BCL

REPORT NO. 7.3

TITLE: STRUCTURAL AND DYNAMIC ASPECTS OF SPEECH

AUTHOR: Ian B. Thomas

DATE: April 1, 1968

SPONSOR: AFOSR 7-67, AF 33(615)-3890, NASA NGR-14-005-111

BIOLOGICAL COMPUTER LABORATORY

DEPARTMENT OF ELECTRICAL ENGINEERING, UNIVERSITY OF ILLINOIS, URBANA, ILLINOIS

STRUCTURAL AND DYNAMIC ASPECTS OF SPEECH SIGNALS

by

Ian B. Thomas
Electrical Engineering Department
University of Massachusetts
Amherst, Massachusetts 01002

BCL Report No. 7.3

April 15, 1968

This work was jointly sponsored at the University of Illinois by

U.S. Air Force Office of Scientific Research
AF-AFOSR Grants 7-66, 7-67

U.S. Air Force Systems Engineering Group
Contract AF 33(615)-3890

National Aeronautics and Space Administration
Grant NGR 14-005-111

BIOLOGICAL COMPUTER LABORATORY
DEPARTMENT OF ELECTRICAL ENGINEERING
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS

ACKNOWLEDGEMENTS

The author wishes to express his thanks to Professors Heinz Von Foerster, W. Ross Ashby, and Murray L. Babcock of the Biological Computer Laboratory at the University of Illinois for their guidance and valuable suggestions.

To the sponsoring agencies of this research - the Air Force Office of Scientific Research, the Air Force Systems Engineering Group, and the National Aeronautics and Space Administration - the author is indeed grateful.

ABSTRACT

It has been shown that distorted natural speech, in which all formant information except the second formant frequency has been suppressed, yields articulation scores as high as 92% for monosyllabic English words out of context. Whistled speech, in which the period of a multivibrator is controlled by a second formant tracker, yields articulation scores near zero. If a random noise source, or a source with period equal to the glottal rate is used to vary the zero-axis crossing times of the whistled speech, the sounds become speech-like and articulation scores increase. The spectrograms for these three approximations to speech are virtually indistinguishable. It appears that the gross formant structure (dynamic information) conveys the word-based cues necessary for distinguishing among different sounds, whereas the excitation (structural information) conveys the cue that the sound is indeed a speech sound.

TABLE OF CONTENTS

INTRODUCTION

The advent of computers has created a need for communicating with machines in natural language. This need has been intensified by the desire of scientists to have direct verbal intercommunication with space vehicles and the sensing equipment used in deep space probes. Finally, the relative scarcity of bandwidth available for communicating with these vehicles makes bandwidth compression of audio signals imperative.

In nearly all studies of human speech made so far, emphasis has been placed on finding those features of a particular speech sound which distinguish it from all other speech sounds. Opinion has differed as to whether a bionic or a non-bionic approach should be used in these investigations. It can be argued, for example, that a machine is not necessarily subject to the same constraints and limitations as the ear-brain system and that the parameters and the processes used by this system to identify speech sounds may not, therefore, be the more efficient and reliable for achieving speech recognition by machine. If the communication channel is one way only, from man to machine, then the above reasoning is entirely justified. If, however, communication is to be established in the other direction, from machine to man, the speech synthesized by the machine must include those features of the speech signal which enable sounds or words to be differentiated by a human listener. But this is not all. It will be shown that before a sound can be processed by the ear-brain system as a speech sound at all, it must contain not only the distinguishing features of a particular word or sound but also contain certain structural features common to all speech sounds.

Before proceeding to the description of an experiment which places the importance of such structural information beyond doubt, it is well to consider the implications of some of the work already performed in speech analysis and speech synthesis. It has been well established that spectrograms of speech contain sufficient information to enable any speech sound to be visually identified in context. More specifically, identification is achieved by observing the relative positions of the formants and by making conclusions about the type of energy source (noise or glottal pulses) by which the spectrum is excited. The standard spectrogram is a poor guide for determining the relative amplitudes of the formants, so that this information, if desired, must be obtained by other means.

The ability of human subjects to recognize speech sounds from their visual representation in spectrograms is no guarantee that the same formant and excitation information is used by the ear for recognition purposes. It merely establishes that sufficient information is contained in the formant structure to permit differentiation of speech sounds to be achieved by eye.

The adequacy of the formant representation in synthesizing speech for recognition by the ear can only be established by showing that synthetic speech having a formant structure similar to that of the original speech is indeed recognizable by human subjects. This has been established beyond doubt in several laboratories by a variety of methods including the use of vocal tract analogs, channel vocoders, and formant vocoders.

In formant vocoders the frequencies and amplitudes of the first three formants plus information about the energy sources of the spectrum are extracted from the input speech in the analyser and are transmitted to the synthesizer. Apart from the difficulties and uncertainties involved in extracting this information from speech signals in real time, the bandwidth required for transmission of such information permits only a modest bandwidth compression when this system is compared to conventional audio channels which transmit speech of comparable quality and intelligibility. But the soundness of the formant approach to speech synthesis is not thereby invalidated. It can only be hoped that much of the information transmitted in the above process is either redundant or otherwise useless.

This is, in fact, the case. A study of data from spectrograms and vocal tract analogs reveals that there is a strong functional interdependence not only between the first three formants, but also between their relative amplitudes[1,2].

Fant has shown that once the first three formant frequencies are specified, the relative amplitudes of these formants are uniquely determined[3]. This functional interdependence is due to physical constraints within the vocal tract which guarantee that certain combinations of formant frequencies and formant amplitudes can never be produced by the human vocal apparatus. Much of the information about frequencies and amplitudes of these formants is then clearly redundant.

An experiment performed by the author has shown that distorted natural speech in which all formant information except the second formant frequency has been suppressed yields articulation scores as high as 92% for monosyllabic English words out of context (4). The highest score in a similar experiment in which all formant information except the first formant frequency was suppressed was 16% (5). In both cases, the amplitude of the formant selected was maintained at a constant level by infinite amplitude clipping. It can be concluded that, for most speech sounds, information about the first and third formants is highly redundant. Consequently, a much greater degree of bandwidth compression can be achieved by sending to the synthesizer information concerning only the second formant frequency and excitation information.

The precise significance of excitation information has yet to be determined. Certainly much of the emotional content of speech is expressed by inflexion and stress and these are merely the linguistic correlates of the pitch and amplitude of the excitation signal. The presence or absence of voicing can be used to distinguish between the so-called voiced and unvoiced consonants but this is obviously a redundant feature since these sounds can be reliably distinguished in whispered speech. From all available evidence it seems unlikely that any of the perceptual cues which enable speech sounds to be distinguished one from the other is carried solely by the excitation information. But is it possible to generate intelligible synthetic speech without making use of this information? The following experiment will elucidate this point and lead us to a discussion of catalytic or structural features of speech which serve merely to identify it as such.

THE GENERATION OF WHISTLED SPEECH

It has been shown above that distorted natural speech in which all formant information except the second formant frequency has been suppressed is still highly intelligible. Spectrograms of this distorted speech show that the behavior of the second formant frequency is undisturbed by the distortion process. Higher frequency bands appearing in the spectrograms are directly attributable to harmonics of the second formant frequency introduced by the clipping process.

Since the second formant frequency contains so much information necessary for distinguishing between speech sounds, is it then possible to generate intelligible synthetic speech whose only similarity to the original speech is the behavior of its second formant frequency? Two experiments were performed in an attempt to answer this question.

A block diagram of the test apparatus for the first of these experiments is shown in Figure 1. Speech from the microphone enters a second formant tracker whose output is a voltage proportional to the second formant frequency of the incident speech. A detailed description of this second formant tracker is available elsewhere (6). Basically, this device computes a running average of the number of axis crossings per second of the filtered and infinitely clipped "second formant speech" mentioned earlier.

The output of the tracker is used to control the repetition frequency of a multivibrator. To achieve this, two additional transitors are included in an otherwise basic multivibrator circuit to provide approximately constant current sources for linearly charging the timing capacitors of the multivibrator. The magnitude of the charging current is controlled by the output of the tracker. Divergence from linearity of less than 3% between the control voltage and the multivibrator frequency was obtained over the entire

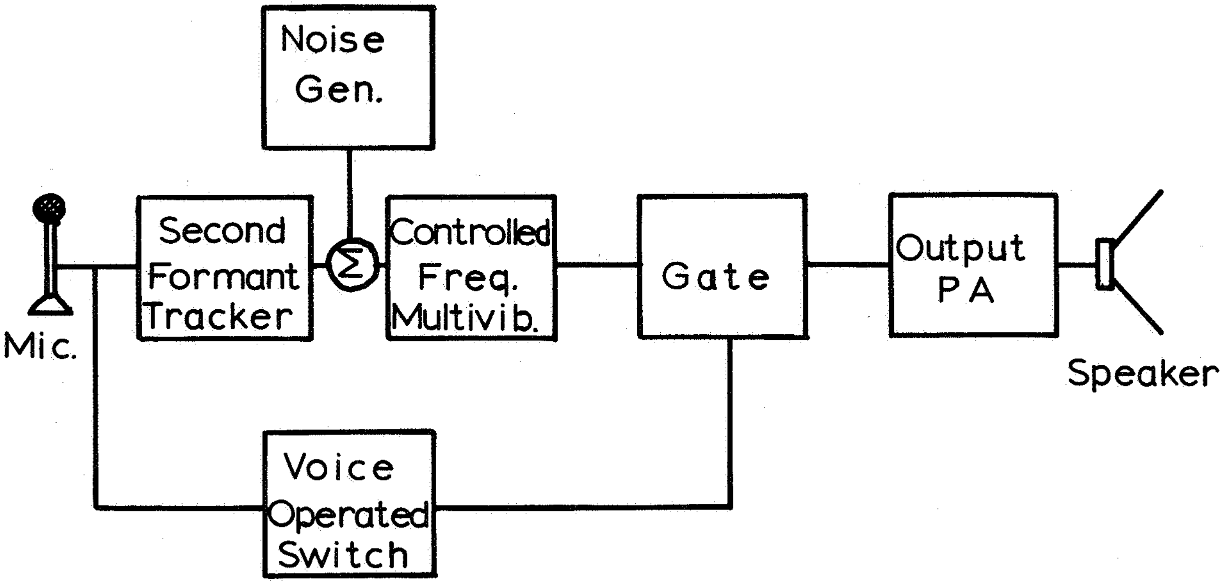

Figure 1. Experimental System for the Generation of Whistled Speech.

range of the second formant frequency.

A gate is used to assure silence during the silent intervals of the speech signal. It is opened and closed by a voice operated switch. This voice operated switch is formed of a combination of two devices which detect the presence of voicing or of noise excitation in the input speech. If either one or both is present, the voice operated switch turns on.

A noise generator is shown in the block diagram, but for the moment it is assumed that its output is set to zero.

RESULTS

The synthetic "speech" produced by this device sounds like an interrupted whistle which, to untrained ears, is completely unintelligible. Yet, superficial examination shows that its spectrogram corresponds closely to that obtained for filtered and clipped "second formant speech" whose intelligibility was as high as 92%. Subjects who heard these sounds, reported that they in no way associated them with speech. Even when told that this was a type of synthetic speech, the subjects had no success in identifying words even out of a small set such as the spoken digits. After considerable practice and with the aid of a long professional acquaintance with the behavior of the second formant frequency in words and phrases, the author was able to classify correctly a few simple words out of a restricted set.

The observation reached on consideration of these results was that, if the second formant frequency does indeed convey much of the information necessary for discriminating between speech sounds, it is presented to the ear in this case in a form which does not allow the normal processing to take place. What then are the factors present in the filtered-clipped speech which are not contained in whistled speech?

An examination of the waveform of the highly intelligible filtered and clipped speech reveals that, during periods of voicing, a basic periodicity at the glottal rate is observable in the waveform. This implies that there is some energy in the spectrum at the glottal rate. It is so small, however, as to be undetected in the spectrograms, Nevertheless, its presence is undeniable and is most easily seen by observing in the waveforms a resynchronization of the second formant oscillation at the beginning of each new glottal period.

An attempt was made to incorporate this feature into the whistled speech by resynchronizing the multivibrator by means of a pulse at typical glottal intervals (here 150 Hz) during voiced periods of the original speech. This step produced no noticeable improvement. A further step was taken to give more prominence to this basic periodicity at the glottal rate. During the voiced intervals of speech the envelope of the multivibrator output was modulated at a glottal rate such that between glottal pulses, the amplitude of the oscillations decreased exponentially. The synthetic sound produced was still totally unacceptable as speech.

A study of waveforms of the filtered-clipped speech for unvoiced or whispered sound reveals that the time intervals between axis crossings show considerable variation about the average interval corresponding to the period of the second formant frequency. On the basis of this evidence it was decided to add white noise to the control signal entering the multivibrator. Addition of noise causes a random variation in the length of each multivibrator half cycle while not altering the average duration of these cycles as determined by the second formant tracker. The action of the noise source can be regarded as a form of frequency modulation or random vibrato about the average output frequency.

This synthetic sound was regarded by listeners as quite "speech-like" and vowel glides were accurately recognized. The overall intelligibility was very low but the listeners clearly associated the sounds with speech. The typical reaction could be summed up as follows: "I know it's speech but I can't quite make it out." It was then decided to add to the formant control voltage a signal repetitive at a typical glottal rate. This produced a frequency modulation of the second formant carrier at the glottal rate. Rectangular waves of different duty ratios as well as sine waves were tested for the glottal modulating signal. In all cases the listeners reported that the sounds were speech-like and had a definite pitch. Again a few simple glides were recognized.

INTERPRETATION OF RESULTS

If this device is to be judged solely on its ability to produce intelligible speech it has certainly proved a spectacular failure. But before too harsh a judgment is pronounced, two facts should be kept firmly in mind. The first of these is that spectrograms of whistled speech and spectrograms of highly intelligible "second formant speech" are virtually indistinguishable. The second is that synthetic speech yielding intelligibility scores as high as 45% has been produced in another machine (see reference 6) which utilizes the very same time variant information used in the production of whistled speech.

Let us consider the similarity observed between the spectrograms of "second formant speech" and whistled speech. It must be remembered that a normal spectrograph is a wide band or low Q device. If a pure tone is used as input, the corresponding frequency band in the spectrogram is not a fine line but a broad band centered around this frequency.

Similarly, for a square wave input (such as whistled speech) there are broad bands at the fundamental and at the various harmonic frequencies. Even when a modest perturbation is imposed on the frequency of a square wave no change is observed in the spectrogram. Yet it is clear from the above tests that it is the fine structure of the spectrum, missing in an ordinary spectrogram, which determines whether or not the ear accepts the sounds as speech.

It appears then that there are two types of information in a speech signal. One type, the dynamic information, conveys the word-based information necessary for distinguishing between different speech sounds. It can clearly be identified with the gross behavior of the formants and, therefore, with the articulatory gestures of speech. The other type, the structural information, conveys one message only: that the sound is indeed a speech sound. The first type has been exhaustively investigated; the second, though obviously present in all forms of intelligible synthetic speech, has yet to be isolated and thoroughly defined. It is hoped that further research in the near future will lead to an accurate description of the structural features. Such knowledge will be of immense value in our understanding of speech perception.

REFERENCES

1. R.K. Potter and G.E. Peterson, "The Representation of Vowels and Their Movements," J. Acoust. Soc. Am., 20, 528 (1948).

2. K.N. Stevens and A.S. House, "Studies of Formant Transitions Using a Vocal Tract Analog," J.Acoust. Soc. Am., 28, 578 (1956).

3. G. Fant et al., "Formant-Amplitude Measurements," J. Acoust. Soc. Am., 35, 1753 (1963).

4. I.B. Thomas, "The Second Formant and Speech Intelligibility," Proceedings of the National Electronics Conference, 23, 544 (1967)

5. I.B. Thomas, "The Influence of First and Second Formanta on the Intelligibility of Clipped Speech," J. Audio Engin. Soc., 16, 18 (1968)

6. I.B. Thomas, "The Significance of the Second Formant in Speech Intelligibility," Tech, Report No. 10, Biological Computer Lab., Elec. Engin. Research Laboratory, Engineering Experiment Station, University of Illinois, Urbana, Illinois, (1966).

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* University of Illinois Biological Computer Laboratory Urbana, Illinois	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

STRUCTURAL AND DYNAMIC ASPECTS OF SPEECH SIGNALS

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
scientific; ;interim

5. AUTHOR(S) *(Last name, first name, initial)*

Thomas, Ian B.

6. REPORT DATE April 1, 1968	7a. TOTAL NO. OF PAGES 13	7b. NO. OF REFS 6

8a. CONTRACT OR GRANT NO. AF-AFOSR 7-66, 7-67 b. PROJECT AND TASK NO. 9769-04 c. 61445014 d. 681304	9a. ORIGINATOR'S REPORT NUMBER(S) BCL Report No. 7.3 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES Partial sponsorship: AF 33(615)-3890, NASA NGR 14-005-111	12. SPONSORING MILITARY ACTIVITY Air Force Office of Scientific Research Directorate of Information Sciences Arlington, Virginia 22209

13. ABSTRACT

It has been shown that distorted natural speech, in which all formant information except the second formant frequency has been suppressed, yields articulation scores as high as 92% for monosyllabic English words out of context. Whistled speech, in which the period of a multivibrator is controlled by a second formant tracker, yields articulation scores near zero. If a random noise source, or a source with period equal to the glottal rate, is used to vary the zero-axis crossing times of the whistled speech, the sounds become speech-like and articulation scores increase. The spectrograms for these three approximations to speech are virtually indistinguishable. It appears that the gross formant structure (dynamic information) conveys the word-based cues necessary for distinguishing among different sounds, whereas the excitation (structural information) conveys the cue that the sound is indeed a speech sound.

Unclassified
Security Classification

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech						
Synthesis						
Cues						
Formant						
Second formant						
Excitation						
Structure						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
_______________________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
_______________________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
_______________________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Unclassified
Security Classification