A LIMITED SPEECH RECOGNITION SYSTEM II

Contract No. NAS 12-138

Final Report

Daniel G. Bobrow

Alice K. Hartley

Dennis H. Klatt

1 April 1969

Submitted to:

National Aeronautics and Space Administration
Electronics Research Center
575 Technology Square
Cambridge, Massachusetts 02139

Attention: Mr. Wayne A. Lea/RCS

I. INTRODUCTION

This report describes results of extended studies based on previous work for National Aeronautics and Space Administration on the feasibility of a voice link with a computer. For this first study, a successful limited speech recognition system, LISPER, was built and tested. LISPER (built in BBN-LISP) has been used as a prototype of a trainable speech pattern recognition system and as a research tool for exploring the acoustic characteristics of speech as sampled by the computer. In this introduction we shall summarize the previous work and indicate the objectives of the current study.

LISPER operates within limitations along a number of dimensions. It is designed to recognize messages with easily demarked beginning and termination points, rather than recognizing a segment of a continuous speech stream. This set of messages is limited in number; at any one time the vocabulary to be distinguished can contain up to approximately one hundred items. However, an item may be any short phrase which has delimited endpoints, and need not be just a single word. The messages illustrated in Table 1, taken from a NASA context, illustrates the complexity of phrases that can actually be utilized and distinguished by the system. That 44 item list was utilized in the original study, achieving a recognition rate approaching 100%. A more extensive list of 109 items shown in Table 2 was utilized in the current study.

The system was originally designed to work well in distinguishing the utterances of a single speaker. Its recognition capability was achieved after a set of training rounds with that speaker. One of the results of the current study provides some data on the possibility of utilizing this system for utterances of new speakers after training by a number of previous speakers. These results were not

one	distance to dock
two	fuel tank content
three	time to sunrise
four	time to sunset
five	orbit apogee
six	orbit perigee
eight	revolution time
nine	closing rate to dock
point	midcourse correction time
plus	micrometeoroid density
stop	radiation count
zero	what is attitude
seven	remaining control pulses
minus	alternate splashdown point
pressure	weather at splashdown point
negative	sea and wind at splashdown point
what is yaw	visibility at splashdown point
what is pitch	temperature at splashdown point
end repeat	skin temperature
affirmative	power consumption
inclination	fuel cell capacity
distance to earth	repeat at intervals

TABLE 1. Message List From NASA Context

about
accumulate
add
address
assemble
at
binary
bite
block
breakpoint
change
choose
chop
close
comma
compare
compile
complement
control
core
cycle
decimal
delete
describe
directive
display
divide
do
down
draw
dump
edit
end
equal
exchange
five
four
from
generate
get
go
greater
half
if
index
initiate
input
insert
intersect
jump
left
less
list
load
location
look
make
memory
minus
move
multiply
name
night
nine
number
octal
of
one
output
overflow
parenthesis
plus
point
print
push
quarter
read
register
replace
restore
revolve
right
save
scale
set
seven
shift
show
shrink
six
skip
space
specify
square
store
subtract
swap
tab
than
think
three
toggle
two
unite
up
whole
word
yield
zero

TABLE 2. Expanded Word List

very encouraging, and we feel that this indicates that the system really learns the idiosyncratic variations of an individual speaker's set of input messages, and the properties developed are relatively speaker dependent. Another goal of this study was to investigate the utility of the speaker dependence of properties for limited speaker recognition. The results are these tests were only mildly encouraging. It appears that this speaker dependence is not really consistent enough to provide reliable classification of the speaker given that you know the message.

The basic structure of the LISPER system is indicated in Figure 1. It consists of three principal components, a digital spectral analyzer to provide the raw data from which the recognition system utilizes a set of programs for extracting properties of the input message from this digitized input spectrum; and training and recognition algorithms, for storing information about the speech signal and making a decision about the identity of an unknown utterance based on earlier experience. The spectrum analyzer has been described in detail in previous reports. It basically consists of 19 bandpass filters whose outputs are rectified, low-pass filtered, sampled by an analog-to-digital converter, and then converted into logarithmic units. The 19 filters consist of 15 filters which are spaced uniformly at 360 Hertz up to about 3,000 Hertz, which is the range encompassed by the first 3 resonances of the male voice; and 4 filters which cover the range from 3,000 to 6,500 Hertz in which there is information about the noise component of speech. The 360 Hertz band width of the low set of filters is sufficiently wide so that one does not see spectral peaks due to individual harmonics of fundamental frequency of the male speaker; a spectral maximum in the filter outputs indicate that the presence of one or more formants (resonances of the vocal tract transfer function). The

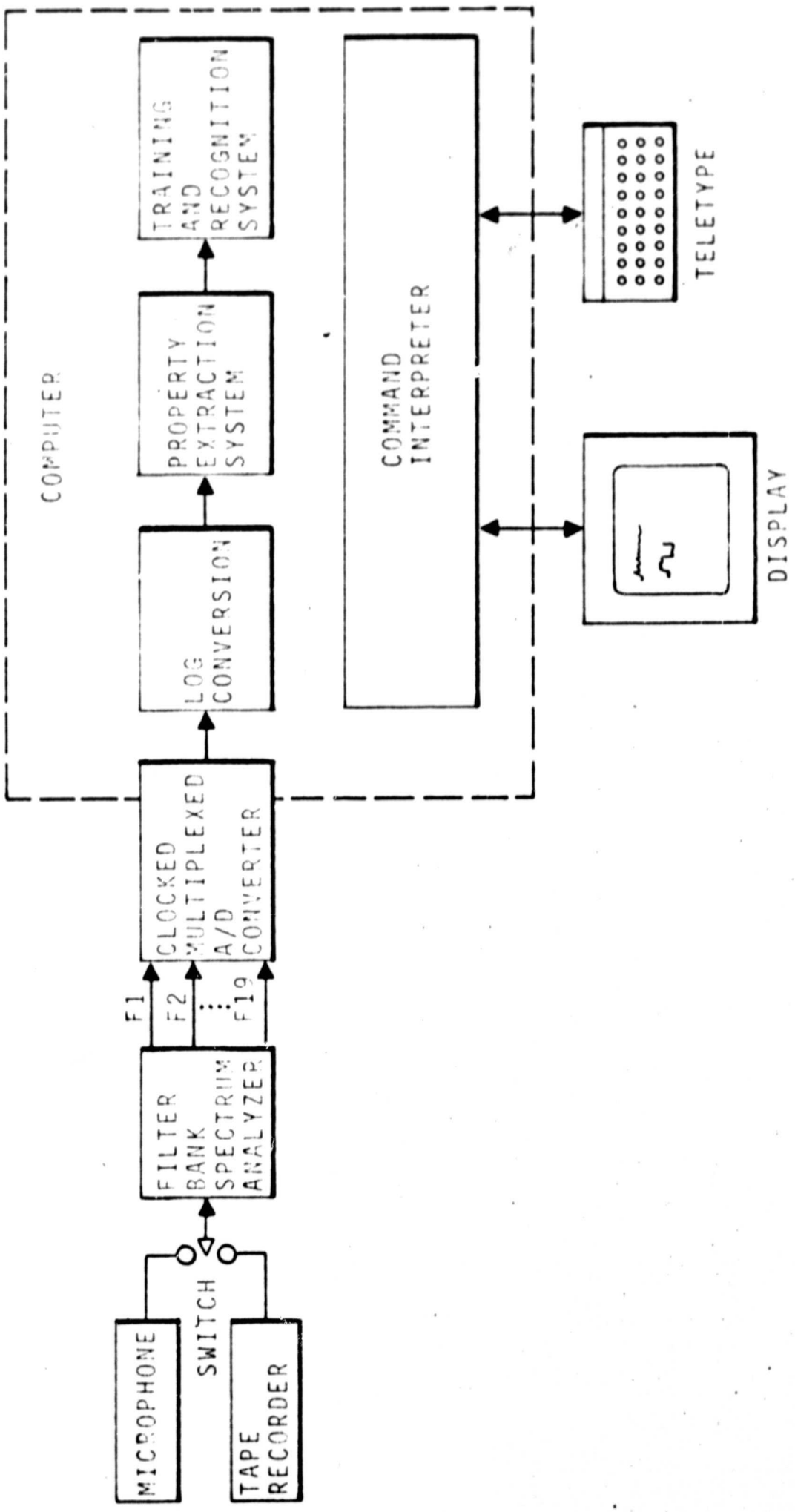

FIG.1 BLOCK DIAGRAM OF THE LISPER SYSTEM

spacing between filters makes resolution of individual formants poor and we do not use formant tracking in our recognition system. The analog to digital sampling rate is 100 Hertz. Each 10 milliseconds the outputs of all 19 band pass filters are sampled simultaneously, giving an approximation to the logarithm of the short time spectrum of the input message.

In the report of our earlier work (Bobrow and Klatt, 1968) we described a number of algorithms we used for extracting features which characterize speech utterances. They are designed to be used with the particular recognition system we have developed which provides high quality message identification in the presence of redundant, inconsistent or incorrect information from these properties. In that report, we present two different sets of properties, one of which is based on the present day knowledge of acoustic phoenetics and distinctive features, the other of which is a more direct reflection of the shape of the input spectral pattern. In this report we describe expanded sets of such linguistic and spectral (abstract) features, and give some measures of their relative effectiveness and information content. In order to provide a broader base for evaluation of the properties we created the new word list of 109 words shown in Table 2 augmenting one of the original experimental lists (the 54 word list originally taken from Ben Gold) with enough additional words so that the expanded word list is approximately phonetically balanced. It also contains many of the common prestressed consonant clusters of English. In addition, we recorded 10 speakers saying this new word list to provide a broader range of speaker characteristics for this evaluation.

The goals of this research can be summarized:

1. To develop and evaluate an improved and expanded set of properties of both linguistic and nonlinguistic variety. These properties are described in detail in the body of the report, and recognition results given comparable to our earlier scores on easier lists.

2. To investigate the relationship between linguistic and nonlinguistic features and evaluate their relative effectiveness. To this end we have compared recognition results using these properties on the expanded word list for a large number of speakers (8 versus 3 for the previous report). We have also computed information and spread measures for these properties.

3. To investigate the utility of speaker dependence of these properties for limited speaker recognition. This would hopefully permit adjustment of the word recognition program to individual speakers. However the results of this study indicate that the consistency of speaker dependence is not sufficient for this purpose. The problem of speaker independent recognition was also explored.

4. To explore some new techniques which utilize the information currently available from the spectral analyzer. These include reanalysis of spectral data for distinguishing between similar words, and application of the ideas used in our property extraction for segmentation of continuous speech.

II. Data Collection

The list of 109 words selected for a final evaluation of the recognition system has been shown. In creating the new word list, the original 54 vocabulary (Gold, 1965) was augmented in such a way that the expanded word list is approximately phonetically balanced and contains many of the common presetressed consonant clusters of English. As an example, there was no occurrence of the sound "sh" in the original list. No attempt was made to exclude word pairs which might be difficult to distinguish such as "add-at", "four-core", etc.

Six recordings of the word list were obtained from each of 10 speakers. A speaker participated in two recording sessions of three rounds on the word list, separated by a one week interval, so that it would be possible to compare short-term and long-term consistency of pronunciation. Speakers sat in a sound-proof room and read a word as it was presented on a flash card (through a window) from the adjoining recording room. The flash cards were shuffled before each reading and presented at a rate of about one word every 5 seconds. Shuffling was done to eliminate any serial effects on word pronunciation.

Recordings were made on 1 1/2 mil Mylar at 7 1/2 ips using an Altec model 685A microphone and an Ampex model AG350 tape recorder. The tape recordings were later played back through the filter bank, sampled, and stored on a large computer disk. The digitized spectra stored on this disk constitute the body of raw data that was used in the final evaluation of the recognition system.

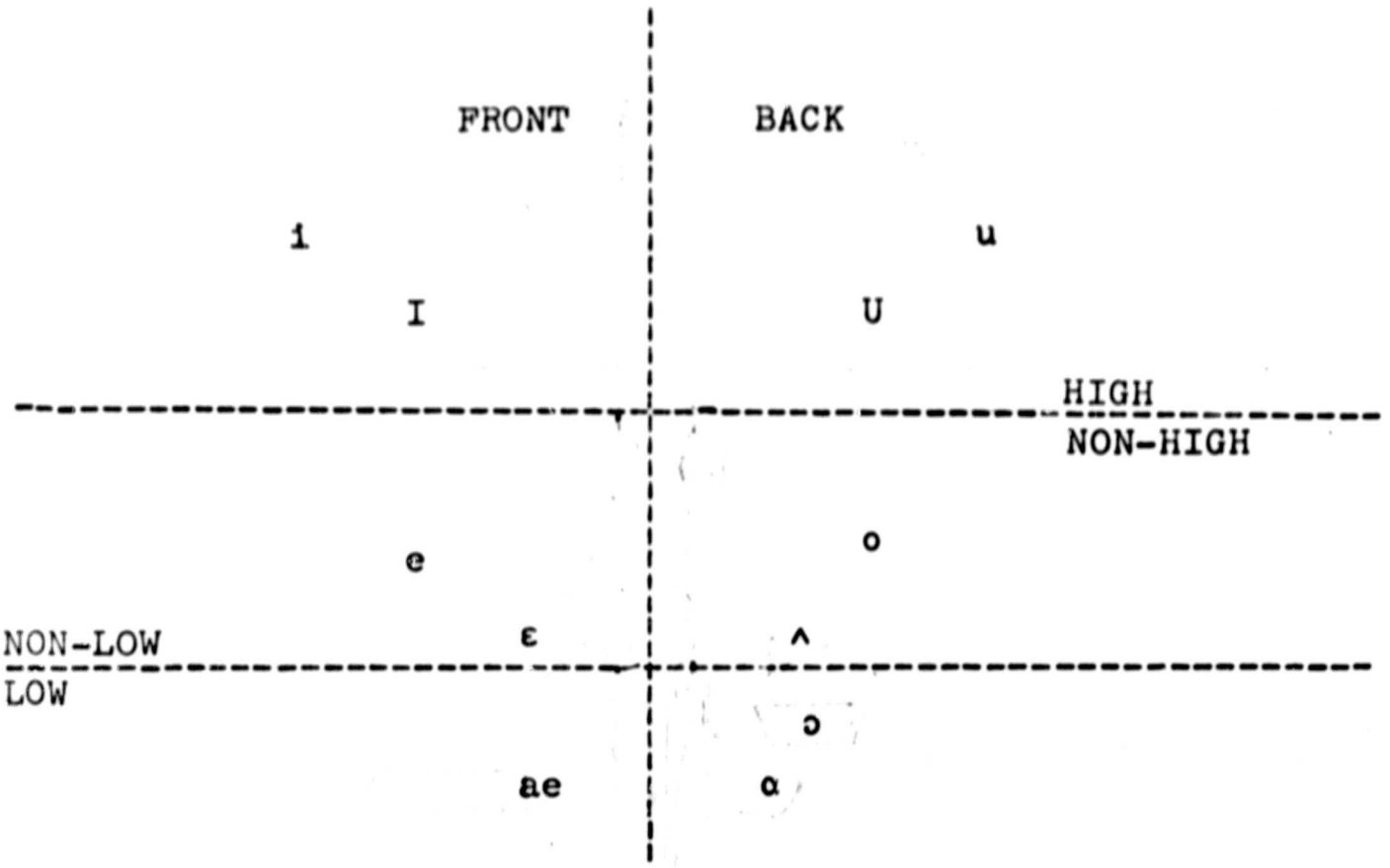

FIGURE 2. Classification of vowels in terms of the phonetic features front, high, and low. These features characterize the position of the body of the tongue and have well-defined acoustic correlates.

p q missing

word can be recognized. In an attempt to improve system performance on the first few learning trials and to provide a desirable form of redundancy, the function FHIGH provides the basis for two features that are identical in form, but have different sets of overlapping thresholds, shown in table 3. The thresholds have been set such that, if a word is treated inconsistently by HIGH, it is likely that it will be treated more consistently by HIGH'.

FEATURE	K1	K2	K3	K4
HIGH	21	15	9	3
HIGH'	18	12	6	0
LOW	60	48	36	24
LOW'	54	42	30	18
FRONT	42	30	18	6
FRONT'	36	24	12	0
BACK	76	64	52	40
BACK'	70	58	46	34

Table 3. Threshold settings for the eight new linguistic features

2. LOW(I) = 2 if [FLOW(I) > K1] or [FLOW(N) > K2 and LOW(I-1) = 2]
1 if [FLOW(I) > K3] or [FLOW(I) > K4 and LOW(I-1) ≠ 0]
0 otherwise

FLOW(I)= F4(I)+F5(I)+F6(I)+F7(I)-F1(I)-F9(I)-F10(I)-F11(I)

This property is designed to indicate the presence of a vowel produced with the tongue body low in the mouth. Low vowels such as /ɔ ɑ æ / have a high first formant frequency; high vowels and consonants do not. If significantly more energy is present in filters 4 through 7 than in filters 1,9,10, and 11, a voiced sound segment with a high first formant is indicated. Thresholds for two features based on the above definition are listed in table

3. FRONT(I) = 2 if [FFRONT(I) > K1] or [FFRONT(I) > K2 and FRONT(I-1)=2]
1 if [FFRONT(I) > K3] or [FFRONT(I) > K4 and FRONT(I-1)≠0]
0 otherwise

FFRONT(I)=F11(I)+F12(I)+F13(I)+F14(I)-F6(I)-F7(I)-F8(I)-F9(I)

This property is designed to indicate the presence of a front vowel or alveolar consonant. Vowels produced with the tongue body forward in the mouth such as /i I e E/ have a high second formant frequency. The locus of the second formant transition of an alveolar consonant is also high except in the environment of a back vowel where alveolar consonants may not be detected. If significantly more energy is present in the high frequency range (filters 11 through 14) than in the mid-frequency range (filters 6 through 9), a sound segment with a high second formant is indicated.

Thresholds for two features based on the above definition are listed in table 3 .

4. BACK(I) = 2 if [FBACK(I) > K1] or [FBACK(I) > K2 and BACK(I-1) = 2]
1 if [FBACK(I) > K3] or [FBACK(I) > K4 and BACK(I-1) ≠ 0]
0 otherwise

FBACK(I) = F1(I)+F2(I)+F5(I)+F6(I)+F7(I)+F8(I)+F9(I)-F3(I)-F4(I)
-F11(I)-F12(I)-F13(I)-F14(I)-F15(I)

This property is designed to indicate the presence of a back vowel or labial consonant. Vowels produced with the tongue body back in the mouth such as /u ʊ o/ have a low second formant frequency. The locus of the second formant transition of a labial consonant is also low in frequency. In the environment of a front vowel, labial consonants may not be detected. If significantly more energy is present in the mid-frequency range (filters 5-8) than in the high-frequency range (filters 11 through 15), a sound segment with a low second formant is indicated. Thresholds for two features based on two definitions above are listed in table 3.

The features HIGH, LOW, FRONT, and BACK are related to the old features AHLIKE, EELIKE, and OOLIKE but the new features differ in several important ways. The new features do not use the feature VOICE in the definition which means that a partial segmentation of the new feature sequences into those states accompanied by voicing and those not accompanied by voicing is not possible. The advantages of this omission are (1) the new features are likely to perform satisfactorily even when the feature VOICE is in error and (2) the timing problems that were created when VOICE changed state a little before or after a function exceeded a threshold do not exist.

The new features are useful in describing the spectral shapes of consonantal segments as well as vowels, whereas AHLIKE, EELIKE, and OOLIKE were carefully restricted to apply only to vowel portions of words. The inclusion of consonants in this way is important because alternative methods* for characterizing consonants are not easy to implement within this feature framework.

5. DHIGH(I) = 0 if VOICE(I) = 0 or VOICE(I-1) = 0 or ...
VOICE(I-5) = 0
1 if [FHIGH(I) > 16] or [FDHIGH(I) > -16 and
DHIGH (I-1) = 1]
-1 if [FDHIGH(I) < -16] or DHIGH(I-1) = -1
2 otherwise

FDHIGH(I) = FHIGH(I) + FHIGH(I-1) + FHIGH(I-2) -
FHIGH(I-3) - FHIGH(I-4) - FHIGH(I-5)

This property is designed to indicate the presence of a rapid change in the location of the first formant. Transitions from consonants to non-high vowels and from non-high vowels to consonants will be detected. The state 0 corresponds to a voiceless sound segment, the state 1 to a consonant-vowel transition, the state -1 to a vowel-consonant transition, and the state 2 to a voiced segment not containing a rapid first formant transition.

---

* The reason for this difficult is that a feature which would segment syllables into consonants and vowels is impossible to define. No general algorithm exists, and the characteristics of consonant-consonant and consonant-vowel transitions are dependent on manner of articulation of the consonant and of features of the vowel.

New Spectral Features

The features added to the set of abstract features include eight features for detecting sudden changes in energy in selected frequency regions These features are defined and motivate as follows:

$$DTn(I) = \begin{cases} 2 & \text{if } [FDTn(I) > K] \text{ or } [FDTn(I) > -2 \text{ and } DTn(I-1) = 2] \\ 0 & \text{if } [FDTn(I) < -6] \text{ or } [FDTn(I) < 2 \text{ and } DTn(I-1) = 0] \\ 1 & \text{otherwise} \end{cases}$$

$$FDTn(I) = Fn(I) - Fn(I-2)$$

These properties are designed to indicate the presence of a sudden change in the outputs of selected filters. The filter numbers, n, and the thresholds, K, are listed in table 4. Previously defined abstract featrues have quantized the output of a filter into 3 levels and have quantized the difference in outputs of adjacent filters into 3 levels. The latter is similar to a derivative of the spectrum with respect to frequency with time held fixed. The eight new features, DTn, are similar to a derivative of the spectrum with respect to time with frequency held constant. State 2 corresponds to a sudden increase in the output from filter n. State 0 corresponds to a sudden decrease, and state 1 corresponds to no significant change. Linguistically, sudden changes in acoustic output signals the onset of voicing or the release of a consonantal closure or constriction. Place of articulation may be determined by observing the filters in which the sudden change appears.

N	K
2	16
5	14
8	12
10	10
13	12
16	14
19	16

TABLE 4. Thresholds for the temporal-derivative abstract features.

IV. Results

Single Speakers

The new sets of properties, both linguistic and spectral, were run on eight speakers. On four of these speakers all six rounds of data were used, providing five recognition scores, and on the other four speakers only three rounds were used. The results are summarized in Figures 3 and 4. The scoring algorithm was simplified from the one described in our earlier report. Each word occuring at a node is given a single vote. The frequency of a word at a node is not used to break ties (and hence need not be stored). This simplification was made to save space and time, with the idea that we would not lose much in the way of decision capability. The hope that very few ties would occur proved true on speaker KNS who provided the data for the design of the original properties, and for early tests of the current more efficient system. Unfortunately, with later speakers ties became much more frequent, (about 50% of number wrong) but to maintain data compatibility and the speed and space saving we did not revert to our original scoring algorithm. Ties were assumed broken by a simple coin flip, and the number unambiguously correct was augmented by one half the number of ties to give the scores shown. As can be seen, the average asymptotic recognition rate is about the same (91% and 94%) on these properties for the extended list as our previous results (96%) for the smaller list.

Cross Speaker Recognition

Trees containing training data from a number of speakers were used as a basis for recognition of new speakers. For both the linguistic and spectral properties two sets of trees were prepared. One set was constructed from six rounds of training from three speakers. The second was constructed from three rounds of training from five speakers. Limitation on the size of the training trees precluded any larger trees. The latter trees were pruned of all branches containing only a single utterance (any one message from any speaker), but this did not effect the recognition rate at all.

TRIAL

Speaker	2	3	4	5	6
KNS	78	92	90	92	98
DHK	80	81	79	92	90
CD	66	85	77	82	91
DGB	63	76	75	83	86
RST	71	86			
CM	75	83			
DD	75	82			
GJ	65	79			
AVG(4 SPKRS)	72%	83%	81%	87%	91%
AVG(8 SPKRS)	72%	83%			

FIGURE 3. Recognition Rate with Linguistic Properties. Each Speaker is Treated Individually.

TRIAL

Speaker	2	3	4	5	6
KNS	83	94	96	93	96
DHK	74	87	81	98	98
CD	77	90	90	96	94
DGB	52	76	81	82	87
RST	76	92			
CM	75	91			
DD	75	90			
GT	71	82			
AVERAGES(4)	73	87	87	92	94
AVERAGES(8)	74	88			

FIGURE 4. Recognition Rate for Spectral Properties

The results of these experiments are summarized in Figure 5. One obvious conclusion is that training by more speakers with fewer rounds is better than the converse. Secondly, the results indicate that except for the first round (obviously), the training on other speakers does not particularly help in recognition. That is, the second round recognition rates are comparable to the results obtained by individual speakers on their own trees. This indicates the idiosyncratic nature of the properties extracted for recognition by this system.

Information Measure of Properties

In the previous final report we defined a measure of information that was used to evaluate the relative usefulness of individual features. For a 109 word vocabulary, a perfect feature would provide $\log_2(109) = 6.7$ bits of information about the word to be recognized, assuming equal word frequencies. A feature that produced the same sequence of states for any word would provide 0 bits of information.

Figures 6 and 7 summarize the information content of the linguistic and abstract (spectral) properties respectively. The information measure is computed for trees containing one speaker, 3 speakers, and 5 speakers. It can be seen that the additional speakers tend to reduce the amount of information about individual word identities in the tree. If speaker 2 is recognized using the combined trees instead of only his own trees, the information measure predicts that his recognition scores will go down.

Figures 6 and 7 also include a measure of the spread, S, in a tree. Spread is given by the formula:

$$S = \frac{N}{M}$$

where N = number of nodes in the tree

and $M = \frac{\text{number of different messages}}{\text{number in the message set} = 109}$

		Round	1	2	3
Linguistic Properties	3 Speakers	DGB	66	79	74
		RST	82	83	90
	5 Speakers	DK	74	80	
		CD	83	82	
Spectral Properties	3 Speakers	DGB	65	78	
		RST	75	86	
	5 Speakers	DK	72	82	
		CD	76	85	

FIGURE 5. Recognition rates (Percent Correct) on Pretrained Trees. On the first round, the trees contain no information about the speaker to be recognized.

FIGURE 6. Information and Spread Measures - Linguistic Properties

I = Information S = Spread

	S2		S2,4,5		S2,7,8,9,11	
	I	S	I	S	I	S
TVOICE	1.12	3.72	.96	5.90	0.87	2.90
TSYL	1.34	3.51	1.15	3.07	1.09	2.56
STRESS	1.68	5.45	1.37	4.48	1.31	4.37
ALIKE	3.50	26.35	3.09	24.08	2.75	16.79
ILIKE	3.37	22.87	2.96	23.86	2.78	17.44
RLIKE	2.28	16.53	2.00	18.16	1.68	15.25
ULIKE	3.25	24.55	2.84	21.12	2.56	17.05
FRICATIVE	3.25	22.03	2.65	20.15	2.50	18.57
STOPBURST	1.81	14.49	1.59	15.00	1.75	15.10
NASAL	2.50	14.28	2.00	14.13	1.75	11.13
TSTRIDE	2.96	22.68	2.37	20.93	2.50	21.06
STRIDESTOP	1.50	7.10	1.06	6.61	1.21	6.63
ASTRESS	1.90	3.52	1.25	1.90	1.15	1.72
ISTRESS	1.90	3.29	1.25	2.01	1.25	1.98
USTRESS	1.78	3.10	1.18	1.86	1.06	1.78
DHIGH	2.93	21.90	2.56	23.04	2.37	19.74
HIGH1	3.65	29.90	3.25	27.87	3.15	23.97
HIGH2	3.87	33.50	3.37	31.36	3.25	25.38
LOW1	3.15	30.68	2.84	27.25	2.59	25.81
LOW2	3.28	30.20	2.96	27.91	2.71	26.24
FRONT1	2.84	21.33	1.96	17.51	2.25	20.60
FRONT2	3.03	23.88	2.15	17.97	2.37	20.65
BACK1	3.56	25.78	2.87	19.76	3.03	21.58
BACK2	3.84	31.68	3.09	22.27	3.21	23.75

FIGURE 7. Information and Spread Measures for Spectral Properties

	S2		S2,4,5		S2,7,8,9,11	
	I	S	I	S	I	S
S1	0.56	3.47	0.53	4.25	0.53	3.51
S2	1.78	8.05	1.46	6.78	1.34	6.25
S3	2.03	14.27	1.93	15.74	1.68	12.77
S4	1.96	11.63	1.81	14.71	1.59	10.77
S5	1.96	12.74	1.75	12.57	1.53	10.42
S6	1.71	5.41	1.43	6.75	1.31	5.51
S7	1.78	7.74	1.50	7.44	1.21	5.54
S8	2.06	11.65	1.75	11.84	1.43	8.00
S9	2.15	12.72	1.84	13.37	1.56	8.86
S10	1.34	5.76	1.25	6.54	0.87	4.85
S11	1.53	6.34	1.21	7.31	0.90	5.39
S12	1.93	11.34	1.43	10.53	1.25	8.21
S13	2.25	16.21	1.65	12.82	1.56	11.92
S14	1.78	9.82	1.28	8.93	1.15	7.55
S15	1.93	13.21	1.34	9.43	1.25	8.42
S16	2.50	18.04	1.62	13.94	1.31	10.29
S17	2.31	14.66	1.28	9.47	1.28	9.04
S18	1.84	9.37	1.28	8.49	1.31	8.96
S19	1.81	8.97	1.37	9.49	1.12	8.05
D2	3.34	18.86	2.87	17.10	2.31	13.73
D3	1.71	13.95	1.09	10.08	1.06	10.58
D4	1.50	11.25	1.12	11.76	1.15	10.75
D5	3.21	19.65	2.46	15.02	2.06	11.01
D6	3.15	22.24	2.59	18.85	2.21	14.36
D7	3.15	19.04	2.68	16.53	2.43	13.47
D8	2.93	16.96	2.34	12.64	2.28	12.36
D9	2.46	16.28	1.81	12.62	1.78	12.42
D10	2.65	15.81	1.93	12.40	1.93	13.03
B2	3.28	25.08	2.65	21.93	2.40	22.21
B5	3.00	22.17	2.53	19.88	2.31	18.46
B8	3.21	26.11	2.59	22.24	2.34	19.27
B10	2.81	28.36	2.34	20.80	2.06	18.45
B19	2.21	13.01	1.62	1.43	1.43	11.64

FIGURE 8. Summary of Information and Spread Measures

	1 Speaker	3 Speakers	5 Speakers
Linguistic Properties			
Average Information	2.68	2.20	2.13
Average Spread	18.4	16.6	15.1
Total Information	64.32	52.80	51.12
Spectral Properties			
Average Information	2.34	1.77	1.58
Average Spread	14.31	12.6	10.8
Total Information	77.22	56.64	52.14

FIGURE 9. Summary of Error Rates. Number of Words Having N Errors for Eleven Test Rounds Taken from Three Speakers

	Number of Errors					
	0	1	2	3	4	5
Linguistic (5 rounds)	19	42	24	20	4	0
Spectral (6 rounds)	27	30	28	18	5	1

Speaker Recognition

The trees generated by 5 of our speakers contain information about word identity and about speaker identity. It would be necessary to reprogram the decision procedure to select the speaker with the most sequence matches for a given word. Unfortunately, the reprogramming effort that would be required in such an undertaking is very great due to the structure of the current program and data.

Instead, we have constructed a theoretical model from which it is possible to predict the applicability of our property detector approach to speaker recognition. The model predicts that, for 5 speakers, our 33 spectral properties, any one of the speakers can speak one word (of the 109 word vocabulary) in to the microphone and have 70 to 90% probability of being correctly identified. Of course, if the set of speakers happens to contain 2 speakers who are very similar, the score will be reduced. On the other hand, requiring the speaker to say more than one word will increase the reliability of the identification.

The model is based on the following decision algorithm. If any speaker n (i = 1,...,5) matches the sequence generated by property i (i = 1,...,32), he receives one vote from that property. If more than one speaker matches the input sequence for property i, the vote is cast for one of the speakers determined randomly with the probability of a speaker getting the vote proportional to the frequency of that speaker at that node with that word. The votes from all properties are summed and the speaker receiving the most votes is selected.

The model predicts an expected recognition score on the basis of the data presented in Figure 10 which indicates the average probability, in percent, that a feature will vote correctly. This average has

P 25 missing

been computed over all 5 speakers and all 3 properties for each word from the spectral property trees described previously.

If we take, for example, a word like "memory", the expected probability that a feature will vote correctly is 32%. If 33 features vote, the expected score will be 33 x .32 = 10.56 votes. The standard deviation in this expected score, $\sigma_{correct} = \sqrt{npq} = \sqrt{33(.32)(.68)} = 2.77$ votes. Assuming that the 4 incorrect speakers are equally dissimilar to the correct speaker, the expected score for an incorrect speaker is 5.5 votes and $\sigma_{incorrect} = 1.75$ votes.

The probability that a specific incorrect speaker will get more votes than the correct speaker is approximately given by integrating the appropriate tails of the convolution of two normal distributions. In this example, the probability is about 8%. The probability of an error is approximately 4 times this value, giving a probability of correct speaker identification using the word "memory" of about 68%. For the slightly better word "breakpoint" the predicted success is aoubt 88%. Pooling the votes from n words of equal quality decreases the probability of error by $\sqrt{n}$ since we effectively multiply the number properties being used by n. Four words like "memory" would thus give a probability correct of about 84%, if all the independence assumptions were reasonably accurate (which we know they are not). This redundancy between properties would increase the error rate, as would the fact that on the average all speakers are not equally dissimilar.

p 27 missing

Segmentation of Continuous Speech

Two features have been used to segment words into smaller units. These are the feature voicing and the feature syllable. The features work very well in the context of this research since a mistake in feature assignment need not be fatal. Therefore, the fact that these features do not always agree with linguistic intuition is easily handled. However, in an application involving continuous speech recognition, a more sophisticated approach will be required. Segmentation errors must be reduced or corrected at later stage of analysis because there is much less redundant information available.

Figure 11 illustrates some of the problems with our current segmentation feature Syllable. As indicated by the right hand side marks, Syllable will create three distinct regions for the five phoneme string (n,I,š,iy,e^{I}). It misses the transition from the nasal n to the vowel I and the transition in the diphthong. Note that the initial i in "initiate" is not present.

There are several ways in which our segmentation feature could be improved or replaced. One way would be to translate the segmentation techniques described by Reddy (1967, 1968) from the time domain to the frequency domain. These methods involve the measurement of spectral change and the construction of a complex decision procedure for determining when a spectral change indicates a new phonetic segment.

In general, however, one would not want to try to do a complete accurate segmentation on an isolated first pass. The recognition process and sementation should interact closely so that phonetic features of individual sound segments can be defined. Thus, segmentation hypotheses must be proposed and evaluated at many stages of the analysis and techniques such as those used by Reddy, form an excellent starting point for future research.

SPEAKER 2
REPETITION 1
WORD INITIATE

FIGURE 11. Segmentation of Speech by SYLLABLE in Part of the Word "initiate"

	FILTER	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Syllable
Phonemes																					
	0	5	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	
	1	16	6	2	1	0	0	0	0	2	0	0	1	1	0	0	0	1	0	0	
	2	22	18	7	4	2	1	0	0	3	1	1	2	2	1	0	0	1	1	0	
	3	27	27	23	13	6	3	0	0	8	9	9	13	16	10	6	9	6	1	0	0
	4	36	37	34	21	11	6	0	4	15	18	14	19	24	22	12	14	13	2	0	
	5	36	39	35	24	17	9	2	7	19	21	18	20	24	21	14	17	13	3	2	1
	6	37	40	35	25	18	11	2	9	22	23	19	22	27	25	18	17	14	2	2	
	7	38	37	32	23	19	14	4	7	18	19	17	20	24	22	14	17	12	2	1	
	8	43	42	32	23	20	15	5	7	16	17	17	21	25	22	13	15	11	1	1	
	9	44	42	31	21	20	16	5	6	12	12	14	20	23	19	10	13	11	1	0	
	10	42	42	31	22	20	15	5	5	10	10	15	20	23	19	10	13	10	1	0	
n	11	37	40	30	22	20	16	5	5	9	9	15	21	25	21	11	14	10	1	0	
	12	43	42	31	25	22	17	6	5	8	8	17	23	27	23	12	14	10	1	0	
	13	47	45	34	27	24	18	8	13	18	13	17	24	29	27	19	18	14	4	2	
	14	43	46	40	30	23	17	9	22	29	28	23	31	42	42	32	27	18	4	4	
	15	44	48	45	35	25	19	13	24	34	33	27	34	45	46	37	29	20	5	8	
	16	47	51	48	38	27	21	17	27	37	38	31	35	46	46	37	35	24	6	13	
	17	45	51	49	39	28	22	18	27	37	39	34	39	49	50	42	45	34	10	20	
I	18	44	50	47	37	25	20	17	26	37	39	34	37	46	47	39	42	34	8	21	
	19	40	47	44	35	25	18	13	21	32	35	30	34	43	44	36	40	32	7	19	
	20	35	42	40	31	20	13	10	18	29	32	27	30	40	41	34	36	29	9	17	
	21	30	38	36	27	18	10	7	14	25	28	23	26	36	37	30	34	29	17	20	
	22	26	34	32	23	13	7	5	10	22	25	21	25	34	35	29	33	38	23	30	
	23	26	30	28	19	9	6	4	7	18	22	19	25	33	34	30	36	40	31	34	1
	24	17	25	24	16	7	4	3	5	15	18	16	22	29	31	29	35	39	30	36	
	25	8	21	19	11	5	4	3	4	12	16	17	24	31	32	29	34	42	34	40	0
š	26	11	19	17	8	4	3	1	2	9	13	14	23	31	31	29	38	42	34	39	
	27	4	13	12	6	3	3	2	2	8	12	15	21	27	28	30	34	42	36	38	
	28	12	11	8	4	3	2	1	1	6	9	13	21	27	28	27	34	40	34	39	
	29	8	7	6	4	2	2	1	1	6	8	12	22	28	28	27	34	39	34	38	
	30	1	4	4	3	2	2	1	1	4	7	12	20	24	26	28	31	38	33	37	
	31	6	3	3	2	2	2	0	1	5	7	11	19	25	26	26	29	36	31	34	
	32	1	2	2	1	1	1	0	0	4	5	9	17	22	23	24	27	34	29	31	
	33	1	1	1	1	1	0	0	0	3	4	7	14	20	21	22	25	34	27	29	
	34	27	20	8	4	3	1	0	0	3	3	6	12	17	18	19	24	30	25	26	0
	35	37	32	17	9	5	3	0	0	4	4	7	12	16	16	18	22	27	22	23	
	36	38	35	22	14	8	4	0	0	5	5	10	15	16	15	19	24	24	18	20	1
iy	37	39	37	25	17	8	5	0	0	5	5	12	17	18	17	22	27	22	15	17	
	38	39	39	28	17	9	5	1	1	5	9	18	21	19	18	24	27	21	11	13	
	39	42	42	31	19	11	7	1	2	7	12	20	24	24	25	31	34	21	9	10	
	40	41	42	32	21	12	7	1	2	7	15	23	26	27	29	35	38	22	7	9	
	41	38	42	32	21	13	7	1	2	7	17	25	28	30	31	35	39	21	5	8	
	42	38	40	34	23	15	8	2	2	9	20	27	31	34	35	36	40	21	5	8	
	43	42	42	35	26	17	10	3	3	11	24	29	34	37	36	35	39	21	4	7	
	44	42	44	38	28	19	12	4	4	15	27	32	36	40	38	36	39	21	5	7	
	45	39	44	39	29	19	12	5	6	20	32	35	38	44	42	37	40	21	7	7	
	46	36	42	40	31	20	12	6	9	23	34	36	40	45	42	36	40	22	8	6	
	47	34	43	42	32	19	13	7	12	26	34	36	40	44	41	35	38	21	6	6	
	48	37	44	43	33	20	13	7	13	27	35	36	40	45	42	35	38	19	5	5	
eI	49	38	44	42	32	19	12	7	14	27	34	35	38	42	38	32	35	18	4	4	
	50	36	42	42	29	17	11	6	10	24	31	32	35	39	36	29	31	14	4	3	
	51	33	41	39	27	16	9	5	8	23	30	31	34	36	34	27	29	11	3	2	
	52	33	41	39	27	15	8	4	7	20	28	29	31	34	31	25	26	9	3	2	
	53	31	39	37	25	14	8	4	5	19	28	29	32	36	34	26	25	8	3	2	
	54	33	38	35	25	14	8	3	5	18	27	30	34	38	36	29	28	9	3	3	
	55	34	36	32	23	12	6	3	4	17	27	31	34	38	37	31	31	11	2	2	
	56	32	34	29	21	10	5	2	3	12	23	26	29	33	31	25	25	8	1	1	1

Reanalysis of Spectral Data

It is plausible that a more efficient recognition system could be built if certain tests were only made if confusions arose in the recognition, or if the certainty of a preliminary analysis were in doubt. The reprogramming necessary to change the system to do this was beyond the scope of the current effort. However, we did run one experiment which tended to verify this hypothesis. We tried the recognition of the sixth round of RST using only 3,6,9, ...,33 of the spectral properties. The results were monotonic; and relatively low scores (or ties) were obtained on errors which were later correctly identified using additional properties. Thus the analysis of the temporal derivatives could have been limited to very few cases.

V. Discussion and Conclusions

In the following paragraphs we discuss reasons why speaker independent features are an illusive goal, and we suggest some linguistic origins of inconsistent pronunciation from any given speaker.

Speaker Invariance

One example of a speaker-dependent parameter that is very difficult to handle when working with filter bank outputs is the overall length of an individual's vocal tract. The length of the vocal tract determines the average spacing between formants. An energy maximum that appears in filter number n for a given speaker may appear in filter number n $\pm$ 1 for another speaker simply because of a length difference in their respective vocal tracts.

Small shifts in the locations of energy maxima can produce large changes in certain feature sequences as defined. For example, features that are based on outputs from single filters or differences in outputs from adjacent filters are sensitive to differences in vocal tract length among speakers. Thus it is not surprising that speakers have highly distinctive feature sequence trees. The differences between speakers are so great that the combined trees from many speakers not only tend to lose their ability to recognize words from a novel speaker, but also degrade in performance for a speaker previously recognized.

One can conclude that, in the process of generalizing across speakers, the recognition trees grows in such a way as to have an unacceptable amount of overlap between words. Other features or other speaker-generalizing methods must be found in order to handle input from an arbitrary unfamiliar speaker.

Sources of Inconsistency

Some speakers approach nearly perfect recognition scores (97%) on the 6th reading of the word list. Other speakers do not reach this high a recognition rate. The performance of our speakers is highly correlated with their experience in reading lists of speech materials under controlled conditions. It is evident that the better speakers are able to render a more consistent pronunciation to repetitions of the same word. However, even these speakers show a significant drop in performance on the 4th round.* The first 3 readings of the word list were recorded successively at one sitting and the last 3 readings were recorded one week later.

Consistency is of course a highly desirable trait in a speaker who is trying to be understood by a recognition algorithm. It is therefore important to identify the various types of speaker inconsistencies encountered in our recordings.

Consider, for example, the word "divide". "Divide" can be pronounced in such a way that the final "d" is released (resulting in a short /də/ syllable of highly variable intensity) or the final "d" may not be released at all. A human listener is not disturbed by this variability, but the recognition algorithm probably will have to be exposed to both types of pronunciation during the learning phase of the program before it can handle the problem.

Another variation encountered in words of this type is an optional prevoicing of the initial "d". Prevoicing of voiced stops is a free variation in English phonology that is discounted by the

---

* It is suggested that being a consistent speaker involves short-term memory, and forgetting is likely to occur after some (possibly small) time interval. Thus our recordings represent higher quality acoustic data than is likely to be encountered in any practical application.

listener. It is, however, a sophisticated decision for a computer to make; our recognition algorithm may again have to be exposed to both pronunciations.

The vowel in the first syllable of "divide" has as an underlying phonetic form the unstressed schwa vowel, /ə/. The acoustic character of a schwa vowel depends on several factors other than the underlying phonetic form. The consonantal environment has the greatest effect on the acoustic character of /ə/, but speaking rate, and, for example, the position of the tongue before the initial "d" can also influence the acoustic output. In the case of the word "divide" the vowel may appear as a short /ə/, /a/, /ä/or even /i/ and the word will still be heard as divide. There is some evidence that an individual speaker will vary between several of these alternatives if he is conciously attempting to help the system by carefully articulating each word.

Numerous other examples of speaker options can be cited.* All of them have the property that a relatively large change in acoustic pattern may occur at places in a word where an English listener is prepared to discount their appearance.

A speech recognition program has only two real alternatives in dealing with the permissible variability in English pronunciation. It can accept each variant as a new pattern to be learned to be learned (as is done in this research) or it must incorporate the

---

* In the vicinity of a nasal consonant, a sound segment may become nasalized without affecting the perception of the word. Certain acoustic correlates of word stress may be absent, or constrastive stress may be applied in situations where the speaker wishes to emphasize the word. Consonants appearing before unstressed vowels are normally rather indistinct, but may become more clearly articulated in slow or emphatic speech.

appropriate rules of English phonology in order to recognize times when specific acoustic features are important and to discount occurrences of acoustic features in free variation. In our view, the latter approach represents the most promising avenue to success in the long run, and the method described in this report has a limited range of practical applications.