of thicker plates, where the assumption of constant temperature across the plate thickness can be maintained. This paper also gives solutions for integral equations which contain a delay factor of the more general type $[1 = (Z/X)^a]^{-b}$ in the kernel, though these solutions show little difference over the results using the delay factor of the type $[1 - (Z/X)]^{-b}$ as done by Sparrow and Lin.

The present results obtained by using the method of iterations, compare within 5 per cent of the results of Sparrow and Lin. Due to a sharp decrease in the value of convective heat transfer coefficient on moving away from the leading edge of the plate, there has to be a sharp increase in the plate temperature, unlike the plots of Sparrow and Lin. The relations given by Cess for plate temperature are not physically consistent over all ranges of parameters, as they do not give a constant increase (or an asymptotic approach to a constant value) in temperature in some cases. The neglect of radiation heat transfer in turbulent flow does not cause as severe an error as in laminar flow, a phenomenon also shown by Cess.

p. 286

REFERENCES

1. R. VISKANTA and R. J GROSH, Heat transfer by simultaneous conduction and radiation in an absorbing medium, *J. Heat Transfer* **84**, 63–72 (1962).
2. J. R. HOWELL, Determination of combined conduction and radiation of heat through absorbing media by the exchange factor approximation, *Chem. Engng Prog Symp. Ser.* **61**, 162–171 (1965).
3. D. DOORNINK and R. G. HERING, Transient combined conduction and radiative heat transfer, ASME Paper No. 71-HT-23 (1971).
4. C. C. OLIVER and P. W. MCFADDEN, The interaction of radiation and convection in the laminar boundary layer, *J. Heat Transfer* **88**, 205–213 (1966)
5. E. M. SPARROW and S. H. LIN, Boundary layers with prescribed heat flux–application to simultaneous convection and radiation, *Int. J Heat Mass Transfer* **8**, 437–448 (1965).
6. R. D. CESS, The effect of radiation upon forced convection heat transfer, *Appl. Sci Res* **10A**, 430–438 (1961)
7. R. D CESS, The interaction of thermal radiation with conduction and convection heat transfer, *Advances in Heat Transfer*, Vol. 1. Academic Press, New York (1964).
8. W. M. KAYS, *Convective Heat and Mass Transfer*, 1st edn. McGraw-Hill, New York (1966).
9. O T HANNA and J. E. MEYERS, Heat transfer in boundary-layer flows past a flat plate with a step in wall-heat flux, *Chem. Engng Sci.* **17**, 1053–1055 (1962).
10. M. S. SOHAL, Scale-up methods for thermal models with coupled conduction, convection and radiation, Ph. D. Dissertation, University of Houston, Houston (1972)

DETERMINATION DE LA TEMPERATURE D'UNE PLAQUE DANS LE CAS D'ECHANGE DE CHALEUR COMBINE PAR CONDUCTION, CONVECTION ET RAYONNEMENT

Résumé—On a résolu numériquement par la méthode itérative l'équation intégro-différentielle singulière pour la température d'une plaque plane avec une génération d'énergie interne et un fluide qui s'écoule sur l'une de ses faces Les résultats présentés se comparent bien à ceux de Sparrow et Lin, sauf pour la région du bord d'attaque de la plaque. On voit aussi que les relations données par Cess pour des problèmes similaires risquent de ne pas donner de solutions convergentes dans tous les cas. L'importance de la conduction dans une plaque à conductivité thermique élevée et du rayonnement dans le cas d'écoulement similaire a été démontrée.

BESTIMMUNG DER PLATTENTEMPERATUR IM FALL DES KOMBINIERTEN WÄRME-AUSTAUSCHS DURCH LEITUNG, KONVEKTION UND STRAHLUNG.

Zusammenfassung—Die Integro-Differentialgleichung für die Temperatur einer ebenen Platte mit inneren Wärmequellen, wobei ein Fluid über eine Plattenfläche strömt, wird numerisch durch Iteration gelöst. Die Ergebnisse lassen sich gut mit denen von Sparrow und Lin vergleichen, außer für den Plattenanfang. Man sieht auch, daß die Gleichung von Cess für ähnliche Probleme nicht für alle Fälle übereinstimmende Lösungen ergibt. Auch wurde die Wichtigkeit der Leitung in einer Platte von hoher thermischer Leitfähigkeit und die der Strahlung im Fall einer laminaren Strömung gezeigt.

ОПРЕДЕЛЕНИЕ ТЕМПЕРАТУРЫ ПЛАСТИНЫ В СЛУЧАЕ СЛОЖНОГО ТЕПЛООБМЕНА ТЕПЛОПРОВОДНОСТЬЮ, КОНВЕКЦИЕЙ И РАДИАЦИЕЙ

Аннотация—Методом итераций проведено численное решение сингулярного интегро-дифференциального уравнения для темпратуры пластины с внутренним источником энергии при обтекании ее с одной стороны. Результаты этой работы хорошо согласуются с даннвми Спэрроу и Лина за исключением передней кромки пластины. Установлено, что соотношения, приведенные Сессом для задач такого типа не всегда дают сходящиеся решения. Показано, что для пластины с высоким коэффициентом теплопроводности и при наличии излучения в случаях ламинарного обтекания теплопроводность играет существенную роль.

p. 287

ORIGINAL PAGE IS OF POOR QUALITY

Project

(A) Project Title: Choice of Honeycomb Cell Size for Best Efficiency of a Flat-Plate Solar Collector

(B) Project Abstract:

Honeycomb size, geometry, and radiative properties are analyzed for their effects on the reduction of convective and radiative losses in solar collectors using honeycomb convection suppressors. It is found the solar absorptance of the honeycomb is a dominant factor in overall collector performance at off-normal solar incidence angles.

(C) Publication: M.S. Thesis in Mechanical Engineering and ASME AICHE Heat Transfer Conference (1976)

(D) Year: 1975

(E) Department: Mechanical Engineering

(F) Student Name: L. Guthrie

(G) Faculty Advisor: Prof. J.R. Howell

an ASME publication

The Society shall not be responsible for statements or opinions advanced in papers or in discussion at meetings of the Society or of its Divisions or Sections, or printed in its publications. *Discussion is printed only if the paper is published in an ASME journal or Proceedings.* Released for general publication upon presentation. Full credit should be given to ASME, the Technical Division, and the author(s).

$3.00 PER COPY
$1.50 TO ASME MEMBERS

Choice of Honeycomb Cell Size for Best Efficiency of a Flat-Plate Solar Collector

L. GUTHRIE

University of Houston, and
Senior Engineer,
WKM Valve,
Houston, Tex
Assoc. Mem. ASME

J. R. HOWELL

Professor,
Mechanical Engineering,
University of Houston,
Houston, Tex.
Mem. ASME

Honeycomb size, geometry, and radiative properties are analyzed for their effects on the reduction of convective and radiative losses in solar collectors using honeycomb convection suppressors It is found that the solar absorptance of the honeycomb is a dominant factor in overall collector performance at off-normal solar incidence angles

Contributed by the Heat Transfer Division of The American Society of Mechanical Engineers for presentation at the ASME-AIChE Heat Transfer Conference, St. Louis, Mo., August 9-11, 1976. Manuscript received at ASME Headquarters April 12, 1976.

Copies will be available until May 1, 1977.

ORIGINAL PAGE IS
OF POOR QUALITY

p. 289

Project

(A) Project Title: An Approximate Solution for the Energy Equation with Radiant Participating Media

(B) Project Abstract:

The analysis of combined modes of heat transfer in a gas becomes increasingly difficult with the introduction of the radiation term in the energy equation. In this paper, the problem of simultaneous convection, conduction, and radiation in the laminar boundary layer over a flat plate is formulated and solved. The solution is obtained analytically by means of the perturbation technique, and the effect of including the radiation term on both the temperature distribution and the heat flux is presented for various conditions. The case of a steady, laminar, two-dimensional flow over a black isothermal flat plate is considered. In addition, the fluid medium is assumed to be gray, isotropic, viscous, radiation absrobing-emmitting, and a thermally conducting perfect gas with most properties independent of the temperature. The nonlinear integral-differential energy equation is reduced to a simpler form through the use of both the Rosseland and the optically thin approximation for the radiation term. The resulting equations clearly illustrate that the problem may be treated as a singular perturbation problem,

particularly if the parameters (N), ( ), and/or their product (N) are small, which is typical for most gases. Results obtained by this approximate method are compared with those obtained numerically for the exact form of the energy equation.

(C) Publication: Ph.D. Dissertation in Mechanical Engineering and AICHE-ASME Heat Transfer Conference, Salt Lake City (Aug. 1977)

(D) Year: 1975

(E) Department: Mechanical Engineering

(F) Student Name: S. Douara

(G) Faculty Advisor: Professor J.R. Howell

C-4

p. 291

an ASME publication

The Society shall not be responsible for statements or opinions advanced in papers or in discussion at meetings of the Society or of its Divisions or Sections, or printed in its publications. *Discussion is printed only if the paper is published in an ASME journal or Proceedings.* Released for general publication upon presentation.
Full credit should be given to ASME, the Technical Division, and the author(s).

$3.00 PER COPY
$1.50 TO ASME MEMBERS

An Approximate Solution for the Energy Equation with Radiant Participating Media

S. DOUARA

Senior Engineer,
Bechtel, Incorporated,
Houston, Texas
Assoc. Mem. ASME

J. R. HOWELL

Professor,
Dept. of Mechanical Engineering,
University of Houston,
Houston, Texas
Mem. ASME

ORIGINAL PAGE IS
OF POOR QUALITY

The analysis of combined modes of heat transfer in a gas becomes increasingly difficult with the introduction of the radiation term in the energy equation. In this paper, the problem of simultaneous convection, conduction, and radiation in the laminar boundary layer over a flat plate is formulated and solved. The solution is obtained analytically by means of the perturbation technique, and the effect of including the radiation term on both the temperature distribution and the heat flux is presented for various conditions. The case of a steady, laminar, two-dimensional flow over a black isothermal flat plate is considered. In addition, the fluid medium is assumed to be gray, isotropic, viscous, radiation absorbing-emitting, and a thermally conducting perfect gas with most properties independent of the temperature. The nonlinear integral-differential energy equation is reduced to a simpler form through the use of both the Rosseland and the optically thin approximation for the radiation term. The resulting equations clearly illustrate that the problem may be treated as a singular perturbation problem, particularly if the parameters (N), (ξ), and/or their product ($N\xi$) are small, which is typical for most gases. Results obtained by this approximate method are compared with those obtained numerically for the exact form of the energy equation.

Contributed by the Heat Transfer Division of The American Society of Mechanical Engineers for presentation at the AIChE-ASME Heat Transfer Conference, Salt Lake City, Utah, August 15–17, 1977. Manuscript received at ASME Headquarters April 20, 1977.

Copies will be available until May 1, 1978.

1.291 -A

THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS, UNITED ENGINEERING CENTER, 345 EAST 47th STREET, NEW YORK, N.Y. 10017

An Approximate Solution for the Energy Equation with Radiant Participating Media

S. DOUARA **J. R. HOWELL**

ABSTRACT

The analysis of combined modes of heat transfer in a gas becomes increasingly difficult with the introduction of the radiation term in the energy equation. In this paper, the problem of simultaneous convection, conduction and radiation in the laminar boundary layer over a flat plate is formulated and solved. The solution is obtained analytically by means of the perturbation technique, and the effect of including the radiation term on both the temperature distribution and the heat flux is presented for various conditions. The case of a steady, laminar, two dimensional flow over a black isothermal flat plate is considered. In addition, the fluid medium is assumed to be gray, isotropic, viscous, radiation absorbing-emitting and a thermally conducting perfect gas with most properties independent of the temperature. The nonlinear integral-differential energy equation is reduced to a simpler form through the use of both the Rosseland and the optically thin approximation for the radiation term. The resulting equations clearly illustrate that the problem may be treated as a singular perturbation problem, particularly if the parameters (N), (ξ), and/or their product (Nξ) are small, which is typical for most gases. Results obtained by this approximate method are compared with those obtained numerically for the exact form of the energy equation.

NOMENCLATURE

a	absorption coefficient
c_p	specific heat at constant temperature
$E_n(t)$	exponential integral function of nth degree, $\int_0^1 \mu^{n-2} \exp(-t/\mu)d\mu$
$f(\eta)$	dimensionless Blasius stream function
$H(\eta)$	Heaviside function
k	thermal conductivity
N	dimensionless conduction-to-radiation parameter, $(ka)/4n^2\sigma T_\infty^3$
n	index of refraction
Pr	Prandtl number, $\mu c_p/k$, ν/α
q	local total heat flux
$\bar{q}^r$	dimensionless local heat flux by radiation, $q^r/\sigma T_\infty^4$
$\bar{q}^c$	dimensionless local heat flux by conduction, $q^c/\sigma T_\infty^4$
T	absolute temperature
u	velocity component in the x-direction, parallel to the plate
U_∞	free stream velocity
v	velocity component in the y-direction, normal to the plate
x	distance measured along surface
y	distance measured normal to surface
α	thermal diffusivity, $k/\rho c_p$
δ	boundary layer thickness
ε	perturbation parameter, $\sqrt{4N\xi Pr}$
η	dimensionless normal coordinate, $y\sqrt{U_\infty/\nu x}$
Θ	dimensionless temperature, inner problem, θN
θ	dimensionless temperature, outer problem, T/T_∞
ξ	dimensionless radiation-to-convection parameter, $a\sigma T_\infty^3 x/\rho c_p U_\infty$, also used as dimensionless axial coordinate
ρ	density
σ	Stefan-Boltzmann constant
τ	optical distance, $a \cdot y$
λ	T_w/T_∞
μ	absolute viscosity
ν	kinematic viscosity

Subscripts

in	inner layer
out	outer layer
w	wall conditions
∞	free stream conditions
0	zeroth order perturbation
1	first order perturbation

Superscripts

c	conduction or convection
r	radiation
t	total
'	denotes differentiation with respect to η

INTRODUCTION

Over the last few years increasing interest has been expressed in problems of combined heat transfer. This interest is attributed to the needs of missile technology, power generation and other industries. The common difficulty encountered in these problems is in solving the governing nonlinear integro-differential energy equation. As known, no general methods have been developed for solving these problems; and it is, therefore, not possible to obtain exact analytical solutions of the combined heat transfer problems without introducing some assumptions and approximations. The most far-reaching assumption which is used in these studies is that the heat transfer by conduction and radiation is one-dimensional. On the other hand, most of the approximations were based on the optical thickness of the gas. The assumption that the gas is either

p. 292

optically thick or optically thin throughout the layer reduces the nonlinear integro-differential equation to an ordinary, but still nonlinear, differential equation which for most cases was solved numerically. Such an approach is physically incorrect, since a radiating gas flowing over a surface should be considered to be optically thin in regions close to the surface and optically thick in regions away from it. Therefore, a solution that combines these two limits is sought. In this study the gas layer is divided into an inner and an outer region. In the inner region the optically thin approximation is used for the radiant term, while the optically thick approximation is used for the outer region. The governing energy equation of both regions is thus reduced to a nonlinear differential equation. The lack of an exact method for solving these types of equations forced us to resort to a form of approximation or to a numerical solution. The latter approach was disregarded since the object of this study is to obtain a closed form analytical solution for the problem at hand. Foremost among these approximation techniques is the systematic method of perturbations in terms of a small or a large parameter present in the equation. The solution of each region is then represented by the first few terms of a perturbation expansion. The solutions are then matched together to form a composite solution valid everywhere in the gas layer.

ASSUMPTIONS

The physical model and coordinate system of the problem is shown in Fig. 1. The case of a radiating, thermally conducting gas flowing in a forced convection over a flat plate is considered. The gas is assumed to be perfect, gray, isotropic, non-scattering and, except for its density, all properties are constant. The plate surface is assumed to be black and isothermal. The flow is assumed to be steady two-dimensional, laminar viscous type. The effect of the viscous dissipation as well as the work of compression are neglected in the energy equation. Also in this study, the gas layer flowing over the plate is divided into two regions: (a) the viscous region, referred to as the inner region, in which the flow is two-dimensional, the gas is optically thin and the temperature gradient is very large.

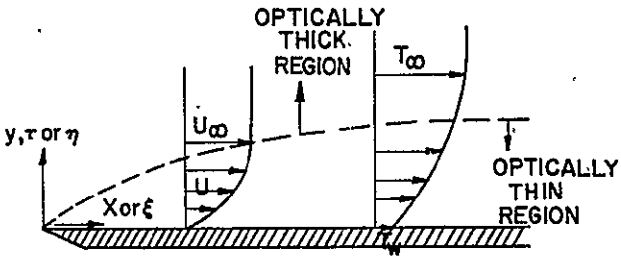


Fig. 1 Physical Model and Coordinate System

These characteristics imply that the radiation emitted by the plate will pass virtually unattenuated through this region, and that conduction is anticipated to play an important role in this region. (b) The free-stream region, referred to as the outer region, where the flow is parallel to the surface, the gas, being considered as infinite in extent, is optically thick and the temperature gradient is not as large as in the vicinity of the surface. Thus, all the radiation emitted by the plate must eventually be absorbed by the gas in this region, and that conduction is not as important as in the previous region. The mathematical model for the assumed physical problem is prescribed by means of the conservation equation of energy, (1) as

$$u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} = \alpha \frac{\partial^2 T}{\partial y^2} - \frac{a}{\rho c_p} \frac{\partial q^r}{\partial \tau} \qquad (1)$$

The terms of the left-hand side of Eq. 1 represent the convection contribution, while the terms on the right-hand side represent the conduction and radiation contribution, respectively. Eq. 1 constitutes a partial nonlinear, integro-differential equation, and even with the simplifications that have been made, a complete solution to the problem is still very difficult. As an alternative to this, Eq. 1 will be applied to each of the two prescribed regions, yielding two limiting solutions that, although restrictive in nature, will yield useful qualitative information.

INNER REGION PROBLEM

The applicable form of the boundary layer energy to the inner region is given by (2) as

$$u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} = \alpha \frac{\partial^2 T}{\partial y^2} + \frac{2a\sigma}{\rho c_p}[T_w^4 + T_\infty^4 - 2T^4] \qquad (2)$$

subject to the following boundary condition:

$$T = T_w \quad ; \quad y = 0 \qquad (3)$$

The other boundary condition is later derived from the matching condition with the outer solution. Since (T_w, T_∞) are constants, it will be convenient to define the following dimensionless temperature ratio

$$\lambda = T_w/T_\infty \quad ; \quad \theta = T/T_\infty$$

then Eq. 2 and 3 become

$$u \frac{\partial \theta}{\partial x} + v \frac{\partial \theta}{\partial y} = \alpha \frac{\partial^2 \theta}{\partial y^2} + \frac{2a\sigma T_\infty^3}{\rho c_p}[\lambda^4 + 1 - 2\theta^4] \qquad (4)$$

$$\theta = \lambda \quad ; \quad y = 0 \qquad (5)$$

Using the familiar Blasius stream function, $f(\eta)$, Eq. 4 reduces to

$$-\frac{f}{2}\frac{d\theta}{d\eta} = \frac{\alpha}{\nu}\frac{d^2\theta}{d\eta^2} + \frac{2a\sigma T_\infty^3}{\rho c_p} \cdot \frac{x}{U_\infty}[\lambda^4 + 1 - 2\theta^4] \qquad (6)$$

Introducing the dimensionless quantities (ξ, Pr, N) into Eq. 6, it reduces to

$$N \frac{d^2\theta}{d\eta^2} + \frac{fNPr}{2}\frac{d\theta}{d\eta} - 4\xi NPr[\theta^4 - \frac{\lambda^4 + 1}{2}] = 0 \qquad (7)$$

Rephrasing Eq. 7 in terms of a new dependent variable, Θ, and the perturbation parameter, ε, which are defined by

$$\Theta = N\theta \quad , \quad \varepsilon = \sqrt{4N\xi Pr} \qquad (8)$$

the following is obtained:

$$\frac{d^2\Theta}{d\eta^2} + \frac{fPr}{2}\frac{d\Theta}{d\eta} - \varepsilon^2[(\frac{\Theta}{N})^4 - (\frac{\lambda^4+1}{2})] = 0 \qquad (9)$$

1.293

The first term on the left-hand side of Eq. 9 represents the conduction contribution, the second term represents the convection contribution and the last term is the radiation contribution. It is clear from Eq. 9 that the radiation contribution is of a smaller order of magnitude with respect to either the conduction or convection contribution. This result was anticipated, since the optically thin approximation was used in this region. The optically thin approximation implies that the radiation contribution from the plate will pass virtually unattenuated through this region, i.e., will be of negligible effect. If $\varepsilon = 0$, Eq. 9 reduces to a pure conduction-convection problem. To determine an improved approximation to the solution of this problem, i.e., to include the effect of the radiation contribution, we seek a perturbation expansion of the form

$$\Theta = \sum_{n=0}^{\infty} (\varepsilon^2)^n \Theta_n = \Theta_0 + \varepsilon^2 \Theta_1 + \ldots \qquad (10)$$

The first term of this expansion, Θ_0, represents the temperature profile for the case of negligible radiation interaction (ε=0), while the second term, Θ_1, denotes the first-order radiation effect upon the temperature profile within the inner region. The series, though not necessarily convergent, is by construction an asymtotic expansion, in that as the perturbation parameter tends to zero, the approximate solution becomes increasingly accurate.

Upon substituting the expansion of Eq. 10 into Eq. 9 and collecting like powers of (ε^2), the following is obtained:

$$\frac{d^2\Theta_0}{d\eta^2} + \frac{fPr}{2}\frac{d\Theta_0}{d\eta} = 0 \qquad (11)$$

$$\frac{d^2\Theta_1}{d\eta^2} + \frac{fPr}{2}\frac{d\Theta_0}{d\eta} = \frac{\Theta_0^4}{N^4} - \frac{\lambda^4+1}{2} \qquad (12)$$

Similarly, the boundary condition of Eq. 5 reduces to the following

$$\Theta_0 = \lambda N \quad ; \quad \eta = 0 \qquad (13)$$

$$\Theta_1 = 0 \quad ; \quad \eta = 0 \qquad (14)$$

Therefore, the problem of the inner region reduces to the solution of Eqs. 11 and 12 subject to the boundary conditions specified in Eqs. 13 and 14, respectively. The solution of the inner region is thus obtained by substituting those results into Eq. 10.

SOLUTION OF THE INNER PROBLEM (ZEROTH ORDER PERTURBATION

The zeroth order perturbation of the inner problem is given by Eqs. 11 and 13, whose solution is

$$\Theta_0 = C_1 \int_0^{\eta} \exp\left(-\frac{Pr}{2}\int_0^{\eta} f\, d\eta\right) d\eta + \lambda N \qquad (15)$$

where C_1 is a constant to be determined from matching with the outer solution.

The Blasius stream function, $f(\eta)$, can be approximately expressed by the following suggested formula:

$$f(\eta) = H_1(3-\eta)(0.166\ \eta^2) + H_2(\eta-3)(\eta-1.7208) \qquad (16)$$

where

$H_1(3-\eta)$ = Heaviside function (step function)
= 0 for $(3-\eta) < 0$ i.e., $\eta > 3$
= 1 for $(3-\eta) \geq 0$ i.e., $\eta \leq 3$
$H_2(\eta-3)$ = Heaviside function
= 0 for $(\eta-3) \leq 0$ i.e., $\eta \leq 3$
= 1 for $(\eta-3) > 0$ i.e., $\eta > 3$

A comparison between the assumed and the actual representation of $f(\eta)$ is given in Fig. 2. From this comparison it is apparent that the analytical representation of $f(\eta)$ is good for all values of η lying between 0 and ∞ except for $\eta = 2.5 \rightarrow 3.8$. Substitute Eq. 16 into Eq. 15 and carry out the integration. The following results are obtained for Θ_0 in terms of the actual variable, θ

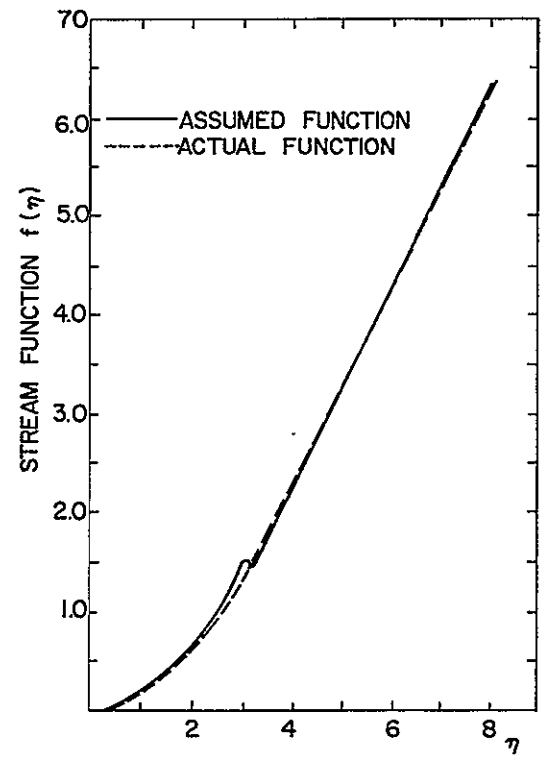


Fig. 2 Comparison Between the Actual and the Assumed Value of $f(\eta)$

$$(\theta_0)_{in} = \lambda + \frac{C_1}{N}\Big(H_1(3-\eta)\Big\{\eta\cdot\exp(-.0276Pr\ \eta^3)\cdot[1 + \frac{.0828Pr\ \eta^3}{4} + \frac{(.0828Pr\ \eta^3)^2}{4*7} + \ldots]\Big\} + H_2(\eta-3)\Big\{\sqrt{\frac{\pi}{Pr}}\cdot\exp(-.337Pr)\cdot[\mathrm{erf}(.86\sqrt{Pr}) + \mathrm{erf}\,\frac{\sqrt{Pr}}{2}(\eta-1.72)]\Big\}\Big) \qquad (17)$$

OUTER REGION PROBLEM

The applicable form of the boundary layer energy to the outer region is given by

$$U_\infty \frac{\partial T}{\partial x} = \alpha \frac{\partial^2 T}{\partial y^2} + \frac{4a\sigma}{3\rho c_p}\frac{\partial^2 T^4}{\partial \tau^2} \qquad (18)$$

Using the dimensionless parameters introduced in the inner region, Eq. 18 reduces to

$$\varepsilon^2 \frac{\partial^2\theta}{\partial\tau^2} + \frac{4}{3}\xi\ Pr\ \frac{\partial^2\theta^4}{\partial\tau^2} - \xi\ Pr\ \frac{\partial\theta}{\partial\xi} = 0 \qquad (19)$$

The first term on the right-hand side of Eq. 19 repre-

ORIGINAL PAGE IS
OF POOR QUALITY

sents the conduction contribution, the second term the radiation contribution and the last term is the convection contribution.

It is clear from Eq. 19 that the conduction contribution is of a smaller order with respect to either the convection or the radiation contribution. This result was anticipated, since in this region the temperature gradient is very small compared to that in the vicinity of the plate. If $\varepsilon=0$, Eq. 19 reduces to a pure radiation-convection problem. To determine an improved approximation to the solution of this problem, i.e., which includes the effect of the conduction contribution, we seek a perturbation expansion of the form:

$$\theta = \sum_{n=0}^{\infty} (\varepsilon^2)^n \theta_n = \theta_0 + \varepsilon^2 \theta_1 + \ldots \tag{20}$$

The first term of this expansion (θ_0) represents the temperature profile for the case of negligible conduction interaction ($\varepsilon=0$), while the second term (θ_1) denotes the first-order conduction effect upon the temperature profile within the outer region.

Upon substituting the expansion of Eq. 20 into Eq. 19 and collecting like powers of ε^2, the following is obtained

$$\frac{4}{3}\frac{\partial^2}{\partial\tau^2}(\theta_0{}^4) = \frac{\partial\theta_0}{\partial\xi} \tag{21}$$

$$\frac{16}{3}\frac{\partial^2}{\partial\tau^2}(\theta_0{}^3\theta_1) = \frac{\partial\theta_1}{\partial\xi} - \frac{1}{\xi\ \mathrm{Pr}}\frac{\partial^2\theta_0}{\partial\tau^2} \tag{22}$$

An additional simplification will now be employed; it consists of assuming that the temperature within the outer region is so close to T_∞, that T^4 may be expressed as a linear function of T_∞. This is accomplished by expanding T^4 in a Taylor series about T_∞ and neglecting higher order terms of T, thus

$$\theta_0^4 \simeq 4\theta_0 - 3$$

similarly

$$\theta_0^3 \simeq 3\theta_0 - 2. \tag{23}$$

Upon substituting Eq. 23 into Eqs. 21 and 22, it reduces to

$$\frac{16}{3}\frac{\partial^2\theta_0}{\partial\tau^2} = \frac{\partial\theta_0}{\partial\xi} \tag{24}$$

$$\frac{16}{3}\frac{\partial^2}{\partial\tau^2}\{\theta_1(3\theta_0 - 2)\} = \frac{\partial\theta_1}{\partial\xi} - \frac{1}{\xi\ \mathrm{Pr}}\frac{\partial^2\theta_0}{\partial\tau^2} \tag{25}$$

The boundary conditions associated with the outer region problem are given by

a) $\theta = 1 \quad ; \quad \xi = 0$ (26)

b) $\theta = 1 \quad ; \quad \tau = \infty$ (27)

c) The second boundary condition in the variable (τ) is the temperature jump at the surface, which for a gray fluid and a black surface is given by:

$$\theta^4 - \lambda^4 = \frac{2}{3}\left(\frac{\partial\theta^4}{\partial\tau}\right) \quad ; \quad \tau = 0$$

or in linearized form it reduces to

$$\theta = \frac{\lambda^4+3}{4} + \frac{2}{3}\frac{\partial\theta}{\partial\tau} \quad ; \quad \tau = 0 \tag{28}$$

Upon substituting the expansion of Eq. 20 into Eqs. 26, 27 and 28, the boundary conditions associated with the outer problem reduce to:

$$\theta_0 = 1 \ \text{ at } \ \xi = 0 \ ; \ \theta_0 = 1 \ \text{ at } \ \tau = \infty \tag{29}$$

$$\theta_0 = \frac{\lambda^4+3}{4} + \frac{2}{3}\frac{\partial\theta_0}{\partial\tau} \ \text{ at } \ \tau = 0$$

and

$$\theta_1 = 0 \ \text{ at } \ \xi = 0 \ ; \ \theta_1 = 0 \ \text{ at } \ \tau = \infty \tag{30}$$

$$\theta_1 = \frac{2}{3}\frac{\partial\theta_1}{\partial\tau} \ \text{ at } \ \tau = 0$$

The problem of the outer region thus reduces to the solution of Eqs. 24, 25, 29 and 30, respectively, and substituting the results into Eq. 20.

SOLUTION OF THE OUTER PROBLEM (ZEROTH ORDER PERTURBATION)

The zeroth order perturbation of the outer problem is given by Eqs. 24 and 29. The solution of these equations is obtained through the use of the Laplace transform, which transforms the governing partial differential equation into an ordinary type,

$$\frac{16}{3}\bar{\theta}_{0_{\tau\tau}} - S\bar{\theta}_0 + 1 = 0 \tag{31}$$

subject to

$$\bar{\theta}_0(\infty,S) = \frac{1}{S}$$

$$\bar{\theta}_0(0,S) = \frac{\lambda^4+3}{4S} + \frac{2}{3}\left.\frac{d\bar{\theta}_0}{d\tau}\right|_{\tau=0} \tag{32}$$

The solution of Eqs. 31 and 32 is given by

$$\bar{\theta}_0 = \frac{1}{S}\left(1 + \frac{\lambda^4-1}{4}\exp\left(-\frac{\sqrt{3S}}{4}\tau\right)\right) - \left(\frac{\lambda^4-1}{4}\frac{1}{\sqrt{S}(2\sqrt{3}+\sqrt{S})}\exp\left(-\frac{\sqrt{3S}}{4}\tau\right)\right) \tag{33}$$

The inverse of Eq. 33 is given by:

$$(\theta_0)_{out} = 1 + \frac{\lambda^4-1}{4}\cdot\left(\mathrm{erfc}\,\frac{\tau\sqrt{3}}{8\sqrt{\xi}} - \exp\left(\frac{3}{2}\tau + 12\,\xi\right)\cdot \mathrm{erfc}\left(\frac{\tau\sqrt{3}}{8\sqrt{\xi}} + 2\sqrt{3\xi}\right)\right)$$

or, in terms of the inner variable, η,

$$(\theta_0)_{out} = 1 + \frac{\lambda^4-1}{4}\left(\mathrm{erfc}\,\frac{\eta\sqrt{3\mathrm{PrN}}}{4} - \exp\left(\frac{3}{2}\eta\xi + 12\xi\right)\cdot \mathrm{erfc}\left(\frac{\eta\sqrt{3\mathrm{PrN}}}{4} + 2\sqrt{3\xi}\right)\right) \tag{34}$$

MATCHING OF THE INNER AND OUTER SOLUTIONS

The method described in (4,5) is used to match the solutions of the inner and outer problem, represented by Eqs. 17 and 34, respectively. This method briefly states that

$$\lim_{\eta\to\infty}(\text{inner solution}) = \lim_{\tau\to 0}(\text{outer solution})$$

where η and τ are the inner and outer independent variables, respectively. The relation between η and τ is given as follows:

since $$\eta = \frac{y}{\delta} = \frac{y}{\sqrt{\frac{\nu x}{U_\infty}}} = \frac{\tau}{a\sqrt{\frac{\nu x}{U_\infty}}} \tag{35}$$

p. 295

and $$N\xi = \frac{ka}{4\sigma T_\infty^3}\frac{a\sigma T_\infty^3}{\rho c_p}\frac{x}{U_\infty} = \frac{1}{4Pr}\left(\frac{\nu x}{U_\infty}\right)a^2 \qquad (36)$$

Substituting the value of $\left(\frac{\nu x}{U_\infty}\right)$ of Eq. 36 into Eq. 35, it reduces to

$$\eta = \frac{\tau}{\sqrt{4N\xi Pr}}$$

therefore,

$$\eta = \frac{\tau}{\varepsilon} \qquad (37)$$

The constant (C_1) of Eq. 17 is obtained through the matching of the two solutions resulting in the following expression for the inner solution:

$$(\theta_0)_{in} = \lambda + \left(\sqrt{\frac{Pr}{\pi}}\frac{\exp(.337\ Pr)}{[1+\mathrm{erf}(.86\ Pr)]}\right)\Big((1-\lambda) + \frac{\lambda^4-1}{4}(1-\exp 12\,\xi\cdot\mathrm{erfc}\ 2\sqrt{3\xi})\Big)\cdot\Big(H_1(3-\eta)\cdot\{\exp(-.0276\ Pr\,\eta^3)\eta\big(1+\frac{.0828\ Pr\,\eta^3}{4} + \frac{(.0828\ Pr\ \eta^3)}{4*7} + \ldots\big)\} + H_2(\eta-3)\{\exp(-.337\ Pr)\cdot\sqrt{\frac{\pi}{Pr}}\big(\mathrm{erf}(.86\sqrt{Pr}) + \mathrm{erf}\ \frac{\sqrt{Pr}}{2}(\eta-1.72)\big)\}\Big) \qquad (38)$$

THE COMPOSITE SOLUTION (ZEROTH ORDER PERTURBATION)

As discussed previously, the outer solution is not valid near the surface of the plate, while the inner solution is not valid in general away from the region $\eta = O(\varepsilon)$. To determine an expansion valid over the whole interval, a composite solution for (θ_0) is formed according to (4,5) as follows:

$$(\theta_0)_{comp} = (\theta_0)_{in} + (\theta_0)_{out} - \big((\theta_0)_{in}\big)^{out} = (\theta_0)_{in} + (\theta_0)_{out} - \big((\theta_0)_{out}\big)^{in} \qquad (39)$$

where

$\big((\theta_0)_{in}\big)^{out} = \lim_{\eta\to\infty}(\theta_0)_{in}$ denotes the outer limit of the inner solution

$\big((\theta_0)_{out}\big)^{in} = \lim_{\tau\to 0}(\theta_0)_{out}$ denotes the inner limit of the outer solution.

The two forms of Eq. 39 are equivalent, since the matching principle requires that

$$\big((\theta_0)_{in}\big)^{out} = \big((\theta_0)_{out}\big)^{in} = 1 + \frac{\lambda^4-1}{4}\big(1-\exp(12\,\xi)\cdot\mathrm{erfc}(2\sqrt{3\xi})\big) \qquad (40)$$

Substituting Eqs. 34, 38 and 40 into 39, the following form of the composite solution of the zero order perturbation is obtained:

$$\theta_0 = \lambda + \left(\sqrt{\frac{Pr}{\pi}}\frac{\exp(.337Pr)}{1+\mathrm{erf}(.86\sqrt{Pr})}\right)\Big((1-\lambda) + \frac{\lambda^4-1}{4}\cdot(1-\exp(12\,\xi)\cdot\mathrm{erfc}\ 2\sqrt{3\xi})\Big)\Big(H_1(3-\eta)\{\exp(-.0276Pr\eta^3)\eta\cdot\big(1+\frac{.0828\ Pr\ \eta^3}{4} + \frac{(.0828\ Pr\ \eta^3)^2}{4*7} + \ldots\big)\} + H_2(\eta-3)\{\exp(-.337Pr)\sqrt{\frac{\pi}{Pr}}\big[\mathrm{erf}(.86\sqrt{Pr}) + \mathrm{erf}\frac{\sqrt{Pr}}{2}(\eta-1.7208)\big]\}\Big) + \Big(1+\frac{\lambda^4-1}{4}\{\mathrm{erfc}\frac{\eta\sqrt{3PrN}}{4} - \exp(\frac{3}{2}\eta\varepsilon + 12\,\xi)\cdot\mathrm{erfc}\frac{\eta\sqrt{3PrN}}{4} + 2\sqrt{3\xi})\}\Big) - \Big(1+\frac{\lambda^4-1}{4}\{1-\exp(12\xi)\cdot\mathrm{erfc}(2\sqrt{3\xi})\}\Big) \qquad (41)$$

ACCURACY OF THE APPROXIMATE METHOD

The problem of the combined mode of heat transfer described above was solved numerically in (6). The exact, rather than the approximate, forms of the radiant term were used in Eq. 1, i.e., the solution was obtained for the governing non-linear integro-differential energy equation. The temperature profile obtained using both the numerical method as well as the approximate solution described by Eq. 41 is plotted in Fig. 3 for the following characteristic parameters: $Pr = 1$, $\xi = .1$, $\lambda = .2$, $N = 2.5\times10^{-6}$. It is clear from Fig. 3 that the two results are in good agreement for values of η less than four but that they diverge from each other for higher values of η. The divergence

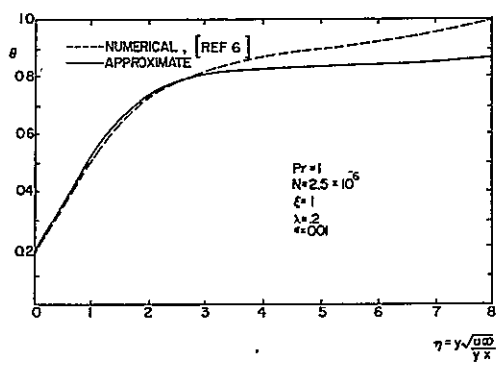


Fig. 3 A Comparison of Temperature Profiles Obtained By the Perturbation Approximation and By the Numerical Analysis

in the temperature profile between the two solutions was anticipated due to the different boundary conditions used in each study. Oliver and McFadden (6) assumed that T_∞ is reached at the edge of the boundary layer, while this study assumed it to be reached at infinity. It is, therefore, evident that the analytical approach developed in this study for problems pertaining to a radiating, thermally conducting medium flowing over a flat plate is accurate. It is also evident, for the problem at hand, that the zeroth order perturbation solution is sufficient to describe the temperature profile in the gas. The results of Eq. 41 were also compared in Fig. 4 with those obtained by (7) using the optically thick approximation throughout the flow region. It is obvious from Fig. 4 that the optically thick approximation overestimates the temperature profile in the region near the plate surface. Having obtained a closed form expression for the temperature distribution in terms of the parameters (Pr, N, ξ, λ), it is now possible to derive the heat flux equations.

HEAT FLUX EQUATIONS

The local heat transfer from the plate surface, being the sum of a conduction and a radiation term, may now be expressed in terms of the temperature pro-

p. 296

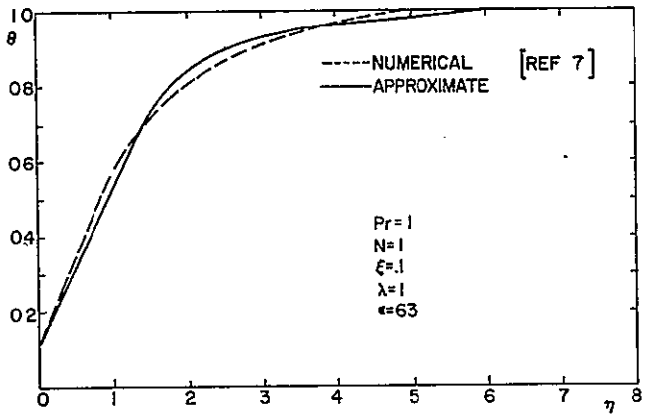


Fig. 4 A Comparison of Temperature Profiles Obtained By the Perturbation Approximation and the Numerical Analysis

file derived in Eq. 40, as follows:

$$q^t = q^c + q^r \tag{42}$$

Dividing both sides of this equation by (σT_∞^4), it reduces to the following non-dimensional form:

$$\bar{q}_w^{\,t} = \bar{q}_w^{\,c} + \bar{q}_w^{\,r} \tag{43}$$

where $\bar{q}_w^{\,r}$ is the normalized heat transfer by radiation and is obtained as follows:

$$\bar{q}_w^{\,r} = \frac{q_w^{\,r}}{\sigma T_\infty^4} = \lambda^4 - 2\int_0^\infty \theta^4(\eta,\xi)\cdot E_2(\eta)\cdot d\eta \tag{44}$$

An exact evaluation of Eq. 44 for the heat transfer by radiation is very tedious and complicated. Two assumptions are made in the evaluation:

a) The exponential integral function is approximated by

$$E_1(\eta) \simeq 2\cdot\exp(-2\eta)$$

therefore,

$$E_2(\eta) = -\int E_1(\eta)d\eta \simeq \exp(-2\eta).$$

Figure 5 shows a comparison between the exact and the approximate values of the function.

b) The function $(\theta)^4$ is very lengthy and complicated not to mention its integral,

$$\int_0^\infty (\theta^4)\cdot E_2\cdot d\eta.$$

To obtain the heat transfer by radiation from the plate surface, approximations become necessary. Since (θ^4) contains both the inner and outer solutions of the temperature distribution, it cannot be automatically linearized with respect to (T_∞), as previously done in the outer solution, $\theta^4_{out} \simeq 4\theta - 3$. However, due to the fact that the temperature distribution of the medium in the inner layer is much closer to Tw than it is to T_∞, its contribution to the total heat radiation might be neglected. This means that θ_{out} can be substituted for θ in the integral and the function $(\theta_{out})^4$ linearized with respect to T_∞, as before. These approximations will result in an overestimation of the amount

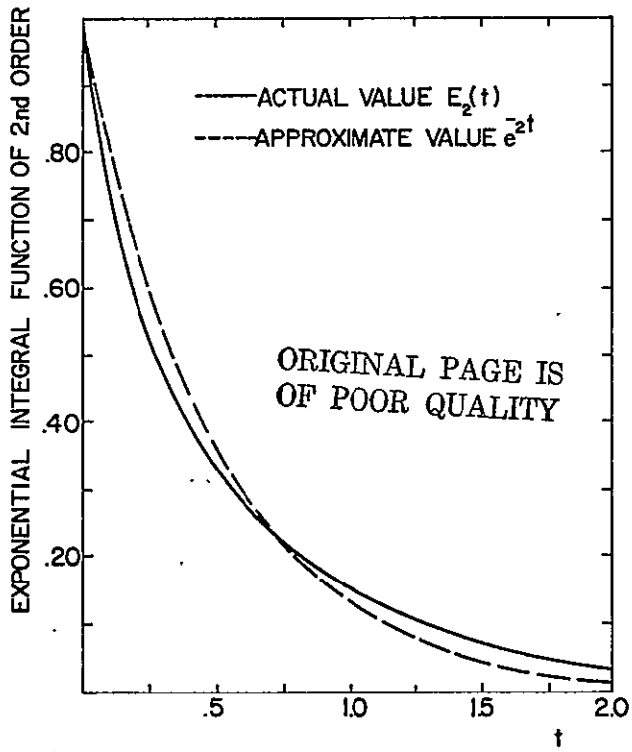


Fig. 5 Comparison Between the Exact and the Assumed Value of $E_2(t)$

of heat radiated from the plane surface.

Equation 44 reduces to

$$\frac{q_w^{\,r}}{\sigma T_\infty^4} = \lambda^4 - 2\int_0^\infty (4\theta_{out} - 3)\cdot\exp(-2\eta)d\eta$$

$$= \lambda^4 - 8\int_0^\infty \theta_{out}\cdot\exp(-2\eta)d\eta + 6\int_0^\infty \exp(-2\eta)d\eta \tag{45}$$

Therefore,

$$\bar{q}_w^{\,r} = (\lambda^4-1)\Big(\frac{2}{a}\exp(12\xi)\operatorname{erfc} 2\sqrt{3\xi} + (1-\frac{2}{a})\exp(\frac{16}{3PrN})\operatorname{erfc}(\frac{4}{\sqrt{3PrN}})\Big) \tag{46}$$

(where $a = 2 - 3\sqrt{PrN\xi}$).

Similarly, the normalized heat transfer by conduction is obtained as

$$\bar{q}_w^{\,c} = \frac{q_w^{\,c}}{\sigma T_\infty^4} = -k\sqrt{\frac{U_\infty}{\nu x}}\cdot\frac{1}{\sigma T_\infty^4}\Big(\frac{\partial T}{\partial \eta}\Big)_{\eta=0}$$

$$= -k\sqrt{\frac{U_\infty}{\nu x}}\cdot\frac{1}{\sigma T_\infty^3}\Big(\frac{\partial \theta}{\partial \eta}\Big)_{\eta=0} \tag{47}$$

p.297

$$= -\sqrt{\frac{4N}{Pr\xi}} \left(\frac{\partial\theta}{\partial\eta}\right)_{\eta=0} \tag{48}$$

Substituting the value of $(\partial\theta/\partial\eta)_{\eta=0}$ into Eq. 48, it reduces to

$$\bar{q}_w^{\,c} = -\sqrt{\frac{4N}{Pr\xi}} \left\{ \left(.141 \frac{\sqrt{Pr}\,\exp(.337Pr)}{[1+\mathrm{erf}(.86\sqrt{Pr})]} (\lambda^4-4\lambda+3)\right) - \left((\lambda^4-1)\cdot\left[\frac{.141\sqrt{Pr}\,\exp(.337Pr)}{1+\mathrm{erf}(.86\sqrt{Pr})} + .75\sqrt{PrN\xi}\right]\cdot \exp(12\xi)\,\mathrm{erfc}(2\sqrt{3\xi})\right)\right\} \tag{49}$$

TEMPERATURE PROFILE

The effect of the characteristic parameters (N, Pr, ξ, λ) on the temperature profile in the boundary layer is given in Figs. 6 through 9. Various curves are drawn in each figure to cover the typical range of those parameters in gases.

In Fig. 6 the dimensionless parameter, N, is varied between 0 and ∞. The temperature profile for N=1 is found to be very close to that of pure conduction $N = \infty$, however, as N decreases, the difference in the temperature profile widens.

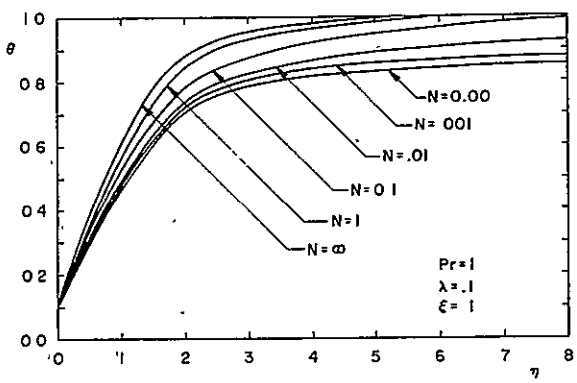


Fig. 6 Variation of the Temperature Profile With the N Parameter

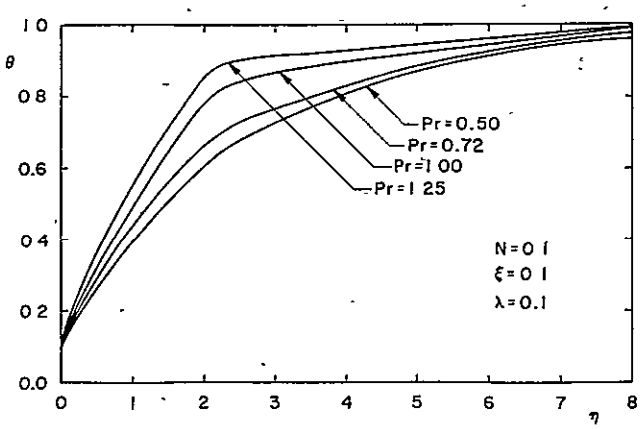


Fig. 7 Variation of the Temperature Profile With the Prandtl Number

In Fig. 8 the dimensionless parameter ξ is varied between .001 and 0.1. The effect of the parameter, ξ, on the temperature profile for the case of N = 1 is found to be on the thickness of the thermal boundary layer. As ξ is increased, i.e., as the radiation mode of heat transfer is increased, it is apparent that the thickness of the thermal boundary layer is decreased.

In Fig. 7 the Prandtl number is varied between 0.5 and 1.25, and in Fig. 9 the parameter λ is varied between 0.5 and 2.

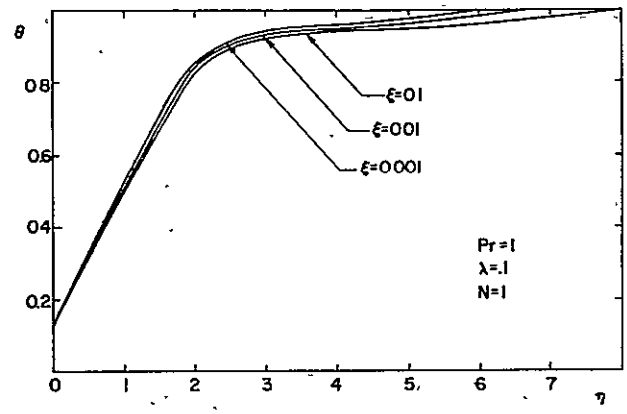


Fig. 8 Variation of the Temperature Profile With the ξ Parameter

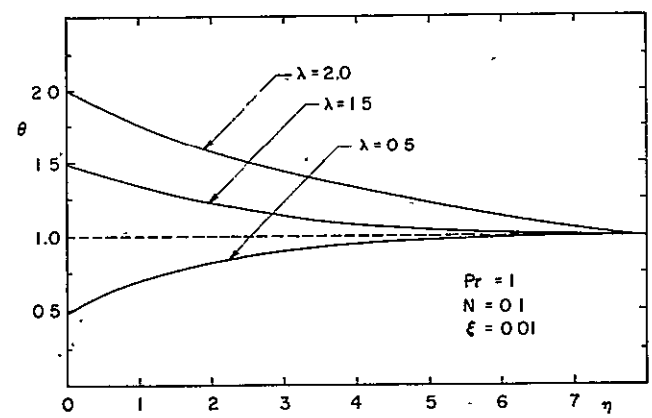


Fig. 9 Variation of the Temperature Profile With the λ Parameter

HEAT FLUX FROM THE SURFACE

The heat transfer calculated for a given set of parameters is illustrated in Figs. 10 through 12.

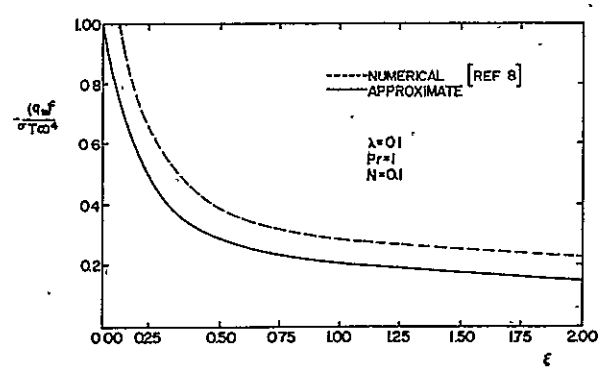


Fig. 10 Conductive Heat Flux From the Wall

p. 298

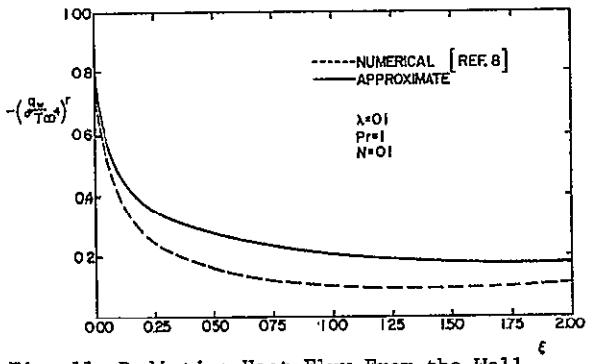


Fig. 11 Radiative Heat Flux From the Wall

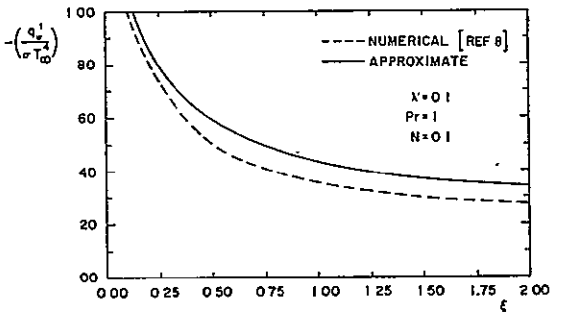


Fig. 12 Total Heat Flux From the Wall

The results obtained for the local conductive and radiative heat fluxes using the approximate method developed in this study are compared with those of (3) for the optically thin conditions. It is apparent that the latter solution predicts consistently less energy transfer by radiation.

CONCLUSION

Analysis of the work done in this study shows that the closed form solutions for the combined heat transfer problems are possible. However, it should be noted that in the process of linearization, only the radiation terms involving $[f(N).\theta^4]$ were linearized while those involving $[f(N).\lambda^4]$ were kept in their non-linear forms. As a result of this partial linearization, the results obtained were valid for all values of λ as long as $N = O(1)$ or larger. However, as (N) decreases, i.e., as radiation becomes progressively dominant, the results obtained are only valid for values of (λ) less than unity. Therefore it is recommended that theoretical studies be continued in order to remove a number of assumptions on which this problem was based. The assumptions to be removed are, in particular, the one-dimensionality of the heat flux due to radiation and conduction, isotropicity and grayness of the media, blackness of the surface and a number of others. The removal of such assumptions shall result in solutions of more practical application.

REFERENCES

1. Schlichting, H., "Boundary Layer Theory", McGraw-Hill Book Co., New York, NY (1960).
2. Howell, J.R., Siegel, R., "Thermal Radiation Heat Transfer", McGraw-Hill Book Co., New York, NY (1972).
3. Viskanta, R., "Radiation Transfer and Interaction of Convection With Radiation Heat Transfer", in Advances in Heat Transfer, Vol. 3 Academic Press, New York, NY (1966).
4. Nayfeh, A.H., "Perturbation Methods", John Wiley and Sons, Inc., New York, NY (1973).
5. Van Dike, M., "Perturbation Methods in Fluid Mechanics", Academic Press, New York, NY (1964).
6. Oliver, C.C., McFadden, P.W., "The Interaction of Radiation and Convection in the Laminar Boundary Layer", Journal of Heat Transfer, (May, 1966), pp. 205-213.
7. Viskanta, R., Grosh, R.J., "Boundary Layer in Thermal Radiation Absorbing and Emitting Media", Int. J. Heat Mass Transfer 5, pp. 795-806 (1962).

ORIGINAL PAGE IS
OF POOR QUALITY

p. 299

Project

(A) Project Title: Solar Energy for Process Heat

(B) Project Abstract:

Maintaining and developing available long and short term energy sources is of great importance. This requirement, coupled with a large potential application for solar heating in the process industry has led to the preparation of this report. Included is a study of the economic feasibility of solar heating for use in the process industry.

This report also includes a discussion of possible configurations and sizing of solar process heating systems including energy storage capacity based on a process heat load and plant operating schedule.

(C) Publication: Thesis for Master of Mechanical Engineering and Presented before International Solar Energy Society Winipeg Conference

(D) Year: 1975

(E) Department: Mechanical Engineering

(F) Student Name: Richard Reimels

(G) Faculty Advisor: Prof. J.R. Howell

SOLAR ENERGY FOR PROCESS HEAT

Presented at ISES Winnipeg Conference, August, 1976.

Richard Reimels
Brown and Root, Inc.
Houston, Texas 77046

John R. Howell
Department of Mechanical Engineering
University of Houston
Houston, Texas 77004

Abstract

Maintaining and developing available long and short term energy sources is of great importance. This requirement, coupled with a large potential application for solar heating in the process industry has led to the preparation of this report. Included is a study of the economic feasibility of solar heating for use in the process industry.

This report also includes a discussion of possible configurations and sizing of solar process heating systems including energy storage capacity based on a process heat load and plant operating schedule.

INTRODUCTION

Solar systems can provide a clean source of energy without the need for burning valuable fossil fuels. Natural gas, oil, etc., are being consumed at increasing rates and as a result becoming scarce and expensive. As these fuels become scarce the need for other energy sources becomes even more critical. Critical in the sense that there are many industrial processes that must use fossil fuels. The production of fertilizer to support our food needs is one example. If we are to conserve our fuels for necessary uses where replacements are not available then we must reduce their use in areas where they can be substituted. Electric power generation and industrial plant process heating are two examples of major fossil fuel uses in which substitutes are available.

Besides solar there are many methods of non-fossil energy generation presently being studied and used, i.e., nuclear power, windpower, geothermal energy, etc. All of these must be further developed and used to meet our future needs. This paper is a study of solar energy in one general application, process plant heating.

This particular project was initiated because it is felt that many process plants throughout the country offer ideal applications

for solar heating systems. Ideal because of the low temperature (low quality) heat processes which are compatible with low temperature (flat-plate collector) solar systems. The chief advantages of solar in these applications are that solar systems are best adapted and more efficient for low temperature heating; solar systems do not have exhaust fumes or pollutants, and maybe most important, solar systems can save our valuable fuels. In support of this idea several process industries have expressed an interest in using solar systems. One such application has been used as a case study for much of the work presented here.

This paper has been prepared as the result of studying and investigating the state-of-the-art of solar design. These investigations help in determining a practical starting point for the basic solar design discussed. This study does not attempt to cover the details of solar theory or collector design. These areas are covered in References [1] and [2] as well as in other literature.

This paper will concentrate on overall solar system design using typical solar collectors in conjunction with standard fluid heating systems. This system is thermodynamically modeled and sized by the computer for a given process heat load to determine economic feasibility of solar systems. Results, for the case study, are presented in graphical form.

PROCESS APPLICATION

Before a model of any heating system can be adequately analyzed an understanding of both the system to be heated (process system) and the system used for heating (solar system) must be obtained. The functional, operating, and design requirements of each system must be known.

The best process application for use with solar systems would be low temperature heating, herein called low quality heat. This is based on the fact that flat-plate solar collectors are limited to temperatures of below 400°F and are more efficient (require less collector area) at lower operating temperatures. The case study application is a good example of this kind of process. The process studied requires heat at a constant 150°F to vaporize a liquid substance.

Although solar systems are adaptable they also have several problems when used in conjunction with process plants that must be evaluated and considered. The two major ones are the process operating schedule and the physical size of the collectors.

Most process plants operate on a twenty-four hour per day schedule and unfortunately the sun does not. Therefore to utilize solar energy in large process systems for total replacement would require extremely large energy storage systems along with extremely large collector systems. As it turns out this is impractical if not impossible.

However, solar can be used as a supplement, and a short-time storage system allows the solar collectors to be oversized for the job. Therefore when the sun is just rising or on partially cloudy days there is extra capacity within the solar system to allow for some operation. In addition on clear sunny days the extra energy collected by the collectors can be added to the storage system. This provides two additional desirable functions. First, on partially cloudy days the storage system will allow continuous operation during periods when the sun is hidden behind the clouds and it also will allow for some additional operating time after sunset. All of these factors are considered in the analyses when sizing a system.

The second major consideration is the required collector area. In many cases although a solar system may be economically justified and functionally capable, the space required for the collectors may not be available or feasible. The size of collectors versus the economic feasibility are also evaluated as described later.

After defining the process and solar systems functional requirements and limitations, the solar system configuration to be evaluated must be determined. Basically there are two general configurations, a direct system (Figure 1) and an indirect system (Figure 2). Both types of systems have advantages. The indirect system is used here for three reasons.

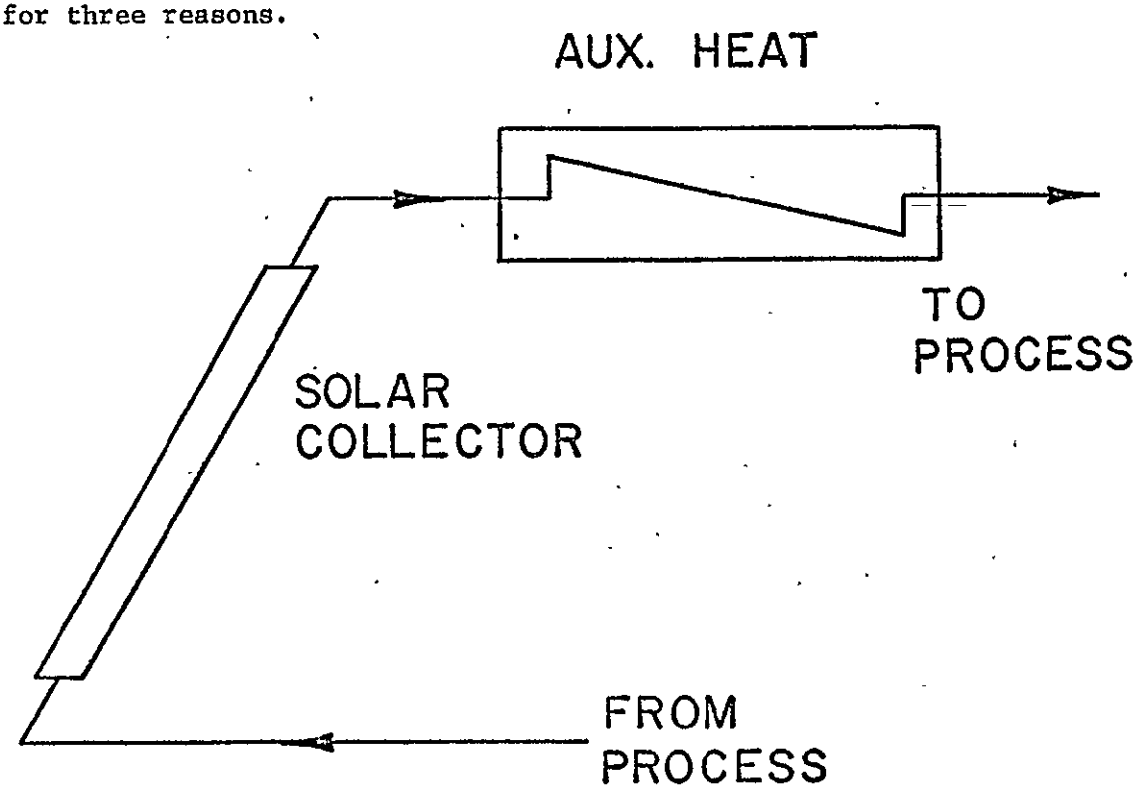


FIGURE 1: DIRECT SOLAR SYSTEM

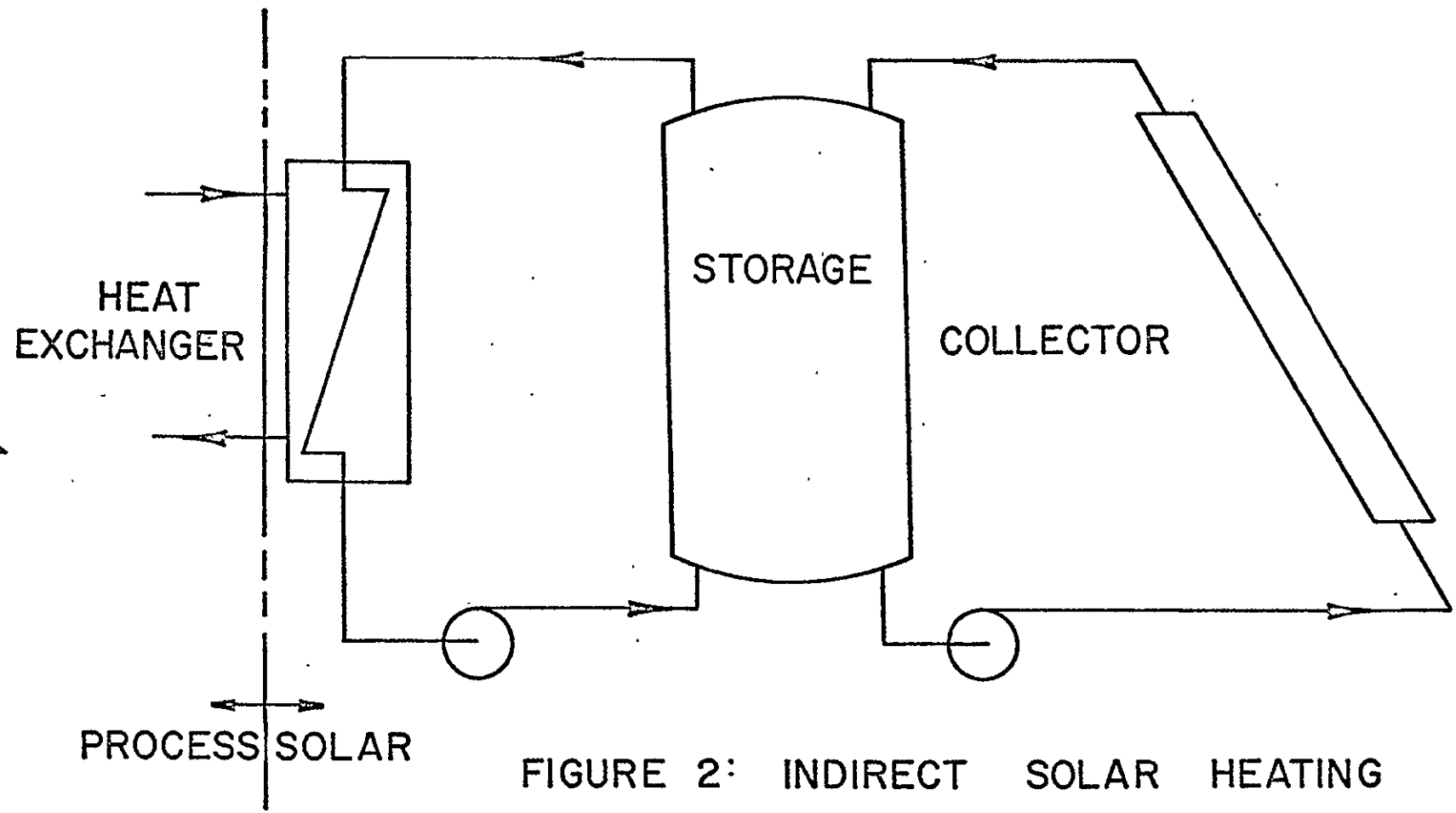


FIGURE 2: INDIRECT SOLAR HEATING

First, the indirect system allows the use of a storage system while the direct system does not. The advantages and reasons for storage for the process applications were discussed above. Second, many process systems, such as that of the case study, are heating substances other than water. Very little collector design or performance data is available for any substances other than water. Therefore, based on present solar technology, the indirect system can be more accurately analyzed. The third reason is that the indirect system is more conservative from the type of equipment needed. The system requires extra pumps, storage tank and solar-to-process heat exchanger.

COMPUTER MODEL AND ANALYSIS

Introduction

The basic system used in the model consists of a flat-plate collector, storage tank, pumps, piping, valves and controls, seen in Figure 3. Water is the pumping medium used to transport heat from the collector to the storage tank and then from the storage tank through a heat exchanger where the heat is transfered for use in the process. The collector-side water is circulated form a water-to-water mixing storage tank, whereas the solar system and process system fluids are separated by a shell-and-tube type heat exchanger. Any auxiliary heat required to supplement the process is added to the process side of the system. The actual computer program is divided into two subprograms, thermal performance analysis and an economic analysis. The results of the thermal performance are in the form of fuel savings, these savings being a primary input to the economic evaluation. The basic features including method of approach, inputs, assumptions and outputs for each subprogram are discussed in the following paragraphs.

Thermal Performance

The indirect solar system, Figure 3, was modeled for use in the case study. For the thermal analysis each component was represented in a system heat balance, eqn 1, with the storage tank temperature being used for system operating limits and controls.

$$Q_s = Q_{sol} - Q_p - Q_{loss} \tag{1}$$

where

Q_p = energy used in the process

Q_{sol} = energy collected by solar system

Q_s = energy in storage

Q_{loss} = energy loss to environment.

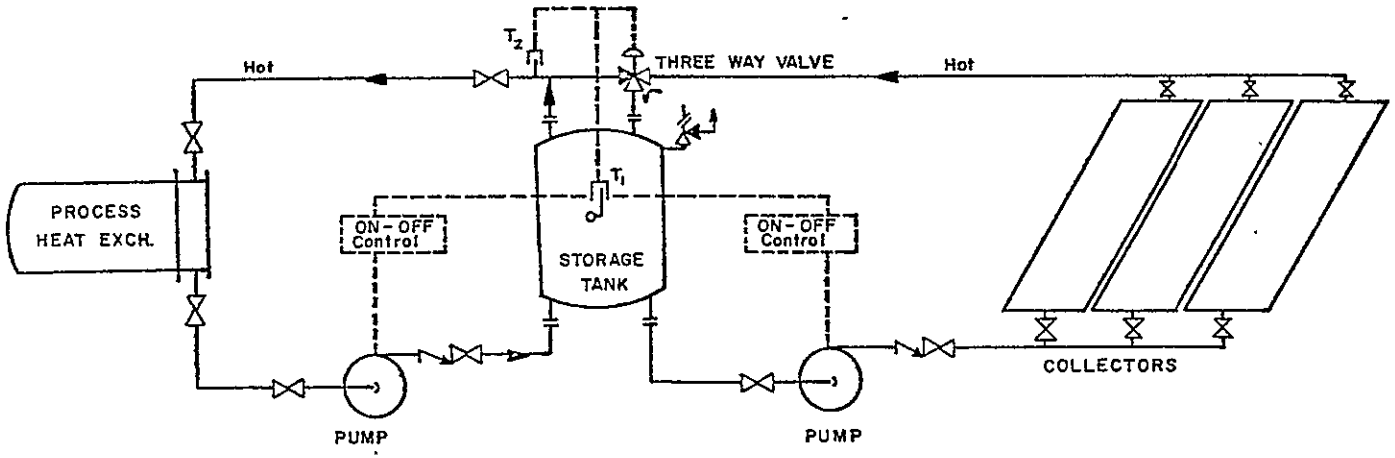


FIGURE 3: INDIRECT SOLAR SYSTEM MODEL

The model control scheme is similar to an actual control scheme. When the storage tank temperature rises above a preset minimum the process-side pump is automatically started and heat is transported via the solar-to-process heat exchanger. This transfer of heat is represented by:

$$Q_p = M_h C_p (T_s - T_{h2}) \tag{2}$$

where

M_h = flow through the heat exchanger

C_p = specific heat of water

T_s = storage tank (heat exchanger inlet) temperature

T_{h2} = heat exchanger outlet temperature.

Whenever the storage tank temperature is below a preset maximum and the solar collector temperature measured at the outlet of the collector is above a preset minimum the collector circulating pump is automatically started to transfer heat to the storage system for use in the process. This transfer of heat is represented by:

$$Q_{sol} = M_c C_p (T_{c2} - T_{c1}) \tag{3}$$

where

M_c = flowrate through the collector

C_p = specific heat of water

T_{c2} = collector outlet temperature

T_{c1} = collector inlet temperature

A more qualitative discussion of the collector performance is given in the next section.

The computer model uses the above heat balance with typical insolation data as input to calculate a useful heat output. The output is the actual amount of solar heat transported and used in the process stream. The heat is then represented as a quantity of fuel saved to determine the economic feasibility of the system.

Storage Model

The storage used with the indirect system is simply a large insulated tank. The heat input from the collector is mixed with cooler water (heat output) from the solar-to-process heat exchanger. Ideally it would be desireable to size the collector and storage system to be able to operate the process system for most, if not all of a twenty-

four day. However, with a large process heat load this would require storage tanks of impractical size. It has been determined as a result of this analysis that for most large heat loads (50×10^6 Btu/hr) the storage system can be used only to stabilize the solar system operation. That is, the storage tank is only sized to operate the process for short periods of times, i.e., one-half hour to two or three hours without any solar input. This time period is based on expected average cloudy periods for a particular area. The tank size is estimated from

$$Q_p = M_s \; C_p \frac{dT}{dt} \tag{4}$$

where

M_s = mass (size) of the storage tank

dT = design storage tank temperature range

dt = design time increment for dT.

It can be seen that the storage tank control temperature limits discussed above have a direct effect on the size of the storage tank.

A physical representation of the storage tank is shown schematically in Figure 4 with the collector outlet and process heat exchanger inlet going to and from the top of the tank respectively. and the collection inlet taken from the bottom of the tank and the heat exchanger outlet returned to the bottom. That is, hot fluids enter and leave at the top and cold at the bottom.

Assuming large storage and hence relatively small velocities within these tanks a degree of stratification will occur, that is, the water temperatures of the tank will vary from top to bottom depending on the density.

In order to analyze this effect the storage tank has been divided into N horizontal sections. With the energy balance of eqn 1 for the ith layer of the tank can be written (Reference [1]):

$$\begin{aligned} M\,C_p = {} & M_c\,C_p[F\,C_i(T_{c_2} - T_i) + (T_{i-1} - T_i) \sum_{j=1}^{i-1} F\,C_j] \\ & + M_h\,C_p[F\,H_i(T_{h_2} - T_i) + (T_{i+1} - T_i) \sum_{j=i+1}^{n} F\,H_j] + Q_{loss,i} \end{aligned} \tag{5}$$

where

$F\,C_i$, $F\,H_i$ = control functions for the collector side and heat exchanger side defined as follows:

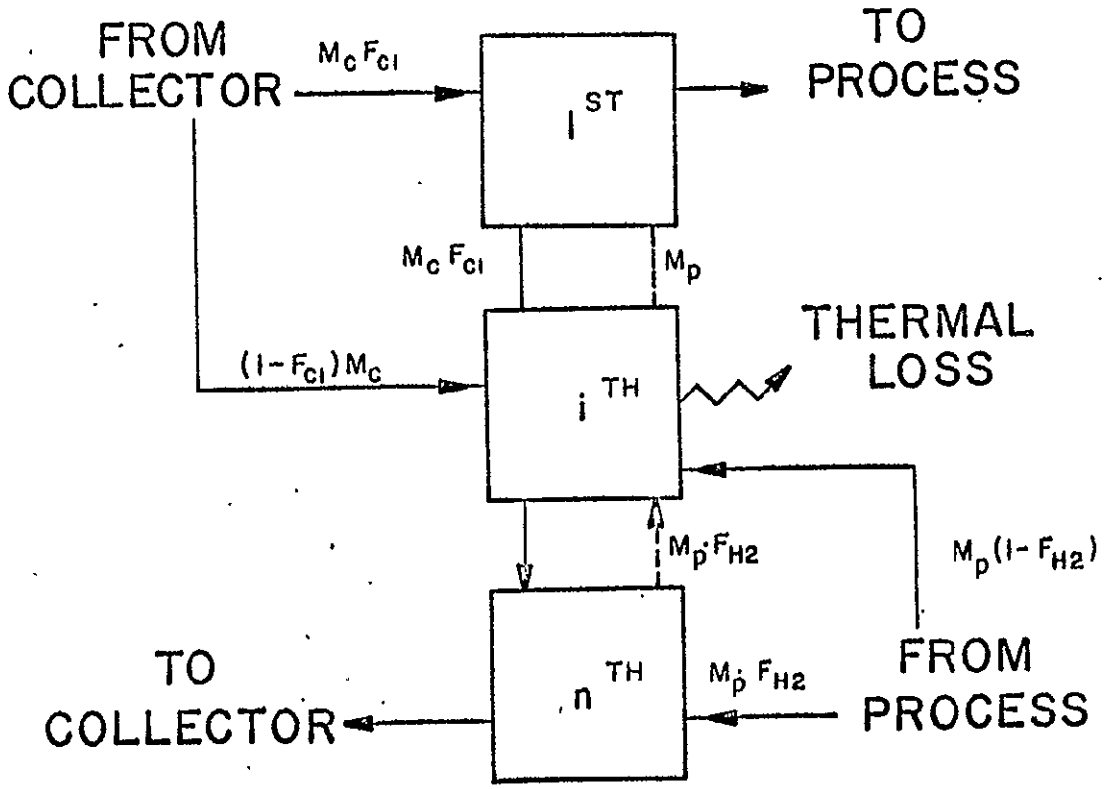


FIGURE 4: STRATIFIED STORAGE MODEL

$$F\,C_i = \begin{cases} 0 \text{ if } T_{i-1} > T_2 > T_i \\ 1 \text{ otherwise} \end{cases} \tag{6}$$

$$F\,H_i = \begin{cases} 1 \text{ if } T_i > T_2 > T_{i+1} \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

and the first term $M_c\ C_p\ \Delta T$ represents the useful solar input $Q_{sol} = M\ C_p\ \Delta T$, the second term represents the heat output $Q_p = M_h\ C_c\ \Delta T$ and Q_{loss} is the system heat loss.

This equation is used to continually calculate the outlet storage temperature based on the inlet solar heat (Q_{sol}) and the output heat (Q_p). As noted above when the storage outlet temperature is less than the minimum allowed the process side is shut off until the tank is reheated.

A typical performance output from the program is shown in Figure 5 (data tabulated in Table 1). For this case, the solar system is assumed to provide the total required process heat as long as the inlet temperature to the heat exchanger is above 150°F. Other data was calculated for the case of the heat exchanger outlet temperature held at 150°F, and the heat to the process being calculated from eqn 2. Again the process heat exchanger inlet temperature was required to be above 150°F. The latter case is more realistic for a reboiler working at 150°F.

Collector Model

The process heat load (Q_p) is a defined number, normally constant for a particular process. The solar energy (Q_{sol}), however, is not constant. Further the solar energy input is dependent on the specific geographical location, time of year, etc. Insolation data is the basic energy input to the collector used to calculate the useful energy collected as follows:

$$Q_{sol} = \eta \times A_{col} \times q_{col} \tag{8}$$

where

Q_{sol} = useful energy collected

A_{sol} = area of the collector

q_{sol} = insolation/area/time

and

$$\eta = A - B(T_p - T_\infty) \tag{9}$$

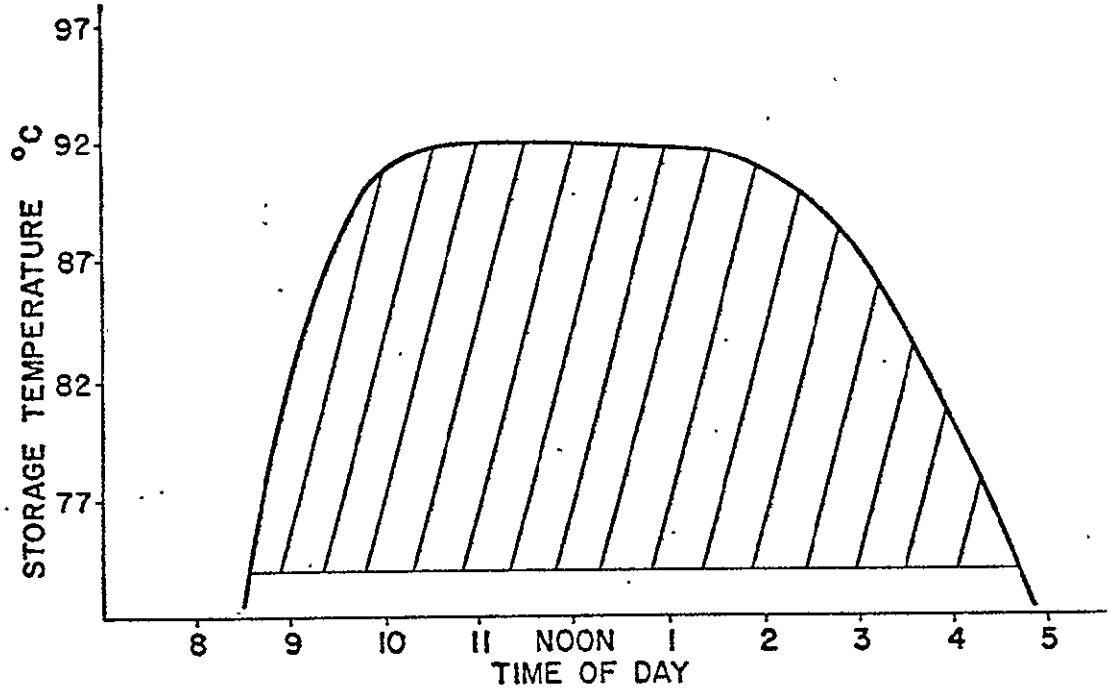


Figure 5
The shaded area represents the area where the process is operated by solar energy.

Table 1 THERMAL PERFORMANCE

Time, Hrs.	Collector Outlet Temp., °F	First Layer Temp.	Last Layer Temp.	Ht.Ex. Outlet Temp.	Solar Heat, $\frac{Btu}{hr/ft^2}$	Collector Efficiency
0.25	46.4	45.7	44.6	45.7	23.5	0.74
0.50	59.9	58.4	56.4	58.4	46.9	0.71
0.75	78.5	76.4	76.3	76.4	70.0	0.67
1.00	101.1	98.6	95.2	98.6	92.7	0.62
1.25	126.5	123.7	119.9	123.7	114.8	0.56
1.50	153.3	150.6	146.3	150.3	135.2	0.58
1.75	157.2	161.8	149.3	111.8	156.7	0.49
2.00	156.6	158.8	147.5	108.3	176.3	0.50
2.25	162.0	162.5	152.2	112.5	194.8	0.49
2.50	165.0	164.7	154.4	114.7	212.1	0.49
2.75	170.3	169.6	159.0	119.6	228.1	0.48
3.00	177.0	176.2	165.4	126.2	242.7	0.47
3.25	184.7	183.7	172.8	133.7	255.8	0.45
3.50	192.7	191.7	180.7	141.7	267.3	0.43
3.75	200.5	199.6	188.6	140.6	277.2	0.42
4.00	200.5	199.7	188.0	149.7	285.3	0.42
4.25	189.2	200.6	189.2	150.6	291.7	0.42
4.50	184.8	200.5	184.8	150.5	296.3	0.44
4.75	198.8	199.8	184.9	149.8	299.1	0.45
5.00	177.4	200.1	177.4	150.1	300.0	0.46
5.25	197.7	199.9	184.1	140.9	299.1	0.44
5.50	201.0	199.7	188.0	149.7	298.3	0.42
5.75	199.3	199.7	186.5	149.7	91.7	0.43
6.00	199.1	199.1	186.7	149.1	285.3	0.42
6.25	184.9	200.1	184.9	150.1	277.2	0.42
6.50	197.4	199.7	185.9	149.7	267.3	0.42
6.75	196.3	200.0	185.4	150.0	255.8	0.41
7.00	199.0	198.9	198.0	148.9	242.7	0.40
7.25	198.8	197.7	187.4	147.7	228.1	0.40
7.50	193.1	195.4	184.3	145.4	212.1	0.41
7.75	187.9	192.0	179.5	142.0	194.8	0.41
8.00	181.1	187.1	173.4	137.1	176.3	0.43
8.25	173.0	181.0	165.8	131.1	156.7	0.44
8.50	173.4	173.5	156.9	123.5	136.2	0.46
8.75	153.4	164.7	147.8	114.7	114.8	0.48
9.00	140.2	155.0	135.3	105.0	92.7	0.51
9.25	142.7	149.3	139.1	149.3	70.0	0.50
9.50	152.4	151.4	150.1	151.4	46.9	0.47
9.75	155.7	155.2	154.6	155.2	23.5	0.46
10.00	154.0	155.5	154.9	155.5	- 0.0	0.45

Flow = 5.00000E 06 lb_m/hr Number of Thermal Layers in Tank = 4

HFlow = 1.00000E 06 lb_m/hr Tank Size = 71899 gallons

where

η = collector efficiency

A,B = constants determined for a particular collector design

T_p = average collector plate temperature

T_∞ = ambient temperature.

This technique of modeling collectors by calculating efficiencies is based on methods similar to those developed in Reference [1] for flat-plate collectors. It has the advantage of allowing close approximations to actual collector performance.

The efficiency of a collector plotted against $(T_p - T_\infty)$ will give a curve as shown in Figure 6 by the dashed line. If a straight line approximation is used for actual or calculated data from a specific collector the coefficient A and B can be determined from the equation of the straight line:

$$y = A - Bx$$

by standard mathematical procedure where $y = \eta$ and $x = (T_p - T_\infty)$.

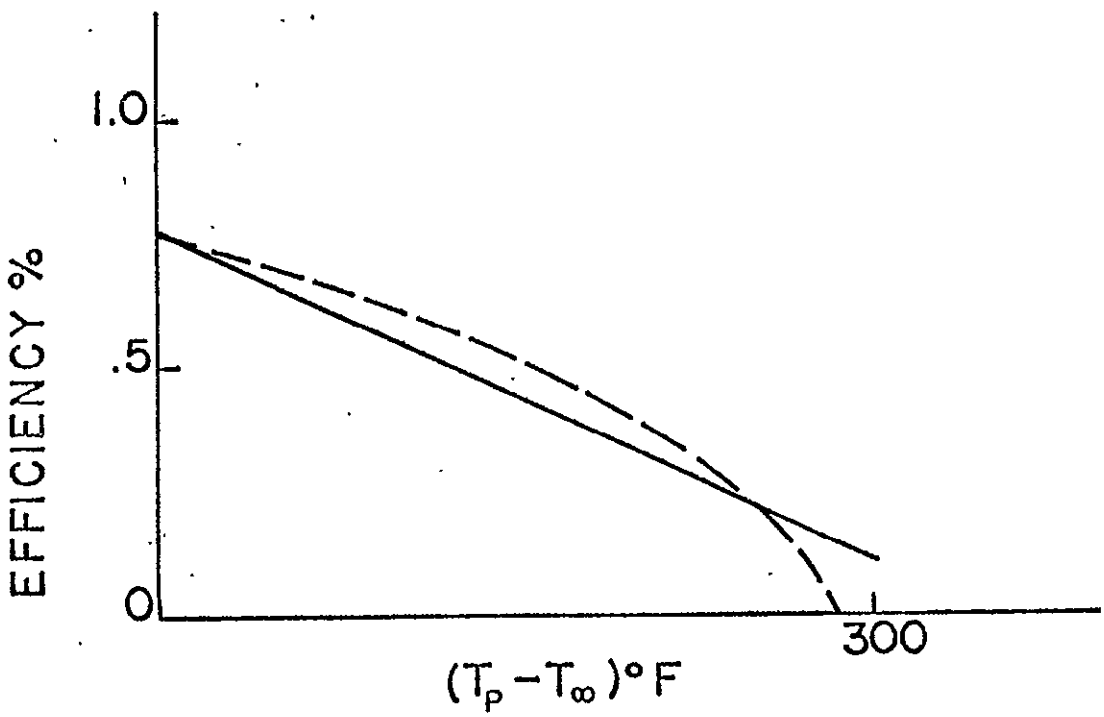


FIGURE 6. COLLECTOR EFFICIENCY

These coefficients are input for the specific type(s) of collector to be analyzed by the program. The computer then calculates collector efficiencies based on $(T_p - T_\infty)$ and uses the results to calculate the useful energy gain from eqn 8.

The useful solar energy, Q_{sol}, is then used in eqn 5 in the general form:

$$Q_{sol} = M_c \, C_p (T_{c2} - T_{c1}) \tag{10}$$

where

M_c = flow through the collector

C_p = specific heat

T_{c2} = collector outlet temperature

T_{c1} = collector inlet temperature.

ECONOMIC EVALUATION

Once the system is sized and the useful energy saved is calculated, the computer program determines the total cost of each solar system and compares it to the cost of fuel saved. Knowing that a system saves money is not necessirly justification alone for installing it. During the technical discussion it was shown that a solar system can only supplement a process system. Therefore, it is also important to know what percentage of the total annual heat load the solar system will handle and also how much fuel it will save. Remember that the saving of valuable fossil fuels is one major objective in using solar. The economic program provides all this data.

In determining the economical feasibility of any system alternative the designer must know the present cost of money (interest rate) and payback period for his company plus he must have available the cost of equipment, fuels, labor, maintenance, etc. In most cases these figures are available. However, when evaluating a technically new product such as solar collectors, cost data may not be readily available or accurate. In performing the analysis, the question of importance becomes what cost per square foot of solar collector can be economically justified and are collectors available at that price? This becomes important because the solar collector costs are the overriding factor in the overall annual costs of a solar process system. This will be shown more clearly in the next sections.

Cost Analysis

The annual total cost of the solar system, including installation, maintenance and operating cost, are calculated by the following expression:

$$\text{Total Cost} = [c_{coll} \times A_{coll} + C_{storage} + C_{pumps} + C_{install} + C_{equip}] \times CRF + E_{pump} + C_{maint} + C_{labor} \quad (11)$$

where

c_{coll} = cost of the collector $\$/ft^2$

A_{coll} = area of the collector

$C_{storage}$ = cost of storage tank

C_{pump} = cost of pump

C_{equip} = cost of piping, valves, instruments, and insulation

$C_{install}$ = cost of installation

CRF = capital recovery factor

E_{pump} = annual energy charge to operate the pumps

C_{maint} = annual cost of maintenance

C_{labor} = annual cost of labor.

The results of the cost analysis show two major items. First, the cost of the solar collectors is by far the major cost of the system. The price of pumps, piping, storage, etc., has a small effect on the results of the analysis. Second, today's cost of fuel requires that the collector costs be lower than $6.00 per square foot for economical justification of the system.

Figure 7 shows results of the economic analysis for the case study. Each curve is plotted for annual total costs for varying sized systems versus the fuel costs savings per million Btu. The horizontal dashed line represents the present day cost of fuel per million Btu (intra-state recent contract price). It can be seen that at collector costs of $4.00 per square foot the solar system can be economically justified. Any point on these curves below the horizontal fuel cost line represents an economically justified system. The low point of the curve is the most economical system. However, the intersection of the two lines on the right side of the figure represents the system which saves the greatest amount of auxiliary fuel. However, the final selection of the solar system size may not be either of these two points but may be based upon the physical space available. For example, on curve II the area varies from 13 to

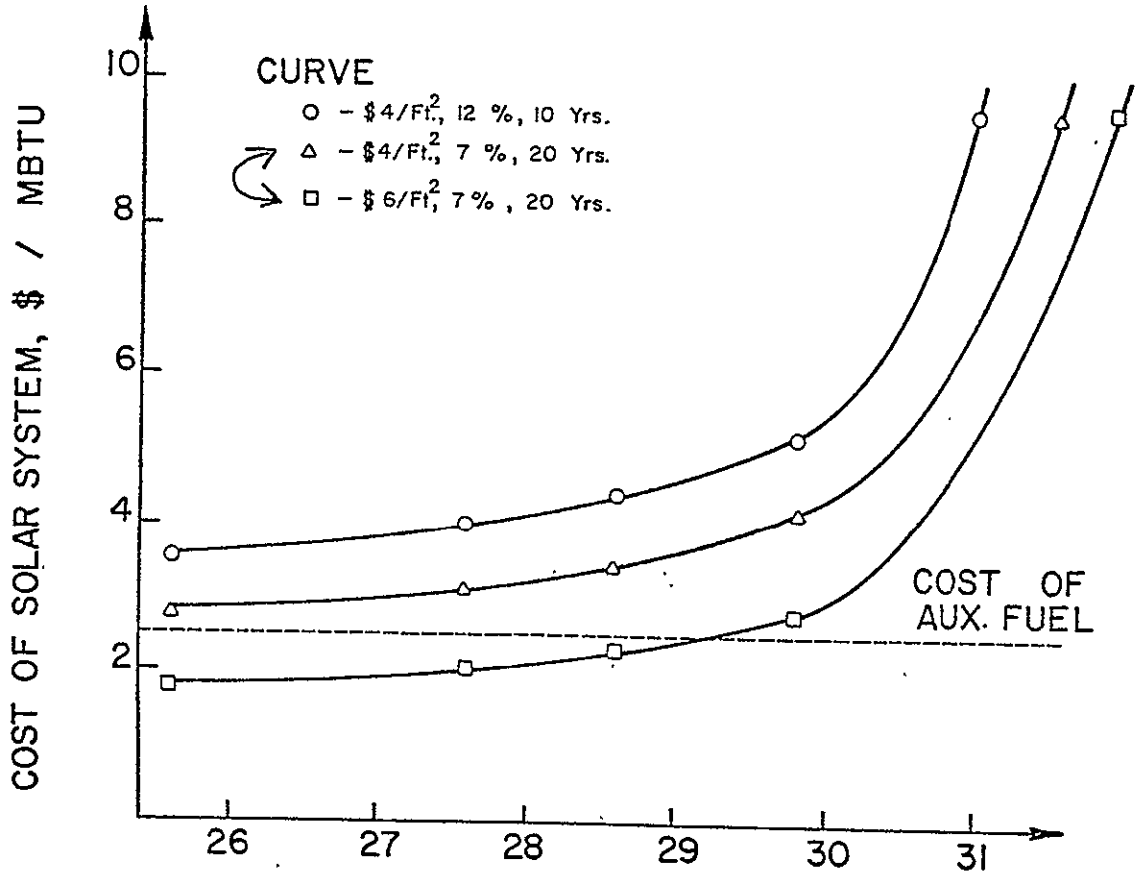


FIGURE 7: $\frac{\text{SOLAR HEAT USED}}{\text{TOTAL HEAT FOR PROCESS}}$, %

p. 316

22 acres for the range of economically justified systems. That is a considerable difference when you think of the space as solar collectors. In addition the present location of a plant may only allow, say for example, 14 acres of collectors. Since this area is justified by economics and will still handle 25% of the total load, 14 acres would be the size system recommended.

The position of the economic curves in Figure 7 with respect to the cost of the fuel line is affected by both the interest rate and the number of years for payback. Curves I and II show this effect. The higher the interest rate and the shorter the payback period the less economical the system. For example, at 12% interest and ten years payoff period the $\$4.00/ft^2$ collector system is not justified, whereas at 7% and 20 years payback it is.

The results do not include the effects of fuel cost escalation. If present trends continue and there is much evidence in the face of fuel shortages that it will, then the cost of conventional fuels should continue to escalate at up to 10 to 15% per year and perhaps even higher. If this is the case, the results are conservative. Therefore, general curves were developed which show the economically justifiable initial investment versus the life of the solar system for various inflation rates. The curve, Figure 8, is based on the method given in Reference [2] for heating systems with increasing fuel costs. The equation used is:

$$P = A_o \frac{(1 + i_{eff})^n - 1}{i_{eff}(1 + i_{eff})^n} \qquad (12)$$

where

P = initial investment

A_o = initial annual fuel saving

N = number of years

$$i_{eff} = \frac{1 - i}{1 + j}$$

where

i = annual interest rate

j = rate of price escalation.

These curves may be used in conjunction with the annual fuel savings calculated in the computer program to determine an economically justifiable system based on a rate of fuel cost escalation.

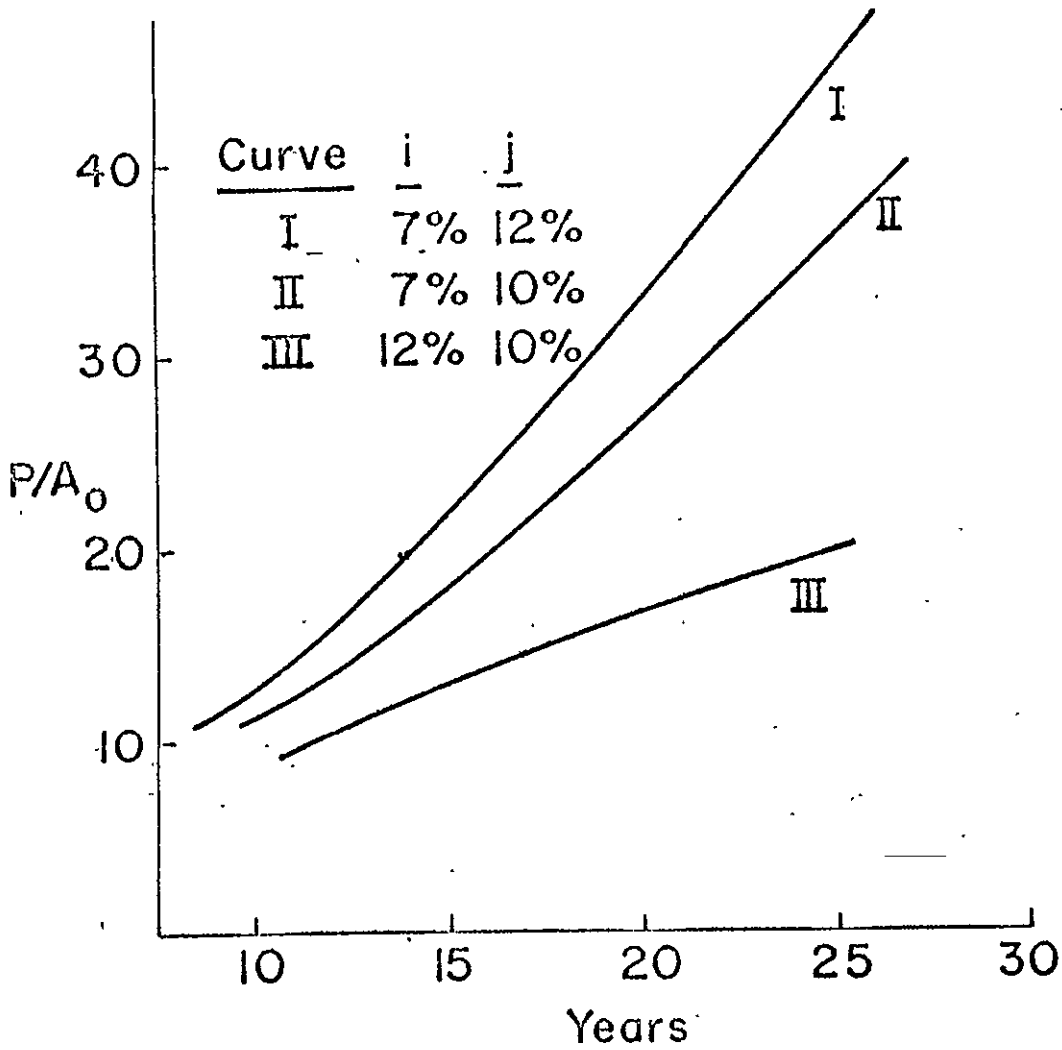


Figure 8

where

A_0 = Initial annual savings from solar system

P = Present justifiable capital expenditure

CONCLUSION

Solar systems can be economically competitive with fossil fuels. However, this report shows some important generalizations in reaching this conclusion. First is that solar systems can only supplement process heat loads. For the case study it was shown that the solar system can supplement 25% of the total load. This is significant when considered as fuel savings. It was also found that 25% to 30% fuel savings could be realized in general for any large heat load. This was assuming that unlimited space is available for the collector area required.

Next it was determined that the main advantage of a storage system was for system stabilization over a given day. The storage tank could provide limited process operation for periods when the sun is hidden behind the clouds, but not for extended periods. A half-hour storage for the case study would require a 70,000 gallon storage tank whereas a two hour storage would require a 200,000 gallon storage tank. The system designer should consider available solar data for the plant location in selecting a size. For a generally cloudy location a larger tank should be selected.

Finally, in the area of economics it was shown that collector costs will determine whether or not the system is justified. Collectors must be available at $\$6.00/ft^2$ or less. All other costs are only a fraction of the total.

As noted, these conclusions are based on a case study with many fixed systems parameters, such as maximum and minimum process system and storage tank operating temperatures. When actually applying these generalizations the designer should keep this in mind. The computer program developed will allow time varying input data for any indirect process solar heating system to aid the designer in final economic verification.

REFERENCES

[1] Duffie, J. A. and Beckman, W. A., "Solar Thermal Processes," Interscience Publishers, New York, 1974.

[2] Kreider, J. F. and Kreith, F., "Solar Heating and Cooling," Hemisphere Publishing Corporation, Washington, D. C., 1975.

[3] Thuesen, H. B., Fabrycky, N. J. and Thuesen, G. J., "Engineering Economy," Prentice Hall, Inc., New Jersey, 1971.

Project

(A) Project Title: Dynamic Behavior of Vapor-Liquid Interphase Mass Transfer Trays

(B) Project Abstract:

This investigation is concerned with a fundamental analysis of the dynamic behavior of vapor-liquid interphase mass transfer trays. A theoretical model is first developed to describe the transient behavior of a single mass transfer tray in response to either a single or multiple disturbance in composition of the entering vapor or liquid stream. The proposed model takes into account the fact of incomplete liquid mixing on the tray by using the eddy diffusion mechanism to describe such liquid mixing behavior. The dynamic interphase mass transfer of a bubbling tray is characterized by two dimensionless parameters, the Peclet number (P_L) and the mass transfer parameter k. The influence of the parameters are determined and discussed.

The concept of partial position transfer function is proposed to provide a concise and general tool for relating the dynamic response of a system of multiple disturbances to those of the same system subjected to single and independent disturbances.

The concept provides the means of breaking the complicated response system into a series of the partial responses which can be obtained from the system with less difficulty.

The concept was applied successfully in obtaining the dynamic response of a tray which is subjected to simultaneous composition disturbances in either pulse or a step or a frequency form.

The dynamic mathematical model for an interphase mass transfer plate column was developed using the fundamental knowledge gained from the study of a single tray. A preliminary attempt has been made and successfully carried out in obtaining the solubion of the model by a numerical method. The solution clearly indicated the importance of the effect of liquid mixing on the dynamic response of an interphase mass transfer column. When the liquid on the trays is not completely mixed, there exists a significant delay between the disturbance and the response.

One of the almost unchallenged assumptions in chemical engineering research and design is that vapor is completely mixed before it moves upward to the tray above in a bubbling plate column. However, this theoretical study has shown that the assumption is not always true unless the interphase mass transfer is fairly rapid.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1968

(E) Department: Chemical Engineering

(F) Student Name: Tu-ching Cheng

(G) Faculty Advisor: Professor C. J. Huang

Project

(A) Project Title: A Dynamic Model of the Circle of Willis

(B) Project Abstract:

A basic pulsatile-flow model of the circle of Willis at the base of the brain was built with integrated lumped circuits. The nonsteady Navier-Stokes equation and one other equation which governs the radial movement of the distensible walls were applied to each of the 22 consecutive short segments into which the circle of Willis was subdivided. The segments were so arranged that the regions where circulation would be influenced by local increase of resistance or altered field of blood supply were located at junctions of the segments. Direct measurement of intravascular blood pressure from a dog's cartoid artery was taken to test the model performance. The computer model compared favorably with the data measured from the animal prototype.

(C) Publication: M.S. Thesis in Civil Engineering and ASME Biomechanical and Human Factors Conference, (1970)

(D) Year: 1970

(E) Department: Civil Engineering

(F) Student Name: J. C. Chao

(G) Faculty Advisor: Professor N. H. C. Hwang

an ASME publication

The Society shall not be responsible for statements or opinions advanced in papers or in discussion at meetings of the Society or of its Divisions or Sections, or printed in its publications.

Discussion is printed only if the paper is published in an ASME journal or Proceedings.

Released for general publication upon presentation

$2.00 PER COPY
$1.00 TO ASME MEMBERS

ORIGINAL PAGE IS
OF POOR QUALITY

A Dynamic Model of the Circle of Willis

J. C. CHAO
Graduate Assistant
N. H. C. HWANG
Associate Professor
Cullen College of Engineering,
University of Houston,
Houston, Texas

A basic pulsatile-flow model of the circle of Willis at the base of the brain was built with integrated lumped circuits. The nonsteady Navier-Stokes equation and one other equation which governs the radial movement of the distensible walls were applied to each of the 22 consecutive short segments into which the circle of Willis was subdivided. The segments were so arranged that the regions where circulation would be influenced by local increase of resistance or altered field of blood supply were located at junctions of the segments. Direct measurement of intravascular blood pressure from a dog's carotid artery was taken to test the model performance. The computer model compared favorably with the data measured from the animal prototype.

Contributed by the Biomechanical & Human Factors Division of The American Society of Mechanical Engineers for presentation at the ASME Biomechanical & Human Factors Conference, Washington, D. C., May 31 - June 3, 1970. Manuscript received at ASME Headquarters, March 16, 1970.
Copies will be available until March 1, 1971.

p. 323

THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS, UNITED ENGINEERING CENTER, 345 EAST 47th STREET, NEW YORK, N.Y. 10017

A Dynamic Model of the Circle of Willis

J. C. CHAO N. H. C. HWANG

INTRODUCTION

Aneurysm on cerebral arteries is a matter of acute concern to neurologists and neurosurgeons alike. Under most circumstances, the aneurysm(s) is (are) found on or near the vicinity of the circle of Willis. Treatment of the cerebral aneurysm has been mostly surgical, either by direct approach to the aneurysm through craniotomy or by elective occlusion of one or more of the afferent cervical arteries.

In the hands of most surgeons, the surgical procedure of cervical artery occlusion carries less risk than the direct intracranial approach to the aneurysm. However, because of the uncertainty of post-operative flow pattern and the pressure distribution in the brain arterial system, it has frequently led to recurrent hemorrhages from the aneurysm(s) for which occlusion was performed and other post-operative bleeding episode. The comparative values of the cervical arterial occlusion treatments has remained a controversial subject.

The basic purpose of cervical artery occlusion is to reduce the local pressure in the weakened portion of the arterial wall in which the aneurysm resides. Therefore, it is extremely important to be able to predict preoperatively the post-occlusive pressure distribution in every part of the brain artery system.

Past attempts to model the circle of Willis have dealt entirely with the study of the much simplified steady flow in rigid tubes using a fluid model (1-4),[1] an electric analog model (5-7) and a computer model (7). Most of these models were comparatively studied in the form of their computer representation recently by Clark, et al. (8). This paper will present a basic model of the circle of Willis with pulsatile flow in distensible tubes.

CONSTRUCTION OF THE MODEL

The measurement of major cerebral arteries of a group of dogs as quoted in reference (4) have been taken as the prototype for our studies. The prototype circle was subdivided into 22 segments of approximately equal volumes. Some irregularities of the sectioning are introduced from the need to have the regions where circulation would be influenced by local increase of resistance or

[1] Underlined numbers in parentheses designate References at end of paper.

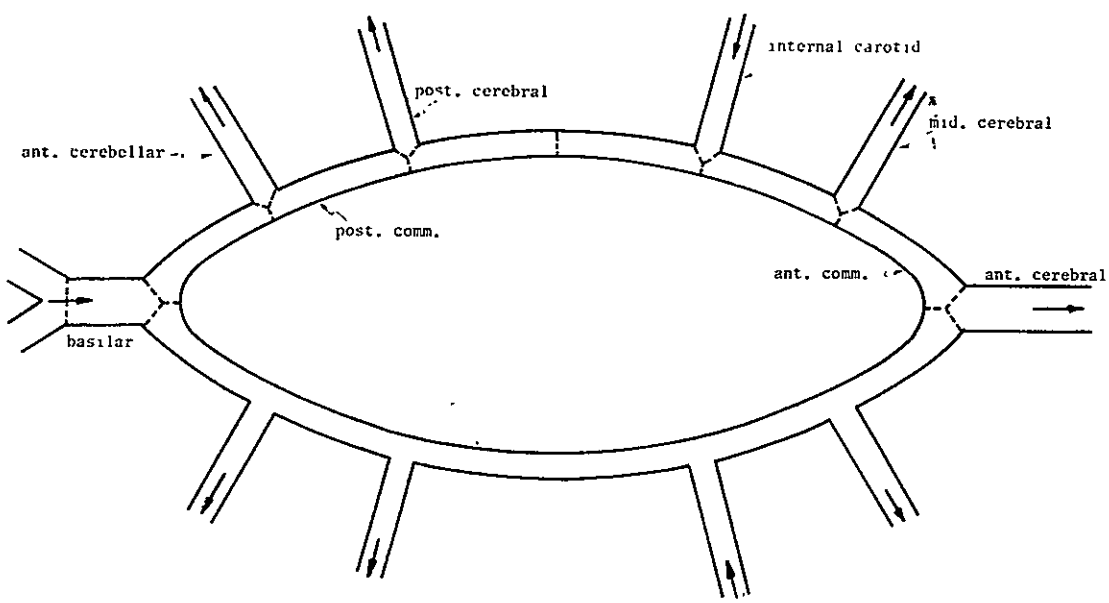


Fig.1 Sketch of the mathematical model

altered field of blood supply coincide with junctions of segments. The mathematical model is schematically shown in Fig.1.

Within each of the 22 segments, blood flow is assumed to be approximately laminar and incompressible. Disregarding body forces and the small tangential motion of the blood, the simplified Navier-Stokes equations in the cylindrical coordinate form can be written as:

$$\frac{\partial p}{\partial z} = -\rho\frac{\partial u}{\partial t} + \mu\left[\frac{1}{r}\frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial r^2}\right] \quad (1)$$

$$\frac{\partial p}{\partial r} = -\rho\frac{\partial w}{\partial t} + \mu\left[\frac{1}{r}\frac{\partial w}{\partial r} + \frac{\partial^2 w}{\partial r^2} - \frac{w}{r^2}\right] \quad (2)$$

Here p is pressure, z is the distance along the axis, ρ is fluid mass density, μ is the ordinary coefficient of viscosity, and u and w are longitudinal and radial velocity components, respectively.

If we further assume that the pressure is independent of r, a reasonable assumption for small radial velocity, equation (2) can be neglected initially. The value of w may then be determined from wall elasticity using the continuity relation.

Equation (1) may be converted from differential to difference form in the radial space dimensions by breaking up the segment radially into N concentric annular shells of equal thickness of Δr for all but the outmout annulus, which was chosen to have thickness $\Delta r/2$. Applying the simplest case, N = 2, Rideout and Dick (9) have converted equation (1) to the following form:

$$p_o - p_i = -\frac{9\rho\Delta z}{4\pi R^2}\frac{dq_i}{dt} - \frac{81\mu\Delta z}{8\pi R^4}q_i \quad (3)$$

Here Δz is the length of the segment, R is the tube radius, and q is the longitudinal flow rate. The subscripts i and o indicate the input and output at the entrance and the exit of the segment, respectively.

The coefficient of the second term on the right-hand side of equation (3) gives the fluid resistance of the segment,

$$R = \frac{81\mu\Delta z}{8\pi R^4}, \quad (4)$$

while the coefficient of the first term on the right-hand side of equation (3) gives the fluid inductance,

$$L = \frac{9\rho\Delta z}{4\pi R^2} \quad (5)$$

At the wall a radial velocity (w_R) is assumed. Here Womersley (10) and Jager, et al., (11) have shown

$$p = \rho h\frac{dw_R}{dt} + \beta w_R + \frac{Eh}{R^2(1-\sigma^2)}\int_0^t w_R dt \quad (6)$$

in which ρ is the wall density, β is the wall damping coefficient, E is Young's modulus of the wall, σ is the Poisson ratio of the wall, and h is wall thickness.

As discussed in reference (9), the first two terms in equation (6) can be ignored because radial velocities are small. Equation (6) then becomes

$$p = \frac{Eh}{R^2(1-\sigma^2)}\int_0^t w_R dt \quad (7)$$

Let the input and output flow at each end of the Δz length of the nth annulus be designated as q_{ni} and q_{no}, then the continuity relates such flows and the radial velocity at r = R as

$$\sum_{n=1}^{N-1}(q_{ni} - q_{no}) = 2\pi R w_R \Delta z \quad (8)$$

From equations (7) and (8),

$$p = \frac{Eh}{2\pi R^3(1-\sigma^2)\Delta z}\int_0^t \sum_{n=1}^{N-1}(q_{ni} - q_{no})dt \quad (9)$$

For the case of N=2, we have,

$$p = \frac{Eh}{2\pi R^3(1-\sigma^2)\Delta z}\int_0^t (q_{ni} - q_{no})dt \quad (10)$$

The fluid capacitance for the segment of length (Δz) is, therefore,

$$C = \frac{2\pi R^3(1-\sigma^2)\Delta z}{Eh} = \frac{3\pi R^3\Delta z}{2Eh}. \quad (11)$$

take $\sigma = 1/2$.

Measuring the dimensions in CGS units and converting the blood pressure from the conventional unit of mm Hg to dyne/sq cm by 1 mm Hg = 1332 dyne/sq cm, the values of R, L, and C are calculated for each of the 22 segments. The values are shown in Table 1.

As shown in Fig.1, all the efferent branches are each represented by two segments in series. The distal part is accounted for 80 percent of the

ORIGINAL PAGE IS
OF POOR QUALITY

p.325

Table 1 Characteristics of Arterial Segments

Segment	Radius (cm)	Length (cm)	R $(\frac{10^3 \text{dyne/cm}}{\text{ml/sec}})$	L $(\frac{10^3 \text{dyne/cm}^2}{\text{ml/sec}})$	C $(10^{-8} \text{ml/dyne/cm})$
Basilar	.05	4.2	68.6	1.130	5.88
Post comm. 1	.035	.50	42.8	.310	.202
Post comm. 2	.035	.25	21.4	.156	.101
Post comm. 3	.035	.36	30.8	.223	.146
Post comm. 4	.035	.36	30.8	.223	.146
Ant. comm. 1	.042	.25	10.4	.107	.175
Ant. comm. 2	.042	.66	27.4	.244	.460
Ant. cerebellar	.032	.73	89.2	.536	.226
Post cerebral	.039	.41	22.7	.202	.228
Int. carotid	.06	4.2	41.6	.880	8.56
Mid. cerebral	.048	.81	19.6	.265	.841
Ant. cerebral	.06	1.12	10.4	.220	2.14

pressure drop. Hence, the lower resistance is four times the upper resistance; the lower compliance is six times the upper compliance.

Applying relations (4), (5), and (11) to a segment (m), which is located between sections (m) and (m+1), equations (3) and (10) can be rewritten as follows:

$$q_m = \frac{1}{L_m} \int_0^t [(p_{m-1} - p_m) - R_m q_m] dt \qquad (12)$$

and

$$p_m = \frac{1}{C_m} \int_0^t (q_m - q_{m+1}) dt \qquad (13)$$

The analog computer setup for the segment is shown in Fig.2.

The entire model including its input function, as mentioned in the next paragraph, has been set on the SS-100 Analog/Hybrid Computer in the Cullen College of Engineering.

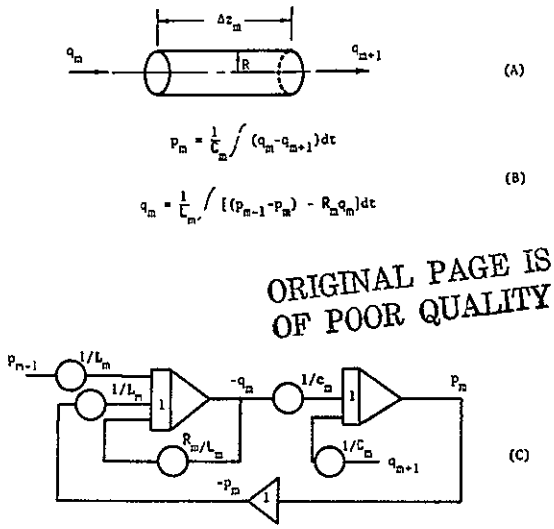
TESTING OF THE MODEL

Intravascular pulse pressure directly measured from a dog's right common carotid artery (12) was taken to test the model performance. At this preliminary phase of the study,[2] several assumptions, which had been commonly made for almost all previous models (1-5,7), were accepted to test our model. These assumptions are as follows: 1) the total flow through the system is estimated at 63 cc/100 gm of dog brain per minute, 2) the afferent flow is divided equally among the two carotids and the basilar, 3) the efferent flow is distributed in accordance with the weight of brain irrigated by each vessel, and 4) the input pressure at all three afferent arteries was the same function of time and there is no phase delay.

Pulse pressure and flow rate at several sections on the model were measured with a Polaroid camera which photographed the traces on the screen of a Textronix dual-beam oscilloscope. The measured pressure and flow rated at each of these sections are shown as functions of time in Figs.3 and 4, respectively.

Comparisons of the model output with the animal data (12,13) were rather encouraging because the relative pressure drop and flow distribution at each of the junctions were comparable

[2] A hybrid computer phase of the model with delayed individual input function for each of the three afferent arteries will be reported in subsequent papers.



ORIGINAL PAGE IS
OF POOR QUALITY

(A) Arterial segment
(B) Integral equations representing (A)
(C) Analog computer setup for (A)

Fig.2

both in magnitude and wave forms. Although a point-to-point comparison was not made in this study due to the fact that different animals were involved, the similarity in trends indicated the possibility of satisfactory refinement of the modeling technique. The comparison of the changes in flow rate and pressure with the previous models was not possible since the flows in all other models were steady in nature. However, the ranges of magnitude showed good agreement between our model and the previous models.

CONCLUSION

The advantages of analog modeling of a complex circulation system, such as the circle of Willis at the base of the brain by integrated lumped circuits, are apparent. Not only does it offer a far greater versatility than a conventional fluid-model, but it also offers the convenience of recording the pressure and flow rates at almost any desired sections in the system; the ease of finding values of functions with varying parameters; and most of all, it permits the possibility of including the nonlinear terms.

The assumption of laminar flow in cerebral arteries may be regarded as a better approximation than is made for larger vessels (such as the aorta) because of the apparently smaller tube diameter and slower longitudinal velocities (14).

Study of brain circulation after various cervical arterial occlusions was not included in this paper. However, it is planned for our next stage of investigation. A hybrid computer extension of the present model is now under construction in our laboratory. In the model, the analog representation described in this paper will be combined with a digital representation of some of the slower acting control loops, such as the peripheral resistance and phase delay of pressure waves in different afferent arteries. Better conformity of model to animal prototype can be expected.

ACKNOWLEDGMENT

The assistance of Drs. Williamina A. Himwich and T. Iwabuchi in furnishing some of the animal data is greatly appreciated.

This work was partially supported by NASA Grant NGL-44-005-084.

REFERENCES

1 Kramer, S. P., "On the Function of the Circle of Willis," Journal of Experimental Medicine, Vol. 15, 1912, pp. 348-363.

2 Rogers, L., "The Function of the Circlus Arteriosus of Willis," Brain, Vol. 70, 1947, pp. 171-178.

3 Avman, N., and Berling, E. A., Jr., "A Plastic Model for the Study of Pressure Changes in the Circle of Willis and Major Cerebral Arteries Following Arterial Occlusion," Journal of Neurosurgery, Vol. 18, 1961, pp. 361-365.

4 Himwich, W. A., et al., "The Circle of Willis as Simulated by an Engineering Model," Arch. Neurol., Vol. 13, 1965, pp. 164-172.

5 Murray, K. D., "Dimensions of Circle of Willis and Dynamic Studies Using Electric Analog," Journal of Neurosurgery, Vol. 21, 1964, pp. 26-34.

6 Fasano, V. A., Portalupi, A., and Broggi, G., "Aspetti Fisici Della Circolazione Nei Vasi Anastomotici Endocranici Basali e Meningo-Cortical: (Studio su un Modello Elettrico Analogico)," Sist. Nerv., Vol. 16, 1964, pp. 271-307.

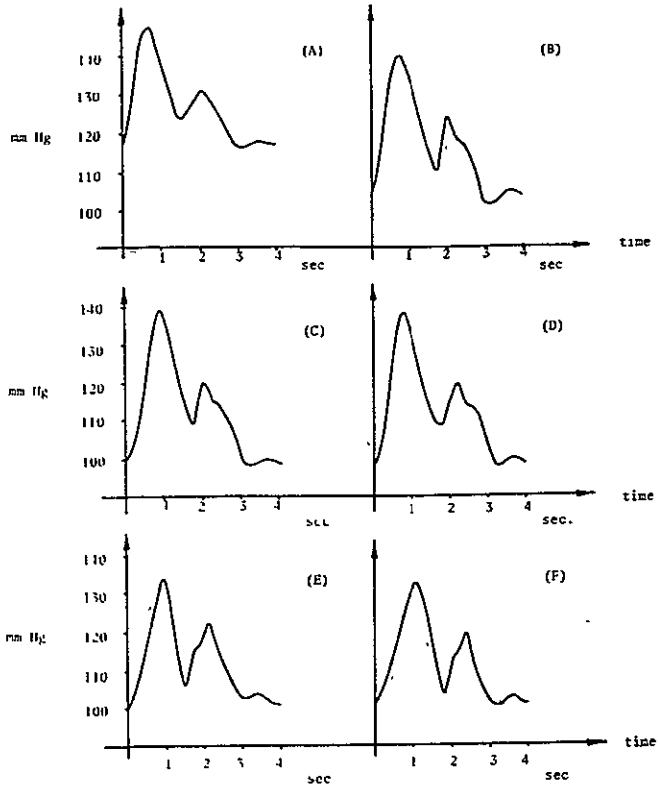
7 Clark, M. E., et al., "Engineering Analysis of the Hemo-dynamics of the Circle of Willis," Arch. Neurol., Vol. 13, 1965, pp. 173-182.

8 Clark, M. E., Himwich, W. A., and Martin, J. D., "A Comparative Examination of Cerebral Circulation Models," Journal of Neurosurgery, Vol. 29, 1968, pp. 484-494.

9 Rideout, V. C., and Dick, D. E., "Difference-differential Equations for Fluid Flow in Distensible Tubes," IEEE Transactions Bio-Med. Engr., BEM-14, 1967, pp. 171-177.

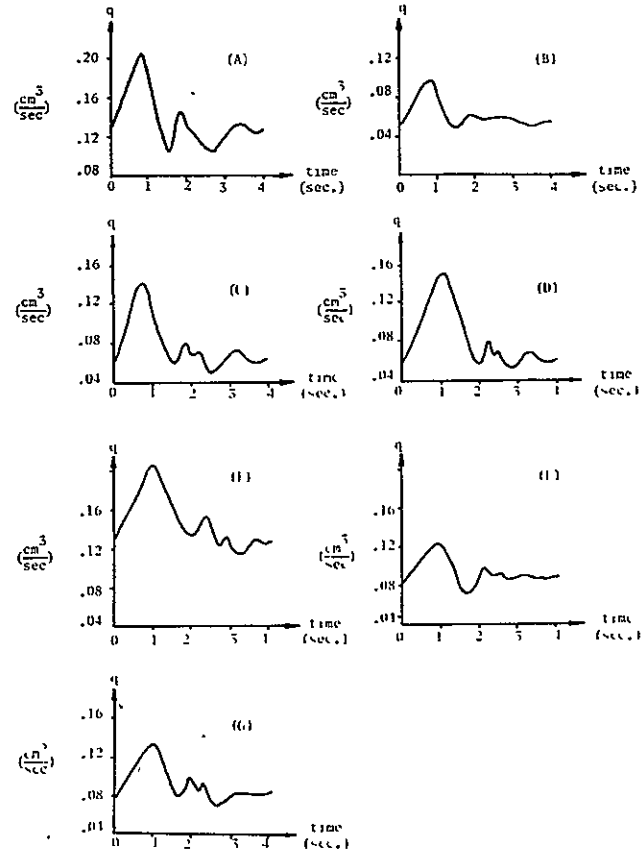
10 Womersley, J. R., "An Elastic Tube Theory for Pulse Transmission and Occillatory Flow in Mammalian Arteries," Wright Air Div. Rept., WADC-TR 56-614, 1957.

1.327



(A) Input pressure function at the junction of vertebral and basilar
(B) Pressure at the junction of post. comm. and basilar
(C) Pressure at the junction of post. comm. and ant. cerebellar
(D) Pressure at the junction of post. comm. and post. cerebral
(E) Pressure at the junction of ant. comm. and mid. cerebral
(F) Pressure at the junction of ant. comm. and ant. cerebral

Fig.3 Pressure distribution at arterial junctions



(A) Flow rate at the junction of basilar and post. comm. (in the direction of post. comm. artery)
(B) Flow rate at the junction of post. comm. and post. cerebral (in the direction of post. cerebral)
(C) Flow rate at the junction of post. comm. and post. cerebral (in the direction of post. comm.)
(D) Flow rate at the mid. point of the post. comm. (between carotid and post. cerebral)
(E) Flow rate at the junction of internal carotid and comm. (in the direction of anterior comm.)
(F) Flow rate at the junction of mid. cerebral and anterior comm. (in the direction of mid. cerebral)
(G) Flow rate at the junction of anterior comm. and anterior cerebral (in the direction of anterior cerebral)

Fig.4 Flow rate distribution in the circle of Willis

11 Jager, G. N., Westerhof, N., and Noordergraaf, A., "Oscillatory Flow Impedance in Electrical Analog of Arterial System," Circ. Res., Vol. 16, 1963, pp. 121-133.

12 Himwich, W. A., and Spurgeon, H. A., "Pulse Pressure Contours in Cerebral Arteries," Acta Neurol. Scandinav., Vol. 44, 1968, pp. 43-56.

13 Personal communication with W. A. Himwich.

14 Hwang, N. H. C., and Chao, J. C., "Turbulent Characteristics of Arterial Flow," Proceedings of the Annual Conference on Engineering in Medicine and Biology, Vol. 10, 1968, 16.5.

p. 328

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: Causes, Magnitude and Effects of Temperature Fluctuations (Flickering) of Catalytic Wires and Gauzes

(B) Project Abstract:

An experimental study has confirmed a theoretical prediction that, when an exothermic mass transfer limited chemical reaction occurs on a single catalytic wire for which the parameter a (Equation 11) is large, temperature fluctuations (flickering) of large amplitude must be induced by concentration fluctuations. A simplified model is presented for predicting the magnitude of flickering in industrial convertors for which the parameter a is usually large. The model should be useful in estimating the influence of improved mising of the reactants on the reduction in precious metal loss from the gauze.

(C) Publication: AIChE Journal, 20, 571 (1974)

(D) Year: Completed in 1973

(E) Department: Chemical Engineering

(F) Student Names: W. M. Edwards and J. E. Zuniga-Chaves

(G) Faculty Advisor: Professors F. L. Worley, Jr. and Dan Luss

Reprinted From . .

AIChE JOURNAL

CAUSES, MAGNITUDE, AND EFFECTS OF TEMPERATURE FLUCTUATIONS (FLICKERING) OF CATALYTIC WIRES AND GAUZES

W. M. Edwards, J. E. Zuniga-Chaves, F. L. Worley, Jr. and Dan Luss

Department of Chemical Engineering, University of Houston
Houston, Texas 77004

VOL. 20, NO. 3 MAY, 1974 PAGES 571-581

6.330

Causes, Magnitude, and Effects of Temperature Fluctuations (Flickering) of Catalytic Wires and Gauzes

An experimental study has confirmed a theoretical prediction that, when an exothermic mass transfer limited chemical reaction occurs on a single catalytic wire for which the parameter a [Equation (11)] is large, temperature fluctuations (flickering) of large amplitude must be induced by concentration fluctuations. A simplified model is presented for predicting the magnitude of flickering in industrial convertors for which the parameter a is usually large. The model should be useful in estimating the influence of improved mixing of the reactants on the reduction in precious metal loss from the gauze.

W. M. EDWARDS
J. E. ZUNIGA-CHAVES
F. L. WORLEY, JR.
and DAN LUSS

Department of Chemical Engineering
University of Houston
Houston, Texas 77004

SCOPE

The industrial ammonia oxidation process utilizes a catalytic gauze consisting of 10 to 40 layers of platinum-rhodium wire screens. Observations of these gauzes reveal localized regions of high luminosity whose intensity fluctuates randomly with time, a phenomenon referred to as *flickering*. Nowak (1969) reported that the rate of precious metal loss in the ammonia oxidation process, which accounts for about 5% of the manufacturing costs, can be represented by an Arrhenius temperature dependence with an activation energy of about 40,000 cal/g mole. Due to this convex temperature dependence, temperature oscillations (flickering) increase the metal loss above the level corresponding to uniform temperature operation at the same average temperature. Thus, information about the magnitude and causes of flickering could lead to a design which reduces these precious metal losses.

Ervin and Luss (1972) used numerical simulation to show that large amplitude flickering can be induced by the fluctuations of the turbulent transport coefficients (caused by velocity fluctuations) if the parameter a, which is the ratio of the characteristic time for changes in wire diameter to the characteristic time for surface concentration changes, was of order unity or less. However, under the conditions prevailing in industrial convertors the parameter a is very large, and for this case numerical simulations indicate that the magnitude of flickering induced by velocity fluctuations is negligible. A model presented here shows that if the parameter a is large and the reactants are not completely mixed on a molecular level, concentration fluctuations can induce flickering of large amplitude. Edwards et al. (1973) measured large amplitude flickering during the catalytic oxidation of butane in air on single platinum wires for which a was rather large. The experiments demonstrated the existence of a correlation between flickering and the turbulence of the reacting gas but did not enable a determination of the specific cause for flickering.

The main objective of this work was to determine the primary cause of flickering of single catalytic wires for which the parameter a is much larger than unity. To this end a high resolution infrared detector was used to measure localized surface temperature fluctuations on single platinum wires during the catalytic oxidation of either hydrogen or ammonia in air. By changing the feed port location and method of injection various intensities of concentration fluctuations were obtained in the same turbulent flow field. The results of this study were used to develop a theoretical model for predicting the causes and magnitude of flickering under the conditions prevailing in industrial convertors.

ORIGINAL PAGE IS
OF POOR QUALITY

CONCLUSIONS AND SIGNIFICANCE

A series of measurements of the root mean square temperature fluctuations in the same turbulent flow field and various levels of concentration fluctuations indicated that concentration fluctuations must be the primary cause for flickering of the magnitude observed in industrial convertors. The theoretical predictions of Equation (23) adequately described the relation between local flickering on a single wire and local concentration fluctuation intensity. A comparison of the probability density functions of the fluctuations about the mean of the wire's temperature with those of the concentration and velocity of the reacting gas further confirm the conclusion that flickering was induced by concentration and not velocity fluctuations. The spectral density functions of the temperature fluctuations do not enable a conclusive determination of the cause of flickering.

A simplified model was developed for estimating the magnitude of flickering in industrial gauze convertors [Equation (41)]. The model predicts that under the conditions prevailing in typical high pressure ammonia convertors imperfect mixing and concentration fluctuations are the main causes for flickering. Thus, equipment modifications which reduce the concentration fluctuations upstream of the gauze will decrease the magnitude of flickering. This in turn will reduce the precious metal loss and the deterioration rate of the catalytic gauze.

Correspondence concerning this paper should be addressed to D. Luss. W. M. Edwards is with the M. W. Kellogg Company, Houston, Texas.

Catalytic gauze convertors deteriorate with time due to precious metal losses (mainly in the form of volatile oxides). The precious metal loss varies from plant to plant and often has a significant impact on the economics of the process. For example, in the high pressure ammonia oxidation process, the metal losses may account for about 5% of the manufacturing costs (Newman and Hulbert, 1971, Gillespie and Kenson, 1971; Heywood, 1973) and, according to Nowak (1969), can be correlated by the empirical expression

$$-\frac{dM}{dt} = f(O_2)\exp(-E_m/RT) \tag{1}$$

where $E_m \simeq 40{,}000$ cal/g mole.

Large amplitude random temperature fluctuations (flickering) have been observed in commercial convertors. One of the main disadvantages of flickering is its deleterious influence on the rate of precious metal losses from catalytic gauzes. An instructive parameter for gauging the influence of flickering is the relative metal loss $r(T)$, defined as the ratio of the metal loss with flickering to that during uniform temperature operation with the same average temperature and average reactant concentration.

$$r(T) = \frac{\left\langle \frac{dM}{dt} \right\rangle}{\left(\frac{dM}{dt}\right)_{T=\langle T\rangle}} = \frac{\left\langle \exp\left(-\frac{E_m}{RT}\right)\right\rangle}{\exp\left(-\frac{E_m}{R\langle T\rangle}\right)} \tag{2}$$

where $\langle\ \rangle$ denotes the time average of a stationary process. Since $\exp(-E/RT)$ is a convex function of the temperature $r(T) \geq 1$. When the probability density function (pdf) of the temperature fluctuations is known, (2) can be used to calculate the effect of flickering on the metal loss. Experiments carried out in this laboratory with single wires indicate that when no information is available about the pdf of the flickering a Gaussian distribution should be good approximation. Typical values of $r(T)$ are reported in Table 1 using this assumption and operating conditions similar to those existing in industrial ammonia oxidation convertors. It is noted that the relative metal loss increases rapidly with increasing root mean square temperature fluctuations.

The temperature and surface concentration of a single catalytic wire on which a single chemical reaction occurs can be described by the equations

$$\frac{d(AS)}{dt} = k_1(S)\cdot C - r \tag{3}$$

$$A\rho c_p \frac{\partial T}{\partial t} = k_w A \frac{\partial^2 T}{\partial s^2} + hP(T_g - T) + P(-\Delta H)r \tag{4}$$

$$(S) + (AS) = L \tag{5}$$

Edwards et al. (1973) have shown that when the reaction is mass transfer controlled the maximal temperature rise is given by

$$\langle \Delta T_{ad}\rangle = \langle \Delta T_{ad}^{\circ}\rangle + \frac{Ak_w}{P\langle h\rangle}\frac{\partial^2 T}{\partial s^2} \tag{6}$$

where

$$\langle \Delta T_{ad}^{\circ}\rangle = \frac{(-\Delta H)\langle C\rangle}{\rho_f c_{p_f}}\left(\frac{N_{Pr}}{N_{Sc}}\right)^{2/3} = \frac{(-\Delta H)\langle x_g\rangle}{c_{p_f} M_f} N_{Le}^{-2/3} \tag{7}$$

The second term on the right-hand side of (6), which accounts for the influence of thermal conduction, is negligible when compared with $\langle \Delta T_{ad}^{\circ}\rangle$ for wires with length to diameter ratio greater than 200.

TABLE 1. VALUES OF $r(T)$ AS A FUNCTION OF THE TIME AVERAGE TEMPERATURE AND THE ROOT MEAN SQUARE TEMPERATURE FLUCTUATIONS ASSUMING A GAUSSIAN TEMPERATURE OSCILLATION pdf AND $E_m = 40{,}000$ CAL/G.MOLE

	$\langle T\rangle$, °C			
T'_{rms}, °C	800	850	900	950
20	1.05	1.04	1.033	1.028
40	1.22	1.18	1.15	1.12
60	1.54	1.43	1.35	1.29

The turbulent mass transfer coefficient to a cylinder can be expressed as (Treybal, p. 59, 1968)

$$N_{Sh} = \frac{k_c\, d}{D_{AB}} = \alpha + \beta u^n \tag{8}$$

where α and β are two empirical constants. Local turbulent transport coefficients fluctuate with time and can be expressed as the sum of their stationary average and a fluctuating component, which is denoted by a prime. When the fluctuating velocity component u' is much smaller than the time averaged velocity $\langle u\rangle$, the ratio between the instantaneous to the time averaged mass transfer coefficient may be approximated by

$$\frac{k_c}{\langle k_c\rangle} = 1 + N\frac{u'}{\langle u\rangle} \tag{9}$$

where

$$N = \frac{n\beta\langle u\rangle^n}{\langle N_{Sh}\rangle} = n\left(1 - \frac{\alpha}{\langle N_{Sh}\rangle}\right) \tag{10}$$

and n is an empirical constant, which is about 0.5. Analogy between heat and mass transfer implies that when the Lewis number is close to unity the right-hand side of (9) describes also the ratio of $h/\langle h\rangle$.

Ervin and Luss (1972) used a numerical simulation to demonstrate that fluctuations of the transport coefficients may induce flickering on a wire. The magnitude of this induced temperature fluctuation was found to be strongly dependent on the parameter

$$a = \frac{A\rho\, c_p\langle k_c\rangle\langle C\rangle}{LP\langle h\rangle} = \frac{A\rho\, c_p\langle \Delta T_{ad}^{\circ}\rangle}{LP(-\Delta H)} = \frac{A\rho\, c_p\langle x_g\rangle N_{Le}^{-2/3}}{LP\, c_{p_f}\langle M_f\rangle} \tag{11}$$

which is the ratio of the characteristic time for changes in wire temperature to the characteristic time for surface concentration changes. The numerical computations indicated that appreciable temperature oscillations were induced by fluctuating transport coefficients only when a was of order one or less.

Estimation of the parameter a requires information about the concentration of active surface sites per unit surface area L. The atomic radius of a platinum atom is 1.39A. Hence, $L \leq 2 \times 10^{-9}$ g atom/cm². A high temperature steady state can exist only if the reactant mole fraction exceeds the value corresponding to extinction x_e. Some estimates of a lower bound on the parameter a are reported in Table 2 for reactions on a 0.025-mm platinum wire, based on the assumptions of $L = 2 \times 10^{-9}$ and $\langle x\rangle = x_e$. These estimates indicate that for all these cases a is at least of order 100. In the industrial ammonia oxidation process $\langle x_g\rangle \simeq 0.11$ and $d_w = 0.075$ mm, yielding 9900 as the lower bound on a. A similar

1.332

TABLE 2. ESTIMATION OF A LOWER BOUND ON THE PARAMETER a FROM EXPERIMENTS WITH A 0.025-mm PLATINUM WIRE

Reactant	N_{Le}	$\Delta T^{\circ}_{ad}/x_g$ [K°]	T_g [°C]	x_c	Min (a)
Butane	2.30	47,500	145	0.0087	140
Hydrogen	0.27	19,600	24	0.010	710
Ammonia	0.895	8,000	24	0.040	1,240

magnitude is attained in the HCN process. The model of Ervin and Luss (1972) predicts that for these high values of a the amplitude of the flickering is negligibly small, and that fluctuations in the transport coefficients cannot induce the large amplitude flickering observed in industrial convertors.

In the following discussion, a model is described which explains flickering for systems having high values of the parameter a. The main assumption of the model is that for mass transfer controlling conditions and large values of a the surface concentration of the adsorbed reactant satisfies the pseudo steady state relation

$$r = k_1(S)C = k_cC \tag{12}$$

Substitution of (12) into (4) and neglecting the effect of axial conduction yields

$$\frac{A\rho\, c_p}{P}\,\frac{dT}{dt} = h(T_g - T) + k_cC(-\Delta H) \tag{13}$$

Equation (13) can be rewritten as

$$\frac{A\rho\, c_p}{P\langle h\rangle\langle T - T_g\rangle}\,\frac{dT'}{dt} = -\left(1 + \frac{h'}{\langle h\rangle}\right)\left(1 + \frac{T'}{\langle T - T_g\rangle} - \frac{T_g'}{\langle T - T_g\rangle}\right) + \left(1 + \frac{k_c'}{\langle k_c\rangle}\right)\left(1 + \frac{C'}{\langle C\rangle}\right) \tag{14}$$

where we have expressed the transport coefficients, the concentrations, and the temperature as a sum of a stationary average and a fluctuating component, and used the relations

$$\frac{d\langle T\rangle}{dt} = 0 \tag{15}$$

ORIGINAL PAGE IS OF POOR QUALITY

$$\frac{(-\Delta H)\langle k_c\rangle\langle C\rangle}{\langle h\rangle\langle T - T_g\rangle} = 1 \tag{16}$$

Substitution of (9) into (14) and neglecting second-order terms yields

$$\frac{A\rho\, c_p}{P\langle h\rangle}\,\frac{dT'}{dt} + T' = T_g' + \frac{\langle T - T_g\rangle C'}{\langle C\rangle} \tag{17}$$

Assuming that the gas temperature is uniform ($T_g' = 0$) Fourier transformation of (17) yields

$$G_{T'}(f) = \frac{\langle T - T_g\rangle^2\, G_{C'}(f)}{\langle C\rangle^2\,[1 + (2\pi f\tau)^2]} \tag{18}$$

where

$$\tau = \frac{A\rho\, c_p}{P\langle h\rangle} \tag{19}$$

and $G_{T'}(f)$ and $G_{C'}(f)$ are the power spectral density functions of the wire temperature fluctuations and of the feed concentration fluctuations, respectively.

The one-dimensional spectral density function of the velocity fluctuations in an isotropic turbulent flow field can be usually approximated very well for low frequencies by (Hinze, 1959)

$$\frac{G_{u'}(f)}{4\langle u'^2\rangle\tau_E} = \frac{1}{1 + (2\pi f\,\tau_E)^2} = \frac{1}{1 + (f/f_{1/2})^2} \tag{20}$$

where τ_E is the Eulerian time scale and $f_{1/2}$ is the half power frequency. The power spectral density of concentration fluctuations in a turbulent field should be rather similar to that of velocity fluctuations and can be approximated by

$$\frac{G_{C'}(f)}{4\langle C'^2\rangle\tau_C} = \frac{1}{1 + (2\pi f\tau_C)^2} \tag{21}$$

where τ_C is the Eulerian concentration time scale. Experimental measurements of the spectral density function of scalar fluctuations in a turbulent flow field have been reported by Becker et al. (1966), Lee and Brodkey (1964), and Freymuth and Uberoi (1971). Substitution of (21) into (18), integration over all possible frequencies and application of the definition of power spectral density function

$$\int_0^\infty G_{T'}(f)\,df = \langle T'^2\rangle, \tag{22}$$

yields

$$\frac{\langle T'^2\rangle}{\langle T - T_g\rangle^2} = \frac{\tau_C\langle C'^2\rangle}{(\tau_C + \tau)\langle C\rangle^2} \tag{23}$$

Substitution of (21) and (23) in (18) yields

$$\frac{G_{T'}(f)}{4\langle T'^2\rangle} = \frac{(\tau_C + \tau)}{[1 + (2\pi f\tau)^2]\,[1 + (2\pi f\tau_C)^2]} \tag{24}$$

Edwards et al. (1973) have shown that if flickering is induced by fluctuations of the transport coefficients then the experimental data should satisfy the relation

$$\frac{G_{T'}(f)}{4\langle T'^2\rangle} = \frac{(\tau_w + \tau_E)}{[1 + (2\pi f\tau_E)^2]\,[1 + (2\pi f\tau_w)^2]} \tag{25}$$

where τ_w is computed from

$$\tan^{-1}(2\pi f_{1/2}\,\tau_w) - \frac{\tau_E}{\tau_w}\tan^{-1}(2\pi f_{1/2}\,\tau_E) - \frac{\pi}{4}\left(1 - \frac{\tau_E}{\tau_w}\right) = 0 \tag{26}$$

Hinze (1959, p. 231) has shown that for isotropic turbulent mixing the ratio τ_E/τ_C is equal to $\sqrt{N_{Sc}}$. For a mixture of a reactant in excess air N_{Sc} is close to unity so the values of τ_E and τ_C should be about the same. In addition, computations indicate that for our experimental conditions τ_w and τ are about equal. Hence, a comparison of (24) and (25) indicates that the temperature spectral density function cannot be used to determine whether flickering is induced by velocity fluctuations or by concentration fluctuations. However, measurements of the temperature fluctuations in the same flow field at various levels of concentration fluctuations can be used to test the validity of (23) and to determine the main cause of flickering. In the first part of this work we report an experimental study of temperature fluctuations on single catalytic wires. The results are then applied to predict the causes and magnitude of flickering in catalytic gauze convertors.

EXPERIMENTAL APPARATUS AND PROCEDURE

The experimental data obtained to determine the cause of flickering include measurements of instantaneous values of

local surface temperature and reactant concentration. Instantaneous wire surface temperatures were measured with an infrared detector while concentration fluctuations were measured with an aspirating probe unit manufactured by Thermo Systems, Inc. The temperature measurements were made on single platinum wires placed normal to the flow direction of a turbulent reactant (either ammonia or hydrogen)-air mixture. The test assembly (Figure 1) which includes gas flow meters, mixing nozzles, flow channel, probe holder, and infrared detection unit is the same as that described by Edwards et al. (1973), but includes a redesigned flow channel, wire probe, and mixing devices.

The rectangular cross section (25 × 152 mm) flow channel (Figure 2) was constructed from two (203 × 1650 mm.) 3.2-mm thick aluminum plates separated by 25-mm square aluminum bar stock. The channel was insulated with a 38-mm thick layer of fiberglass. The feed gas, air, or nitrogen, entered the flow channel through a 12.5-mm diameter opening into a short (127 mm long) diverging section. In order to break up the incoming jet and to provide a uniform velocity distribution, the gas was passed through two 30-mesh stainless steel screens (100 mm apart) followed by a section packed with 6.3 mm I.D. 102-mm long stainless steel tubes. The flow inside these tubes was laminar for the experimental conditions used.

The catalytic wire probe (Figure 3) was positioned 700 mm downstream from the exit of the stainless steel tubes (56 channel half heights). The probe opening (12.5 × 50 mm) in the bottom of the channel was sealed with a spring loaded fiber plate to prevent gas leakage.

A 25-mm. diameter sapphire window transmitting radiation in the 1-8μ range was installed in the top plate to permit direct viewing of the platinum wire with the infrared unit. This unit measured the instantaneous local wire surface temperature of a 0.04 mm. diameter spot. The root mean square noise level of the measured temperature fluctuations was 0.7°C. The gas temperature was measured with a mercury in glass thermometer inserted through a port in the side of the channel near the catalytic wire. Details of the method of measurement, calibration, recording, and data processing were described by Edwards et al. (1973) and Edwards (1973).

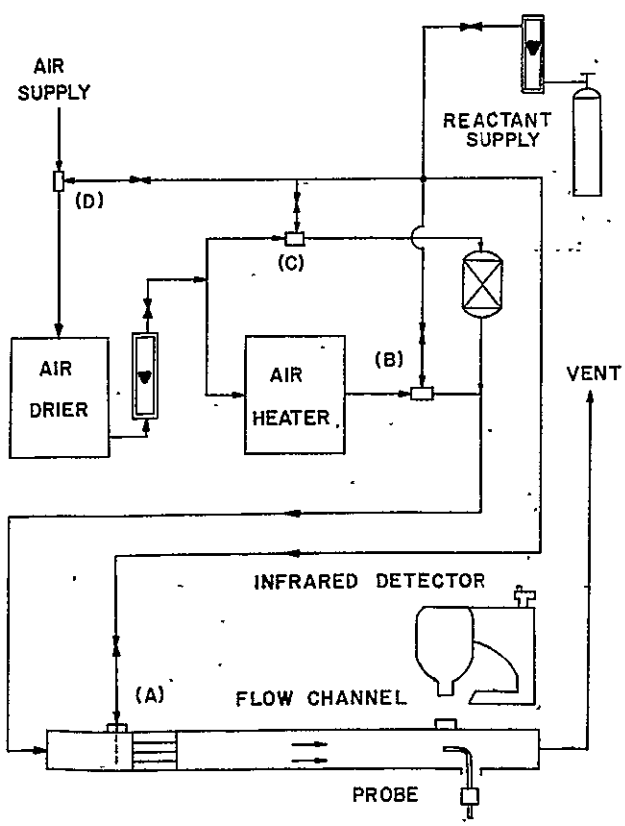


Fig. 1. Schematic of test assembly.

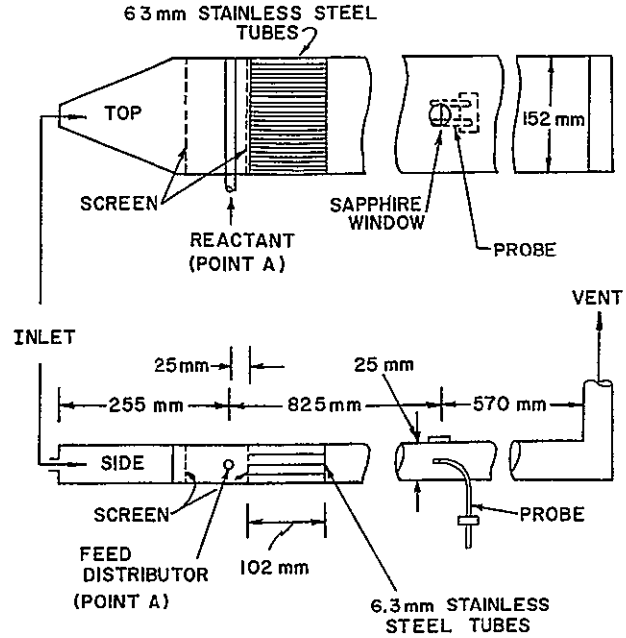


Fig. 2. Schematic of flow channel.

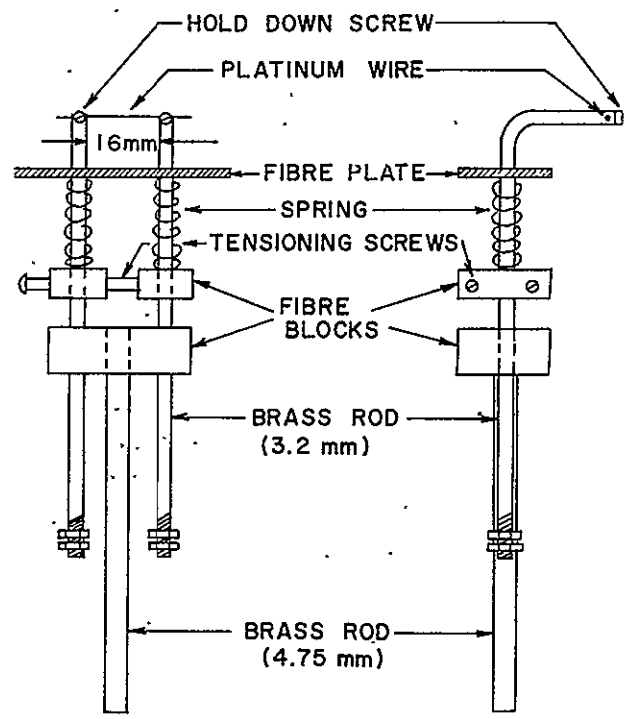


Fig. 3. Catalytic wire probe.

A wide range of reactant concentration fluctuations was obtained by varying the method of reactant injection and/or by changing the injection point (labeled A, B, C, and D in Figure 1). The reactant injector at point A consisted of a 6.3-mm stainless steel tube placed normal to the air flow in the channel between the two 30-mesh screens. The gas was injected through eleven 0.34-mm diameter holes spaced 12.5 mm apart. This distributor could be rotated so that the reactant could be injected at any prescribed angle relative to the bulk flow axis. Feed point B was a tee connection and the reactant was injected through nine 0.24-mm diameter holes in a steel plate. Feed point C was similar to B and contained nine 0.5-mm

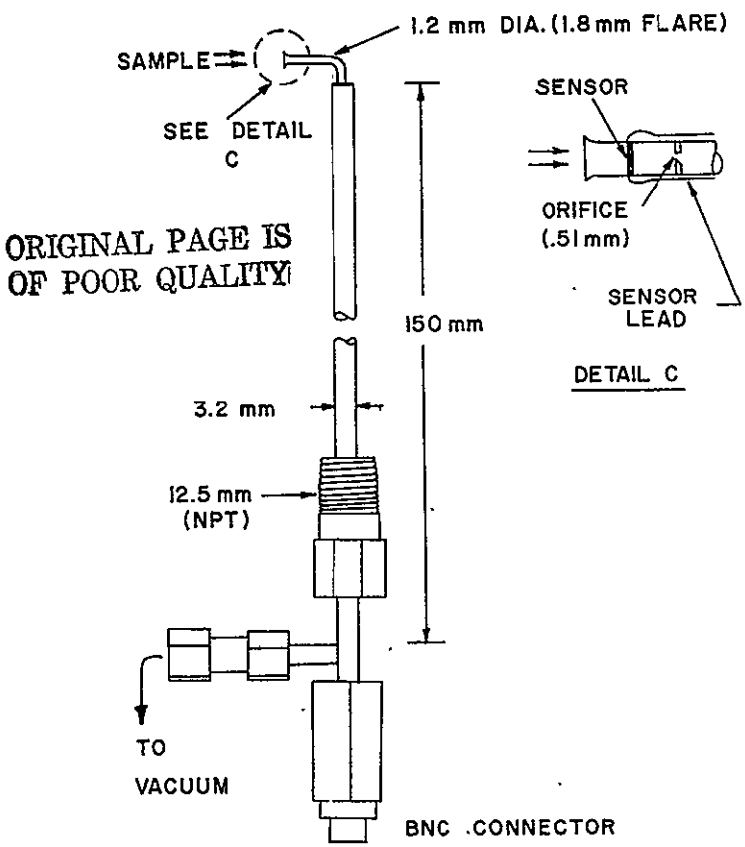
diameter holes in a plastic plate. A 50-mm I.D. packed bed containing 100 cm³ of 0.065-mm diameter glass beads topped by 100 cm³ of 3-mm diameter glass beads was placed between points C and the channel inlet in order to improve the mixing and therefore reduce the magnitude of concentration fluctuations. Feed point D consisted of a tee connection ahead of the 1m. high packed bed air drier shown in Figure 1.

Instantaneous concentration fluctuations for nonreacting turbulent helium-air or hydrogen-nitrogen mixtures were measured with a Thermo Systems Model 1441A aspirating probe (Figure 4) in conjunction with a Thermo Systems Model 1010A constant temperature anemometer system. The design and principles of operation of this probe have been described by Blackshear and Fingerson (1962). Edwards (1973) discussed the conditions under which the measurements of the aspirating probe can be considered valid.

EXPERIMENTAL RESULTS AND DISCUSSION

Edwards et al. (1973) have shown that for butane oxidation the temperature fluctuations behave as a stationary process for times much longer than those required for a single experiment and that conduction due to end effects had a negligible influence on the time averaged temperature and root mean square temperature fluctuations at the center of wires of diameter 0.075 mm or less. In the experiments reported here, it has been assumed that the same is also true for either ammonia or hydrogen oxidation.

Preliminary experiments demonstrated that when the reactant was injected at either one of the feed points labeled as B, C and D in Figure 1, there existed no gradient in the time averaged concentration at the wire station. However, when either hydrogen or ammonia were injected at point A, gradients in the time averaged concentration existed at the wire station. Moreover, it was found that the concentration fluctuation intensity, the



ORIGINAL PAGE IS OF POOR QUALITY

Fig. 4. Schematic of the aspirating probe.

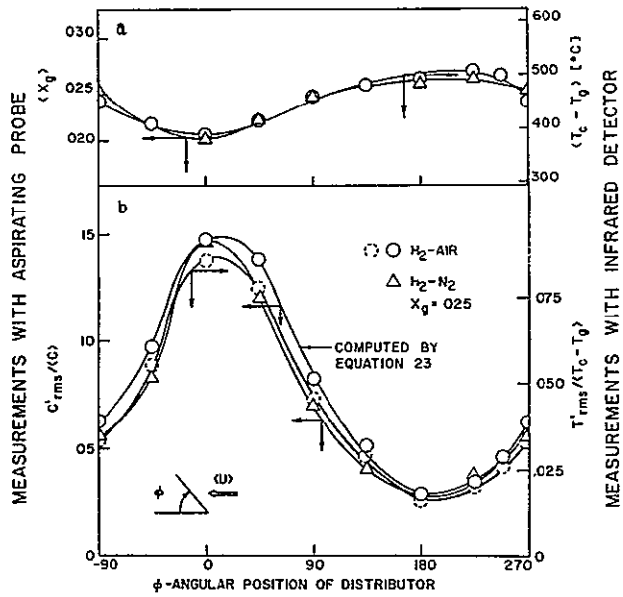


Fig. 5. The effect of the angular position of the feed distributor on: $\langle X_g \rangle$ and $C'_{rms}/\langle C \rangle$ as measured by the aspirating probe, on $\langle T_c - T_g \rangle$ and $T'_{rms}/\langle T_c - T_g \rangle$ for a 0.075-mm wire as measured by the infrared detector, and on $C'_{rms}/\langle C \rangle$ as computed from $T'_{rms}/\langle T_c - T_g \rangle$ by Equation (23).

probability density function (pdf), and the spectral density function (sdf) were sensitive to the angular position of the distributor at point A. The intensity of the concentration fluctuations at the catalytic wire station depended on the feed port location and was in the following order A >> B > C > D.

A series of experiments were carried out to determine the relation between the concentration fluctuations (as measured by the aspirating probe for a hydrogen-nitrogen mixture) and the root mean square temperature fluctuations (as measured by the infrared detector) at the center of platinum wires during the oxidation of hydrogen in air. In all the experiments reported here, unless otherwise specified, the channel Reynolds number was 4000, the turbulent intensity ($u'_{rms}/\langle u \rangle$) was 0.11, $T_g = 24°C$, and $d = 0.075$ mm.

Experiments with a mixture of 3%v H_2 in air yielded values of T'_{rms} equal to 0.84, 1.10, and 2.7°C and $C'_{rms}/\langle C \rangle$ equal to 0.002, 0.004, and 0.006 when the hydrogen was injected from feed points D, C, and B, respectively. These results indicate a direct correlation between the concentration and temperature fluctuations and point out that in the absence of concentration fluctuations T'_{rms} will be very small (the noise level of the detector was 0.7°C). The above measured values of $C'_{rms}/\langle C \rangle$ are accurate to only one significant digit due to the poor signal to noise ratio. Thus, in order to obtain data suitable for a critical examination of the theoretical predictions it was necessary to use injection point A, which yields high values of concentration fluctuations at the wire station.

Experiments with the aspirating probe as well as chromatographic analysis of samples taken at various points across the channel revealed that gradients in the time averaged concentration existed next to the wire when injection point A was used to obtain a mixture of 2.5%v H_2 in air. Moreover, the time averaged concentration next to the wire depended on the angular position of the feed distributor. A comparison of $\langle x_g \rangle$ and $\langle T_c - T_g \rangle$ shown in Figure 5a indicates that the variations in the time averaged concentration next to the wire were responsible for

the variation of the time averaged temperature rise of the wire with ϕ and its deviation from 475°C. (If mixing of the reactants were such that no gradients in the time averaged concentration existed then for this mass transfer limited reaction $\langle T_c - T_g \rangle$ would have been 475°C. The injection angle ϕ is defined as 0 for injection in the direction of air flow and positive when measured in the clockwise direction, for example, vertical upwards injection corresponds to 90°.)

Figure 5b describes the effect of the angular position of the feed distributor (point A) on the measured concentration and temperature fluctuations. The results indicate a close similarity between $C'_{rms}/\langle C \rangle$ and $T'_{rms}/\langle T_c - T_g \rangle$. Thus, while $C'_{rms}/\langle C \rangle$ changes by a factor of about five as the distributor is rotated, the ratio of $(C'_{rms}/\langle C \rangle)/(T'_{rms}/\langle T_c - T_g \rangle)$ changes by no more than 8%. The values of T'_{rms} obtained in these experiments were quite large compared to the values obtained at points B, C, and D, and ranged from 8.6° to 33.4°C. Since peak to peak temperature fluctuations are about six times larger than T'_{rms}, the temperature fluctuations could be observed visually.

The assumption that flickering is induced by concentration fluctuations when $a >> 1$ led to the development of Equation (23) in the theoretical section. In order to test its validity it was used to compute $C'_{rms}/\langle C \rangle$ from the measured values of $T'_{rms}/\langle T_c - T_g \rangle$. To accomplish these computations, values of τ and τ_c must be known. The value of τ was determined from heat transfer experiments as 0.070 sec. Measurements with the aspirating probe showed that τ_c varied between 0.030 and 0.039 s depending on the injection angle. The computed values of $C'_{rms}/\langle C \rangle$ shown in Figure 5b agree very well with the values measured by the aspirating probe, and support the validity of Equation (23).

Figure 6a describes the measured values of $T'_{rms}/\langle T_c - T_g \rangle$ as a function of the wire diameter for a mixture of 2.5% vol. H_2 in air. The mixture was injected at point A using two different angular positions of the distributor. The temperature fluctuations in these cases almost doubled as the wire diameter was decreased from 0.125 to 0.025 mm. The magnitude of $\langle T_c - T_g \rangle$ in these experiments was about 475°C, and the flickering was quite large in some cases.

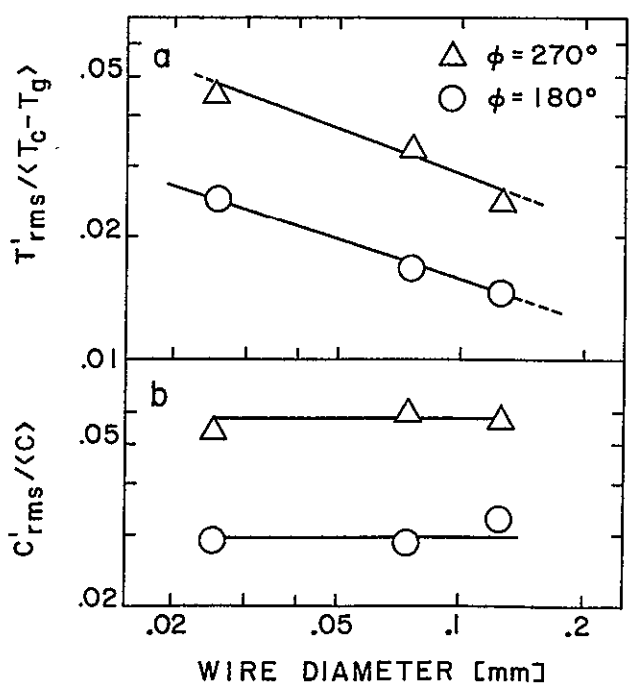


Fig. 6. (a) The effect of wire diameter and position of distributor on the temperature fluctuation for a mixture containing 2.5% H_2 in air; (b) Computed values of $C'_{rms}/\langle C \rangle$ from $T'_{rms}/\langle T_c - T_g \rangle$ using Equation (23).

The above measurements of T'_{rms} were used as a further test of the validity of Equation (23). The values of τ as measured by heat transfer experiments were 0.011, 0.070, and 0.141 s for wires of 0.025, 0.075, and 0.125 mm diameter, respectively. Measurements with the aspirating probe determined τ_c as 0.039 and 0.030 s for $\phi = 180°$ and 270°, respectively. Figure 6b presents the values of $C'_{rms}/\langle C \rangle$ calculated by Equation (23) from the measured temperature fluctuations and the measured values of τ and τ_c. The computed concentration fluctuations are independent of the diameter of the wire used for measuring the temperature fluctuations and agree within 4% with values of $C'_{rms}/\langle C \rangle$ measured directly by the aspirating probe. These results strongly support Equation (23) and the notion that when the parameter a is large compared to unity temperature fluctuations are induced mainly by concentration fluctuations in the reacting gas mixture.

It should be noted that Equation (23) was derived assuming that Equation (21) is an adequate representation of the concentration spectral density function (sdf). Fortunately, the constants appearing in (23) are rather insensitive to the exact form of the concentration sdf as is often the case with results based on integrals of assumed profiles. This insensitivity to the sdf enables application of (23) even when Equation (21) is no more a proper representation of the concentration sdf. On the other hand, the temperature fluctuation sdf is rather sensitive to the exact form of the concentration fluctuation sdf (especially when $\tau_c > \tau$) and Equation (24) cannot be expected to be valid when (21) is not. However, Equation (18) should be valid regardless of the form of the sdf of the concentration fluctuations, provided that the hypothesis about the cause of flickering is correct.

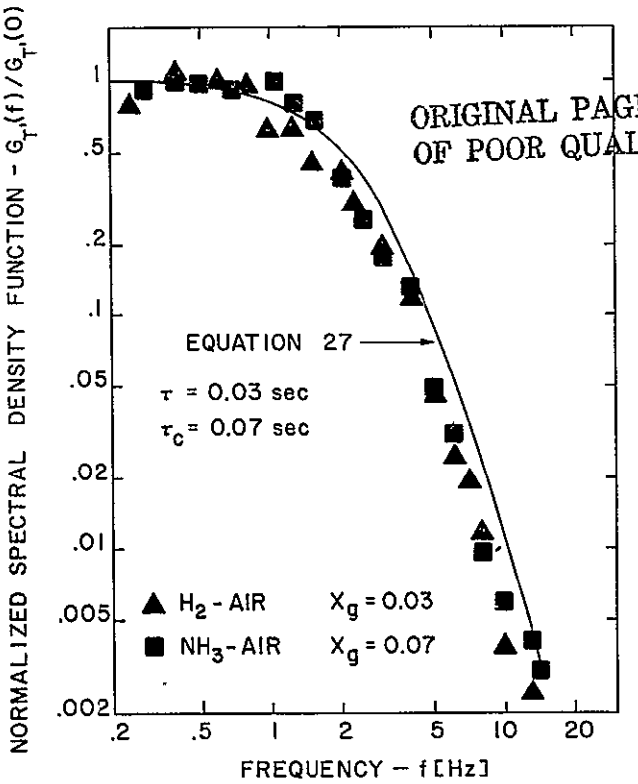
The sdf of the temperature fluctuations was measured for various reactions, wires, and operating conditions. Figure 7 describes the normalized sdf $G_{T'}(f)/G_{T'}(0)$ for a 0.075-mm wire using mixtures of either ammonia or hydrogen in air injected at point C. According to Equation (24) this normalized sdf should satisfy the relation

$$\frac{G_{T'}(f)}{G_{T'}(0)} = \frac{1}{[1 + (2\pi f\tau)^2]\,[1 + (2\pi f\tau_c)^2]} \qquad (27)$$

For the experiments shown in Figure 7, τ was determined to be 0.07 s, while τ_c was measured with the aspirating probe (H_2-N_2 mixture) and found to be 0.03 s. The experimental results agree well with the theoretical curve which uses the value of τ_c for the H_2-N_2 mixture. This agreement is not a critical test of the validity of (27) since $\tau > \tau_c$ and the theoretical curve is rather insensitive to τ_c in the range of frequencies for which the measurements were made. A critical test of the theory requires experiments with very thin wires for which $\tau < \tau_c$. Experiments with a 0.025-mm diameter wire ($\tau = 0.011$ s) have been inconclusive in this respect due to vibrations of the wire which introduced large errors in the measurements. Thus, the sdf of the temperature fluctuations, in contrast to the root mean square, was not useful in determining the major cause for flickering. Additional information concerning the sdf is presented by Zuniga (1974).

Instantaneous velocity measurements by a hot wire indicated that the pdf of the velocity fluctuations at the catalytic wire station was skewed with the mode smaller than the average. Changes of the angular position of the feed distributor did not affect the magnitude of the veloc-

ity fluctuations or the pdf. However, rotation of the distributor affected both the magnitude and the shape of the pdf of the temperature fluctuation on the wire. Figure 8 describes the temperature fluctuation pdf on a 0.075-mm wire during the oxidation of a mixture of 2.5% vol. hydrogen in air. The pdf's are skewed, and the position of the mode relative to the average depends on the angular position of the distributor. Measurements of the concentration fluctuation pdf of 2.5% vol. hydrogen in nitrogen showed that they were also skewed and very similar to those of the temperature fluctuation pdf. The similarity between the temperature and concentration fluctuation pdf's, and



ORIGINAL PAGE IS
OF POOR QUALITY

Fig. 7. The normalized spectral density function of temperature fluctuations at the center of a 0.075-mm diameter wire during the oxidation of either ammonia or hydrogen in air.

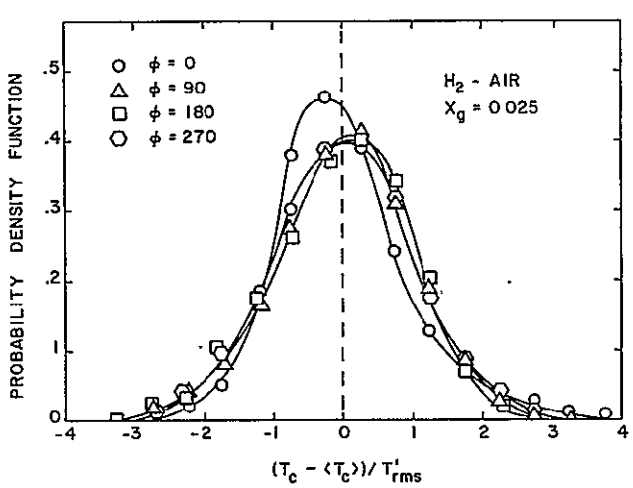


Fig. 8. The probability density function of temperature fluctuations on a 0.075-mm wire as a function of the angular position of the feed distributor.

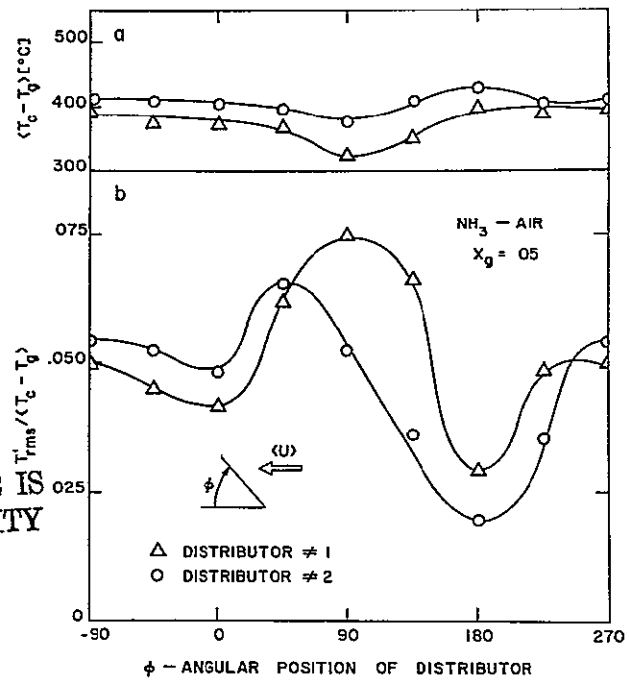


Fig. 9. The influence of the angular position of the distributor and the size of the holes in the distributor on the magnitude of: (a) $\langle T_c - T_g \rangle$ and (b) $T'_{rms}/\langle T_c - T_g \rangle$ during the oxidation of ammonia on a 0.075-mm wire.

the lack of any correlation between the velocity and temperature fluctuations further supports the argument that temperature fluctuations are induced mainly by the concentration fluctuations and not by velocity fluctuations.

Temperature fluctuations were measured during the oxidation of ammonia in air on a 0.075-mm diameter wire using two geometrically similar distributors, one of which (#2) had 0.51-mm diameter injection holes as compared with the 0.34-mm diameter holes for the standard distributor. The results (Figure 9) indicate that both T'_{rms} and $\langle T_c - T_g \rangle$ depend on the size of the holes as well as on the angular position of the distributor. This result is of course similar to that obtained with hydrogen. A comparison of Figures 9 and 5 indicates that the maxima in T'_{rms} were attained at different angular positions of the standard distributor during the oxidation of either hydrogen or ammonia. This must be due to the difference in the effect of ϕ on the magnitude of the concentration fluctuations for different reactants. Unfortunately, $C'_{rms}/\langle C \rangle$ for ammonia cannot be measured using the aspirating probe due to the small difference in the thermal conductivities of air and ammonia. When the distributor at point A was used to obtain the 2.5% vol. mixture of H_2 the injection velocity of the jets was 1.39×10^4 cm/s. However, when a 5% vol. mixture of NH_3 was prepared the injection velocities were 2.84×10^4 and 1.29×10^4 cm/s for feed distributors 1 and 2, respectively. This difference in the injection velocities and the density of H_2 and NH_3 is responsible for the variation in the effect of ϕ on $C'_{rms}/\langle C \rangle$.

There is no question, however, that concentration fluctuations are the main factor affecting the magnitude of temperature fluctuations in both the hydrogen and ammonia oxidation reactions. This conclusion was further confirmed by passing a premixed mixture (5.0% vol.) of ammonia in air over the catalytic wire at $N_{Re} = 3200$ and $T_g = 24°C$. The measured $T'_{rms}(0.76°C)$ was very close

to the root mean square noise level of the detector (0.70°C). Thus, we conclude from our experiments that large temperature fluctuations are caused mainly by concentration fluctuations of the reactants when the parameter a is large.

CAUSES AND MAGNITUDE OF FLICKERING IN GAUZE CONVERTORS

A theoretical model of gauze convertors will be used here to predict the main causes of flickering and to estimate its magnitude. The model assumes that: (1) the reaction rate is limited by mass transfer; (2) the parameter a is large; (3) plug flow of the gas through each gauze; (4) the length of the effective flow path through a single gauze is l and the surface area of a single gauze is $a_v l$ per unit cross section of the reactor; (5) the local temperature of each gauze layer is uniform in the flow direction and the rate of heat transfer by conduction and radiation between successive gauze layers is negligible; and (6) the effect of axial concentration dispersion can be ignored.

The concentration of the gaseous reactant flowing through gauze i is described by the equation

$$u_i \frac{dC}{dz} = -k_{C,i}\, a_v C \tag{28}$$

Defining C_i as the concentration of the gas stream leaving gauze i we obtain from (28)

$$\frac{C_i}{C_{i-1}} = \exp\left(-\frac{k_{C,i}\, a_v l}{u_i}\right) = \exp(-N_{St,i}\, a_v l) = \xi_i \tag{29}$$

The mass transfer coefficient for a long cylinder can be described by the correlation (Treybal, 1968)

$$N_{Sh} = 0.43 + 0.53\, N_{Sc}^{0.31}\, N_{Re,w}^{0.5} \tag{30}$$

For a typical high pressure ammonia oxidation convertor the gauze wire Reynolds number is about 35 and the Stanton number is essentially a constant changing by less than 15% as the gas average film temperature increases from 500° to 900°C. Thus, it will be assumed that ξ_i is the same for all the gauze layers, yielding

$$\frac{C_i}{C_o} = \xi^i \tag{31}$$

At steady state the heat loss from any gauze is equal to the heat generated by the reaction. Hence, the temperature of each gauze satisfies

$$T_i - T_{gi,lm} = \frac{(-\Delta H) k_c\, C_{i,lm}}{h} = \frac{(-\Delta H) C_o}{\rho_f C_{p_f}} \frac{C_{i,lm}}{C_o} N_{Le}^{-2/3} \tag{32}$$

where we have used the heat and mass transfer analogy

$$\frac{k_c\, \rho_f C_{p_f}}{h} = N_{Le}^{-2/3} \tag{33}$$

and the subscript i, lm denotes the logarithmic average across gauze i. An enthalpy balance yields

$$T_{gi,lm} - T_{go} = \frac{(-\Delta H)(C_o - C_{i,lm})}{\rho_f C_{p_f}} = \frac{(-\Delta H) C_o}{\rho_f C_{p_f}} \left(1 - \frac{C_{i,lm}}{C_o}\right) \tag{34}$$

Addition of (34) and (32) yields

$$T_i - T_{go} = \frac{(-\Delta H) C_o}{\rho_f C_{p_f}} \left[1 + (1 - N_{Le}^{-2/3}) \frac{(1-\xi)\xi^{i-1}}{\ln \xi}\right] \tag{35}$$

If $\tau << \tau_c$ and τ_E this relation can be used to predict the temperature fluctuation T_i'. Substitution of

$$C_o = \langle C_o \rangle + C' \tag{36}$$

$$\xi = \langle \xi \rangle + \xi' \tag{37}$$

into (35) yields after discarding second-order terms

$$\frac{T_i'}{\langle T_i \rangle - T_{go}} = \frac{C_o'}{\langle C_o \rangle} + K_i \frac{\xi'}{\langle \xi \rangle} \tag{38}$$

where

$$K_i = \frac{(-\Delta H)\langle C_o \rangle (1 - N_{Le}^{-2/3})}{\rho_f C_{p_f} (\langle T_i \rangle - T_{go})} \left[\left(\frac{(i-1)\langle \xi \rangle^{i-1} - i\langle \xi \rangle^i}{\ln \langle \xi \rangle}\right) - \frac{\langle \xi \rangle^{i-1} - \langle \xi \rangle^i}{(\ln \langle \xi \rangle)^2}\right] \tag{39}$$

The fluctuations in ξ are induced by the turbulent velocity fluctuations and thus cause fluctuations in the mass transfer coefficient. Substitution of (9) into (29) yields

$$\frac{\xi'}{\langle \xi \rangle} = 1 - \langle \xi \rangle^{(1-N)u'/\langle u \rangle} \tag{40}$$

Substitution of (40) into (38), squaring, and time averaging yields

$$\frac{\langle T_i'^2 \rangle}{(\langle T_i \rangle - T_{go})^2} = \frac{\langle C_o'^2 \rangle}{\langle C_o \rangle^2} + 2K_i \frac{\langle C_o' \rangle (1 - \langle \xi \rangle^{(1-N)u'/\langle u \rangle})}{\langle C_o \rangle} + K_i^2 \langle (1 - \langle \xi \rangle^{(1-N)u'/\langle u \rangle})^2 \rangle \tag{41}$$

Computation of K_i for typical industrial conditions indicate that its value is usually much smaller than unity. This indicates that the main cause of flickering in industrial gauzes is the concentration fluctuations and not the turbulent velocity fluctuations. Numerical computations of the amplitude of temperature fluctuations caused either by velocity fluctuations or concentration fluctuations for a typical ammonia oxidation convertor for which $x_o = 0.11$, $N_{Le} = 0.895$, $\langle \xi \rangle = 0.5$, $T_{go} = 25°$, and $N = 0.45$ presented in Table 3 demonstrate this point.

When the time constant of a single wire is about equal to or larger than τ_C or τ_E the amplitude of the flickering will be attenuated and (41) will overestimate $\langle T_i'^2 \rangle$. In this case a numerical simulation of the transient behavior of the various gauze layers is required for predicting T_i'. Comparison of (23) and (41) indicates that T'_{rms} will be attenuated by a factor of $1/\sqrt{1 + \tau/\tau_c}$. For a typical high pressure ammonia convertor $d_w = 0.075$ mm and the wire Reynolds number is 35, yielding $\tau = 0.03$ s. The reactor diameter is about 1 m and the average gas velocity 2 m/s. Since $\Lambda_f \approx D/3$ then $\tau_E = \Lambda_f \langle u \rangle \approx \tau_c$ is approximately 0.166 s and $\tau/\tau_c = 0.18$. Thus, the attenuation of T'_{rms} due to the heat capacity of the wire should be very small in industrial convertors and (41) should yield a good prediction of T'_{rms}. It should be noted that in some of our single wire experiments $\tau > \tau_c$ and the attenuation was not negligible.

The above model predicts that flickering is a localized phenomenon induced mainly by concentration fluctuations.

J. 338

TABLE 3. COMPUTATIONS OF K_i AND T'_{rms} FOR AN AMMONIA CONVERTER FOR WHICH $\langle x_0 \rangle = 0.11$, $N_{Le} = 0.895$, $\langle \xi \rangle = 0.5$, $T_{go} = 25°$, and $N = 0.45$

Layer number i	$\langle T_{gi} \rangle$ [C°]	$\langle T_i \rangle$ [C°]	$10^2 K_i$	T'_{rms} $u'_{rms} = 0.1 \langle u \rangle$ $C'_{rms} = 0$	T'_{rms} $u'_{rms} = 0$ $C'_{rms} = 0.05 \langle C \rangle$
1	253	887	2.3	0.7	43.1
2	547	865	3.9	1.2	42.0
3	695	853	3.3	1.0	41.4
4	768	848	2.4	0.7	41.2
5	805	845	1.5	0.5	41.0
10	841	842	0.1	0.03	40.9

We will now examine briefly the influence of conduction on flickering. The characteristic time for conduction along a single wire is

$$\tau_k = \rho c_p s^2/k \tag{42}$$

while that for heat losses by forced convection from a wire is

$$\tau_k = \frac{\rho c_p d}{4h} = \frac{\rho c_p d^2}{4 N_{Nu} k_f} \tag{43}$$

These two time scales are equal when

ORIGINAL PAGE IS OF POOR QUALITY

$$\frac{s}{d} = \left[\frac{k_w}{4 k_f N_{Nu}} \right]^{0.5} \tag{44}$$

Equation (44) predicts that under typical operating conditions the two time scales are equal when the wire length is about ten diameters. Hence, the influence of conduction is rather localized and is not expected to attenuate to a large extent local temperature fluctuations induced by instantaneous concentration fluctuations.

The above results indicate that flickering is induced primarily by concentration fluctuations. Hence, it is important to estimate the magnitude of concentration fluctuations in industrial convertors. The only theoretical correlations developed so far are restricted to a homogeneous isotropic turbulent flow field. Corrsin (1957) suggested that for an idealized mixer

$$\frac{\langle C'^2(t) \rangle}{\langle C'^2(0) \rangle} = \exp\left(-\frac{6\nu t}{\lambda^2} \right) \tag{45}$$

According to Hinze (1958, p. 186)

$$\frac{L_e}{\lambda} = \frac{4}{3} \frac{\Lambda_f}{\lambda} = \frac{\gamma}{15} \frac{u'_{rms} \lambda}{\nu} \tag{46}$$

where γ is a constant of about one. Substitution of $t = X/\langle u \rangle$ and (43) into (44) yields

$$\frac{\langle C'^2(t) \rangle}{\langle C'^2(0) \rangle} = \exp(-0.4\gamma\, m) \tag{47}$$

where

$$m = \frac{3}{4} \frac{X}{\Lambda_f} \frac{u'_{rms}}{\langle u \rangle} \tag{48}$$

Beek and Miller (1959) carried out a numerical computation of the spectral transfer of concentration fluctuations in the wave number space. Their graphical results indicate that after a short distance downstream from the mixer the concentration fluctuations decay exponentially as a function of m. Their design charts show a significant improvement in mixing when the number of inlet nozzles is increased. However, this result should be applied with care since it is based on the yet experimentally unproven conjecture that the size of the concentration eddies in the inlet is inversely proportional to the square root of the number of injection nozzles. Moreover, the calculations are restricted to the special cases in which the injection velocity is equal to final mixed stream velocity.

In a typical ammonia convertor the distance between the mixer and the gauze $X \simeq 20D$, $u'_{rms}/\langle u \rangle = 0.03$, and $\Lambda_f \simeq 0.4D$ so that $m \simeq 1.13$. For this case (47) predicts that $\langle C'^2(t) \rangle / \langle C'^2(0) \rangle = 0.64$, while the graphs of Beek and Miller predict 0.57 and 0.014 for the cases of one and 64 injection nozzles, respectively. Although the above predictions are a rough approximation, they indicate that under commercial conditions the distance between the mixer and the gauze is not sufficient to eliminate concentration fluctuations.

One of the main difficulties in estimating $C'_{rms}(t)$ is the lack of a reliable estimation technique for $C'_{rms}(0)$ in industrial mixers. Keeler et al. (1965) suggested for an ideal one-dimensional system the approximation

$$\frac{C'_{rms}(0)}{\langle C(0) \rangle} \simeq \sqrt{\frac{G}{I}} \tag{49}$$

where G and I are the main stream and tracer volumetric flow rates, respectively. Toor (1969) obtained the same result for an idealized mixer in pipe flow. This prediction is at best a crude estimate for an industrial gas mixer for which gradients in the time averaged concentrations usually exist next to the injection nozzles. For these mixers, the initial root mean square concentration has to be determined experimentally. Fortunately, in most practical cases the time averaged concentration gradients decay faster than the concentration fluctuations, and the graphs of Beek and Miller (1959) should be a good approximation for the decay of $\langle C'^2(t) \rangle$ at the center of the pipe at points where no gradients in the time averaged concentration exist. The decay rate next to the wall is expected to exceed that predicted by their graphs.

The prediction that C'_{rms} decays exponentially as a function of m is most useful for scale-up purposes and for equipment modifications. Thus, measurements of C'_{rms} at several points downstream from the mixer should enable one to predict the effect of changing the length of the inlet pipe on C'_{rms} at the gauze. This should enable design or modification of equipment to reduce concentration fluctuations and the accompanying flickering.

CONCLUDING REMARKS

The present study was concerned with the causes and magnitude of flickering of single wires and gauze convertors on which a single catalytic reaction occurs. Flickering has a deleterious influence on gauze convertors since it increases the precious metal loss. Dorawla and Douglas (1971) have shown that in complex reaction networks the yield of a desired intermediate can be increased by oscillatory operation. Wandrey and Renken (1973) have recently shown that periodic variation of the concentration of the feed to a gauze converter (frequency 0.5-2 per min.) has a significant influence on the selectivity (CO/

CO_2 ratio) during the oxidation of cyclohexane on a catalytic gauze. Therefore, it is most likely that concentration fluctuations which induce flickering affect the yield in industrial gauze convertors used for the production of HCN.

Our analysis indicates that the parameter a is usually large in industrial reactors and that flickering of the magnitude observed in industrial ammonia convertors must be caused by imperfect mixing of the reactants. It is desirable to minimize the amplitude of the flickering in order to reduce the precious metal loss, and our analysis indicates that this can be accomplished by improving the mixing of the reactants. This prediction merits a test under commercial conditions and is supported by the observation of Heywood (1973) that when a catchment gauze was placed below the platinum gauze there was evidence for improved gas mixing and a reduction in the rate of deterioration of the gauze.

At the present time, mixing theory is restricted to certain idealized flow fields and mixing devices and is not suitable for predicting the performance of various industrial mixing devices. This study points out the need of improving our knowledge and understanding of mixing under industrial conditions. One of the difficulties in investigating mixing is that most probes for concentration fluctuations can be operated only with a very limited number of substances which are not normally encountered in industrial reactors. The fact that flickering of a single wire is directly related to the concentration fluctuations has led to the development of a new probe which can measure the intensity of concentration fluctuations of a number of gaseous mixtures of practical interest. This probe should be useful in many applications and will be described elsewhere.

ACKNOWLEDGMENT

The financial support of the NSF through grant GK 18226 is gratefully acknowledged.

NOTATION

a = capacity term defined by Equation (11)
a_v = gauze surface area per unit volume, cm^{-1}
A = wire cross-sectional area, cm^2
AS = occupied catalyst sites, g mole cm^{-2}
C = reactant concentration in fluid, g mole cm^{-3}
c_p = wire heat capacity, cal g^{-1} °K^{-1}
c_{pf} = fluid heat capacity, cal g^{-1} °K^{-1}
d = wire diameter, cm
D = diameter, cm
D_{AB} = binary diffusion coefficient, cm^2 s^{-1}
E_m = metal loss activation energy, cal g $mole^{-1}$
f = frequency, Hz
$f_{1/2}$ = half-power frequency, Hz
G = main stream volumetric flow rate, cm^3 s^{-1}
$G_y(f)$ = one-sided power spectral density of random process y
h = heat transfer coefficient, cal s^{-1} cm^{-2} °K^{-1}
ΔH = heat of reaction, cal g $mole^{-1}$
i = catalytic gauze layer number, integer
I = tracer injection volumetric flow rate, cm^3 s^{-1}
k_1 = mass transfer and adsorption rate constant, cm^3 g $mole^{-1}$ s^{-1}
k_c = mass transfer coefficient, cm s^{-1}
k = thermal conductivity, cal s^{-1} cm^{-1} °K^{-1}
K_i = constant defined by Equation (39)
l = flow path length through a single gauze layer, cm
L = total catalytic sites, g mole cm^{-2}
L_e = characteristic scale of turbulence, cm
M = weight of precious metal gauze, g
m = parameter defined by Equation (48)
M_f = molecular weight of fluid, g g $mole^{-1}$
n = empirical velocity exponent in Equation (10)
N = constant defined by Equation (10)
N_{Le} = Lewis number, $k_f/\rho_f D_{AB} c_{pf}$
N_{Nu} = Nusselt number, hD/k_f
N_{Pr} = Prandtl number, $c_{pf}\mu_f/k_f$
N_{Re} = Reynolds number, $D u/\nu_f$
N_{Sc} = Schmidt number, ν_f/D_{AB}
N_{Sh} = Sherwood number, $k_c D/D_{AB}$
N_{St} = Stanton number, k_c/u
P = wire perimeter, cm
r = surface reaction rate, g mole cm^{-2} s^{-1}
$r(T)$ = relative metal loss, Equation (1)
R = gas constant, cal g $mole^{-1}$ °K^{-1}
S = unoccupied catalytic sites, g mole cm^{-2}
s = distance along wire axis, cm
t = time, s
T = temperature, °K
u = fluid velocity, cm s^{-1}
X = distance from mixer, cm
x_g = mole fraction reactant in fluid
z = axial distance in catalytic gauze, cm

Greek Letters

α = constant in Equation (8)
β = constant in Equation (8)
γ = constant used in Equation (47)
λ = concentration microscale, cm
Λ_c = concentration macroscale, cm
Λ_f = velocity longitudinal macroscale, cm
μ = viscosity, g cm^{-1} s^{-1}
ν = kinematic viscosity, μ/ρ, cm^2 s^{-1}
ξ = defined by Equation (29)
ρ = wire density, g cm^{-3}
ρ_f = fluid density, g cm^{-3}
τ = wire time constant, Equation (19), s
τ_c = Eulerian time scale of concentration fluctuation, s
τ_E = Eulerian time scale of velocity fluctuations, s
τ_k = characteristic time for conduction along wire axis, s
τ_w = wire time constant defined by Equation (26)

Subscripts

ad = adiabatic
f = fluid
g = gas
o = inlet condition
w = wire

Superscripts

$'$ = fluctuating component

LITERATURE CITED

Becker, H. A., R. E. Rosenweig, and J. R. Gwozdz, "Turbulent Dispersion in a Pipe Flow," *AIChE J.*, 12, 964 (1966).

Beek, J., Jr., and R. S. Miller, "Turbulent Transport in Chemical Reactors," *Chem. Eng. Progr. Symp. Ser. No. 25*, 55, 23 (1959).

Blackshear, P. L., Jr., and L. Fingerson, "Rapid Response Heat Flux Probe for High Temperature Gases," *ARS J.*, 1709 (1962).

Corrsin, S., "Simple Theory of an Idealized Turbulent Mixer" *AIChE J.*, 3, 329 (1957).

———., "The Isotropic Turbulent Mixer: Part II. Arbitrary Schmidt Number," *ibid.*, 10, 870 (1964).

Dorawala, T. G., and J. M. Douglas, "Complex Reactions in Oscillating Reactors," *ibid*, 17, 974 (1971).

Edwards, W. M., "Temperature Fluctuations on Catalytic Wires and Gauzes," Ph.D. thesis, Univ. of Houston, Texas (1973).

Edwards, W. M., F. L. Worley, Jr., and D. Luss, "Temperature Fluctuations (Flickering) on Catalytic Wires and Gauzes—II Experimental Study of Butane Oxidation on Platinum Wires," *Chem. Eng. Sci.*, 28, 1479 (1973).

J.340

Ervin, M. A., and D. Luss, "Temperature Fluctuations (Flickering) of Catalytic Wires and Gauzes—I Theoretical Investigation," *ibid.*, 27, 339 (1972).

Freymuth, P., and M. S. Uberoi, "Structure of Temperature Fluctuations in the Turbulent Wake Behind a Heated Cylinder," *Phys. Fluds*, 14, 2574 (1971).

Gillespie, G. R., and R. E. Kenson, "Catalyst System for Oxidation of Ammonia to Nitric Acid," *Chem. Tech.*, 1, 627 (1971).

Heywood, A. E., "Platinum Recovery in Ammonia Oxidation Plants," *Plat. Met. Rev.*, 17, 118 (1973).

Hinze, J. O, *Turbulence*, McGraw Hill, New York (1959).

Keller, R. N., E. E. Petersen, and J. M. Prausnitz, "Mixing and Chemical Reaction in Turbulent Flow Reactors," *AIChE J.*, 11, 221 (1965).

Lee, J., and R. S. Brodkey, "Turbulent Motion and Mixing in a Pipe," *ibid.*, 10, 187 (1964).

Newman, D. J., and B. J. Hulbert, "Nitric Acid Development Aims at Cutting Pollution and Catalyst Costs," *Europ. Chem. News*, Large Plants Supp., (Sept. 24, 26, 1971).

Nowak, E. J., "Prediction of Platinum Losses during Ammonia Oxidation," *Chem. Eng. Sci.*, 24, 421 (1969).

Toor, H. L., "Turbulent Mixing of Two Species with and without Chemical Reactions," *Ind. Eng. Chem. Fundamentals*, 8, 655 (1969).

Treybal, R. E., *Mass Transfer Operation*, 2nd edit., McGraw Hill, New York (1968).

Wandrey, C. and A. Renken, "Zur Beeinflussung der Produktverteilung durch periodische konzentrationsschwankungen bei der Oxidation von Kohlenwasserstoffen," *Chim. Ing. Teckn.*, 12, 854 (1973).

Zuniga-Chaves, J. E., "The Causes for Temperature Fluctuations on Catalytic Wires," M.Sc. thesis, Univ. Houston, Texas (1974).

Manuscript received December 20, 1973; revision received January 29 and accepted January 30, 1974.

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: Measurements of Concentration Fluctuations in Gaseous Mixtures

(B) Project Abstract:

A new experimental technique is presented for measuring concentration fluctuations in gaseous mixtures using a catalytic wire on which a mass-transfer-limited exothermic reaction occurs. This method utilizes a catalytic sensor in conjunction with a constant temperature anemometer unit. It should be a useful tool for studies to improve the design of gas mixing equipment and of chemical reactors in which the yield and/or conversion depend on the degree of mixing.

(C) Publication: M.S. Thesis in Chemical Engineering and Ind. Eng. Chem. Fundamentals, 15, p. 341, (1976)

(D) Year: Completed in 1974

(E) Department: Chemical Engineering

(F) Student Names: J. E. Zuniga Chaves and W. M. Edwards

(F) Faculty Advisor: Professors F. L. Worley, Jr. and Dan Luss

Reprinted from I&EC FUNDAMENTALS, Vol. 15, Page 341, November 1976
Copyright 1976 by the American Chemical Society and reprinted by permission of the copyright owner

ORIGINAL PAGE IS
OF POOR QUALITY

Measurements of Concentration Fluctuations in Gaseous Mixtures

William M. Edwards, Jorge E. Zuniga-Chaves, Frank L. Worley, Jr., and Dan Luss*

Department of Chemical Engineering, University of Houston, Houston, Texas 77004

A new experimental technique is presented for measuring concentration fluctuations in gaseous mixtures using a catalytic wire on which a mass-transfer-limited exothermic reaction occurs. This method utilizes a catalytic sensor in conjunction with a constant-temperature anemometer unit. It should be a useful tool for studies to improve the design of gas mixing equipment and of chemical reactors in which the yield and/or conversion depend on the degree of mixing.

Information about instantaneous concentration fluctuations in gaseous mixtures is very useful for the design of industrial gas mixing equipment and of reactors in which the yield and/or conversion are sensitive to the mixing of the reactants. The techniques now available for measuring these fluctuations (Corrsin, 1949; Blackshear and Fingerson, 1962; McQuaid and Wright, 1973) are either suitable only for mixtures in which the heat transfer properties (e.g., k, C_p) of the two gases are significantly different, or require the simultaneous use of several sensors. Recently, optical techniques were developed which use interferometer or crossed Schlieren optical systems to detect selective index gradients for mixtures of gases with large density differences (Wilson and Prosser, 1971). These can be related to density fluctuations and hence concentration fluctuations for isothermal gas flow.

We describe here a novel technique for measuring root-mean-square concentration fluctuations. The method utilizes a catalytic sensor in conjunction with a constant-temperature anemometer unit. It is applicable to gaseous mixtures containing species which can react rapidly and exothermally on a platinum wire, such as hydrocarbons and oxygen. A major advantage of this technique is that it can be applied to mixtures of gases with similar physical properties.

Theoretical Background

Consider a catalytic wire whose capacity parameter, defined as the ratio of the characteristic time for changes in wire temperature to the characteristic time for surface concentration changes due to changes in the rate of mass transfer, is large. When a mass-transfer-limited exothermic reaction occurs on such a wire its temperature may be described by the equation (Edwards et al., 1974).

$$A\rho C_p \frac{\partial T}{\partial t} = k_w A \frac{\partial^2 T}{\partial x^2} + hP(T_g - T) + P(-\Delta H)k_c C \quad (1)$$

For wires with negligible end effects, the first term in the right-hand side of (1) can be discarded (this term is normally ignored when dealing with hot wires with length-to-diameter ratios larger than 200 (Hinze, 1959)). If in addition the wire is heated by means of an electric current, eq 1 takes the form

$$A\rho C_p \frac{dT}{dt} = hP(T_g - T) + P(-\Delta H)k_c C + JI^2R \quad (2)$$

The temperature of the wire can be maintained at a constant level by using a constant-temperature anemometer to control the heating current. Therefore, the time derivative in (2) vanishes. Moreover, the transport coefficients, the limiting reactant concentration, and the electric current can be expressed as the sum of a stationary time average and a fluctuating component thus enabling the reduction of (2) into

$$\langle h\rangle (T - T_g)\left(1 + \frac{h'}{\langle h\rangle}\right) = (-\Delta H)\langle k_c\rangle\langle C\rangle \times \left(1 + \frac{k_c'}{\langle k_c\rangle} + \frac{C'}{\langle C\rangle}\right) + J\langle I\rangle^2 R\left(1 + \frac{2I'}{\langle I\rangle}\right)\Big/P \quad (3)$$

In this equation all second-order terms have been neglected and $\langle\ \rangle$ denotes a stationary time average.

Time averaging of (3) yields

$$\langle h\rangle (T - T_g) = (-\Delta H)\langle k_c\rangle\langle C\rangle + J\langle I\rangle^2 R/P \quad (4)$$

When no electrical heating is used

$$T^* - T_g \triangleq (-\Delta H)\langle k_c\rangle\langle C\rangle/\langle h\rangle \quad (5)$$

and $T^* - T_g$ is defined as the adiabatic temperature rise. Subtracting (4) from (3) and division by (5) yields

$$\frac{h'}{\langle h\rangle} + m\left(\frac{h'}{\langle h\rangle} - \frac{k_c'}{\langle k_c\rangle}\right) = m\frac{C'}{\langle C\rangle} + \frac{2I'}{\langle I\rangle} \quad (6)$$

p.343

where

$$m = \frac{T^* - T_g}{T - T^*} = \frac{R^* - R_g}{R - R^*} \tag{7}$$

and

$$\frac{J\langle I\rangle^2 R}{P\langle h\rangle(T - T^*)} = 1 \tag{8}$$

For flow normal to a cylindrical wire the heat and mass transfer coefficients may be approximated by (Treybal, 1968; p 63)

$$N_{Nu} = C_1 + C_2 N_{Re}{}^n N_{Pr}{}^{0.31} \tag{9}$$

$$N_{Sh} = C_1 + C_2 N_{Re}{}^n N_{Sc}{}^{0.31} \tag{10}$$

where C_1, C_2, and n are experimentally determined constants. When the transport coefficients and the velocities in (9) and (10) are expressed as the sum of time-averaged and fluctuating quantities, and the result is substituted into (6) the following equation is obtained

$$\frac{2I'}{\langle I\rangle} = \frac{2E'}{\langle E\rangle} = \alpha f n \frac{U'}{\langle U\rangle} - m\frac{C'}{\langle C\rangle} \tag{11}$$

where

$$f = \frac{C_2\langle N_{Re}\rangle^n N_{Pr}{}^{0.31}}{\langle N_{Nu}\rangle} \tag{12}$$

$$\alpha = \left[1 + m\left(1 - \frac{\langle N_{Nu}\rangle}{\langle N_{Sh}\rangle} N_{Le}{}^{0.31}\right)\right] \tag{13}$$

Squaring and time averaging (11) yields

$$\left\langle\left(\frac{2E'}{\langle E\rangle}\right)^2\right\rangle = \left\langle\left(\alpha f n \frac{U'}{\langle U\rangle}\right)^2\right\rangle - 2\alpha m n f \frac{\langle U'C'\rangle}{\langle U\rangle\langle C\rangle} + m^2\left\langle\left(\frac{C'}{\langle C\rangle}\right)^2\right\rangle \tag{14}$$

It follows from eq 14 that by operating the constant-temperature anemometer at three different wire temperatures (values of m) the resulting simultaneous equations can in principle be solved to determine $U'_{rms}/\langle U\rangle$, $\langle U'C'\rangle/(\langle U\rangle\langle C\rangle)$ and $C'_{rms}/\langle C\rangle$, where the subscript rms (root-mean-square) refers to the square root of the time averaged square values that appear in eq 14.

Equation 14 can be rewritten as

$$\frac{2E'_{rms}}{m\langle E\rangle} = \frac{C'_{rms}}{\langle C\rangle}\sqrt{1 + B^2 - 2BR_{UC}} = \frac{C'_{rms}}{\langle C\rangle}F \tag{15}$$

where

$$B = \frac{\alpha f n U'_{rms}/\langle U\rangle}{m C'_{rms}/\langle C\rangle} \tag{16}$$

$$R_{UC} = \frac{\langle U'C'\rangle}{U'_{rms}C'_{rms}} \tag{17}$$

The correlation coefficient R_{UC} is bounded between −1 and +1. Consequently, for $B < 1$

$$1 - B \le F \le 1 + B \tag{18}$$

This suggests that if B can be made sufficiently small, the parameter F in eq 15 may be taken as unity and $C'_{rms}/\langle C\rangle$ can then be determined from a single measurement of $E'_{rms}/\langle E\rangle$. This condition can be met if the system is operated with a large value of m.

Note that the catalytic probe technique described above should not be used for mixtures in which the average reactant concentration exceeds the lower explosive limit (LEL). On the other hand, the limiting reactant concentration should not be lower than the extinction concentration below which the reaction cannot be sustained on the sensor without electric heating. Typical values of these bounds for several mixtures of reactants in air are reported in Table 1.

Table I. Extinction Concentrations and Lower Explosive Limits for Various Reactants in Air

Reactant	Extinction concn, mol %	L.E.C.,[a] mol %
Ammonia	3.9	16
Methane	1.5	5
Butane	0.8	1.9
Hydrogen	1.3	4

[a] Values from Steere (1967).

Experimental System and Procedure

A catalytic wire probe similar to that described by Edwards et al. (1974, Figure 3) was used to measure the point concentration fluctuations. The sensor consisted of a 6 mm long by 0.025 mm diameter platinum wire (l/d = 240) supported by two 3.2-mm diameter brass rods. For the aspect ratio used it was assumed that the conduction end effects were small, in accordance with the normally accepted practice in hot wire anemometry.

Constant wire temperatures were attained by connecting the probe to a Thermo Systems Model 1010A constant-temperature anemometer. This unit provided a convenient method for maintaining desired values of the overheat parameter, defined as

$$m = (R^* - R_g)/(R - R^*) \tag{7}$$

or

$$m = (R - \Delta R - R_g)/\Delta R \tag{19}$$

where

$$\Delta R \triangleq \text{overheat resistance} = (R - R^*) \tag{20}$$

The instantaneous anemometer bridge voltage was conditioned by filtering out all frequencies above a predetermined cutoff frequency of f_c with a Krohn–Hite Model 3750 low-pass filter and then recorded on magnetic tape using a Hewlett-Packard Model 3960 FM recorder. The signal was subsequently processed by digitation at a sampling rate of $2f_c$ to obtain $\langle E\rangle$ and E'_{rms}. A calibration of $\langle E\rangle$ vs. ΔR was utilized in interpreting the tape recorder signals.

While the above procedure was found to be the most accurate, reasonably good values could also be obtained by direct measurements without tape recording. In this case the instantaneous bridge voltage was passed through a dc offset unit (two operational amplifiers in series) to remove the dc component of the original signal. The fluctuating quantity was then passed through a band-pass filter and finally connected to a Thermo Systems Inc. Model 1060 RMS voltmeter. This procedure enabled a simultaneous measurement of both $\langle E\rangle$ and E'_{rms} and these values agreed to within 6% of those determined by processing of the tape recorded data.

To determine the maximum error in the concentration fluctuations prediction of eq 15 (single point procedure), it is necessary to estimate values of the parameter B via eq 16. To do this, an independent measurement of the turbulence intensity, $U'_{rms}/\langle U\rangle$, was made using the constant-temperature anemometer unit and a Thermo Systems hot film sensor. In practice this measurement can be made with the catalytic wire by operating at a sensor temperature below the ignition temperature or with a nonreactive mixture.

Initial experiments indicated that the electrical resistance of the platinum sensor could increase up to 10% after several hours of operation due to a progressive roughening of its surface. To eliminate this effect, the wire was first pretreated

p.344

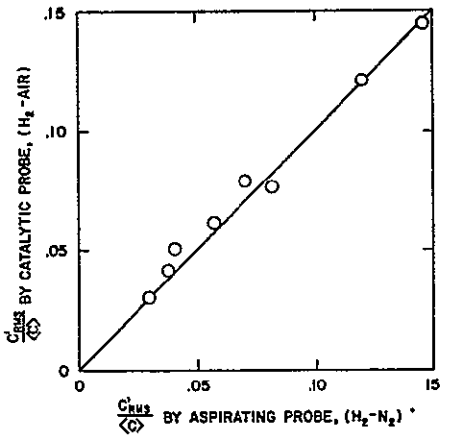


Figure 1. Comparison of $C'_{RMS}/\langle C \rangle$ measurements by the catalytic probe with those made by the aspirating probe. (T_g = 24 °C, channel Reynolds number = 4000).

(activated) for a period of 8–10 h in a stream of 3%v hydrogen in air. During this period the resistance reached an asymptotic value and did not change for several weeks.

Experimental Results

In order to test the validity of the technique and its accuracy, measurements of concentration fluctuations were made in a mixture of hydrogen and air. These results were compared with those for the nonreactive hydrogen in nitrogen mixture. The latter measurements were carried out with a Thermo Systems Model 1441 aspirating probe in conjunction with a Thermo Systems Model 1010A constant-temperature anemometer. The aspirating probe has been described by Blackshear and Fingerson (1962) and a discussion of its applicability was presented by Edwards (1973).

The measurements were carried out in the rectangular (25 × 152 mm) flow channel described by Edwards et al. (1974). A wide range of hydrogen concentration intensities, $C'_{rms}/\langle C \rangle$, were obtainable by using eight different angles of hydrogen injection relative to the main (air or nitrogen) stream, as shown by Edwards et al. (1974). The hydrogen was injected into the main stream (N_{Re} = 4000, T_g = 24 °C) at a rate corresponding to a final average mixture concentration of 2.5%v for all experiments. After measuring the hydrogen in nitrogen concentration fluctuations, a stream of air was substituted for the identical flow conditions (N_{Re} = 4000, T_g = 24 °C), and the aspirating probe was replaced by the catalytic wire probe. The chemical reaction was then initiated on the catalytic sensor and a series of measurements were made at each of the injector angles used above. In all the experiments the signal was conditioned by filtering out all frequencies above 50 Hz. Values of the hydrogen concentration fluctuation intensity were calculated using eq 15 with F = 1. The results of these two sets of measurements are summarized in Figure 1, and they indicate that there was good agreement between the two independent measuring techniques.

Independent measurements of the turbulence intensity at the catalytic wire station with no hydrogen injection showed that $U'_{rms}/\langle U \rangle$ = 0.11, and that values of f = 0.31 and n = 0.5 could be employed for estimating the error term B, eq 16. The factor α was calculated from eq 13 using C_1 = 0.43 and C_2 = 0.532 (Treybal, 1968), in conjunction with the experimentally chosen values of m. When the values of B were calculated for each experiment it was found that the largest value of B was 0.12, and the average of B for all experiments was 0.07. Thus the average error in $C'_{rms}/\langle C \rangle$ due to the assumption that F = 1, was of the order of 7%.

The experiments indicate that the proposed technique is suitable for a rapid evaluation of mixing characteristics of gaseous streams. Its applicability to a variety of systems, such as mixtures of hydrogen, ammonia, and hydrocarbon species in air, should make it useful for evaluation of industrial mixers and in scale-up of reactors in which fast chemical reactions occur.

Nomenclature

A = wire cross-sectional area, cm^2
B = parameter defined by eq 16
C = reactant concentration in fluid, g-mol/cm^3
C_p = heat capacity, cal/g K
d = wire diameter, cm
D_{AB} = binary diffusion coefficient, cm^2/s
E = bridge voltage, V
f = parameter defined by eq 12
F = parameter defined by eq 15
h = heat transfer coefficient, cal/s cm^2 K
ΔH = heat of reaction, cal/g-mol
I = wire current, A
J = conversion factor, cal/s W
k = thermal conductivity, cal/s cm K
k_c = mass transfer coefficient, cm/s
m = parameter defined by eq 7
n = empirical velocity exponent in eq 9
N_{Nu} = Nusselt number, hd/k_f
N_{Pr} = Prandtl number, $C_{pf}\nu_f/k_f$
N_{Re} = Reynolds number, Ud/ν_f
N_{Sc} = Schmidt number, ν_f/D_{AB}
N_{Sh} = Sherwood number, $k_c d/D_{AB}$
P = wire perimeter, cm
R = wire resistance, ohms
R_{UC} = correlation coefficient (eq 17)
t = time, s
T = temperature, K
U = gas velocity, cm/s
x = distance along the wire, cm

ORIGINAL PAGE IS OF POOR QUALITY

Greek Letters

α = parameter defined by eq 13
μ = viscosity, g/cm s
ν = kinematic viscosity, cm^2/s
ρ = wire density, g/cm^3

Subscripts

f = fluid
g = gas phase
rms = root-mean-square
w = wire

Superscripts

′ = fluctuating component
* = with reaction only

Literature Cited

Blackshear, P. L., Fingerson, L., *ARS J.*, 1709 (1962).
Corrsin, S., NACA Technical Note 1864 (1949).
Edwards, W. M., Zuniga-Chaves, J. E., Worley, F. L., Luss, D., *AIChE J.*, **20**, 571 (1974).
Edwards, W. M., Ph D. Thesis, University of Houston, 1973.
Hinze, J. O., "Turbulence", McGraw-Hill, New York, N.Y., 1959.
McQuaid, J., Wright, W., *Int. J. Heat Mass Transfer*, **16**, 819 (1973).
Steere, N. V., Ed., "Handbook of Laboratory Safety", The Chemical Rubber Co., Cleveland, Ohio, 1967.
Treybal, R. E., "Mass Transfer Operations", McGraw-Hill, New York, N.Y., 1968.
Wilson, L. N., Prosser, D. W., Proceedings of Symposium on Turbulence Measurements in Liquids, 1971.

Received for review October 10, 1975
Accepted May 24, 1976

The financial support of the National Science Foundation through grant GK 18226 is gratefully acknowledged.

p. 345

Project

(A) Project Title: Forced Longitudinal Vibration of Prismatic Slender Bars under the Influence of Quadratic and Equivalent Viscous Damping

(B) Project Abstract:

Forced longitudinal vibration of a prismatic, slender bar with quadratic skin damping and equivalent viscous damping are analyzed. The differential equation of motion with quadratic damping is solved numerically by the method of characteristics. The bar is assumed to be free of stress at one end and subjected to sinusoidal motion at the other. Frequency-response curves are plotted for several values of the nonlinear damping coefficient. The equivalent linear system is analyzed by Laplace Transforms and the results are compared with those for the nonlinear system. The damping is linearized by equating the energies dissipated per cycle by quadratic and viscous damping.

(C) Publication: M.S. Thesis in Mechanical Engineering

(D) Year: 1970

(E) Department: Mechanical Engineering

(F) Student Name: A.D. Wilkinson, Jr.

(G) Faculty Advisor: Prof. C.D. Michalopoulos

Project

(A) Project Title: Axisymmetric Contact Problems Involving the Half-Space and the Elastic Layer with a Circular Cylindrical Hole

(B) Project Abstract:

The present investigation is an analytical study of axisymmetric contact problems involving the half-space and the elastic layer with a transverse circular cylindrical hole. A bolt-shaped, rigid punch is pressed into the elastic medium under the action of an axial load. Extended Hankel transforms are employed to transform the Navier differential equations of equilibrium in cylindrical coordinates with axial symmetry.

Two sets of dual integration equations are derived: One for the half-space and one for the layer. Both are shown to reduce to a singular Fredholm integral equation of the first kind with symmetric kernel. The integral equation for the half-space is solved numerically and results for stresses and displacements are presented in graphical form.

(C) Publication: Ph.D. Dissertation in Mechanical Engineering

(D) Year: 1971

(E) Department: Department of Mechanical Engineering

(F) Student Name: Solon C. Parlas

(G) Faculty Advisor: Professor C. D. Michalopoulos

Project

(A) Project Title: The Effect of Tension on the Dynamic Behavior of Eccentric Shafts Rotating in Fluid Medium

(B) Project Abstract:

Using Euler-Bernoulli beam theory an investigation of the dynamic behavior of an eccentric rotating shaft, subject to linearly varying or constant tension, was made. The shaft has distributed mass and elasticity and is suspended in a fluid. Initial lack of straightness was also included in the analysis. The local mass eccentricity is assumed to be a deterministic function of the axial coordinate.

For the variable-tension case the response was determined for a vertical shaft simply supported at the top and vertically guided at the bottom. The constant-tension case was analyzed for a shaft simply supported at its ends. The solution was obtained using modal analysis. It is in series form and is expressed in terms of characteristic functions of the free vibration shaft.

External damping was linearized by equating the energy dissipated per revolution by quadratic and equivalent viscous dampling.

Displacements and stresses were computed along the shaft at a specific speed of rotation. Also maximum stress

ORIGINAL PAGE IS
OF POOR QUALITY

and displacement were computed for speeds in the neighborhood of a natural frequency. Results are given in graphical form for several values of the tension and different eccentricity functions.

(C) Publication: Ph.D. Dissertation in Mechanical Engineerig

(D) Year: 1972

(E) Department: Mechanical Engineering

(F) Student Name: Victor Prodonoff

(G) Faculty Advisor: Prof. C.D. Michalopoulos

Project

(A) Project Title: The Effect of Transverse Shear Deformation on the Large-Amplitude Free Vibrations of Plates and Cylindrical Panels

(B) Project Abstract:

In the present investigation, the equations of nonlinear flexural vibrations of plates and cylindrical panels are derived with the effects of rotatory inertia and transverse shear deformation taken into account. The Galerkin technique is used to solve the nonlinear differential equations. The one-term approximate solutions of the free vibrations of a rectangular plate and cylindrical panel is found. With the assumption that the in-plane inertia and rotatory inertia effects are negligible, it is found that the transverse shear deformation has greater influence on the free vibrations of transversely isotropic plates with large $\frac{E}{Ge}$ ratio (such as pyrolytic graphite material) than on isotropic plates. The same is true for a cylindrical panel.

(C) Publication: Ph.D. Dissertation in Mechanical Engineering

(D) Year: 1972

(E) Department: Mechanical Engineering

(F) Student Name: Chao-Hwang Su

(G) Faculty Advisor: Prof. C.D. Michalopoulos

Project

(A) Project Title: A Hybrid Computer Study of the Dynamics of a Tubular Chemical Reactor

(B) Project Abstract:

The stable hybrid computer solution of a time-dependent tubular chemical reactor represented by a system of parabolic or elliptic-parabolic partial differential equations is studied. In the classical approach to the serial hybrid solution of the one-dimensional diffusion equation using the continuous-space-discrete-time (CSDT) technique, there exists an undesirably large amount of positive analog loop feedback. This makes the classical hybrid method highly unstable in the study of higher frequency transient behavior.

The serial decomposition method used in this study replaces the linear second order differential operator by two stable first order operators integrating in opposite directions and yields one-pass solutions instead of the usual iterative solutions. Thus, considerable computation economy can be expected.

Application of the serial decomposition method to the two-space dimension problem is also possible. The results of the analysis of tubular reactor dynamics in two space dimensions using the continuous-space-discrete-space-discrete-time (CSDSDT) approach not only support the findings of the

previous digital computer steady-state simulation of a homogeneous turbulent flow gas phase $SOCl_2$ decomposition problem, but also give insight to the transient behavior which is not so easy to obtain otherwise.

Dynamic studies of chemical processes appears promising with the hybrid decomposition method. Special interest may be in the area of process sensitivity and stability analysis.

The difficulties and major errors involved in hybrid computation of partial differential equations are also discussed.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1970

(E) Department: Chemical Engineering

(F) Student Name: Hong-Mou Lee

(G) Faculty Advisor: Professor R. L. Motard

Project

(A) Project Title: Information Structures and Order Concepts for Chemical Process Computations

(B) Project Abstract:

A study was made to determine how to carry out chemical process computations on a digital computer in a more effective manner. The concepts and algorithms developed during this study were then used in the development of a more powerful and more efficient chemical process simulation system.

More effective use of computer memory was desired, along with a minimal sacrifice in computational efficiency. This was achieved by storing process data on secondary storage devices and retrieving this data only as needed. The retrieval and storage function was done simultaneously while computations were proceeding. This feature minimizes the computational delays normally associated with the retrieval of data from the comparatively slow secondary storage devices.

An algorithm was developed for ordering process computations in such a way to minimize the number of recycle parameters. There is strong evidence to indicate that computational efficiency can be significantly improved by ordering process computations in this manner. A different approach was used in devising this algorithm that that used

by previous investigators. This approach is intuitively similar to the signal flow diagram concept used in control system theory. The algorithm which resulted from this study is effective and conceptually much simpler than other existing algorithms.

Lastly, a method was proposed to permit a user of a process simulator to vary design parameters during the actual execution of a simulation. This will permit a good design engineer limited interaction with the simulator and will be of considerable value.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1970

(E) Department: Chemical Engineering

(F) Student Name: Rich Walter Barkley

(F) Faculty Advisor: Professor R. L. Motard

Project

(A) Project Title: Simulation of Photochemical Smog

(B) Project Abstract:

A hybrid computer simulation study of the Eschenroeder and Martinez (General Research Corporation, March 1971) photochemical smog model generated more accurate prediction of ground level pollutant concentrations. Advantage was taken of the speed of computation of the analog-digital system to adjust the photochemical kinetic constants in real time within the diffusion-kinetic model.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1973

(E) Department: Chemical Engineering

(F) Student Name: Koki Goto

(G) Faculty Advisor: Professor R. L. Motard

SIMULATION OF PHOTOCHEMICAL SMOG

Koki Goto*

and

Rodolphe L. Motard†

Department of Chemical Engineering
University of Houston
Houston, Texas 77004

* Present address: Far East Mercantile Co., Ltd.
2-1, 2-Chome, Otemachi
Chiyoda-ku, Tokyo, Japan

† Author to whom correspondence should be addressed

p.356

ABSTRACT

A hybrid computer simulation study of the Eschenroeder and Martinez (General Research Corporation, March 1971) photochemical smog model generated more accurate prediction of ground level pollutant concentrations. Advantage was taken of the speed of computation of the analog-digital system to adjust the photochemical kinetic constants in real time within the diffusion-kinetic model.

Table 1

Photochemical Reaction Mechanism for Propylene Kinetic Rate Coefficients (ppm^{-1} min^{-1})

	Reaction Step	From Diffusion Model (1)	Laboratory
+1.	$NO_2 + (h\nu) \rightarrow NO + O$	.4*	.4*
**2.	$O + O_2 + (M) \rightarrow O_3 + (M)$	1.32×10^{-5}	1.32×10^{-5}
3.	$O_3 + NO \rightarrow NO_2 + O_2$	40	22 - 44
4.	$O + HC \rightarrow 2RO_2$	6100	6100
5.	$OH + HC \rightarrow 2RO_2$	80	244
6.	$RO_2 + NO \rightarrow NO_2 + .5OH$	1500	122
7.	$RO_2 + NO_2 \rightarrow (PAN)$	6	122
**8.	$OH + NO + (M) \rightarrow HNO_2 + (M)$	10	99
**9.	$OH + NO_2 + (M) \rightarrow HNO_3 + (M)$	30	300
10.	$O_3 + HC \rightarrow RO_2$	.0125	.0093 - .125
≠11.	$NO + NO_2 + H_2O \rightarrow 2HNO_2$	.01	
+12.	$HNO_2 + (h\nu) \rightarrow NO + OH$	.001	

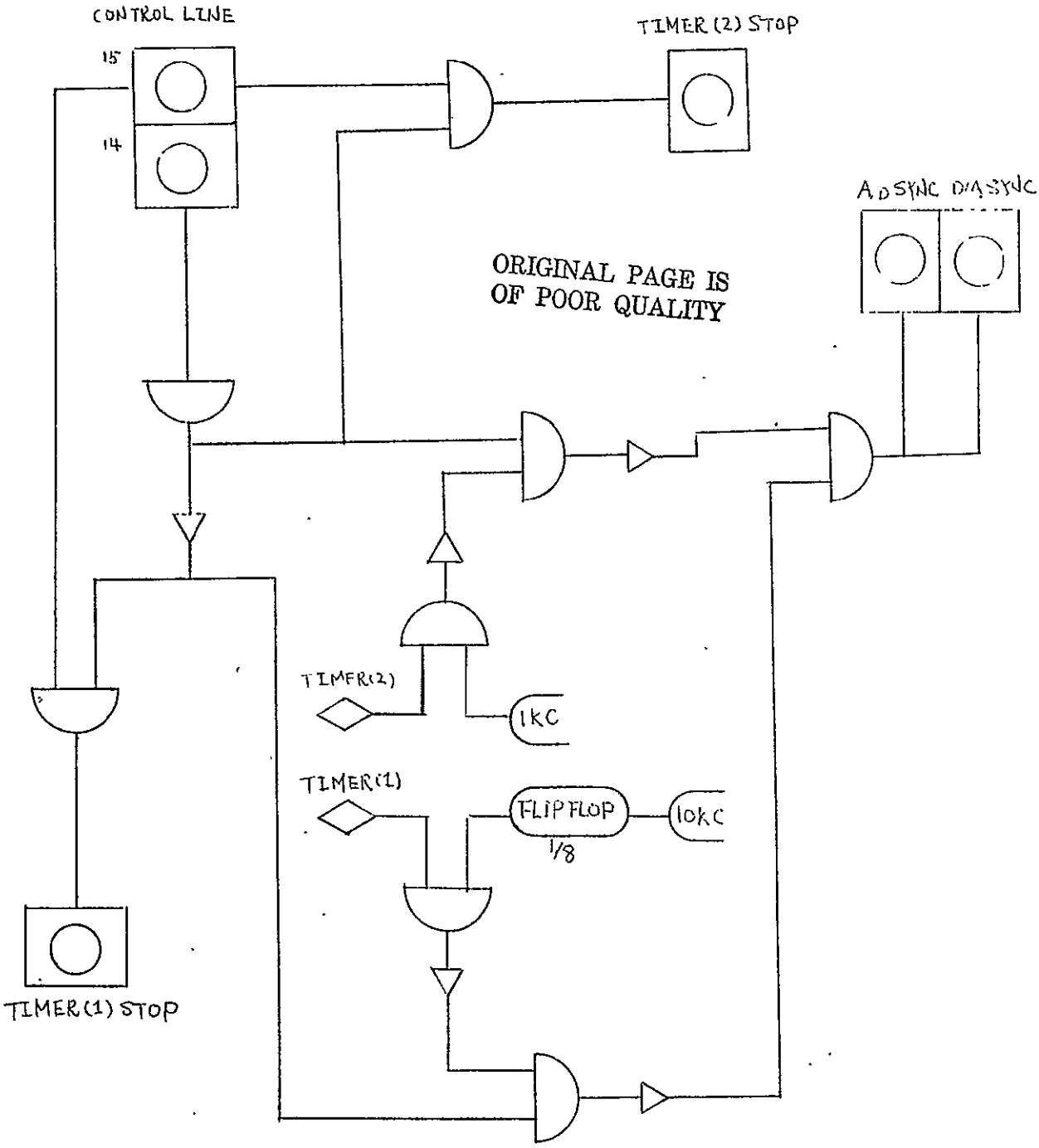
* min^{-1}

\+ Incident light effect lumped into first order rate coefficient

** Third body concentration lumped into second order rate coefficient

≠ Water vapor concentration lumped into second order rate coefficient

Control Line (14) Logic True ---- Forward Integration
Logic False ----Backward Integration

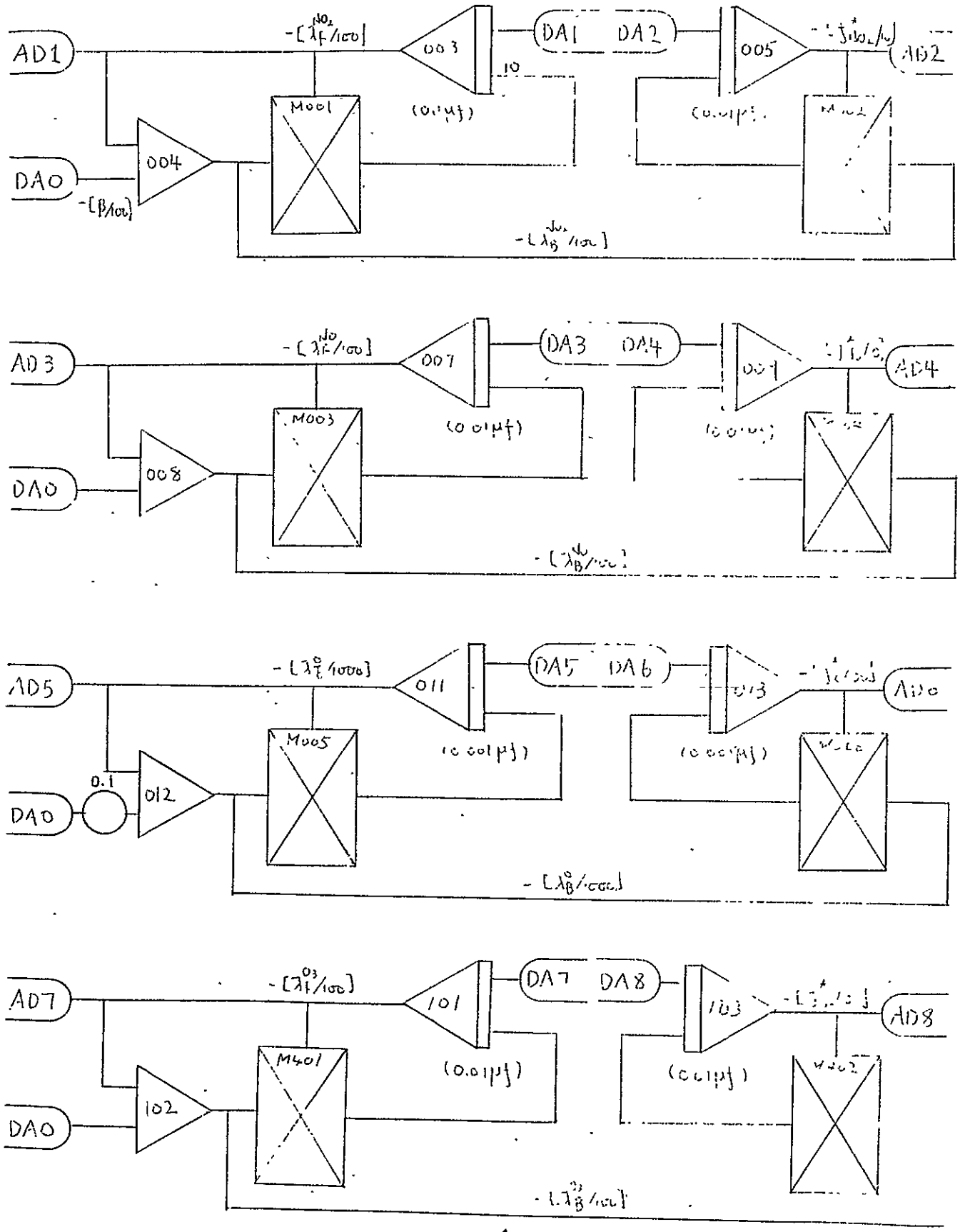


(Timer (1) ⟶ 2msec IC, 1 sec OP
Timer (2) ⟶ 2msec IC, 100 msec OP)

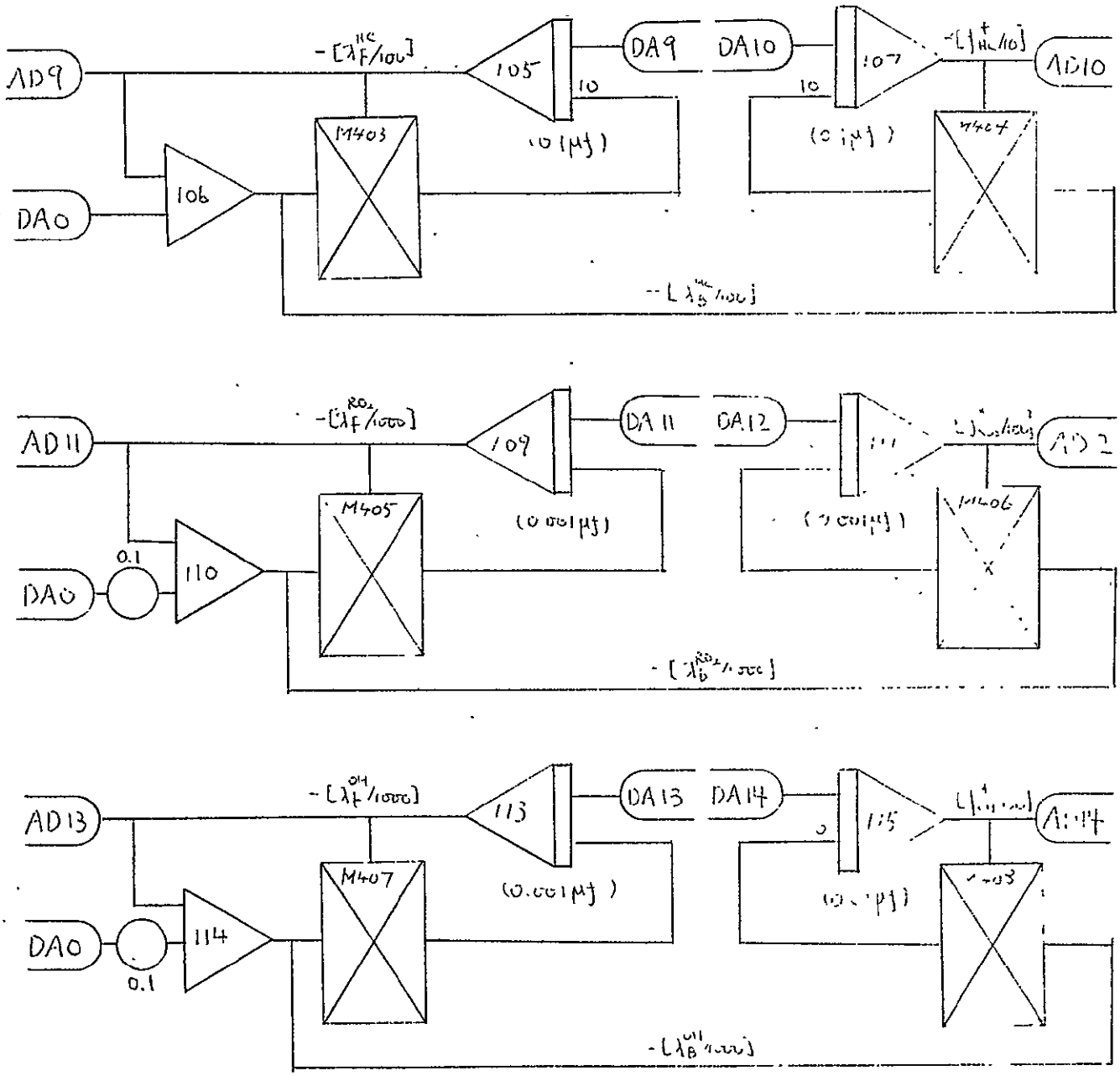
Backward Integration (f^* and λ_F) ---- Timer(1), 1KC Sample rate
Forward Integration (f) ------Timer(2), 1.25KC Sample rate.

LOGIC BOARD DIAGRAM

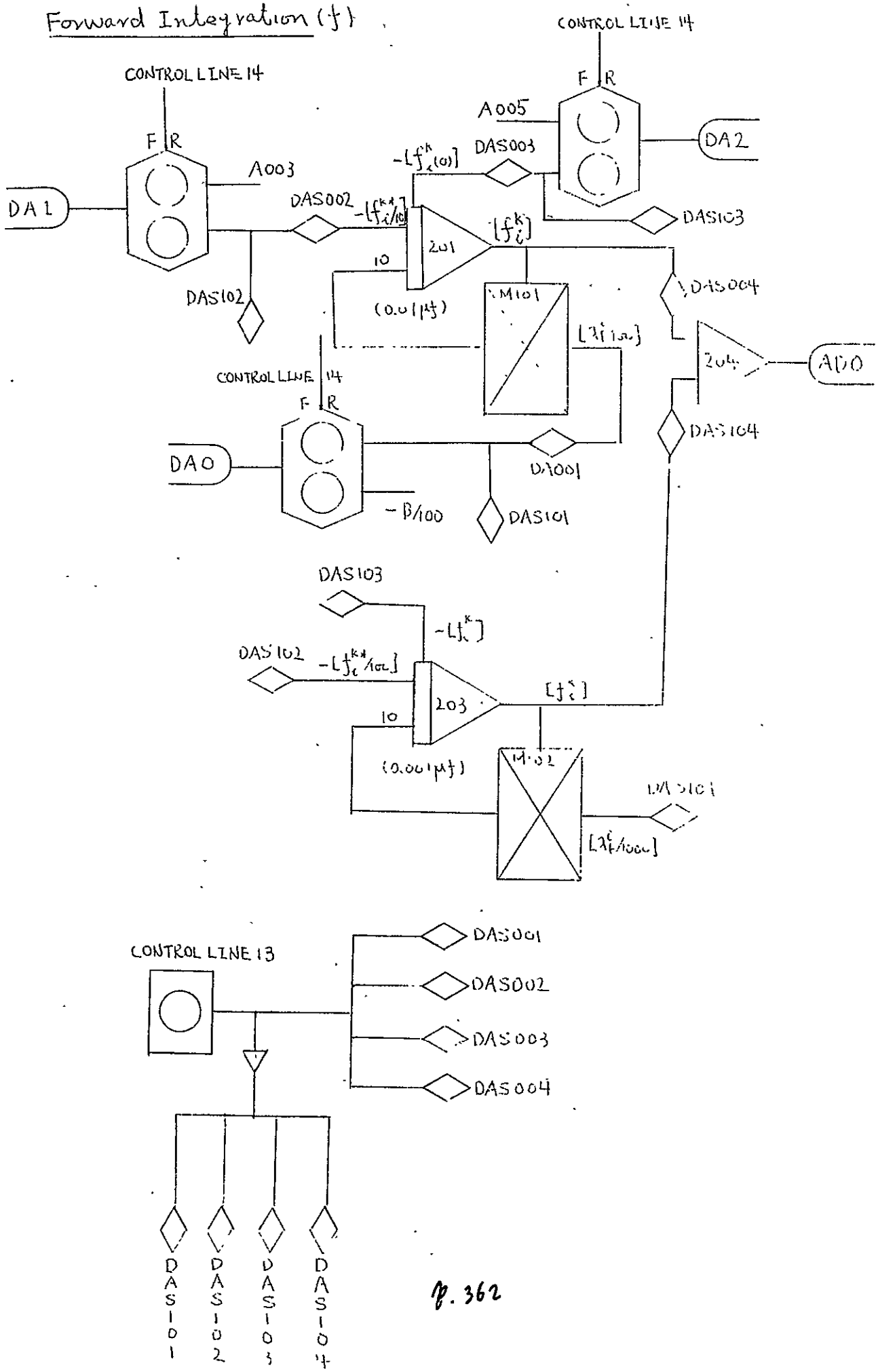
Backward Integration (f^* and λ_F)



ORIGINAL PAGE IS
OF POOR QUALITY



Forward Integration (f)



p. 362

[APPENDIX C]

ORIGINAL PAGE IS
OF POOR QUALITY

Digital Computer Program

MAIN	PSDS
SUBROUTINE	DAADCC
	RUN
	LOCDAC
	EXC
	PRINT
	STCHCK
	UHPLOT
	XMODE
	XSTCTL
	XRSCTL
	XRDAN
	XWRTDA
BLOCK	DATA

```
C **** A HYBRID SIMULATION FOR PHOTOCHEMICAL SMOG DIFFUSION SYSTEM **** PSDS
C                 ------- ( LOS ANGELES BASIN ) -------                 PSDS
C & THE LAGRANGIAN SMOG-DIFFUSION EQUATIONS WITH PHOTOCHEMICAL REACTION PSDS
C     (DC#/DZ)=(E(Z)*C#')'+R(C#)                                        PSDS
C  < C# ; CONCENTRATION FOR EACH SPECIES #      Z ; HEIGHT >            PSDS
C  <  ' ; (D/DT)     T ; TIME     R(C#) ; REACTION TERM >               PSDS
C  < E(Z) ; DIFFUSIVITY >                                               PSDS
C & SPECIES                                                             PSDS
C     11 (7) COMPONENTS                                                 PSDS
C ---> NO2  NO  O  O3  HC(HYDROCARBON)  RO2(PEROXYACYL)  OH             PSDS
C ---> PAN(PEROXYACYL NITRATES)  <HNO2>  <HNO3>  <O2>                   PSDS
C & REACTION EQUATIONS  --- K (PPM-MIN. UNIT) ---                       PSDS
C  ( 1) NO2 + (LIGHT) --> NO + O          K1 = .6                       PSDS
C  ( 2) O + O2 + (M) --> O3 + (M)         K2 = .00264                   PSDS
C  ( 3) O3 + NO --> NO2 + O2             K3 = 40                       PSDS
C  ( 4) O + HC --> 2RO2                  K4 = 305                      PSDS
C  ( 5) OH + HC --> 2RO2                 K5 = 4                        PSDS
C  ( 6) RO2 + NO --> NO2 + .5OH          K6 = 1500                     PSDS
C  ( 7) RO2 + NO2 --> PAN                K7 = 6                        PSDS
C  ( 8) OH + NO --> HNO2                 K8 = 10                       PSDS
C  ( 9) OH + NO2 --> HNO3                K9 = 30                       PSDS
C  (10) O3 + HC --> RO2                  K10= .0125                    PSDS
C  (11) NO + NO2 + (H2O) --> 2HNO2       K11= .01                      PSDS
C //// CONDITION                                                       PSDS
C  DIFFUSIVITY E(Z) = 40 + 3.34Z (M**2/MIN.) .... *                    PSDS
C     AT AVERAGE WIND SPEEDS  U= 1 (METER/SEC.)                         PSDS
C                          * A.Q.ESCHENROEDER, G.R.C., NOV. 1971        PSDS
C  & AUTOMOTIVE EMMISIONS DATA .... SYSTEMS APPLICATIONS, INC., DA.C. 1971
C  < Q# ; AUTOMOBILE EMISSIONS FACTORS FOR EACH COMPONENT # >           PSDS
C  < M(D,N,T)=D(L)(N(1)*T(1)+N(2)*T(2)+.....+N(S)*T(S)) >               PSDS
C   EMISSIONS Q@(GRAMS/HR)=Q#(GRAMS/VEHICLE)*M(D,N,T)(VEHICLE MILES/HR) PSDS
C  < N ; VEHICLES PER DAY(GIVEN AS TRAFFIC COUNTS AT A POINT,           PSDS
C                          ASSIGNABLE TO A SEGMENT OF ROAD) >           PSDS
C  < D(L) ; FRACTION OF DAILY TRAFFIC COUNT ASSIGNABLE TO HOURLY (L) >  PSDS
C  < T ; MILES OF ROAD SEGMENT TO WHICH COUNT (N) IS ASSIGNED >         PSDS
C  < S ; NUMBER OF ROAD SEGMENTS CONTAINED IN EACH DIVIDED SECTION >    PSDS
C   .... Q@ IS ESTIMATED FOR SURFACE STREETS AND FREEWAYS               PSDS
C*********************************************************************PSDS
      COMMON/DAC/F1,F2,ALPH,KHR                                         PSDS
      COMMON/SM/SCAL,MAX1,DA                                            PSDS
      COMMON/PS/IVAL,NPOT,KVAL                                          PSDS
      COMMON/PRN/LOCA,NAME                                              PSDS
      COMMON/QQ/QHC,QNO2,QNO,DTIME,KMIN,TIME,KTIME,MIN,TLND,VIC,VFIN    PSDS
      COMMON/FO/FONO2,FONO,FOO,FOO3, FOHC,FORO2,FOOH                    PSDS
      REAL*8 MODCCW(2),MODRCB(4),STPCCW(2),STPRCB(4)                    PSDS
      REAL*8 PTSCCW(2),PTSRCB(4),DACCW(2),DARCB(4),ADCCW(2),ADRCB(4)    PSDS
      INTEGER*2 LOCP(28),LOCIC(3),IVAL(28),MIC/16/,KVAL(28)             PSDS
      INTEGER*2 LOCA(1100),LOCB(1100),LOCC(150),LOCD(150),ISHDR(11 00)  PSDS
      INTEGER*4 IHOLD,NPOT(28)                                          PSDS
```

```
      INTEGER NO(20),VERT(5),HORZ(5),NAME(7),ISCL(7)                    PSDS  5
      REAL F2(7,32),F1(7,32),G(7,32),Y(32)                              PSDS  5
      REAL KHR(11),KMIN(11),SCAL(7),DA(14),VIC(4),VFIN(4)               PSDS  5
      REAL LAMF(7,32),FSTAR(7,32)                                       PSDS  5
      REAL QHC(12),QNO2(12),QNO(12),ALPH(32),BETA(32),FOOH(32)          PSDS  5
      REAL FONO2(32),FONO(32),FOO(32),FOO3(32),FOHC(32),FORO2(32)       PSDS  5
      REAL A(10,9)                                                      PSDS  5
      DATA HORZ/' PPM',' ',' ',' ',' '/                                 PSDS  5
      DATA VERT/' OF ','THE ','CLOC','K ',' '/                          PSDS  5
      DATA NO/'PHOT','OCHE','MICA','L SM','OG H','ISTO','RY A','LON',   PSDS  5
     1' 11:','30-E','L MO','NTE ','( 1-','NO2,',' 3-N','O, 5','-O3,',   PSDS  6
     2' 7-H','C )',' '/                                                 PSDS  6
C                                                                       PSDS  6
C     ID NUMBER OF EACH SPECIES                                         PSDS  6
C     1...NO2, 2...NO, 3...O, 4...O3, 5...HC, 6...RO2, 7...OH           PSDS  6
C                                                                       PSDS  6
C     NO(9)-(12) MAY BE CHANGED FOR EACH WIND TRAJECTORY                PSDS  6
C                                                                       PSDS  6
      DO 303 I=1,7                                                      PSDS  6
  303 ISCL(I)=SCAL(I)                                                   PSDS  6
      N=TEND/DTIME                                                      PSDS  7
CC                                                                      PSDS  7
C     A(I,J) MATRIX FOR PLOT SUBROUTINE                                 PSDS  7
C     DIMENSION I=N2, J=9                                               PSDS  7
C                                                                       PSDS  7
      N1=N+1                                                            PSDS  7
      N2=N+2                                                            PSDS  7
      DO 125 I=1,9                                                      PSDS  7
      DO 125 J=1,N2                                                     PSDS  7
  125 A(J,I)=0.0                                                        PSDS  7
      A(1,1)=TIME-TEND                                                  PSDS  8
      A(1,2)=VIC(1)                                                     PSDS  8
      A(1,4)=VIC(2)                                                     PSDS  8
      A(1,6)=VIC(3)                                                     PSDS  8
      A(1,8)=VIC(4)                                                     PSDS  8
      A(N1,3)=VFIN(1)                                                   PSDS  8
      A(N1,5)=VFIN(2)                                                   PSDS  8
      A(N1,7)=VFIN(3)                                                   PSDS  8
      A(N1,9)=VFIN(4)                                                   PSDS  8
C                                                                       PSDS  8
C     INITIAL CONDITION SET FOR EACH SPECIES                            PSDS  9
C                                                                       PSDS  9
      FONO2(1)=VIC(1)*SCAL(1)                                           PSDS  9
      FONO(1)=VIC(2)*SCAL(2)                                            PSDS  9
      FOO3(1)=VIC(3)*SCAL(4)                                            PSDS  9
      FOHC(1)=VIC(4)*SCAL(5)                                            PSDS  9
C                                                                       PSDS  9
C     STATIC CHECK FOR ANALOG BOARD                                     PSDS  9
C                                                                       PSDS  9
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
      WRITE(15,111)
  111 FORMAT('NEED STATIC CHECK ? EOB,NO... 1')
      READ(15,106) K2
      IF(K2.EQ.1) GO TO 41
      CALL STCHCK
C
C     ALPH(Z)=100*100/(5*E(Z))   BETA(Z)=(1/E(Z))*(DE(Z)/DZ)
C
   41 WRITE(6,101) TIME
  101 FORMAT(//5X,'***** PHOTOCHEMICAL SMOG DIFFUSION *****'//'----- (',
     1F5.2,'-EL MONTE TRAJECTORY OF AIR MASS) ------'//)
      MX=MAX1+1
      Z=0.0
      DZ=1.0/(MAX1-1)
      DO 1 I=1,MAX1
      IF(Z.GE. 0.75) GO TO 50
      ALPH(I)=10000./(1800.+20800.*Z)/5.
      BETA(I)= 20800./(1800.+20800.*Z)/10.
      GO TO 51
   50 ALPH(I)=10000./18000./5.
      BETA(I)=0.
   51 Y(I)=100.*Z
    1 Z=Z+DZ
C
C     F1 IS EACH COMPONENT CONCENTRATION OF PREVIOUS STEP
C
      DO 3 I=1,MAX1
      F1(1,I)=FONO2(I)
      F1(2,I)=FONO(I)
      F1(3,I)=FOO(I)
      F1(4,I)=FOO3(I)
      F1(5,I)=FOHC(I)
      F1(6,I)=FORO2(I)
    3 F1(7,I)=FOOH(I)
C
C     KHR(PPM-HOUR UNIT),KMIN(PPM-MIN. UNIT)...REACTION RATE COEFFICIENTS
C
      DO 300 I=1,11
  300 KHR(I)=KMIN(I)*60.
      IHOLD=0
      CALL XSTCTL(IHOLD,15)
C
C     INITIAL CONDITION OF EACH INTEGRATER IS ZERO
C
      WRITE(15,112)
  112 FORMAT('NEED POT SETS ? EOB,NO... 1')
      READ(15,106) K2
      IF(K2.EQ.1) GO TO 42
    6 CALL FRCBSU(PTSRCB,28,PTSCCW)
```

```
      CALL POTSS(PTSCCW,28,NPOT,KVAL,PTSRCB)                            PSDS 14
      CALL EXC(PTSRCB)                                                  PSDS 14
      CALL XRDAN(NPOT,28,LOCP)                                          PSDS 15
      DO 4 I=1,28                                                       PSDS 1
    4 LOCP(I)=LOCP(I)/.8191                                             PSDS 1
      WRITE(6,103) (NPOT(I),LOCP(I),KVAL(I),I=1,28)                     PSDS 1
  103 FORMAT(//20X,'***** POTS. TEST ...ADDRESS = ACTUAL (DISIRED) VALUEPSDS 1
     1 *****'///  (5(3X,A4,'=',I6,' ( ',I4,' ) ')/))                   PSDS 1
      DO 5 I=1,28                                                       PSDS 1
       ERROR=LOCP(I)-KVAL(I)                                            PSDS 1
      IERROR=ABS(ERROR)                                                 PSDS 1
      IF(IERROR.LT.10) GO TO 5                                          PSDS 1
      WRITE(6,104) NPOT(I),IERROR                                       PSDS 1
    5 CONTINUE                                                          PSDS 1
      WRITE(15,105)                                                     PSDS 1
  104 FORMAT(5X,'D(',A4,')...',I5)                                      PSDS 1
  105 FORMAT('WANT TO POTSET AGAIN ? TYPE IN 1 WITH I2')                PSDS 1
      READ(15,106) K1                                                   PSDS 1
  106 FORMAT(I2)                                                        PSDS 1
      IF(K1.EQ.1) GO TO 6                                               PSDS 1
C                                                                       PSDS 1
C     POT SETS ARE OVER                                                 PSDS 1
C                                                                       PSDS 1
      WRITE(6,121) (I,KMIN(I),I=1,11)                                   PSDS 1
  121 FORMAT(//20X,'..... REACTION RATE COEFFICIENTS FOR THIS SYSTEM (PPPSDS 1
     1M-MIN. UNIT) .....'//(10X,'---> ',5('K(',I2,')=',E11.4,5X)/)////) PSDS 1
C                                                                       PSDS 1
C     PRINT OUTS OF INITIAL CONDITIONS FOR THIS SYSTEM                  PSDS 1
C                                                                       PSDS 1
   42 WRITE(6,102) TIME,KTIME,MIN,(NAME(I),I=1,7),(ISCL(I),I=1,7),      PSDS 1
     1(Y(MX-I),(F1(K,MX-I),K=1,7),I=1,MAX1)                             PSDS 1
  102 FORMAT(//23X,'***** PROFILE OF SMOG CONCENTRATION ALONG ',F5.2,   PSDS 1
     1'-EL MONTE IN LOS ANGELES BASIN *****'////'***** AT ',I2,':',I2,' PSDS 1
     2*****'///1X,'HEIGHTS(M)',1X,7(9X,A4,2X),3X//12X,7(9X,'*',I4,1X)// PSDS 1
     3100(2X,F6.2,4X,7(F15.6),3X/)//)                                   PSDS 1
      WRITE(15,109)                                                     PSDS 1
  109 FORMAT('PRINT OUT DA-VALUE ? YES...1')                            PSDS 1
      CALL MODE(MODCCW,MIC)                                             PSDS 1
      CALL FRCBSU(MODRCB,28,MODCCW)                                     PSDS 1
      CALL EXC(MODRCB)                                                  PSDS 1
      READ(15,106) LPR                                                  PSDS 1
      F1(3,1)=0.10                                                      PSDS 1
      F1(7,1)=0.10                                                      PSDS 1
      F1(6,1)=0.10                                                      PSDS 1
       DO 31 K=1,7                                                      PSDS 1
      Z=F1(K,1)                                                         PSDS 1
      DZ=F1(K,1)/(MAX1-1)                                               PSDS 1
      DO 31 I=1,MAX1                                                    PSDS 1
      F1(K,I)=Z                                                         PSDS 1
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
   31 Z=Z-DZ                                                            PSDS 1
      DO 29 K=1,7                                                       PSDS 1
      DO 29 I=1,MAX1                                                    PSDS 1
   29 G(K,I)=F1(K,I)                                                    PSDS 2
       WRITE(6,110) KTIME,MIN,(NAME(I),I=1,7),(ISCL(I),I=1,7),(Y(MX-I), PSDS 2
     1(F1(K,MX-I),K=1,7),I=1,MAX1)                                      PSDS 2
  110 FORMAT(//10X,'/// ASSUMED INITIAL CONDITIONS ///'                 PSDS 2
     1                                        //'***** AT ',I2,':',I2,' PSDS 2
     2*****'///1X,'HEIGHTS(M)',1X,7(9X,A4,2X),3X//12X,/(9X,'*',I4,1X)// PSDS 2
     3100(2X,F6.2,4X,7(F15.6),3X/)//)                                   PSDS 2
      DO 52 I=1,15                                                      PSDS 2
   52 LOCA(I+1)=0.0                                                     PSDS 2
      CALL XWRTDA(16,0,0,15,LOCA)                                       PSDS 2
C                                                                       PSDS 2
C     BETA(Z) IS PREPARED FOR DAC(0)                                    PSDS 2
C                                                                       PSDS 2
      DO 9 I=1,MAX1                                                     PSDS 2
      IB=MAX1+1-I                                                       PSDS 2
      LOCA(30*I-28)=BETA(IB)*8191./10.                                  PSDS 2
    9 LOCA(30*I-13)=LOCA(30*I-28)                                       PSDS 2
C-----------------------------------------------------------------------PSDS 2
C                                                                       PSDS 2
C     SYSTEM STARTS FOR EACH TIME STEP                                  PSDS 2
C                                                                       PSDS 2
      DO 7 M=1,N                                                        PSDS 2
C                                                                       PSDS 2
C     F2 IS ASSUMED CURRENT CONCENTRATION FOR EACH ITERATION PROCESS    PSDS 2
C                                                                       PSDS 2
      DO 8 K=1,7                                                        PSDS 2
      DO 8 I=1,MAX1                                                     PSDS 2
      F1(K,I)=G(K,I)                                                    PSDS 2
    8 F2(K,I)=G(K,I)                                                    PSDS 2
C                                                                       PSDS 2
C     POLLUTANT SOURCE AT GROUND LEVEL                                  PSDS 2
C                                                                       PSDS 2
   27 FQHC=-QHC(M)*SCAL(5)/3.05                                         PSDS 2
      FQNO2=-QNO2(M)*SCAL(1)/3.05                                       PSDS 2
      FQNO=-QNO(M)*SCAL(2)/3.05                                         PSDS 2
      ICOUNT=1                                                          PSDS 2
C* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *PSDS 2
C                                                                       PSDS 2
C     ITERATION PROCESS FOR F STARTS                                    PSDS 2
C                                                                       PSDS 2
   23 CALL XRSCTL(IHOLD,14)                                             PSDS 2
C                                                                       PSDS 2
C     INPUT DATA TO EVERY LAMDA-F AND F* INTEGRATIONS FOR DAC           PSDS 2
C                                                                       PSDS 2
      CALL LOCDAC                                                       PSDS 2
      IF(LPR.NE.1) GO TO 113                                            PSDS 2
```

```
      CALL PRINT(MAX1)
  113 ND=30*MAX1
C
C     ANALOG COMPUTER STARTS FOR INTEGRATIONS OF LAMDA-F AND F* EQUATIONS
C
      CALL DAADCC(DACCW,DARCB,LOCA,ADCCW,ADRCB,LOCB,I4,ND)
      CALL RUN(IHOLD,DARCB,ADRCB)
C
C     ANALOG COMPUTER STOPS
C
C     CHECK OVERLOADS OF SCALED VARIABLES FROM ADC
C
      DO 26 I=1,ND
      ISTOR(I)=-LOCB(I+1)
      IF(ISTOR(I).LE.8191.AND.ISTOR(I).GE.-8191) GO TO 26
       J=I-1
  302 IF(J.LT.15) GO TO 301
      J=J-15
      GO TO 302
  301 STOR=ISTOR(I)*100./8191.
      WRITE(6,500) J,STOR
      IF(ISTOR(I).LT.-8191) ISTOR(I)=-8191
      IF(ISTOR(I).GT. 8191) ISTOR(I)=8191
   26 CONTINUE
  500 FORMAT(/50X,'OVERLOAD AD(',I2,')=',F10.4,'VOLTS',4X,'ASSUMED II + 
     1UR - 100 VOLTS')
C
C     LAMDA-F AND F* ARE CONVERTED FROM ADC VALUE
C
      DO 45 K=1,7
      IF(K.EQ.3.OR.K.EQ.6.OR.K.EQ.7) GO TO 43
      DO 44 I=1,MAX1
      IB=2*K+30*(MAX1-I)
      FSTAR(K,I)=ISTOR(IB+1)*10./8191.
   44 LAMF(K,I)=ISTOR(IB)*100./8191.
      GO TO 45
   43 DO 2 I=1,MAX1
      IB=2* K+30*(MAX1-I)
      FSTAR(K,I)= ISTOR(IB+1)*100./8191.
    2 LAMF(K,I)=ISTOR(IB)*1000./8191.
   45 CONTINUE
  114 CALL XSTCTL(IHOLD,14)
C.......................................................................
C
C     DO LOOP 13 FOR FORWARD INTEGRATION
C
      DO 13 K=1,7
      CALL XRSCTL(IHOLD,12)
      DO 14 I=1,MAX1
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
      IB= 2*K+30*(MAX1-I)
      LOCC(4*I-2)=-ISTOR(IB)
      LOCC(4*I)=-ISTOR(IB)
      LOCC(4*I-1)= ISTOR(IB+1)
   14 LOCC(4*I+1)= ISTOR(IB+1)
      LOCC(4*MAX1)=0
      LOCC(4*MAX1+1)=0
C
C     COMPUTATION OF INITIAL VALUE FOR F-INTEGRATION
C
      FIC= LOCC(3)/(LOCC(2)*10.)
      IF(K.EQ.1) GO TO 15
      IF(K.EQ.2) GO TO 16
      IF(K.EQ.5) GO TO 17
C       IF(K.EQ.4.OR.K.EQ.6.OR.K.EQ.7) CALL XSTCTL(IHOLD,12)
      GO TO 18
   15 FIC=FIC-FQNO2*8191./(LOCC(2)*100.)
      GO TO 18
   16 FIC=FIC-FQNO*8191./(LOCC(2)*100.)
      GO TO 18
   17 FIC=FIC-FQHC*8191./(LOCC(2)*100.)
   18 IF(FIC.GT. 1.0) FIC=1.0
  115 LOCIC(2)=FIC*8191.
      LOCIC(1)=514
      IF(K.EQ.3.OR.K.EQ.6.OR.K.EQ.7) GO TO 46
      CALL XSTCTL(IHOLD,13)
      GO TO 47
   46 CALL XRSCTL(IHOLD,13)
C
C     SET INITIAL CONDITION FOR F-INTEGRATION
C
   47 CALL XWRTDA(1,0,2,2,LOCIC)
      ND=4*MAX1
C
C     ANALOG COMPUTER STARTS FOR F-INTEGRATION
C
      CALL DAADCC(DACCW,DARCB,LOCC,ADCCW,ADRCB,LOCD,1,ND)
      CALL RUN(IHOLD,DARCB,ADRCB)
C
C     ANALOG COMPUTER STOPS
C
      DO 19 I=1,MAX1
   19 G(K,I)= LOCD(4*I-2)/8191.
   13 CONTINUE
C
C     FORWARD INTEGRATION OVER
C
C.....................................................................
C
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
C     CHECK ITERATION CRITERION                                         PSDS 34
C                                                                       PSDS 34
      DO 30 I=1,15                                                      PSDS 34
   30 LOCA(I+1)=0.0                                                     PSDS 34
      CALL XWRTDA(16,0,0,15,LOCA)                                       PSDS 34
      NCOUNT=0                                                          PSDS 34
      DO 200 K=1,7                                                      PSDS 3
      SUM=0.0                                                           PSDS 3
      DMAX=0.0                                                          PSDS 3
      DO 20 I=1,MAX1                                                    PSDS 3
      IF(G(K,I).LE..0001) G(K,I)=0.0001                                 PSDS 3
       DERIV   =ABS(G(K,I)-F2(K,I))/ABS(G(K,I))                         PSDS 3
      SUM=SUM+DERIV                                                     PSDS 3
   20 IF(DERIV.GT.DMAX) DMAX=DERIV                                      PSDS 3
      AVE=SUM*100./FLOAT(MAX1)                                          PSDS 3
      IF(AVE.GT.10.0) NCOUNT=NCOUNT +1                                  PSDS 3
  200 CONTINUE                                                          PSDS 3
      IF(NCOUNT.EQ. 0) GO TO 21                                         PSDS 3
      IF(LPR.NE.1) GO TO 116                                            PSDS 3
      WRITE(6,201) ICOUNT,   (NAME(I),I=1,7),(Y(MX-I),( G(K,MX-I),I=1,7)PSDS 3
     1,I=1,MAX1)                                                        PSDS 3
  201 FORMAT(//23X,'UNDER ITERATION (',I2,'-TIME)'///1X,'HEIGHTS(M)',2X,PSDS 3
     17(9X,A4,2X),3X//100(2X,F6.2,4X,7(F15.6),3X/)//)                   PSDS 3
  116 DO 22 K=1,7                                                       PSDS 3
      DO 22 I=1,MAX1                                                    PSDS 3
   22 F2(K,I)=(G(K,I)+F2(K,I))/2.                                       PSDS 3
      ICOUNT=ICOUNT+1                                                   PSDS 3
      GO TO 23                                                          PSDS 3
C                                                                       PSDS 3
C     ITERATION PROCESS FOR F ENDS                                      PSDS 3
C                                                                       PSDS 3
C* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *PSDS 3
   21 MIN=MIN+60*DTIME                                                  PSDS 3
      IF(MIN.LT.60) GO TO 24                                            PSDS 3
      KTIME=KTIME+1                                                     PSDS 3
      MIN=MIN-60                                                        PSDS 3
   24 DO 25 K=1,7                                                       PSDS 3
      DO 25 I=1,MAX1                                                    PSDS 3
   25 F1(K,I)=G(K,I)                                                    PSDS 3
      A(M+1,1)=FLOAT(KTIME)+FLOAT(MIN)/100.                             PSDS 3
      A(M+1,2)=F1(1,1)/SCAL(1)                                          PSDS 3
      A(M+1,4)=F1(2,1)/SCAL(2)                                          PSDS 3
      A(M+1,6)=F1(4,1)/SCAL(4)                                          PSDS 3
      A(M+1,8)=F1(5,1)/SCAL(5)                                          PSDS 3
      IF(M.EQ.N.AND.LPR.EQ.0) GO TO 118                                 3
      IF(LPR.NE.2.AND.LPR.NE.1) GO TO 7                                 PSDS 3
      WRITE(6,1111) ICOUNT                                              PSDS 3
 1111 FORMAT(//10X,'AFTER ',I2,'-TIME ITERATION'//)                     PSDS 3
  118 WRITE(6,107) KTIME,MIN,(NAME(I),I=1,7),(ISCL(I),I=1,7),(Y(MX-I),  PSDS 3
```

```
     1(F1(K,MX-I),K=1,7),I=1,MAX1)                                      PSDS 3
  107 FORMAT(//'***** AT ',I2,':',I2,' *****'///1X,'HEIGHTS(M)',2X,7(9X,PSDS 3
     1A4,2X),3X//12X,7(9X,'*',I4,1X)//100(2X,F6.2,4X,7(F15.6),3X/)//)   PSDS 3
    7 CONTINUE                                                          PSDS 3
C                                                                       PSDS 3
C     COMPUTATION ENDS FOR EACH TIME STEP                               PSDS 3
C                                                                       PSDS 3
C-----------------------------------------------------------------------PSDS 4
      WRITE(6,108)                                                      PSDS 4
  108 FORMAT(  //T60,'..... AT EL MONTE '//)                            PSDS 4
      CALL UHPLOT(NO,A,N2,9,VERT,HORZ,0)                                PSDS 4
      WRITE(6,119)                                                      PSDS 4
  119 FORMAT(/30X,'CALC.--->',3X,'1--> NO2   3--> NO   5--> O3   7--> HC.PSDS 4
     1'/30X,'OBS. --->',3X,                                             PSDS 4
     2            '2--> NO2   4--> NO   6--> O3   8--> HC'////1X,'*****PSDS 4
     3 END OF SIMULATION ***** ')                                       PSDS 4
      STOP                                                              PSDS 4
      END                                                               PSDS 4
```

```
      SUBROUTINE DAADCC(DACCW,DARCB,LOCA,ADCCW,ADRCB,LOCB,NCON,ND)
      REAL*8 DACCW(1),DARCB(1),ADCCW(1),ADRCB(1)
      INTEGER*2 LOCA(1),LOCB(1)
      LOCA(1)=NCON
      LOCB(1)=NCON
      CALL WRITDA(DACCW, 3,ND,LOCA)
      CALL FRCBSU(DARCB,30,DACCW)
      CALL READAD(ADCCW,ND, 3,LOCB)
      CALL FRCBSU(ADRCB,29,ADCCW)
      RETURN
      END
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
      SUBROUTINE RUN(IHOLD,DA,AD)                                       RUN
      REAL*8 DA(1),AD(1)                                                RUN
      CALL XRSCTL(IHOLD,15)                                             RUN
      CALL FRTIO(DA,IR)                                                 RUN
      CALL FRTIO(AD,IR)                                                 RUN
      CALL FCHECK(AD,IR,1)                                              RUN
      CALL XSTCTL(IHOLD,15)                                             RUN
      RETURN                                                            RUN
      END                                                               RUN
```

```
      SUBROUTINE EXC(RCB)                                          EXC
      REAL*8 RCB(1)                                                EXC
      CALL FRTIO(RCB,IR)                                           EXC
      CALL FCHECK(RCB,IR, 1)                                       EXC
      RETURN                                                       EXC
      END                                                          EXC
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
      SUBROUTINE LOCDAC
      COMMON/DAC/F1,F2,ALPH,KHR
      COMMON/PRN/LOCA,NAME
      COMMON/SM/SCAL,MAX1,DA
      INTEGER NAME(7)
      REAL F1(7,32),F2(7,32),ALPH(32),KHR(11),SCAL(7),DA(14)
      INTEGER*2 LOCA(1100)
      M=MAX1
      DO 1 I=1,M
      IB=M+1-I
      KF1=30*I-28
      KF2=30*I-13
      F=ALPH(IB)*8191.
      LOCA(KF1+1)=-F*(0.1+0.05*(KHR(1)+KHR(7)*F2(6,IB)/SCAL(6)+
     1KHR(9)*F2(7,IB)/SCAL(7)+KHR(11)*F2(2,IB)/SCAL(2)))*DA(1)
      LOCA(KF1+2)=F*(F1(1,IB)+0.5*SCAL(1)*(KHR(3)*F2(4,IB)/SCAL(4)+
     1KHR(6)*F2(6,IB)/SCAL(6))*F2(2,IB)/SCAL(2))*DA(2)
      LOCA(KF1+3)=-F*(0.1+0.05*(KHR(3)*F2(4,IB)/SCAL(4)+KHR(5)*F2(6,IB)
     1/SCAL(6)+KHR(8)*F2(7,IB)/SCAL(7)+KHR(11)*F2(1,IB)/SCAL(1)))*DA(3)
      LOCA(KF1+4)=F*(F1(2,IB)+0.5*SCAL(2)*KHR(1)*F2(1,IB)/SCAL(1))*DA(4)
      LOCA(KF1+5)=-F*(0.01+0.005*(KHR(2)*0.232E6+KHR(4)*F2(3,IB)/SCAL(3)
     1))*DA(5)
      LOCA(KF1+6)=F*(F1(3,IB)*0.1+0.05*SCAL(3)*KHR(1)*F2(1,IB)/SCAL(1))
     1*DA(6)
      LOCA(KF1+7)=-F*(0.1+0.05*(KHR(3)*F2(2,IB)/SCAL(2)+KHR(10)*F2(5,IB)
     1/SCAL(5)))*DA(7)
      LOCA(KF1+8)=F*(F1(4,IB)+0.5*SCAL(4)*KHR(2)*0.232E6*F2(3,IB)/
     1SCAL(3))*DA(8)
      LOCA(KF1+9)=-F*(0.1+0.05*(KHR(4)*F2(3,IB)/SCAL(3)
     1+KHR(5)*F2(7,IB)/SCAL(7)+KHR(10)*F2(4,IB)/SCAL(4)))*DA(9)
      LOCA(KF1+10)=F*F1(5,IB)*DA(10)
      LOCA(KF1+11)=-F*(.01+.005*(KHR(6)*F2(2,IB)/SCAL(2)+KHR(7)*F2(1,IB)
     1/SCAL(1)))*DA(11)
      LOCA(KF1+12)=F*(          0.05*SCAL(6)*(2.0*KHR(4)*F2(3,IB)/SCAL(3)+
     12.0*KHR(5)*F2(7,IB)/SCAL(7)+KHR(10)*F2(4,IB)/SCAL(4))*F2(5,IB)/
     2SCAL(5)+0.1*F1(6,IB))*DA(12)
      LOCA(KF1+13)=-F*(.01+.005*(KHR(5)*F2(5,IB)/SCAL(5)+KHR(6)*F2(2,IB)
     1/SCAL(2)+KHR(9)*F2(1,IB)/SCAL(1)))*DA(13)
      LOCA(KF1+14)=F*(          .05*SCAL(7)*0.5*KHR(6)*F2(6,IB)*F2(2,IB)
     1/SCAL(6)/SCAL(2)+0.1*F1(7,IB))*DA(14)
      DO 6 J=1,7
      JO=KF1+2*J-1
      JE=KF1+2*J
      IF(LOCA(JO).GT.-1) LOCA(JO)=-1
    6 IF(LOCA(JE).LT.1) LOCA(JE)=1
      DO 2 K=1,14
    2 LOCA(KF2+K)=LOCA(KF1+K)
    1  CONTINUE
      N=30*M-14
```

```
      DO 4 I=1,N                                                        LOC  50
      IF(LOCA(I+1).LE.8191.AND.LOCA(I+1).GE.-8191) GO TO 4              LOC  51
      J=I-1                                                             LOC  52
    7 IF(J.LT.15) GO TO 8                                               LOC  53
      J=J-15                                                            LOC  54
      GO TO 7                                                           LOC  55
    8 OVER=LOCA(I+1)*100./8191.                                         LOC  56
      WRITE(6,100) J,OVER                                               LOC  57
      IF(LOCA(I+1).LT. -8191) LOCA(I+1)=-8191                           LOC  58
      IF(LOCA(I+1).GT. 8191) LOCA(I+1)= 8191                            LOC  59
    4 CONTINUE                                                          LOC  60
  100 FORMAT(/10X,'OVERLOAD DA(',I2,')=',F10.4,'VOLTS',4X,'ASSUMED IN '  LOC  61
     1OR - 100 VOLTS')                                                  LOC  62
  101 FORMAT(/10X,'OVERLOAD LOC(',I3,')=',F10.4,'VOLTS',4X,'ASSUMED IN  LOC  63
     1-100 VOLTS')                                                      LOC  64
      MAXA=30*M-14                                                      LOC  65
      DO 3 I=1,15                                                       LOC  66
    3  LOCA(MAXA+I)=0                                                   LOC  67
C                                                                       LOC  68
C     DA(2I-1)...INPUT TO LAMDA-F INTEGRATION FOR EACH SPECIES I        LOC  69
C     DA(2I)  ... INPUT TO F* INTEGRATION FOR EACH SPECI                LOC  70
C                                                                       LOC  71
      RETURN                                                            LOC  72
      END                                                               LOC  73
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
      SUBROUTINE PRINT(M)                                               PRN   1
      COMMON/PRN/LOCA,NAME                                              PRN   2
      INTEGER*2 LOCA(1100)                                              PRN   3
      INTEGER NAME(7)                                                   PRN   4
      REAL PLAM(7,32),PFST(7,32)                                        PRN   5
      DO 1 I=1,M                                                        PRN   6
      KF=30*I-28                                                        PRN   7
      DO 1 K=1,7                                                        PRN   8
      PLAM(K,I)=LOCA(KF+2*K-1)/8191.                                    PRN   9
    1 PFST(K,I)=LOCA(KF+2*K)/8191.                                      PRN  10
      WRITE(6,100) (NAME(I),I=1,7),((PLAM(K,I),PFST(K,I),K=1,7),I=1,M)  PRN  11
  100 FORMAT(//10X,'LIST OF LOCA IN DAC' ///7(9X,A4,4X)//100(7(1X, F6.3)PRN  12
     1/))                                                               PRN  13
      RETURN                                                            PRN  14
      END                                                               PRN  15
```

```
      SUBROUTINE STCHCK
C     ****** STATIC CHECK FOR PHOTOCHEMICAL SMOG DIFFUSION SYSTEM ******
      COMMON/PS/IVAL,NPOT,KVAL
      INTEGER*4 IHOLD,NPOT(28)
      INTEGER*2 IVAL(28),KVAL(28)
      INTEGER*2 PVAL(28),AVAL(49),LOCA(17)
      REAL*8 PTSCCW(2),PTSRCB(4)
      REAL CAMP(55),X(55),ERROR(55)
      INTEGER PX(28),PERROR(28),IAMP(55)
      DATA IAMP    /'A003','A004','A005','A007','A008','A009','A011',
     1'A012','A013','A101','A102','A103','A105','A106','A107','A1[illegible]',
     2'A110','A111','A113','A114','A115','M001','M002','M003','M[illegible]',
     3'M005','M006','M401','M402','M403','M404','M405','M406','M4[illegible]',
     4'M408','C003','C005','C007','C009','C011','C013','C101','C1[illegible]',
     5'C105','C107','C109','C111','C113','C115','A201','C201','M101',
     6'A203','C203','M102'/
      INTEGER FAMP(3)/'A201','C201','M101'/
      INTEGER GAMP(3)/'A203','C203','M102'/
      CALL XPS
      CALL FRCBSU(PTSRCB,29,PTSCCW)
      CALL POTSS(PTSCCW,29,NPOT,IVAL,PTSRCB)
      CALL EXC(PTSRCB)
      CALL XRDAN(NPOT,28,PVAL)
      DO 20 I=1,28
      PX(I)=PVAL(I)/0.8191
      A=PX(I)
      B=IVAL(I)
   20 PERROR(I)=ABS(A-B)
      WRITE(6,100) (NPOT(I),PX(I),IVAL(I),PERROR(I),I=1,28)
  100 FORMAT(//20X,'***** STATIC CHECK FOR PHOTOCHEMICAL SMOG DIFFUSION
     1******'            10X,' *.... POTS. TEST '//20(4(3X,A4,1X,'=',I7,
     21X,'(',I4,') ...',I4)/)//)
      WRITE(15,200)
  200 FORMAT('WANT TO POTSET AGAIN ? TYPE IN 1(I2)')
      READ(15,300)J1
  300 FORMAT(I2)
      IF(J1.EQ.1) GO TO 1
      CAMP(1)=-(-100.*0.1)
      CAMP(2)=-(5.+CAMP(1))
      CAMP(3)=-(100.*0.1)
      CAMP(22)=-(CAMP(1)*CAMP(2))/100.
      CAMP(23)=-(CAMP(2)*CAMP(3))/100.
      CAMP(36)=-(CAMP(22)*10.+5.)/10.
      CAMP(37)=-(CAMP(23)*1.0+10.)/10.
      CAMP(4)=-(-100.*0.1)
      CAMP(5)=-(5.+CAMP(4))
      CAMP(6)=-(100.*0.1)
      CAMP(24)=-(CAMP(4)*CAMP(5))/100.
      CAMP(25)=-(CAMP(5)*CAMP(6))/100.
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
      CAMP(38)=-(CAMP(24)+15.)/10.
      CAMP(39)=-(CAMP(25)+20.)/10.
      CAMP(7)=-(-100.*0.1)
      CAMP(8)=-(5.*0.1+CAMP(7))
      CAMP(9)=-(100.*0.1)
      CAMP(26)=-(CAMP(7)*CAMP(8))/100.
      CAMP(27)=-(CAMP(8)*CAMP(9))/100.
      CAMP(40)=-(CAMP(26)+25.)/10.
      CAMP(41)=-(CAMP(27)+30.)/10.
      CAMP(10)=-(-100.*0.1)
      CAMP(11)=-(5.+CAMP(10))
      CAMP(12)=-(100.*0.1)
      CAMP(28)=-(CAMP(10)*CAMP(11))/100.
      CAMP(29)=-(CAMP(11)*CAMP(12))/100.
      CAMP(42)=-(CAMP(28)*1.0+35.)/10.
      CAMP(43)=-(CAMP(29)     +40.)/10.
      CAMP(13)=-(-100.*0.1)
      CAMP(14)=-(5.+CAMP(13))
      CAMP(15)=-(100.*0.1)
      CAMP(30)=-(CAMP(13)*CAMP(14))/100.
      CAMP(31)=-(CAMP(14)*CAMP(15))/100.
      CAMP(44)=-(CAMP(30)*10.+45.)/10.
       CAMP(45)=-(CAMP(31)*10.+50.)/10.
      CAMP(16)=-(-100.*0.1)
      CAMP(17)=-(5.0+CAMP(16))
      CAMP(18)=-(100.*0.1)
       CAMP(32)=-(CAMP(16)*CAMP(17))/100.
      CAMP(33)=-(CAMP(17)*CAMP(18))/100.
       CAMP(46)=-(CAMP(32)*1.0+55.)/10.
      CAMP(47)=-(CAMP(33)*10.+60.)/10.
       CAMP(19)=-(-100.*0.1)
      CAMP(20)=-(5.+CAMP(19))
      CAMP(21)=-(100.*0.1)
      CAMP(34)=-(CAMP(19)*CAMP(20))/100.
      CAMP(35)=-(CAMP(20)*CAMP(21))/100.
      CAMP(48)=-(CAMP(34)*10.+65.)/10.
       CAMP(49)=-(CAMP(35)*10.+70.)/10.
       CAMP(50)=-(-20.)
       CAMP(52)=-(CAMP(50)*(-20.))/100.
      CAMP(51)=-(-20.+10.*CAMP(52))/10.
      CAMP(53)=-(-20.)
       CAMP(55)=-(CAMP(53)*(-20.))/100.
      CAMP(54)=-(-20.+10.*CAMP(55))/10.
60    CALL XSTCTL(IHOLD,15)
      CALL XRSCTL(IHOLD,14)
      LOCA(2)=0.05*8191.
      DATEST=0.05
      DO 30 I=1,14
      LOCA(I+2)=DATEST*8191.
```

```
30    DATEST=DATEST+0.05
      CALL XIC
      CALL XWRTDA(15,0,0,14,LOCA)
      CALL XRDAN(IAMP,49,AVAL)
      DO 70 I=1,49
 70   X(I)=AVAL(I)*100./8191.
      CALL XSTCTL(IHOLD,14)
      CALL XSTCTL(IHOLD,13)
      DO 50 I=1,3
 50    LOCA(I+1)=.2*8191.
      CALL XWRTDA(3,0,0,2,LOCA)
      CALL XRDAN(FAMP,3,AVAL)
      DO 90 I=1,3
 90    X(49+I)=AVAL(I)*100./8191.
      CALL XRSCTL(IHOLD,13)
      CALL XRDAN(GAMP,4,AVAL)
      DO 92 I=1,3
 92   X(52+I)=AVAL(I)*100./8191.
      DO 40 I=1,55
 40   ERROR(I)=ABS(X(I)-CAMP(I))
      WRITE(6,400) (IAMP(I),CAMP(I),X(I),ERROR(I),I=1,55)
 400  FORMAT(/5X,'COMPONENT',5X,'CALCULATION',10X,'REAL',9X,'ERROR'//
     1(7X,A4,9X,F10.4,4X,F10.4,4X,F10.4/))
      WRITE(15,500)
 500  FORMAT('REPEAT ? 1,READ OUT AGAIN ? 2')
      READ(15,300) J2
      IF(J2.EQ.1) GO TO 1
      IF(J2.EQ.2) GO TO 60
      WRITE(6,600)
 600  FORMAT(/20X,'***** END OF STATIC TEST ***** ')
      RETURN
      END
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
      SUBROUTINE UHPLOT(NO,A,NROW,NCOL,VERT,HORZ,BSORT)
      IMPLICIT INTEGER(A-Z)
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C     ..................................................................
C
C        SUBROUTINE UHPLOT
C
C        PURPOSE
C           PLOT SEVERAL CROSS-VARIABLES VERSUS A BASE VARIABLE
C               OVER-PRINTING OVERLAPPING POINTS.
C
C        USAGE
C           CALL UHPLOT(NO,A,NROW,NCOL,VERT,HORZ,BSORT)
C
C        DESCRIPTION OF PARAMETERS
C
C        NO = PLOT HEADING IDENTIFICATION - 20 WORDS OF FORMAT A4
C             INFORMATION PRINTED AT TOP OF PLOT. (80 CHARACTERS)
C        A = MATRIX OF DATA TO BE PLOTTED.  FIRST COLUMN REPRESENTS
C             BASE VARIABLE AND SUCCESSIVE COLUMNS ARE THE CROSS OR
C             DEPENDENT VARIABLES (MAXIMUM IS 9)
C             CAUTION:  MATRIX MUST HAVE DIMENSIONS (NROW,NCOL) ......
C       CONTINUE
C       NROW = NUMBER OF ROWS IN MATRIX A
C       NCOL = NUMBER OF COLUMNS IN MATRIX A.  MAX. IS 10.
C       VERT = VECTOR OF 5 WORDS IN A4 FORMAT DESCRIBING VERTICAL
C             SCALE UNITS.
C       HORZ = VECTOR OF 5 WORDS IN A4 FORMAT DESCRIBING HORIZONTAL
C             SCALE UNITS.
C       BSORT = CODE FOR SORTING BASE VARIABLE.
C             0 = DO NOT SORT.
C             1 = SORT BASE VARIABLE IN ASCENDING ORDER
C                 (I.E. COLUMN 1 OF MATRIX A).
C
C             NONE
C
C     ..................................................................
C
      REAL A,FLOAT,YSCAL,YMAX,YMIN,YPR,XPR,F
      DIMENSION A(1),NO(20),VERT(5),HORZ(5),YPR(11),ANG(9)
      DIMENSION OUT(101),OUTOVR(101)
C
      DATA BLANK,PLUS,ANG/' ','+','1','2','3','4','5','6','7','8','9'/
C
      M1 = NROW + 1
      YMIN = A(M1)
      MN = NROW * NCOL
```

```
 251  YMAX=YMIN
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C        COMPUTE YMAX, YMIN, YSCALE
C
      DO 205 J=M1, MN
      IF(YMAX .GE. A(J)) GO TO 200
      YMAX = A(J)
200   IF( YMIN .GE. A(J) ) YMIN = A(J)
205   CONTINUE
      YSCAL = (YMAX - YMIN) / 100.0
      YPR(1) = YMIN
      DO 208 J=1, 9
208   YPR(J+1) = YPR(J) + YSCAL * 10.0
      YPR(11) = YMAX
C        SET UP FIRST VERTICAL SCALE VALUE FOR LATER USE.
      XPR = A(1)
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C        SORT BASE VARIABLE DATA IN ASCENDING ORDER OF BSORT = 1
C
      IF(BSORT) 300,300,210
C
210   DO 215 I=1, NROW
      DO 214 J=1,NROW
      IF( A(I) - A(J)) 214,214,211
211   L = I - NROW
      LL = J - NROW
      DO 212 K=1, NCOL
      L = L + NROW
      LL = LL + NROW
      F = A(L)
      A(L) = A(LL)
212   A(LL) = F
214   CONTINUE
215   CONTINUE
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C        PRINT HEADING , SCALES AND 1ST LINE OF PLOT
C
300   WRITE(6,100)
      WRITE(6,101) NO
      WRITE(6,102) VERT
       WRITE (6,103) HORZ
      WRITE(6,104) (YPR(J),J=2,10,2)
      WRITE(6,105)
      WRITE(6,106) (YPR(J),J=1,11,2)
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C        PRINT BODY OF GRAPH
C
C        PERFORM INITIAL SPACE -- SPACING IS DONE BY A SEPERATE
C              WRITE STATEMENT TO ALLOW OVERPRINTING.
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
      WRITE(6,107)
C
      MY = NCOL - 1
      L = 1
340   DO 390 N=2, NROW
C        BLANK BUFFER - OUT
      DO 355 IX=1, 101
355   OUT(IX) = BLANK
C        INSERT PLUS GRID MARKS
      DO 358 K=1, 101, 10
358   OUT(K) = PLUS
C        FILL BUFFER WITH ALL POINTS - OVERPRINTING WHERE OVERLAPS
C              OCCUR.
C
      DO 370 J=1, MY
      LL = L + J*NROW
      JP = ( (A(LL) - YMIN) / YSCAL) + 1.0
      IF( OUT(JP) .NE. BLANK .AND. OUT(JP) .NE. PLUS) GO TO 361
      OUT(JP) = ANG(J)
      GO TO 370
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C        OVER-PRINT SECTION
C
361   DO 362 K=1, 101
362   OUTOVR(K) = BLANK
      OUTOVR(JP) = ANG(J)
      WRITE(6,108) OUTOVR
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
370   CONTINUE
C        PRINT REGULAR LINE THEN SPACE
C
      WRITE(6,109) XPR,(OUT(I),I=1,101)
      WRITE(6,107)
      XPR = A(N)
      L = L + 1
390   CONTINUE
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C        PRINT LAST LINES
C
      WRITE(6,104) (YPR(J),J=2,10,2)
      WRITE(6,105)
      WRITE(6,106) (YPR(J),J=1,11,2)
C
      RETURN
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
100   FORMAT(1H1)
101   FORMAT(///,5X,'---- ',20A4,' ----')
```

```
102   FORMAT(//,10X,'VERTICAL SCALE UNITS    = ',5A4)                    UHPL14
103   FORMAT(10X,'HORIZONTAL SCALE UNITS = ',5A4,////)                   UHPL15
104   FORMAT(T19,5(G17.6,3X))                                            UHPL15
105   FORMAT(T28,'|',4(19X,'|'))                                         UHPL15
106   FORMAT(9X,G17.6,T28,'|',1X,G17.6,T48,'|',1X,G17.6,T68,'|',1X,      UHPL15
     1  G17.6,T88,'|',1X,G17.6,T108,'|',1X,G17.6)                        UHPL15
107   FORMAT(1H0/)                                                       UHPL15
108   FORMAT(1H+,16X,101A1)                                              UHPL15
109   FORMAT(1H+,G12.6,4X,101A1)                                         UHPL15
      END                                                                UHPL15
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
/&
//ADD  JOB  ,SD2004G5DJAHLK          ADD XSERIES
//    ACCESS  SDSREL
//  DELETE  SDSREL(XMODE,XIC,XOP,XPS,XHD,XRT,XST)
//SYS000 ACCESS SDSREL(XMODE,XIC,XOP,XPS,XHD,XRT,XST),NEW
//XMODE   EXEC  FORTRAN(MAP)
      SUBROUTINE XMODE
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C               SUBROUTINE TO SET THE ANALOG MODE.
C               TO SET THE DESIRED MODE CALL WITH THE PROPER ENTRY
C               POINT.
C
C               NOTE: IF CALLED AS XMODE NO MODE CHANGE OCCURS.
C
      REAL*8 MODRCB(4),MODCCW
      INTEGER*2 M,COMBTE/23/,OP/4/
      INTEGER*4 CONTRL/28/,SUBNAM(2)/'XMOD','E   '/
      EXTERNAL FRCBSU,FRTIO,FCHECK,CCWCB,XERRAN
      RETURN
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
      ENTRY XIC
      M = 16
      GO TO 300
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
      ENTRY XOP
      M = 8
      GO TO 300
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
      ENTRY XPS
      M = 32
      GO TO 300
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
      ENTRY XHD
      M = 4
      GO TO 300
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
      ENTRY XRT
      M = 10
      GO TO 300
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
      ENTRY XST
      M = 17
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
300   CONTINUE
```

```
      CALL CCWCB(MODCCW,COMBTE,OP,M,2)                                  XMOD 54
      CALL FRCBSU(MODRCB,CONTRL,MODCCW)                                 XMOD 55
      CALL FRTIO(MODRCB,IRET)                                           XMOD 56
      IF( IRET .NE. 0 ) CALL XERRAN(1,IRET,SUBNAM)                      XMOD 57
320   CALL FCHECK(MODRCB,IRET,0)                                        XMOD 58
      IF ( IRET .EQ. 4 ) GO TO 320                                      XMOD 59
      IF(IRET .NE. 0 ) CALL XERRAN(2,IRET,SUBNAM)                       XMOD 60
      RETURN                                                            XMOD 61
      END                                                               XMOD 62
```

```
/&
//ADD  JOB  ,SD2004G5DJAHLK          ADD XSERIES
//     ACCESS  SDSREL
//  DELETE  SDSREL(XSTCTL,XRSCTL)
//SYS000 ACCESS SDSREL(XSTCTL,XRSCTL),NEW
//XSTCTL  EXEC  FORTRAN(MAP)
      SUBROUTINE  XSTCTL(IHOLD,LINE)                                    XSCL
C               IHOLD = INTEGER*2 TEMPORY WORK AREA THAT HAS BITS SET   XSCL
C               SUBROUTINE TO SET A GIVEN SINGLE CONTROL LINE.          XSCL
C               IHOLD = INTEGER*4 TEMPORY WORK AREA THAT HAS BITS SET   XSCL
C               IN IT ACCORDING TO THE PRIOR REQUESTS THAT HAVE BEEN    XSCL
C               PROCESSED.                                              XSCL
                                                                        XSCL
C               IF IHOLD IS SET TO ZERO AND ANOTHER CALL ON XSTCTL IS   XSCL
C               MADE ALL OTHER LINES WILL BE RESET EXCEPT FOR THE       XSCL
C               REQUESTED LINE.                                         XSCL
C               LINE = INTEGER*4 THE LINE NUMBER TO BE SET.             XSCL
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---   XSCL
      REAL*8  XRCB(4),XCCW                                              XSCL
      INTEGER*2   OP/4/,COMBTE/31/,IHOLD                                XSCL
      INTEGER*4    CONTRL/28/,SUBNAM(2)/'XWTC','TL  '/,ONE/1/,MONE/-1/  XSCL
      EXTERNAL     FRTIO,FCHECK,FRCBSU,CCWCB,XERRAN                     XSCL
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---   XSCL
C                                                                       XSCL
C               THESE SUBROUTINES DO NOT DESTROY THE VALUE OF ANY OTHER XSCL
C               CONTROL LINES.  THE ONLY LINE CHANGED IS THE ONE        XSCL
C               REQUESTED.  THIS IS DUE TO IHOLD HOLDING THE PAST RECORD XSCL
C               OF ALL CALLS ON XSTCTL AND XRSCTL.                      XSCL
C                                                                       XSCL
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---   XSCL
      ITEMP = ISL(ONE,(15-LINE))                                        XSCL
      MTEMP = IHOLD                                                     XSCL
      IHOLD = IOR(MTEMP,ITEMP)                                          XSCL
      GO TO 290                                                         XSCL
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---   XSCL
C                                                                       XSCL
      ENTRY  XRSCTL (IHOLD,LINE)                                        XSCL
C                                                                       XSCL
C               SUBROUTINE TO RESET A GIVEN SINGLE CONTROL LINE.        XSCL
C                                                                       XSCL
      ITEMP = ISL(ONE,(15-LINE))                                        XSCL
      ITEMP = IEOR(ITEMP,MONE)                                          XSCL
      MTEMP = IHOLD                                                     XSCL
      IHOLD = IAND(MTEMP,ITEMP)                                         XSCL
290   IF (LINE .LT. 0  .OR.  LINE .GT. 15)  GO TO 600                   XSCL
      CALL CCWCB(XCCW,COMBTE,OP,IHOLD,2)                                XSCL
      CALL  FRCBSU   (XRCB,CONTRL,XCCW)                                 XSCL
      CALL  FRTIO(XRCB,IRET)                                            XSCL
      IF(IRET .NE. 0)  CALL XERRAN(1,IRET,SUBNAM)                       XSCL
300   CALL FCHECK(XRCB,IRET,0)                                          XSCL
      IF (IRET .EQ. 4) GO TO 300                                        XSCL
      IF (IRET .NE. 0) CALL XERRAN(2,IRET,SUBNAM)                       XSCL
320   RETURN                                                            XSCL
600   WRITE(6,100) LINE                                                 XSCL
100   FORMAT(3X,'$$$$***- ERROR IN XWTCTL - '/' CONTROL LINE = ',15,    XSCL
     1' .LT. 0 OR .GT. 15',/' LINE SET TO -9999')                       XSCL
      LINE =-9999                                                       XSCL
      GO TO 320                                                         XSCL
      END                                                               XSCL
```

p. 388

```
//ADD  JOB  ,SD200465DJAHLK          ADD XSERIES
//     ACCESS  SDSREL
//  DELETE  SDSREL(XRDAN)
//SYS000 ACCESS  SDSREL(XRDAN),NEW
//XRDAN   EXEC  FORTRAN(MAP)
      SUBROUTINE XRDAN(NAME,N,IVAL )
C
C  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C              SUBROUTINE TO READ ANY ANALOG COMPONENT ON THE HSI SS-100
C              NAME = VECTOR OF UP TO 100 A4 FORMAT ANALOG COMPONENT
C              NAMES....P101, A005    ETC.
C              N = NUMBER OF COMPONENTS TO BE READ
C              INTEGER*2 VECTOR IN WHICH THE READ VALUES WILL BE
C              RETURNED.
C
C  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
      REAL*8 ANARCB(8),ANACCW(2)
      DIMENSION NAME(1)
      INTEGER*4 CONTRL/28/,READAD/29/
      INTEGER*2 IVAL(1),ANADDR(100)
      EXTERNAL ANALOG,FRCBSU,FRTIO,FCHECK,XERRAN
C  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
      IF( N .GT. 100 ) GO TO 600
      CALL ANALOG(ANACCW,N,NAME,ANADDR,IVAL)
      CALL FRCBSU(ANARCB,CONTRL,ANACCW,4,READAD,ANACCW(2))
      CALL FRTIO(ANARCB,IRET)
      IF( IRET .NE. 0) CALL XERRAN(1,IRET,'XRDAN   ')
C              FCHECK OPTI = 0 SO CAN BE USED IN RTT
200   CALL FCHECK(ANARCB,IRET,0)
      IF( IRET .EQ. 4) GO TO 200
      IF( IRET .NE. 0 ) CALL XERRAN(2,IRET,'XRDAN   ')
220   RETURN
C
C  --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C              ERROR REPORT SECTION
C
600   WRITE(8,100) N
100   FOR AT(3X,'$$$$***- N .GT. 100 IN -XRDAN, N= ',I10)
      GO TO 220
      END
```

ORIGINAL PAGE IS
OF POOR QUALITY

```
/&
//ADD  JOB  ,SD2004G5DJAHLK          ADD XSERIES
//    ACCESS  SDSREL
//  DELETE  SDSREL(XWRTDA)
//SYS000 ACCESS  SDSREL(XWRTDA),NEW
//XWRTDA   EXEC  FORTRAN(MAP)
      SUBROUTINE XWRTDA(N,ICON,SA,EA,LOCA)                              XWDA  1
C                                                                       XWDA  2
C               SUBROUTINE TO DO A DIGITAL TO ANALOG WRITE.             XWDA  3
C               ****  WORKS IN THE SEQUENTIAL MODE ONLY  ****           XWDA  4
C               N = NUMBER OF D/A TO BE WRITTEN.                        XWDA  5
C               ICON = MODE CONTROL                                     XWDA  6
C               SA = INTEGER*4 STARTING ADDRESS OF D/A                  XWDA  7
C               EA = INTEGER*4 ENDING ADDRESS OF D/A                    XWDA  8
C               LOCA = INTEGER*2 VECTOR OF VALUES TO BE WRITTEN.        XWDA  9
C               LOCA(1) CAN NOT BE USED AS IT IS USED BY THE SYSTEM.    XWDA 10
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  XWDA 11
      REAL*8 WTRCB(4),WTCCW(2)                                          XWDA 12
      INTEGER*4 SA,EA                                                   XWDA 13
      INTEGER*4 WRTDAL/30/,SUBNAM(2)/'XWRT','DA  '/,ONE/1/,             XWDA 14
     1 M8ASK/ZFFFFFFF7/                                                 XWDA 15
      INTEGER*2 LOCA(1)                                                 XWDA 16
      EXTERNAL FRTIO,FCHECK,XERRAN,WRITDA,FRCBSU                        XWDA 17
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  XWDA 18
C                                                                       XWDA 19
      LOCA(1) = SA*256 + EA                                             XWDA 20
200   IF( IAND( ISL(ICON,-2) , ONE ) .EQ. ONE ) GO TO 300               XWDA 21
      GO TO 350                                                         XWDA 22
300   WRITE(8,100) ICON                                                 XWDA 23
100   FORMAT(3X,'$$$$***- ICON = ',I5,' IN XWRTDA INVALID'/' RANDOM MODEXWDA 24
     1 NOT SUPPORTED'/' RETYPE ICON - I2...EOB')                        XWDA 25
      READ(8,101) ICON                                                  XWDA 26
101   FORMAT(I2)                                                        XWDA 27
      GO TO 200                                                         XWDA 28
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  XWDA 29
C                                                                       XWDA 30
C               THE FOLLOWING IS INSERTED SINCE THE HSI 1044 HAS NOT    XWDA 31
C               BEEN WORKING FOR A WRITE D/A IN THE BURST MODE WITHOUT  XWDA 32
C               DOUBLE BUFFERING.  THE STATEMENT AT 350 RESETS THE MODE XWDA 33
C               TO MULTIPLEXED IF BURST MODE HAS BEEN REQUESTED         XWDA 34
C               WITHOUT DOUBLE BUFFERING.                               XWDA 35
C                                                                       XWDA 36
350   IF (IAND (ICON,8) .EQ. 8 .AND. IAND( ICON,ONE) .EQ. 0)            XWDA 37
     1ICON = IAND(ICON,M8ASK)                                           XWDA 38
C                                                                       XWDA 39
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---  XWDA 40
      CALL  WRITDA ( WTCCW,ICON,N,LOCA)                                 XWDA 41
      CALL  FRCBSU ( WTRCB,WRTDAL,WTCCW)                                XWDA 42
      CALL FRTIO  (WTRCB,IRET)                                          XWDA 43
      IF(IRET .NE. 0)  CALL XERRAN(1,IRET,SUBNAM)                       XWDA 44
360   CALL FCHECK(WTRCB,IRET,0)                                         XWDA 45
      IF(IRET .EQ. 4)  GO TO 360                                        XWDA 46
      IF(IRET .NE. 0) CALL XERRAN(2,IRET,SUBNAM)                        XWDA 47
      RETURN                                                            XWDA 48
      END                                                               XWDA 49
```

```
/&
//ADD  JOB  ,SD200465DJAHLK          ADD XSERIES
//     ACCESS  SDSREL
//  DELETE  SDSREL(XWRTDA)
//SYS000 ACCESS  SDSREL(XWRTDA),NEW
//XWRTDA   EXEC  FORTRAN(MAP)
      SUBROUTINE XWRTDA(N,ICON,SA,EA,LOCA)
C
C               SUBROUTINE TO DO A DIGITAL TO ANALOG WRITE.
C               ****  WORKS IN THE SEQUENTIAL MODE ONLY  ****
C               N = NUMBER OF D/A TO BE WRITTEN.
C               ICON = MODE CONTROL
C               SA = INTEGER*4 STARTING ADDRESS OF D/A
C               EA = INTEGER*4 ENDING ADDRESS OF D/A
C               LOCA = INTEGER*2 VECTOR OF VALUES TO BE WRITTEN.
C               LOCA(1) CAN NOT BE USED AS IT IS USED BY THE SYSTEM.
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
      REAL*8 WTRCB(4),WTCCW(2)
      INTEGER*4 SA,EA
      INTEGER*4 WRTDAL/30/,SUBNAM(2)/'XWRT','DA  '/,ONE/1/,
     1 M8ASK/ZFFFFFFF7/
      INTEGER*2 LOCA(1)
      EXTERNAL FRTIO,FCHECK,XERRAN,WRITDA,FRCBSU
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
      LOCA(1) = SA*256 + EA
200   IF( IAND( ISL(ICON,-2) , ONE ) .EQ. ONE ) GO TO 300
      GO TO 350
300   WRITE(8,100) ICON
100   FORMAT(3X,'$$$$***- ICON = ',I5,' IN XWRTDA INVALID'/' RANDOM MODE
     1 NOT SUPPORTED'/' RETYPE ICON - I2...EOB')
      READ(8,101) ICON
101   FORMAT(I2)
      GO TO 200
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
C
C               THE FOLLOWING IS INSERTED SINCE THE HSI 1044 HAS NOT
C               BEEN WORKING FOR A WRITE D/A IN THE BURST MODE WITHOUT
C               DOUBLE BUFFERING.  THE STATEMENT AT 350 RESETS THE MODE
C               TO MULTIPLEXED IF BURST MODE HAS BEEN REQUESTED
C               WITHOUT DOUBLE BUFFERING.
C
350   IF (IAND (ICON,8) .EQ. 8 .AND. IAND( ICON,ONE) .EQ. 0)
     1ICON = IAND(ICON,M8ASK)
C
C --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
      CALL  WRITDA ( WTCCW,ICON,N,LOCA)
      CALL  FRCBSU ( WTRCB,WRTDAL,WTCCW)
      CALL FRTIO  (WTRCB,IRET)
      IF(IRET .NE. 0)  CALL XERRAN(1,IRET,SUBNAM)
360   CALL FCHECK(WTRCB,IRET,0)
      IF(IRET .EQ. 4)  GO TO 360
      IF(IRET .NE. 0) CALL XERRAN(2,IRET,SUBNAM)
      RETURN
      END
```

ORIGINAL PAGE IS OF POOR QUALITY

```
      BLOCK DATA
      COMMON/PS/IVAL,NPOT,KVAL
      COMMON/SM/SCAL,MAX1,DA
      COMMON/QQ/QHC,QNO2,QNO,DTIME,KMIN,TIME,KTIME,MIN,TEND,VIC,VFIN
      COMMON/PRN/LOCA,NAME
      COMMON/FO/FONO2,FONO,FOO,FOO3, FOHC,FORO2,FOOH
      REAL QHC(12),QNO2(12),QNO(12),FOOH(32),KMIN(11),DA(14),SCAL(7)
      REAL FONO2(32),FONO(32),FOO(32),FOO3(32),FOHC(32),FORO2(32)
      REAL VIC(4),VFIN(4),DTIME/.5/,TEND/4.0/,TIME/11.30/
      INTEGER*4 NPOT(28)
      INTEGER*2 IVAL(28),KVAL(28),LOCA(1100)
      INTEGER NAME(7),MAX1/32/,          KTIME/7/,MIN/30/
      DATA SCAL/1., 2., 10.0,2.,1.,10.0,10.0/
      DATA NAME/'NO2','NO','O','O3','HC','RO2','OH'/
      DATA QNO2/0.430,0.334,0.265,0.305,0.299,0.230,0.092,0.242,4*0.0/
      DATA QNO/1.290,1.0,0.795,0.915,0.896,0.690,0.276,0.725,4*0.0/
      DATA QHC/3.40,2.64,1.86,2.39,2.54,0.909,1.62,2.15,4*0.0/
      DATA FONO/32*0.0/
      DATA FONO2/32*0.0/
      DATA FOO/32*0./
      DATA FOO3/32*0.0/
      DATA FOHC/32*0.0/
      DATA FORO2/32*0./
      DATA FOOH/32*0./
      DATA IVAL/19*1000,7*10,2*50/
      DATA KVAL/0,1000,0,0,1000,0,0,1000,6*0,1000,0,0,1000,0,20,
     12*10,20,25,2*10,2*50/
      DATA NPOT/'P003','P004','P005','P007','P008','P009','P011','P012',
     1'P013','P101','P103','P105','P107','P109','P110','P111','P112',
     2'P114','P115','P201','P202','P203','P204','P205','P206','P207',
     3'P208','P209'/
      DATA DA/0.01,0.01,0.001,0.01,2*0.001,3*0.01,0.1,3*0.001,0.01/
      DATA KMIN/.6,0.264E-02,40.,305.,4.,1500.,6.,10.,30.,.0125,0.01/
      DATA VIC/0.174,0.439,0.100,0.945/
      DATA VFIN/0.213,0.032,0.236,0.64/
      END
```

```
      BLOCK DATA
      COMMON/PS/IVAL,NPOT,KVAL
      COMMON/SM/SCAL,MAXI,DA
      COMMON/QQ/QHC,QNO2,QNO,DTIME,KMIN,TIME,KTIME,MIN,TEND,VIC,VF..
      COMMON/PRN/LOCA,NAME
      COMMON/FO/FONO2,FONO,FOO,FOO3, FOHC,FORO2,FOOH
      REAL QHC(12),QNO2(12),QNO(12),FOOH(32),KMIN(11),DA(14),SCAL(7)
      REAL FONO2(32),FONO(32),FOO(32),FOO3(32),FOHC(32),FORO2(32)
      REAL VIC(4),VFIN(4),DTIME/.5/,TEND/4.0/,TIME/11.30/
      INTEGER*4 NPOT(28)
      INTEGER*2 IVAL(28),KVAL(28),LOCA(1100)
      INTEGER NAME(7),MAXI/32/,          KTIME/77/,MIN/30/
      DATA SCAL/1., 2., 10.0,2.,1.,10.0,10.0/
      DATA NAME/'NO2','NO','O','O3','HC','RO2','OH'/
      DATA QNO2/0.430,0.334,0.265,0.305,0.299,0.23[illegible],0.[illegible]92,0.[illegible]42,4*[illegible]/
      DATA QNO/1.290,1.0,0.795,0.915,0.896,0.690,0.276,0.725,4*[illegible]/
      DATA QHC/3.40,2.64,1.86,2.39,2.54,0.909,1.62,2.15,4*0.[illegible]/
      DATA FONO/32*0.0/
      DATA FONO2/32*0.0/
      DATA FOO/32*0./
      DATA FOO3/32*0.0/
      DATA FOHC/32*0.0/
      DATA FORO2/32*0./
      DATA FOOH/32*0./
      DATA IVAL/19*1000,7*10,2*50/
      DATA KVAL/0,1000,0,0,1000,0,0,1000,6*0,1000,0,0,1000, ,20,
     12*10,20,25,2*10,2*50/
      DATA NPOT/'P003','P004','P005','P007','P008','P009','P011','P012',
     1'P013','P101','P103','P105','P107','P109','P110','P111','P11[illegible]',
     2'P114','P115','P201','P202','P203','P204','P205','P206','P207',
     3'P208','P209'/
      DATA DA/0.01,0.01,0.001,0.01,2*0.001,3*0.01,0.1,3*0.001,0.01/
      DATA KMIN/.6,0.264E-02,40.,305.,4.,1500.,6.,10.,30.,.0125, . 1/
      DATA VIC/0.174,0.439,0.100,0.945/
      DATA VFIN/0.213,0.032,0.236,0.64/
      END
```

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: Periodic and Cut-Off Responses for Linear State-Space Equations

(B) Project Abstract:

Two problems are investigated in this paper:

1) The complete response to periodic inputs of a linear time-varying system with periodically varying parameters is synthesized. When the periodic input is continuous or piecewise continuous, a computational method is established to decompose the complete response of a linear system into two components: the periodic response and the transient response.

2) For inputs which cut off at time t_1, the cut-off response of linear time-invariant or time-varying systems is also analysed. A computational method to obtain the cut-off response for such inputs is derived.

Several examples which illustrate the methods are included.

(C) Publication: International Journal of Control, 16, 369 (1972)

(D) Year: 1972

(E) Department: Electrical Engineering

(F) Student Name: L. S. Shieh

(G) Faculty Advisor: Professors G. F. Paskusz and D. R. Williams

INT. J. CONTROL, 1972, VOL. 16, NO. 2, 369–377

Periodic and cut-off responses for linear state-space equations†

ORIGINAL PAGE IS
OF POOR QUALITY

L. S. SHIEH, G. F. PASKUSZ and D. R. WILLIAMS
Department of Electrical Engineering, University of Houston,
Houston, Texas 77004

[Received 6 September 1971]

Two problems are investigated in this paper:

(1) The complete response to periodic inputs of a linear time-invariant system or time-varying system with periodically varying parameters is synthesized. When the periodic input is continuous or piecewise continuous, a computational method is established to decompose the complete response of a linear system into two components; the periodic response and the transient response.

(2) For inputs which cut off at time t_1, the cut-off response of linear time-invariant or time-varying systems is also analysed. A computational method to obtain the cut-off response for such inputs is derived.

Several examples which illustrate the methods are included.

1. Introduction

The complete response of a linear time-invariant system to continuous periodic inputs may be obtained by classical methods, for example Laplace transform and Fourier transform. When the input functions are discontinuous, one usually needs to refer to stepwise integration.

For a linear time-varying system, the analytical closed form response is not necessarily obtainable. However, the complete response may be obtained by means of numerical integration techniques.

The recent emphasis on practical applications has led to a focusing on the study of the steady-state response as a component of the complete response of a linear system to continuous periodic inputs. For a linear time-invariant system the subject has been analysed by the methods of Brand (1968), Susman (1969), Blackman (1961). The steady-state response of linear time-invariant systems to periodic non-sinusoidal input functions has been investigated by the methods of Davison and Smith (1971), and Miyata (1964). The methods of this paper are applicable to time-invariant and a class of time-varying linear systems as well.

In many practical applications the input function cuts off at some time t_1. The response of a linear systems to such an input may cease at $t \geqslant t_1$ if the proper set of initial conditions is chosen. This response is called the cut-off response and has been investigated by the method of Brand (1969), but that paper again treats linear time-invariant systems only.

The computational techniques established in this paper are valid for the analysis of the complete response and of the cut-off response for time-invariant and a class of time-varying linear systems. When the input is a periodic continuous function, or a piecewise continuous function whose pieces are of class

† Communicated by the Authors.

p.395

C^{n-1}, then the complete response of a linear system can be decomposed into a periodic and a transient component.

For the input which cuts off at a time t_1, a set of initial conditions can be chosen which will yield the cut-off response.

The approach rests on the use of the superposition principle.

2. Evaluation of transition matrix and zero-state response

A linear system can be described by the state equation

$$\dot{X}(t) = A(t)X(t) + B(t)U(t), \quad X(0) = X(t_0), \tag{1}$$

where $X(t)$ is the n-dimensional state vector, $A(t)$ and $B(t)$ are $n \times n$ and $n \times m$ matrices respectively, and $U(t)$ is the m-dimensional input vector.

The solution of (1) is given by

$$X(t) = \Phi(t, t_0)X(t_0) + P(t, t_0), \tag{2}$$

where

$$P(t, t_0) = \int_{t_0}^{t} \Phi(t, \lambda)B(\lambda)U(\lambda)d\lambda$$

and $\Phi(t, t_0)$ is the state transition matrix. Although the explicit forms of $\Phi(t, t_0)$ and $P(t, t_0)$ in (2) are not always obtainable, numerical solutions for them can be generated by numerical integration. Let $\Phi(t, t_0)$ have columns $\phi_1, \phi_2, \ldots, \phi_n$ or

$$\Phi(t, t_0) = (\phi_j) \quad (j = 1, 2, \ldots, n) \tag{3}$$

and choose $X(t_0)$ as the natural basis, or

$$X_j(t_0) = \delta_j^n,$$

where

$$\delta_j^n = (0, \ldots, 0, 1, 0, \ldots, 0)^T \quad (j = 1, 2, \ldots, n). \tag{4}$$

The numerical values of $\Phi(t, t_0)$ and $P(t, t_0)$ can be evaluated for any time t, using the following procedure :

Step A : let $U(t) = 0$, and apply the Runge–Kutta integration to (1) to obtain

$$\begin{aligned} X_j(t) &= \Phi(t, t_0)X_j(t_0) = \Phi(t, t_0)\delta_j^n \\ &= \phi_j \quad (j = 1, 2, \ldots, n). \end{aligned} \tag{5}$$

Evaluation of $\Phi(t, t_0)$ for any time t requires n numerical integrations.

Step B : let $U(t) \neq 0$ and apply numerical integration again to obtain :

$$\begin{aligned} X(t) &= \Phi(t, t_0)X_1(t_0) + P(t, t_0) \\ &= \Phi(t, t_0)\delta_1^n + P(t, t_0) \\ &= \phi_1 + P(t, t_0). \end{aligned} \tag{6}$$

Since ϕ_1 has been obtained previously, eqn. (6) may be solved for $P(t, t_0)$:

$$P(t, t_0) = X(t) - \phi_1. \tag{7}$$

Thus, the zero-state response $P(t, t_0)$ is obtained.

p.396

3. Synthesis of complete response

Let $U(t)$ be a periodic input function of period T and be either continuous or piecewise continuous with pieces of class C^{n-1}, and let $X_p(t)$ be the periodic response, $X_{tr}(t)$ the transient response. It is well known that the complete response $X(t)$ is a linear combination of $X_p(t)$ and $X_{tr}(t)$. The explicit equations are as follows:

$$U(t)=U(t+KT), \quad t_0 \leqslant t \leqslant T+t_0 \quad (K=1, 2, \ldots), \tag{8}$$

$$X_p(t)=X_p(t+KT), \quad t_0 \leqslant t \leqslant T+t_0 \quad (K=1, 2, \ldots) \tag{9}$$

and

$$X(t)=X_p(t)+X_{tr}(t) \quad t_0 \leqslant t < \infty. \tag{10}$$

Rewriting eqns. (1) and (2) as (11) and (12) respectively we have

$$\dot{X}(t)=A(t)X(t)+B(t)U(t) \tag{11}$$

and

$$X(t)=\Phi(t, t_0)X(t_0)+P(t, t_0). \tag{12}$$

Where the linear system is restricted to be a system of time invariant or a system with periodically varying parameters. Our goal is to decompose the complete response $X(t)$ of (12) into two components: the periodic response $X_p(t)$ and the transient response $X_{tr}(t)$ or

$$X(t)=X_p(t)+X_{tr}(t). \tag{13}$$

If the periodic response exists, a set of initial conditions, defined as periodic conditions and denoted by $X_p(t_0)$, can be chosen such that the solution of (12) yields the periodic response or

$$\Phi(t, t_0)X_p(t_0)+P(t, t_0)=X_p(t), \tag{14}$$

$$\Phi(t+T, t_0)X_p(t_0)+P(t+T, t_0)=X_p(t+T) \tag{15}$$

and

$$X_p(t)=X_p(t+T), \quad t_0+KT \leqslant t \leqslant t_0+(K+1)T. \tag{16}$$

Replacing t_0 for t in (14), (15), and (16) yields

$$\begin{aligned} X_p(t_0) &= \Phi(t_0, t_0)X_p(t_0)+P(t_0, t_0) \\ &= IX_p(t_0), \end{aligned} \tag{17}$$

$$X_p(t_0+T)=\Phi(t_0+T, t_0)X_p(t_0)+P(t_0+T, t_0) \tag{18}$$

and

$$X_p(t_0)=X_p(t_0+T), \tag{19}$$

where

$$\Phi(t_0, t_0)=I \quad \text{and} \quad P(t_0, t_0)=0.$$

Equating (17) and (18) yield

$$X_p(t_0)=[I-\Phi(t_0+T, t_0)]^{-1}P(t_0+T, t_0), \tag{20}$$

if the inverse matrix in the right side of (20) exists. Equation (20) shows us that we can adjust the initial conditions $X(t_0)$ of (12) to $X_p(t_0)$ of (20) for which the complete response of (12) will be the periodic response $X_p(t)$. In other words, by using the set of initial conditions $X_p(t_0)$ and applying the numerical

p. 397

integration technique to the given system (11) the periodic response $X_p(t)$ can be obtained. The periodic response of (12) is therefore

$$X_p(t) = \Phi(t, t_0)[I - \Phi(t_0+T, t_0)]^{-1}P(t_0+T, t_0) + P(t, t_0)$$

and

$$X_p(t) = X_p(t+KT), \quad t_0+KT \leqslant t \leqslant t_0+(K+1)T \quad (K=1, 2, \ldots). \tag{21}$$

Recall that the numerical values of $\Phi(t_0+T, t_0)$ and $P(t_0+T, t_0)$ in (21) can be obtained from (5) and (7). The transient response $X_{tr}(t)$ can be evaluated from (10) or

$$X_{tr}(t) = X(t) - X_p(t). \tag{22}$$

Substituting (12) and (14) in (22) yields

$$X_{tr}(t) = \Phi(t, t_0)[X(t_0) - X_p(t_0)], \quad t_0 \leqslant t < \infty. \tag{23}$$

4. Analysis of cut-off response

In many practical applications the input function exists in a finite interval T only or

$$U(t) = 0, \quad T \leqslant t < \infty. \tag{24}$$

The response of a linear system to this input depends on the initial conditions chosen, and a set of initial conditions can be chosen for which the complete response will vanish for $t \geqslant T$. This is called the cut-off response.

We again rewrite eqns. (1) and (2) as (25) and (26) respectively

$$\dot{X}(t) = A(t)X(t) + B(t)U(t), \tag{25}$$

$$X(t) = \Phi(t, t_0)X(t_0) + P(t, t_0), \tag{26}$$

where the linear system is time invariant or time varying. A set of initial conditions, defined as cut-off conditions and denoted by $X_c(t_0)$, can be chosen for which at $t = T + t_0$ the summation of the right side of (26) yields the cut-off response or

$$\Phi(t_0+T, t_0)X_c(t_0) + P(t_0+T, t_0) = X(t_0+T) = 0. \tag{27}$$

Since

$$U(t) = 0, \quad t_0+T \leqslant t$$

and

$$X(t_0+T) = 0. \tag{28}$$

The uniqueness theorem (Brand 1966) shows that $X(t)$ vanishes when $t \geqslant t_0+T$ and the $X(t)$ of (26) is a cut-off response. Rearranging eqn. (27) yields

$$X_c(t_0) = -\Phi(t_0+T, t_0)^{-1}P(t_0+T, t_0), \tag{29}$$

if $\Phi(t_0+T, t_0)^{-1}$ exists. Recall again that $\Phi(t_0+T, t_0)$ and $P(t_0+T, t_0)$ can be obtained from (5) and (7). By using $X_c(t_0)$ and applying a numerical integration technique the cut-off response can be obtained. The cut-off response is therefore

$$X_c(t) = -\Phi(t, t_0)\Phi(t_0+T, t_0)^{-1}P(t_0+T, t_0) + P(t, t_0). \tag{30}$$

1.398

5. Illustrative examples

5.1. *Examples of periodic response*

Example 1

Consider the following unstable system with a sinusoidal input of period 2π and initial conditions as given below :

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \sin t \tag{31}$$

and

$$\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{32}$$

The numerical values of the transition matrix and that of the zero-state response evaluated at $t = 2\pi$ can be obtained from (5) and (7). However, for this example the explicit form of the response can be written as follows :

$$\Phi(t) = \Phi(t)X(0) + P(t), \tag{33}$$

where

$$\Phi(t) = \begin{bmatrix} 2e^t - e^{2t} & -e^t + e^{2t} \\ 2e^t - 2e^{2t} & -e^t + 2e^{2t} \end{bmatrix} \tag{34}$$

and

$$P(t) = \begin{bmatrix} -\frac{1}{2}e^t + \frac{1}{5}e^{2t} + \frac{3}{10}\cos t + \frac{1}{10}\sin t \\ -\frac{1}{2}e^t + \frac{2}{5}e^{2t} - \frac{3}{10}\sin t + \frac{1}{10}\cos t \end{bmatrix}. \tag{35}$$

A set of initial conditions $X(0)$ in (33) can be chosen for which the transient component on the right side of (33) vanishes. The procedure for finding $X_p(0)$ is as follows. Replacing 2π for t in (34) and (35) or applying the results of (5) and (7) we have

$$\Phi(2\pi) = \begin{bmatrix} 2e^{2\pi} - e^{4\pi} & -e^{2\pi} + e^{4\pi} \\ 2e^{2\pi} - 2e^{4\pi} & -e^{2\pi} + 2e^{4\pi} \end{bmatrix} \tag{36}$$

and

$$P(2\pi) = \begin{bmatrix} -\frac{1}{2}e^{2\pi} + \frac{1}{5}e^{4\pi} + \frac{3}{10} \\ -\frac{1}{2}e^{2\pi} + \frac{2}{5}e^{4\pi} + \frac{1}{10} \end{bmatrix}. \tag{37}$$

Substituting (36) and (37) in (20) finally yields

$$\begin{bmatrix} x_{p1}(0) \\ x_{p2}(0) \end{bmatrix} = \begin{bmatrix} 1 - 2e^{2\pi} + e^{4\pi} & e^{2\pi} - e^{4\pi} \\ -2e^{2\pi} + 2e^{4\pi} & 1 + e^{2\pi} - 2e^{4\pi} \end{bmatrix}^{-1} \begin{bmatrix} -\frac{1}{2}e^{2\pi} + \frac{1}{5}e^{4\pi} + \frac{3}{10} \\ -\frac{1}{2}e^{2\pi} + \frac{2}{5}e^{4\pi} + \frac{1}{10} \end{bmatrix} = \begin{bmatrix} \frac{3}{10} \\ \frac{1}{10} \end{bmatrix}. \tag{38}$$

Substituting (38) in (21) and (23) we have the periodic response

$$\begin{bmatrix} x_{p1}(t) \\ x_{p2}(t) \end{bmatrix} = \begin{bmatrix} \frac{3}{10}\cos t + \frac{1}{10}\sin t \\ -\frac{3}{10}\sin t + \frac{1}{10}\cos t \end{bmatrix} \tag{39}$$

p. 399

and the transient response

$$\begin{bmatrix} x_{tr1}(t) \\ x_{tr2}(t) \end{bmatrix} = \begin{bmatrix} 2e^t - e^{2t} & -e^t + e^{2t} \\ 2e^t - 2e^{2t} & -e^t + 2e^{2t} \end{bmatrix} \begin{bmatrix} x_1(0) - x_{p1}(0) \\ x_2(0) - x_{p2}(0) \end{bmatrix} = \begin{bmatrix} \frac{1}{2} e^t + \frac{1}{5} e^{2t} \\ \frac{1}{2} e^t + \frac{2}{5} e^{2t} \end{bmatrix}. \tag{40}$$

The initial conditions $x_{p1}(0)$ and $x_{p2}(0)$ are in fact one point on the oscillatory solution circle in the phase plane, and can be used to generate the oscillatory circle.

Example 2

Consider the following system with a piecewise continuous input of period 4 sec

$$\dot{x} + x = g(t). \tag{41}$$

The first period of input $g(t)$ is as follows:

$$g(t) = t - 2h(t-2) - 2(t-2)h(t-2) + (t-4)h(t-4) + 2h(t-4), \tag{42}$$

where $h(t)$ is a Heaviside function.

The explicit solution of (41) and (42) for the first period of input is

$$x(t) = \Phi(t)x(0) + p(t), \tag{43}$$

where

$$\Phi(t) = e^{-t} \quad \text{and} \quad p(t) = (t - 1 + e^{-t}) - 2(t-2)h(t-2).$$

Evaluating $\Phi(t)$ and $p(t)$ at $t = 4$ yields

$$\Phi(4) = e^{-4} \quad \text{and} \quad P(4) = -(1 - e^{-4}). \tag{44}$$

Substituting (44) in (20) we have the periodic conditions $x_p(0)$ or

$$x_p(0) = [1 - \Phi(4)]^{-1} P(4) = -1$$

and the first period solution is therefore

$$x_p(t) = (t-1) - 2(t-2)h(t-2), \quad 0 \leqslant t < 4,$$

or

$$\begin{aligned} x_p(t) &= t - 1 \qquad t \in [0, 2] \\ &= -t - 3 \quad t \in [2, 4] \end{aligned}$$

and the solution for the other periods is

$$x_p(t) = x_p(t + 4K), \quad 0 \leqslant t < 4, \quad (K = 1, 2, \ldots)$$

and the transient solution is

$$x_{tr}(t) = e^{-t}[x(0) - x_p(0)], \quad 0 \leqslant t < \infty.$$

5.2. *Examples of cut-off response*

Example 1

Consider a harmonic oscillatory system with step-function input as follows :

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} g(t), \tag{45}$$

where

$$g(t) = 1 + h(t-\pi) - 3h(t-2\pi) + h(t-3\pi).$$

For a properly chosen set of initial conditions denoted by $x_c(0)$ the response of (45) vanishes ('cut-off') at $t = 3\pi$. The transition matrix and the zero-state response evaluated at $t = 3\pi$ can be obtained from eqns. (5) and (7). However, for this example the explicit response can be obtained as follows :

$$X(t) = \Phi(t) X_c(0) + P(t), \tag{46}$$

where

$$\Phi(t) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \tag{47}$$

and

$$P(t) = \begin{bmatrix} [1-\cos t] + [1-\cos(t-\pi)]h(t-\pi) - 3[1-\cos(t-2\pi)]h(t-2\pi) \\ \qquad + [1-\cos(t-3\pi)]h(t-3\pi) \\ \sin t + \sin(t-\pi)h(t-\pi) - 3\sin(t-2\pi)h(t-2\pi) \\ \qquad + \sin(t-3\pi)h(t-3\pi) \end{bmatrix}. \tag{48}$$

ORIGINAL PAGE IS
OF POOR QUALITY

We put $t = 3\pi$ in (47) and (48)

$$\Phi(3\pi) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \quad \text{and} \quad P(3\pi) = \begin{bmatrix} -4 \\ 0 \end{bmatrix}. \tag{49}$$

Substituting (49) in (29) yields

$$\begin{bmatrix} x_{c1}(0) \\ x_{c2}(0) \end{bmatrix} = -\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}^{-1} \begin{bmatrix} -4 \\ 0 \end{bmatrix} = \begin{bmatrix} -4 \\ 0 \end{bmatrix}. \tag{50}$$

Equation (50) is the cut-off conditions. With these values and from eqn. (45) we have the cut-off response

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 1 - 5\cos t \\ 5\sin t \end{bmatrix}, \quad t \in [0, \pi],$$

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 2 - 4\cos t \\ 4\sin t \end{bmatrix}, \quad t \in [\pi, 2\pi],$$

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} -1-\cos t \\ \sin t \end{bmatrix}, \quad t \in [2\pi, 3\pi]$$

and

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad t \in [3\pi, \infty].$$

If $x_{c1}(t)$ and $x_{c2}(t)$ are the power absorbed by resistors, then integrating $x_{c1}(t)$ and $x_{c2}(t)$ with respect to t, the total energy consumed in resistors can be evaluated.

Example 2

Consider a time-varying system with a cut-off input as follows :

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \dfrac{1}{t+1} & -\dfrac{t}{t+1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t), \tag{51}$$

where $u(t) = (t+1)[h(t) - h(t-1)]$ which cuts off at $t = 1$ sec. The numerical values of the transition matrix and that of the zero-state response evaluated at $t = 1$ can be obtained from (5) and (7) or

$$\Phi(1) = \begin{bmatrix} 1+e^{-1} & 1 \\ 1-e^{-1} & 1 \end{bmatrix} \quad \text{and} \quad P(1) = \begin{bmatrix} 1-e^{-1} \\ 1+e^{-1} \end{bmatrix}. \tag{52}$$

Substituting (52) in (29) we get the cut-off conditions

$$\begin{bmatrix} x_{c1}(0) \\ x_{c2}(0) \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

The cut-off response is therefore

$$\begin{bmatrix} x_{c1}(t) \\ x_{c2}(t) \end{bmatrix} = \begin{bmatrix} (t-1)^2 - (t-1)^2 h(t-1) \\ 2(t-1) - 2(t-1) h(t-1) \end{bmatrix}, \quad t \in [0, 1] \tag{53}$$

and

$$\begin{bmatrix} x_{c1}(t) \\ x_{c2}(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad 1 \leqslant t < \infty.$$

6. Conclusion

Two problems are discussed in this paper. When the periodic input is continuous or piecewise continuous, the complete response of a linear time-invariant or of a class of time-varying system can be decomposed into two components : the periodic response and the transient response. The periodic conditions can be immediately obtained from the developed formula and

p. 402

numerical computation. In the phase plane the periodic conditions developed in this paper turn out to be a point to generate the oscillatory circle in the phase plane if the linear system has a closed trajectory.

When the input is continuous or piecewise continuous and vanishes at time t_1, the cut-off conditions can be obtained by the derived formula and numerical computation. Thus the corresponding cut-off response can be obtained. Since the closed curve will be obtained by the cut-off conditions chosen, then energy or power for a linear system may be evaluated.

References

Blackman, R. B., 1961, *I.R.E. Trans. Circuit Theory*, **8**, 371.

Brand, L., 1966, *Differential and Difference Equations* (John Wiley), p. 58; 1968, *Archs ration. Mech. Analysis*, **5**, 365; 1969, *Ibid.*, **3**, 238.

Davison, E. J., and Smith, H. W., 1971, *I.E.E.E. Trans. Circuit Theory*, **18**, 181.

Miyata, F., 1964, *I.E.E.E. Trans. Circuit Theory*, **11**, 283.

Susman, L., 1969, *I.E.E.E. Trans. Circuit Theory*, **16**, 546.

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: Equipment for High Pressure Infrared Absorption, PVT, and Configurational Properties of Liquids - Measurements of Cis-Pentene-2 and Mixtures

(B) Project Abstract:

The continued study of the liquid state as presented in this work includes the development of a Perkins type PVT apparatus for measurements up to 10,000 atmospheres, a semiconductor strain gauge pressure transducer, and a 0-1500 atmospheres variable path length IR cell. PVT measurements were made and correlated, and certain configurational thermodynamic properties calculated.

The strain gage pressure transducer measures pressure with an accuracy of 5.5 parts per 10,000 at 10,000 atmospheres. The Perkins PVT apparatus determines volume with a maximum uncertainty of 3-4 parts per 1000.

The PVT data covers the range 0-6000 atmospheres and 24 to 2000°C. Measurements were made on cis-pentene-2, and mixtures of cis-pentene-2 and carbon tetrachloride, and cis-pentene-2 and chloroform. A nonlinear least squares method, which included weighting factors, was used to correlate the data by means of the Tait-Whol equation of state. The cyclohexane, 2,3 dimethylbutane, and n-hexane data of Shaver were also refit. It was observed that data determined with the

Beattie PVT apparatus could not be fit by the Tait-Wohl equation to the limit of the precision of the data, 3-5 parts per 10,000, suggesting that further improvements in the equation are required.

A procedure for calculating the configurational properties of liquids was extended and used to determine the configurational energy and entropy of aforementioned systems.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1968

(E) Department: Chemical Engineering

(F) Student Name: Jerry Van Fox

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: Mathematical Evaluation of a Dynamic Method for Determining V-L Equilibria

(B) Project Abstract:

A dynamic method proposed by Prengle and Curtice to determine the vapor-liquid equilibrium data at constant pressure for a binary mixture using only the initial composition and the temperature-liquid weight data has been explored and proved feasible. The method obviates the necessity of having to determine experimentally the liquid or vapor composition.

By combining the equations proposed with the Rayleigh equation an integral equation is produced. The value of the integral can be calculated from the liquid weight at any instant and intially. The lower limit of the integral is the initial weight fraction of the more volatile component and which is a known value. The upper limit is the weight fraction at any instant which is to be determined. A computer program was written using Gauss' quadrature numerical method for solving the integral equation.

Ramalho and Tiller's experimental data for Rayleigh distillation of methanol-water system at 760 mm Hg, including measured liquid composition data, were used to check the method. The results obtained indicate good agreement.

Extension of the dynamic method into an experimental

procedure is discussed.

The average excess enthalpy was determined from literature experimental constant pressure V-L data and found to be in good agreement with individually measured excess enthalpy data.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1969

(E) Department: Chemical Engineering

(F) Student Name: Ling-Kun Huang

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: Kinetics of Djels-Alder Reactions Preliminary Work on the Prediction of Rate Constants by Transition State and Molecular Orbital Methods

(B) Project Abstract:

Selected data on various Diels-Alder reactions in the gas and liquid phases were analysed by the transition state theory, and equations for roughly estimating the enthalpy and entropy of activation were obtained for these. The activation enthalpy, ΔH^*, was found to be the significant variable in comparing reaction rate constants, k, between similar reactions, catalyzed or uncatalyzed.

Solvent effects on reaction rates and possible correlation methods were reviewed. The effect was found to be small in most cases and subject to correlation by regular solution or dipole-interaction methods.

The activation enthalpy was broken up into various parts, the significant variable being a π-electron energy change ΔE_{π}, which can be roughly evaluated from perturbational molecular orbital calculations.

A molecular orbital program based on the Hückel method was developed and parameters for calculations on 1,3-butadiene and acrolein derivatives were evaluated by comparing calculated physical properties against experimental

ORIGINAL PAGE IS
OF POOR QUALITY

values. The parameters were used in perturbation calculations and values for $\Delta E_{\pi}/\gamma^2$ were evaluated for different reactant pairs, γ being an interaction term. Plots of ΔH^* or RTlnk vs. $\Delta E_{\pi}/\gamma^2$ were found to be linear for various sets of reactions.

Simplified correlations of ΔH^* or RTlnk vs. some calculated properties were found to give definite trends.

The influence of an electron-withdrawing effect at the polar group of the dienophile was studied. It was concluded that a likely explanation for the reduction in ΔH^* during catalysis lies in a reduction of the π-electron delocalisation and dipole-induction energies.

A combination of the transition state and molecular orbital methods was thus found to be a suitable starting point toward predicting rate constants in Diels-Alder systems.

Suggestions for future work toward a more coherent approach are presented.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1970

(E) Department: Chemical Engineering

(F) Student Name: Dhiren Bhailal Patel

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

Project

(A) Project Title: Measurement of Infrared Absorption Coefficients of Pollutant Gas Species

(B) Project Abstract:

Quantitative analysis of gas mixtures by means of infrared spectroscopy must be based on a knowledge of the absorption coefficients, k_ν, of the species involved at the actual conditions of optical path length and temperature of the sample.

For the majority of pollutant gases, information about these coefficients is scarce or incomplete and the objective of this work was to determine such coefficients for the following pollutants, CH_4, C_2H_4, CO_2, CO, SO_2, NO_2, NO, H_2S. Measurements were made at room temperature for certain of their wavelengths, and the absorption coefficients k_ν, determined over the widest possible range of optical path length, pl. The wavelengths chosen were those free from interference with absorption by water vapor.

The validity of Beer's law was confirmed for a limiting value of pl, which was different for each wavelength. A correlation between k and pl was made of the following form,

$$\ln k_\nu = \ln k_{\nu o} \left[\sum_{i=0}^{m} C_i (\ln pl)^i \right]$$

p.410

ORIGINAL PAGE IS
OF POOR QUALITY

For three gases, C_2H_4, CO_2, SO_2 at one characteristic wavelength, $k_{\tilde{\nu}}$ was determined as a function of temperature, at 125°C, and 200°C. A correlation was found to be

$$\ln k_{\tilde{\nu}} = a + s \ln (T/T_o)$$

Combining the two correlations of $k_{\tilde{\nu}}$ as a function of pl and T the following form results,

$$\ln k_{\tilde{\nu}} = \ln k_{\tilde{\nu}o} \left[\sum_{i=0}^{m} C_i (\ln pl)^i \right] + s \ln(T/T_o)$$

and permits computation of $k_{\tilde{\nu}}$ at any desired condition.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1971

(E) Department: Chemical Engineering

(F) Student Name: Paolo Campani

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

Project

(A) Project Title: Configurational Properties of Liquids

(B) Project Abstract:

A project has been underway in the Chemical Engineering Department to investigate properties of liquids and solutions; the work presented in this thesis is one part of the program. The objective was to determine the configurational thermodynamic properties of certain substances as a function of liquid density and molecular size and shape, which could be used later in theoretical molecular models for, 1) the configurational properties of pure components as a function of density, and 2) the excess thermodynamic functions for species in solution.

Fifteen nonpolar and polar substances were chosen covering a wide range of molecular weights and acentric factors:

Hydrocarbons: methane, cis-pentene-2, cyclohexane, benzene, n-hexane, 2,3 dimethylbutane, 2,2,4 trimethylpentane, n-octane, n-decane.

Hydrogen bonded substances: water, methyl alcohol, isopropyl alcohol.

Others: argon, nitrogen, carbon tetrachloride.

Using the approach of molecular statistical thermodynamics, a "sun of contributions" method was employed, involving translation, external rotation, internal vibration

plus rotation, and intermolecular configuration; also the total property can be visualized as made up of two parts: one temperature dependent and the other density dependent.

The results obtained for the configurational energy indicate a very decided effect of molecular size and shape for all classes of substances, and similarly for the configurational entropy; however for the latter, the trends are somewhat obscurred, and not as apparent as for the energy.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1971

(E) Department: Chemical Engineering

(F) Student Name: Kenneth Earl Bush

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: Measurement of Infrared Absorption Coefficients for Certain Pollutant Gases

(B) Project Abstract:

Concentrations of air pollutants in the atmosphere can be determined by long path absorption spectroscopy, and by remote emission spectroscopy, but require a knowledge of the absorption coefficients k_ν as a function of optical path length and temperature T. Literature search revealed that such information is scarce and incomplete for most pollutant gases. In a previous investigation absorption coefficients were determined at room temperature for CH_4, C_2H_4, CO_2, CO, SO_2, NO_2, NO and H_2S, and at elevated temperatures for C_2H_4 and SO_2. The objective of this work was to extend the measurements at elevated temperatures for CO_2, CH_4, NO, NO_2, and CH_3CHO, and at room temperature for CH_3CHO. Data from both investigations were compatible with the Bouguer-Beer Law for a limiting range of the optical path length x, and for larger values of x, k_ν decreased, and was correlated as a function of x and T. A method was developed and used to correct for the finite spectral slit width which has considerable effect on the experimentally determined absorption coefficient.

A model was proposed for spectral band absorption which reduces to the Square Root Absorption Law for large

ORIGINAL PAGE IS
OF POOR QUALITY

values of x. The correlation of the k_ν values indicated that the above law is a very rough approximation of the trend of actual data.

A method was developed and used to calculate the Einstein Transition Probability from the k_ν values.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1972

(E) Department: Chemical Engineering

(F) Student Name: Uday Mahagaokar

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

Project

(A) Project Title: Exploration of Intermolecular Effects in Liquids by Infrared Spectroscopy

(B) Project Abstract:

Intermolecular effects in liquids, field potential and decay of rotation, were explored by means of infrared spectroscopy.

The shift and broadening of the C=C stretching band of cis-pentene-2 were studied from 473°K to 104°K, covering the liquid from the critical point to the triple point, and some measurements in the solid. Pure cis-pentene-2 and in solution with four different solvents: n-pentane, toluene, diethyl ether, and acentronitrile, were studied.

The shift in the vibrational frequency was related to the second derivative of the intermolecular function by a mathematical model, and used to obtain the field potential, $\bar{z}U/k$ as a function of density, and the parameters in a potential function, A, γ, and $\bar{z}_m \varepsilon_o/k$. Assuming $\bar{z}_m \simeq 14$, ε_o/k values are in reasonable agreement with other experimental values.

The vibrational-rotational band width was related to the rotational energy of the molecule. It was found that the rotational energy of cis-pentene in all five systems decayed as a function of the density ρ, according to the relationship.

p. 416

$$E_{rot}/RT = (3/2)(1 - a\rho/\rho_o)^2$$

decreasing to approximately 10% of its fully developed value at the triple point.

The decay in translational and internal mode energy was also estimated from the data and found to be substantial.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1974

(E) Department: Chemical Engineering

(F) Student Name: Stanley Curtice

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

ORIGINAL PAGE IS
OF POOR QUALITY

Project

(A) Project Title: Equilibrium Theory of Liquids - An Exponential Perturbation Theory

(B) Project Abstract:

In this work, the equilibrium theory of liquids, molecular correlation functions, radial distribution function, and perturbation theories are reviewed and discussed. A new exponential perturbation theory was developed to calculate the radial distribution function and the thermodynamic properties of liquids.

A function called the potential difference function is defined as the difference of the molecular interaction potential of average force, and expanded into a Taylor series of the reciprocal of the temperature. By a topographical reduction method, an integral equation is obtained for the function. The integral equation provides a generalization of the perturbation, and the various possibilities of approximation were analyzed. With certain assumptions, the integral equation can be solved by knowing the interaction potential and the hard sphere radial distribution function. A simple procedure was proposed to correlate the hard sphere diameter as a function of temperature and density.

The theory was tested against the computer simulation results of Verlet for a Lennard-Jones (6,12) fluid. The

agreement was excellent, the radial distribution function and thermodynamic properties are reproduced with high accuracy. The computation time is reasonable.

Calculations were also extended to argon, and it was confirmed that the Lennard-Jones (6,12) fluid is a good approximation for argon. With suitable choices of potential parameters, the pressure-volume-temperature relations of liquid argon can be reproduced successfully.

A new cell model was constructed to calculate the Helmholtz free energy as the sum of various contributions: the Helmholtz free energy of the hard spheres, the average cell potential energy and the fluctuation of the cell potential energy. The average cell potential energy was assumed linear with density. The fluctuation term was evaluated from the results of the exponential perturbation theory. The formulation was simple and gave good results for Helmholtz free energy and internal energy. The model might be useful in predicting the excess mixing properties. Compressibility factor calculations were satisfactory.

Suggestions and formulations are given to extend the exponential perturbation theory to polyatomic molecules and mixtures of simple liquids, but no numerical calculations were attempted. Using published data for a hard sphere mixture, the present theory was shown to be better than the existing perturbation theories.

Suggestions for the future work have been made. Results of the present work are to be correlated in terms of hard sphere properties and it is hoped that this correlation will make the perturbation theory feasible for engineering calculations.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1974

(E) Department: Chemical Engineering

(F) Student Name: Tsung Teh Tseng

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

Project

(A) Project Title: Correlation and Measurement of Stack Gas Pollutants by Infrared Absorption

(B) Project Abstract:

Stack gas pollutants can be determined by use of long path absorption spectroscopy for stack samples and by remote emission spectroscopy on stack plumes. However, a relationship between the absorption coefficient and the optical density is required in order to do so. Previous investigations were conducted on CH_4, C_2H_4, CO_2, CO, SO_2, NO_2, NO, and H_2S, but mathematical and experimental errors rendered the results inaccurate. The objective of this work was to improve the previous measurements and extend them to partially oxidized hydrocarbons (HCHO).

Data obtained from this investigation was seen to approach the constant limit of the Beer's Law absorption coefficient at low values of optical density while at high values the absorption coefficients decreased with optical density. A model was proposed for the correlation and a least squares program was used to obtain the correlation model constants for each pollutant.

A stack sampling system was built to obtain samples from a boiler furnace. Samples were collected and analyzed using the previously correlated data.

p. 421

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1975

(E) Department: Chemical Engineering

(F) Student Name: Kerry Lynn Rock

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

p.422

Project

(A) Project Title: Molecular Thermodynamics - Exploration of Translational and Rotational Decay in Condensed Phase

(B) Project Abstract:

In this work the Nearest Neighbor-Distribution-Fluctuation Cell model was improved and extended, and the decay of the dynamical modes - rotation and translation - in condensed phase was investigated.

For the substances, argon, methane, cyclohexand and cispentene-2 a wide range of conditions was studied, from perfect gas state at the critical temperature down to the $0^{o}K$ solid. It was found that the external potential field not only influences configurational properties, but also changes the dynamical behavior of molecules.

The configurational properties are determined by the cell potential, a value evaluated from the minimum energy (at $0^{o}K$), and the average number of effective nearest neighbors, a function of density. The average number of effective nearest neighbors was found to have a negative deviation from a linear relationship. The deviation is larger for more complex molecules and was well correlated by a mass-dispersion parameter of molecule, the acentric factor.

A parameter called the normalized maximum kinetic energy was defined for the dynamical modes, and is used in

the partition function integral to obtain the effect of decay on all properties.

It was found that translation decays in some condensed phases. For argon and methane, translation is essentially fully developed and maintains approximately 60 - 70% developed at triple-point solid. For more complex molecules, the decay of translation is larger and is proportional to the mass-dispersion of the molecule.

The decay of non-preferential rotation for spherical-top molecules is very similar in magnitude to that of translation. The decay of preferential rotation was represented by superimposing the effects of rotational moments, the ratio of moment of inertia to the smallest one, on that of non-preferential rotation.

A generalized theory for all modes was constructed to provide a method for prediction of the thermodynamic properties. Ethane was chosen as a test substance for the proposed models and the agreement with experimental data was quite satisfactory.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1976

(E) Department: Chemical Engineering

(F) Student Name: Eddy Chi-Hua Sun

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

Project

(A) Project Title: Remote Sensing of Temperature and Composition of Gas Plumes by Infrared Emission Interferometry-Spectroscopy

(B) Project Abstract:

Quantitative aspects of the technique of remote sensing of stationary sources by infrared emission spectroscopy have been studied in this investigation. A method is developed for independently determining the temperature and composition of the remote gas plume from data collected with a rapid scan interferometer-spectrometer system. The analysis accounts for atmospheric attenuation caused mainly by CO_2 and H_2O background radiation originating from various sources in the background of the plume. The background radiation factor has been semiempirically correlated for application to transparent and opaque targets.

Measurements were made, using the above technique, to observe temperature gradients and fluctuations in an emerging plume in a generally calm atmosphere. The axial profile data indicated a decay in the mean plume temperature from 507°K at the source to 425°K within three meters downstream from the source. The experimental data showed that near the stack, the axial gradient is small, whereas Priestley's theoretical model predicts a large gradient. The data does not agree with Priestley's model because the

temperature decay near the stack is primarily dependent on the temperature and energy content of the plume at the point of emergence and the effect of the surrounding fluid is neglible. Further downstream, turbulent diffusion and air entrainment become appreciable causing the temperature to decay more rapidly. The experimental data when extrapolated along the theoretical model predicted a decay of the mean plume temperature to near ambient levels within 45 meters from the source. Radial temperature data were also taken and showed an approximate Gaussian Profile. Temperature fluctuations of the order of $\pm$ 25°K were observed. Intensities of turbulent temperature fluctuations were determined axially and radially and compared to existing data on laboratory scale jets.

The remote sensing technique has been developed for quantitative determination of the pollutants - CO, NO, NO_2, CH_4, C_2H_4, HCHO, CH_3CHO, C_2H_6, H_2S and SO_2. Field measurements were made at 68 meters distance on gas fired power plant plumes. NO_2 and HCHO were consistently found while small quantities of CO and CH_4 were also detected under normal furnace operation. A sharp rise in the CO and CH_4 concentrations was detected when the excess air in the furnace was lowered.

The remote temperature measurements were found to be accurate to within $\pm$ 5°K. Concentration results over a large number of runs showed good consistency. NO_2 and HCHO

concentrations showed a per cent deviation of 21% each. The per cent deviations for CH_4 and CO concentrations were 76% and 51% respectively; a large part of these deviations were due to natural fluctuations in the normal operation of the furnace.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1976

(E) Department: Chemical Engineering

(F) Student Name: Uday Mahagaokar

(G) Faculty Advisor: Professor H. W. Prengle, Jr.

Project

(A) Project Title: Biofeedback Training of 40Hz EEG and Behavior

(B) Project Abstract:

A specific pattern of brain electricity, a narrow frequency band centering at 40Hz, reflects a state of circumscribed cortical excitability or focused arousal which is "optimal" for consolidation in short-term store. On-line control procedures have been developed to reliably record and digitally count this low-amplitude EEG activity independent of muscle artifact.

A high degree of operant control of the 40Hz EEG can be achieved by biofeedback training in both conditioning and suppression. In control testing sessions following conditioning and suppression training, some degree of voluntary control has been demonstrated when subjects alternately turned the 40Hz on and off only upon instructional sets. The generality and stability of relationships shown between the conditioned 40Hz EEG and problem-solving behavior requires further systematic verification.

(C) Publication: Behavior and Brain Electrical Activity

(D) Year: 1974

(E) Department: Psychology

(F) Faculty Advisor: Professor Daniel E. Sheer

Reprinted from: BEHAVIOR AND BRAIN ELECTRICAL ACTIVITY
Edited by N. Burch and H.L. Altshuler
Book available from: Plenum Publishing Corporation
227 West 17th Street, New York, New York 10011

biofeedback training of 40-hz eeg and behavior

ORIGINAL PAGE IS
OF POOR QUALITY

Daniel E. Sheer, Ph.D.

Department of Psychology
University of Houston
Houston, Texas

The use of operant training techniques to control electrical brain activity is relatively recent, dating back some ten years to the beginning of the present series of publications (Mulholland 1968). As is generally true of new research, there is a focus on questions about the basic process itself—definitive training procedures, transfer and concomitant effects and, not the least of it, skepticism about the reliability of the phenomenon itself. The ever-accelerating literature reflected, in part, in a handbook (Barber et al. 1971a), in a recent bibliography (Butler and Stoyva 1973), and in a series of annual reviews (Barber et al. 1971b, Stoyva et al. 1972, Shapiro et al. 1973), has both clarified some issues and raised additional questions.

In spite of an amorphous surround of "mind control" distorting the question, it is now clear that patterns of electrical activity in the brain can be brought under operant control through response-reinforcement contingencies. This conditioning has been demonstrated in humans through a range of different EEG patterns: theta (Green, Green, and Walter 1972; Beatty et al. 1974), alpha (Brown 1970, Nowlis and Kamiya 1970, Mulholland and Peper 1971), a sensorimotor rhythm at 12 to 14 Hz (Sterman 1972), and beta (Beatty 1971). There is also some indication that asymmetrical control can be achieved with occipital alpha by differential feedback to homologous scalp areas of the two hemispheres (Peper 1971, 1972). Such localized control raises the possibility of more specific training, because the functional

1.428-a

significance of laterality has also been demonstrated with EEG measures. Galin and Ornstein (1972) recorded from normal subjects during the performance of predominantly right (spatial) and left (verbal) hemisphere tasks. They found a significantly higher average spectral power (more alpha) in the right hemisphere, and thus more alpha blocking and beta in the left hemisphere, during performance of the verbal as compared with the spatial tasks.

Systematic information on optimal training procedures and conditions under which operant control will generalize beyond the training laboratory is simply not available. Among some relevant observations, Black (1971) has shown that peripheral skeletal muscle mediation is not an essential condition for operant control of CNS electrical activity. Hippocampal theta and non-theta waves can be conditioned in dogs whose skeletal musculature was paralyzed by gallamine. The possibility still remains, however, that hippocampal theta waves may be related to central circuits that control skeletal muscle activity because such drugs do not block central circuits. Indeed, rats who are free to move can be more quickly conditioned to theta than can immobile rats (Black 1972, and personal communication). This would imply a relationship between the circuitry for skeletal movement and hippocampal theta and/or demonstrate the importance of response state for conditioning theta.

At the human level the problem of mediative processing, particularly cognitive, becomes more difficult to dissect. Beatty (1972) compared alpha conditioning during one-hour training periods in one group of subjects who received feedback on EEG response-reinforcement contingencies, in another group who had only pretrial information on the behavioral state associated with the required alpha responses, in a third group who received both, and in a fourth group who received neither. Groups one, two, and three showed exactly the same magnitude and development of operant conditioning as compared with the fourth group, which showed no change. Prior information evidently produced the same effects as response-reinforcement contingencies, and a combination of these made no difference. Upon questioning, subjects in the "information only" group readily reported the typical correlates of alpha—relaxation, calmness, etc.—while the "feedback only" group gave a wide range of subjective reports. Apparently, in the absence of an external feedback system, the "information only" group of subjects cognitively monitored their own internal states to reinforce the required performance.

Subtle reinforcers can be established in human subjects, through cognitive processing, by the history of the subject, by instructional sets, and by momentary motivational states of the subject—to specify but a few conditions. Obviously it would be useful to have systematic data on cognitively mediated reinforcement and, when performance deteriorates for no apparent

1.428-1

reason, on nonreinforcement as well. On the positive side, this cognitive processing at the human level may inevitably accompany neural response-reinforcement contingencies. It may thus provide a more effective method for achieving voluntary control, as compared with such other methods of control as reinforcement of overt behaviors, drugs, or direct stimulation (Black 1972).

It is difficult to demonstrate clear connections between changes in patterns of brain electricity and changes in concomitant behaviors, but there are encouraging signs. The first approach to the problem was subjective reports, in which alpha states were variously described as "pleasant feelings" (Brown 1970) and relaxing and "letting go" (Nowlis and Kamiya 1970). Mulholland and Peper (1971) associated alpha with passive observation of clearly visible targets and its attenuation with visual efferent control processes concerned with orienting and tracking. Galin (personal communication) recorded from symmetrical right and left central leads referred to the vertex while subjects were performing a mirror-tracking task that required visual spatial abilities. From computer analysis of the left-right ratios of EEG alpha he consistently found a comparative increase of alpha in the right (spatial) hemisphere in the 0.25-second period preceding the occurrence of an error.

An extensive series of studies by Sterman and his colleagues (Sterman, MacDonald, and Stone 1974) focused on a 12- to 14-Hz EEG rhythm recorded from sensorimotor cortex (SMR) in cats and man, which is associated with relaxation and inhibition of movement. The operant conditioning with cats showed that SMR-trained animals had enhanced EEG sleep-spindle activity, reduced motor disturbances during sleep, and increased resistance to seizures induced by convulsant doses of monomethyl hydrazine (Sterman 1972). Subsequently, seven human subjects—four epileptic and three normal—were trained for 6 to 18 months. After two to three months of regular and continuous training, the four epileptic patients began to show reductions in abnormal EEG signs and seizures, which were sustained throughout the training (Sterman, MacDonald, and Stone 1974).

Theta activity recorded from cortical leads has been associated with a subjective state somewhere between the relaxed wakefulness of alpha and the sleep state of delta (Green, Green, and Walter 1972). Subjective reports by persons trained in theta, include descriptions of hypnagogic-like imagery and reverie or wandering imagination which might be characterized as creative thought patterns. Beatty et al. (1974) investigated the effect of theta states associated with a very low level of arousal on a monotonous visual monitoring task requiring a high level of vigilance to maintain efficiency. They trained one group of subjects to augment occipital theta waves and another group to suppress them. Then both groups performed the contin-

ORIGINAL PAGE IS
OF POOR QUALITY

uous monitoring task in which contingent reinforcement was given for the group-appropriate response. The conditioned theta augmentation, also reinforced throughout the task, produced a significant deterioration in monitoring efficiency, while theta suppression produced a significant increase.

The substantial relationships shown in these studies between brain electricity and behavior are restricted entirely to inhibitory behavioral functions. The SMR rhythm is associated with relaxation and inhibition of movement. The alpha rhythm shows significant comparative increases in one hemisphere during a time when it is not maximally operative—the right hemisphere during verbal performance and the left during spatial performance. Again, the alpha shows significant comparative increases in the functional hemisphere—the right hemisphere during mirror-tracing performance—but only at the time when an error is made. The augmentation of occipital theta significantly depresses vigilant performance in a visual monitoring task, which shows improvement when the theta is suppressed.

For all these behaviors the following questions may well be asked. What are the patterns of brain electricity associated with facilitatory behavioral functions? What rhythm occurs in the sensorimotor cortex with facilitation of movement, and not inhibition? What waves show up in the right hemisphere during performance of spatial tasks, and in the left during verbal performance? What happens in the right hemisphere with mirror tracing during correct responding and not when an error is made? What brain electricity is in the occipital cortex when theta is suppressed with concomitant monitoring efficiency?

The first approximate answer is that the EEG is desynchronized at a very low amplitude with mixed fast frequencies. With the recording techniques and computer resolution now available, it is time that we took a much closer look at this low-amplitude, fast-frequency "desynchronized" EEG usually represented in charts as irregularly thickened black lines. The single designation, "desynchronized" or "arousal," for an EEG clearly refers to a number of different electrical patterns. In their study of conditioning in monkeys, Morrell and Jasper (1956) found that a generalized desynchronization was diffusely present in the cortex only during the first stage of sensory-sensory conditioning. When conditioning was established, a "stable, well-localized desynchronization" was limited to relevant cortical areas. Further, Morrell (1961) reported differences in units recorded from the brainstem reticular formation, hippocampus, and visual cortex during generalized cortical desynchronization as compared with the localized desynchronization in visual cortex.

The first stage of conditioning, diffuse cortical desynchronization, represents initial responding to novel stimuli within the complex matrix of irrelevant environmental stimuli. It is an oscillatory, unstable state of the

J. 428-d

organism, in which many different subassemblies of the intrinsic electrical activity are firing in different spatiotemporal patterns, that is, nonsynchronously. When connections become established, through spatiotemporal patterning of inputs, as in conditioned stimulus-unconditioned stimulus (CS-UCS) pairings, subassemblies of the electrical activity now fire in synchronous organizations restricted to the relevant circuitry. The desynchronized EEG is then no longer diffuse, but it still may appear as desynchronized without finer grained analysis because the synchronous subassemblies restricted to limited cortical areas are submerged within the total ongoing electrical activity. The limited subassemblies, defined by the relevant environmental inputs and the contingent reinforcement, are now firing at a synchronous frequency "optimal" for consolidation. We have proposed that a specific pattern of brain electricity, a narrow frequency band centered at 40 Hz, reflects this state of circumscribed cortical excitability or focused arousal (Sheer 1970, Sheer and Grandstaff 1970, Sheer 1972). The association of focused arousal-40 Hz represents an extension of the continuum from sleep-delta, through wakefulness-alpha, and diffuse arousal-beta.

Our focus on the 40-Hz EEG had its beginning in the large-amplitude, highly synchronous bursts recorded from the olfactory bulbs and other rhinencephalic structures of cats during sniffing, exploring, and orienting behaviors (Sheer, Grandstaff, and Benignus 1966). In quadruped animals, particularly, olfaction is an important distance receptor, and the associated motor feedback of sniffing is a highly adaptive orienting response for exploratory, feeding, and sexual behavior. This pronounced electrical rhythm occurs in rhinencephalic structures throughout the phylogenetic scale from catfish to man (Sheer and Grandstaff 1970).

At the neocortical level, where laminar structure is far more complex, the 40-Hz rhythm is at a much lower amplitude in a more complicated electrical background, but it can still be observed visually on the oscillograph from epidural leads at fast paper speeds. For systematic reliable data, however, computer analysis is clearly necessary. In a series of studies with the cat in a successive visual discrimination task (Sheer 1970, Sheer and Grandstaff 1970), consistent relationships were observed between 40 Hz and the acquisition phase of learning. In the 10-sec epoch of a 7/sec flickering light cue (^{S}D), a burst of 40 Hz occurred in visual and motor cortex about 0.5 sec before, and continuously for about 1.5 sec after, a correct bar-press response.

The few references to 40 Hz in the literature present a rather consistent picture. Galambos (1958), recording from the caudate nucleus and globus pallidus, observed 40 Hz when cats had learned that the last in a series of 11 clicks led to an unavoidable electric shock. Rowland (1958), recording from ectosylvian and lateral cortex and medial geniculate nucleus in cats, observed

p. 428-e

ORIGINAL PAGE IS
OF POOR QUALITY

this activity during acquisition, when an auditory CS was paired with an electric-shock UCS. Killam and Killam (1967) also observed it in the lateral geniculate of cats when they were fully trained to discriminate a correct visual pattern from three presented. Pribram, Spinelli, and Kamback (1967) reported 40 Hz in the striate cortex of monkeys just after they made incorrect differential responses in a very difficult visual task. In the Russian literature, Dumenko (1961), recording from the auditory, somesthetic, and motor cortex of dogs, observed 40 Hz when limb responses were conditioned to tone as the CS and excitation of skin with induction current as the UCS. Sakhiulina (1961) observed it in the sensorimotor cortex of dogs when conditioned flexion of the contralateral leg was paired with different conditioned stimuli.

In recording from the scalp in humans, the meditation state provides a unique situation in which subjects are immobile and the recordings relatively free of muscle artifact. Das and Gastaut (1955), recording from occipital leads in seven trained yogis, reported high-amplitude levels of 40 Hz activity during the samadhi state, which is the final, most intense concentration stage in this form of meditation. Just recently Banquet (1973), studying 12 subjects practicing transcendental meditation with recording from left occipital and frontal leads, also observed 40 Hz during the third deep stage of meditation. Giannitrapani (1969), recording from scalp leads in middle- and high-IQ subjects, compared the EEG during mental multiplication activity and a resting condition. A 40-Hz rhythm occurred during the multiplication behavior just prior to the subjects' answering.

Considerable attention has been focused in recent years on a rather heterogeneous clinical grouping variously termed learning disability, minimal brain injury, and minimal brain dysfunction in children. One relatively clear subgrouping of such children can be characterized by no hard neurological signs, no primary sensory or motor defects, no apparent primary emotional disturbances, and low-normal to normal IQ. The main presenting problem is that these children cannot learn; they are either retarded in grade level or are in special classes. They are unable to assimilate new material and solve problems at the level of their chronological peers. Significant decrements were obtained specifically in the 40-Hz EEG band during problem-solving tasks in a carefully selected group of such children, as compared with matched children at normal grade level (Sheer and Hix 1971; Sheer 1974). One purpose of the present series of studies is to develop biofeedback training procedures for conditioning 40-Hz EEG in learning-disabled children.

Experimental Procedures and Controls

Reliable, consistent EEG recording of 40 Hz from the intact scalp is the

p.428-9

first essential procedure to be accomplished, and it is by no means a simple task. This low-amplitude, fast-frequency part of the EEG spectrum, recorded from the human scalp, is of the order of 5 μV. In addition, it completely overlaps with the muscle spectrum, which is broad and highly polyphasic, with generally a peak at 50 Hz (Chaffin 1969). Thus one must contend at the same time with a low signal level and muscle artifact at a much higher amplitude, both in the same frequency range. Over the past several years we have developed recording and analysis procedures that have been used both in the controlled laboratory situation and in such field situations as primary grade schools. The procedures have been control-tested with analog and digital computer analyses and implemented in portable hardware for on-line corrections and digital counts.

Our standard experimental procedure for operant conditioning of the 40-Hz EEG signal in both the field and laboratory situations is as follows: The subject is seated in a lounge chair in a slightly reclining position in front of a screen. His instructions are to turn on as many slides as possible in the conditioning period. The slides are colored, detailed, and cover a wide range of subjects with the interest level pegged for the particular group under investigation, adults or children. The slide-projector is automatically triggered through a stimulus control unit by pre-set criteria of the 40-Hz EEG signal, recorded from a specified set of bipolar leads.

The equipment and experimental set-up for biofeedback training in both the field and laboratory situations are shown in Figures 1A and 1B. In the former situation, the special electrode assembly consists of a Bionetics 10-mm porous silver chloride pellet (Kanter Associates, Santa Anna, Cal.) held in close reference (<¼") to a Texas Instruments TIS 58 Silent Field-Effect Transistor (FET) or Motorola equivalent configurated with a ZN 5089 NPN transistor to provide a current source for the FET and a 33K resistor to limit the current into the FET at 1 mA. This assembly is insulated with Insulex and encapsulated in a mold with No. 8751 epoxilite. The leads from the FET are a 50-ohm coaxial cable about four feet in length.

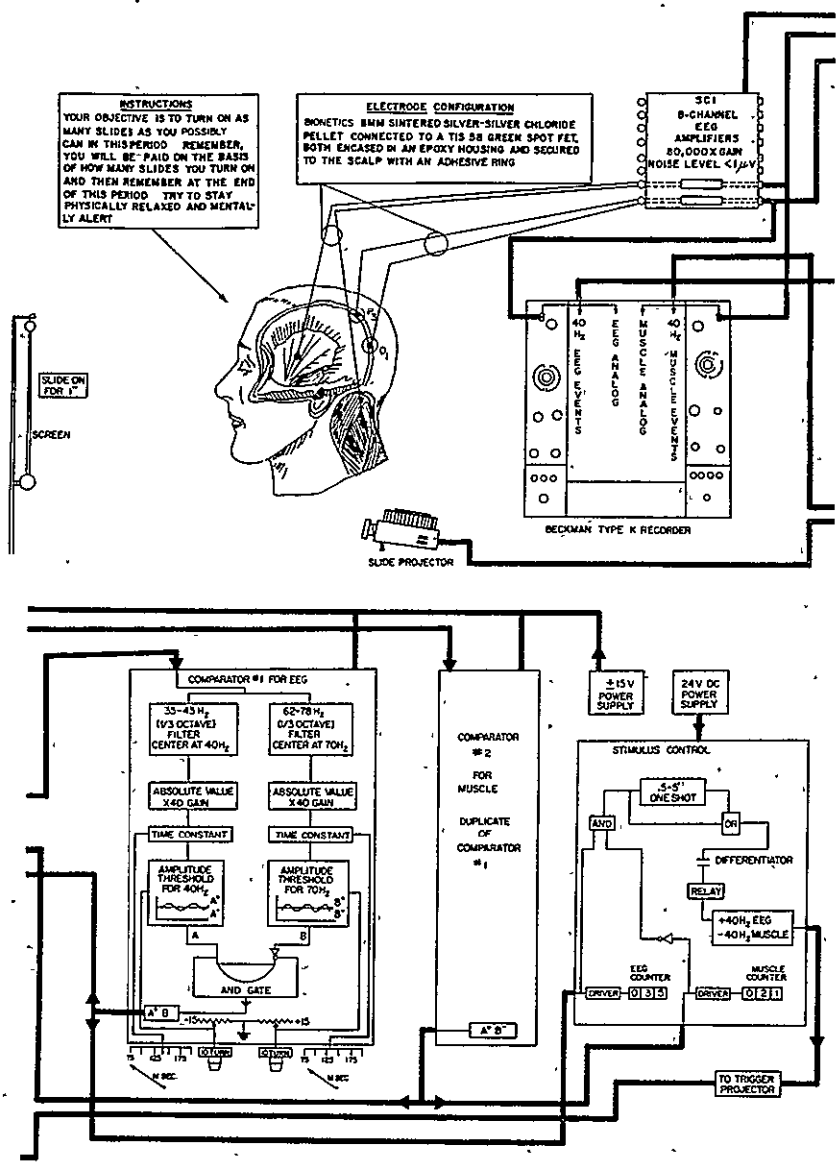
The effect of placing an FET at the immediate electrode site is to provide a low source impedance from the point of the FET of about 300 ohms which produces an excellent shunt to ground for all cable-induced or EMF noise detected along the length of the cable.

The electrode assemblies are applied to the scalp with adhesive rings, which are first attached to the rim of the epoxy mold and then fixed to the scalp. Standard electrode paste is applied between the pellet and the scalp with a syringe needle through holes in the rim of the mold. This electrode assembly fixes firmly to the scalp and can be easily removed by applying acetone.

The electrode leads go into a portable eight-channel differential AC

p.428-8

ORIGINAL PAGE IS
OF POOR QUALITY



Figures 1A and 1B. Equipment and set-up for conditioning 40-Hz in the field situation. The special electrode-configuration changes the source impedance at the scalp to about 300 ohms in transmission over a 50-ohm cable to the first stage of amplification. This configuration does not require shielding and attenuates movement artifact. The special SCI amplifiers are high-gain, low-noise, with narrow-frequency windows. Their outputs from the EEG and muscle leads go on to the two comparators and then to the paper and FM tape recorders. At the same time they also activate the slide projector through the stimulus control unit. In the laboratory situation, the recordings are made with standard Grass electrodes and a 10-channel, Model 78 Grass polygraph with the same comparators and feedback loop.

p.428-2

amplifier assembly at 80,000 gain, with a narrow frequency window between 16 and 80 Hz and a common mode rejection above 100 db (Model MC 128E-4, SCI Systems, Houston, Texas). The power supply for this entire configuration is regulated at ± 15 volts and the internal noise level is of the order of 2 μV. The output from these amplifiers is monitored on-line with a two-channel Beckman Type-K Dynograph.

In the laboratory the recordings are made with standard Grass electrodes and a ten-channel Model 78 Grass polygraph. The Grass amplifiers are set at or close to maximum sensitivity; the high-pass filter cut-off is set at 10 Hz and the low-pass at 90 Hz, with the 60-Hz notch filter cut out. EEG records are monitored on-line on the Grass oscillograph by running the paper speed at 100 mm/sec during critical trial periods or for periodic samples during baseline conditions.

Electrode placements, in both laboratory and field situations, follow the standard Ten-Twenty System; in addition, a set of bipolar leads from the neck and temporal muscles are recorded from the side of the head on which the EEG signal is conditioned. Outputs from the Grass polygraph are stored on a seven-channel FM tape recorder for further computer processing.

Comparators. On-line, the EEG and muscle leads go into identical comparators or coincidence detection units, which are the hardware developed for feedback control of muscle artifact. A schematic drawing of these units is shown in Figures 2A and 2B. Each unit consists of two high Q, narrow-band twin-T analog filters (Model 3385, White Instrument Co., Austin, Texas) with rectified output compared against a DC level to develop a digital output. Both filters have a 23-percent band, one tuned at a center frequency of 40 Hz, the other at a center frequency of 70 Hz. The filter outputs are integrated with adjustable time constants and their threshold levels are set with amplitude comparators.

For the EEG leads an anion gate circuit allows the 40-Hz output to trigger a reinforcement only when it is not coincident with the 70-Hz output, which is used as an index of the polyphasic muscle. In addition, when a 40-Hz muscle signal from the muscle comparator coincides with a 40-Hz EEG signal, the slide projector will again not trigger.

When the output from the EEG comparator is neither coincident with the 70-Hz EEG signal nor with the 40-Hz muscle signal, it activates the stimulus control unit which triggers the slide projector. The outputs from the EEG and muscle comparators also go to digital counters, which keep a consecutive count of 40-Hz EEG bursts and 40-Hz muscle bursts for any specified time period.

The criteria of amplitude levels and burst durations are set by adjustable gain potentiometers and time constants for both the EEG and muscle comparators. The time constants are set at 75 msec, which represents three

1.428-i

ORIGINAL PAGE IS
OF POOR QUALITY

cycles of EEG at 40 Hz and five cycles of muscle bursts at 70 Hz. The amplitude levels are empirically determined for each subject at a preconditioning baseline session and are adjusted to allow for a low to moderate level of free operants. If there is any partial overlap of the three cycles of 40 Hz and the five cycles of 70 Hz from the same EEG leads, the slide is not triggered. This overlap can occur from a minimum of 20 msec before to 20 msec after the 40-Hz filter output because of the different time delays of the 40-Hz filter (54 msec) and the 70-Hz filter (31 msec), and a 40-msec one-shot time delay after the 70-Hz time constant circuit. These stringent criteria are probably an overcorrection for muscle artifact, resulting in a conservative estimate of the EEG signal.

Control procedures. The on-line control for muscle artifact with the comparators is essentially nonparametric contingency detection of the coincidence between EEG and muscle within the threshold limits specified. It is based on a correction for muscle using a parametric analysis of covariance, shown in Figure 3, developed with computer analysis on data obtained with matched groups of normal and learning-disabled children (Sheer and Hix 1971, Sheer 1974).

As can be noted in Figure 3, electrical activity in the range from 62 to 78 Hz with a center frequency at 70 Hz is specified as muscle (Σx^2) and, from the same leads, activity in the range from 36 to 44 Hz with a center at 40 Hz is specified as EEG (Σy^2). The corrected power function

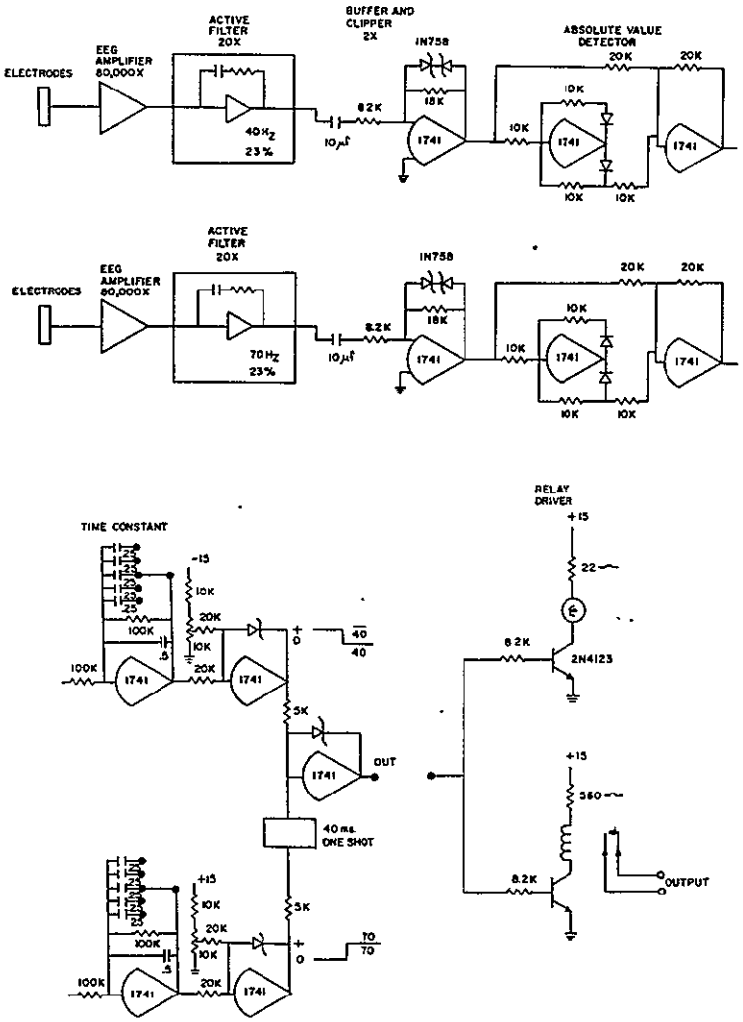
$$\Sigma y^2 = \frac{(\Sigma xy)^2}{\Sigma x^2}$$

represents the variance or power of the EEG independent of muscle. We have carried out a number of independent control checks on this power equation to confirm its independence from muscle.

An analog-computer analysis procedure (Sheer 1970) was used to obtain these corrected spectral power functions for three 23-percent frequency bands, centered at 31.5, 40, and 50 Hz, on normal and learning-disabled children during control and problem-solving situations. Using the corrected power functions, there were significant increases in the 40-Hz power bands for the normal children during problem-solving situations but not in the bordering 31.5 and 50 Hz bands set up as controls (Sheer and Hix 1971, Sheer 1974). There is no reason why polyphasic muscle, with a relatively higher amplitude at 50 Hz, should show up differentially in the 40-Hz band but not in the 31.5- and 50-Hz bands.

Using a hybrid-computer analysis procedure with an IBM 360, high-resolution spectral-density functions were obtained with a Fast Fourier Transform program modified to provide covariance power functions. At the

p.428-j



ORIGINAL PAGE IS
OF POOR QUALITY

Figures 2A and 2B. Schematic of the comparators used in the conditioning set-up to control for muscle artifact. The EEG leads to be conditioned go into one comparator and the muscle leads into another. Each comparator splits the input signal into a 40-Hz and a 70-Hz output with adjustable gain levels and burst durations, which are empirically determined for each subject. The EEG comparator will trigger the stimulus control unit; that is, count a 40-Hz burst and turn on the slide projector only when there is not a coincident 70-Hz burst. Also the muscle comparator and EEG comparator are connected so that when there is a 40-Hz muscle burst coincident with a 40-Hz EEG burst, the stimulus control will not trigger. The time constant or burst duration for both the EEG and muscle comparator is set at 75 msec, which represents three cycles of EEG at 40 Hz and about five cycles of muscle at 70 Hz. With a time delay in the 23% 40-Hz filter of 54 msec and in the 23% 70-Hz muscle filter of 31 msec, and a 40-msec one-shot after the time constant circuit, the 40-Hz EEG signal will not trigger the stimulus control unit if the 70-Hz muscle signal occurs from a minimum of 20 msec before to 20 msec after the 40-Hz filter output.

1.428-k

first stage of this analysis, during digitization of the analog signals, bursts of high-frequency components were automatically blanked out at pre-set levels by detecting the slope of the line of baseline crosses as the amplitude of the first derivative. The amplitude levels for blanking were empirically determined for each record and pre-set for automatically digitizing the analog EEG by balancing the maximal blanking of high-frequency bursts with the minimal effect on frequencies of interest. This procedure was the first step in the hybrid processing, primarily to keep the standard deviations of the EEG distributions within a homogeneous range for the subsequent covariance analyses. It is a gross correction, analagous to deleting obvious muscle bursts by visual inspection.

STATISTICAL CONTROL OF MUSCLE

x = Observed Muscle (70 H_z Filter)

y = Observed EEG (40 H_z Filter)

y' = Predicted EEG from x

y - y' = Difference between Observed and Predicted EEG
---- the EEG Independent of Muscle

$(y-y')^2$ = Power Function

$$y' = \beta x$$

$$y - y' = y - \beta x$$

$$(y-y')^2 = (y-\beta x)^2$$

where $\beta = r\frac{\sigma_y}{\sigma_x}$; $r = \frac{\mathcal{E}xy}{\sqrt{\mathcal{E}x^2 \cdot \mathcal{E}y^2}}$; and $\frac{\sigma_y}{\sigma_x} = \frac{\sqrt{\mathcal{E}y^2}}{\sqrt{\mathcal{E}x^2}}$

then $\beta = \frac{\mathcal{E}xy}{\sqrt{\mathcal{E}x^2 \cdot \cancel{\mathcal{E}y^2}}} \cdot \frac{\cancel{\sqrt{\mathcal{E}y^2}}}{\sqrt{\mathcal{E}x^2}} = \frac{\mathcal{E}xy}{\mathcal{E}x^2}$

therefore $\mathcal{E}(y-y')^2 = \mathcal{E}y^2 - \frac{(\mathcal{E}xy)^2}{(\mathcal{E}x^2)^{\cancel{2}}} \cdot \frac{\cancel{\mathcal{E}x^2}}{1}$

$$= \mathcal{E}y^2 - \frac{(\mathcal{E}xy)^2}{\mathcal{E}x^2}$$

Figure 3. Corrected power function for the 40-Hz EEG signal is shown at bottom line. It is essentially a covariance analysis, in which the variance of the errors of estimate are determined for the 40-Hz frequency band when the spectral power functions are computer-analyzed.

p.428-l

A high-resolution print-out, shown in Figure 4, compares the corrected spectral-density function for an EEG signal with a spectral function for concurrent muscle. At one-half Hz resolution, it was possible to show clear differences in the spectral density distributions for EEG and muscle when the corrected power function was used for the EEG signal (Sheer 1973).

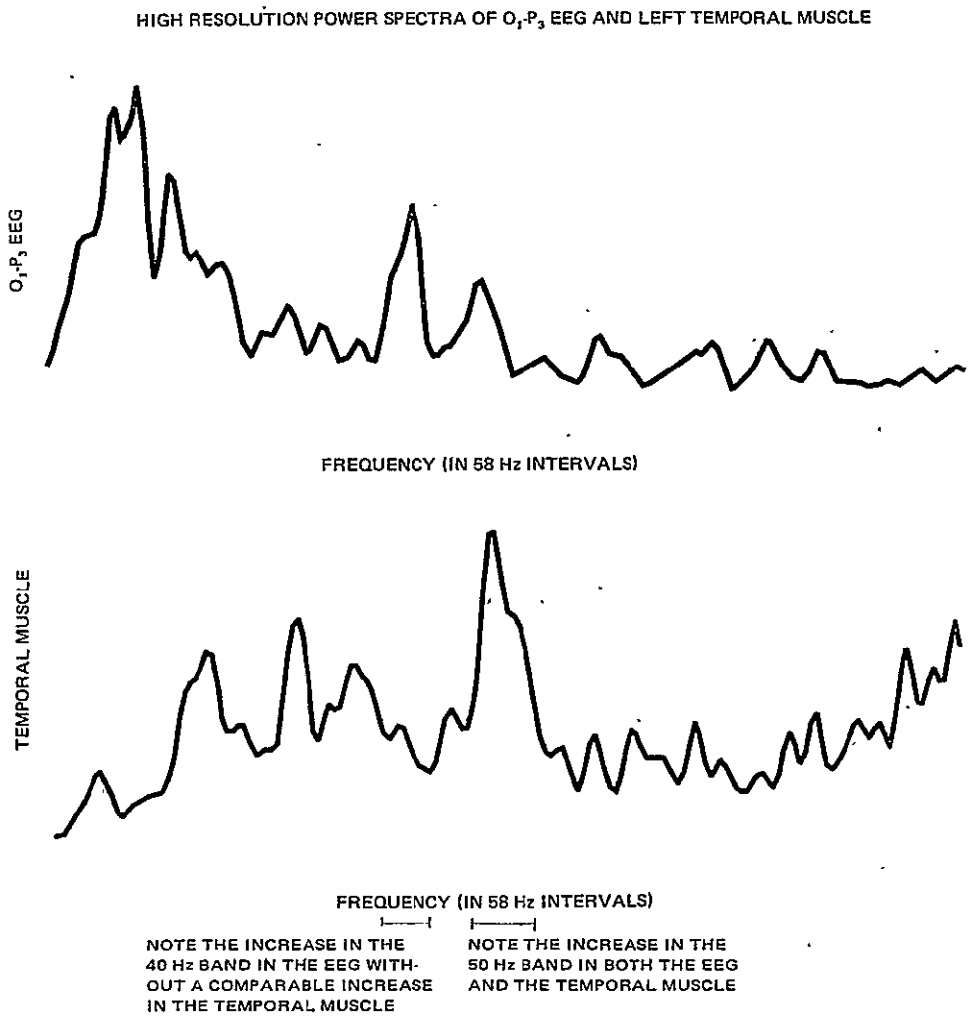


Figure 4. High-resolution spectral-density function print-out of concurrent EEG and muscle activity using the FFT digital-computer analysis. Ordinate is relative power; abscissa is the frequency spectrum in one-half Hz intervals.

1.428-m

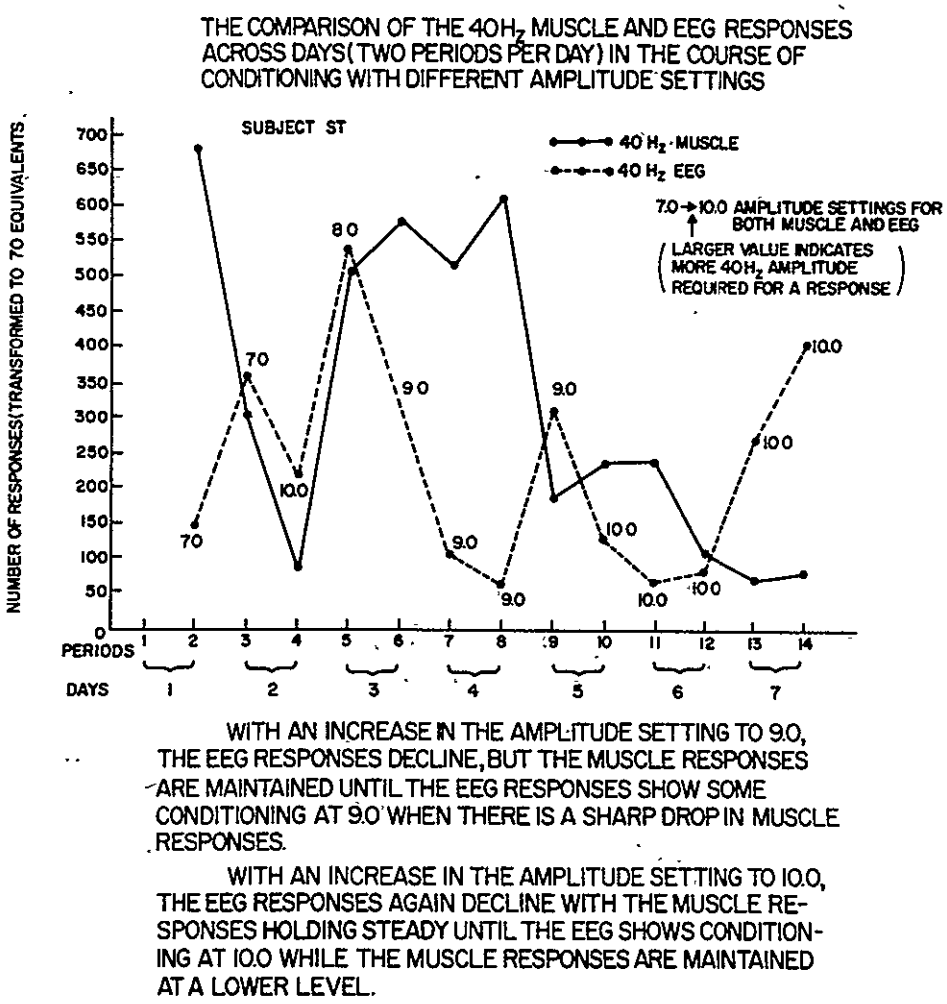


Figure 5. Dissociation of EEG and muscle response during the course of conditioning in one subject. Ordinate represents the number of bursts at each of two 15-minute conditioning periods per day over seven days. It is clear that the EEG and muscle responses do not follow the same pattern with changes in amplitude settings and conditioning.

During the course of conditioning in biofeedback training sessions, learning curves for the EEG and muscle 40-Hz responses, obtained from the comparators, were compared. The EEG leads were a bipolar recording from O_1-P_3 and the muscle leads were a biopolar recording from the neck and temporal muscles on the same side. The responses were corrected 40-Hz bursts, 75 msec in duration, at the same amplitude threshold for both EEG and muscle. Figure 5 shows the learning curves for one subject when the

p.428-n

ORIGINAL PAGE IS OF POOR QUALITY

amplitude thresholds were varied during the course of conditioning to emphasize the dissociation of EEG and muscle.

There were two 15-minute conditioning periods per day for seven successive days with the reinforcement contingent on the EEG responses only. At day 3, with an increase in the amplitude threshold from 8.0 to 9.0, there is a decline in EEG responses for three periods until the conditioning effect begins to show up as an increase in responses at the first period on day 5. The muscle responses follow a quite different pattern. Beginning at the second period on day 5, with an increase in amplitude threshold to 10.0, the EEG responses again show a decline for three periods until the conditioning increases the EEG responses for two periods on day 7. The muscle responses do not show the same decline with an amplitude increase to 10.0 on day 5, and actually show a decrease in responses on the two EEG conditioning periods of day 7. The differential pattern of these curves for EEG and muscle obtained with this subject typifies the consistent dissociation between EEG and muscle responses obtained during the course of conditioning with the on-line comparators.

Results

Conditioning and suppression. The data presented here are based on two groups of five adult subjects each who were trained to condition 40 Hz and one group of five who were trained to suppress 40 Hz. All subjects had one baseline session, during which the amplitude thresholds of both the EEG and muscle comparators were adjusted for each individual subject to allow a low to moderate level of EEG operants. They then received eight conditioning or suppression sessions with two 15-minute periods in each session. The subjects were instructed as follows: "The task is to learn to control your own brain waves. The best way to do this is to remain physically relaxed and mentally alert. You will know how well you are succeeding by how many slides you are able to turn on. The money you earn in these sessions will be based on the increased number of slides you turn on and remember, from session to session. After each session you will be asked to describe the slides you saw."

For the conditioning sessions the subjects were told that increases in a brain wave would turn on the slide projector. For the suppression sessions the subjects were told that decreases in a brain wave would keep a tone off, and that for each 30 seconds the tone remained off the slide projector would turn on.

In addition to the digital counts of 40-Hz EEG and muscle bursts, another comparator—set at the same burst duration and amplitude level but with a filter in the frequency range of 21 to 30 Hz—also counted beta bursts.

7. 428-6

The EEG leads that were either conditioned or suppressed were O_1-P_3, the muscle leads were from the left neck and temporal muscles, and the beta responses were counted from the same O_1-P_3 leads.

The EEG from the O_1-P_3 leads, left neck and temporal muscle, and triangulated combinations of these were continuously monitored during the conditioning and suppression sessions. An EEG record from a 40-Hz conditioning session and another record from a beta conditioning session are shown in Figure 6. Beta conditioning at 21 to 30 Hz from the O_1-P_3 leads was carried out on additional subjects as control procedures.

In the 40-Hz conditioning session (Figure 6), the first-event pen indicates the occurrence of beta; the second pen, 40 Hz; and the third, muscle. For the 40 Hz on event pen 2 the 60-Hz marker only above the baseline indicates that the 40-Hz EEG is contingent with the 40-Hz muscle and thus not counted. It is only counted when the 60-Hz marker is above and below the line, indicating noncontingency with both 70-Hz EEG and 40-Hz muscle. In the beta conditioning sessions, the first event pen, now a 60-Hz marker, indicates the occurrence of beta; the second pen, muscle.

From the triangulation of these leads it is sometimes possible to infer a more specific locus for muscle bursts or for distinctive trains of the EEG. It is interesting to note that, with the paper speed at 100 mm/sec, it becomes clear that what is being conditioned as beta (21 to 30 Hz) is not "desynchronization" but quite synchronous bursts.

EEG records from the first and seventh suppression sessions are shown in Figure 7. The seventh session is distinctly different from the first, with the appearance of alpha and the absence of 40 Hz, beta, and muscle. Note also the spread of the high-amplitude alpha activity into the muscle leads, NM_L-TM_L. Apparently there can also be EEG artifact when recording muscle activity.

The conditioning data on the ten subjects and suppression data on five subjects are presented in Table 1. With session 1 as a baseline, the percentage changes on 40 Hz, beta, and muscle are shown for these two groups across sessions as a function of conditioning and suppression.

For the 40-Hz conditioning, Friedman signed-ranks analyses of variance with N = 10 were computed. On the 40-Hz EEG there was a significant difference across sessions at .01; on the beta there was a significant difference at .05; on the muscle the difference was not significant.

The same analyses with N = 5 were computed for the 40-Hz suppression. The only significant difference was on the 40 Hz at .05; beta and muscle were not significant.

On the 40-Hz conditioning all ten subjects showed a consistent trend toward conditioning. The group had a 160-percent increase in 40-Hz responses from session 1 to 8. On beta responses there was an increase of 65

J. 428-J

percent. On muscle responses the session changes were more variable and the percentage change from session 1 to 8 of 16 percent was not significant.

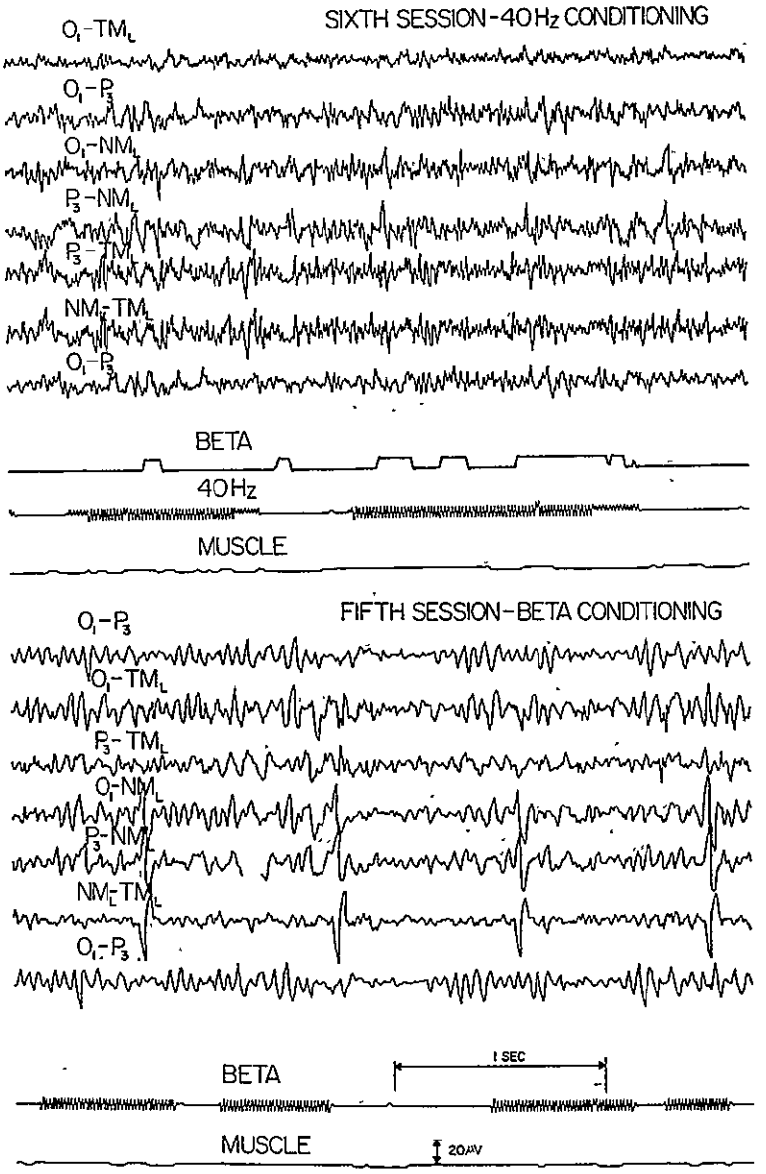


Figure 6. EEG records of 40 Hz and beta conditioning sessions, showing the bipolar leads recorded and three events in the top record and two events in the bottom record, indicating the occurrence of beta, 40 Hz, and muscle from the conditioned bipolar leads, O_1-P_3. NM_L = left neck muscle, TM_L = left temporal muscle. Note that with the paper speed at 100 mm/sec the conditioned beta (21 to 30 Hz) shows up as synchronous bursts instead of desynchronization.

p. 428-8

ORIGINAL PAGE IS
OF POOR QUALITY

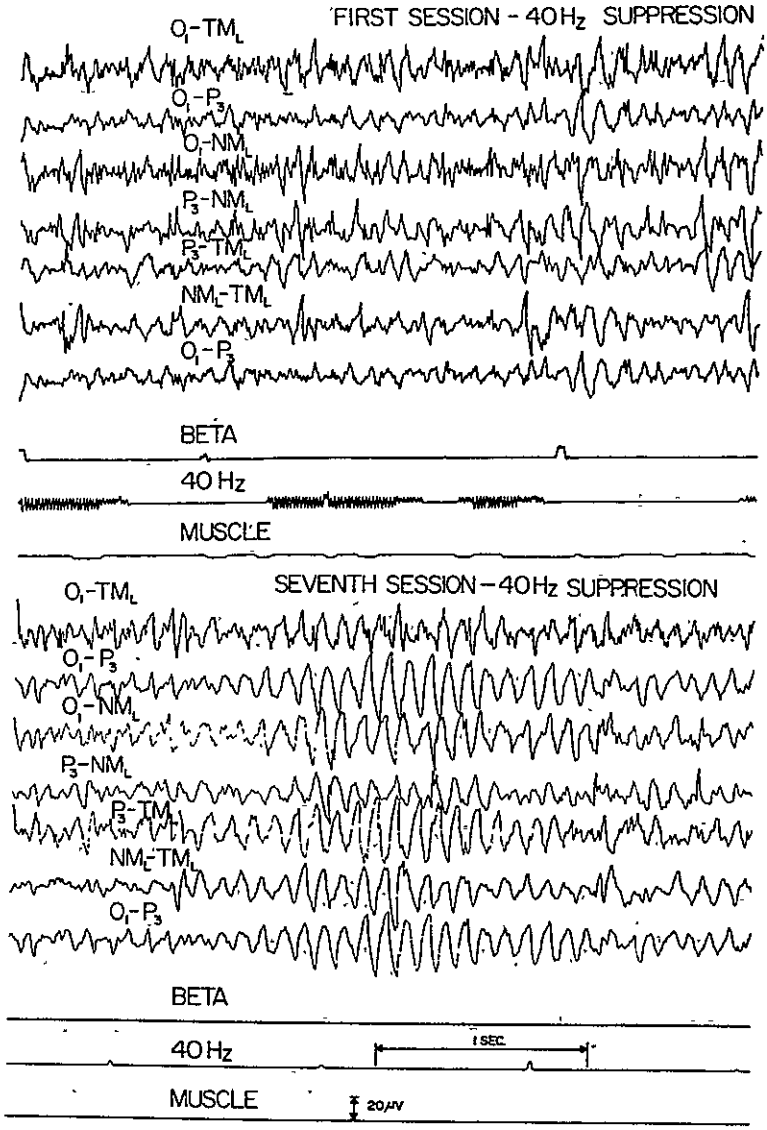


Figure 7. EEG records of 40-Hz suppression sessions, showing the bipolar leads recorded and three events, beta, 40 Hz, and muscle, from the suppressed bipolar leads, O_1-P_3. NM_L = left neck muscle and TM_L = left temporal muscle. Note the marked difference in the record during the seventh suppression session with the presence of alpha and an absence of beta, 40 Hz, and muscle.

J.428-2

TABLE 1.
Percentage Changes in 40 Hz, Beta, and Muscle Responses from Baseline During the Course of 40-Hz Conditioning and Suppression
(Conditioning N=10; Suppression N=5)

40 Hz-Conditioning

		% Changes from Session 1						
	Session 1 Means	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7	Session 8
40 Hz	186.4	+6%	+60%	+107%	+97%	+100%	+77%	+160% *
Beta	1017.6	-33%	+33%	+66%	+42%	+20%	+54%	+65% **
Muscle	3191.8	-2%	+.5%	+3%	+8%	+1%	+8%	+16%

40 Hz-Suppression

		% Changes from Session 1						
	Session 1 Means	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7	Session 8
40 Hz	213.3	-23%	-2%	-3%	-22%	-64%	-75%	-79% **
Beta	1587.6	+10%	+11%	-3%	-14%	-19%	-15%	-18%
Muscle	3538.5	-21%	-34%	-17%	-20%	-17%	-23%	-15%

*significant at .01 level
**significant at .05 level

ORIGINAL PAGE IS
OF POOR QUALITY

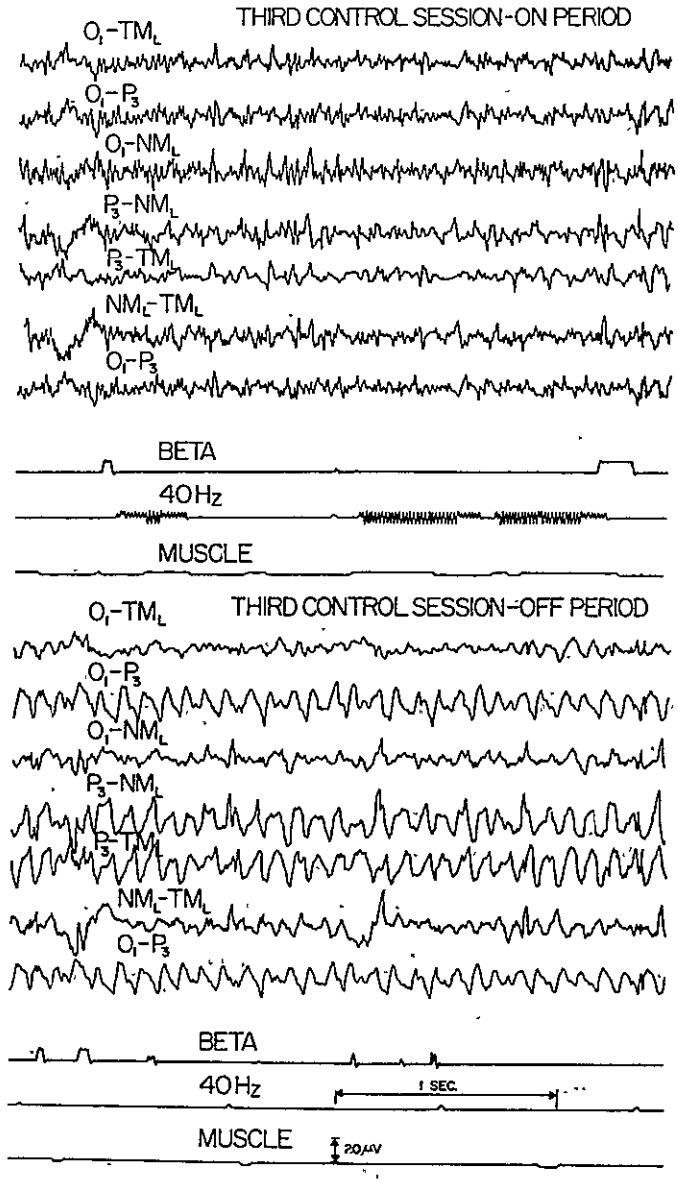


Figure 8. EEG records of alternate 2-minute control sessions in which subjects were required to turn 40 Hz "on" and "off" without reinforcement feedback. Note the absence of 40 Hz during the "off" period and marked presence of alpha with some beta and muscle in the O_1-P_3 lead. During the "on" period there is considerable 40 Hz and little beta.

β.428-t

On the 40-Hz suppression all five subjects showed a consistent trend toward suppression. For the group the percentage decrease was 79 percent on 40 Hz from session 1 to 8, 18 percent on beta, and 15 percent on muscle.

From these data it appears that a high degree of operant control of 40 Hz activity can be achieved by biofeedback training in both conditioning and suppression. With proper controls the conditioning of 40-Hz EEG can be dissociated from muscle activity. There is a significant but low degree of common variance between 40-Hz activity and beta (21 to 30 Hz). The distribution of correlations between 40 Hz and beta for the different sessions, combining conditioning and suppression, generally ranged from .35 to .45, which indicates about a 20 percent common variance. It is understandable that there should be a significant common variance—perhaps larger if error variance were reduced—because beta and 40 Hz represent different aspects or functioning of a common arousal process, diffuse and focused. At the same time it should be recognized that the different functions must have other parameters that are distinct and significant because there is a considerable variance which is not common.

Control testing. From one to three weeks after completion of the conditioning and suppression sessions, control-testing procedures were instituted to examine how much voluntary control the subjects had over the 40-Hz EEG. The test sessions were given once a week and consisted of one 8-minute-45-second warm-up period in which the same reinforcement feedback was provided as in the original conditioning and suppression. This was followed by ten consecutive 2-minute control periods consisting of alternate five "on" periods and five "off" periods, during which reinforcement was not given.

For the conditioning subjects the instructions for the "on" periods were to turn on the brain rhythm that had turned on the slides, and for the "off" periods to turn this brain rhythm off.

For the suppression subjects the instructions for the "on" periods were to turn on the brain rhythm (40 Hz) that kept the slides off. For the "off" periods they were told to turn on the brain rhythm (non-40 Hz) that turned on the slides.

EEG recordings from an "on" period and from the alternate "off" period during the third control session for one subject are shown in Figure 8. Differences between these two records are clearly evident. During the "on" period there is considerable 40 Hz present, little beta, and a good deal of 40-Hz muscle activity from the NM_L-TM_L leads. During the alternate "off" period the EEG picture had changed considerably. There is a complete absence of 40 Hz, about the same beta and less muscle, but now we see a straight run of alpha activity.

p. 428-a

ORIGINAL PAGE IS
OF POOR QUALITY

The EEG records during the sixth control session for the same subject are shown in Figure 9. Again the differences between alternate "on" and "off" periods are clear. During the "on" period the presence of 40-Hz EEG is strong, with some beta and 40-Hz muscle activity. During the "off" period there is now no 40-Hz EEG or beta, but about the same muscle. The pronounced run of polyphasic muscle activity in the NM_L-TM_L leads, primarily due to the neck muscle, does not appreciably affect the O_1-P_3 EEG leads.

The data on 40 Hz, beta, and muscle responses for one subject, carried through with control-testing sessions from the third to seventh postconditioning week, are shown in Table 2. There is consistent pronounced control over the 40 Hz activity without reinforcement feedback throughout this five-week period. On the 40-Hz EEG the mean for the "on" periods was 37.4 responses, with only two 2-minute periods at zero responses. The mean for the "off" periods was 0.48 responses with 17 out of 25 2-minute periods at zero responses. On beta the mean for the "on" periods was 151.04 responses and the mean for the "off" periods was 55.80 with no zero responses in either period. On muscle activity the mean for the "on" periods was 246.18 responses and the mean for the "off" periods was 180.56, with no zero responses in either period.

For the "on" periods a rank-order correlation (N = 25) between 40-Hz EEG and beta was .22; between 40 Hz EEG and muscle it was -.16; and between beta and muscle it was .34. For the "off" periods the 40-Hz EEG distribution had 19 zeros out of 25 scores and so correlations could not be computed.

On the eighth postconditioning week the subject was given the same control session with the same instructions, but he was also required to solve a series of problems during both the "on" and "off" periods. The subject was instructed, as in the previous five weeks, to turn on brain waves during the 2-minute "on" periods and to turn them off during the 2-minute "off" periods, but now he was also given three problems to solve during each period.

On the ninth postconditioning week the control session was repeated the same as before without problem-solving. The data for the eighth and ninth postconditioning weeks are shown in Table 3.

When the subject was required to solve the series of problems noted in Table 3, he could not turn off the 40 Hz activity as he had done for the previous five weeks and as he successfully did for the ninth postconditioning week without problem-solving. For the eighth postconditioning week the means for 40-Hz EEG were 41.2 responses for the "on" periods and 48.5 for the "off" periods. For the ninth postconditioning week they were 25.8 responses for the "on" periods and 0.4 for the "off" periods.

This experimental situation seems to be very sensitive to behavioral effects on the 40-Hz activity. This subject had achieved quite remarkable

p. 428-v

ORIGINAL PAGE IS
OF POOR QUALITY

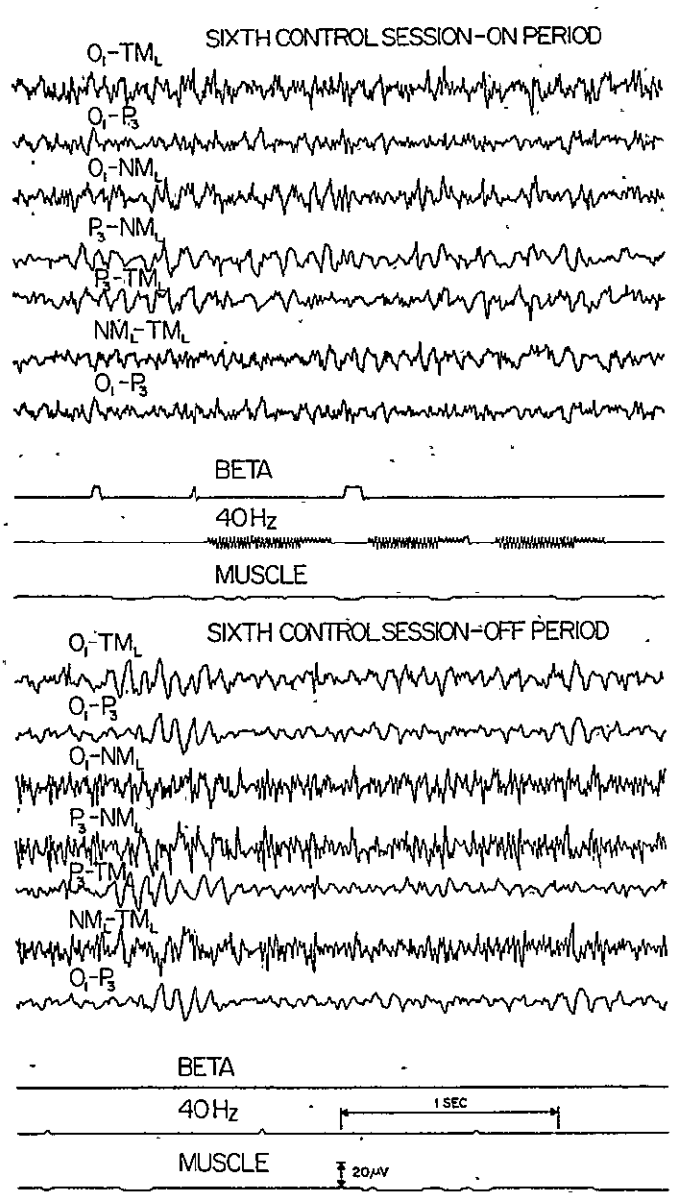


Figure 9. EEG records of alternate 2-minute control sessions in which subjects were required to turn 40 Hz "on" and "off" without reinforcement feedback. Note the absence of beta and 40 Hz during the "off" period and marked neck-muscle artifact which does not show up in the O_1-P_3 leads.

p. 428-w

control over his EEG activity on the basis of conditioning and instructional set. He could not maintain this control when given what were apparently conflicting instructions to solve problems.

The voluntary control of 40 Hz activity has generality beyond this one subject. Data are presented in Table 4 for four additional postconditioning subjects and two postsuppression subjects who completed the control testing during the first postconditioning week.

Comparisons between the pairs of alternate "on" and "off" periods for the four postconditioning subjects show a definite trend for a higher level of 40-Hz responding during the "on" periods. However, they also clearly show the important effect of individual differences in motivation level when subjects attempt to maintain voluntary control over their own brain rhythms on the basis simply of instructional set. Subject 1 had a comparatively low level of 40-Hz but was able to maintain the distinction between alternate "on" and "off" periods except for the third pair, where he produced only one response for each period. Subject 4 maintained a consistently strong distinction throughout. Subjects 2 and 3 started out, for the first 3 alternate pairs, with very high levels of 40-Hz responses and clear distinctions between "on" and "off" periods, but they reversed on alternate pairs 4 and 5 and produced high levels of 40-Hz responses during the last "off" period in the session.

The two suppression subjects followed a consistent pattern. As instructed, they were able to suppress 40-Hz responses during all "off" periods as compared with their alternate "on" periods. The mean number of responses during the "off" periods was 0.6; for the "on" periods it was 4.2 responses. These can be compared with the means for the four conditioned subjects for whom the mean was 9.15 for the "off" periods and 33.25 for the "on" periods.

Concomitant behaviors. A number of different behavioral probes were tried in this training series to see what techniques might be effectively used for demonstrating relationships between 40 Hz change and behavior.

One series of measures focused on remembering the slides used as reinforcers. Detailed descriptions of these slides were obtained from subjects after each conditioning session. Quantitative and qualitative categorizations of this descriptive material as related to various measures of 40 Hz change failed to reveal any consistent trends.

Many different forms of self-reports—interviews, Q-sorts, adjective checklists, etc.—have been used with biofeedback training to try to establish connections with psychological variables. In the present series extensive structured and unstructured interviews were conducted with subjects after each conditioning session. From the voluminous material obtained almost any hypothesis could be partially substantiated, depending upon the classifications made and inferences drawn from these classifications.

1.428-x

TABLE 2.
Number of Responses of 40 Hz, Beta, and Muscle for Ten Consecutive Control Periods of Alternate "On" and "Off"

		On	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*
Third	40 Hz	41	0	36	2	42	0	0	0	15	0
post	Beta	213	24	175	50	111	24	16	33	117	46
week	Muscle	273	201	259	217	350	168	216	134	279	182
Fourth	40 Hz	0	0	87	0	5	0	41	0	75	0
post	Beta	13	7	137	11	33	15	54	7	139	15
week	Muscle	290	163	107	109	271	182	109	115	85	208
Fifth	40 Hz	52	0	1	0	65	0	27	1	8	2
post	Beta	268	155	209	170	167	174	195	203	207	202
week	Muscle	273	156	240	216	274	164	169	113	251	270
Sixth	40 Hz	57	0	15	1	50	0	46	1	39	3
post	Beta	186	11	83	41	219	17	167	39	204	23
week	Muscle	240	277	194	195	326	106	286	278	332	258
Seventh	40 Hz	37	0	44	1	52	1	35	0	55	0
post	Beta	156	32	185	29	172	26	210	17	140	24
week	Muscle	307	211	285	135	233	144	287	176	223	136

Note: Subjects were instructed to turn on the brain rhythm that turned the slides on or to keep the rhythm off. Sessions were given once a week after 40-Hz conditioning. Data represent sessions from the third through the seventh post-conditioning weeks for one subject.

7.428-4

TABLE 3.
Number of Responses of 40 Hz, Beta, and Muscle for
Alternate "On" and "Off" Control Periods at Eighth and Ninth Postconditioning Weeks
for Same Subject Shown in Table 2.

		On	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*
Eighth	40 Hz	35	55	57	43	37	56	46	48	31	40
post	Beta	150	180	206	139	159	172	168	193	115	166
week	Muscle	287	226	210	158	293	184	278	276	281	240
Correct answers		2	1	3	1	3	2	1	2	0	1
Test items (N=3 in each period).		verbal analogies		verbal opposites		math. problems		math. problems		ravens matrices	
		On	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*
Ninth	40 Hz	4	1	27	0	32	1	19	0	47	0
post	Beta	74	36	102	17	140	30	122	17	158	19
week	Muscle	269	197	229	184	275	203	230	270	245	203

Note: At the eighth week the subject was also required to solve problems as shown, during both "on" and "off" periods. While solving problems he could not maintain suppression of 40 Hz during "off" periods. At the ninth week, under standard conditions, he again had control of the 40 Hz during "on" and "off" periods.

p. 428-3

TABLE 4.
Number of 40-Hz Responses for Alternate "On" and "Off" Control Periods at First Postconditioning Week for Four Additional 40-Hz-Conditioned Subjects and Two 40-Hz-Suppressed Subjects

	40-Hz Conditioning—First Post Week—40 Hz Bursts									
Subjects	*On*	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*
1	13	1	1	0	1	1	4	0	2	1
2	35	13	66	11	86	6	6	15	5	31
3	56	2	72	15	61	5	2	9	20	61
4	30	1	26	2	30	1	22	4	27	4

	40-Hz Suppression—First Post Week—40 Hz Bursts									
Subjects	*On*	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*	*On*	*Off*
1	0	6	3	4	1	2	1	1	1	2
2	0	4	0	10	0	2	0	5	0	6

Note: Instructions to suppressed subjects during "off" periods were to turn on brain rhythm (non-40 Hz) that kept slides on. During "on" periods they were instructed to turn on brain rhythm (40 Hz) that kept slides off.

7. 428-aa

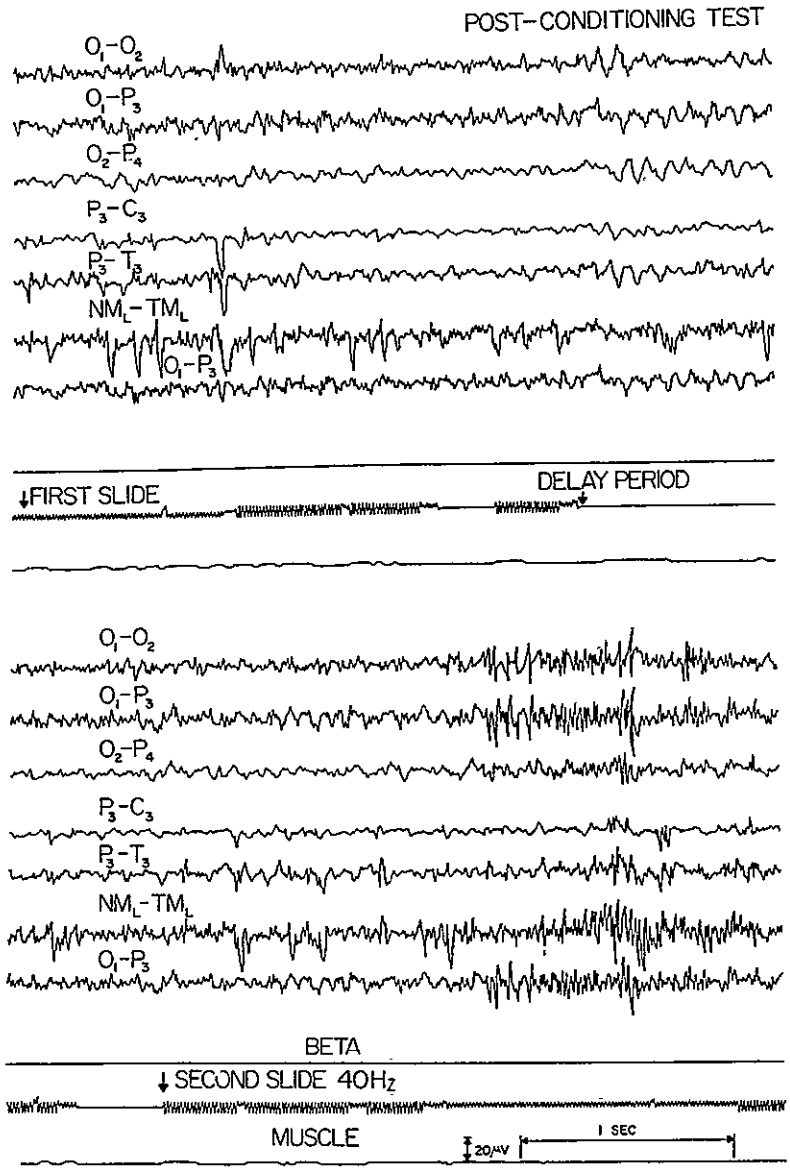


Figure 10. EEG record of a postconditioning test session in which subjects were required to solve problems presented as questions (first slide) and multiple choice answers (second slide). Note the marked occurrence of 40 Hz and the absence of beta with the presentation of the first slide and particularly at the beginning of the second slide in the O_1-P_3 leads. When a large muscle burst occurs with the verbal answer at the end of the second slide in the muscle leads as well as the O_1-P_3 leads, the 40-Hz bursts are not counted on the 40-Hz event pen because of comparator contingencies. Note also that the conditioning of the O_1-P_3 leads does not appear to carry over to the O_2-P_4 leads.

ORIGINAL PAGE IS
OF POOR QUALITY

Based on the literature and our previous work on 40 Hz, we set up pre- and postconditioning test sessions on problem-solving tasks for five conditioned subjects. Each test session consisted of 5-minutes pre-baseline, two 10-minute sets of problems, a 10-minute interpolated activity, and 5 minutes post-baseline.

During each 10-minute problem-solving period five test items were presented. Each item consisted of a problem presented on a slide for 30 seconds, followed by a 15-second pause, and then another slide with multiple-choice answers on for 60 seconds. An EEG record of a postconditioning test trial is shown in Figure 10.

For the pre- and postconditioning test sessions, the electrode leads recorded include O_2-P_4, P_3-C_3, and P_3-T_3 to obtain some idea of what spread may have occurred from conditioning the O_1-P_3 leads. As can be seen in the figure, the marked occurrence of 40 Hz in the O_1-P_3 leads toward the end of the first slide and the beginning of the second slide does not generalize to the O_2-P_4 leads on the opposite side and also does not appear to spread in either the P_3-C_3 or P_3-T_3 direction. Of interest is the complete absence of beta even though there is considerable 40 Hz present during this problem-solving situation. Note also the high amplitude polyphasic muscle burst which represents the subject's verbal answer. It spreads into the O_1-P_3 leads, but the 40-Hz EEG event pen does not register any responses because of the comparator contingencies.

The data on 40-Hz responses and test performance, comparing pre- and postconditioning test sessions, are shown in Table 5. The pre- and post-baselines and memory for words were controls for the two problem-solving tasks, modified forms of the Minnesota Paper Form and Differential Aptitude Test. Mean number of 40 Hz responses before and after conditioning showed no significant differences on the baseline and "words" conditions. There were significant mean increases in 40-Hz responses on the two problem tasks, with all five subjects showing a rise during both problem-solving periods. On the test performance measure, that is, the number of correct responses on both sets of problems, there was a significant improvement by all five subjects after conditioning.

The data on behavioral correlates of the 40-Hz EEG appear promising. The results show that, with proper controls, this electrical activity can be conditioned and that some degree of voluntary control can be achieved, although with a good deal of individual variability. The important question still remains: can changes in this brain electrical activity bring along changes in behavior that have sufficient generality and stability to be significant? Systematic behavioral analysis, with a focus on generalization and transfer effects through a range of subject populations, is certainly required before firm conclusions can be drawn.

1.428-cc

TABLE 5.
Data on Pre- and Postconditioning Test Sessions for Five Subjects Conditioned on 40-Hz EEG

Mean number of 40-Hz bursts during baseline and test sessions before and after conditioning for five subjects

	Pre-Baseline	*MPFB Problems*	*DAT Problems*	*Words*	*Post-Baseline*
Before conditioning	21	37	41	77	58
After conditioning	23	141	158	121	22
Number subjects increasing	2	5	5	3	3

Mean number of correct responses in test sessions before and after conditioning for five subjects

	MPFB Problems	*DAT Problems*	*Words*
Before conditioning	2.6	1.6	11
After conditioning	4.0	4.4	16.6
Number subjects improving	5	5	3

Notes: Pre- and post-baseline: Five minutes of EEG recording during rest before and after test sessions. *MPFB problems:* Two equated sets of five items each from the Minnesota Paper Form Board Test, one set given before and another set after conditioning. For each item the problem was presented on a slide for 30 seconds, followed by a 15-second pause, and then followed by a multiple-choice answer slide which was on until the subject answered, or for 60 seconds if there was no answer. Score was the total number of correct answers. *DAT problems:* Two equated sets of five items each from the Differential Aptitude Test, one set given before and another set after conditioning. The procedure was the same as for the first group of problems. Score was the total number of correct answers. *Words:* Thirty words were presented on slides, one word per slide on for one second, with a 15-second pause between slides. Score was the number of words correctly remembered after the session.

p.428-dd

Theory

The background for biofeedback training of 40 Hz has been reviewed in the introduction. Briefly, learning and problem-solving behavior must depend on short-term memory processing; and, in turn, memory traces in short-term store—dynamic organizations which are time-dependent, state-dependent, and context-dependent—must take the form of patterned electrical brain activity (Sheer 1970, Sheer 1972). The 40-Hz EEG reflects a state of localized cortical excitability or focused arousal, which is "optimal" for consolidation in short-term store.

There is a unique value in the EEG because it provides an index of the combined activity of masses of cells, and it is precisely this combination that is of major interest in the analysis of molar brain function. The effective use of EEG data lies in synthesizing the framework within which molecular analysis can be carried on, in the fashion that Sherrington reflex physiology provided the essential molar concepts on which the analysis of the electrical properties of the spinal motoneuron has been based. From our correlations between 40 Hz and problem-solving behavior, we can go on to a consideration of the mechanisms for this electrical activity.

The strongest, most pronounced 40 Hz occurs in rhinencephalic structures, particularly the olfactory bulb, through a range of species from catfish to man (Sheer and Grandstaff 1970). At the olfactory bulb the essential and sufficient stimulus is airflow; at the amygdala airflow is essential but a certain level of arousal is also necessary (Pagano 1966, Sheer, Grandstaff, and Benignus 1966); and at prepyriform cortex 40 Hz can be conditioned to neutral stimuli (Freeman 1963). All this is very interesting because in quadruped animals olfaction is a distance receptor, and sniffing—taking in stimuli—is an important orienting response.

In the olfactory bulb laminar structure is first encountered in a much simpler form than in neocortex and an analysis of mechanism should be easier to come by. The evidence is strong that the 40-Hz waves in the olfactory bulb are standing potentials derived from synaptic and postsynaptic events, with the synchrony attributable to successive trains of excitation and recurrent inhibition. The property of recurrent inhibition is essentially negative feedback which functionally contributes to the phasing of rhythmic discharges (Andersen, Eccles, and Loyning 1963; Granit 1963; Eccles 1965; Andersen and Andersen 1968).

Rall and Shepherd (1968) and Shepherd (1970) made an elegant detailed analysis of a dendrodendritic synaptic interaction as a mechanism for this rhythmic activity in the olfactory bulb. Air flow excites bipolar receptor cells in the olfactory mucosa, whose axons form the olfactory nerves synapsing within encapsulated glomeruli with dendrites of tufted and mitral cells.

P.428-ee

Impulse discharge in mitral cells results in synaptic excitation of granule cells; these granule cells then deliver graded inhibition to mitral cells. This inhibition cuts off the source of synaptic excitatory input to the granule cells. As the granule-cell activity subsides, the amount of inhibition delivered to the mitral cells is reduced, which permits the mitral cells to respond again to excitatory input from the glomeruli. In this way a sustained excitatory input to the mitral cells would be converted into a rhythmic sequence of excitation followed by inhibition, locked in timing to a rhythmic activation of the granule-cell pool.

At the neocortex there exists a more complicated situation for which details on mechanism are lacking. Stefanis and Jasper (1964) and Jasper and Stefanis (1965) reported that the axon collaterals of cortical pyramidal cells in cats are facilitated at relatively high frequencies of repetitive excitation. The negative feedback of recurrent inhibition provides an automatic control of level of excitation—the greater the excitation, the more the feedback of inhibition.

The point is that the 40-Hz EEG reflects repetitive stimulation at a constant frequency for a limited time over a limited circuitry. The circuitry is defined behaviorally by the spatial-temporal patterning of sensory inputs, motor outputs, and reinforcement contingencies. It is "optimal" for consolidation because repetitive synchronous excitation of cells maximizes the efficiency of synaptic transmission over the limited circuitry.

Eccles (1964) has well documented this property of frequency potentiation by measuring the size of excitatory postsynaptic potentials that develop with repetitive activation at different frequencies. The duration of constant repetitive discharges is probably as significant in the transfer of information as is the intensity of neuronal firing or the number of cells involved. From quantitative studies in the spinal cord, Granit (1963) concluded that frequency of firing was the main code determining rate of continuous discharge in control of tonic motoneurons. At the human somatosensory cortex, Libet et al. (1967) have shown that subthreshold stimulus pulses could elicit conscious sensory experience only if they were delivered repetitively at 20 to 60 pulses per second.

A significant outcome of this feedback loop—wherein frequency potentiation sets up recurrent inhibition which sets up synchrony which sets up frequency potentiation—is the property of contrast. The negative feedback of recurrent inhibition particularly depresses those synapses that are weakly excited in the "surround" and so serves to further sharpen the focus of excitation. Behaviorally it is reflected, on the input side, in sharpening of attention to relevant stimuli and, on the output side, in decreasing extraneous responses and greater precision of relevant movements. The operation of contrast as a function of surround inhibition has been detailed for the somes-

p. 428-ff

thetic system (Mountcastle 1961), the visual system (Hubel and Wiesel 1962), and audition (Whitfield 1965).

There is evidence that this repetitive synchronous excitation of cortical cells is dependent upon cholinergic pathways approaching the pyramidal cells of *layer V*. These pathways provide the final stage in the ascending reticular activating system that arises from the tegmental and reticular nuclei of the brain stem (Shute and Lewis 1967). Direct stimulation of the mesencephalic reticular formation enhances the release of acetylcholine and is correlated with an EEG activation pattern (Celesia and Jasper 1966).

With the exception of its nicotinic action on Renshaw cells, acetylcholine's effect on neurons in the cortex is slow in onset and offset, not suitable for rapid "detonation" transmission, but functional in the modulation of the excitability of cortical input cells (Krnjevic 1969). In their early work Dempsey and his colleagues (Chatfield and Dempsey 1942, Morrison and Dempsey 1943) reported that repetitive excitation after single stimulation of a peripheral nerve was greatly increased after application of acetylcholine to the cortex of cats. This work has been extended by Krnjevic and Phillis (1963), who found that acetylcholine enhances rhythmic afterdischarges following sensory volleys, and by Spehlman (1971) who found that it facilitates the firing rate of cortical units activated by reticular stimulation.

Behaviorally, the importance of the reticular activating system in the consolidation process has been detailed by Block and his colleagues (Block 1970). They showed that direct reticular stimulation, when applied immediately after registration of information, considerably facilitates learning; the effect is less when the stimulation is delayed until 90 seconds after the learning trial. In addition, under certain conditions, posttrial reticular stimulation annuls the effect of fluothane anesthesia, which by itself prevents consolidation.

A series of studies has shown further that learning may be impaired by drugs that inhibit and facilitated by drugs that increase acetylcholine action. Atropine blocks the cortical postsynaptic effects of acetylcholine release and induces a slow-wave, high-voltage EEG pattern. Although behaviorally there is no concomitant appearance of drowsiness or sleep, atropine does affect learning. It depressed the performance of learned avoidance responses (Herz 1959) and auditory discrimination learning in rats (Michelson 1961), but, in both studies, only when administered during the early stages of training. Other impaired tasks include successive discrimination learning (Whitehouse 1964) and learned alternation and complex multiple-choice discrimination (Carlton 1963). On the other hand, physostigmine, which increases acetylcholine action, facilitated one-trial avoidance by rats when administered a few minutes before training trials (Bures, Bohdanecky, and Weiss 1962). It

p.428-gg

also improved the rats' learning of a Lashley III maze when administered 30 seconds after each daily trial (Stratton and Petrinovich 1963). Physostigmine impaired learning in both situations when given in larger doses, which recalls the U-shaped function between activation level and behavioral efficiency.

The network we have woven here is a long way from the correlation between 40-Hz EEG and problem-solving behavior, and its value is primarily heuristic. It makes connections between a number of related research areas in a context that should provide choice points for experimental testing.* The research strategy for such complex functions as memory processing would seem to require a two-step procedure as was followed here: first, to establish correlations between behavior and critical patterns of organized electricity; then, to focus attention on the physical and chemical mechanisms that are the basis for these organizations. The two steps seem required because, on the one hand, units and mechanisms in isolation are unlikely to be directly correlated with the complex behavior represented by memory traces; on the other hand, correlations between behavior and electrical organizations are only a first step toward an analysis of underlying mechanisms.

Summary

A specific pattern of brain electricity, a narrow frequency band centering at 40 Hz, reflects a state of circumscribed cortical excitability or focused arousal which is "optimal" for consolidation in short-term store. On-line control procedures have been developed to reliably record and digitally count this low-amplitude EEG activity independent of muscle artifact.

A high degree of operant control of the 40-Hz EEG can be achieved by biofeedback training in both conditioning and suppression. In control testing sessions following conditioning and suppression training, some degree of voluntary control has been demonstrated when subjects alternately turned the 40 Hz on and off only upon instructional sets. The generality and stability of relationships shown between the conditioned 40-Hz EEG and problem-solving behavior requires further systematic verification.

*Work in progress in our laboratory shows a significant spectral coherence, a lock-in at 40 Hz, between mesencephalic reticular and cortical visual and motor areas when the cat is beginning to meet learning criteria in a successive visual discrimination task.

p. 428-22

Acknowledgments

Research reported here has been supported, during its equipment development phase, under NASA Grant No. NAS 9-1462. I would like to acknowledge the collaboration of Bruce Bird and Fred Newton, who assumed major responsibility for carrying out training experiments; Dr. Lyllian Hix for the hybrid computer programming; and David Ballard and William Ellis for technical assistance.

References

Andersen, P.O., and Andersen, S.A. 1968. *Physiological Basis of Alpha Rhythm.* New York: Appleton-Century-Crofts.

Andersen, P., Eccles, J.C., and Loyning, Y. 1963. Recurrent inhibition in the hippocampus with identification of the inhibitory cell and its synapses. *Nature* 198:540.

Banquet, J.P. 1973. Spectral analysis of the EEG in meditation. *Electroencephalogr. Clin. Neurophysiol.* 35:143.

Barber, T., DiCara, L.V., Kamiya, J., Miller, N.E., Shapiro, D., and Stoyva, J. (eds.) 1971a. *Biofeedback and Self-Control.* Chicago: Aldine-Atherton.

Barber, T., DiCara, L.V., Kamiya, J., Miller, N.E., Shapiro, D., and Stoyva, J. (eds.) 1971b. *Biofeedback and Self-Control 1970.* Chicago: Aldine-Atherton.

Beatty, J.T. 1971. Effects of initial alpha wave abundance and operant training procedures on occipital alpha and beta activity. *Psychonom. Sci.* 23:197.

Beatty, J.T. 1972. Similar effects of feedback signals and instructional information on EEG activity. *Physiol. and Behav.* 9:151.

Beatty, J. Greenberg, A., Derbler, W.P., and O'Hanlon, J.F. 1974. Operant control of occipital theta rhythm affects performance in a radar monitoring task. *Science* 183:871.

Black, A.H. 1971. The direct control of neural processes by reward and punishment. *Am. Sci.* 59:238.

Black, A.H. 1972. The operant conditioning of central nervous system electrical activity. In G.H. Bower (ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory,* p. 47. New York: Academic Press.

Block, V. 1970. Facts and hypotheses concerning memory consolidation processes. *Brain Res.* 24:561.

Brown, B.B. 1970. Recognition of aspects of consciousness through association with EEG alpha activity represented by a light signal. *Psychophysiology* 6:442.

Bures, J., Bohdanecky, Z., and Weiss, T. 1962. Physostigmine induced hippocampal theta activity and learning in rats. *Psychopharmacologia* 3:254.

Butler, F., and Stoyva, J. 1973. *Biofeedback and Self-Regulation: A Bibliography.* Denver, Col.: Biofeedback Research Society.

Carlton, P.L. 1963. Cholinergic mechanisms in the control of behavior by the brain. *Psychol. Rev.* 70:19.

Chaffin, D.B. 1969. Surface electromyography frequency analysis as a diagnostic tool. *J. Occup. Med.* 11:109.

p.428-ii

Chatfield, P.O., and Dempsey, E.W. 1942. Some effects of prostigmine and acetylcholine on cortical potentials. *Am. J. Physiol.* 135:633.

Celesia, G.G., and Jasper, H.H. 1966. Acetylcholine released from cerebral cortex in relation to state of activation. *Neurology* 16:1053.

Das, N.N., and Gastaut, H. 1955. Variations de l'activité electrique du cerveau, du coeur et des muscles squellettiques au cours de la méditation et de l'extase yogique. *Electroencephalogr. Clin. Neurophysiol.* Suppl. 6:211.

Dumenko, V.N. 1961. Changes in the electroencephalogram of the dog during formation of a motor conditioned reflex stereotype. *Pavlov Journal of Higher Nervous Activity* 11:64.

Eccles, J.C. 1964. *The Physiology of Synapses.* New York: Springer-Verlag.

Eccles, J.C. 1965. The control of neuronal activity by postsynaptic inhibitory action. *23rd International Congress of Physiological Sciences,* p. 84. Amsterdam: Excerpta Medica Foundation.

Freeman, W.J. 1963. The electrical activity of a primary sensory cortex: Analysis of EEG waves. *Int. Rev. Neurobiol.* 5:53.

Galambos, R. 1958. Electrical correlates of conditioned learning. In M. Brazier (ed.), *The Central Nervous System and Behavior,* p. 375. New York: Josiah Macy, Jr. Foundation.

Galin, D., and Ornstein, R. 1972. Lateral specialization of cognitive mode: An EEG study. *Psychophysiology* 9:412.

Giannitrapani, D. 1969. EEG average frequency and intelligence. *Electroencephalogr. Clin. Neurophysiol.* 27:480.

Granit, R. 1963. Recurrent inhibition as a mechanism of control. In A. Moruzzi, A. Fessard, and H.H. Jasper (eds.), *Brain Mechanisms,* vol. I, *Progress in Brain Research,* p. 23. New York: Elsevier.

Green, E.E., Green, A., and Walter, E.D. 1972. Biofeedback for mind-body self-regulation: Healing and creativity. *Fields Within Fields . . . Within Fields* 25:131.

Herz, A. 1959. Effects of atropine on early stages of learning in the rat. *Arch. Exp. Pathol. Pharmacol.* 236:110.

Hubel, D.H. and Wiesel, T.N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160:106.

Jasper, H. and Stefanis, C. 1965. Intracellular oscillatory rhythms in pyramidal tract neurones in the cat. *Electroencephalogr. Clin. Neurophysiol.* 18:541.

Killam, K.F. and Killam, E.K. 1967. Rhinencephalic activity during acquisition and performance of conditional behavior and its modification by pharmacological agents. In W.R. Adey and T. Tokizane (eds.), *Progress in Brain Research.* vol. 27, *Structure and Function of the Limbic System,* p. 338. New York: Elsevier.

Krnjevic, K. 1969. Central cholinergic pathways. *Fed. Proc.* 28:113.

Krnjevic, K., and Phillis, J.W. 1963. Acetylcholine-sensitive cells in the cerebral cortex. *J. Physiol.* 166:296.

Libet, B., Alberts, W.W., Wright, E.W., Jr., and Feinstein, B. 1967. Responses of human somatosensory cortex to stimuli below threshold for conscious sensation. *Science* 158:1597.

Michelson, M.J. 1961. Effects of atropine on an auditory discrimination task in the rat. *Activatis Nervosa Superior* 3:140.

1.428-ii

Morrell, F. 1961. Lasting changes in synaptic organization produced by continual neuronal bombardment. In J.F. Delefresnaye (ed.), *Brain Mechanisms and Learning.* p. 91. Oxford: Blackwell.

Morrell, F., and Jasper, H.H. 1956. Electrographic studies of the formation of temporary connections in the brain. *Electroencephalogr. Clin. Neurophsyiol.* 8:201.

Morrison, R.S., and Dempsey, E.W. 1943. Mechanisms of thalmocortical augmentation and repetition. *Am. J. Physiol.* 138:297.

Mountcastle, V.B. 1961. Duality of function in the somatic afferent system. In M.A. Brazier (ed.), *Brain and Behavior,* vol. 1, p. 67. Washington, D.C.: American Institute of Biological Sciences.

Mulholland, T. 1968. Feedback electroencephalography. *Activitas Nervosa Superior* 10:4.

Mulholland, T., and Peper, E. 1971. Occipital alpha and accommodative vergence, pursuit tracking and fast eye movements. *Psychophysiology* 8:556.

Nowlis, D.P. and Kamiya, J. 1970. The control of electroencephalographic alpha rhythm through auditory feedback and the associated mental activity. *Psychophysiology* 6:496.

Pagano, R.R. 1966. The effects of central stimulation and nasal airflow on induced activity of olfactory structures. *Electroencephalogr. Clin. Neurophysiol.* 21:269.

Peper, E. 1971. Comment on feedback training of parietal-occipital alpha asymmetry in normal and human subjects. *Kybernetik* 9:156.

Peper, E. 1972. Localized EEG alpha feedback training: A possible technique for mapping subjective, conscious, and behavioral experiences. *Kybernetik* 11:166.

Pribram, K.H., Spinelli, D.N., and Kamback, M.C. 1967. Electrocortical correlates of stimulus response and reinforcement. *Science* 157:94.

Rall, W., and Shepherd, G.M. 1968. Theoretical construction of field potentials and dendrodentritic synaptic interactions in the olfactory bulb. *J. Neurophysiol.* 6:884.

Rowland, V. 1958. Discussion under electroencephalographic studies of conditioned learning. In M. Brazier (ed.), *The Central Nervous System and Behavior,* p. 347. New York: Josiah Macy, Jr. Foundation.

Sakhiulina, G.T. 1961. EEG manifestations of tonic cortical activity accompanying conditioned reflexes. *Pavlov Journal of Higher Nervous Activity* 11:48.

Shapiro, D., Barber, T.X., DiCara, L.V., Kamiya, J., Miller, N.E., and Stoyva, J. (eds.) 1973. *Biofeedback and Self-Control 1972.* Chicago: Aldine-Atherton.

Sheer, D.E. 1970. Electrophysiological correlates of memory consolidation. In A. Ungar (ed.), *Molecular Mechanisms in Memory and Learning,* p. 177. New York: Plenum Press.

Sheer, D.E. 1972. Neurobiology of memory storage processes. *Winter Conference on Brain Research,* Vail, Colorado.

Sheer, D.E. 1973. Biofeedback training of 40 Hertz in animals and man. *Symposium on Learned Control of Brain Rhythms,* American Psychological Association, Montreal, Canada.

Sheer, D.E. 1974. Electroencephalographic studies in learning disabilities. In H. Eichenwald and A. Talbot (eds.), *The Learning Disabled Child.* Austin: University of Texas Press.

Sheer, D.E., and Grandstaff, N.W. 1970. Computer-analysis of electrical activity in the brain and its relation to behavior. In H.T. Wycis (ed.), *Current Research in Neurosciences,* vol. 10, *Topical Problems in Psychiatry and Neurology,* p. 160. New York:

p. 428-pb

Karger.

Sheer, D.E., Grandstaff, N.W., and Benignus, V.A. 1966. 40 c/sec electrical activity in the brain of the cat. *Symposium on Higher Nervous Activity*, Fourth World Congress on Psychiatry, Madrid, Spain.

Sheer, D.E., and Hix, L. 1971. Computer-analysis of 40 Hz EEG in normal and MBI children. *Society of Neurosciences Meeting*, Washington, D.C.

Shepherd, G.M. 1970. The olfactory bulb as a sample cortical system: Experimental analysis and function implications. In F.O. Schmidtt (ed.), *The Neurosciences, Second Study Program*, p. 539. New York: Rockefeller Press.

Shute, C.C.D., and Lewis, P.R. 1967. The ascending cholinergic reticular system: Neocortical, olfactory and subcortical projections. *Brain* 90:497.

Spehlman, R. 1971. Acetylcholine and the synaptic transmission of non-specific impulses to the visual cortex. *Brain* 94:139.

Stefanis, C., and Jasper, H. 1964. Recurrent collateral inhibition in pyramidal tract neurons. *J. Neurophysiol.* 27:855.

Sterman, M.B. 1972. Studies of EEG biofeedback training in man and cats. *Highlights of 17th Annual Conference: VA Cooperative Studies in Mental Health and Behavioral Sciences*, Washington, D.C.: Veterans Administration.

Sterman, M.B., MacDonald, L.R., and Stone, R.K. 1974. Biofeedback training of the sensorimotor EEG rhythms in man: Effects on epilepsy. *Epilepsia* (in press).

Stoyva, J., Barber, T.X., DiCara, L.V., Kamiya, J., Miller, N.E., and Shapiro, D. (eds.) 1972. *Biofeedback and Self-Control 1971*. Chicago: Aldine-Atherton.

Stratton, L.O., and Petrinovich, L. 1963. Post-trial injections of an anti-cholinesterase drug and maze learning in two strains of rats. *Psychopharmacologia* 5:47.

Whitehouse, J. 1964. Effects of atropine on discrimination learning in the rat. *J. Comp. Physiol. Psychol.* 57:13.

Whitfield, J.C. 1965. "Edges" in auditory information processing. In *23rd International Congress of Physiological Sciences*, p. 245. Amsterdam: Excerpta Medica Foundation.

J. 428-ll

Project

(A) Project Title: Biofeedback Training of 40Hz EEG Activity In Humans: Relationships to Cognitive Performance, Voluntary Control, Subjective States and Autonomic Activity

(B) Project Abstract:

Forty Hertz EEG was examined in a group of 49 male university students during problem-solving tasks and during resting baseline conditions. Following the completion of the Pre-Test problem-solving phase subjects were then trained using biofeedback procedures, to either increase or suppress 40Hz EEG and were again tested on the Post-Test which was an equivalent form of the Pre-Test. Thirty-two subjects completed this phase of the study.

Forty Hertz EEG and beta (21-30Hz) were recorded from the O_1-P_3 lead in 25 subjects during the Pre-Test condition. Forty Hertz EEG was recorded from the C_z-O_1 and C_z-O_2 leads, and theta, alpha and beta were recorded from either C_z-O_1 or C_z-O_2 leads in 24 subjects. Forty Hertz muscle activity was monitored from the left or the left and right neck-temporal muscle for all subjects. Forty Hertz and 70Hz from the muscle leads were compared using coincidence detectors to prevent counting 40Hz generated from the muscle leads as EEG responses. EEG and muscle filters were set to trigger a digital output when three cycles of electrical activity

was present. Heart rate was recorded from a total of 14 subjects.

Thirteen subjects were given a Control Test following biofeedback training to measure the degree of voluntary control in the absence of feedback.

Subjective state analysis was performed on a total of 31 subjects at various stages in the experimental sequence. Q-Sort items, representing descriptions of theta, alpha, beta and 40Hz EEG states were sorted into four subjective feeling categories.

The following results were obtained:

1. Forth Hertz EEG was significantly greater during problem solving as compared to a resting baseline.
2. Following biofeedback training, subjects who were conditioned to increase 40Hz EEG had significantly more of this activity during the tasks condition of the Post-Test as compared to the tasks condition of the Pre-Test. Those subjects who could not be conditioned to increase 40Hz EEG had significantly less of this activity during the tasks condition of the Post-Test as compared to the tasks condition of the Pre-Test.
3. Strong and consistent dissociation between 40Hz EEG and 40Hz muscle activity, and 40Hz EEG and beta (21-30Hz) were demonstrated.
4. Significant increases in cognitive performance occurred across all groups between Pre- and Post-Tests and was

probably due to a general practice effect. However, one group, trained to increase 40Hz from the O_1-P_3 lead, showed increases in performance that could not be completely attributed to practice effect alone.

5. Voluntary control of 40Hz EEG in the absence of feedback information was shown for nine subjects, and this control persisted at least nine weeks after training.

6. Subjective state analysis indicated high arousal and mental concentration during 40Hz increase state, and low arousal and little mental effort during the 40Hz suppress state.

(C) Publication: Ph.D. Dissertation in Psychology

(D) Year: 1975

(E) Department: Psychology

(F) Student Name: Frederick A. Newton

(G) Faculty Advisor: Professor Daniel E. Sheer

Project

(A) Project Title: Biofeedback Training of 40 Hz EEG in Humans: Effects on Other EEG Rhythms and Autonomic Activity

(B) Project Abstract:

Biofeedback procedures have been employed to establish control over several EEG rhythms in humans and animals. This study demonstrated successful training of human subjects to increase and suppress a high-frequency EEG rhythm, centered at 40 Hz, using biofeedback procedures. Successful biofeedback training of EEG in one hemisphere also produced comparable changes in the opposite, untrained hemisphere. Substantial dissociation of 40Hz EEG from potential muscle contaminators of EEG and from Beta EEG occured in biofeedback training. No significant changes were found for Alpha and Theta EEG, and heart rate. However, objective measures of Subjective correlates of 40 Hz EEG showed that changes in subjective awareness followed biofeedback-produced changes in EEG. Data suggested that procedurea variables and variables related to individual differences significantly influenced EEG biofeedback learning. The potential of EEG biofeedback for research on EEG-behavior relations was discussed.

(C) Publication: Ph.D. Dissertation in Psychology

(D) Year: 1975

(E) Department: Psychology

(F) Student Name: Bruce L. Bird

(G) Faculty Advisor: Professor Daniel E. Sheer

J.433

Project

(A) Project Title: Biofeedback Training of 40 Hz EEG in Humans: Follow-Up on Control, Generalization of Effect, and Maintenance of Control During Problem Solving

(B) Project Abstract:

Long-term voluntary control of 40Hz EEG activity was investigated in six subjects, originally trained to increase and suppress 40Hz EEG in a previous study. The elapsed time between initial biofeedback training and follow-up control testing varied from one to three years. No practice sessions were held during this period. Subjects were first instructed to alternately produce and suppress 40Hz EEG with feedback. Feedback was terminated for subsequent periods if and when consistent control was shown. During the final session, subjects were given a battery of test items and were instructed to alternately produce and suppress 40Hz EEG while solving problems. Forty Hertz EEG was monitored from the O_1-C_z, O_2-C_z, P_3-C_z, P_4-C_z leads during training and problem solving periods. Forth Hertz EMG was recorded from neck-temporal muscles. On-line comparator circuits prevented counting 40Hz EMG as 40Hz EEG.

Significant control of 40Hz EEG, without feedback, was shown for five of the six subjects. One subject was erratic only in the production of 40Hz EEG. Significant control was

shown to generalize to the O_1-C_z, O_2-C_z, and P_3-C_z leads, regardless of which lead had been reinforced. The amount of 40Hz EEG during the suppression periods, while solving problems, was significantly greater than during the suppression periods without feedback.

It was concluded that, following biofeedback training, long-term voluntary control of 40Hz EEG can be maintained for long periods of time. Furthermore, though the greatest control was demonstrated at the conditioned lead, the effects did generalize to other nonconditioned leads, indicating that it is an overall state that is learned. Finally, 40Hz EEG could not be suppressed during problem solving periods as compared to suppression periods without feedback. This further supports the association between 40Hz EEG and mental activity.

Other topics which were investigated were the changes in alpha and beta production during the training and problem solving sessions and the performance aspect during the 40Hz EEG production and suppression periods while problem solving. Relevance of the results of this study to the training of MBD children was discussed.

(C) Publication: M.A. Thesis in Psychology

(D) Year: 1976

(E) Department: Psychology

(F) Student Name: Martin R. Ford

(G) Faculty Advisor: Professor Daniel Sheer

Project

(A) Project Title: Theoretical Aspects of (1) Hydraulic Deliquoring, (2) Rotary Drum Filtration, and (3) Disc Filtration

(B) Project Abstract:

This work deals with theoretical aspects of compressible cake filtration and continuous rotary vacuum filters.

Reduction of moisture in filter cakes is important when heat requirements for drying are excessive. During filtration, the porosity of the cake is lowest at the points of maximum accumulative drag or lowest hydraulic pressure. By reversing the flow through the cake, it is possible to reduce the average porosity and hence the moisture content of the cake.

An equation is developed for calculating filtration rates in a rotary vacuum drum filter, taking into account finite partitioning of the drum, variable hydrostatic head and medium resistance. Constant average specific resistance of the cake is assumed. Formula based on sectioning gives rates which vary as much as 15% from those presently in use. The emerging cake varies in thickness in a periodic manner because portions of a section are subjected to vacuum for different cake formation times. This variation in thickness of the emerging cake can be as much as 20% and

therefore accounts for the difference in the conventional and the proposed formula.

Analytical formulas for the overall filtration rate through a continuous rotary disc filter are developed. Previously theoretical equations have not been available for design purposes. In this derivation the following are taken into account:

1. Division of the filtration surface into N equal radial sections with inner and outer radii of R_1 and R_2
2. Separation of sections by a blank strip representing a dead area for flow
3. Variable hydrostatic head
4. Medium resistance.

It is shown that the inner radius R_1 can be optimized to produce a maximum rate of filtration. A simple rule is presented for calculating the inner radius. The flow rates of disc and rotary drum filters are compared.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1973

(E) Department: Chemical Engineering

(F) Student Name: Hemant Risbud

(G) Faculty Advisor: Professor F. M. Tiller

Project

(A) Project Title: Mechanical Expression, Stresses at Cake Boundaries and New CP Cell

(B) Project Abstract:

Expression or squeezing operation under constant pressure was analyzed taking into account the medium resistance. A non-linear partial integro-differential equation representint the squeezing operation was solved using numerical methods. A computer program was developed to calculate transient pressure profiles and the cake thickness as a function of time. The problem was solved in two different coordinate systems.

Apparatuses were developed to measure the stress distribution on the boundaries of filter cakes compacted under mechanical pressure. It was discovered that for thin cakes the stress distributions on the bottom and at top are bell-shaped and not flat profiles as assumed by investigators in this field.

An improved compression-permeability cell was developed. In this apparatus, the hydraulic pressure profile was measured inside the cake at the center and at the wall. The filtration resistance was determined from slopes of the curved profiles. Previous calculations have been made on the assumption of linearity of hydraulic pressure.

(C) Publication: Ph.D. Dissertation in Chemical Engineering

(D) Year: 1974

(E) Department: Chemical Engineering

(F) Student Name: Hemant Madhav Risbud

(G) Faculty Advisor: Professor F. M. Tiller

Project

(A) Project Title: A Chain of Factored Matrices for Routh Array Inversion and Continued Fraction Inversion

(B) Project Abstract:

A chain of factored matrices is derived for formulating the Routh array if the first column of the array is known. The factored matrices may be used to perform the conversion of a continued fraction into a rational function. A set of tables based on the second Cauer form of continued fraction expansion is also included. These tables are the approximated rational functions for the commonly used irrational function $\sqrt{s}$ and the transcendental function e^s.

(C) Publication: International Journal of Control, 13, No. 4, 691 (1971)

(D) Year: 1971

(E) Department: Electrical Engineering

(F) Student Name: L. S. Shieh

(G) Faculty Advisor: Professors W. D. Schneider and D. R. Williams

INT. J. CONTROL, 1971, VOL. 13, NO. 4, 691–703

A chain of factored matrices for Routh array inversion and continued fraction inversion†

L. S. SHIEH, W. P. SCHNEIDER and D. R. WILLIAMS

Department of Electrical Engineering, University of Houston, Houston, Texas 77004

[Received 17 March 1970]

A chain of factored matrices is derived for formulating the Routh array if the first column of the array is known. The factored matrices may be used to perform the conversion of a continued fraction into a rational function. A set of tables based on the second Cauer form of continued fraction expansion is also included. These tables are the approximated rational functions for the commonly used irrational function $\sqrt{s}$ and the transcendental function e^s.

1. Introduction

In 1877, Routh supplied the sufficiency condition for asymptotic stability of a linear, lumped, stationary system when the characteristic equation of the system is known. This condition can be obtained by checking the sign change of the first column of a Routh array. Kalman and Bertram (1960) directly constructed the Lyapunov function in Schwarz's coordinate by using the coefficients of the characteristic equation of the linear, lumped, stationary system. Chen and Chu (1966) linked the relationship between the Lyapunov function of Kalman and Bertram with the terms of the first column of the Routh array, and revealed the fact that the linear transformation matrix between Schwarz's coordinate and the phase variable coordinate could be constructed with the elements of the first column and those of the other elements of the array. Csaki and Lehoczky (1967) connected the relationship between the elements of the first column and weighting factors without considering the initial conditions.

Routh's algorithm and continued fractions were associated by Wall (1945) and Frank (1946). In these works, the elements of the first row are those of the denominator of a rational function, while the row two elements are replaced by the coefficients of the numerator of the rational function. The other row elements are obtained using Routh's algorithm. In this case the formulation of the zig-zag pattern no longer gives the true Routh array. It cannot be used to judge stability according to the sign changes in the first column. In order to simplify this discussion, we can define a problem in the following manner.

The elements of the first and second row of an array are given according to the Wall and Frank formulation. The other rows are determined according to the Routh algorithm. For simplicity we still call such a pattern a Routh array.

The problem to be discussed is that of determining the elements of the other columns of such a Routh array when the elements of the first column are known. We call this an inversion problem for the Routh array. Recently Chen and Shieh (1969) proposed an algorithm to systematically treat this inversion and they revealed the fact that the third, fifth, ..., rows of the Routh array can be

† Communicated by Professor D. R. Williams.

7.441

used to formulate a linear transformation matrix between the second Cauer form in state-space coordinates and the phase-variable coordinate. If this approach is used, many determinant calculations are required. In a letter related to this topic, Chen (1969) offers a formula for continued fraction inversion. In the method shown by Chen, many iterative calculations are required. Recently Chen and Shieh (1970) derived a matrix for control system design; this paper proposes an alternate expression for the matrix such that the inversion of the Routh array and continued fraction can be performed.

2. Analysis and observations

Consider a Routh array where the Routh elements are given and identified by double subscript notation:

$$\left.\begin{array}{lllll} A_{11} & A_{12} & A_{13} & A_{14} & \cdots \\ A_{21} & A_{22} & A_{23} & \cdots & \\ A_{31} & A_{32} & \cdots & & \\ A_{41} & \cdots & & & \\ \cdots & & & & \end{array}\right\} \tag{1}$$

The general formulation between the elements can be written:

$$A_{j,k} = A_{j-2,k+1} - \frac{A_{j-2,1}A_{j-1,k+1}}{A_{j-1,1}}, \quad j = 3, 4, \ldots, n+1, \quad k = 1, 2, \ldots, n-1. \tag{2}$$

Now let us define a new constant:

$$h_p = \frac{A_{p,1}}{A_{(p+1),1}}, \quad p = 1, 2, \ldots, \quad h_p \neq 0, \tag{3}$$

where the h_p is a quotient of the neighbouring terms of the first column. In order to observe the relationship between these h terms and the elements of the first two rows, we can express the Routh array elements

$$A_{j,k} \quad (k = 1, 2, \ldots; j = 3, 4, \ldots)$$

in terms of the h_p terms, $p = 1, 2, \ldots$, and the $A_{j,k}$ $(k = 1, 2, \ldots; j = 1, 2)$ terms and insert an extra column for the h_p terms in the Routh array. We will then have the following alternate form for eqn. (1):

$$\left.\begin{array}{lllll} & A_{11} & A_{12} & A_{13} & \cdots \\ h_1 < & & & & \\ & A_{21} & A_{22} & A_{23} & \cdots \\ h_2 < & & & & \\ & A_{12}-h_1A_{22} & A_{13}-h_1A_{23} & A_{14}-h_1A_{24} & \cdots \\ h_3 < & & & & \\ & A_{22}-h_2(A_{13}-h_1A_{23}) & A_{23}-h_2(A_{14}-h_1A_{24}) & \cdots & \\ h_4 < & & & & \\ & A_{13}-h_1A_{23}-h_3(A_{23}-h_2[A_{14}-h_1A_{24}]) & \cdots & & \\ \cdot < & & & & \\ & \cdots & & & \\ \cdot < & & & & \end{array}\right\} \tag{4}$$

A.442

From eqn. (3), it may be noted that

$$\left.\begin{aligned} h_1 &= \frac{A_{11}}{A_{21}}, \quad h_2 = \frac{A_{21}}{A_{12}-h_1 A_{22}}, \\ h_3 &= \frac{A_{12}-h_1 A_{22}}{A_{22}-h_2(A_{13}-h_1 A_{23})}, \\ h_4 &= \frac{A_{23}-h_2(A_{13}-h_1 A_{23})}{A_{13}-h_1 A_{23}-h_3[A_{23}-h_2(A_{14}-h_1 A_{24})]}, \\ &\ldots \end{aligned}\right\} \quad (5)$$

ORIGINAL PAGE IS
OF POOR QUALITY

We can rearrange eqn. (5) to obtain:

$$\left.\begin{aligned} A_{11} &= h_1 A_{21}, \\ h_2 A_{12} &= A_{21}+h_1 h_2 A_{22}, \\ A_{12}+h_2 h_3 A_{13} &= (h_1+h_3) A_{22}+h_1 h_2 h_3 A_{23}, \\ (h_2+h_4) A_{13}+h_2 h_3 h_4 A_{14} &= A_{22}+(h_1 h_2+h_1 h_4+h_3 h_4) A_{23}+h_1 h_2 h_3 h_4 A_{24} \\ &\ldots \end{aligned}\right\} \quad (6)$$

A matrix form for eqn. (6) can be formulated as follows:

$$\begin{bmatrix} \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & h_2 h_3 h_4 & (h_2+h_4) & 0 & 0 \\ \ldots & 0 & h_2 h_3 & 1 & 0 \\ \ldots & 0 & 0 & h_2 & 0 \\ \ldots & 0 & 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} \ldots \\ A_{14} \\ A_{13} \\ A_{12} \\ A_{11} \end{bmatrix} =$$

$$\begin{bmatrix} \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & h_1 h_2 h_3 h_4 & (h_1 h_2+h_1 h_4+h_3 h_4) & 1 & 0 \\ \ldots & 0 & h_1 h_2 h_3 & (h_1+h_3) & 0 \\ \ldots & 0 & 0 & h_1 h_2 & 1 \\ \ldots & 0 & 0 & 0 & h_1 \end{bmatrix} \begin{bmatrix} \ldots \\ A_{24} \\ A_{23} \\ A_{22} \\ A_{21} \end{bmatrix}. \quad (7)$$

In order to reduce the symbological complexity, we can define the $n \times n$ matrix of the left-hand side of eqn. (7) as $H_L{}^n$, and the $n \times n$ matrix of the right-hand side of the equation as $H_R{}^n$. With this notation the superscript n indicates the size of the square matrix while the subscript L or R indicates which side of the equation it occupies.

$$H_L{}^n = \begin{bmatrix} \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & h_2 h_3 h_4 & (h_2+h_4) & 0 & 0 \\ \ldots & 0 & h_2 h_3 & 1 & 0 \\ \ldots & 0 & 0 & h_2 & 0 \\ \ldots & 0 & 0 & 0 & 1 \end{bmatrix} \quad (7\,a)$$

and

$$H_R^n = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & h_1h_2h_3h_4 & (h_1h_2+h_1h_4+h_3h_4) & 1 & 0 \\ \cdots & 0 & h_1h_2h_3 & (h_1+h_3) & 0 \\ \cdots & 0 & 0 & h_1h_2 & 1 \\ \cdots & 0 & 0 & 0 & h_1 \end{bmatrix} \tag{7 b}$$

In general the H_L^n and H_R^n are not easy to remember, particularly when n is high; however, we can decompose H_L^n and H_R^n into a chain of matrices which may be called the factored form. To demonstrate this procedure we choose $n = 4$ this means that, h_1, h_2, h_3, h_4 being known, the 4×4 square matrix can be written as H_L^4 and H_R^4 which are obtained by a partitioning of the general matrix H_L^n and H_R^n. By taking the lower right-hand corner of the H_L^n and H_R^n matrices we determine:

$$H_L^4 = \begin{bmatrix} h_2h_3h_4 & (h_2+h_4) & 0 & 0 \\ 0 & h_2h_3 & 1 & 0 \\ 0 & 0 & h_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7 c}$$

and

$$H_R^4 = \begin{bmatrix} h_1h_2h_3h_4 & (h_1h_2+h_1h_4+h_3h_4) & 1 & 0 \\ 0 & h_1h_2h_3 & (h_1+h_3) & 0 \\ 0 & 0 & h_1h_2 & 1 \\ 0 & 0 & 0 & h_1 \end{bmatrix}. \tag{7 d}$$

Each H_L^4 and H_R^4 can be decomposed into four matrices as follows:

$$H_L^4 = \begin{bmatrix} h_4 & 1 & & \\ & h_3 & 1 & \\ & & h_2 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} h_3 & 1 & & \\ & h_2 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} h_2 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \tag{8}$$

and

$$H_R^4 = \begin{bmatrix} h_4 & 1 & & \\ & h_3 & 1 & \\ & & h_2 & 1 \\ & & & h_1 \end{bmatrix} \begin{bmatrix} h_3 & 1 & & \\ & h_2 & 1 & \\ & & h_1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} h_2 & 1 & & \\ & h_1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \begin{bmatrix} h_1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \tag{9}$$

We observed that H_L^4 can be obtained from eqn. (9) by taking the last three matrices from the right-hand side of eqn. (9) and substituting $j = 1, 2, \ldots$, by

p. 444

h_{j+1} for h_j. The last three matrices of eqn. (9) are then:

ORIGINAL PAGE IS
OF POOR QUALITY

$$\begin{bmatrix} h_3 & 1 & & \\ & h_2 & 1 & \\ & & h_1 & \\ & & & 1 \end{bmatrix}\begin{bmatrix} h_2 & 1 & & \\ & h_1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix}\begin{bmatrix} h_1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} h_j \to h_{j+1}, \quad j = 1, 2, \ldots.$$

$$= \begin{bmatrix} h_4 & 1 & & \\ & h_3 & 1 & \\ & & h_2 & \\ & & & 1 \end{bmatrix}\begin{bmatrix} h_3 & 1 & & \\ & h_2 & & \\ & & 1 & \\ & & & 1 \end{bmatrix}\begin{bmatrix} h_2 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix}$$

$$= H_L^4. \tag{10}$$

We also observed that the sequential distribution of h's in eqn. (9) is more regular than that of eqn. (8), therefore we will concentrate on eqn. (9).

In general if n terms of h's or h_j $(j = 1, 2, \ldots, n)$ are known then we have n factored matrices such that

$$H_R^n = \prod_{j=1}^{n} H_j = H_n . H_{n-1} . \ldots . H_1, \tag{11}$$

where H_j is a $n \times n$ matrix where the elements are such that if we define $H_j(l, k)$ as an element at lth row and kth column in H_j matrix then:

$$\left.\begin{aligned} H_j(l, l) &= h_{(j+1-l)}, \quad l = 1, 2, \ldots, j, \\ H_j(l, l+1) &= 1, \quad l = 1, 2, \ldots, j-1, \\ H_j(l, l) &= 1, \quad l = j+1, \ldots, n \quad \text{if} \quad j < n. \end{aligned}\right\} \tag{11 a}$$

All other elements in H_j are zero. Of course:

$$H_L^n = \prod_{j=1}^{n-1} H_j \Big|_{\text{substitute } h_{j+1} \text{ for } h_j \, (j = 1, 2, \ldots).} \tag{11 b}$$

From eqns. (7 *c*, *d*) we can make the following observations:

(1) The first row of the Routh array is in a one-to-one correspondence to the first row of the H_R^n matrix.

(2) The second row of the Routh array can be obtained by:
(i) substituting h_{j+1} for h_j $(j = 1, 2, \ldots)$ in the second row of the H_R^n matrix;
(ii) dropping the zero first element and shifting the complete second row one column to the left.

(3) The third row of the Routh array can be obtained by:
(i) substituting h_{j+2} for h_j $(j = 1, 2, \ldots)$ in the third row of the H_R^n matrix;
(ii) dropping the zeros of the first two elements and shifting the complete third row two columns to the left.

In general we discard the zeros at the lower triangular matrix of H_R^n and shift all the elements to the left. After this step we modify H_R^n by substituting h_{j+l-1} for h_j $(j = 1, \ldots, n; l = 2, \ldots, n)$ into each lth row of H_R^n. From this we

observe that the new formulated matrix, called [A], is the required Routh array, for example $n = 4$:

$$H_R{}^4 = \begin{bmatrix} h_1h_2h_3h_4 & (h_1h_2+h_1h_4+h_3h_4) & 1 & 0 \\ 0 & h_1h_2h_3 & (h_1+h_3) & 0 \\ 0 & 0 & h_1h_2 & 1 \\ 0 & 0 & 0 & h_1 \end{bmatrix}$$

$$(h_1 \to h_2, h_2 \to h_3, h_3 \to h_4) \quad (h_1 \to h_3, h_2 \to h_4) \quad (h_1 \to h_4).$$

and shift whole elements to the left:

$$\begin{bmatrix} h_1h_2h_3h_4 & (h_1h_2+h_1h_4+h_3h_4) & 1 & 0 \\ h_2h_3h_4 & (h_2+h_4) & 0 & 0 \\ h_3h_4 & 1 & 0 & 0 \\ h_4 & 0 & 0 & 0 \end{bmatrix}$$

$$= [A]$$

$$= \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix}. \tag{12}$$

Note that in this case the element $A_{(n+1,1)} = 1$ or $A_{5,1} = 1$ if the given $A_{(n+1,1)} \neq 1$ then all elements of $[A]$ must be multiplied by a normalization factor $A_{(n+1,1)}$. Thus we complete the Routh array inversion.

In case we are not given the first column, but all the h's are given, we know that the first row and the second row of the Routh array are in one-to-one correspondence to the coefficients of the denominator and the numerator of a respectively rational function which was arranged in ascending order. A continued fraction expansion can be formulated by these h's. If a rational function $f(s)$ is given:

$$f(s) = \frac{b_1+b_2s+b_3s^2+\ldots+b_{m-1}s^{m-1}}{a_1+a_2s+a_3s^2+\ldots+s^m}, \tag{13}$$

then we arrange the coefficient of the denominator of eqn. (13) as the first row of the Routh array or $A_{11} = a_1, A_{12} = a_2, \ldots$, and that of the numerator of eqn. (13) as the second row of the Routh array, i.e. $A_{21} = b_1, \ldots$, the Routh array, is:

$$\left.\begin{matrix} A_{11} & A_{12} & A_{13} & \ldots \\ A_{21} & A_{22} & A_{23} & \ldots \end{matrix}\right\} \tag{13 a}$$

following eqns. (2) and (3) respectively we have h_p $(p = 1, 2, \ldots)$ thus we can express eqn. (13) in a second Cauer form continued fraction expansion as follows:

$$f(s) = \cfrac{1}{h_1+\cfrac{s}{h_2+\cfrac{s}{\vdots \cfrac{s}{h_j}}}} \triangleq C(h_1, h_2, \ldots, h_j). \tag{13 b}$$

6

9.446

We are interested in the continued fraction inversion. In the case where all h's have the same values Shieh and Chen (1969) developed a simple algorithm to convert this kind of continued fraction into a rational function. In most cases all h's have different values. For the general case we can make a second observation.

The rational function corresponding to this continued fraction can be obtained from $H_R{}^j$ and $H_L{}^j$ if j terms of h's are given, i.e. the corresponding rational function is:

$$= \frac{H_L{}^j(1,1)+H_L{}^j(1,2)\,S+H_L{}^j(1,3)\,S^2+\ldots+H_L{}^j(1,j)\,S^{j-1}}{H_R{}^j(1,1)+H_R{}^j(1,2)\,S+H_R{}^j(1,3)\,S^2+\ldots+H_R{}^j(1,j)\,S^{j-1}}. \tag{14}$$

Recall that $H_R{}^j$ or $H_L{}^j$ is a $j\times j$ matrix which is obtained by taking the lower right-hand corner of the $H_R{}^n$ or $H_L{}^n$ matrix.

The above two observations can be proved by induction. The proof is elementary and the check is easily made by performing the Routh algorithm to the Routh array obtained. This proof is omitted.

Example I

A numerical example is given for illustrating the above steps. Given: the first column of Routh's array:

$$\begin{aligned} A_{11} &= 120,\\ A_{21} &= 120,\\ A_{31} &= 60,\\ A_{41} &= 20,\\ A_{51} &= 5,\\ A_{61} &= 1. \end{aligned}$$

Find: All the elements of Routh's array.

From eqn. (3) we have $h_1 = 1$, $h_2 = 2$, $h_3 = 3$, $h_4 = 4$ and $h_5 = 5$, we are given 5 h's therefore $n = 5$; formulate $H_R{}^5$ matrix by following eqn. (11):

$$H_R{}^5 = \begin{bmatrix} h_5 & 1 & & & \\ & h_4 & 1 & & \\ & & h_3 & 1 & \\ & & & h_2 & 1 \\ & & & & h_1 \end{bmatrix} \begin{bmatrix} h_4 & 1 & & & \\ & h_3 & 1 & & \\ & & h_2 & 1 & \\ & & & h_1 & \\ & & & & 1 \end{bmatrix} \begin{bmatrix} h_3 & 1 & & & \\ & h_2 & 1 & & \\ & & h_1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

$$\times \begin{bmatrix} h_2 & 1 & & & \\ & h_1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix} \begin{bmatrix} h_1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} h_1h_2h_3h_4h_5 & (h_1h_2h_3+h_1h_2h_5+h_1h_4h_5+h_3h_4h_5) & (h_1+h_3+h_5) & 0 & 0 \\ 0 & h_1h_2h_3h_4 & (h_1h_2+h_1h_4+h_3h_4) & 1 & 0 \\ 0 & 0 & h_1h_2h_3 & (h_1+h_3) & 0 \\ 0 & 0 & 0 & h_1h_2 & 1 \\ 0 & 0 & 0 & 0 & h_1 \end{bmatrix}, \tag{15}$$

p. 447

replacing h_j by h_{j+l-1} ($j = 1, 2, 3, 4, 5$; $l = 2, 3, 4, 5$) into each lth row of $H_R{}^5$ matrix we have:

$$\begin{bmatrix} h_1 h_2 h_3 h_4 h_5 & (h_1 h_2 h_3 + h_1 h_2 h_5 + h_1 h_4 h_5 + h_3 h_4 h_5) & (h_1 + h_3 + h_5) & 0 & 0 \\ 0 & h_2 h_3 h_4 h_5 & (h_2 h_3 + h_2 h_5 + h_4 h_5) & 1 & 0 \\ 0 & 0 & h_3 h_4 h_5 & (h_3 + h_5) & 0 \\ 0 & 0 & 0 & h_4 h_5 & 1 \\ 0 & 0 & 0 & 0 & h_5 \end{bmatrix},$$

then shift all elements to the left, and finally we have the required Routh array:

$$\begin{bmatrix} h_1 h_2 h_3 h_4 h_5 & (h_1 h_2 h_3 + h_1 h_2 h_5 + h_1 h_4 h_5 + h_3 h_4 h_5) & (h_1 + h_3 + h_5) & 0 & 0 \\ h_2 h_3 h_4 h_5 & (h_2 h_3 + h_2 h_5 + h_4 h_5) & 1 & 0 & 0 \\ h_3 h_4 h_5 & (h_3 + h_5) & 0 & 0 & 0 \\ h_4 h_5 & 1 & 0 & 0 & 0 \\ h_5 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Substituting all h values into it the required Routh array is:

$$\begin{array}{ccc} 120 & 96 & 9 \\ 120 & 36 & 1 \\ 60 & 8 & \\ 20 & 1 & \\ 5 & & \\ 1 & & \end{array}$$

ORIGINAL PAGE IS
OF POOR QUALITY

Example II

Consider a unit step input applied to a pure time delay system. The Laplace transform of the delay function is the transcendental function e^{-s}. The approximate rational functions of this function are required.

First we expand e^s as a Taylor series about $s = 0$:

$$e^s = 1 + S + \frac{1}{2!}S^2 + \frac{1}{3!}S^3 + \frac{1}{4!}S^3 + \ldots, \tag{16}$$

then:

$$e^{-s} = \frac{1 + 0S + 0S^2 + 0S^3 + \ldots}{1 + S + 1/2!\,S^2 + 1/3!\,S^3 + 1/4!\,S^4 + \ldots}, \tag{17}$$

we arrange eqn. (17) into the following Routh array

$$\begin{array}{cccccc} 1 & 1 & 1/2! & 1/3! & 1/4! & \ldots \\ 1 & 0 & 0 & 0 & 0 & 0 \quad \ldots. \end{array}$$

Apply eqns. (2) and (3) respectively to obtain:

$$h_1 = 1, \quad h_2 = 1, \quad h_3 = -2, \quad h_4 = -3, \quad \ldots.$$

Equation (17) can be expressed by a continued fraction in the form of eqn. (13 *b*) or:

$$e^{-s} \triangleq C[1, 1, -2, -3, 2, 5, -2, -7, \ldots, 2, 2M-1, -2, -(2M+1), \ldots],$$

if the first $4h$'s are taken then we formulate H_R^4 by eqn. (9) and construct H_L^4 by eqn. (10):

$$H_R^4 = \begin{bmatrix} h_1h_2h_3h_4 & (h_1h_2+h_1h_4+h_3h_4) & 1 & 0 \\ 0 & h_1h_2h_3 & (h_1+h_3) & 0 \\ 0 & 0 & (h_1h_2) & 1 \\ 0 & 0 & 0 & h_1 \end{bmatrix}$$

and

$$H_L^4 = \begin{bmatrix} h_2h_3h_4 & (h_2+h_4) & 0 & 0 \\ 0 & h_2h_3 & 1 & 0 \\ 0 & 0 & h_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

ORIGINAL PAGE IS
OF POOR QUALITY

since $j = 4$; by following eqn. (14) we have:

$$\begin{aligned} e^{-s} &\doteq \frac{H_L^4(1,1)+H_L^4(1,2)S+H_L^4(1,3)S^2+H_L^4(1,4)S^3}{H_R^4(1,1)+H_R^4(1,2)S+H_R^4(1,3)S^2+H_R^4(1,4)S^3} \\ &= \frac{h_2h_3h_4+(h_2+h_4)S}{h_1h_2h_3h_4+(h_1h_2+h_1h_4+h_3h_4)S+S^2} \\ &= \frac{6-2s}{6+4s+s^2}; \end{aligned}$$

if $3h$'s are taken, in this case $j = 3$:

$$\begin{aligned} e^{-s} &\doteq \frac{H_L^3(1,1)+H_L^3(1,2)S+H_L^3(1,3)S^2}{H_R^3(1,1)+H_R^3(1,2)S+H_R^3(1,3)S^2} \\ &= \frac{h_2h_3+S}{h_1h_2h_3+(h_1+h_3)S} \\ &= \frac{-2+S}{-2-S}; \end{aligned}$$

if only h_1 and h_2 are taken:

$$\begin{aligned} e^{-s} &\doteq \frac{H_L^2(1,1)+H_L^2(1,2)S}{H_R^2(1,1)+H_R^2(1,2)S} \\ &= \frac{h_2}{h_1h_2+S} \\ &= \frac{1}{1+S}. \end{aligned}$$

Example III

Consider a system where an irrational function $[1/\sqrt[3]{(s+1)}]$ is to be synthesized. Expanding the irrational function $\sqrt[3]{(s+1)}$ into a Taylor series about $s = 0$, we have:

$$\sqrt[3]{(s+1)} = (1+s)^{1/3} = 1+\frac{s}{3}-\frac{s^2}{9}+\frac{5s^3}{81}-\frac{10s^4}{243}+\ldots; \tag{18}$$

p.449

assume that $[1/\sqrt[3]{(s+1)}]$ can be expanded into the following continued fraction:

$$\frac{1}{\sqrt[3]{(s+1)}} = \cfrac{1}{h_1 + \cfrac{S}{h_2 + \cfrac{S}{h_3 + \cfrac{S}{\vdots}}}}, \tag{19}$$

where the h terms are unknown with their values to be determined. Instead of evaluating these constants by the Routh array we evaluate them by the following approach which has a more significant physical meaning.

The left-hand side of eqn. (19) is a system. Equating eqns. (18) and (19) we have:

$$\frac{1}{1+(s/3)-(s^2/9)+(5s^3/81)-(10/243)s^4+\ldots} = \cfrac{1}{h_1 + \cfrac{S}{h_2 + \cfrac{S}{h_3 + \cfrac{S}{\vdots}}}}; \tag{20}$$

let $s = 0$ and $h_1 = 1$ is immediately found. This can be interpreted as the addition of a unit step function to the given system; then apply the final value theorem to the system and from eqn. (20) we have the final value $1/1 = 1$ which is the reciprocal of h_1.

Substituting h_1 into eqn. (20) and taking the reciprocal of eqn. (20) we have:

$$1+\frac{s}{3}-\frac{s^2}{9}+\frac{5s^3}{81}-\frac{10s^4}{243}+\ldots = 1+\cfrac{S}{h_2 + \cfrac{S}{h_3 + \cfrac{S}{\vdots}}}. \tag{21}$$

Rewriting eqn. (21):

$$\frac{s}{3}-\frac{s^2}{9}+\frac{5s^3}{81}-\frac{10s^4}{243}+\ldots = \cfrac{S}{h_2 + \cfrac{S}{h_3 + \cfrac{S}{\vdots}}} \tag{22}$$

and adding a unit step function to eqn. (22) we have:

$$\frac{1}{3}-\frac{s}{9}+\frac{5s^2}{81}-\frac{10s^3}{243}+\ldots = \cfrac{1}{h_2 + \cfrac{S}{h_3 + \cfrac{S}{\vdots}}}. \tag{23}$$

Again setting $S = 0$, we have $h_2 = 3$. Repeating these processes gives:

$$h_3 = 1, \quad h_4 = \frac{9}{2}, \quad h_5 = \frac{4}{5}, \quad \ldots.$$

7.450

ORIGINAL PAGE IS
OF POOR QUALITY

Consequently, the complete expansion has been obtained:

$$\frac{1}{\sqrt[3]{(s+1)}} = \cfrac{1}{1+\cfrac{S}{3+\cfrac{S}{1+\cfrac{S}{\frac{9}{2}+\cfrac{S}{\vdots}}}}}. \tag{24}$$

We see that eqn. (24) is in the form of eqn. (13 *b*), therefore we can apply eqn. (14) to find the corresponding rational function. For instance if $h_1 = 1$, $h_2 = 3$, $h_3 = 1$ and $h_4 = \frac{9}{2}$ are given, and the corresponding rational function is required, from eqns. (7 *c*, *d*) and (14) we have:

$$\frac{1}{\sqrt[3]{(S+1)}} \doteq \frac{h_2 h_3 h_4 + (h_2+h_4)S}{h_1 h_2 h_3 h_4 + (h_1 h_2 + h_1 h_4 + h_3 h_4)S + S^2}$$

$$= \frac{13{\cdot}5 + 7{\cdot}5S}{13{\cdot}5 + 12S + S^2}$$

3. Conclusions

A chain of factored matrices are developed to find the inversion of Routh's array. Once the first column of the Routh array is given, the whole Routh array can be formulated by these factored matrices, and these factored matrices can be applied to continued fraction inversion.

It is believed that this approach is new and simpler than the methods (Chen and Shieh 1969, Chen 1969) mentioned above.

Appendix

Based on the second Cauer form's continued fraction expansion, the approximated rational functions for the often-used irrational function $\sqrt{s}$ and transcendental function e^s are listed in the following tables.

Define the symbol as follows:

$$h_1 + \cfrac{S}{h_2 + \cfrac{S}{h_3 + \cfrac{S}{\cfrac{\vdots}{h_j}}}} \triangleq h_1 + S(h_2, h_3, \ldots, h_j, \ldots).$$

p. 451

Table 1. Irrational function $\sqrt{s}$

$$\sqrt{s} = 1+(s-1)\,(2, 2, 2, \ldots, 2, 2, \ldots,) \quad |s|<\infty$$

$$\sqrt{s} \doteq \frac{s+1}{2}$$

$$= \frac{3s+1}{s+3}$$

$$= \frac{s^2+6s+1}{4s+4}$$

$$= \frac{5s^2+10s+1}{s^2+10s+5}$$

$$= \frac{s^3+15s^2+15s+1}{6s^2+20s+6}$$

$$= \frac{7s^3+35s^2+21s+1}{s^3+21s^2+35s+7}$$

$$= \frac{s^4+28s^3+70s^2+28s+1}{8s^3+56s^2+56s+8}$$

$$= \frac{9s^4+84s^3+126s^2+36s+1}{s^4+36s^3+126s^2+84s+9}$$

$$= \frac{s^5+45s^4+210s^3+210s^2+45s+1}{10s^4+120s^3+252s^2+120s+10}$$

$$= \frac{11s^5+165s^4+462s^3+330s^2+55s+1}{s^5+55s^4+330s^3+462s^2+165s+11}$$

$$= \frac{s^6+66s^5+495s^4+924s^3+495s^2+66s+1}{12s^5+220s^4+792s^3+792s^2+220s+12}$$

$$= \frac{13s^6+286s^5+1287s^4+1716s^3+715s^2+78s+1}{s^6+78s^5+715s^4+1716s^3+1287s^2+286s+13}$$

$$= \frac{s^7+91s^6+1001s^5+3003s^4+3003s^3+1001s^2+91s+1}{14s^6+364s^5+2002s^4+3432s^3+2002s^2+364s+14}$$

$$= \frac{15s^7+455s^6+3003s^5+6435s^4+5005s^3+1365s^2+105s+1}{s^7+105s^6+1365s^5+5005s^4+6435s^3+3003s^2+455s+15}$$

$$= \frac{s^8+120s^7+1820s^6+8008s^5+12\,870s^4+8008s^3+1820s^2+120s+1}{16s^7+560s^6+4368s^5+11\,440s^4+11\,440s^3+4368s^2+560s+16}$$

$$= \frac{17s^8+680s^7+6188s^6+19\,448s^5+24\,310s^4+12\,376s^3+2380s^2+136s+1}{s^8+136s^7+2380s^6+12\,376s^5+24\,310s^4+19\,448s^3+6188s^2+680s+17}$$

$$= \frac{s^9+153s^8+3060s^7+18\,564s^6+43\,758s^5+43\,758s^4+18\,564s^3+3060s^2+153s+1}{18s^8+816s^7+8568s^6+31\,824s^5+48\,620s^4+31\,824s^3+8568s^2+816s+18}$$

$$= \frac{19s^9+969s^8+11\,628s^7+50\,388s^6+92\,378s^5+75\,582s^4+27\,132s^3+3876s^2+171s+1}{s^9+171s^8+3876s^7+27\,132s^6+75\,582s^5+92\,378s^4+50\,388s^3+11\,628s^2+969s+19}$$

$$= \frac{s^{10}+190s^9+4845s^8+38\,760s^7+125\,970s^6+184\,756s^5+125\,970s^4+38\,760s^3+4845s^2+190s+1}{20s^9+1140s^8+15\,504s^7+77\,520s^6+167\,960s^5+167\,960s^4+77\,520s^3+15\,504s^2+1140s+20}$$

p. 452.

Table 2. Transcendental function

If e^{TS} is required we simply replace all the 'S' in table 2 by 'TS'

$$e^s = 1+s[1, -2, -3, 2, 5, -2, -7, \ldots, 2, 2n-1, -2, -(2n+1), \ldots] \quad |s|<\infty$$

$$e^s \doteq s+1$$

$$= \frac{-s-2}{s-2}$$

$$= \frac{s^2+4s+6}{-2s+6}$$

$$= \frac{s^2+6s+12}{s^2-6s+12}$$

$$= \frac{s^3+9s^2+36s+60}{3s^2-24s+60}$$

$$= \frac{-s^3-12s^2-60s-120}{s^3-12s^2+60s-120}$$

$$= \frac{s^4+16s^3+120s^2+480s+840}{-4s^3+60s^2-360s+840}$$

$$= \frac{s^4+20s^3+180s^2+840s+1680}{s^4-20s^3+180s^2-840s+1680}$$

$$= \frac{s^5+25s^4+300s^3+2100s^2+8400s+15\,120}{5s^4-120s^3+1260s^2-6720s+15\,120}$$

$$= \frac{-s^5-30s^4-420s^3-3360s^2-15\,120s-30\,240}{s^5-30s^4+420s^3-3360s^2+15\,120s-30\,240}$$

$$= \frac{s^6+36s^5+630s^4+6720s^3+45\,360s^2+181\,440s+332\,640}{-6s^5+210s^4-3360s^3+30\,240s^2-151\,200s+332\,640}$$

$$= \frac{s^6+42s^5+840s^4+10\,080s^3+75\,600s^2+332\,640s+665\,280}{s^6-42s^5+840s^4-10\,080s^3+75\,600s^2-332\,640s+665\,280}$$

$$= \frac{s^7+49s^6+1176s^5+17\,640s^4+17\,640s^3+1\,164\,240s^2+4\,656\,960s+8\,648\,640}{7s^6-336s^5+7560s^4-100\,800s^3+831\,600s^2-3\,991\,680s+8\,648\,640}$$

$$= \frac{-s^7-56s^6-1512s^5-25\,200s^4-277\,200s^3-1\,995\,840s^2-8\,648\,640s-17\,297\,280}{s^7-56s^6+1512s^5-25\,200s^4+277\,200s^3-1\,995\,840s^2+8\,648\,640s-17\,297\,280}$$

$$= \frac{s^8+64s^7+2016s^6+40\,320s^5+554\,400s^4+5\,322\,240s^3+34\,594\,560s^2+138\,378\,240s+259\,459\,200}{-8s^7+504s^6-15\,120s^5+277\,200s^4-3\,326\,400s^3+25\,945\,920s^2-121\,080\,960s+259\,459\,200}$$

ORIGINAL PAGE IS
OF POOR QUALITY

References

Chen, C. F., and Chu, H., 1966, *I.E.E.E. Trans. autom. Control*, **11**, 303.

Chen, C. F., and Shieh, L. S., 1969, *I.E.E.E. Trans. Circuit Theory*, **16**, 197; 1970, *Int. J. Control*, **11**, 717.

Chen, C.-T., 1969, *Proc. Instn elect. electron. Engrs*, **57**, 1780.

Csaki, F., and Lehoczky, J., 1967, *I.E.E.E. Trans. Autom. Control*, **12**, 462.

Frank, E., 1946, *Am. math. Mon.*, **53**, 144.

Kalman, R. E., and Bertram, J. E., 1960, *Trans. Am. Soc. mech. Engrs*, Series D, 371.

Shieh, L. S., and Chen, C. F., 1969, *I.E.E.E. Trans. Aerospace electron. Syst.*, **5**, 967.

Wall, H. S., 1945, *Am. Math. Monthly*, **52**, 308.

p.453

Project

(A) Project Title: Computer Aided Design and Economic Evaluation of Chemical Processes

(B) Project Abstract:

Economic evaluation is the tool used to minimize the risks involved in the development of projects. A computer program was developed which makes the economic evaluation for chemical processes. The program computes the cost of the individual pieces of equipment and the fixed-capital investment (±15%) by Miller's refined factored method. Manufacturing cost is also computed by the program. Costs of raw materials and catalysts must be supplied to the program for this purpose; utilities costs are computed by means of material and energy balance. These data are used to compute a return on investment.

Routines were developed for the design and/or cost estimation of the fo-lowing pieces of equipment: distillation and absorption columns, heat exchangers (single phase, condenser, kettle reboilers), reactors, furnaces, pumps, compressors, and tanks.

The program uses the thermodynamic package contained in CHESS. Flow rates, compositions and state of all process streams must be supplied for the use of the program. This information can be obtained by the use of a simulation program such as CHESS. It is possible to integrate the program

with CHESS so that the material and energy balance, and the economic evaluation of the process would be computed by the same program.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1970

(E) Department: Chemical Engineering

(F) Student Name: Jorge Enrique Villalobos

(G) Faculty Advisor: Professor F. L. Worley, Jr.

J. 455

Project

(A) Project Title: Optimization Techniques in Computer Aided Design of Chemical Processes

(B) Project Abstract:

This work is a continuation of an earlier one developed by J. E. Villalobos in computer aided process design.

It is a study about a chemical engineering design and economic evaluation package named CHEEP, and optimization techniques in order to incorporate the compatible ones to the package.

The purpose of this project is to make available to undergraduate students in chemical engineering process design class, optimization techniques which can be used as tools in obtaining a better design.

These optimization techniques written in FORTRAN-IV are available either on the IBM/360 or the UNIVAC 1108 computers.

The optimization techniques selected were Golden section search (SUBROUTINE MAD3) and Rosenbrock's method (SUBROUTINE MAD2) for unidimensional and multidimensional cases, respectively.

During the incorporation of the optimization subroutines some modifications and adjustments were made to the design and economic evaluation package.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1971

(E) Department: Chemical Engineering

(F) Student Name: Benito Lopez Nakazono

(G) Faculty Advisor: Professor F. L. Worley, Jr.

B.457

Project

(A) Project Title: Control Equipment Design and Analysis

(B) Project Abstract:

I. Introduction

II. System Description

III. Equipment Module Descriptions

IV. Coding Instructions

Bibliography

Appendix A Logic Diagrams and Sample Calculations of Equipment Modules

Appendix B Logic Diagram, Sample Calculations and Subroutine Glossary for the System

Appendix C Example Problems

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1972

(E) Department: Chemical Engineering

(F) Student Name: Irvin Simon Fisch

(G) Faculty Advisor: Professor F. L. Worley, Jr.

Project

(A) Project Title: Mechanisms for the Removal of Sulfur Dioxide from the Atmosphere

(B) Project Abstract:

The removal of sulfur dioxide from the atmosphere by photochemical oxidation and by absorption into rain or fog droplets with possible subsequent catalytic oxidation in solution has been simulated. Photochemical oxidation of sulfur dioxide is represented by a first order reaction mechanism with a maximum rate constant of 10 percent per hour. Absorption of sulfur dioxide into rain or fog droplets is characterized using conventional mass-transfer mechanisms which account for the reversibility of the absorption and for the liquid phase mass-transfer resistance. Catalytic oxidation is represented by a mechanism relating oxidation rate to sulfur dioxide and metal oxide concentration in solution. The initial sulfur dioxide distribution in the atmosphere was determined using the binomial continuous plume equation and Holland's equation for plume rise. Simulation of these mechanisms was made by use of a digital computer.

Results of the simulation show that absorption of sulfur dioxide into rain droplets is a more efficient sulfur dioxide removal process than is photochemical oxidation at all rainfall rates and raindrop sizes studied. Absorption

of sulfur dioxide into liquid fog droplets was found to be an ineffective sulfur dioxide removal process due to the small mass of liquid water present in fogs.

Oxidation of sulfur dioxide in solution in the presence of metal oxide catalysts is found to increase the overall removal of sulfur dioxide from the atmosphere and to be a reasonable mechanism to explain the production of sulfates in precipitation.

It is proposed that the reported reduction in the rate of catalytic oxidation in solution with time is due to the reduction in aqueous sulfur dioxide solubility caused by the sulfates produced in the catalytic oxidation reaction lowering the pH of the solution.

(C) Publication: M.S. Thesis in Chemical Engineering

(D) Year: 1976

(E) Department: Chemical Engineering

(F) Student Name: George William Maltsberger, Jr.

(G) Faculty Advisor: Professor F. L. Worley, Jr.

APPENDIX A

Names of Departments and Faculty Members

Chemical Engineering

	Ph.D.	M.S.
Prof. A.E. Dukler	2	1
Prof. R.W. Flumerfelt	1	-
Prof. E.J. Henley	3	9
Prof. C.J. Huang	1	-
Prof. Dan Luss	1	1
Prof. R.L Motard	2	1
Prof. H.W. Prengle, Jr.	4	7
Prof. F.M. Tiller	1	1
Prof. F.L. Worley, Jr.	-	4

Civil Engineering

Prof. N.H.C. Hwang	-	1

Electrical Engineering

Prof. C.J. Chen	6	-
Prof. L.S. Shieh*	-	-
Prof. D.R. Williams	1	-
Prof. J.F. Pasknsz*		
Prof. W.P. Schneider*		

Mechanical Engineering

Prof. B.D. Cook	1	-
Prof. Charles Dalton	3	3
Prof. R.D. Finch	2	-

	Ph.D.	M.S.
Prof. J.R. Howell	3	2
Prof. C.D. Michalopoulous	3	1
Psycology		
Prof. D.E. Sheer	2	1
Total	36	32

* Faculty Research Projects

APPENDIX B

Typical Laboratory/Homework Assignments

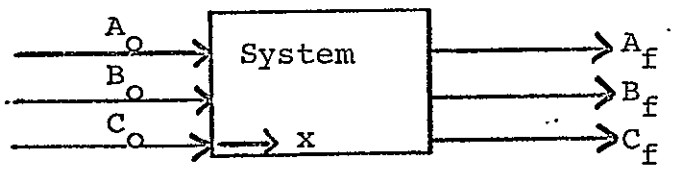
EGR 630 HYBRID COMPUTATION

UNIVERSITY OF HOUSTON

EGR 630 Hybrid Computation

System Identification Problem

A distributed parameter system receives as input three variables A_o, B_o, C_o and produces as output A_f, B_f, C_f. The governing differential equations for the system are assumed to be:



$x_{max} = 1$

$$\frac{dA}{dx} = -k_1 A + k_2 B$$

$$\frac{dB}{dx} = k_1 A - k_2 B + k_4 C - k_3 B$$

$$\frac{dC}{dx} = k_3 B - k_4 C$$

where x is the scaled independent space variable.

The maximum values of A, B, C are all 1.0, but scaling should be carefully considered anyway.

The unknown parameters k_1, k_2, k_3, k_4 are to be determined from a set of ten experimental runs on the physical system.

OBSERVED DATA

Run	A_o	B_o	C_o	A_f	B_f	C_f
1	.80	.10	.10	.0286	.1333	.8392
2	.75	.15	.10	.0277	.1317	.8415
3	1.0	0.	0	.0323	.1410	.8282
4	.10	0	.90	.0111	.1032	.8860
5	0	.50	.50	.0111	.1042	.8850
6	.25	.50	.25	.0176	.1150	.8685
7	.95	0	.05	.0317	.1390	.8300
8	.40	.20	.40	.0200	.1192	.8610
9	0	0	1.	.0094	.0988	.8923
10	.50	0	.50	.0211	.1200	.8600

Develop a hybrid program which iteratively generates the ten experimental conditions and computes a global error criterion ($y_i = A_f$, B_f or C_f)

$$\Sigma(y_i \text{ obs} - y_i \text{ calc})^2$$

This error is transmitted through a DAC to the analog computer digital voltmeter. DVM mode on PP and DAC output is wired to DVM IN. Initial conditions A_o, B_o, C_o are reset by the digital between every repetitive operation cycle. The "best" set of k_i can be found by manually adjusting the pots which represent them and minimizing the error.

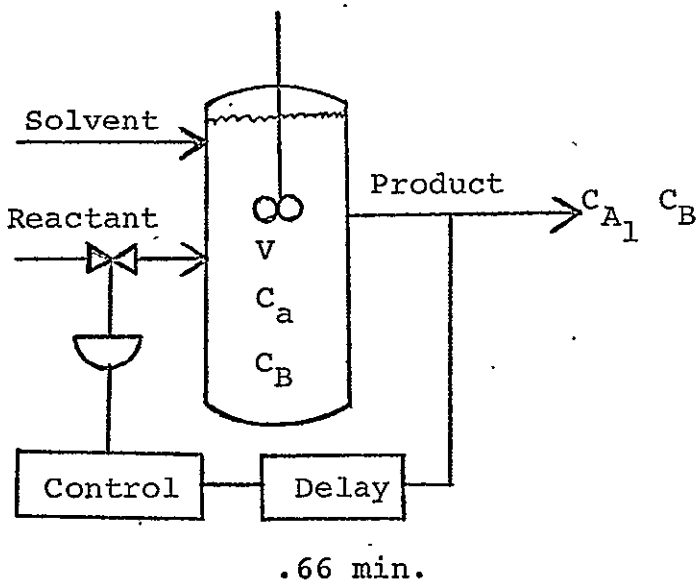
Develop a hybrid program which automatically searches for the best set of k_i.

University of Houston

EGR 630 HYBRID COMPUTATION

Problem 2

A chemical reactor receives a controlled input of a solvent, and the flow of reactant is controlled by the concentration of product, B which must not exceed 1.3 lb/gal.



Dynamic material balances on A and B are as follows:

$$Q_S C_{A_o} + \rho_A Q_A - C_A (Q_S + Q_A) - kC_A^2 V = V \frac{dC_A}{dt} - C_B (Q_S + Q_A)$$

$$+ kC_A^2 V = V \frac{dC_B}{dt}$$

where:

Q_S = Flow of solvent = 50 gal/min

Q_A = flow of reactant = 10 gal/min

C_{A_o} = solvent concentration of A = 0.4 lb/gal

ρ_A = reactant density = 9.04 lb/gal

k = reaction rate constant = 0.1418 gal/(lb - min)

V = reactor volume = 600 gal

Steady State Values:	Maximum Values:
Q_S = 50 gal/min	Q_S = 70 gal/min
Q_A = 10 gal/min	Q_A = 20 gal/min
C_A = 0.84 lb/gal	C_A = 1.5 lb/gal
C_B = 1.0 lb/gal	C_B = 1.3 lb/gal
C_{A_o} = 0.4 lb/gal	C_{A_o} = 2.0 lb/gal

The controller is a derivative-integral controller where

$$\frac{M(s)}{E(s)} = \frac{K_c(T_D s + 1)}{(T_I s + 1)}$$

and

$$M(s) = L\ \{m(t)$$
$$E(s) = L\ \ e(t)$$
$$m(t) = Q_A - A_A \text{ (steady state)}$$
$$e(t) = C_B(\text{set}) - C_B$$

$K_c = 100$; $T_I = 15.$; $T_o = 3.3$

The delay due to sampling may be approximated by a second order Padé form (p. 227 EAI Handbook of Analog Computation).

Plot the system response, C_B , on a strip chart recorder for a step change in C_B(set) from 1.0 to 1.1 . Also for a change in solvent flow from 50. to 70 gal/min.

APPENDIX C

Engineering Systems Simulation Laboratory

Cullen College of Engineering

University of Houston

The Cullen College of Engineering at the University of Houston houses one of the most advanced hybrid system in existence. The equipment is described in the following sections. This facility has the hardware and software to allow computer usage which covers the spectrum from pure analog to pure digital.

In general, the hybrid computer provides the following advantages:

- Combines the speed of the analog computer with the accuracy of the digital computer.
- Permits the use of system hardware or a"real world" analog of this hardware in a digital simulation.
- Increase the flexibility of an analog simulation by using digital memory and control.
- Increase the speed of digital computation by utilizing analog subroutines.
- Provides real time system simulation with man-made interfacing.

A partial list of the areas of application are;

- Simulation of physical systems.
- Analyses of sampled data systems.
- Random process simulation.
- System optimization.
- Simulation of distributed parameter systems.
- Analyses and studies in guidance and control of high speed devices.
- Simulation of man-machine systems.
- Process control system analysis and optimization.

The computing equipment in the Engineering Systems Simulation Laboratory of the Cullen College of Engineering, University of Houston, consists of an IBM Model 44 digital computer which has 128K bytes of core memory, two high-speed multiplexer channels, a low-speed multiplexer channel and the 32-level priority interrupt feature. Three 2311 disc drives are switchable between the two high-speed multiplexer channels. The single-disc storage drive in the model 44 processing unit is attached to a subchannel on one high-speed multiplex channel. There are two 9-track tape drives, a card read punch, a line printer, and a console typewriter attached to the multiplexer channel. The digital computer configuration is shown in Figure 1.

Communication between the 360/44 and the analog computer is through a Hybrid Systems Model 1044 Hybrid Linkage unit. Data transfer between the linkage and the digital computer is split between the two high-speed multiplexer channel. Digital data for conversion at the interface is transmitted from the 360 to the linkage over a subchannel of the second high-speed multiplex channel. The "split configuration" of the linkage, insofar as A-to-D and D-to-A activity is concerned is probably unique to this installation. The purpose of such a split is that much higher effective data rates are possible with the input and output activity operating asynchronously over two separate channels.

ENGINEERING SYSTEMS SIMULATION LABORATORY

IBM System 360/44
Digital Computing System

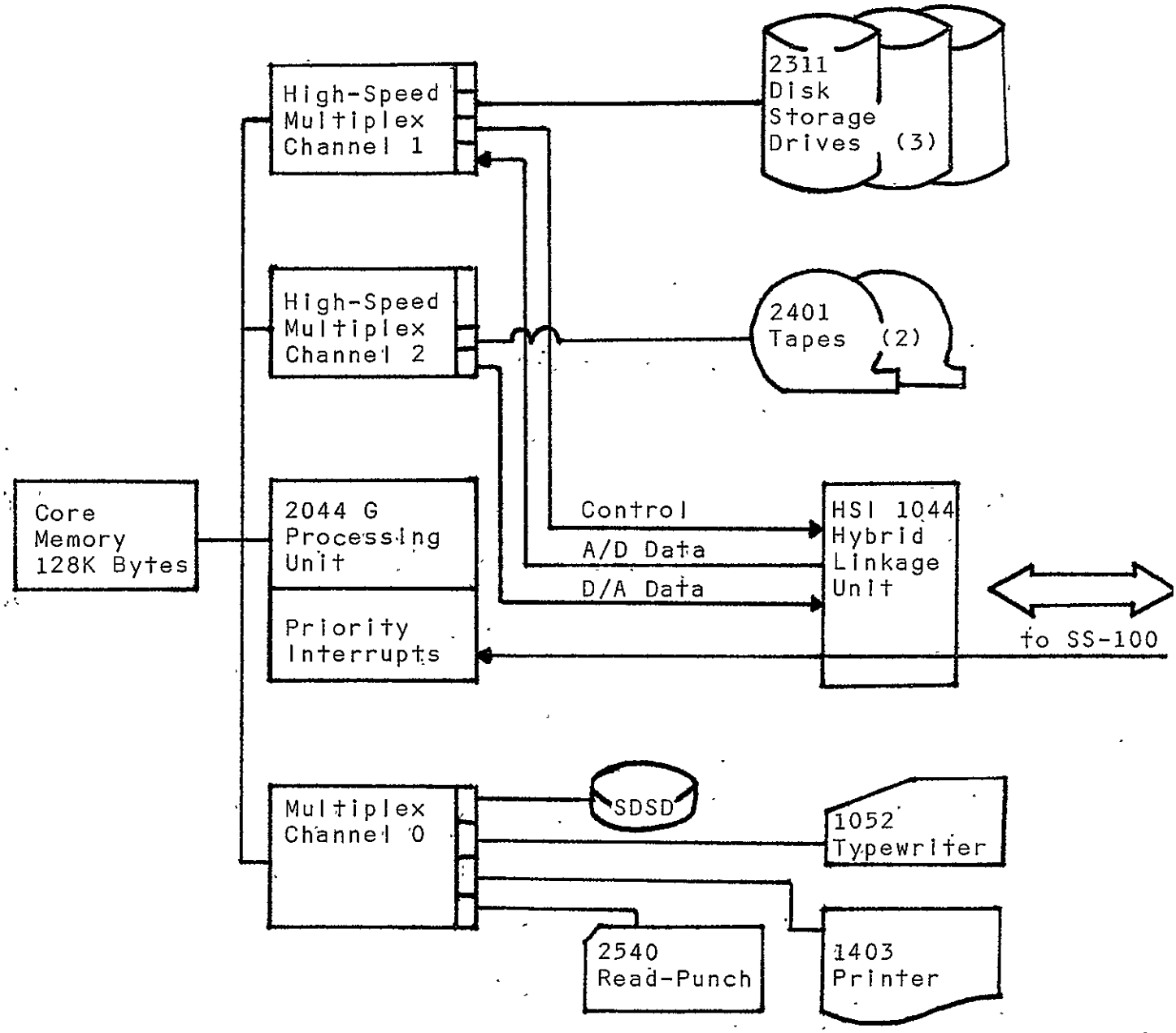


FIGURE 1

p.472

The analog computer is a Hybrid Systems Incorporated, Model SS100. Components of the analog computer summarized in Figure 2 include 54 summer/inverter amplifiers, 36 summer/integrator amplifiers, 80 servo-set potentiometers, manual potentiometers, D-to-A switches, sample-and-hold amplifiers, comparators, multipliers, sine-cosine generators and most of the lesser analog and logic components associated with an analog computer of this size and of recent manufacture. Presently Analog output is being handled by photo-recording from an oscilloscope or by digitizing and line printing.

Major software elements for the system are shown in Figure 3. The operating system for the 360/44 used in this hybrid environment is the Data Acquisition Multiprogramming System (DAMPS-II). This system is a rather elaborate extension of the programming system for the 360/44, 44PS and is furnished by IBM for the 360/44 in real-time applications. The functions of DAMPS are supplemented and extended by a system of programs, run under DAMPS, known as the Hybrid Executive. These extensions were developed by Hybrid Systems, and serve to expand the functions of DAMPS in a manner appropriate to hybrid computation.

Two additional software packages are also available. The first of these is called ANASET. Facilities are provided in ANASET for setup, static check, dynamic check, and execution monitoring of programs for the hybrid computer, utilizing a

Hybrid Systems SS-100 Analog Computing System

54	Summer-inverter amplifiers
36	Summer-integrator amplifiers
6	Dual inverter amplifiers
80	Servo-set potentiometers (with slew control)
12	Manual potentiometers
16	Digital-analog switches
16	Sample-and-hold amplifiers
12	Comparators
2	Sine-cosine generators
5	Feedback limiters
8	Quarter-square multipliers
20	Log(x) diode function generators
5	Bridge limiters
12	SPDT function relays
8	Shift registers (32 flip-flops)
50	AND/NAND gates
2	Three-mode timers
2	Single-shot multivibrators
8	Logic pushbuttons
16	Logic-level indicators
16	Analog to digital conversion multiplexer channels (13 bits + sign)
16	Multiplying digital to analog converters
16	Sense lines
32	Control lines
16	Priority interrupts
	Mode and timer control features
	Analog address and readout system
	6 Channel recorder
	x-y Plotter 8 1/2" x 11"
	Variable persistence oscilloscope

FIGURE 2

HYBRID PROGRAMMING SUPPORT SOFTWARE

DAMPS - II

Hybrid Executive

ANASET

AutoTrak

FIGURE 3

1.475

specifically restricted FORTRAN-IV statement structure for reference to analog computer components. The second subsidiary software package is called AutoTrak. AutoTrak permits the display, monitoring and modification of variables at execution-time in FORTRAN IV-coded programs at the console typewriter. Reference to these FORTRAN variables may be made either by variable name, absolute address, or by relative address. Combination of ANASET with AutoTrak allows the modification of potentiometer settings, scale factors, and other variables pertinent to the analog computer from the digital computer at the time the ANASET program is being executed.

The memory of the 360/44 under DAMPS is divided into two user partitions, a background partition and a real-time partition. DAMPS functions are summarized in Figure 4. Processing in the background partition is in conventional batch mode. 44PS jobs coded in FORTRAN-IV or assembler language can be executed in the roughly 49,000 byte area provided in the background partition. All the facilities of 44PS including overlay are available to programs being processed in the background under DAMPS. Activity in the real-time partition, the execution of real-time jobs (RTJs), falls essentially in two categories. The processing associated directly with priority interrupts is handled by so-called

DAMPS Functions

(Hybrid Programming Support)

Identify RTJs

Queue FGTs

Dequeue FGTs

Terminate FGTs

Attach RTTs to PILs

Terminate RTTs

Enable PILs

Disable PILs

Terminate RTJs

Real-Time Input-Output

- Set up controls for RTIO
- Perform (overlapped) RTIO
- Check RTIO

FIGURE 4

real-time tasks, (RTTs) one RTT being attached to each priority interrupt level (PIL). RTTs may be coded either in FORTRAN-IV or in assembler language, but in either case are routines which execute very rapidly in response to interrupts. More lengthy computational tasks, such as non-real-time input/output, etc. may be queued by RTTs for execution in the real-time partition upon completion of all pending priority-interrupt processing. These foreground tasks (FGT's) have all facilities of the DAMPS supervisor available. DAMPS-II provides a real-time partition of approximately 37,000 bytes of core storage.

The batch jobs constituting the background job stream are input through the card reader. In this two-partition mode, the loading of RTJs is control-led from the console typewriter, the programs making up the RTJ being loaded from a previously compiled and linkage-edited program library. The initiation of an RTJ consists essentially of entry of appropriate job control statements from the typewriter. In the event that a real-time job requires more storage than that available in the real-time partition, DAMPS-II may be operated in a single-partition mode, providing both the space of the real-time partition and that of the background partition to the real-time job. Control statements for the real-time job executed in this one-partition mode are then entered through the card reader.

DAMPS provides a number of functions to the real-time problem programmer in the form of subroutines which, as are appropriate, may be called from RTTs or from FGTs. Activities falling into this category include the identification of RTJs, the queueing, dequeueing, and termination of FGTs, the enabling and disabling of PILs, the attachment of RTTs to PILs, the termination of RTTs, non-real-time I/O from FGTs, real-time I/O from RTTs, (both overlapped and not overlapped with other processing), and the termination of RTJs. Functions provided in a similar manner by the Hybrid Executive shown in Figure 5 include the dynamic allocation of RTTs to PILs, display of the FGT queues, the construction of so-called uncontrolled RTTs, overlapped non-real-time I/O, manipulation of the interval timer, programmable delay, internal triggering of PILs, a set of pseudo-sense switches, and execution-time accounting. Time and space do not permit a detailed treatment here of features and operation of DAMPS and the Hybrid Executive, nor for that matter, of ANASET and AutoTrak.

The linkage interface unit, HSI 1044, is described in Figure 6. As mentioned earlier the linkage for this particular system includes dual control logic and dual bi-directional registers to allow independent, asynchronous operation of analog input (A/D) and analog output (D/A). The control functions shown in Figure 6 are actually duplicated in the interface.

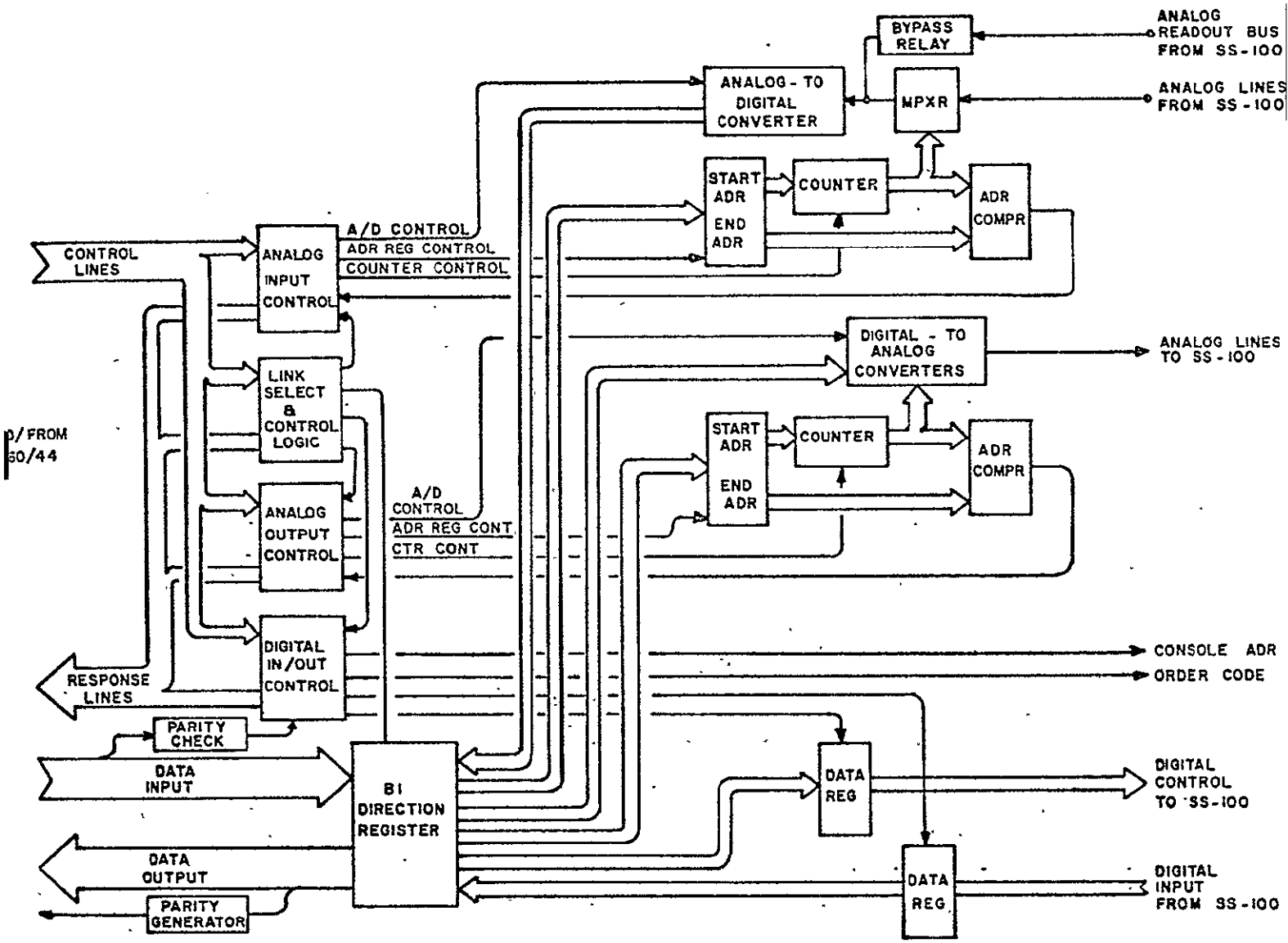
HYBRID EXECUTIVE FUNCTIONS

(Supplementing DAMPS)

Dynamic PIL - RTT allocation
Display FGT Queues
Trigger PILs internally
Attach uncontrolled RTTs
Overlapped non-RTIO
Interval timer handling
Pseudo-sense switches
Programmable delay
Execution-time accounting
Data conversion functions and control
Analog mode functions
Single bit communication
Addressing and read-out of analog components
Potentiometer setting
Timer Control

FIGURE 5

ORIGINAL PAGE IS
OF POOR QUALITY



BLOCK DIAGRAM OF 1044

FIGURE 6

p.481

The interface includes the control features for sequential and random addressing of A/D multiplexer channels and the multiplying D/A converters. In addition, these data transmission activities can be operated with the CPU channel in either the burst or multiplex mode. When the channel is in the multiplex mode (essential for random addressing) the transmission of A/D multiplexer or D/A converter addresses is interleaved with the input or output of analog data. The interface operation can also be controlled in either mode or direction by external synchronization where the conversion of each sample is under the control of a logic signal from a clock, gate, comparator or timer at the analog console. In the sequential, internally synchronized mode the data transfer rate approaches 100,000 samples per second concurrently on A/D and D/A channels. This is equivalent to a transfer rate of 400,000 bytes per second.

Single bit data transfers for analog control (control lines) or status sensing at the analog (sense lines) can also be controlled through the interface. All other modes of analog control, addressing and readout are available to the digital program. Thus, all analog computer elements may be addressed and read out over the A/D converter, the various modes such as IC, OP, HOLD, POTSET, RATE TEST, STATIS TEST can be controlled from the digital computer. Four decade timer intervals, clock rate and mode such as 2-mode repetitive

operation (IC, OP, HOLD) can also be initialized from the IBM 360. Analog read-out is done through the A/D converter automatically without patching at the analog console in preference to reading the digital voltmeter which would slow down the process considerably.

The analog elements are designed with a band width of roughly 100 KH for diode function generators (multipliers, log and sine-cosine functions) and 300 to 500 KH for the linear equipment (amplifiers). Switching times are on the order of one half micro-second or less for logic elements, D/A switches and mode control switches. Noise levels, unfiltered, are in the 1-5 multivolt range depending on component class. All switching is electronic with liberal use of FET switches. Two or three-mode repetitive operation at rates approaching 1000 problems solutions per second are available. Comparators have latching controls, sample and hold amplifiers may be used either as an interface between the A/D multiplexer or in stand-alone operation, D/A converters may be used as digital altenators. Sample and hold amplifiers may also be used as high speed integrators.

Four capacitor selected integration time scales are available; 1.0, .1, .01 and .001 microfarad.