SPEAKER-MACHINE INTERACTION IN AUTOMATIC SPEECH RECOGNITION

JOHN I. MAKHOUL

TECHNICAL REPORT 480

DECEMBER 15, 1970

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
RESEARCH LABORATORY OF ELECTRONICS
CAMBRIDGE, MASSACHUSETTS 02139

The Research Laboratory of Electronics is an interdepartmental laboratory in which faculty members and graduate students from numerous academic departments conduct research.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, by the JOINT SERVICES ELECTRONICS PROGRAMS (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract No. DA 28-043-AMC-02536(E), and by the National Institutes of Health (Grants 5 PO1 GM15006-03 and GM14940-04).

Requestors having DOD contracts or grants should apply for copies of technical reports to the Defense Documentation Center, Cameron Station, Alexandria, Virginia 22314; all others should apply to the Clearinghouse for Federal Scientific and Technical Information, Sills Building, 5285 Port Royal Road, Springfield, Virginia 22151.

71-403

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 480 December 15, 1970

SPEAKER-MACHINE INTERACTION IN AUTOMATIC SPEECH RECOGNITION

John I. Makhoul

Submitted to the Department of Electrical Engineering at the Massachusetts Institute of Technology in May 1970 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

(Manuscript received September 29, 1970)

Abstract

In this study the feasibility and limitations of speaker adaptation in improving the performance of a fixed (speaker-independent) automatic speech recognition system are examined. A fixed vocabulary of 55 [əCVd] syllables is used in the recognition system, where C is one of eleven stops and fricatives [p], [t], [k], [b], [d], [g], [f], [s], [š], [v], and [z], and V is one of 5 tense vowels [i], [e], [a], [o], and [u]. In the recognition of the [CV] syllable a highly interactive computer system is employed. The experiments were conducted in the computer room (where the signal-to-noise ratio rarely exceeded 25 dB) with the subject speaking on-line into a directional microphone whose output was amplified, sampled, and stored in the computer memory. The sampled speech waveform served as a source for playback and as raw data for analysis and recognition. The recognized [CV] syllable was displayed on a cathode-ray tube.

The results of the experiment on speaker adaptation, performed with 6 male and 6 female adult speakers, show that speakers can learn to change their articulations to improve recognition scores. The initial average error rate was 18.3%. As a result of the experiment it was predicted that the average error rate could decrease to 2.3%. The preliminary results also indicate that most of the necessary adaptation can be achieved in a relatively short time, provided the speakers are instructed about how to change their articulations to produce the desired effects. Several methods of changes in articulation were found useful, for example, lip rounding, diphthongization, deliberate efforts at voicing and/or frication. Errors between nonlabials were the most difficult to correct by changes in articulation.

The recognition scheme is based on the extraction of several acoustic features from the speech signal. This is accomplished by a hierarchy of decisions made on carefully selected parameters that are computed from a spectral description of the speech signal by means of a set of energoids (energy centroids), each energoid representing the center of energy concentration in a particular spectral energy band. Short-time spectra were obtained either from a bank of 36 bandpass filters covering the range 150-7025 Hz, or by directly computing the Fast Fourier Transform of portions of the sampled speech signal.

TABLE OF CONTENTS

CONTENTS

I. INTRODUCTION

In spite of the many methods now at his disposal, man still depends on speech for most of his everyday communication with other humans. This is not surprising because it is relatively easy for man to produce and perceive speech, even as a child. One of the reasons for the ease of speech communication between humans may be that man is born with the predisposition to manipulate patterns, symbols, and features such as those in speech, and to create rules that govern these manipulations (Keyser and Halle[1]). This attribute of man does not seem to extend, in its fullest manifestations, to other living beings. It is possible to train certain animals to produce and to respond to a very limited set of speech words, but it appears that it is not possible for those animals to deduce relationships or create rules that would help them perceive new utterances or produce the already learned words in some ordered fashion (so as to form very simple phrases, for example). The limitations seem to lie both in their articulatory and perceptual mechanisms.

Ever since the advent of the electronic digital computer man has been trying to improve his means of communication with "the machine." This is the well-known problem of input-output, how to make it faster, more efficient and meaningful. From punch cards to on-line typewriters, to graphical displays, there is an unending search for better communication. Why not, then, a speech interface with the machine?

A man-machine interface would require the machine to produce and "perceive" speech sounds. Basically, the machine is capable of storing information and performing certain mathematical and logical operations on that information in a manner that is precisely specified by an algorithm. Such an algorithm is ordinarily supplied by the user, and the performance of the machine is dependent almost entirely on the algorithm supplied to it. If the machine is to be able to produce and perceive or simply recognize speech sounds, it is the human, the user, who must specify the exact algorithms based on his knowledge of his own speech production and perception mechanisms. In man-animal speech communication the limitations are mainly those of the animals. In man-machine speech communication, however, the greatest limitation seems to be the human's knowledge of his own perceptual mechanisms. Recent speech perception theories provide for closely related processes of speech production and speech perception with certain components and operations that are common to both (Halle and Stevens,[2] Liberman et al.,[3] review articles by Stevens and House,[4] and by Lindgren[5, 6]).

At present, any man-machine speech interface can only be of a very elementary nature. Although machine generation of intelligible speech has become a reality, the machine recognition of speech is still in its most elementary stages and has been restricted mainly to a small set of isolated words or phrases spoken by a very limited number of speakers. The auditory perceptual mechanism operates at several hierarchical levels simultaneously, including the phonetic, morphemic, syntactic, and semantic levels, all of which are closely related. It is at the phonetic level that most of

the recent research has been directed, although there has been a considerable effort to study speech perception at the other levels also. As a result, most of the speech-recognition systems developed during the last two decades have worked at the acoustic and phonetic levels of speech recognition. Comprehensive surveys of attempts before 1968 may be found in Flanagan,[7] Lindgren,[5,6] and Hyde[8]; more recent attempts are those of Bobrow and Klatt[9] and Medress.[10]

Human speech communication is highly interactive. Therefore, it would be desirable to include some form of interaction in a man-machine speech interface. The ideal interaction would include adaptation by both the machine and the speaker. The problem of machine adaptation and learning forms a major area of research in artificial intelligence today, although a relatively small but increasing effort is devoted to speech. On the other hand, the problem of speaker adaptation to a particular recognition system or machine has received even less attention. This is surprising, since the human is known to adapt himself rather well to changing conditions and complicated tasks.

Almost all of the previous attempts at automatic speech recognition used the human simply as an input device that is capable of producing strings of acoustic (speech) waveforms to be recognized by the machine. There was always an attempt to adapt the machine to the speech of the human, but there was no deliberate attempt to adapt the human to the machine. If we speak in order that we may be understood, then should we not take into consideration the object of our speech? After all, when we speak to little children do we not adapt our speech in order that they might understand us better? Then, it seems reasonable that the human speaker should try to adapt his speech to the machine just as the machine is made to adapt its procedures to fit the speech of a particular speaker.

During the past decade it has become fashionable to talk about the "single-speaker approach" to automatic speech recognition. The recognition system is initially designed to recognize the speech of one particular speaker. In order to recognize the speech of a new speaker the system adapts its parameter values to the new speaker. This adaptation is often referred to as "tuning in" or "speaker normalization." In principle, this approach should lead to good recognition results. And it would, provided the parameters are well chosen. For the parameters to be well chosen, however, one must examine the speech characteristics of several speakers. That will ensure that the parameters are significant for a wide range of speakers. This is what might be called the "multispeaker approach" to automatic speech recognition. Here, the recognition system design is based on parameters that are applicable to many speakers. Then, during recognition, the system adapts its parameter values to each speaker. The recognition system in this report is designed with the use of the multispeaker approach. The system is essentially fixed and does not adapt to each speaker; i.e., there is no speaker normalization. The reasons for this lie in the objectives of this work, which are the following.

1. To test the feasibility of a very limited speaker-independent speech recognition system.

2. To find out the degrees of speaker-independence of different speech parameters. This would give an indication of which measurements are more reliable.

3. To see if speakers can help improve the performance of a particular recognition system by properly adapting their articulations without significantly affecting the human perception of the utterances.

How well a speaker can adapt or change his articulation to effect a certain change in the acoustic output is still not very well understood. For steady-state vowels and consonants there is evidence that significant changes in articulation do not always produce a correspondingly significant change in the acoustic output (Stevens[11]). The extent to which a speaker can modify the acoustic output during the transitions is not known, however. It is worth noticing that much of the information necessary for recognition lies in the transitional segments of speech, especially for the stop consonants, the nasals and the glides, when followed by vowels.

In Section II automatic speech recognition is discussed as a problem in pattern recognition, and feature extraction is argued as the principal tool that should be used in speech recognition. Section III goes through the rationale for the choice of the particular recognition problem and the associated methods of analysis. In Section IV there is a discussion of the acoustic correlates for the different vowel and consonant features that were analyzed using the methods of analysis described in Section III. The details of the recognition system are described in Section V. The performance results of the recognition system and of speaker adaptation to the system are reported in Section VI. Different methods of changes in articulation are also discussed.

The reader might wish to skip Sections IV and V on the first reading of this report. The detailed nature of these two sections might obscure the main thrust of the report, which is that of speaker adaptation.

II. PATTERN RECOGNITION OF SPEECH

Every spoken word can be considered as a two-dimensional pattern (amplitude vs time: time waveform), as a three-dimensional pattern (amplitude vs frequency vs time: spectrogram) after appropriate frequency analysis, or as any n-dimensional pattern that might result from a transformation applied to the time waveform. In each case, the problem of speech recognition can be viewed as a problem in pattern recognition. This view does not in any way change the character of the problem at hand, but it does tie the problem of speech recognition to other pattern-recognition problems, although admittedly this tie is just as loose and as general as that among many other pattern-recognition problems.

Among other definitions, pattern recognition comprises "the detection, perception, or recognition of any kind of regularity or relation between things. Both things and the relations between them are patterns ... Indeed, the recognition even of things requires some formulation of the abstract structure that characterizes them, which is already a formulation of the relations between their identifiable parts" (Kolers and Eden[12]). The problem of pattern recognition, then, is (i) to classify things and relations into patterns or pattern classes; (ii) given a pattern, to abstract the corresponding pattern class.

In a recent paper, Nagy[13] reviewed the state of the art in pattern recognition. Of the several methods discussed feature extraction seems to be the one appropriate for speech recognition. In feature extraction, according to Nagy, "one attempts to transform the sample space in such a manner that the members of each class exhibit less variability and the relative separation between the classes is increased, thus allowing the use of a simpler decision mechanism ... ." The invariance of a human's perception of the same word uttered by several speakers is an indication that this invariance must, in part, exist somewhere in the acoustic signal. A quick look at the different acoustic manifestations of the same word shows, however, that the invariance is by no means apparent; it must, somehow, be extracted from the acoustic signal. Furthermore, the acoustic signal does not always supply all the necessary information for the detection of that invariance. The human listener takes an active part in supplying some of the missing information. The words of Cherry[14] concerning human recognition in general are appropriate: "For human recognition is a psychophysiological problem, involving a relationship between a person and a physical stimulus; it is a phenomenon which can scarcely be explained solely in terms of properties of the object or pattern alone." It is not surprising to see that pattern-recognition schemes that employed brute-force template matching of multidimensional patterns have met with very limited success. The large variance and dimensionality of the acoustic speech signal have frustrated all attempts to automatically recognize speech by template matching of spectrographic patterns, for example. Such methods neglect the fact that the information in the speech signal is highly redundant. If the

redundancy in speech is ignored, then according to Shannon's formula, a channel capacity of up to 166,000 bits/sec (10-kHz bandwidth and a 50-dB signal-to-noise ratio) is needed to process that information. On the other hand, the human's ability to process information is very limited. Several experiments have been conducted to assess man's informational capacity and, according to Flanagan,[15] "None of the experiments show the human to be capable of processing information at rates greater than the order of 50 bits/sec." Some have maintained that "for sustained communication beyond a few seconds the practicable rates for most human activities have been closer to 10 bits per second than to 40" (Karlin and Alexander[16]). Since speech is produced and intended to be perceived by humans, it seems reasonable to assume that the information content in speech is closer to, and most probably less than, 50 bits/sec rather than 166,000 bits/sec. Our knowledge of how humans achieve this enormous reduction in information is still very limited. Needless to say, whatever knowledge is at hand ought to be helpful and should be employed in developing automatic recognition systems. There has been much evidence on the linguistic, articulatory, and perceptual levels for the existence of certain invariant distinctive features in speech (Jakobson, Fant, and Halle,[17] Chomsky and Halle,[18] and Wickelgren[19]). These features have been hypothesized at several levels of language. Most of the studies thus far have concentrated on the phonetic level. The acoustic correlates of many phonetic features are currently under study. (See Section IV for a limited exposition of the acoustic correlates of some of those features.)

The purpose of this discussion has been to indicate that feature extraction should indeed be helpful in automatic speech recognition. It is this writer's opinion that successful, moderately ambitious recognition systems will have to depend on a hierarchy of decisions based on the extraction of several features from the speech signal. No complicated mathematical classification techniques can obviate the necessity for such feature extraction. Classification techniques should serve as an aid rather than a substitute for feature extraction. A brief review of some of the automatic speech recognition systems employing some form of feature extraction is presented here.

Perhaps the first speech-recognition system using feature extraction methods was the electronically implemented automatic speech recognizer designed by Wiren and Stubbs.[20] Their system was based on the idea of distinctive features of Jakobson, Fant, and Halle,[17] and it functioned well for vowels in short words spoken by 21 speakers. Most of the features used were those of manner (e.g., voicing) rather than place of articulation.

As more was known about the acoustic correlates of place of articulation, speech recognition systems were developed using this information. The systems were either electronically implemented or simulated on an electronic digital computer. Martin, Nelson, and Zadell[21,22] attempted the recognition of several consonants and vowels in [CVd] syllables in real time using a bank of broad bandpass filters that simulate the auditory nerve in some fashion. According to their findings: "The features that are

more invariant and more easily abstracted by machine are the spectral regions of increasing and decreasing energy." With six male speakers the recognition scores were best for fricatives and worst for stops.

In 1965, Hughes and Hemdal[23] published a report on a computer recognition system for vowels in isolated words using four distinctive features. They used the single-speaker approach. For a single speaker they obtained results comparable with human performance.

Gold's word-recognition computer program attempted the recognition of 54 practically oriented English words.[24] A segmentation procedure was developed which depended on changes in energy level, voicing and over-all spectrum. The final probabilistic decision making was applied to a set of parameters that were computed from the spectral description of the speech signal as obtained from a 16-filter spectrum analyzer. Instead of specific formant measurements, the program computed three centers of energy concentration that corresponded very roughly to the first three formants. Only preliminary recognition results for one repetition by 10 male speakers were reported.

Reddy[25] attempted the computer recognition of connected speech using the sampled time waveform and some pitch-synchronous harmonic analysis. Briefly, he segmented the acoustic waveform into segments that corresponded roughly to phonemes, which he recognized by a set of heuristically developed features. The recognition scheme worked well mainly for manner-of-articulation features. Apparently the most difficult problem encountered was that of segmentation. Reddy also used the single-speaker approach.

A limited speech-recognition system with good results was developed by Bobrow and Klatt.[9] The vocabulary that was tested comprised a limited set of short phrases and words. The recognition scheme used both feature extraction and template matching, the template being the set of features for each utterance. Thus, the utterance is analyzed according to a set of features the presence or absence of which determines a "code pattern" which is then matched to the stored patterns. There is no attempt at segmentation into phonemes. The scheme is speaker-dependent; hence, each speaker has to train the system separately on the particular vocabulary. This recognition system is a good example of how pattern matching is used to "fill in the gap" where our knowledge of the speech recognition process is still lacking. This is an unavoidable aspect of an operative system at this time. A more recent version of this recognition system (Bobrow, Hartley, and Klatt[26]) included new features for place of articulation and an expanded vocabulary list of more than 100 words.

The most recent attempt at automatic speech recognition using feature extraction is that of Medress.[10] His method was developed to "abstract distinctive features information [using gross spectral properties] about vowels, stops, fricatives, and some sonorants from a filter-bank representation of single-syllable, single-morpheme English words." An important feature of Medress' scheme is that it exploits some rules

of English phonology. This approach to speech recognition of single words is the most promising thus far.

The approach that will be used in this report is that of feature extraction, which will depend upon decisions made on a number of computed parameters. The attempt is made to specify features and parameters that are speaker-independent to as large a degree as possible, both for male and female speakers.

III. PROBLEM FORMULATION AND DATA ANALYSIS

3.1 RECOGNITION PROBLEM

The utterances chosen for recognition are a set of CV (consonant-vowel) syllables. This choice allows for a close examination of the effect of changes in articulation on the consonant, the vowel, and the consonant-vowel transition. The vowels and consonants that were used are shown in Fig. 1. A set of 55 CV syllables results.

Five vowels: [i], [e]', [a], [o]', [u];

Eleven consonants:

six stops: [p], [t], [k], [b], [d], [g];

five fricatives: [f], [s], [š], [v], [z].

' The diphthongized vowels [e^{I}] and [o^{U}] are actually used, but they will be represented in this report by [e] and [o], respectively.

Fig. 1. Vowels and consonants in recognition problem.

The feature distinctions that are used with this set of vowels and consonants are (i) Vowels: front-back, high-mid-low, and (ii) Consonants: (a) Manner: stop-fricative, voiced-unvoiced, and (b) Place: labial-dental-alveolar-velar.

Figure 2 shows how the vowels and consonants in Fig. 1 are divided according to the above-mentioned features. For a specification of these features in terms of distinctive features and a comparison with the Chomsky and Halle distinctive features (Chomsky and Halle[27]) see Appendix A.

The vowels chosen are all tense, since these are the only vowels that can occur in final position in English. Lax vowels were avoided in any case because they tend to be reduced and are affected more by coarticulation (Öhman,[28] Lindblom,[29,30] Stevens and House,[31] and Stevens, House, and Paul[32]). The fricatives [θ] and [$\eth$] were not included in the consonant set because they are very often confused with [f] and [v], even for humans (Miller and Nicely[33]). The phoneme [$\overset{v}{z}$], the voiced correlative of [$\overset{v}{s}$], was not included because no English words can start with that sound. Other consonant classes such as nasals, liquids, and glides were not included in the consonant set because they would greatly expand the recognition problem. Also, knowing, for example, that the spectra of [m] and [n] are very useful in speaker-recognition systems (Glenn and Kleiner,[34] Wolf[35]) makes one hesitate to use nasals in the vocabulary of a speaker-independent automatic speech-recognition system.

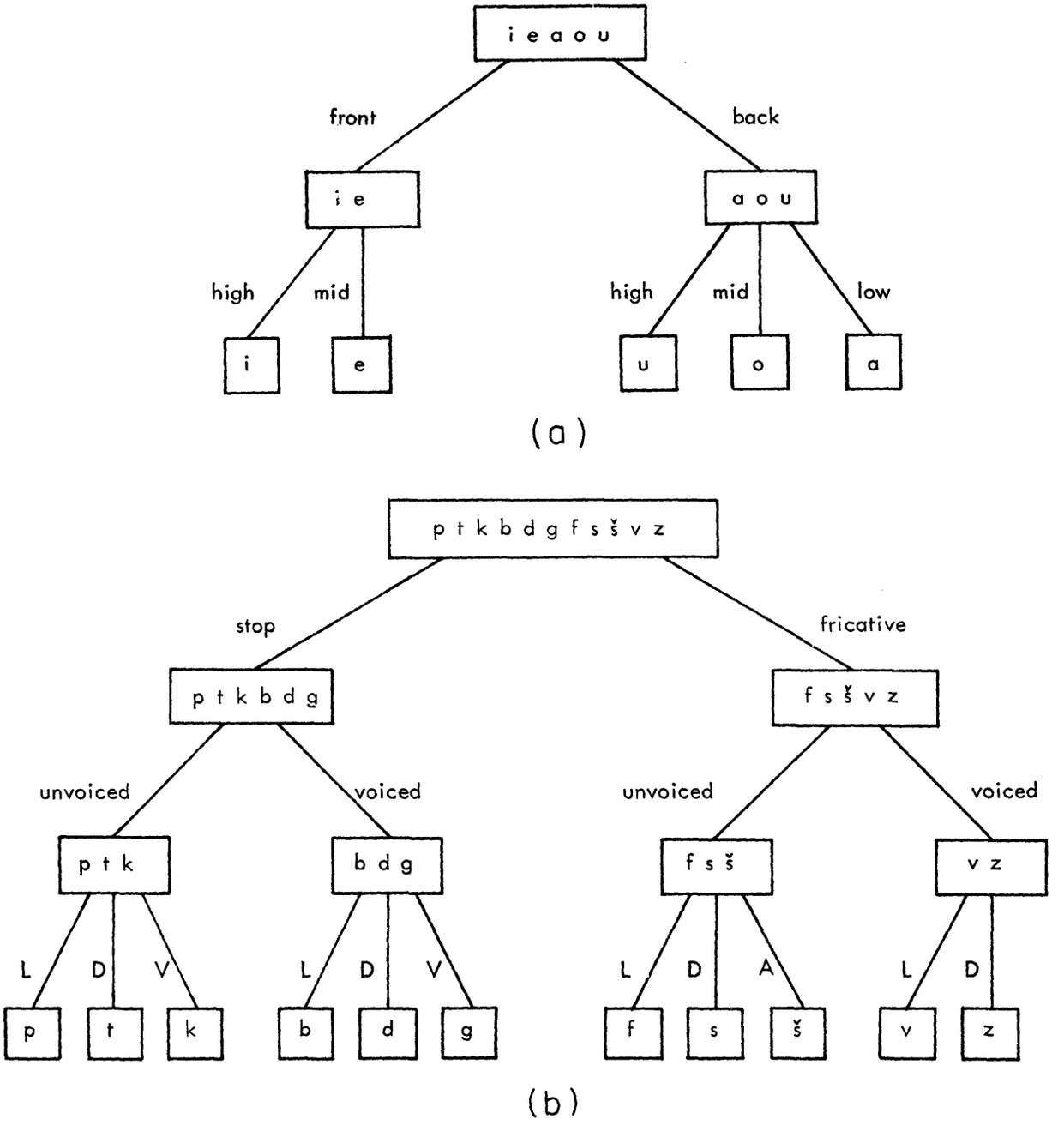


Fig. 2. (a) Vowels divided according to features.
(b) Consonants divided according to features.
L = labial, D = dental, A = alveolar-palatal,
V = velar.

	i	e	a	o	u
p	a peak a peep a pea appeal	a pate opaque	a pot a pop a pod	a pope a poke	a pool
t	a teak a tease	a take a tape	a tot a top a tog	a tote a toad	a toot a tube a tool
k	a keep	a cape a cake a cage	a cot a cop a cock a cod	a Coke a coat a code	a kook a coop a coot
b	a beep a bead a beat a beak	a bait a bake a babe obeyed	a bog a bop a bob	abode a boat	a boot a boor
d	a deed a deep a deer	a date	a dog a daub a dot	a dope	a dude a duke ado
g	a gear	agape a gate a gauge	a god	a goat a goad ago	a goop a ghoul
f	a feat a feed a fee a feel	a fake a phase	a fog		a food
s	a seed a seep a seat	a sake a sage assayed	a sop	a soap a soak	a soup a suit
š	a sheet a sheep a sheik	a shake a shade a shape	a shot a shop a shock	a show	a shoot a shoe
v	a veep a vee	evade		a vote evoke	
z	a zee		a czar		

Fig. 3. List of $[\partial C_1 V C_2]$ words whose $[C_1 V]$ syllables can be correctly recognized by the recognition scheme.

Previous research shows that initial and final phonemes in a word introduce special problems that do not exist intervocalically. For example,

1. The silence before the burst in an initial stop consonant cannot be easily distinguished from the silence that exists before closure.

2. The transitory period from silence into the initial sound is of such changing nature that it makes it difficult to analyze.

3. A final vowel tends to be very variable in length and in the way it approaches silence. Depending on the preceding consonant, a final vowel may not even assume a steady-state position.

Furthermore, the acoustic correlates of some features have been investigated mainly for intervocalic sounds.

Since we were seeking features that are invariant, it was desirable to embed the CV syllable in a constant environment, namely attach phonemes before the consonant and after the vowel. We hoped that such a constant environment would reduce the variability for each speaker. If a schwa, [ə], is placed before C, and [d] is placed after V, the utterance assumes the form [əCVd], where V is stressed and [ə] is unstressed. The schwa was helpful in reducing speaker variability, and the [əC] transition was sometimes helpful in place-of-articulation recognition. The choice of [d] to end the utterances was quite arbitrary. The [Vd] transition was actually never used in the recognition scheme. (As a matter of fact, replacing [d] with any other stop consonant did not affect the recognition results. Figure 3 is a table of English words whose CV syllables could be correctly recognized by the recognition system described in Section V. This statement is based on an informal test given to a few male and female subjects.)

The recognition problem, then, is the automatic recognition of 55 CV syllables in utterances of the form [əCVd], where [ə] is unstressed, as spoken by a group of adult males and females; the vowels and consonants are those shown in Fig. 1.

3.2 DATA COLLECTION

The final experiments to test the recognition scheme and to investigate the speakers' abilities to change their articulations were performed with subjects speaking on-line. For purposes of analysis, however, it was much more practical to have at hand recorded samples from several speakers. Those recordings were used in selecting the parameters that are important to specify the different acoustic features. Having the utterances recorded also helped in testing the relative performance of different values of various parameters.

Initially, the 55 [əCVd] utterances were recorded once through by each of 10 adult speakers (5 male and 5 female). Four of the speakers were recorded in an anechoic chamber and the rest were recorded in the PDP-9 computer room, of the Speech Communications Group of the Research Laboratory of Electronics, where the final experiments took place. The computer room is quite noisy: The maximum signal-to-noise ratio $(S/N)^1$ of an utterance varied from 12 dB for a soft-spoken female to 30 dB

for a loud-spoken male speaking at approximately 2 in. from a directional, dynamic microphone. (Henceforth, the S/N ratio of an utterance will refer to the maximum signal-to-noise ratio of the utterance. This maximum usually occurs around the beginning of the stressed vowel.)

It was soon clear that a very low S/N ratio was going to adversely affect the recognition results, especially those of the consonants. After experimenting with several subjects I decided that a minimum S/N ratio of 20 dB was reasonable to require of the subjects; if necessary, by speaking louder, or by coming closer to the microphone, or both. Problems associated with speaking close to the microphone are discussed in Section VI.

Even by placing the minimum S/N ratio at 20 dB we were still faced with a large range of S/N ratios: 20-40 dB (the 40 dB being normal for the recordings in the anechoic chamber). Since the final experiments were going to take place in the computer room we decided to eliminate the recordings made in the anechoic chamber from the data set. This reduced the S/N range to 20-30 dB. By requiring loud-spoken subjects to keep their distance from the microphone, it was possible to restrict the S/N range to 20-25 dB. In this manner, we hoped that the recognition scheme would have enough tolerance so that it would be insensitive to an S/N ratio within the range 20-25 dB. In practice, the upper-limit requirement of 25 dB was often surpassed, with no noticeable effects on recognition performance. On the other hand, the lower limit of 20 dB was a more stringent requirement; lower values ($\leq$18 dB) often caused recognition errors.

As a result of the 20 dB S/N ratio minimum restriction, a large portion of the recordings made in the computer room were rendered useless. New recordings were made in the computer room, often using different speakers from those used in the first recording. The resulting data base consisted of recordings of the 55 [əCVd] utterances by each of 6 adult speakers: 3 male and 3 female.

3.3 PROCESSING TECHNIQUES

The speech processing in this investigation was all performed on the highly flexible Speech Computer Facility (Henke[36]) of the Speech Communications Group of the Research Laboratory of Electronics. The particular configuration used is shown in schematic form in Fig. 4.

There are two sources of speech input to the system: (i) a program-controlled tape recorder, and (ii) a directional, dynamic microphone conveniently placed next to the CRT display. The input speech goes through an analog-to-digital (A/D) converter that samples the speech at 20 kHz into 9-bit samples which are stored two-per-word in the PDP-9 computer 24 K memory. The program is now able to store 810 ms of speech, which is enough for one [əCVd] utterance. All subsequent processing is performed on this sampled and stored data, which will be referred to as the time waveform.

The time waveform has been previously used directly in detecting vowel regions and consonant manner-of-articulation (Reddy[25]). In this system the time waveform is used in three ways.

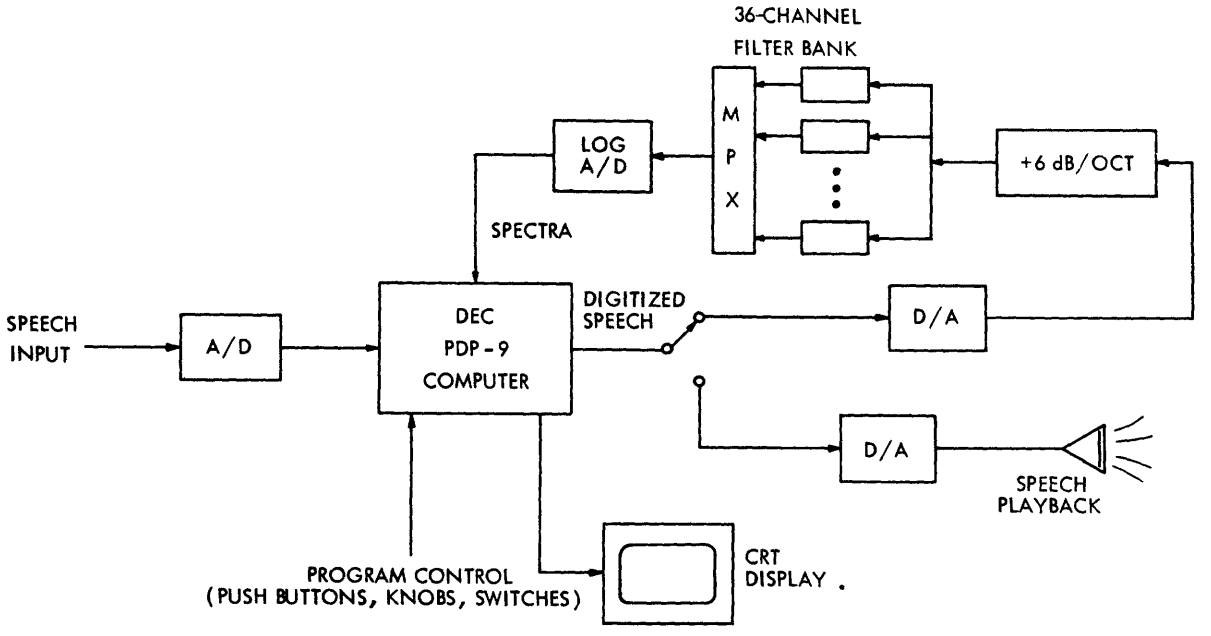


Fig. 4. Speech recognition system.

1. As a source of playback for checking purposes and as auditory feedback for subjects. The subject has access to a push-button that initiates the desired playback.

2. As the "raw material" for further processing. Different parts of the waveform can be processed differently depending on past processing.

3. As a time reference for specific speech events. Having the time waveform accessible makes it easy to accurately locate the burst in stop consonants.

These uses will be clearer further on.

Spectral analysis is used exclusively for processing the time waveform. Two methods are used, one employing hardware and the other software.

(a) Spectrum Analyzer

The spectrum analyzer (Fig. 4) comprises a bank of 36 simple-tuned bandpass filters (Fig. 5) whose outputs are rectified and lowpass-filtered (simple one-pole RC circuit with a time constant of 10 ms). The outputs of the lowpass filters are sampled, digitized on a logarithmic scale into 64 levels (6 bits) and stored in the computer memory. Using a 36-channel multiplexer, an automatic scan of the filter outputs is made that runs from lowest to highest frequency in approximately 1.3 ms (which is small compared with the 10-ms time constant of the lowpass filters). Frequency characteristics

Filter Number i	Center Frequency (Hz)	Bandwidth (Hz)
1	150	100
2	250	100
3	350	100
4	450	100
5	550	100
6	650	100
7	750	100
8	850	100
9	950	100
10	1050	100
11	1150	100
12	1250	100
13	1350	100
14	1450	100
15	1550	100
16	1650	125
17	1775	125
18	1900	150
19	2050	150
20	2200	175
21	2375	175
22	2550	200
23	2750	200
24	2950	225
25	3175	225
26	3400	250
27	3650	275
28	3925	300
29	4225	325
30	4550	350
31	4900	375
32	5275	400
33	5675	425
34	6100	425
35	6550	475
36	7025	475

Fig. 5. Filter set data.

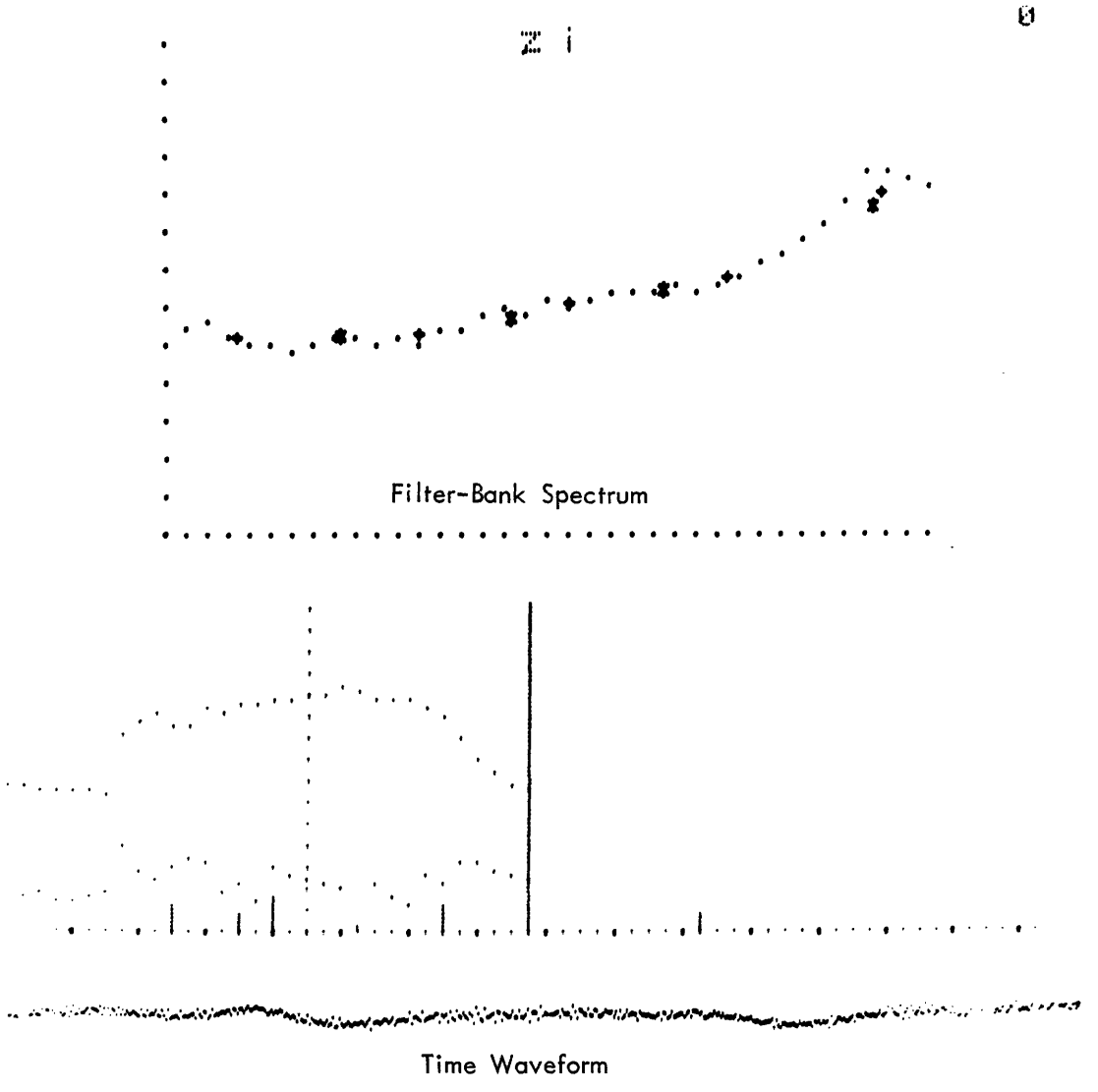


Fig. 6. Time waveform display and the corresponding filter-bank spectrum for [z] from the utterance [əzid].

of adjacent filters intersect at the -3 dB points. Before processing by the filter bank the speech is pre-emphasized by a filter with a rising characteristic of +6 dB/octave.

In order to find the spectrum at any time t, the time waveform is played through a D/A converter into the spectrum analyzer (Fig. 4) starting at time t-40 ms until time t when the filter outputs are sampled and stored as mentioned above. The lower part of Fig. 6 shows 25.6 ms of the time waveform corresponding to the [z] portion of [əzid]. The most right-hand point in the time waveform corresponds to time t. The corresponding spectrum, as obtained from the filter bank, is shown in the upper part of Fig. 6. The filter number and the corresponding center frequency are marked on the abscissa; and the relative energy in dB at each frequency is plotted on the ordinate (see also Fig. 9). The other portions of Fig. 6 will be explained later.

(b) Fast Fourier Transform (FFT)

The short-time spectrum of any portion of the time waveform can be calculated using the FFT technique (Cochran, et al.[37]). An example of such a spectrum is shown in Fig. 7. Note that the time waveform is exactly the same as that in Fig. 6. The spectrum coordinate scales in both figures are identical. (The 512-point FFT that is actually used gives the spectrum at equal frequency intervals of 39 Hz. So, a regrouping of energies was necessary to obtain the energies on the new scale.) Note that the spectra obtained by the two methods are different. Figure 7 shows that the FFT spectrum gives a better indication of voicing and has a better frequency resolution than that obtained by using the filter bank. The computation of the FFT is a very slow process as compared with the essentially instantaneous access to the filter-band spectrum, however, and the

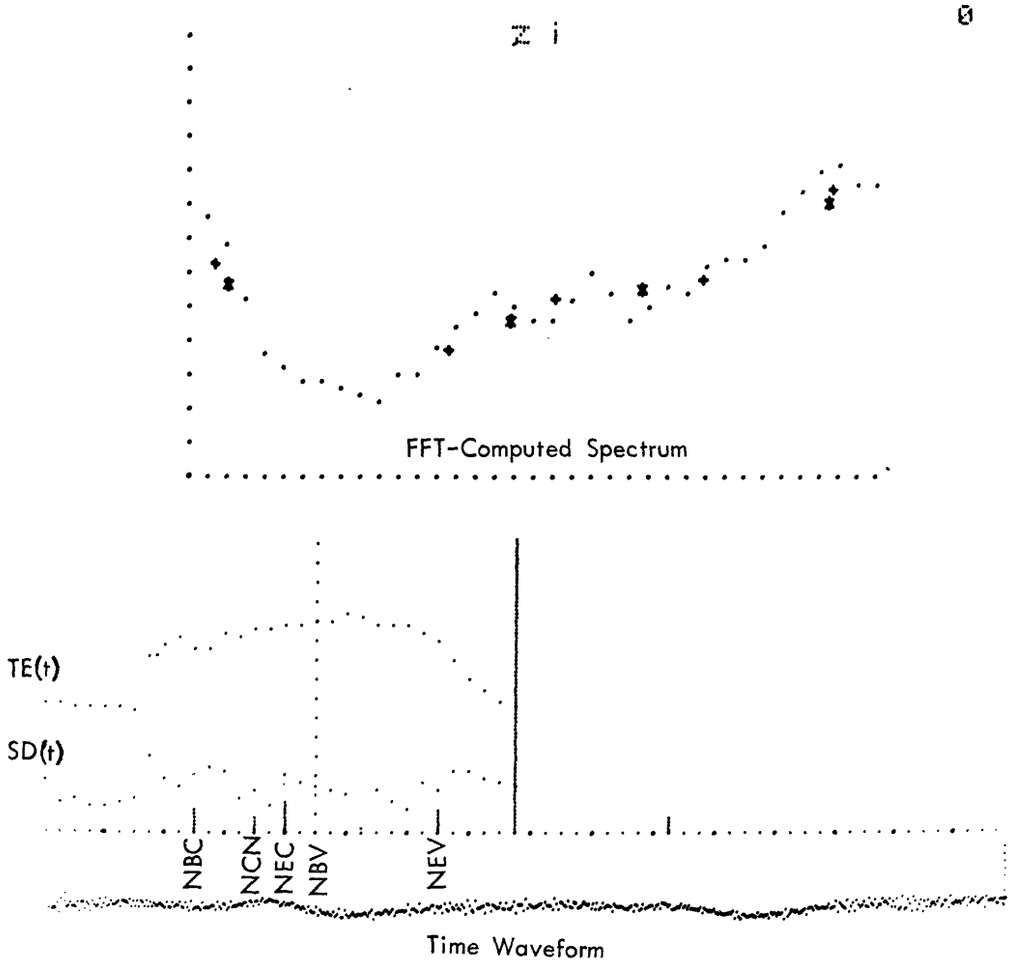


Fig. 7. Time waveform and corresponding FFT-computed spectrum for [z] from the utterance [əzid]. Also displays of the total energy TE(t) and the spectral derivative SD(t) and the spectral derivative SD(t) for the whole utterance.

factor of speed is very important if a man-machine interactive system is to be meaningful and effective.

For more on the particular FFT algorithm used, see Appendix B. A discussion of FFT vs filter-bank spectra is also included.

3.4 ANALYSIS METHODOLOGY

In order to motivate the methods of analysis used in this recognition scheme, I shall first discuss some of the relevant methods of vowel and consonant analysis that have been used by other researchers.

3.4.1 Previous Vowel Analysis

By far the most extensively studied phonemes have been the vowels. The first three formants were measured for a large number of speakers and reported by Peterson and Barney.[38] A result of that experiment was that it is not possible to classify vowels reliably for a large number of speakers according to the locations of the first and second formants only. Using the data of Peterson and Barney, Gerstman[39] applied a normalization technique based on the locations of the formants F1 and F2 for the three vowels [i], [a], and [u] for each individual speaker, and plotted the rest of the vowels on this self-normalized space. The results show some improvement over the original data as far as classification of vowels is concerned, but the improvement was not adequate for absolute automatic recognition of vowels.

An attempt at vowel recognition by linear discriminant function techniques was performed by Suzuki, Kasuya, and Kido[40] for the Japanese vowels [i], [e], [a], [o], and [u] spoken by 113 speakers. Their conclusion was that together with F1 and F2, "F3 or F0 is an indispensable parameter in order to recognize the vowels without distinction of the sex and age of the speakers." Correct recognition scores improved from 92.9% to 97.2% when either F3 or F0 was used in the recognition scheme.

A different approach to vowel recognition was that of Hughes and Hemdal.[23] Their criteria for recognition were based on the distinctive features of Jakobson, Fant, and Halle[17] rather than on absolute discrimination of each vowel. For example, the ratio (F2-F1)/(F3-F2) was used as one measure of acuteness; the ratio is larger for acute vowels than the corresponding grave ones. They used a total of 4 features to classify 10 vowels: compact-diffuse, acute-grave, tense-lax and flat-plain. For a single speaker the recognition accuracy was 97% for the tense vowels and 84% for the lax vowels.

Most vowel-recognition systems up to date have used the formant frequencies almost exclusively in one form or another. There have been problems associated with locating formant frequencies (Flanagan,[41] Ladefoged[42]). One of the problems is that often formants are so close to one another that they appear to merge into one broad formant. This occurs frequently with F1 and F2 in the vowel [a]. The reasons that formant frequencies have been so extensively used are twofold. First, historically, the spectrogram provided a clear display of vowel formants as early as

1947 (Potter, Kopp, and Green[43]). Second, the transfer function of the vocal tract in terms of poles and zeroes for particular configurations have been mathematically related to the formant frequencies (Fant[44, 45]). Thus formants have been used mainly in terms of speech production rather than speech perception, hence the relative success of speech synthesizers. Perceptually, there is evidence suggesting that although particular formant frequencies may be important for the perception of steady-state vowels, the dynamic mode of perception of speech is substantially different (Fujimura[46]). The preliminary results of recent perceptual experiments show that the ear is more sensitive to movements of broad energy concentrations than to movements of individual formants.[47]

This discussion indicates that it might be desirable to process vowels in terms of regions of major energy concentrations as they appear in the spectrum. Those regions may consist of one or more formants, and there may be more than one such region in a particular spectrum.

3.4.2 Previous Consonant Analysis

The discussion here will be restricted to the stop and fricative consonants.

The acoustic correlates of the fricatives have been investigated by several researchers (Hughes and Halle,[48] Strevens,[49] Heinz and Stevens,[50] and Stevens[51]). The position of the major spectral energy concentration in the "steady-state" portion of the fricative carries most of the information as to its place of articulation. The dependence on the following vowel is usually minimal (except, for example, when the vowel is rounded there is often a noticeable downward shift in the position of the energy concentration). Hence the relative success of systems attempting to recognize fricatives (Martin et al.[22] achieved recognition rates of 98% for 6 male speakers).

In contrast to the fricatives, the stop consonants have been some of the most difficult consonants to recognize automatically (Martin, et al.[22]). The reasons for this are twofold.

1. Acoustically, the stop is a period of silence, corresponding to the articulatory closure, followed by a burst after the release, followed by a period of aspiration (for unvoiced stops only) and then voicing of the following vowel starts. The silence is, by far, the longest part of this process, and it conveys no information other than that the phoneme is a stop. So, most of the information is contained in the period following the release of the stop, which is a short period compared with the whole duration of the stop. Most recognition systems have been designed to cope with phonemes that possess a more or less steady-state condition between transitions. These systems have not performed well for stop consonants where the "steady-state" part is the silence, which carries no place-of-articulation information, and almost all of the information lies in the transitional part.

2. Until recently the stop consonants had been investigated by measuring burst frequencies and formant transitions (Fischer-Jorgensen,[52] Halle, Hughes, and Radley,[53]

and Fant[45]). Experiments on the perception of synthesized stop consonants were also analyzed in terms of burst frequencies and formant transitions (Cooper, et al.,[54] Liberman, et al.,[55] Harris, et al.,[56] and Hoffman[57]). The main conclusion that can be drawn from these experiments is that the burst frequencies and formant transitions of stop consonant-vowel syllables are very context-dependent; they depend very much on the vowel following the stop, and some stops are more context-dependent than others. Theories like the "hub" theory (Potter, Kopp, and Green[43]) and, more importantly, the "locus" theory (Delattre, Liberman and Cooper[58]) have attempted to provide a framework to describe the results of the experiments in a concise manner. The "locus" theory is a very attractive one and, in fact, it does apply to a large chunk of the available data; but the counterexamples are too many to be disregarded, so that one must be careful not to overgeneralize. At any rate, it would be possible to have a recognition system which would handle the different context dependencies. Such a system would be both cumbersome and unattractive.

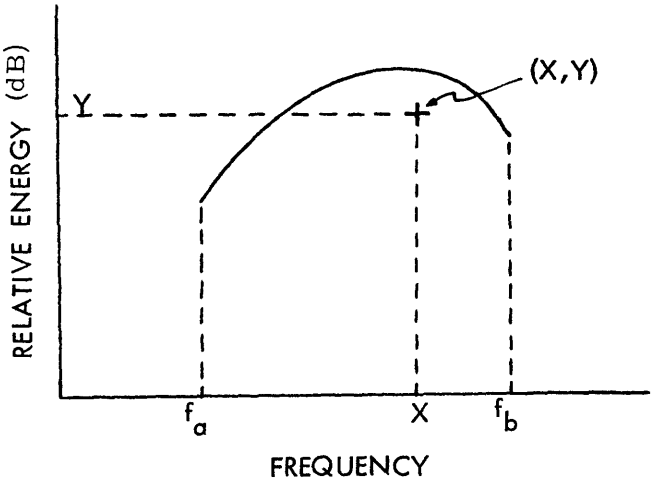
What appears to be a more promising framework for the specification of the acoustic correlates of place of articulation for consonants was put forth by Stevens.[59] It deals with the build-up of major energy concentrations in different regions of the spectrum relative to those of the following vowel. A more analytical analysis of the acoustic correlates of place of articulation for stop and fricative consonants in terms of "approximately coincident formants" is given by Stevens.[51] And it is well known that when two (or more) formants become approximately coincident, an increase in their amplitudes results (Fant[44]), thereby giving rise to a major spectral energy concentration. We shall dwell more on the acoustic correlates of consonants in Section IV.

3.4.3 Analysis with Energoids

The discussions in sections 3.4.1 and 3.4.2 point to one idea, namely, that processing in terms of major spectral energy concentrations might be desirable for both vowels and consonants.

Another reason for working with energy concentrations rather than with formant frequencies is a practical one: The first 19 filters (up to 2 kHz) of the filter bank (Fig. 5) have bandwidths of 150 Hz or less; hence, for a female speaker, the spectrum tends to show the individual harmonics instead of the desired spectrum envelope. This causes multiple peaks to occur in the spectrum, of which only a few can be associated with formants.

How does one characterize the concentration of energy in a portion of the spectrum, for example, the band of spectral energy shown in Fig. 8? The method that is used in this investigation is to compute the center of mass of spectral energy in the frequency band under consideration, shown as a plus sign in Fig. 8. Henceforth, each center of mass will be called an energoid (energy centroid). The X coordinate of the energoid gives the approximate location of the major energy concentration in the frequency band, and the Y coordinate is simply the "average energy" in that band.



(a)

	Energoid	Filter Number Range	Frequency Range (Hz)
+ energoids	1	1 - 7	150 - 750
	2	8 - 15	850 - 1550
	3	15 - 22	1550 - 2550
	4	22 - 29	2550 - 4225
	5	29 - 36	4225 - 7025
* energoids	A	1 - 15	150 - 1550
	B	10 - 20	1050 - 2200
	C	16 - 28	1650 - 3925
	D	24 - 36	2950 - 7025

(b)

Fig. 8. Various energoids and their ranges.

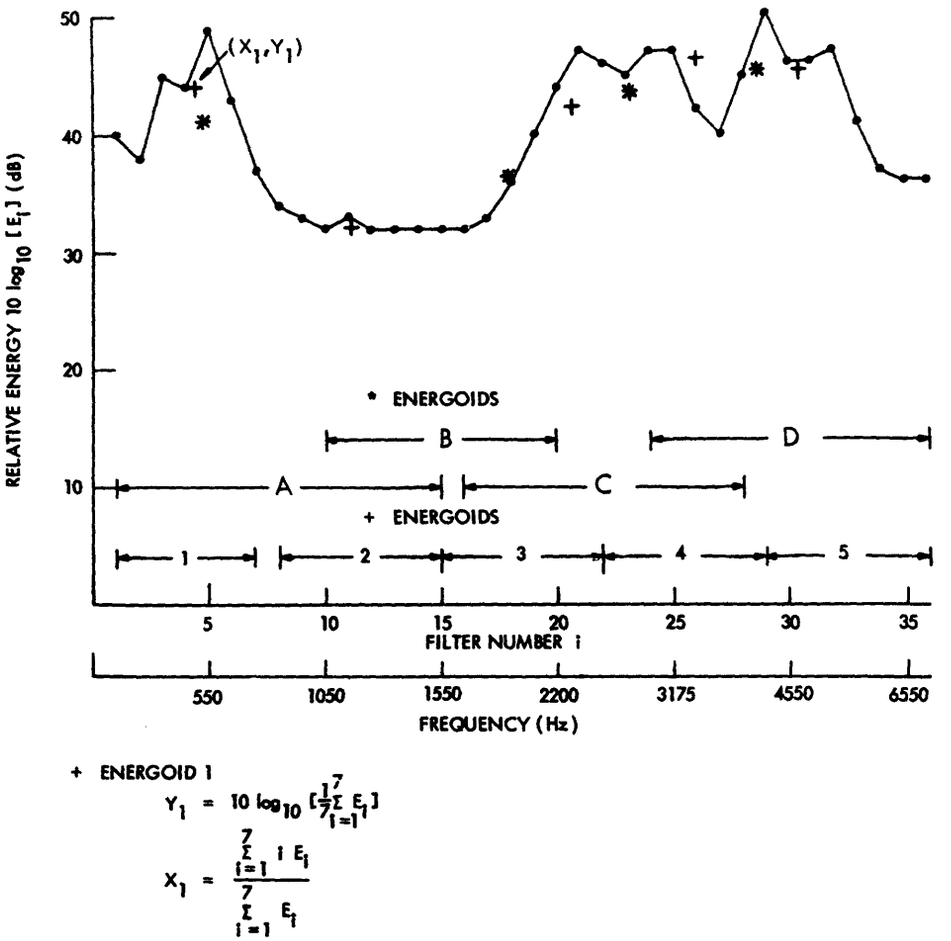


Fig. 9. Distribution and definition of energoids.

The spectrum was divided into several regions (Figs. 8b and 9). The energoid of each region is marked in Fig. 9 by either a plus (+) or a star (*). The regions with the + energoids form one set of 5 nonoverlapping regions and those with the * energoids form another set of 4 overlapping regions. Note that each of the two sets of regions covers the whole spectrum and that a region with a * energoid covers at most two adjacent regions with + energoids. The equations for computing the coordinates of each energoid are as follows (see example for + energoid 1 in Fig. 9):

$$X_n = \frac{\sum_{i=a}^{b} iE_i}{\sum_{i=a}^{b} E_i}$$

$$Y_n = 10 \log_{10} \left[\frac{1}{b-a+1} \sum_{i=a}^{b} E_i \right] \text{dB}, \qquad (1)$$

where E_i is the energy, in linear units, at the frequency corresponding to the filter number i, a = first filter number in energoid range, b = last filter number in energoid range, and n = energoid number = 1, 2, 3, 4, 5, A, B, C, or D. Note that all energoid ordinates are measured in decibels so that each energoid can be directly plotted on the spectrum which is also plotted in decibels. It should be emphasized that E_i in Eq. 1 is measured in linear units and not in decibels. The corresponding decibel measurement is defined by $10 \log_{10} [E_i]$, as shown on the ordinate of Fig. 9.

Figure 9 shows the spectrum of [i] as spoken by a female speaker. Note the position of the + and * energoids. They give a rather good general description of the spectrum. Changes with time of energy concentration in a certain region can now be monitored by simply tracking the corresponding energoid. Relative changes in energy concentration between regions can also be monitored by tracking the corresponding energoids and noting differences in their X and Y positions.

There is no doubt that information about the positions of particular formants is all but lost in this method, but instead we have a good idea about where general energy concentrations are located. One major property of the method of energoids is its insensitivity to small "bumps" or spurious peaks which are very common with female speakers. One example (and one could cite more dramatic examples) is the region of the first formant in Fig. 9, where there are two peaks instead of one: Which of the two is the first formant, or is it somewhere in between? Note the position of energoid 1; X_1 seems to be a fairly good indicator about where F1 might be. For many other cases X_1 is a good approximation to the location of the first formant. This is not a coincidence: For front vowels one expects the center of energy concentration for frequencies up to 750 Hz to be quite close to F1, since that is the only energy concentration in that region. Although energoid 1 bears a close relation to F1, the other energoids do not have any particular significance in terms of formant frequencies.

The choice of the frequency ranges for the different regions in Fig. 8b was not arbitrary. They were decided upon only after examining sample spectra for several male and female speakers. The regions had to be important in terms of energy movements for particular features and had to apply to as many speakers as possible (in this case 6 speakers were used in the analysis). The ranges shown in Fig. 8b are by no means optimal. As a matter of fact, other ranges are used for specific purposes. For example, the energoids used in vowel analysis are energoids 1, 2, A, and C, with energoid C having a range of filters 16-30 instead of 16-28. Also, during the analysis of the schwa a new energoid is defined that depends on the location of the energy concentration corresponding to the second formant. For all other purposes, however, the ranges are as shown in Fig. 8b. Note that each of the regions of the * energoids covers no more than two adjacent regions of the + energoids (see Fig. 9). After some inspection it can be seen that the following condition must hold between X coordinates of the + and * energoids:

$$X_1 < X_A < X_2 < X_B < X_3 < X_C < X_4 < X_D < X_5. \quad (2)$$

This X-inequality is exemplified in the positions of the energoids in Figs. 6, 7, and 9.

3.5 ANALYSIS PARAMETERS

Following is a list of parameters that were available for analysis. Each parameter could be computed at any point in the time waveform, plotted on the CRT display, and its value displayed. Means were available to automatically evaluate and plot a parameter at equal intervals in time for the whole utterance. All of the parameters listed below are directly derived from spectra that are obtained from the time waveform either using the filter bank or the FFT, as explained in section 3.3. Every spectrum consists of the relative energy present at each of 36 filters in the filter bank (Fig. 9).

For reasons of computing facility, only positive integers were allowed as parameter values. All negative values were placed equal to zero, and all fractional parts were truncated. This meant that some parameters had to be scaled appropriately. The use of the following parameters in recognition is described in Section V.

(a) Total Energy

$$TE(t) = 10 \log_{10} \sum_{i=1}^{36} E_i(t), \quad (3)$$

where $E_i(t)$ is the energy, in linear units, at the frequency corresponding to filter number i, at time t. This parameter was intended to give some indication of the "short-time" total energy in the waveform at time t. Note, however, that because of the +6 dB/octave pre-emphasis (Fig. 4) TE(t) emphasizes the energy at higher frequencies and, hence, does not reflect the "true" total energy. Figure 7 is a plot of TE(t) computed every 25.6 ms for the utterance [əzid].

(b) Spectral Derivative

$$SD(t) = A * 10 \log_{10} \frac{\sum_{i=1}^{36} |E_i(t)-E_i(t_1)|}{\sum_{i=1}^{36} |E_i(t)+E_i(t_1)|} + B, \quad (4)$$

where $E_i(t)$ is defined above, and $E_i(t_1)$ is similarly defined at time t_1. A is a positive scale factor, and B is an additive positive constant. Note that SD(t) is also a function of t_1; however, if $(t-t_1)$ is fixed, as will usually be the case, then it is enough to state the value of $(t-t_1)$ and simply concentrate on SD(t).

This parameter assumes its least values when the spectrum changes very little, as in the steady-state portions of vowels and consonants. Any change either in the amplitude

or in the spectral shape causes SD(t) to increase. The constants A and B were experimentally chosen such that the significant values of SD(t) were positive and conveniently expanded to allow for a useful display. Several values were tried, but the final values chosen were A = 1.5, and B = 20. Figure 7 is a plot of SD(t) computed every 25.6 ms for the utterance [əzid].

(c) Energoid Coordinates

The definition and properties of energoids have already been discussed in section 3.4.3. The X and Y coordinates of an energoid, defined over a range of filters, are given by Eq. 1. Because of the truncation problem, both Eqs. 1 and 2 are multiplied by 10. The new definitions become

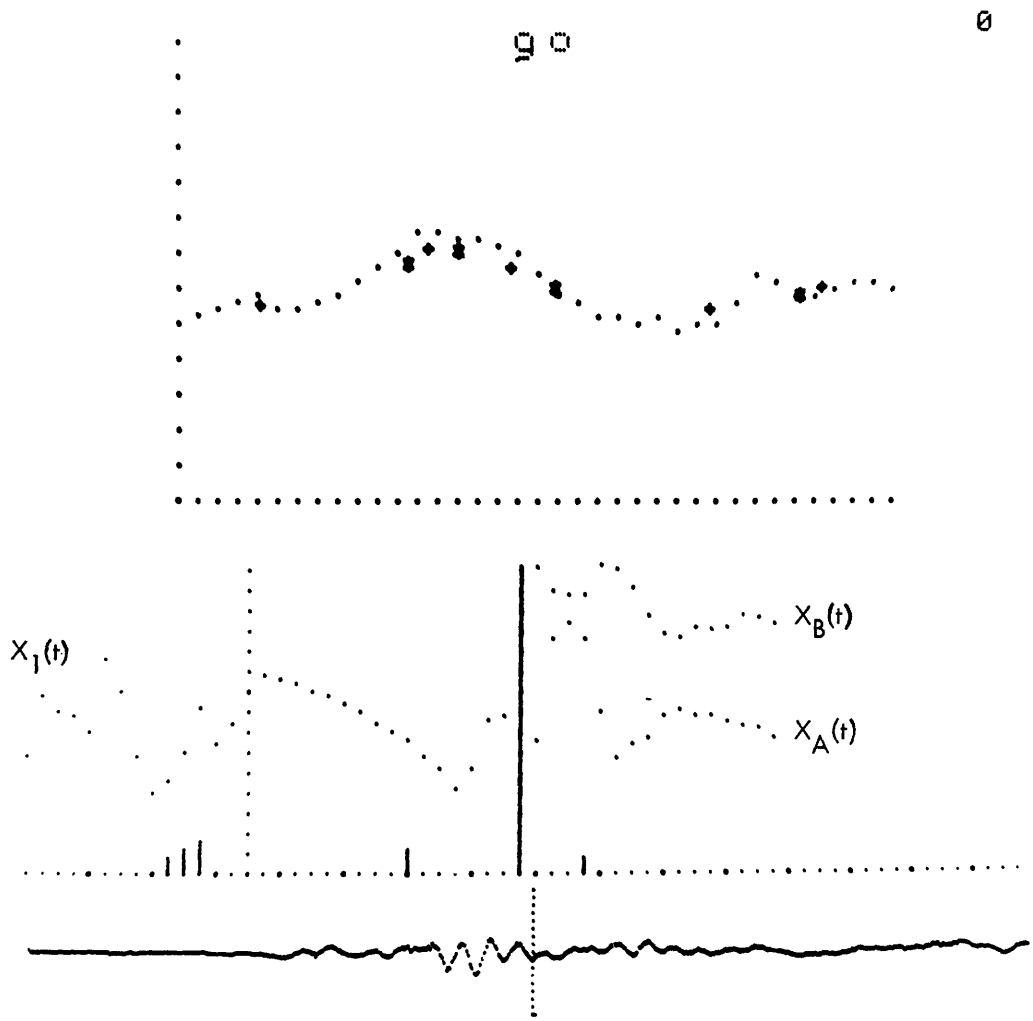


Fig. 10. Time waveform and corresponding filter-bank spectrum of the [g] burst in [əgod]. Also a display of $X_1(t)$ for the whole utterance, and displays of $X_A(t)$ and $X_B(t)$ beginning at the stop burst.

$$X_n(t) = 10 \frac{\sum_{i=a}^{b} iE_i(t)}{\sum_{i=a}^{b} E_i(t)}$$

$$Y_n(t) = 100 \log_{10} \left[\frac{1}{b-a+1} \sum_{i=a}^{b} E_i(t) \right], \tag{5}$$

where a, b, and n are as defined in Eq. 1.

A display of $X_1(t)$ is shown in Fig. 10 computed every 25.6 ms for the utterance [əgod]. Also shown are displays of $X_A(t)$ and $X_B(t)$ starting from the beginning of the stop burst and into the following vowel.

Energoid parameters which will also be useful are the differences of energoid coordinates defined as follows:

$$XD_{mn} = X_m - X_n$$

$$YD_{mn} = Y_m - Y_n, \tag{6}$$

where m and n are energoid numbers, and X_n and Y_n are defined by Eq. 5.

(d) Parameter LMAXY(t)

For a particular spectral frame, LMAXY is the filter number at which the first local maximum occurs after the point of maximum positive slope in the spectrum. This parameter will be useful in consonant recognition (sec. 5.4). The computation of LMAXY proceeds as follows:

Let

$$YD_{rs}(t) = \max [YD_{mn}(T), \text{ mn} = A1, B2, C3, D4, 2A, 3B, 4C, 5D], \tag{7}$$

where $YD_{rs}(t)$ is an approximation to the maximum positive slope of the spectrum at time t, in terms of differences of the Y coordinates of adjacent energoids. (Adjacent energoids are defined by Eq. 2.) As a result, either r or s must be a numeral (corresponding to a + energoid) and the other must be a letter (corresponding to a * energoid).

Let

$$k = \begin{cases} r, & \text{if r is a numeral} \\ s + 1, & \text{if s is a numeral.} \end{cases}$$

Then

$$\text{LMAXY}(t) = \min i \left[E_i(t) \text{ is local max, } i \geq X_k(t)/10 - 2, \right.$$
$$\left. \text{and } 100 \log_{10} E_i(t) > Y_k(t) \right], \tag{8}$$

where min i [condition] ≡ minimum i such that the condition is satisfied.

The analysis methodology described above will be useful in determining the acoustic correlates of the different features, reported in Section IV. The analysis parameters are essential in the recognition scheme that is detailed in Section V.

IV. ACOUSTIC CORRELATES

Before turning to the details of the recognition scheme it would be very useful to discuss some of the more important acoustic correlates of the different features that are used in this recognition system.

The discussion will be limited to features corresponding to the vowels and consonants shown in Figs. 1 and 2 and mainly in the context [əCVd]. These results were obtained from an extensive analysis of the recorded utterances of the six speakers using the processing and analysis techniques described in Chapter III. Professor Dennis Klatt's collection of spectrograms for all [əCV] English syllables was very useful, especially in the detection of the relative importance of the [əC] transition in place-of-articulation recognition.

4.1 ACOUSTIC CORRELATES OF VOWEL FEATURES

The pertinent vowel feature oppositions are front-back and high-mid-low. Each of these features refers to an articulatory position of the tongue body.

4.1.1 Front-Back

The vowels [i] and [e] are front, while [a], [o], and [u] are back.

The main acoustic correlate of this feature is the position of the major spectral energy concentration, or what will be called henceforth the region of prominence. For a vowel, the region of prominence is the spectral region associated with the second formant F2. The region of prominence for a back vowel includes F1 and F2, and that for a front vowel includes F2, F3, and perhaps higher formants.

For many adult male and female speakers the magical dividing frequency is around 1600 Hz: The region of prominence lies below 1600 Hz for back vowels, and above 1600 Hz for front vowels. This applies to the 5 vowels both under steady-state and coarticulated conditions. Care must be taken with coarticulations of back vowels like [u] and [o] with postdentals like [t], [d], [s], and [z]. In those cases, F2 starts quite high and may be closer to F3 than to F1 and the tendency would be to classify the vowel as front. It is well known, however, that F2 decreases with time to a value approaching that of the steady-state vowel. So, it is safer to determine the region of prominence by examining the latter portion of the vowel.

4.1.2 High-Mid-Low

The vowels [i] and [u] are high, [e] and [o] are mid, and [a] is low.

The position of the first formant is the only consistent acoustic correlate for "highness." F1 increases as the tongue body is moved from high to mid to low, ceteris paribus. It is true that the position of F2 can be correlated with highness, but the correlation is a function of at least one other feature: front-back. Even then, changes in F2 are in general not very consistent.

Now, although the first formant is a reliable correlate of highness, one must be able to distinguish three levels of F1 corresponding to the three positions: high, mid, and low. This is not a simple task especially across many speakers. It is rather easy to establish a test on F1 that would separate high from low vowels, and would apply to a large number of speakers. The real problem is to be able to make the high-mid and mid-low distinctions. Across many speakers, the F1 space for mid vowels partially overlaps both that of the high vowels and that of the low vowels. The solution to this problem makes use of the fact that the mid vowels [e] and [o] are usually diphthongized in American English, and are pronounced as [e^I] and [o^U], respectively. Both vowels are diphthongized from a mid vowel toward a high vowel. This means that F1 decreases as a function of time. Therefore we now have two parameters for differentiating high-mid-low vowels: the level of F1, and the relative decrease of F1 with time.

4.2 ACOUSTIC CORRELATES OF CONSONANT FEATURES

The features for stop and fricative consonants are usually divided into two classes: features for the manner of articulation, and features for the place of articulation.

4.2.1 Manner-of-Articulation Features

This class includes the features stop-fricative and voiced-unvoiced.

(a) Stop-Fricative

The consonants [p], [t], [k], [b], and [g] are stop consonants, while [f], [s], [$\overset{\vee}{s}$], [v], and [z] are fricative consonants.

The main acoustic correlate for stop consonants is the relative absence of energy during the closure before the burst. This period of closure contains no energy (except for ambient noise) for unvoiced stops, and some voicing energy for voiced stops. The amount of voicing energy present is very speaker-dependent and it always decreases with time. In contradistinction to stops, fricatives are characterized by relatively more energy that is due to frication noise. The amount of frication energy present depends both on the place of articulation and on the particular speaker. Between speakers, it is possible to have a voiced stop with more energy than that in a weak labial fricative, and with little difference in their spectra. (This problem will be discussed further in Section VI.)

(b) Voiced-Unvoiced

The consonants [b], [d], [g], [v], and [z] are voiced, while [p], [t], [k], [f], [s], and [$\overset{\vee}{s}$] are unvoiced.

One of the more important acoustic correlates of voicing is the presence of low-frequency energy caused by voicing. This is true for both stops and fricatives. Also, the presence of aspiration after a stop burst indicates a voiceless stop. Under certain contexts the lengthening of the preceding vowel can be an indication of a voiced stop or

fricative, but this is not important in the context [əCVd], and hence will not be further discussed.

4.2.2 Place-of-Articulation Features

As the name implies, a place-of-articulation feature indicates the place along the vocal tract where the consonant is considered to be articulated. For a fricative, this refers to the place of the small constriction at which the turbulent noise or frication is produced. For a stop, it is the point of complete closure along the vocal tract, which is also the place where the turbulent noise is generated after the release of the stop.

The places of articulation for the different consonants under consideration are

1. Labial: [p], [b], [f], [v];
2. Postdental: [t], [d], [s], [z];
3. Alveolar: [s̬];
4. Velar: [k], [g].

Refer to Appendix A for a specification of places of articulation in terms of distinctive features.

Instead of listing the acoustic correlates for each place of articulation separately, we shall discuss several acoustic events and the roles that they play in connection with the different places of articulation. These will include the [əC] transition, the turbulent noise energy, the burst length and aspiration for stops, and, most importantly, the [C] spectrum and the [CV] transition.

(a) [əC] Transition

The transition from the schwa to the consonant in the context [əCVd] carries some information as to the place of articulation of the consonant. Figure 11 shows a typical representation of the movements of F1 and F2 for the schwa in [əCVd], where C is labial in Fig. 11a and nonlabial in Fig. 11b. The significant difference between the two figures is that F1 and F2 move parallel to each other with labials and are divergent with nonlabials. A simpler, but less general, way of stating this is that F2 decreases with labials and increases with nonlabials. (Note that for some nonlabials such as [θ], F2 could decrease but F1 and F2 would still be divergent.) This generality holds for most of the utterances that have been examined. Exceptions are bound to happen because [ə] is initial and unstressed in the utterance [əCVd]. Needless to say, the amount of decrease or increase in F2 is highly variable between speakers and for different places of articulation. Still, the overriding characteristic is that F2 decreases for the schwa before labials.

(b) Turbulent Noise Energy

In general, the turbulent noise generated at the lips, when either fricatives are produced or stops are released, is weaker and contains less energy than that produced by

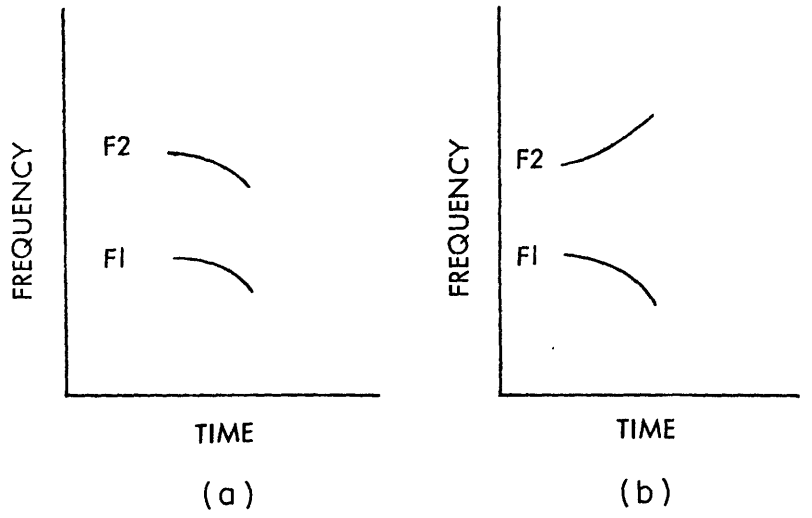


Fig. 11. Traces of F1 and F2 for the schwa in [əCVd] for (a) labial [C] and (b) nonlabial [C].

nonlabials. The frication energy for [f] and [v] is usually weaker than that for [s], [$\overset{v}{s}$] or [z]. The bursts of [p] and [b] are weak when compared with the bursts of [t], [k], and [d].

(c) Burst Length and Aspiration Length

For stops, the burst length is well correlated with the place of articulation. The burst, short and often nonexistent for labials, is longer for postdentals and longest for velars. For voiced stops in the context [əCVd], the burst length is defined as the distance in time between the beginning of the burst and the initiation of voicing of the following vowel. For unvoiced stops the burst is directly followed by, and assimilated into, a period of aspiration which terminates upon the initiation of voicing. Since both the burst and the aspiration are noiselike, it is usually difficult to measure the burst length for unvoiced stops. Therefore, the burst length is potentially useful only for voiced stops.

Although the burst length is a reliable acoustic correlate of place for any single speaker, it is very difficult to make use of that fact when one tries to set absolute limits of burst length for many speakers. Between speakers, and for the same stop, a typical ratio of longest to shortest burst lengths is 4:1. Compare this with the ratio of the burst length for [g] to that for [d], which is typically in the order of 2:1 for any one speaker. The problem of differentiating [g] from [d] in terms of burst length is then evident. One could still make use of extreme circumstances, however, to aid in recognition. For example, a burst length greater than 30 ms is almost certainly that of a [g] and not a [d]. Also, a nonexistent burst usually indicates a labial. This test is reliable only if the algorithm for determining burst lengths is accurate within a couple of milliseconds because burst lengths of 5 ms for [d] are not uncommon.

The existence of aspiration after a burst is an indication that the stop is unvoiced. The absence of aspiration, however, does not necessarily indicate that the stop is voiced (for example, the [k] in [ski] is not aspirated in American English). In the context [əCVd] an unvoiced stop consonant is aspirated. The length of aspiration (including the burst length) can be roughly correlated with the place of articulation. The aspiration length of labials is generally shorter than that of nonlabials.

(d) [C] Spectrum and [CV] Transition – (Spectral Correlate)

With each consonant we shall associate a characteristic consonant spectrum or [C] spectrum. For fricatives, the [C] spectrum is simply the frication spectrum; for stops, the [C] spectrum is the spectrum of the stop burst. The [CV] transition will be discussed in terms of spectral energy movements from the [C] spectrum into the following vowel. The [C] spectrum and the [CV] transition together contain much of the information necessary for place recognition. The relative importance of each depends largely on the particular consonant.

The acoustic correlate under consideration (the spectral correlate) is the position, on the frequency scale, of the major energy concentration of the [C] spectrum for one place of articulation relative to:

1. The position of the major energy concentration of the [C] spectrum for other places of articulation, keeping the vowel the same;
2. The position of the region of prominence of the vowel.

Figure 12 shows the relative ordering of the places of articulation according to the position of the major energy concentration on the frequency scale. From highest to lowest the ordering is postdental, alveolar, velar, then labial. Also listed are the formants associated with each place of articulation. This means that during the [CV] transition the major energy concentration of the [C] spectrum splits into either or both of the formants listed (see Stevens[51]). In the case of velars, the major spectral peak coincides with F2 before back vowels and with F3 before front vowels. Keeping in mind that F2 and F3 are very close to each other for front vowels and that the region of prominence is that associated with F2, we can make the following generalization: The major energy concentration for the [C] spectrum of velars always lies close to the region of prominence of the following vowel. This statement and the ordering given in Fig. 12 completely specifies, in general descriptive terms, the spectral acoustic correlate for place of articulation. All of this presumes the existence of a major energy concentration in the [C] spectrum. For postdentals, alveolars, and velars such an energy concentration is almost always present, but not so for labials. In many cases the [C] spectrum for labials is quite flat with no energy prominence anywhere. When a major energy concentration does exist, however, it is usually at lower frequencies and it obeys the ordering in Fig. 12. This is certainly true for [p] and [b]. Also, for several speakers tested, [v] displayed the low-frequency energy concentration (around 1000 Hz). The only possible anomaly is [f] which, in many cases, has energy

Place of Articulation	Position of major energy concentration of the [C] spectrum on the frequency scale	Associated Formants
Postdental	highest	F4, F5
Alveolar	↓	F3, F4
Velar	↓	F2, F3
Labial	lowest	- - -

Fig. 12. Relative positions of major energy concentrations for different places of consonant articulations and their associated formants.

prominence at higher frequencies, but not at the lower frequencies expected. One could argue that if indeed [v] does, at times, exhibit low-frequency energy prominence, then so should [f]. But since this has not been the case, we must conclude that the low-frequency resonances have been severely damped. (Low-frequency energy concentrations for [f] have been reported by Hughes and Halle.[48]) The low-frequency resonances for labials are possibly due to half-wave resonances of the back cavity, and it is not too surprising that these resonances are not always excited, since the noise source is at a small orifice anterior to the mouth cavity. In any case, the acoustic correlate under discussion does not seem to be too reliable for labials. This should be no cause for alarm because we already have three other acoustic correlates for the labial-nonlabial opposition: the [əC] transition, turbulent noise energy and burst length (for stops only). Henceforth, the spectral correlate for place of articulation will be discussed mainly for nonlabials.

Let us study further some of the implications of the formants associated with the different places of articulation according to Fig. 12. Typically for vowel spectra, the formants that change most, on a scale such as that in Fig. 9, are F1 and F2, and to a much lesser extent F3; F4 and F5 seem to be resistant to large variations. Furthermore, F3, F4, and F5 are usually bunched together at the higher frequencies. Therefore, one would expect the positions of energy peaks for postdentals and alveolars to be quite stable and not change much with the following vowel. On the other hand, one would expect the positions of energy peaks for velars to change appreciably, depending on the following vowel. The lack of dependence of the [C] spectrum of postdentals and alveolars, and the close dependence of velars on the following vowel are borne out under analysis and also in experiments on the perception of natural stops (Schatz[61]) and fricatives (Harris[62]). Now, since the only nonlabial fricatives in this study are postdentals and alveolars, it should be possible to recognize the place of articulation using only the [C] spectrum with complete disregard of the following vowel. The region of

prominence of the vowel, however, should play a role in the recognition of stops, since they include velar consonants.

It is rather fortunate, for recognition purposes, that neither the stops nor the fricatives in English employ all three places of articulation: postdental, alveolar, and velar.

V. RECOGNITION SCHEME

We shall now give most of the details of the recognition scheme, except for some minute details. The scheme that was developed is basically a quantification of the acoustic correlates described in Section IV in addition to other less important correlates and some ad hoc procedures.

There were two major guidelines in developing this recognition scheme. First, the total computation time for the recognition of one CV syllable should be as close to real-time as possible because it is necessary that the speakers get feedback on the recognition result as soon as possible. With the PDP-9 computer facility that was used, this meant that several special subroutines had to be developed to cut down execution time. Also, some computations had to be completely eliminated because of time considerations. The final execution time for recognition was 4-5 sec, depending on whether the consonant was a fricative or a stop. The execution time could have been reduced further with more complicated programming, but subjects did not complain about the delay, so the program was not changed.

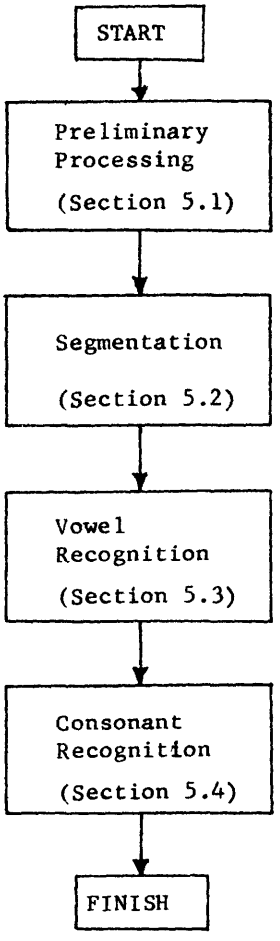


Fig. 13. The recognition scheme.

The second guideline was that the speaker would have a chance to change his articulation to effect correct recognition. This meant that requirements on the recognition scheme would not be as severe, since we hoped that, in case of recognition errors, speakers would be able to provide the necessary articulations for correct recognition. The result was a simpler recognition algorithm, at least in certain portions of it.

A block diagram of the recognition scheme is shown in Fig. 13. The preliminary processing starts by sampling the speech waveform and storing the samples in the computer memory. The stored time waveform is then used to compute a set of functions at equal intervals of 25.6 ms. These functions include the total energy, the spectral derivative, the coordinates of some of the energoids, and front-back decisions that are used later in the vowel recognition. Using the total energy and spectral derivative functions, the segmentation of the utterance into a schwa, a consonant and a vowel proceeds (see block diagram in Fig. 16). Following segmentation, the vowel is recognized (see block diagram in Fig. 21).

Table 1. List of parameters and locations of their definitions.

I. Functions of Time or of Frame Number

Parameter	Definition	Parameter	Definition
TE(t)	Eq. 3	$XD_{mn}(t)$	Eq. 6
SD(t)	Eq. 4	$YD_{mn}(t)$	Eq. 6
$X_n(t)$	Eq. 5	LMAXY(t)	Eq. 8
$Y_n(t)$	Eq. 5		

[Energoid numbers and ranges are shown in Fig. 8.]

II. Specific Frame Numbers or Time Markers

Frame Number	Corresponding Time Value	Definition
NSCH	TSCH	Eq. 12
NBC	TBC	Eq. 13
NBV	TBV	Eq. 16
NEC	TEC	Eq. 18
NEV	TEV	Eq. 20, 21
NCN	TCN	Eq. 26
QBBR	TBBR	Eq. 33
–	TBRST	Eq. 34

In general:

Time value = [Frame Number] * [fixed time interval]
+ initial time offset

III. Other Parameters

Parameter	Definition
APBL	Eq. 32
BG	Eq. 35
DEN	Eq. 36
INMD	Eq. 25
KMVM	Eq. 32
LBRST	Eq. 38
LCON	Eq. 19
MAMIN	Eq. 30
MAXV	Eq. 17
MINIM	Eq. 28
MNXB	Eq. 39
MSXB	Eq. 40
PAL	Eq. 36
SD12	Eq. 41
SMMD	Eq. 24
VC	Eq. 27

A decision about whether the vowel is front or back is made first. The vowel recognition is then completed by deciding whether the vowel is high, mid or low. The recognition of the consonant follows (see block diagram in Fig. 25). Based on the minimum energy level during the consonant portion of the utterance, the consonant is recognized as either a stop or a fricative. If the consonant is recognized as a fricative, an FFT-derived spectrum computed in the fricative portion of the utterance is used to determine whether the fricative is voiced or unvoiced, and also the place of articulation of the fricative. If the consonant is recognized as a stop (see block diagram in Fig. 31), the voiced-unvoiced decision is made using both an FFT-derived spectrum taken during the closure portion of the stop and a rough measure of the length of the burst and aspiration. The burst spectrum and parameters computed during the [CV] spectral transition are then used in the place-of-articulation recognition. Such recognition depends on whether the following vowel is front or back.

Because many parameters will be defined in the course of the algorithm specification, it is very useful to tabulate the different parameters and where they are defined, so as to provide easy cross reference. Table 1 provides partial lists of such parameters.

5.1 PRELIMINARY PROCESSING

5.1.1 Computation of TE(n), SD(n), $X_1(n)$, $X_A(n)$, FB(n)

The speech input is sampled into 9-bit samples at 20 kHz and stored in the computer memory. The stored time waveform is normalized so that the absolute value of the peak is set equal to 255_8 (the maximum absolute value allowed by the 9 bits). This, of course, does not improve the S/N ratio, but it does help the experimenter in the inspection of the time waveform.

At fixed intervals of time T = 25.6 ms (as measured along the time waveform) several parameters are computed from the spectrum, which is obtained from the filter bank (see section 3.3). Every such set of parameters computed at one point in time will be referred to as a frame. For a time waveform of 810 ms (limited by the buffer size) a total of 31 frames is computed. The parameters that are computed for each frame (see section 3.5 for definitions) are the following.

1. Total energy TE(n).
2. Spectral derivative SD(n).
3. The X and Y coordinates of energoids 1, 2, A and C, where the range of energoid C is modified to a filter range of 16-30.
4. Front-back vowel parameter FB(n).

In each parameter

$$t = nT + t_o, \tag{9}$$

with n the frame number, and t_o an initial time offset approximately equal to 30 ms.

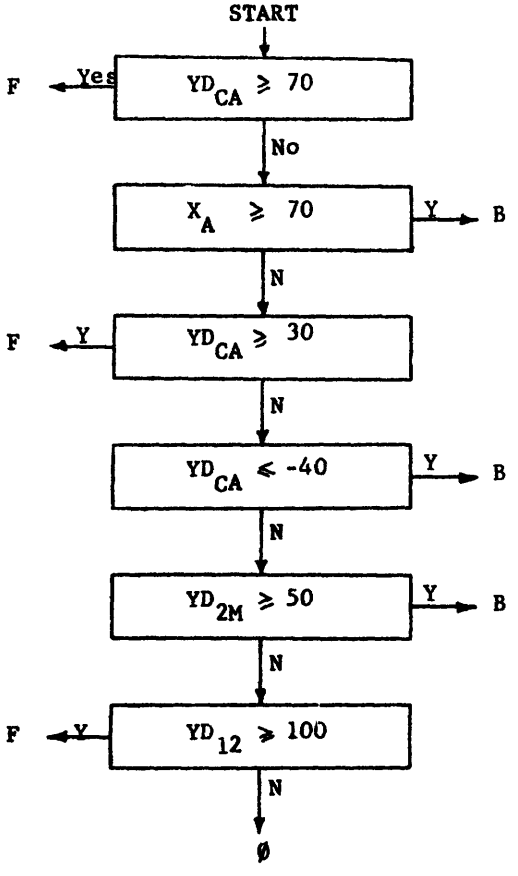
TE(n) and SD(n) will be useful mainly for segmentation purposes. The energoid

FB(n)	F-B-Ø Decision
0	Ø (don't know)
1	F (front)
2	B (back)

Fig 14. Front-back variable FB(n).

computations are intended for later use in the vowel recognition, but are computed here for all 31 frames.

Now, under the assumption that every spectral frame belongs to the vowel portion of the utterance, a test is made to see whether the spectrum is that of a front or a back vowel. The resulting decision will be known as the F-B-ϕ decision, the ϕ indicating a don't-know decision. A variable FB(n) is assigned a value depending on the decision made and according to the table in Fig. 14. The values of FB(n) during the vowel portion will determine whether the vowel is front or back.



$YD_{2M} = Y_2 - 100 \log_{10}[\min(E_8 \ldots E_{15})]$

Fig. 15. F-B-ϕ Test – Flow chart for front-back decisions.

The only parameters that are stored for further usage are TE(n), SD(n), $X_1(n)$, $X_A(n)$, and FB(n). Each of these functions can be displayed and the values examined on the CRT display. Figure 7 shows TE(n) and SD(n) for the utterance [əzid]. Figure 10 shows $X_1(n)$ for the utterance [ə god]. Other examples are provided by Figs. 18-20.

5.1.2 Front-Back Decisions

Figure 15 shows a flow chart for the F-B-ϕ decisions. YD_{CA} and YD_{12} are defined by Eq. 6. YD_{2M} is defined in Fig. 15.

The main acoustic correlate of the feature front-back is the level of energy

prominence at higher frequencies vs that at lower frequencies (section 4.1.1). This is exemplified quantitatively by the tests on YD_{CA} in Fig. 15. The test on X_A is intended to isolate back vowels that exhibit a good deal of energy at high frequencies, for example, [a]. The test on YD_{2M} is a rough test for the existence of a second formant in the region of energoid 2. The test on YD_{12} isolates front vowels with a well-defined energy dip after the first formant. Failing all the tests shown in Fig. 15 a verdict of "don't know" is given.

A final decision about whether the vowel is front or back depends on the F-B-ϕ decisions made along the vowel portion of the utterance. The final front-back decision is an important one, since it influences the course of several portions of the recognition scheme, as will be evident later on. For this reason, the F-B-ϕ test shown in Fig. 15 is very important. The simplicity of the F-B-ϕ test is reinforced by the fact that several of those tests are employed in determining the final front-back decision.

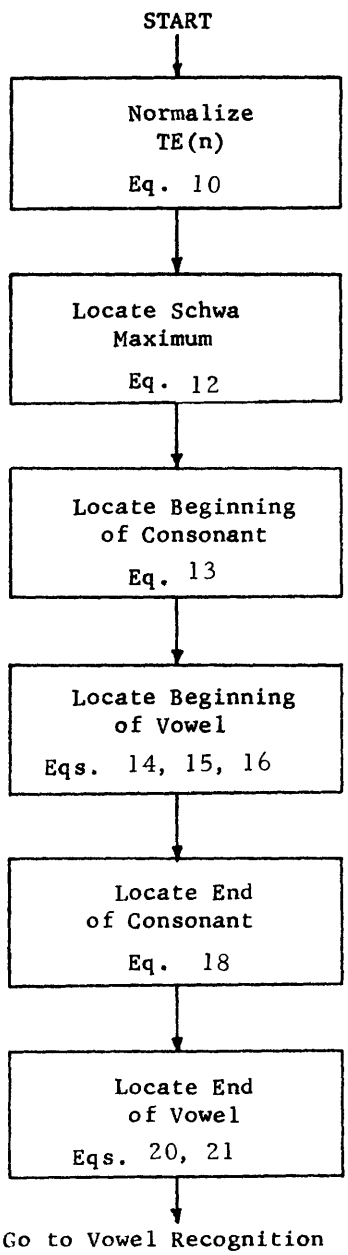


Fig. 16.
The segmentation procedure.

5.2 SEGMENTATION

The problem of segmentation is one of the stickiest problems in speech recognition. A satisfactory solution has not yet been found, even for isolated words. Some reasons for this are discussed by Fant.[63] In the recognition problem at hand, however, for which all utterances are of the form [əCVd], the segmentation problem is not quite so difficult. What is required is to locate the schwa, the consonant, and the vowel. The [Vd] transition is never used in the recognition scheme and, therefore, will be ignored henceforth. Actually, the scheme works just as well if [d] is replaced by any other plosive or is omitted from the utterance.

The only two parameters that are used in this portion of the recognition scheme are the total energy TE(n) and the spectral derivative SD(n). In order to render TE(n) independent of noise level, it is normalized by subtracting the noise energy level in decibels from each of the 31 computed values of TE(n) for the whole utterance. That is,

$$\text{Noise Energy Level} = \min [TE(1) \ldots TE(31)]$$

$$TE(n) \leftarrow TE(n) - \text{Noise Energy level}, \quad n = 1 \ldots, 31 \tag{10}$$

so that a normalized TE(n) = 0 dB would correspond to the absence of the speech signal. Henceforth, TE(n) will represent the normalized total energy as defined by Eq. 10. It might be of interest to note that the spectrum of the noise in the computer room is essentially flat, with the +6 dB/octave pre-emphasis included.

A block diagram of the segmentation procedure is shown in Fig. 16. After TE(n) is normalized, the following events are located in order: schwa maximum, beginning of consonant, beginning of vowel, end of consonant, and end of vowel. A mnemonic is given to the frame number associated with each of these locations as shown in Fig. 17. The following inequalities always hold, as will be seen below:

$$\text{NSCH} < \text{NBC} \leq \text{NEC} < \text{NBV} < \text{NEV}. \tag{11}$$

Examples of NBC, NEC, NBV, and NEV are shown superimposed on the total energy and spectral derivative displays in Fig. 7 for the utterance [əzid]. Other examples

Frame Number Mnemonic	Event
NSCH	Schwa Maximum
NBC	Beginning of consonant
NEC	End of consonant
NBV	Beginning of vowel
NEV	End of vowel

Fig. 17. Mnemonics for certain frame numbers.

are shown in Figs. 18-20 for the utterances [əgod], [ətud], [əvad], [əsid] and [əfid]; the total energy and spectral derivative plots were taken from displays similar to that in Fig. 7. The plot of $X_1(n)$ in Fig. 18 is relevant to high-mid-low recognition (section 5.3.2). NSCH is not shown in Figs. 18, 19, and 20, but can be detected as the schwa maximum in each case. NCN, shown in these figures, will be defined in section 5.4.2.

(a) Computation of NSCH – (Schwa maximum)

$$\text{NSCH} = \min n\,[\text{TE}(n) \text{ is a local max, and } \text{TE}(n) > 7 \text{ dB}], \tag{12}$$

where min n [condition] ≡ minimum n such that the condition is satisfied.

The 7 dB threshold is necessary to avoid local maxima that often occur because of a temporary interruption of voicing immediately following initiation of voicing. In the time waveform this is usually seen as a couple of pitch periods followed by a short period of silence comparable to one or two pitch periods. The schwa maximum location will be useful in the schwa analysis (section 5.4.1).

(b) Computation of NBC – (Beginning of consonant)

$$\text{NBC} = \min n\,[n > \text{NSCH}, \text{ and } \text{TE}(n) \text{ is local min}]. \tag{13}$$

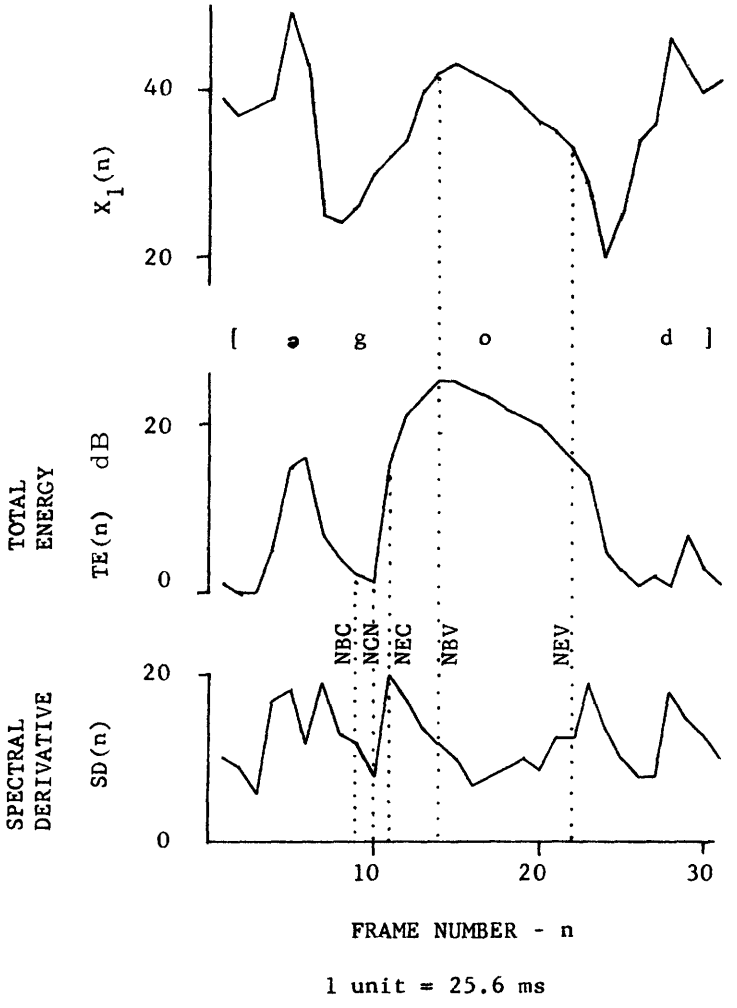


Fig. 18. The X coordinate of energoid 1, the total energy, and the spectral derivative for the utterance [ə god].

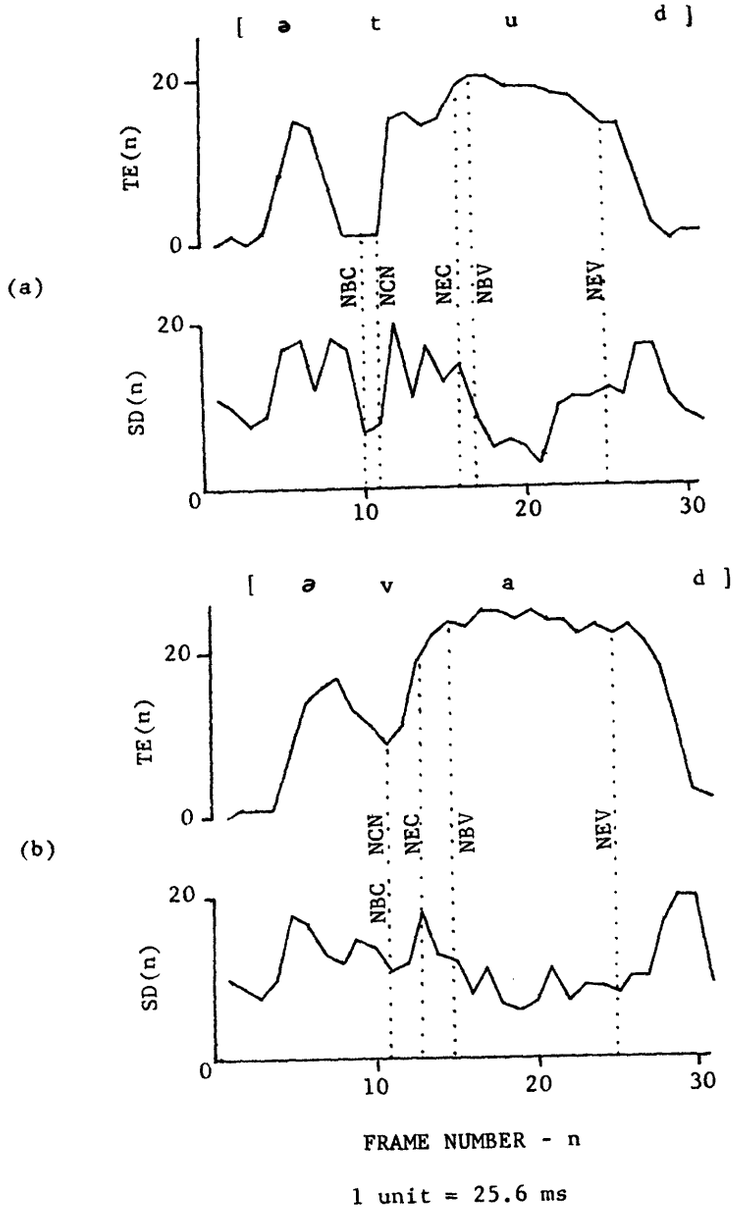


Fig. 19. Total energy and spectral derivative for the utterances [ətud] and [əvad].

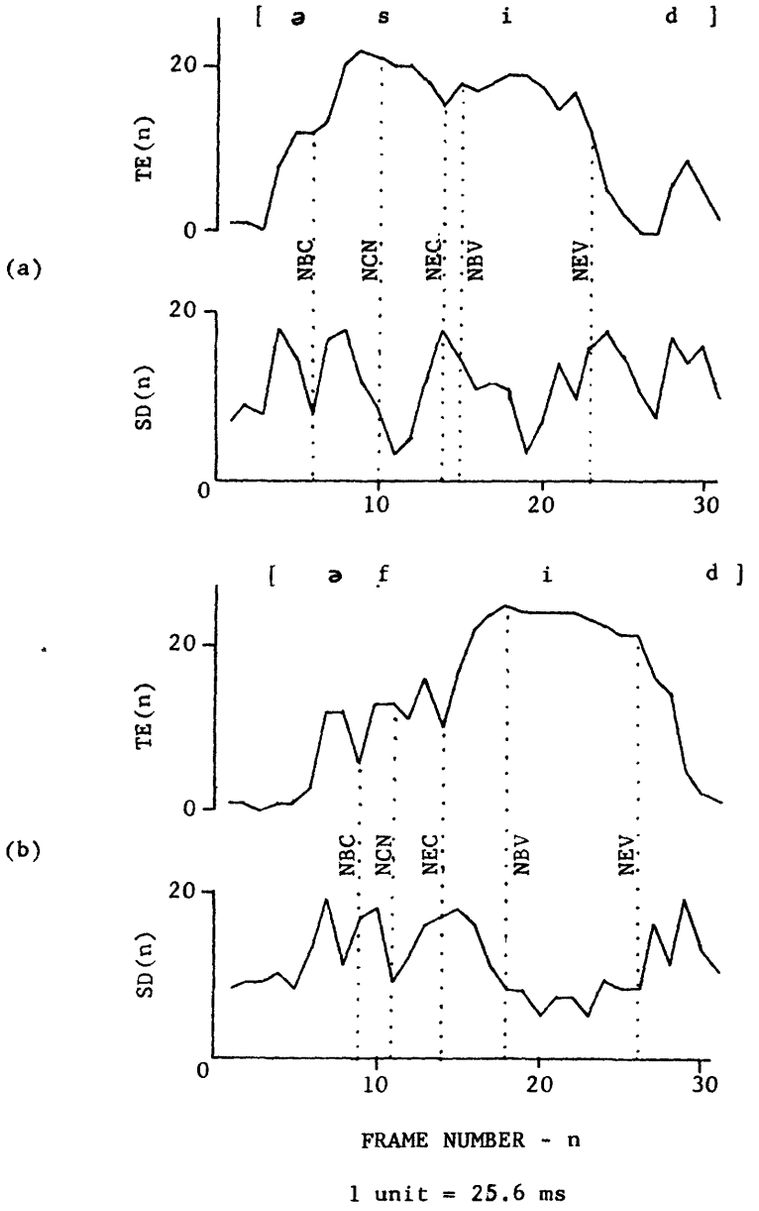


Fig. 20. Total energy and spectral derivative for the utterances [əsid] and [əfid].

For most fricatives NBC approximately coincides with the beginning of the consonant, while for the stops NBC usually points to some place toward the end of the stop gap and before the burst. So, instead of thinking of NBC as pointing to the beginning of the consonant it is perhaps more accurate to think of it as pointing to the point of minimum energy during the consonant portion of the utterance.

(c) Computation of NBV – (Beginning of vowel)

Let

$$\text{NMSD} = \min n \left[n \geq \text{NSCH} + 11, \text{ and SD}(n) \text{ is local min}\right], \tag{14}$$

and

$$\text{NXSD} = \max n \left[n < \text{NMSD}, \text{ and SD}(n) \text{ is local max} \geq 15\right]. \tag{15}$$

Then

$$\text{NBV} = \min n \left[\text{NXSD} + 1 \leq n \leq \text{NMSD}, \text{ and TE}(n) \text{ is local max}\right]. \tag{16}$$

Also,

$$\text{MAXV} = \text{TE}(\text{NBV}). \tag{17}$$

The beginning of the vowel is defined as the frame of maximum total energy after the initiation of voicing, which is usually accompanied by a large value of the spectral derivative. Equation 14 simply leads to a frame in the middle of the vowel region where the spectral derivative is at a minimum. The test $n \geq \text{NSCH} + 11$ indicates that the vowel must be at least 11 frames (≈280 ms) away from the schwa maximum. This is a restriction on the speaker, but is simple to by-pass simply by articulating the vowel a sufficient length (see Section VI). Equations 15 and 16 apply the definition above to compute NBV.

(d) Computation of NEC – (End of consonant)

$$\text{NEC} = \begin{cases} \text{NXSD} - 1, & \text{if SD(NXSD-1)} > 13 \\ \text{NXSD}, & \text{otherwise.} \end{cases} \tag{18}$$

$$\text{LCON} = \text{NEC} - \text{NBC}, \tag{19}$$

where NXSD is defined by Eq. 15. The end of the consonant is defined as the point in time before the initiation of voicing for the vowel. The computation of NEC approximates that, since a maximum of the spectral derivative occurs in that region. LCON is known as the "length" of the consonant.

(e) Computation of NEV – (End of vowel)

$$\text{NEV} = \min n \left[n > \text{NBV}, \text{ and MAXV} - \text{TE}(n) \geq \text{THR}\right] - 2 \tag{20}$$

$$NEV \leftarrow NEV - 1, \qquad \text{if } (NEV - NBV + 1) \text{ is odd} \tag{21}$$

$$NF = NEV - NBV + 1, \tag{22}$$

where

$$THR = \begin{cases} 15, & MAXV > 18 \text{ dB} \\ MAXV-4, & MAXV \leq 18 \text{ dB} \end{cases}$$

and NF is the number of frames in the vowel region starting with NBV and ending in NEV. The end of the vowel is simply located by a threshold THR on the total energy function. Equation 21 renders the number of frames in the vowel region, NF, an even number; the reason for this is given in the vowel-recognition procedure.

5.3 VOWEL RECOGNITION

A block diagram of the vowel recognition algorithm is shown in Fig. 21. Details of the algorithm are shown in the following figures and are briefly described below.

The frames that are used in vowel recognition are those included in the vowel region, which is bounded by frame numbers NBV and NEV as defined by Eqs. 16,

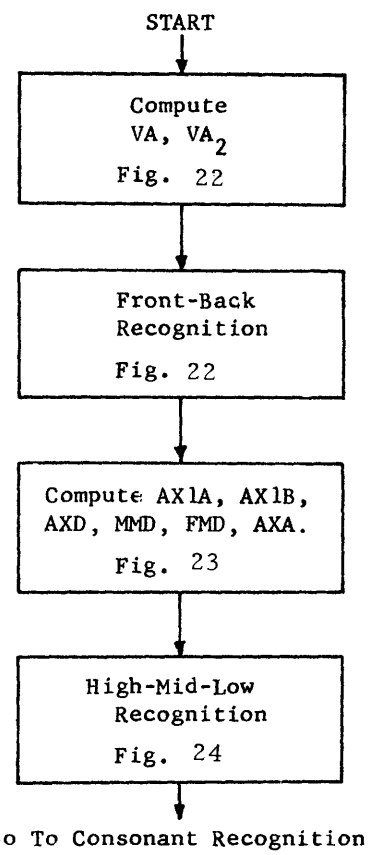


Fig. 21. Vowel recognition.

Let NZF = number of frames for all FB(m) $\neq 0$, m = 1, ..., NF, and NZF2 = number of frames for all FB(m) $\neq 0$, $m = \frac{NF}{2} + 1, \ldots, NF$.

Then

$$VA = \begin{cases} \left[\sum_{m=1}^{NF} FB(m)\right] * 100/NZF, & NZF > 1 \\ 200 \quad , & NZF \leq 1 \end{cases}$$

and

$$VA2 = \begin{cases} \left[\sum_{m=\frac{NF}{2}+1}^{NF} FB(m)\right] * 100/NZF2, & NZF2 > 0 \\ 200 \quad , & NZF2 = 0 \end{cases}$$

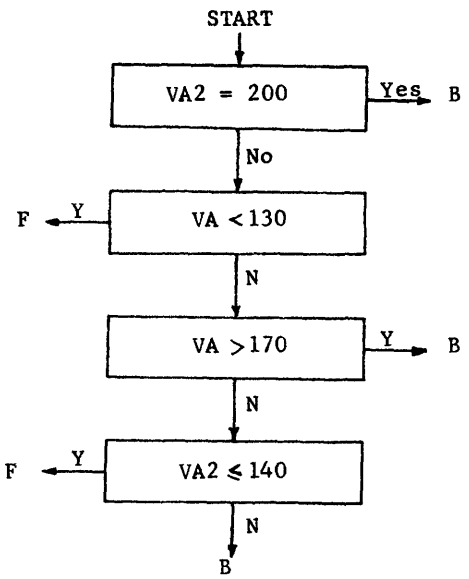


Fig. 22. Algorithm for front-back vowel recognition.

20 and 21. The vowel region is divided into two equal subregions, each having NF/2 frames, where NF is the total number of frames in the vowel region. (NF is even in order to render NF/2 an integer.) The recognition will depend on the properties of several parameters which are computed in each of the two subregions. A new frame number m is used, where

$$m = n - NBV + 1$$

so that the vowel region extends from m = 1 to m = NF.

5.3.1 Front-Back Recognition

The algorithm for the final front-back decision is shown in Fig. 22. FB(m) represents the F-B-ϕ decisions made in the vowel region (see section 5.1). This algorithm is one of the simplest and most successful in the whole recognition scheme. Some cases occur for which most or all of the values of FB(m) are zero, indicating ϕ (don't know) decisions. Such cases are taken as special cases, as shown in the formulas for VA and VA2 in Fig. 22, and are recognized as back vowels (a purely empirical decision, given that a binary decision had to be made). The values of FB(m) are either 0, 1 or 2 (see Fig. 14); from the equations in Fig. 22 we may conclude that VA and VA2 always obey the following inequality:

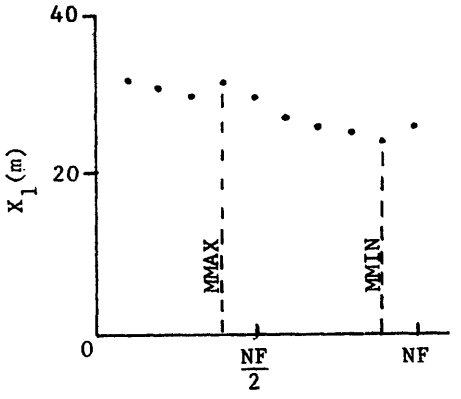
$$100 \leq VA, \qquad VA2 \leq 200.$$

Values closer to 100 signal a front vowel, while values closer to 200 indicate a back vowel. Note that the algorithm in Fig. 22 assigns special weight to the last part of the vowel. This was done purposely because, for any one utterance, the spectral changes are minimal and the spectrum is more indicative of the nature of the vowel in the second half of the vowel than in the first. This is especially important for back vowels following dental consonants.

5.3.2 High-Mid-Low Recognition

The different parameters used in the high-mid-low recognition are defined in Fig. 23. They are all derived from $X_1(m)$, the X coordinate of energoid 1, except for one minor parameter, AXA, which is derived from $X_A(m)$. MMAX and MMIN point to the frames at which $X_1(m)$ is maximum and minimum, respectively, in the vowel region. AX1A and AX1B are average values of $X_1(m)$ in the first half and second half of the vowel, respectively. (Figure 18 is a plot of $X_1(n)$ for the utterance [əgod]; the vowel region, which is used in high-mid-low recognition, is delineated by dashed lines corresponding to the frame numbers n = NBV and n = NEV, which correspond to m = 1 and m = NF, respectively.)

The recognition algorithm is shown in Fig. 24. The initial tests isolate the low



FRAME NUMBER – m

[m = n – NBV + 1]

1 unit = 25.6 ms

$$AX1A = \frac{2}{NF} \sum_{m=1}^{NF/2} X_1(m) \qquad AX1B = \frac{2}{NF} \sum_{m=\frac{NF}{2}+1}^{NF} X_1(m)$$

$$AXD = AX1A - AX1B$$

$$MMD = X_1(MMAX) - X_1(MMIN)$$

$$FMD = X_1(NF) - X_1(MMIN)$$

$$AXA = \frac{1}{NF} \sum_{m=1}^{NF} X_A(m)$$

Fig. 23. Computation of parameters for high-mid-low vowel recognition.

vowel [a] mainly by recognizing that the value of AX1B is large. (In this discussion it is helpful to remember that X_1 is a good approximation of the first formant F1.) The rest of the vowel-recognition algorithm tries to separate high and mid vowels. The scheme is based on the fact that mid vowels have a higher F1, and hence a higher X_1, which exhibits greater change in time than the corresponding F1 for the high vowels. The tests on AX1A depend on whether the vowel was recognized as front or back. The test in the dashed box was eliminated in the final version of the algorithm, as will be seen in section 6.5. The rest of the algorithm simply tests the change of X_1 throughout the vowel region.

5.4 CONSONANT RECOGNITION

A block diagram of the consonant recognition procedure is shown in Fig. 25. The computation of schwa parameters, the "consonant frame" location, and the corresponding FFT computation are all independent of the consonant to be recognized.

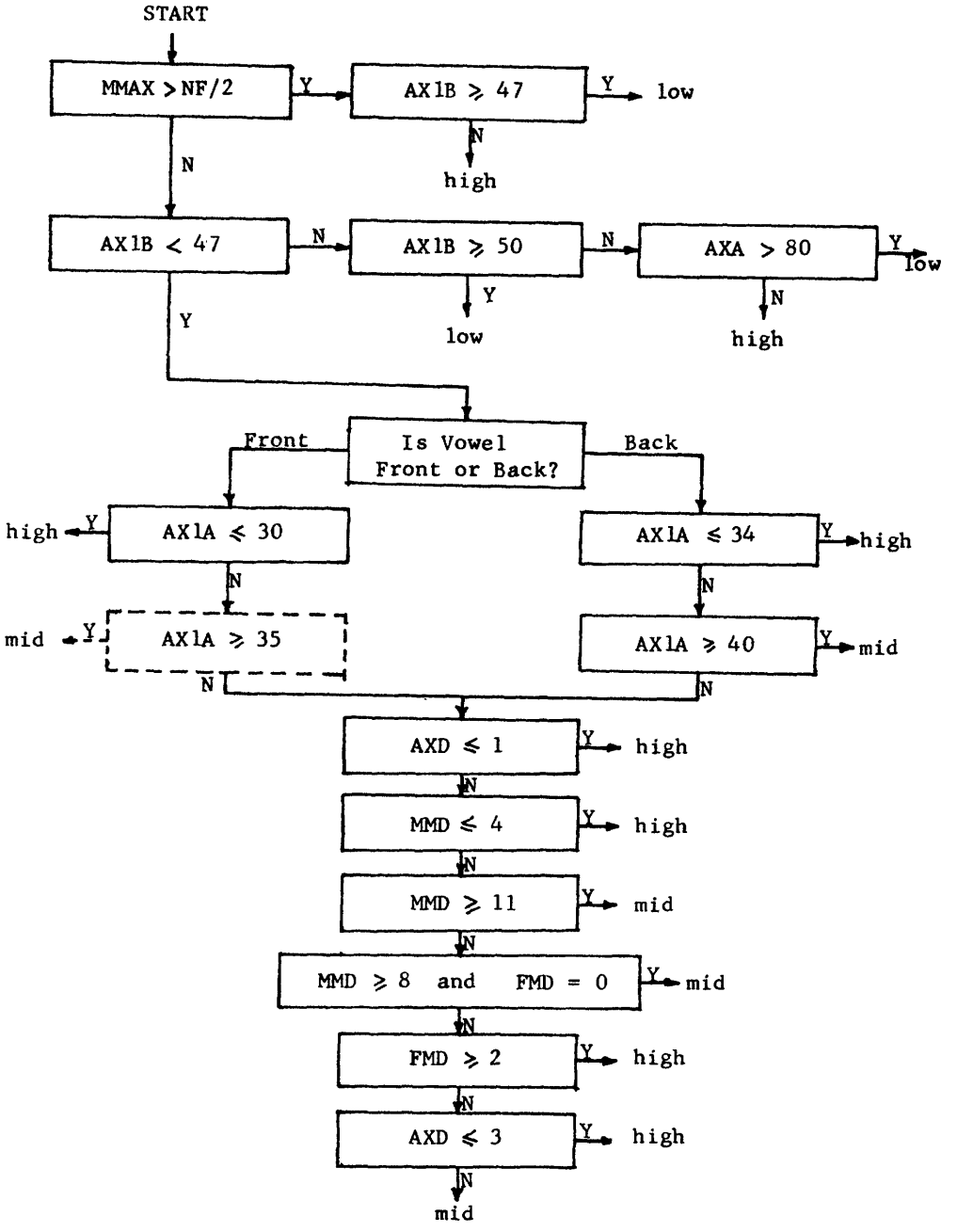


Fig. 24. Algorithm for high-mid-low vowel recognition.

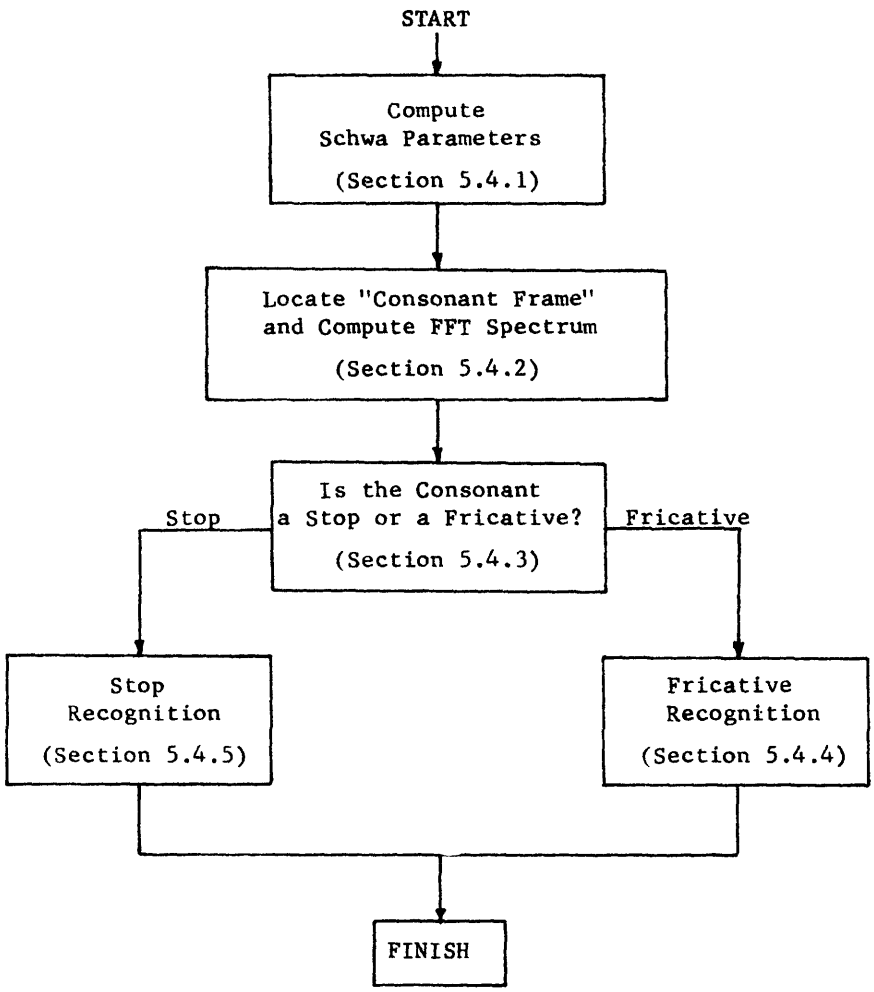


Fig. 25. Consonant recognition procedure.

Consonant-dependent parameters are computed in corresponding portions of the recognition scheme. Fricative recognition depends mainly on the FFT-derived spectrum, which is computed once during the fricative portion. Stop recognition depends mainly on the burst spectrum and, to a lesser extent, on the [CV] transition.

5.4.1 Schwa Parameters Computation

As mentioned in section 4.2.4, the [əC] transition carries information about whether the consonant is labial or not. In particular, F2 decreases with labials and increases with nonlabials. The problem is to track F2 throughout the schwa vowel. F2 can be approximated by X_S, the X coordinate of an energoid S whose range is defined in the neighborhood of F2 and is dependent on each specific utterance. The definition of the range of energoid S is given below.

The filter-bank spectrum is obtained for frame number NSCH, which points to the schwa maximum as defined by Eq. 12. The range of energoid S, starting with filter

number IIN and ending in filter number IFIN, is computed from the spectral frame NSCH as follows.

Let

$$IMAX = i\ [E_i(NSCH) \text{ is max},\ 10 \leq i \leq 19].$$

Then

$$IFIN = i\ [E_i(NSCH) \text{ is min},\ IMAX \leq i \leq 22], \tag{23}$$

and

$$IIN = \max i\ [i < IMAX \text{ and } E_i(NSCH) \text{ is local min}],$$

where i is the filter number, and what is included between brackets is the condition that must be satisfied in each case.

X_S is computed by using Eq. 5 with a = IIN and b = IFIN. Because the duration of the schwa in the context [əCVd] is usually short, it is not enough to compute X_S every 25.6 ms. So, X_S is computed at points in time corresponding to the frame numbers $n \leq NBC$ plus values computed half way in between. That is, a new frame number p is defined, where p = 2n, so that two p frames are separated by 12.8 ms.

Another fact of importance is that at points of low total energy, values of X_S are meaningless. Therefore, the only values of X_S that have been considered important are those that are computed at points where $TE(p) \geq [TE(2*NSCH)-6\ dB]$, where the arguments in parentheses are given in terms of p. The schwa parameters are computed as follows.

Let

$$PMAX = p\ [TE(p) \text{ is max},\ p \leq 2*NBC \text{ and } TE(p) \geq TE(2*NSCH) - 6]$$

and

$$PMIN = p\ [TE(p) \text{ is min},\ p \leq 2*NBC \text{ and } TE(p) \geq TE(2*NSCH) - 6],$$

where p is the new frame number such that p = 2n.

Then

$$SMMD = \begin{cases} X_S(PMAX) - X_S(PMIN), & PMAX \geq PMIN \\ X_S(PMIN) - X_S(PMAX), & PMIN > PMAX. \end{cases} \tag{24}$$

Let

$$PIN = \min p\ [TE(p) \geq TE(2*NSCH) - 6].$$

Then (25)

$$INMD = X_S(PIN) - X_S(PMIN).$$

Both SMMD and INMD will be used to isolate labials. A labial is usually characterized by a negative SMMD.

5.4.2 "Consonant Frame" Location and FFT Computation

For the detection of voicing it was necessary to employ the FFT; the filter bank did not give the low-energy peak necessary for such detection. Another reason for using the FFT was its superior filtering characteristics as compared with the filter bank (see Appendix B). This is particularly important for the recognition of fricatives. The one important drawback, however, was the excessive time needed for the computation of the FFT (more than 1 sec, including multiplying by a window, transforming, and taking logarithms). Since the system was supposed to operate as close to real time as possible, I decided to compute only one FFT for the recognition of a single consonant. This means that the frame at which the FFT is computed must be judiciously chosen. This will be called the "consonant frame" location, and NCN will point to that location.

A logical place to compute the single FFT is somewhere during the "steady-state" portion of the consonant. This would certainly be very appropriate for fricatives. For stops, however, it would be useful only in the recognition of voicing, and further analysis would be necessary to determine the place of articulation. A simple way to determine the steady-state portion is to look for places where the spectral derivative is minimum. An important fact to be kept in mind is that the amplitude of voicing decreases with time during a consonant (this is especially true for the voiced stops), and therefore one must compute the FFT in the earlier portion of the consonant. This would certainly be very appropriate for fricatives. For stops it would be useful only in the recognition of voicing, and further analysis would be necessary to determine the place of articulation. A simple way to determine the steady-state portion is to look for places where the spectral derivative is minimum. An important thing to be kept in mind is that the amplitude of voicing decreases with time during a consonant (this is especially true for the voiced stops), and therefore one must compute the FFT in the earlier portion of the consonant to ensure a strong low-energy peak in the FFT computed spectrum. The compromise reached is shown by the following equations.

Let

$$NM = n\,[SD(n) \text{ is min, } NSCH + 2 \leq n \leq NEC].$$

Then

$$NCN = \begin{cases} NM-1, & \text{if } SD(NM-1) \leq 10 \text{ or } [SD(NM-1) - SD(NM) < 6 \text{ and } SD(NM-1) < 17] \\ NM, & \text{otherwise} \end{cases} \tag{26}$$

where NM is the frame at which SD(n) is minimum.

Vertical lines showing the location of NCN for some utterances are shown in Figs. 18-20. A 512-point FFT, with a Hamming window weighting, is computed for 25.6 ms of the time waveform starting and ending at the times corresponding to frame

numbers NCN-1 and NCN, respectively. A grouping of energies is performed on the linearly spaced spectral energy values, so that the frequency scale is the same as that for the filter bank (see Appendix B). Also, all the energoids listed in Fig. 8 are computed. As an example, see the CRT display shown in Fig. 7 for the utterance [əzid]. The 25.6 ms portion of the time waveform corresponding to [z] that is used for the FFT computation is displayed at the bottom. The corresponding FFT-computed spectrum is shown at the top of the figure with the + and * energoids. The frame number NCN is indicated below the plot of SD(t).

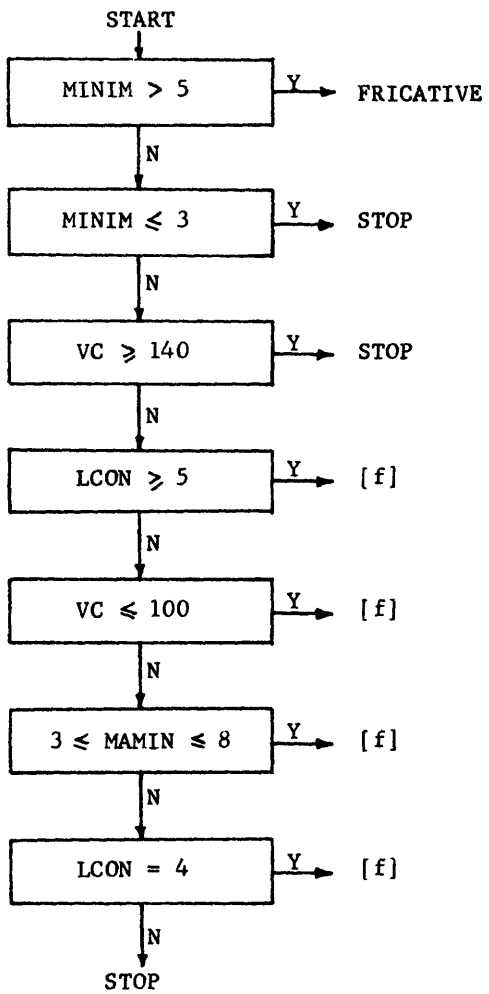


Fig. 26. Flow chart for stop-fricative recognition.

The single parameter used for the recognition of voicing is computed as follows.

$$VC = \left[100 \log_{10} \sum_{i=1}^{3} E_i\right] - Y_2, \qquad (27)$$

where it is understood that the spectrum under consideration is that computed by the FFT, E_i is the energy at the frequency corresponding to filter number i, and Y_2 is the Y coordinate of energoid 2.

Other parameters derived from the FFT spectrum for fricative recognition will appear in section 5.4.4.

5.4.3 Stop-Fricative Recognition

The flow chart describing how stops and fricatives are distinguished from one another is shown in Fig. 26. The major differentiating parameter is MINIM, which is usually the minimum total energy (in dB) along the consonant.

It is defined as

$$MINIM = TE(NBC), \qquad (28)$$

where TE(NBC) is the total energy at frame number NBC which is defined by Eq. 13. MINIM is measured in decibels.

The only fricative allowed to have $3 < MINIM \leq 5$ is [f], which is an unvoiced fricative. So, except for the first two tests on MINIM, the rest of the flow chart tries to isolate a possible [f]. The remaining undefined parameter is MAMIN, which is defined as follows.

Let

$$NCM = \min n \,[TE(n) \text{ is local max, and } NBC \leq n \leq NEC]. \tag{29}$$

Then

$$MAMIN = \begin{cases} TE(NCM) - MINIM, & \text{if } NCN \neq NEC \\ 0 \quad , & \text{if } NCM = NEC \end{cases} \tag{30}$$

It is important to note that the limits on MINIM are absolute values in dB, and it is the speaker's responsibility to articulate his consonants appropriately. (This will be discussed further in Section VI.)

5.4.4 Fricative Recognition

The flow chart for fricative recognition is shown in Figs. 27-30. The only spectrum that is used in computing the different parameters is that computed using the FFT, as

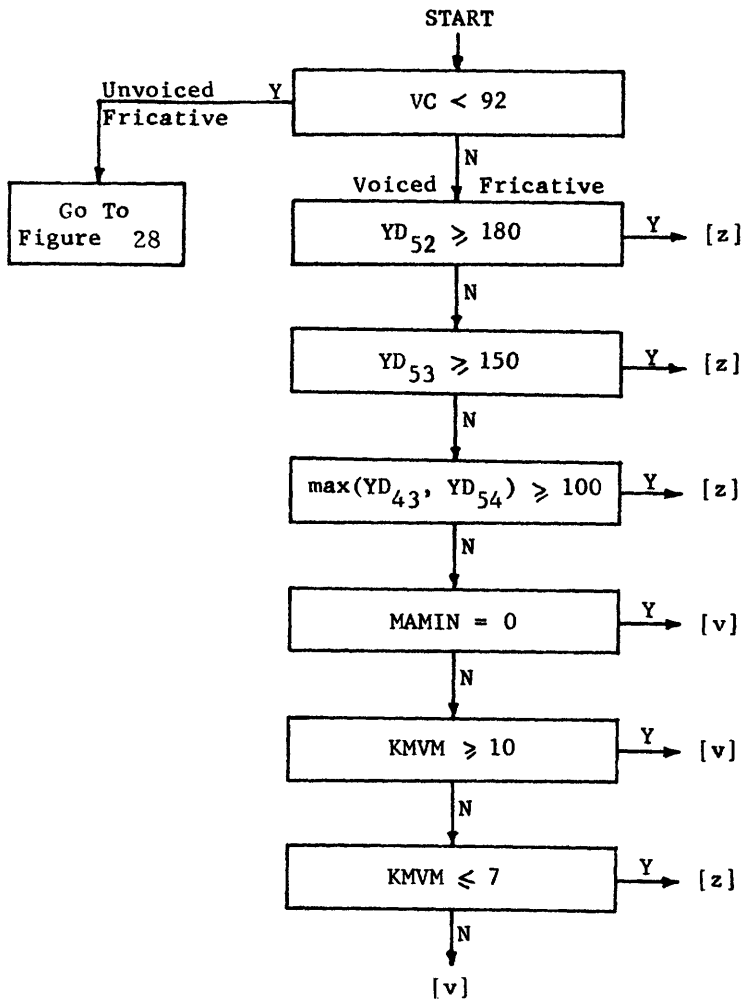


Fig. 27. Flow chart for fricative recognition. Recognition of voiced fricatives.

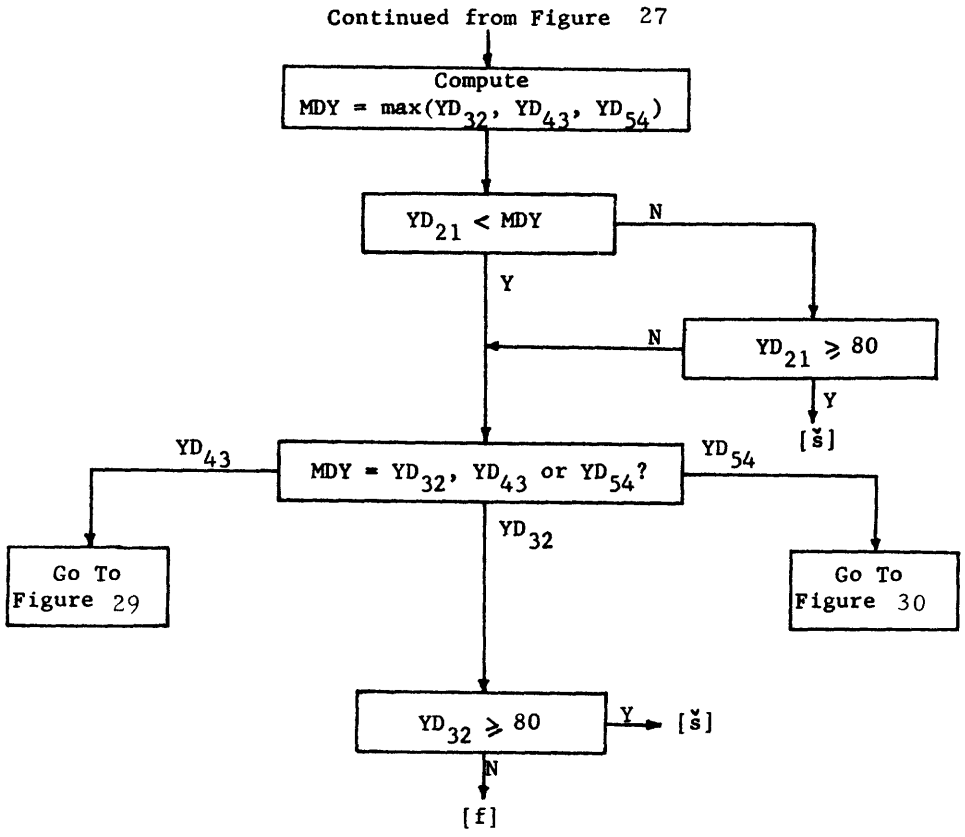


Fig. 28. Flow chart for unvoiced fricative recognition.

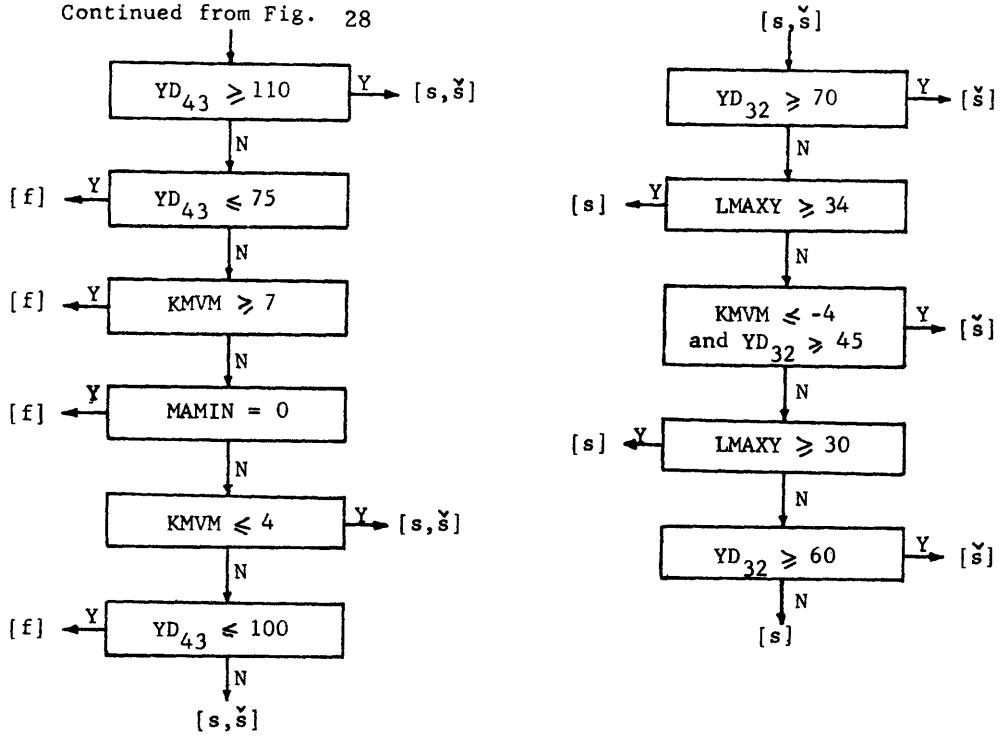


Fig. 29. Continuation of flow chart for unvoiced fricative recognition.

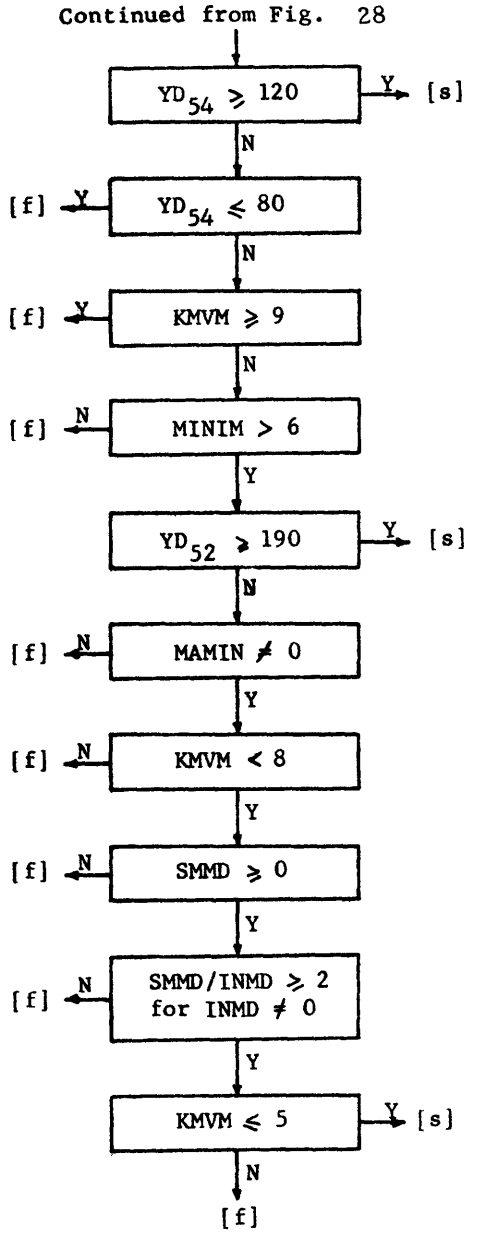


Fig. 30. Continuation of flow chart for unvoiced fricative recognition.

explained in section 5.4.2. There is one parameter, KMVM, that has not been defined yet. KMVM is a measure of the relative total energy between the consonant and the following vowel, and is defined as

$$KMVM = TE(NBV) - TE(NCM), \tag{31}$$

where NBV is defined by Eq. 16 and NCM by Eq. 29.

Only one test is used to determine whether the fricative is voiced or unvoiced, as shown in Fig. 27. Consequences of this single test as related to recognition performance are discussed in Section VI.

The different factors that were used in place-of-articulation recognition are the location of the major energy concentration in the FFT-computed spectrum, the total energy level of the consonant, the relative energy between the vowel and the consonant, and the [əC] transition. (Refer to section 4.2.2 to see how these factors relate to the place of articulation.) The [CV] transition was not used in fricative recognition.

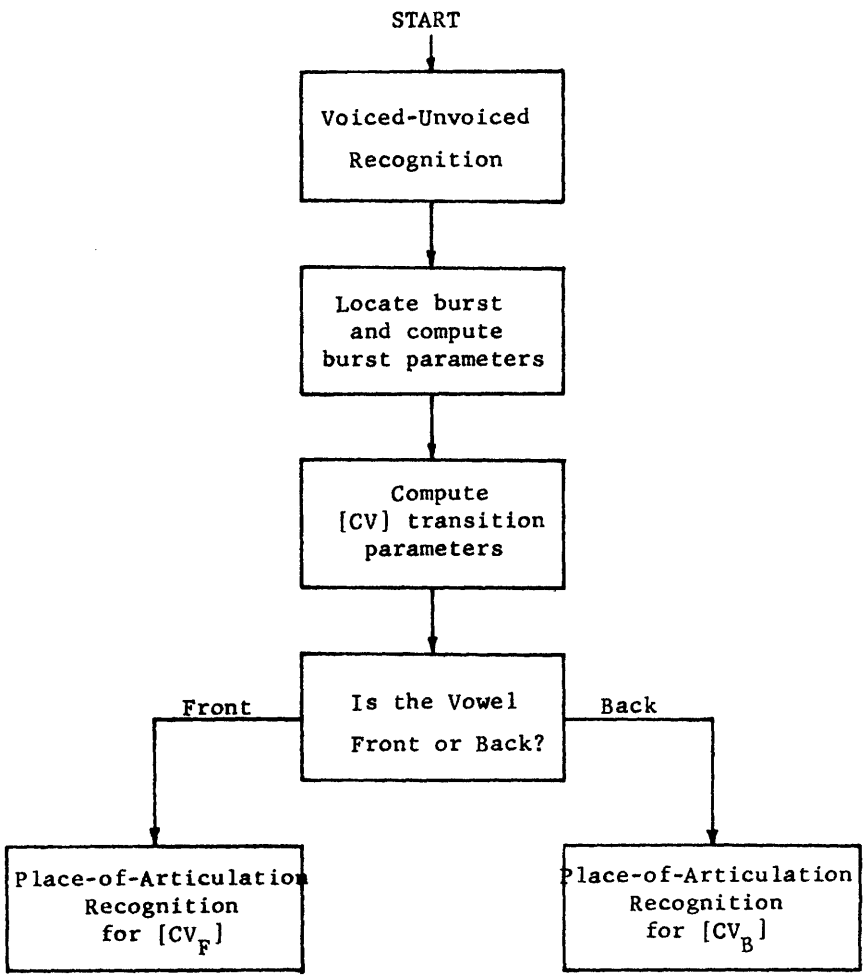


Fig. 31. Stop recognition.

5.4.5 Stop Recognition

A block diagram for stop recognition is shown in Fig. 31. Voiced-unvoiced recognition is determined from the FFT-computed spectrum and the approximate burst length. The burst is then accurately located and the burst spectrum is used as a major source of information for place recognition. The recognition algorithm will sometimes depend upon whether the stop is voiced or unvoiced and whether the following vowel is front or back. Also, the [CV] transition is employed for place recognition of stops preceding back vowels. (Refer to section 4.2.2 to see how the burst spectrum and the [CV] transition relate to the place of articulation.)

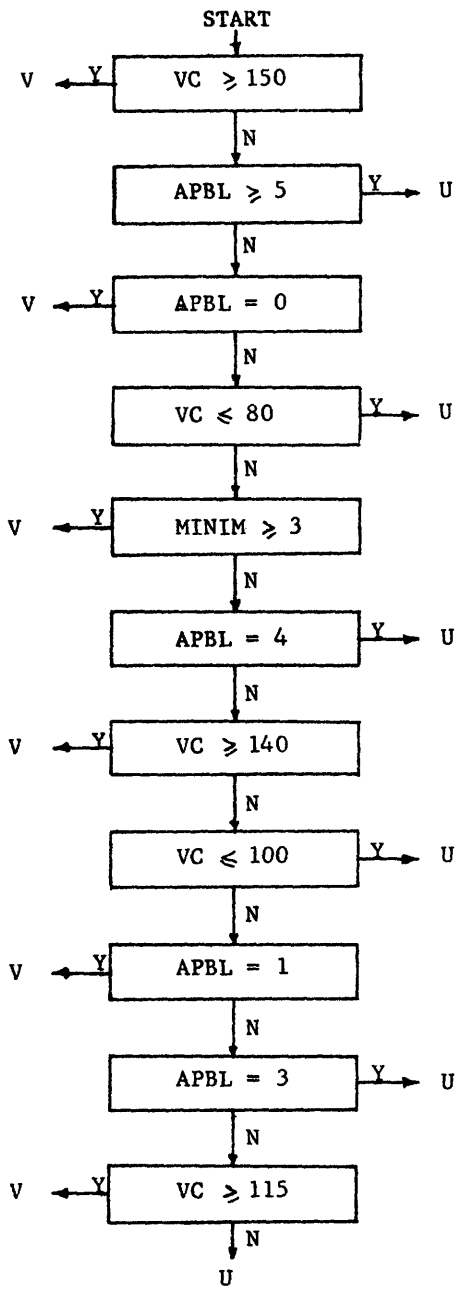


Fig. 32. Flow chart for the recognition of voicing with stop consonants. V = Voiced, U = Unvoiced.

(a) Voiced-Unvoiced Recognition

The flow chart for the recognition of voicing is shown in Fig. 32. APBL, which stands for approximate burst length, is computed as follows.

$$\text{APBL} = \begin{cases} \text{NEC} - \text{NM}, & \text{if result} \leq \text{LCON} \\ \text{LCON}, & \text{otherwise} \end{cases} \qquad (32)$$

where NM is defined by Eq. 26. APBL is supposed to give a rough indication of whether there is a large period of aspiration or not.

Note that VC is computed by Eq. 27 from the FFT-computed spectrum as explained in section 5.4.2.

(b) Burst Location and Burst Parameters

The location of the burst is accomplished by detecting a sudden rise in total energy TE(t) starting from frame number NBC, the "beginning" of the consonant. In order to accurately locate the burst, one must compute TE(t) at steps much smaller than 25.6 ms, since some burst lengths are of the order of 6 ms. It was decided to compute TE(t) at a spacing of T = 1.6 ms, starting with TBC, the time corresponding to frame number NBC. Let the new frame numbers be denoted q,

where q = 0 corresponds to t = TBC; that is, TE(n = NBC) = TE(q = 0). Then the location of the beginning of the burst in terms of q is computed in a manner similar to the following.

Let

$$QMIN = q\ [TE(q) \text{ is min},\ 0 \leq q \leq 16 * LCON].$$

Then

$$QBBR = \min q\ [TE(q) - TE(QMIN) \geq 3] - 1, \tag{33}$$

where QBBR points to the beginning of the burst within 1.6 ms. The algorithm that is actually used in computing QMIN is much more difficult to describe than that given above, but the end result is identical. The actual algorithm results in less execution time as compared with the definition given in (33).

The burst spectrum is obtained through the filter bank at t = TBRST, defined as

$$TBRST = \begin{cases} TBBR + 6.4 \text{ ms}, & \text{when [C] is voiced} \\ TBBR + 12.8 \text{ ms}, & \text{when [C] is unvoiced.} \end{cases} \tag{34}$$

All energoid coordinates are computed, and so is LMAXY.

A parameter that will be useful in differentiating labials (b, p) from palatals (g, k) is BG, defined as

$$BG = \begin{cases} (YD_{21},\ \text{iff} > 0) + (YD_{32},\ \text{iff} > 0) \\ 0, \qquad \text{otherwise} \end{cases} \tag{35}$$

Two parameters, PAL and DEN, will be used in velar vs dental recognition with back vowels only. They are computed as follows.

Let

$$DX_{mn} = \begin{cases} 20 - XD_{mn}, & \text{if result} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where mn = 2A, B2, 3B, C3, 4C, D4, 5D.

Then

$$PAL = DX_{2A} + 2*DX_{B2} + 3*DX_{3B} + 4*DX_{C3}$$

$$DEN = 3*DX_{4C} + 2*DX_{D4} + DX_{5D}. \tag{36}$$

Other parameters derived from the burst spectrum have already been defined elsewhere.

(c) [CV] Transition

The program was capable of tracking all energoids throughout the transition. Two energoids were found most useful: X_A and X_B, the X coordinates of energoids A and B whose ranges are 150-1550 Hz and 1050-2200 Hz. Starting at the beginning of the burst (t = TBBR) and continuing into the vowel, the spectrum was obtained from the filter bank at fixed intervals of time T, where

$$T = \begin{cases} 6.4 \text{ ms}, & \text{for [C] voiced,} \\ 12.8 \text{ ms}, & \text{for [C] unvoiced.} \end{cases} \tag{37}$$

$X_A(t)$ and $X_B(t)$ were computed for each of those frames.

The first parameter computed is LBRST which is a much better estimate of the burst length than APBL. The computation of LBRST depends on the fact that when the voicing of the vowel commences, a sudden increase in low-frequency energy causes a sharp drop in X_A. This sharp drop occurs in many of the utterances considered, but certainly not all. Therefore, the use of LBRST is used in the recognition scheme only as a secondary parameter. The value of LBRST is defined as

$$\text{LBRST} = \text{number of frames between TBBR and the sharp decrease in } X_A(t). \tag{38}$$

No detailed flow chart will be given for the computation of LBRST. Note that each unit of LBRST represents T ms as defined by Eq. 37.

The two other parameters are derived from $X_B(t)$. They are

$$\text{MNXB} = \max X_B(t) - \min X_B(t) \tag{39}$$

and

$$\text{MSXB} = \max [X_B(t) - X_B(t+T)], \tag{40}$$

where TBRST $\leq$ t < TBRST + 192 ms, and T is defined by Eq. 37. A display of $X_A(t)$ and $X_B(t)$ for the utterance [əgod] starting at the [g] burst is shown in Fig. 10. Note the small vertical line which indicates the end of the burst.

There is still another parameter that compares the spectral derivative at t = TBRST with that T ms later. It is

$$\text{SD12} = \text{SD(TBRST)} - \text{SD(TBRST} + \text{T}), \tag{41}$$

where T is defined by (37). SD12 is used in separating labials from nonlabials; it is greater for nonlabials than for labials.

(d) Place-of-Articulation Recognition

Place recognition for stops depends on whether the following vowel is front or back.

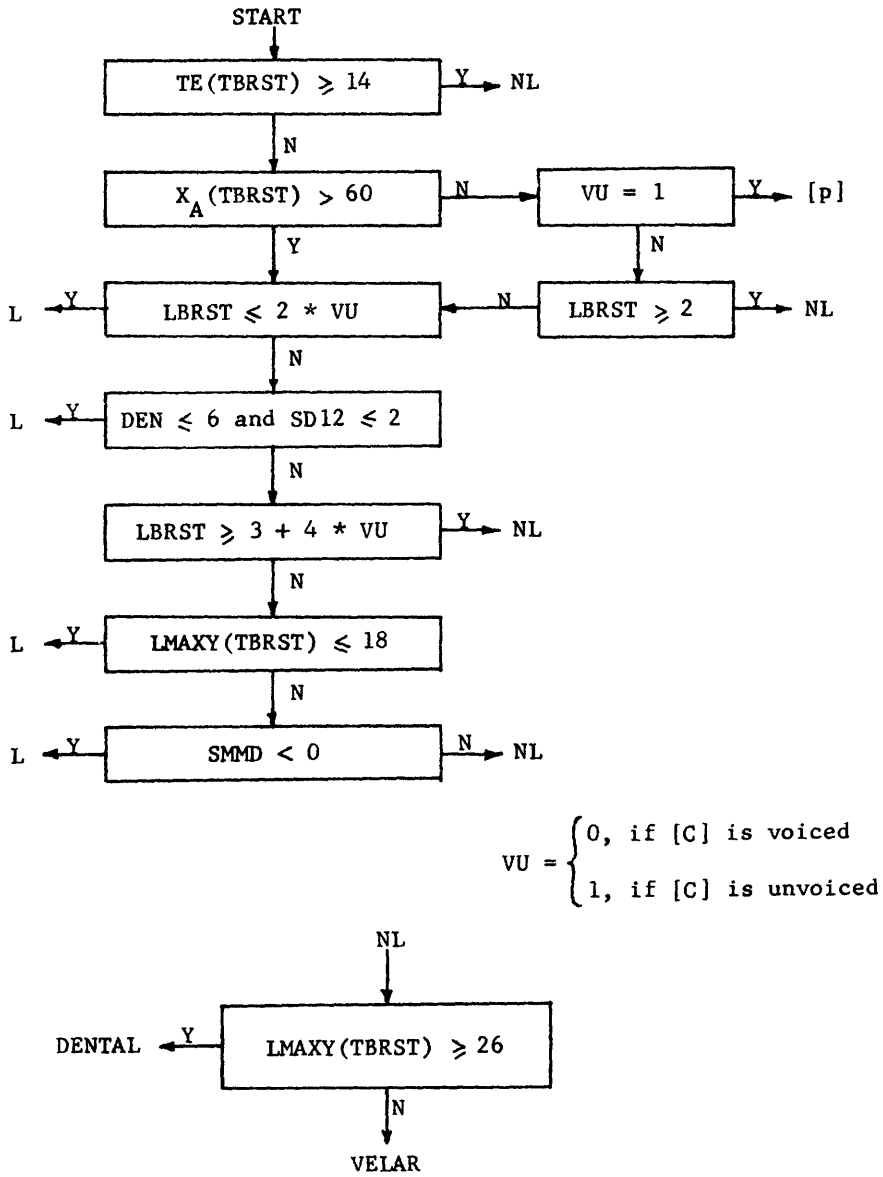


Fig. 33. Flow chart for place-of-articulation recognition of stops before front vowels. L = Labial, NL = Non-labial.

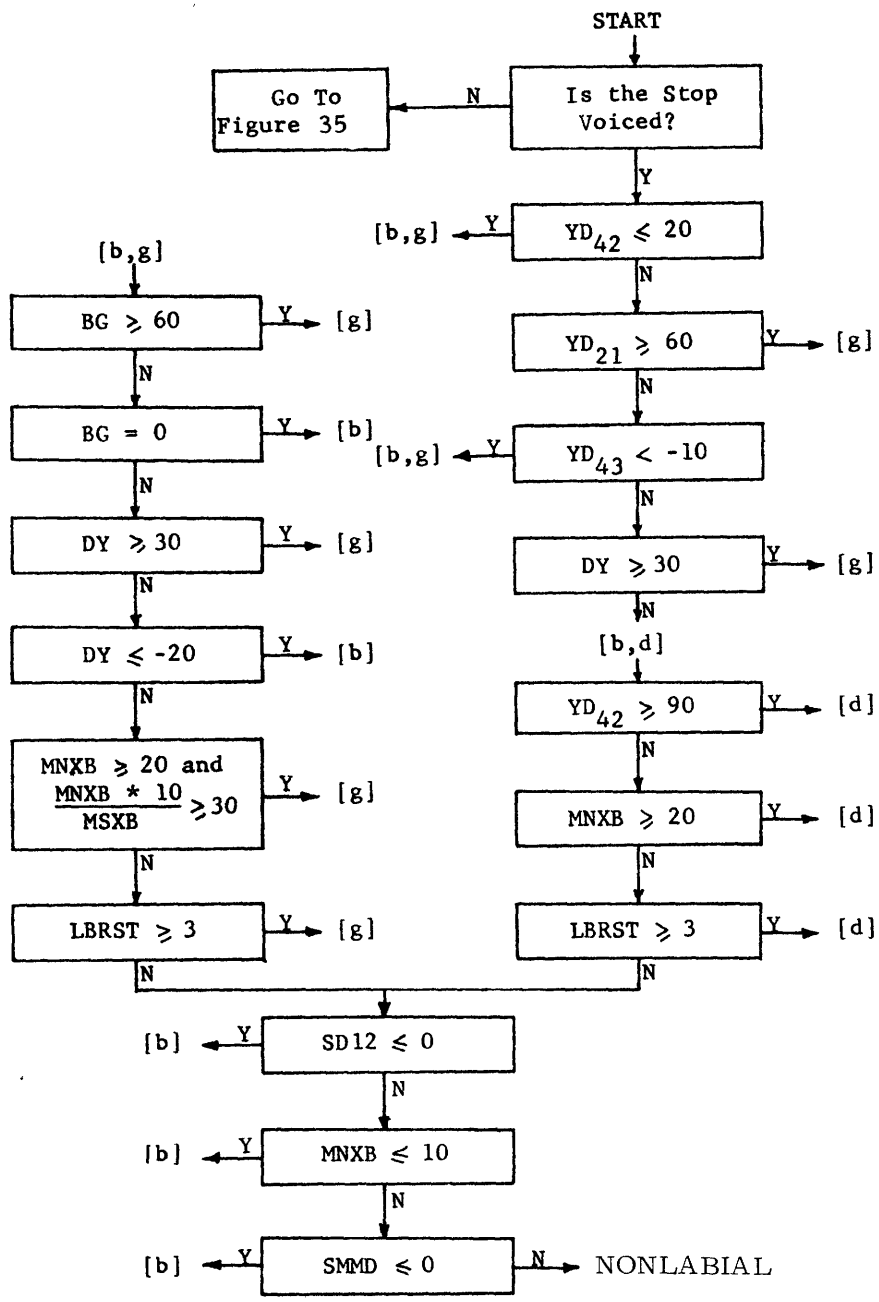


Fig. 34. Flow chart for place-of-articulation recognition of voiced stops before back vowels. DY = $YD_{21} - YD_{43}$.

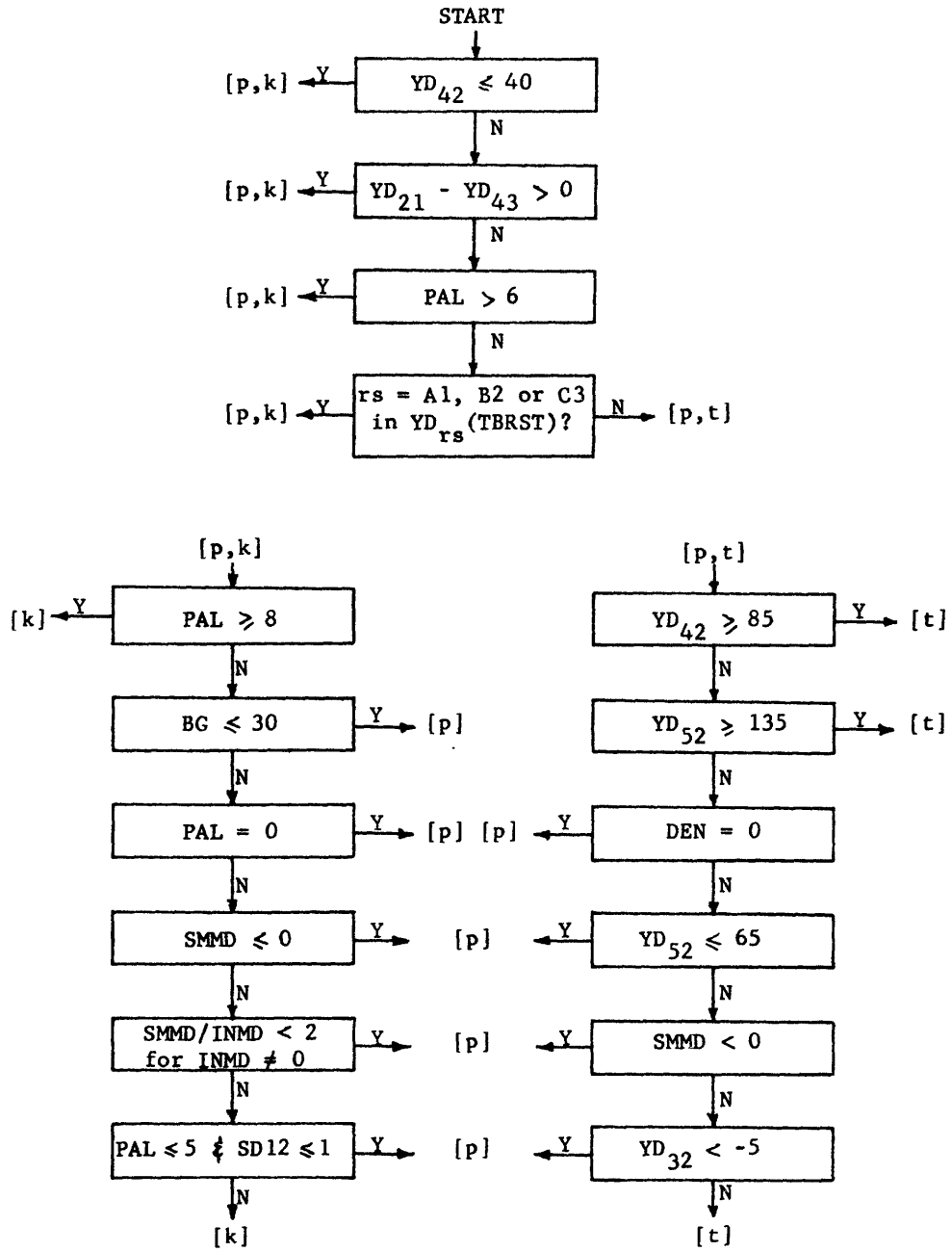


Fig. 35. Flow chart for place-of-articulation recognition of unvoiced stops before back vowels.

The flow chart for stop recognition before front vowels is shown in Fig. 33. The same algorithm is useful for voiced, as well as unvoiced stops. Note that there is only one test that separates dentals from velars, namely the position of the major energy peak in terms of the corresponding filter number. The fact that only one test exists to separate dentals from valars before front vowels is an important factor in explaining the corresponding recognition results in Section VI.

The recognition of stops before back vowels depends on whether the stop is voiced or unvoiced, as shown in Figs. 34 and 35. The tests in both cases are extensive compared with those in Fig. 33.

VI. EXPERIMENTATION AND RESULTS

6.1 EXPERIMENTAL ARRANGEMENT

Photographs of the experimental arrangement are shown in Fig. 36. The subject and the experimenter sit next to each other with full view of the CRT display which is visible in Fig. 36b. The position of the dynamic, directional microphone is controlled by the subject by means of a flexible arm attached to it. The subject usually speaks at a distance of only a few inches from the microphone, because of the noise in the computer room. Program control through knobs, switches, and push buttons is performed solely by the experimenter. The knobs and switches are on the panel opposite the experimenter. The knobs and switches are on the panel opposite the experimenter in Fig. 36b. The push-button box, shown in Fig. 36a with the experimenter's hand on it, is also accessible to the subject. By pressing one of the push-buttons, the subject is able to play back the sampled utterance as recorded in the computer memory.

An example of what the subject sees after the recognition of an utterance is shown in Fig. 7. The only portion of the display that is of concern to the subject is the recognized [CV] syllable, [zi] for the utterance [əzid] in the case of Fig. 7. The rest of the display is of concern only to the experimenter. Most of the subjects were M. I. T. students and their curiosity was satisfied only after it was explained to them what the different plots on the display meant. A few were actually able to tell what went wrong when certain errors in recognition occurred.

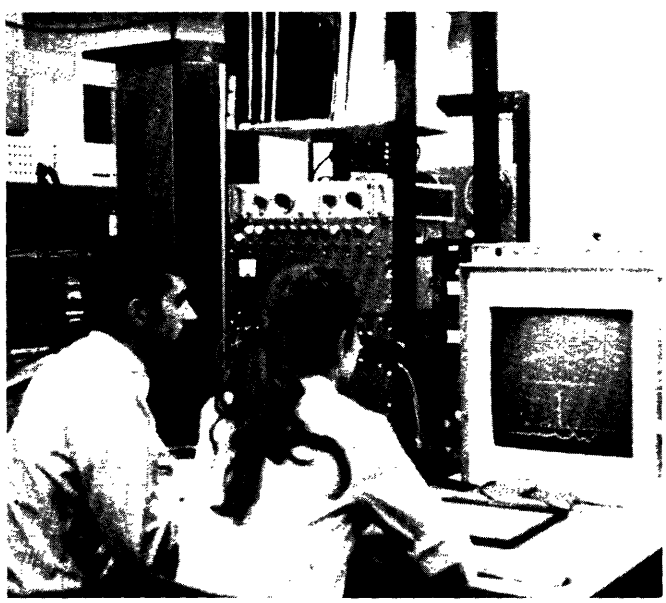
The subjects received all of their instructions orally from the experimenter. Different methods of changes in articulation were also usually suggested by the experimenter. In a practical system, the experimenter might be replaced by a training program, so that the interaction would take place wholly with the machine. It must be kept in mind, however, that the recognition system developed in this research is experimental, and that this experiment should be looked upon as a pilot study in which the experimenter's subjective participation is essential.

6.2 EXPERIMENT

The recognition scheme was tested on 12 speakers: 6 males and 6 females ranging from 17 to 30 years of age. Three of the speakers had been used in the analysis that led to this particular recognition scheme. The other nine speakers were the first to respond to an advertisement requesting subjects. No screening of subjects took place, although three of them had definite peculiar speech characteristics (which might have been more crucial in an experiment with continuous speech recognition). As a result of the randomness of speaker selection, the speakers had different accents: they came from Massachusetts, New York, Maryland, Ohio, Chicago, California, Montreal (Canada), and one had French and Arabic as his first and second languages. At no time did I (the experimenter) participate as a subject.

The randomized list of 55 [əCVd] utterances was written in phonetic form. It

(a)



(b)

Fig. 36. Experimental arrangement.

took only a few minutes to explain to the subject how to pronounce the phonetic forms. The vowels [i], [e], [a], [o], [u], the schwa [ə], and [$\overset{\vee}{s}$] were the only unfamiliar phonetic symbols to them. They were told to stress the vowel [V] and pronounce the schwa unstressed. For some it was necessary to give examples from written English, e.g., [əbid] is pronounced as a bead. I emphasized to them that it was very important to pronounce the utterances as naturally as possible. Most of them were able to overcome the initial fear quickly, actually a double fear: fear of speaking into a microphone, and a fear of the "machine" that was about to recognize what they were saying! The latter fear was quite real in the beginning: Whenever the machine incorrectly recognized what they said, the immediate reaction was, "What did I do wrong?," and I had to assure them that they were speaking very intelligible sounds and "proved" it to them by playing it back by means of the auditory feedback option. That phase of initial fear was overcome quickly by all except one male speaker who was always self-conscious.

We shall use the term a run to mean a single reading of the 55 [əCVd] syllables. In a run, each utterance is processed by the program and the recognized [CV] syllable is displayed. Those syllables that are incorrectly recognized are recorded by the experimenter. Of course, the subject is also in full view of the CRT, which displays the recognized syllable. The only recognition errors that were not recorded were those caused by incorrect segmentation. This was done because it was not our purpose to concern ourselves in any serious manner with the problem of segmentation. Therefore, only those utterances that were segmented correctly (a condition which was easily detected from a display such as that in Fig. 37) were considered for recognition. Sometimes the syllable was correctly recognized with incorrect segmentation; that was not accepted either. In each case of incorrect segmentation, the subject was instructed to repeat the utterance. Sometimes it was necessary to give hints to the subject to effect correct recognition. During a run, no instructions were given to the subject other than those concerned with correct segmentation. In general, the reading of the list ran smoothly with a few repetitions here and there.

The experiment was conducted in two runs, one subject at a time, with each run followed by a learning session. Ideally, one would have a large number of runs, each followed by a learning session, and the experiment would end when recognition rates assumed an asymptotic value. This would have been quite impractical, however, with 12 speakers and a very limited available computer time. So, instead, only two runs and two learning sessions were performed with the second learning session having a formal format that will be described later. Because the second learning session was run differently from the first, we shall call the first run and the first learning session Experiment 1; similarly, the second run and the second learning session will be called collectively Experiment 2.

Experiment 1 was in two sessions separated by a week. The first session started by familiarizing the subject with the system and instructing him or her how to

pronounce the different utterances. The session ended after going through one run and recording the incorrectly recognized utterances. The session took anywhere from 40 minutes to 1 hour, depending on the subject. The second session was completely spent as a learning session. The session was very informal, the intention being that several methods of changes in articulation would be developed in an atmosphere of a minimum of psychological pressures. The subject was instructed to change his or her articulation in such a way as to effect correct recognition for those words that had been incorrectly recognized in the first run. Other utterances were also tried out upon the discretion of the experimenter. Most of the methods used in the changing of articulation were suggested by myself, since I had a knowledge of which part of the algorithm had caused the incorrect recognition. The methods used will be described in more detail.

Experiment 2 was completed in a 1-hour session, usually 1 week after the first learning session. At the beginning of the session the subject was briefly reminded of the methods of changes in articulation that had been useful in the first learning session. That was followed by a run through the 55 utterances, and the incorrectly recognized syllables were recorded in an error list. The utterances in the error list were then each repeated twice by the subject. Those utterances that were correctly recognized twice in a row were eliminated from the error list, thereby resulting in a reduced error list. Then the subject underwent a formal learning session with each of the utterances in the reduced error list, and the proceedings were recorded. The subject would repeat each utterance as many times as necessary to reveal to the experimenter either that the "correct" articulation had been achieved or that it was a hopeless case. (The criteria for these decisions will be discussed in section 6.4.) Between repetitions of each utterance I felt free to give the subject any instructions that I felt would effect the desired recognition. The subject was also encouraged to try things on his own.

The reasons for the specific format described for the second experiment will become evident in section 6.5, which also contains the recognition results of Experiments 1 and 2.

6.3 METHODS FOR CHANGES IN ARTICULATION

Most of the methods for changes in articulation that were used by subjects stemmed directly from the specific recognition algorithm that is described in Section V. The experimenter's intimate knowledge of the recognition algorithm put him in an ideal position to suggest those methods. Whenever an utterance was incorrectly recognized, I was able to find out very quickly (by checking parameter values that were displayed at the push of a button) which part of the algorithm was responsible for the incorrect decision, and supply the appropriate changes in articulation necessary for correct recognition. After a short while, I had learned what kinds of articulations were necessary for most of the errors that were occurring. Often the subjects were able

to figure out what to do without any coaching on my part.

The methods of changes in articulation that are reported in this section are those that were found to be successful in achieving the desired results. Several other methods were tried but failed. One such method that is worth mentioning is tongue positioning. In general, subjects were not aware how their tongues were positioned for a given phoneme (except, perhaps, for dentals) so that when they were asked to move their tongues in a certain direction they were unable to do it. Dentalizing of postdentals [t] and [d] was the only adjustment in tongue position they were able to perform on request.

6.3.1 Changes in Articulation for Structural Errors

Structural errors are all errors that resulted in unsatisfactory conditions for the application of the vowel and consonant recognition algorithm. Upon the detection of such errors the subject was instructed to repeat the utterance with an appropriate change in articulation, if necessary. Structural errors were completely ignored in recognition results. One of the reasons is that it was always possible to correct for these errors in a very simple manner.

Included among the structural errors were the following.

1. Errors caused by fractional recording of the utterance, i.e., where a portion of the utterance is truncated because the push button to terminate the recording was pressed either too early or too late. (Remember that the buffer was able to handle only 810 ms of speech.) These errors were the responsibility of the experimenter, who pressed the push button, and they were easily detected from the total energy plot on the CRT display or by a playback of the recorded utterance. The subject was told that the error was the experimenter's fault and was instructed to simply repeat the utterance.

2. Segmentation errors attributable either to incorrect schwa location or incorrect vowel location. Errors in schwa location occurred because either the schwa was pronounced in a very weak manner or it had a very short duration. Corresponding instructions to pronounce a louder or a longer schwa were given. Errors in vowel location usually occurred because the vowel was very short in duration, so the subject was simply instructed to lengthen the vowel. Segmentation errors were very common initially but dwindled very fast as soon as the subject learned what sort of general pronunciation the program expected. The initial learning period usually took approximately 15 min of the first session of Experiment 1.

3. Errors resulting from a low signal-to-noise ratio. The subject was instructed either to speak louder or come closer to the microphone, sometimes as close as 1 inch. Speaking very close to the microphone created another problem, however. With certain sounds like [p], and sometimes [t], [f] and [s] as well, the microphone was driven into a nonlinear region, because of the strong puff of air that is sometimes associated with those sounds for some speakers. This caused large voltage excursions which sounded like "thuds" on playback. The "thud problem" was eliminated in a crude manner by having the subject place a piece of paper, held tight by both hands, before

the microphone. This piece of paper acted as a highpass acoustic filter and eliminated the low-frequency "thud." Another way to treat this problem required the subject to consciously try to reduce the strength of the puff of air for those consonants. This method succeeded with some subjects, but not completely.

6.3.2 Changes in Articulation for Vowel Errors

The two major methods used with vowel errors were rounding and diphthongization.

1. Rounding. A conscious rounding and protruding of the lips causes a drop in the first formant F1, and hence in X_1, the X coordinate of energoid 1. This is basically due to the effective lengthening of the vocal tract. The drop in X_1 reflects itself as a drop in AX1A and AX1B, the average values of X_1 over the first half and second half of the vowel region, respectively (see Fig. 23). From the recognition algorithm in Fig. 24 it is clear that this form of adaptation in articulation can change the vowel recognition from low to mid or from mid to high. This method was used with the back vowels only, since rounding is not common with front vowels in the English language. Therefore, rounding was useful when [o] was incorrectly recognized as [a] and when [u] was incorrectly recognized as [o].

Rounding was also successful as a means of de-diphthongization of [u] in certain contexts. In American English, it is quite common to diphthongize [u] after dentals, so that [ədud] is often pronounced [ədjud]. The recognition algorithm recognized such vowels as front instead of back. Instructing the subject to protrude his lips as he pronounced the [u] prevented the diphthongization from materializing, and hence resulted in correct recognition.

2. Diphthongization. In general, the mid vowels [e] and [o] are more diphthongized than [i] and [u]. This fact was taken advantage of in the separation between mid and high vowels (see Fig. 24). An obvious method used to effect a change in vowel recognition from high to mid, for example, was to diphthongize the vowel more. The reverse process was also used, but it was easier for subjects to consciously diphthongize the vowel rather than avoid or prevent diphthongization. Both methods were used successfully to effect changes in recognition from high to mid, mid to high and low to mid, for all vowels.

6.3.3 Changes in Articulation for Consonant Errors

Consonant errors can be divided into manner-of-articulation errors and place-of-articulation errors.

1. Manner-of-Articulation Errors. These were errors that could be corrected by supplying the proper amount of voicing or frication or both. Figure 26 shows that the stop-fricative decision depends mainly on MINIM, the minimum total energy in the consonant region. All fricatives must have MINIM > 5 (except [f] which can have

MINIM > 3) and all stops must have MINIM ≤ 5. Errors usually occurred when a voiced fricative such as [v] was not fricated enough and was recognized as a voiced stop, or when a voiced stop was intensely voiced and was recognized as a fricative. The last problem was more acute with females than with males. The solution was to instruct the subject to either de-emphasize the articulation of the voiced stop or lengthen the stop gap, which would allow the voicing energy to diminish with time and result in a lower value of MINIM. The problem of [v] being recognized as a voiced stop was solved either by voicing and frication emphasis or by shortening the duration of the voiced fricative. Most subjects were able to control the strength of voicing and frication, but not all.

Some voiced fricatives were incorrectly recognized as unvoiced. This was more prevalent with [z] than with [v]. A conscious effort to emphasize the voicing in the fricative usually resulted in correct recognition.

2. Place-of-Articulation Errors. Errors in labial-nonlabial distinction were corrected by changing the total energy levels of the consonants. In case of errors, subjects were instructed to emphasize the stop burst for nonlabial stops, to de-emphasize the burst for labial stops, to strongly fricate a nonlabial fricative, and to reduce the amount of frication for labial fricatives. Some subjects were not able to sufficiently reduce the frication in [f], so they were instructed instead to greatly emphasize the following vowel. Apparently, shifting their emphasis to the vowel resulted in an unconscious reduction in frication energy of the fricative, and the method was successful.

Errors between nonlabials were the most difficult to correct by changes in articulation. Often, errors with dentals were corrected by "dentalizing" the articulation more.

6.4 SPEAKER ADAPTATION AND LEARNING

The speaker's adaptation to the recognition scheme can be divided into two major forms of adaptation: structural adaptation and phonetic adaptation.

Structural adaptation includes articulating an utterance [əCVd] with an unstressed schwa and a stressed long vowel V with a signal-to-noise ratio of at least 20 dB, such that proper segmentation of the utterance ensues. We have discussed some methods for changes in articulation that were sometimes necessary for this form of adaptation. Subjects adapted very well to this portion of the recognition algorithm. Although no record of structural errors was kept, it was clear that subjects were well adapted within 5-15 minutes. For the most part, speakers (other than the 12 subjects) who just walked up to the microphone were able to adapt their articulations appropriately with a few instructions and in just a few trials.

Adaptation to the recognition algorithms for the vowels and consonants is called phonetic adaptation. In the first learning session a record was kept of the progress of adaptation only for the interesting cases. In the second learning session, however, the

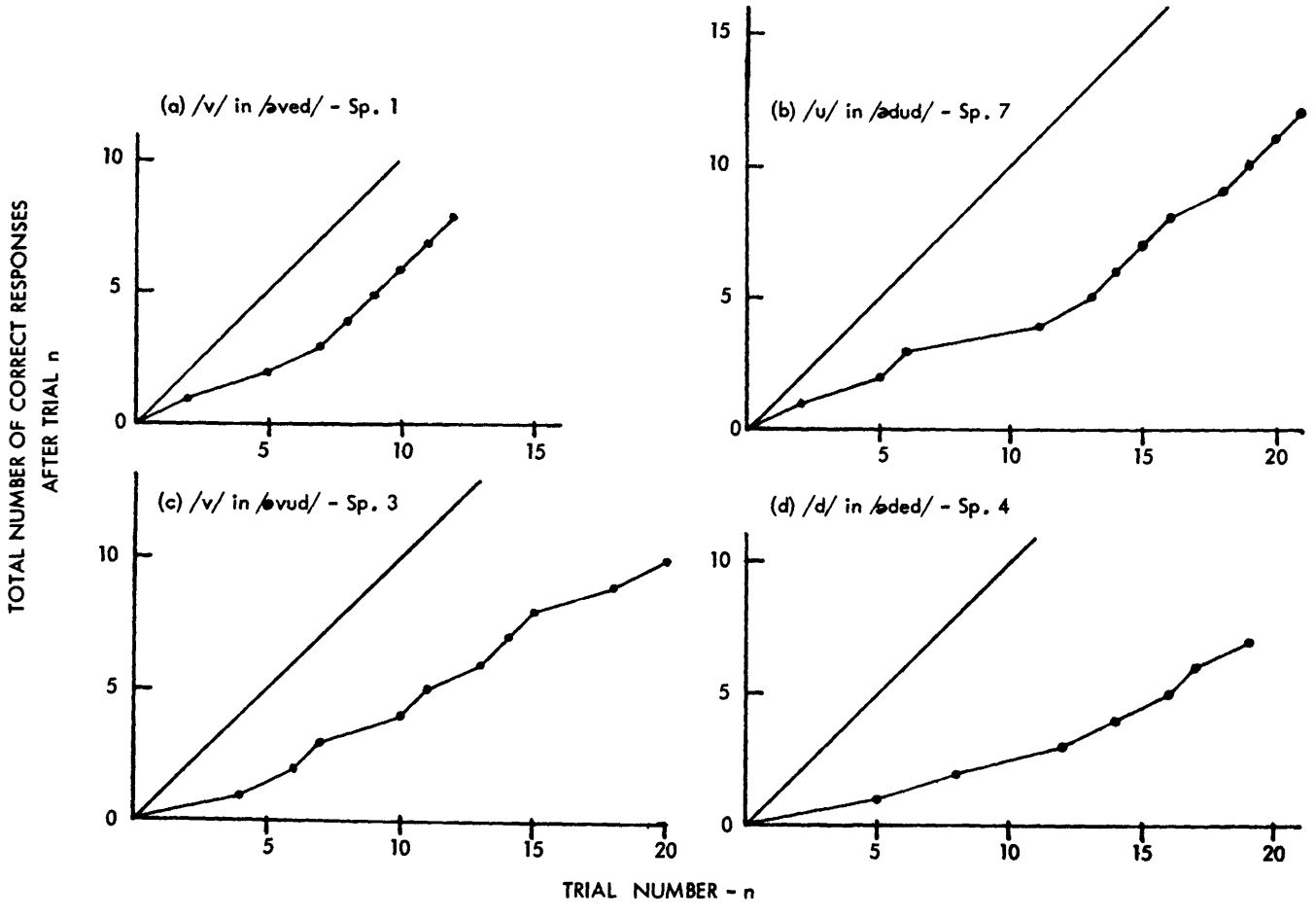


Fig. 37. Learning curves.

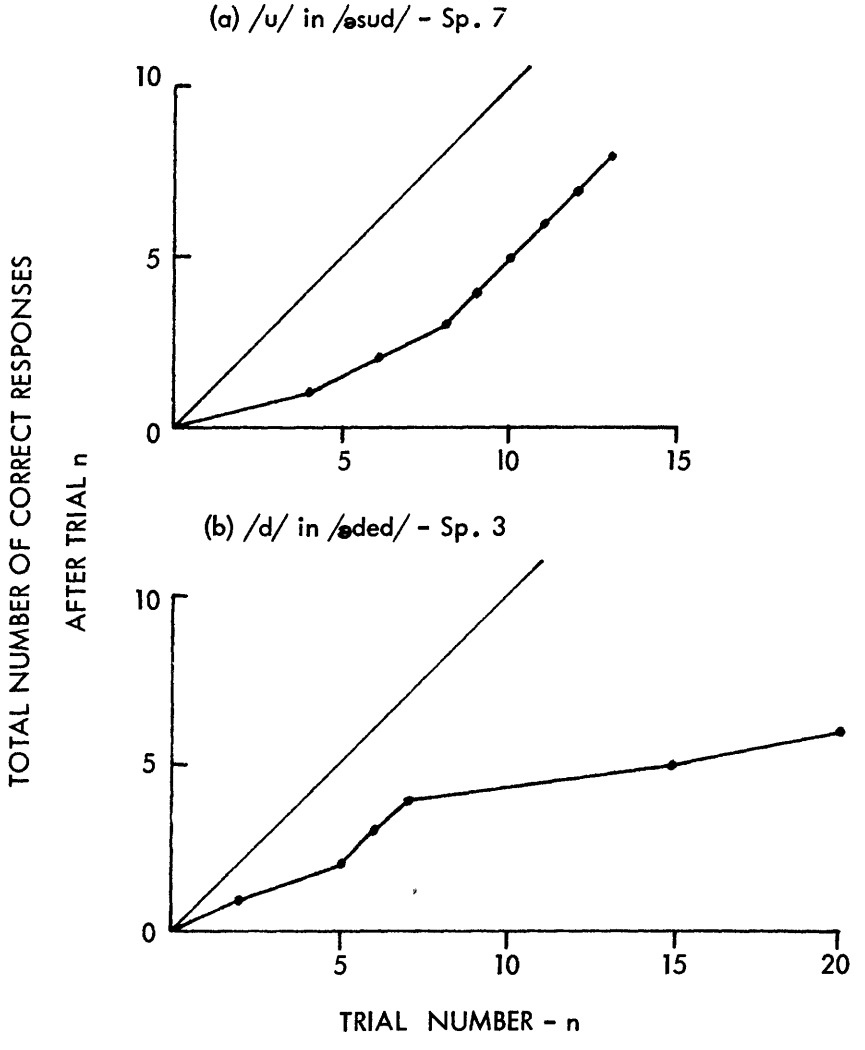


Fig. 38. Learning curves.

record was kept for all utterances for which subjects had to undergo a process of adaptation. Different methods for changes in articulation have been discussed. The adaptation process for the majority of the incorrectly recognized utterances was successful in very few trials, and it was difficult to judge the type of learning that the subjects were going through. There was a handful of cases, however, in which the learning process extended to more than 10 trials, and it is worthwhile to examine those cases. Figures 37 and 38 show learning curves for some of those utterances.

The trial number n is plotted on the abscissa, and the total number of correct responses after trial n is plotted on the ordinate. The black dots, representing points of correct recognition, are joined by straight lines. The ideal perfect recognition would be represented in each case by the 45° diagonal line. The learning process is considered successful if the learning curve becomes parallel to the ideal diagonal line, as it does in Figs. 37a, 37b, and 38a. Note that all three learning curves are typified by a sudden change in slope (at trial 7 in Fig. 37a, at trial 13 in Fig. 37b, and at trial 8 in Fig. 38a). In most cases for which the learning process was successful, this sudden change in slope occurred much earlier, around the second trial. This indicates that the learning is not a gradual process, but rather comes about as a result of a sudden conscious realization on the part of the speaker about the articulation needed to effect the correct recognition. This is attested to by the verbal remarks of the speakers. For example in Fig. 37b, the [u] in [ədud] was being incorrectly recognized as a front vowel because it was pronounced as a diphthong [ədjud] instead of [ədud]. This was immediately explained to the speaker, but it was not until the 12th trial that she suddenly realized what she was doing, and her remark was "Oh, I see! It's because of the [d]" (meaning that the [d] was causing her to diphthongize [u]), and she immediately proceeded to change her articulation accordingly.

If the learning curve did not become parallel to the ideal diagonal line, the adaptation process was considered unsuccessful, as in Figs. 37c, 37d, and 38b. In these cases we could say that the over-all slope of the learning curve approximates the probability of correct recognition for the particular utterance. For example, the probability of correctly recognizing [əvud] as spoken by Speaker 3 is approximately 50% (as computed from Fig. 37c), and the probability of correctly recognizing [əded] as spoken by Speaker 4 is approximately 30% (Fig. 37d). It is possible, in principle, that in further learning sessions these speakers could improve on their recognition probabilities for those particular utterances. It appears that as a rule of thumb if a speaker was going to be successful in an adaptation process it usually occurred within the first 20 trials. The adaptation process in Fig. 37c might indeed have been an exception to this rule, but that would have been very unlikely for the adaptation process in Fig. 37d.

6.5 RECOGNITION AND ADAPTATION RESULTS

We shall now discuss the results of the experiment described in section 6.2. To recapitulate: the vocabulary or recognition set contained 55 [əCVd] utterances with

C and V as shown in Fig. 1. The set was tested on 12 adult subjects, 6 males and 6 females. The experiment ran as follows for each subject.

Experiment 1. (A) First run (i.e., reading of the list) and errors recorded in error list 1.

(B) Informal learning session.

Experiment 2. (A) Second run and errors recorded in error list 2.

(B) Utterances in error list 2 were each repeated twice; those utterances correctly recognized twice in a row were eliminated from error list 2, which resulted in a reduced error list.

(C) Formal learning session for utterances in the reduced error list. Those resulting in successful learning curves (as discussed in section 6.4) are eliminated from the reduced error list, which resulted in a final error list.

The results that will be discussed are error list 1 from Experiment 1A, error list 2 from Experiment 2A, reduced error list from Experiment 2B, and the final error list from Experiment 2C.

6.5.1 Recognition Results

Figure 39 shows the recognition results of Experiments 1A and 2A. Bar graphs of the error rates are displayed against the speaker number. The average error rates for all 12 speakers are shown to the far right, above the word "Average." The upper graph shows the consonant error rates, the middle graph shows the vowel error rates, and the total word error rates are shown in the lower graph. The white bars show the results of Experiment 1A, before any learning had occurred. The solid bars show the results of Experiment 2A, which took place a week after the first learning session. Speakers 1-6 are male; Speakers 7-12 are female. Speakers 6, 8, and 11 were used in the original analysis, and of these, Speakers 8 and 11 were phonetically aware. Note that their performance is not very different from that of the other speakers.

In general, the vowel error rates were much lower than the consonant error rates. Speaker 10 had a very high vowel error rate in the first experiment, because of exceptionally high first formants for high vowels, for which there was no hope of a cure through learning. So, instead of quietly eliminating this subject from the experiment, the decision algorithm was slightly changed (the test in the dashed box in Fig. 24 was eliminated), and as a result, the recognition improved considerably in the second run. That change was responsible, however, for the increased vowel error rate for Subject 9, but these errors were successfully eliminated in the second learning session.

On the average, the recognition results were better in the second run than in the first, but not much. A possible reason for this is that the second run took place a week after the first learning session, and the subjects did not apply the

changes in articulation which they learned in the first session, although they were reminded of these changes before the run. Another possible reason is that it was difficult for a subject to concentrate on each utterance separately when reading the whole list of 55 utterances.

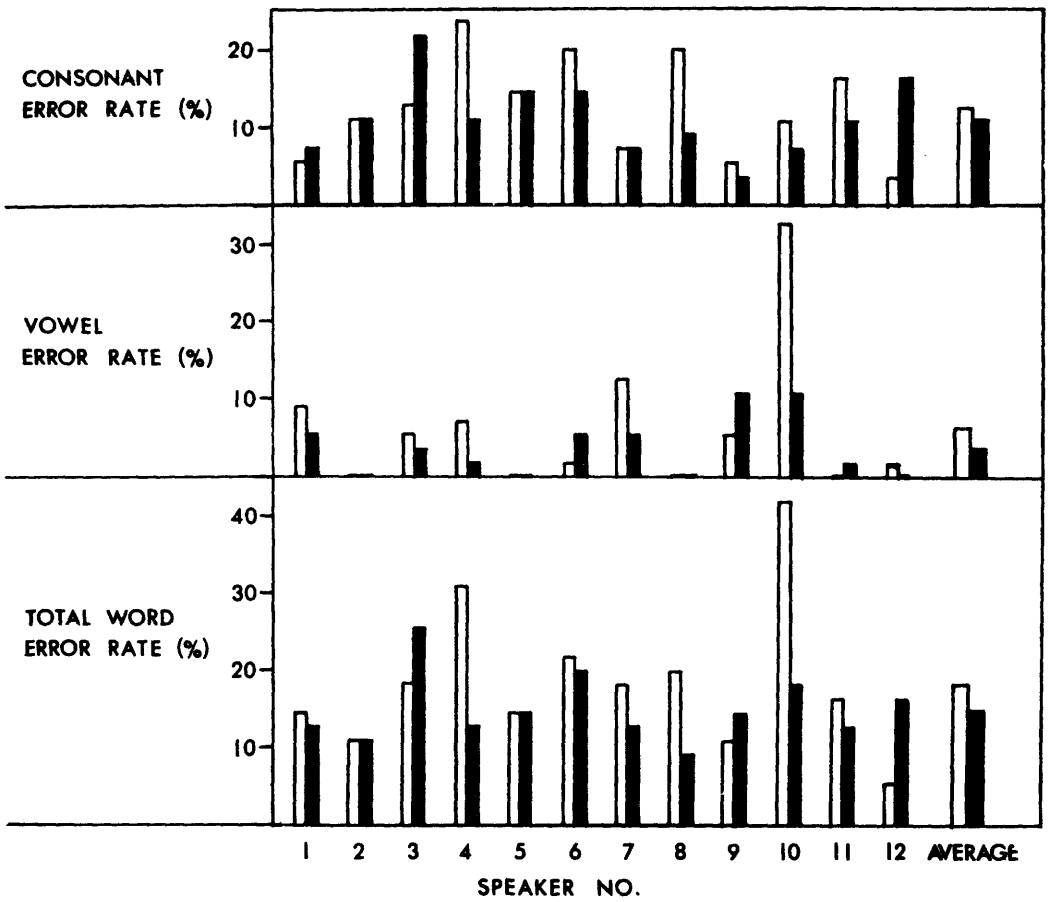


Fig. 39. Recognition results. Exp. 1A □ male speakers 1-6.
Exp. 2A ■ female speakers 7-12.

In order to examine the types of errors more closely, let us first take a look at the vowel errors in the second run. (The vowel errors from the first run will not be discussed, because of the change in the vowel algorithm after the first run.) Figure 40a shows the vowel confusion matrix. The errors are regrouped in terms of vowel features in Fig. 40b and 40c, where every error indicates a single feature error. Figure 40 shows that the total number of feature errors is equal to the number of vowel errors, which means that no vowel recognition was incorrect by more than a single feature. The 6 [e]/[o] confusions were the only front/back confusions and they were all made by Speaker 9, because of unusually low energy at high frequencies. Out of the 10 [u]/[i] confusions, 7 resulted from the utterance [ədud]. All [u] resulting in the 10 confusions were being diphthongized as [ju]. The

MACHINE RESPONSE

STIMULUS	i	e	a	o	u	Vowel Error Rate (%)
i	130	2				1.5
e	1	125		6		5.3
a			132			0.0
o			1	129	2	2.3
u	10			3	119	9.8

(a)

RESPONSE

STIMULUS	Front	Back
Front	258/264	6
Back	10	386/396

(b)

RESPONSE

STIMULUS	High	Mid	Low
High	259/264	5	
Mid	3	260/264	1
Low			132/132

(c)

Fig. 40. Vowel recognition results for Experiment 2A. Vocabulary: 55 [CVd] utterances, 11 consonants with each vowel. One repetition by 6 male and 6 female speakers. Total number of utterances presented: 660.
(a) Vowel confusion matrix.
(b) and (c) Vowel feature-confusion matrices.

STIMULUS \ MACHINE RESPONSE	p	t	k	b	d	g	f	s	š	v	z
p	108	1	9	1		1					
t	3	114	3								
k	1	7	110			2					
b				110		6				4	
d				9	98	12				1	
g	1			7	7	103				2	
f	1		1				108	8		2	
s							1	111	8		
š							1	5	114		
v			5	15	3		3			89	5
z							1	18		3	98

Fig. 41. Consonant confusion matrix for Experiments 1A and 2A.
5 vowels with each consonant.
Two repetitions by 6 male and 6 female speakers.
Total number of utterances presented: 1320.
(See percentage error rates in Fig. 43.)

high-mid-low confusions were quite random in nature. (Vowel adaptation results will be given in section 6.5.2.)

Now let us examine the consonant confusions. Since the errors in the two experiments were not very different, we shall combine the confusions from both runs and discuss the resultant. Figure 41 shows the combined consonant confusions for Experiments 1A and 2A. The four largest error rates are due to the consonants [v], [z], [d], and [g]. The errors are regrouped as shown in Fig. 42. The numbers along the diagonal represent place-of-articulation errors, and the rest represent manner-of-articulation errors. Note that for stop and unvoiced fricative consonants, most of the errors were place-of-articulation errors, while most of the errors for voiced fricatives were manner-of-articulation errors.

STIMULUS \ MACHINE RESPONSE		STOP		FRICATIVE	
		unvoiced	voiced	unvoiced	voiced
STOP	unvoiced	24	4		
	voiced	1	41		7
FRICATIVE	unvoiced	2		23	2
	voiced	5	18	22	8

Fig. 42. Recognition errors for Experiments 1A and 2A.
Total number of stimuli: 1320 utterances.

It is of interest to examine the number of consonant errors for consonants followed by either back vowels [CV_b] or front vowels [CV_f]. Figure 43a shows the corresponding error rates. For stop consonants, [CV_f] error rates are consistently greater than [CV_b] error rates, the differences being very marked for [d] and [g]. Recall that the algorithm for stop recognition was different, depending on whether the stop was followed by a front or a back vowel (see Figs. 33-35). Figure 33 shows that there is only one test to determine whether a nonlabial stop is dental or velar when followed by a front vowel, while several tests exist in Figs. 34 and 35 to perform the same task. In short, the algorithm for stop recognition is much more thorough for [CV_b] syllables than for [CV_f] syllables. Returning to Fig. 43a, the error rates for fricatives do not show any consistent tendencies with regard to whether the following vowel is front

	ERROR RATES (%)				
	With Back Vowels	With Front Vowels	Male Speakers	Female Speakers	Total
p	6.9	14.6	18.3	1.7	10.0
t	2.8	8.3	5.0	5.0	5.0
k	2.8	16.7	6.7	10.0	8.3
b	6.9	10.4	6.7	10.0	8.3
d	1.4	43.3	25.0	11.7	18.3
g	4.2	29.2	13.3	15.0	14.2
f	6.9	14.6	10.0	10.0	10.0
s	8.3	6.3	15.0	0.0	7.5
š	6.9	2.1	8.3	1.7	5.0
v	22.2	31.2	35.0	16.7	25.0
z	16.7	20.8	11.7	25.0	18.3
Average	7.8	18.0	14.1	9.7	11.9
Total number of utterances	792	528	660	660	1320
	(a)		(b)		(c)

Fig. 43. Consonant error rates from Experiments 1A and 2A.
(a) Error rates for consonants followed by back or front vowels.
(b) Error rates according to male or female speakers.
(c) Total error rates.

or back. Again, recall that the algorithm for fricative recognition makes no distinction about the following vowel. The conclusion that can be drawn is that, indeed, fricative spectra (during the "steady-state" portion) are to a large extent independent of the particular following vowel.

Figure 43b shows the consonant error rates averaged separately for the male and female speakers. On the average, the female speakers (Speakers 7-12 in Fig. 39) scored better than the male speakers (Speakers 1-6). Some of the discrepancies in error rates between the male and female speakers are discussed below.

[p]: The errors by the males were mostly due to strong [p] bursts that had energy peaks at frequencies close to those of [k] bursts.

[d], [g]: As evident from Fig. 43a, most of the errors were due to [CV_f] syllables. This is due to the closeness in the frequency peaks of the burst spectra for [d] and [g] before front vowels. It was observed that, on the average, the females exhibited larger separation in the frequency peaks between [d] and [g] than did the males, and that may be one of the reasons for better performance by the females.

[s], [$\overset{v}{s}$]: Figure 41 shows that most [s] and [$\overset{v}{s}$] errors were [s]/[$\overset{v}{s}$] and [$\overset{v}{s}$]/[s] confusions. Now, the distinction between [s] and [$\overset{v}{s}$] was based mainly on the position of energy prominence in the spectrum. If the position of prominence was at very high frequencies (6-7 kHz) the fricative was recognized as [s], if the position was around 2-3.5 kHz it was recognized as [$\overset{v}{s}$], and if in between it could have been either. As a rule, the spectral energy prominence for either [s] or [$\overset{v}{s}$] was at higher frequencies for the females than for the males. Also, frequency separation between [s] energy prominence and [$\overset{v}{s}$] energy prominence was larger for the females than for the males. Because for females the energy prominence of [s] was always at very high frequencies and because of the relatively large frequency separation between [s] and [$\overset{v}{s}$] energy prominences, the error rates for females were negligible. On the other hand, the males exhibited a wider range of frequency positions for [s] and [$\overset{v}{s}$] energy prominences and hence resulted the error rates in Fig. 43b.

[v]: Most of the [v] errors were fricative/stop confusions attributable to the low energy level of [v] voicing and frication. In general, the females were able to produce stronger [v]'s than the males, which resulted in lower error rates for the females. But the error rates were high for both males and females. This was a direct result of the algorithm for stop-fricative recognition (Fig. 26) which depended mainly on one simple test for the energy level in the consonant. We hoped that the speakers would easily adapt to that simple test. (The results of adaptation are given in section 6.5.2.)

[z]: Most of the [z] errors were [z]/[s] confusions, i.e., voicing errors. Again, there was only one test for voicing with fricatives (Fig. 27) based on only one FFT-computed spectrum. Voicing was detected by the high energy level at the first three filter numbers. Such voicing energy did not always materialize in the spectrum, in spite of the voicing evident from the time waveform display. The main reason for this is that the FFT-computed spectrum is often phase-sensitive, so that a shift in the time

window by 2 ms would cause considerable differences in voicing energy levels. The voicing energy for the females would often be restricted to the first filter number, so that the average energy at the first three filter numbers would be considerably reduced, thereby resulting in an unvoiced decision, and hence higher error rates for females. Note that the problems of voicing with fricatives did not occur to the same degree with stops because the voiced-unvoiced recognition algorithm for stops (Fig. 32) also used the approximate burst length and minimum total energy level to improve the recognition.

6.5.2 Adaptation Results

First, the rationale for the specific format of Experiment 2 will be given, followed by the results of speaker adaptation. The purpose of Experiment 2B and 2C was to approximate the optimum recognition results that are possible with the present system and with speaker adaptation, without having to perform several runs with the 12 speakers.

Assume that with each utterance $\underline{u}$ and Speaker $\underline{s}$ we associate a probability $P_s(u)$:

$$P_s(u) \equiv \text{probability of } \underline{\text{incorrect}} \text{ recognition of utterance } \underline{u} \text{ spoken by Speaker } \underline{s} \tag{42}$$

where u is one of the 55 [əCVd] utterances and s is the speaker number, $1 \leq s \leq 12$ in this experiment. An approximation to $P_s(u)$ can be obtained experimentally by recording the recognition errors from many repetitions of u by Speaker s.

$$P_s(u) \cong \text{error rate for utterance u spoken by Speaker s.} \tag{43}$$

The ultimate goal in adaptation is to reduce $P_s(u)$ until it approaches zero:

$$P_s(u) \to 0, \qquad \text{for all u and s.} \tag{44}$$

($P_s(u) = 0$ is practically impossible to attain.) Acknowledging that the goal in (44) is unattainable for all u and s, we would like to see how well it could be approximated.

Examples of changes in $P_s(u)$ through adaptation can be approximately obtained from learning curves such as those in Figs. 37 and 38 as follows:

$$P_s(u) \cong 1 - \text{slope of learning curve.} \tag{45}$$

For example, $P_3(vu) \cong 0.5$ in Fig. 37c, and $P_4(de) \cong 0.7$ in Fig. 37d, while in Fig. 37a, $P_1(ve)$ changes from a value of $P_1(ve) \cong 0.5$ to $P_1(ve) \cong 0.0$, thereby indicating "successful" adaptation.

It is important to know what is meant by "successful" adaptation in terms of future recognition performance. A learning curve exhibiting successful adaptation indicates that the subject is consistently able to produce the proper articulation necessary

for correct recognition of a specific utterance u_i at the time of the experiment. This does not necessarily mean that the subject had actually learned the proper articulation for u_i in the sense that a week later he would automatically be able to reproduce the articulation of u_i which he had learned. What it does mean is that after several runs and learning sessions the proper articulation would indeed have been learned. Therefore, a learning curve with successful adaptation is interpreted as a prediction that the subject will eventually be able to produce the learned articulation such that condition (44) is met. On the other hand, a learning curve for which the adaptation process is unsuccessful indicates that, so far, it is not possible to predict whether or not the subject will be able to produce the articulation proper for correct recognition.

After several runs and learning sessions one could approximate $P_s(u)$ for all s and u simply by computing the error rates that had been reached. If $P_{s_j}(u_i)$ is "very small," that is, condition (44) is met for a particular u_i and s_j, then we conclude that the proper articulation for u_i had actually been attained (learned) by speaker s_j. How small is "very small"? This depends very much on the particular application of the recognition system. In our system, we shall consider that the proper articulation for u_i had been learned if, given that a certain instance of u_i by s_j results in incorrect recognition, then the speaker is immediately able to provide, with almost 100% certainty, another instance of u_i with the proper articulation resulting in correct recognition. How small $P_s(u)$ should be to meet the preceding condition can be calculated only after an appropriate amount of data is compiled.

In light of the discussion above let us consider the implications of the testing sequence in Experiments 1 and 2, as outlined at the beginning of section 6. 5, and the results thereof. The purpose of the first run (Experiment 1A) was to get an idea about how well the system performs with new speakers. The average recognition rates for the 12 speakers are shown in Fig. 44 and also in the bar graph of Fig. 39. The first learning session took place one week after the first run and was devoted completely to an informal

Experiment	Consonant Error Rate (%)	Vowel Error Rate (%)	Total Word Error Rate (%)
1A	12.6	6.4	18.3
2A	11.2	3.8	15.0
2B	5.0	0.8	5.8
2C	2.3	0.0	2.3

Fig. 44. Over-all recognition results for the 12 speakers.

exploration of different possibilities of changes in articulation that might lead to better recognition performance. It was evident from the learning session that some success was already being achieved.

Approximately one week later the second run took place. By that time the subjects had forgotten what they might have learned in the first learning session. Although they were reminded of the changes in articulation that were useful for each of them, they had a hard time remembering them during the run, and the results of the second run (Experiment 2A in Fig. 39) were only slightly better than those of the first run. The consonant error rate improved from 12.6% to 11.2%, a decrease of only 1.4%. The significantly large decrease in the vowel error rate (from 6.4% to 3.8%) is attributable mainly to the change in the vowel recognition algorithm (see section 6.5.1). The effects of the first learning session (Experiment 1B) on the second run (Experiment 2A) were minimal. Although the individual recognition errors in Experiment 2A, in general, were different from those in Experiment 1A, the combined distribution of feature errors for the 12 speakers in Experiment 1A remained essentially the same in Experiment 2A.

Immediately following the second run the utterances in error list 2 were each repeated twice in a row with specific instructions to the subject to try to articulate the utterances properly. Hints about the changes in articulation that might be helpful were supplied by the experimenter. For those utterances that were correctly recognized twice in a row the interpretation was as follows: Since the subject was able to deliberately articulate those utterances and effect correct recognition twice in a row, then it was highly likely that the subject would indeed learn those proper articulations, in the sense discussed before, if he had not done so already. Since our aim was to find out the optimum results possible with this system, it was legitimate to subtract the twice-correctly-recognized utterances from error list 2, thereby resulting in a reduced error list. The average error rates in the reduced error list are shown in Fig. 44 under Experiment 2B. The reduction in the total word error rate from 15.0% to 5.8% is appreciable. It shows that, on the average, most of the errors were due to utterances whose proper articulations were either already learned or could easily be learned by the subjects.

For each of the utterances in the reduced error list the subject underwent a learning process, and a learning curve was plotted, as in Figs. 37 and 38. In light of the discussion concerning the relation between successful adaptation and learning, it would be fair to eliminate those utterances resulting in successful adaptation from the reduced error list, since they would be learned in the future. The error rates from the final error list are shown under Experiment 2C in Fig. 44. The vowel error rate went to zero, and the consonant error rate decreased to 2.3%. This means that after several runs and learning sessions for the 12 speakers it is possible that the average recognition results could so improve that no vowel errors and only 2.3% consonant errors would occur. It is possible that for some of the utterances in the final error list, the speakers might later on adapt their articulations properly. In that case the

	ERROR RATES (%)			
Subject Number	Experiment 1A	Experiment 2A	Experiment 2B	Experiment 2C
1	14.5	12.7	1.8	0.0
2	10.9	10.9	7.3	0.0
3	18.2	25.5	12.7	7.3
4	30.9	12.7	3.6	1.8
5	14.5	14.5	7.3	7.3
6	21.8	20.0	9.1	1.8
7	18.2	12.7	1.8	0.0
8	20.0	9.1	5.5	1.8
9	10.9	14.5	7.3	1.8
10	41.8	18.2	0.0	0.0
11	16.4	12.7	3.6	0.0
12	5.5	16.4	9.1	5.5
AVERAGE	18.3	15.0	5.8	2.3

Fig. 45. Total-word error rates for each of the 12 speakers. Subjects 1-6 are male and 7-12 are female.

STIMULUS \ MACHINE RESPONSE	p	t	k	b	d	g	f	s	š	v	z
p	54		5			1					
t	1	57	2								
k		4	55			1					
b				57		2				1	
d				6	47	7					
g	1			4	4	50				1	
f	1						56	3			
s								57	3		
š								3	57		
v				6	2		3			48	1
z							1	8		3	48

Fig. 46. Consonant confusion matrix for second run (Exp. 2A). 6 male and 6 female speakers.

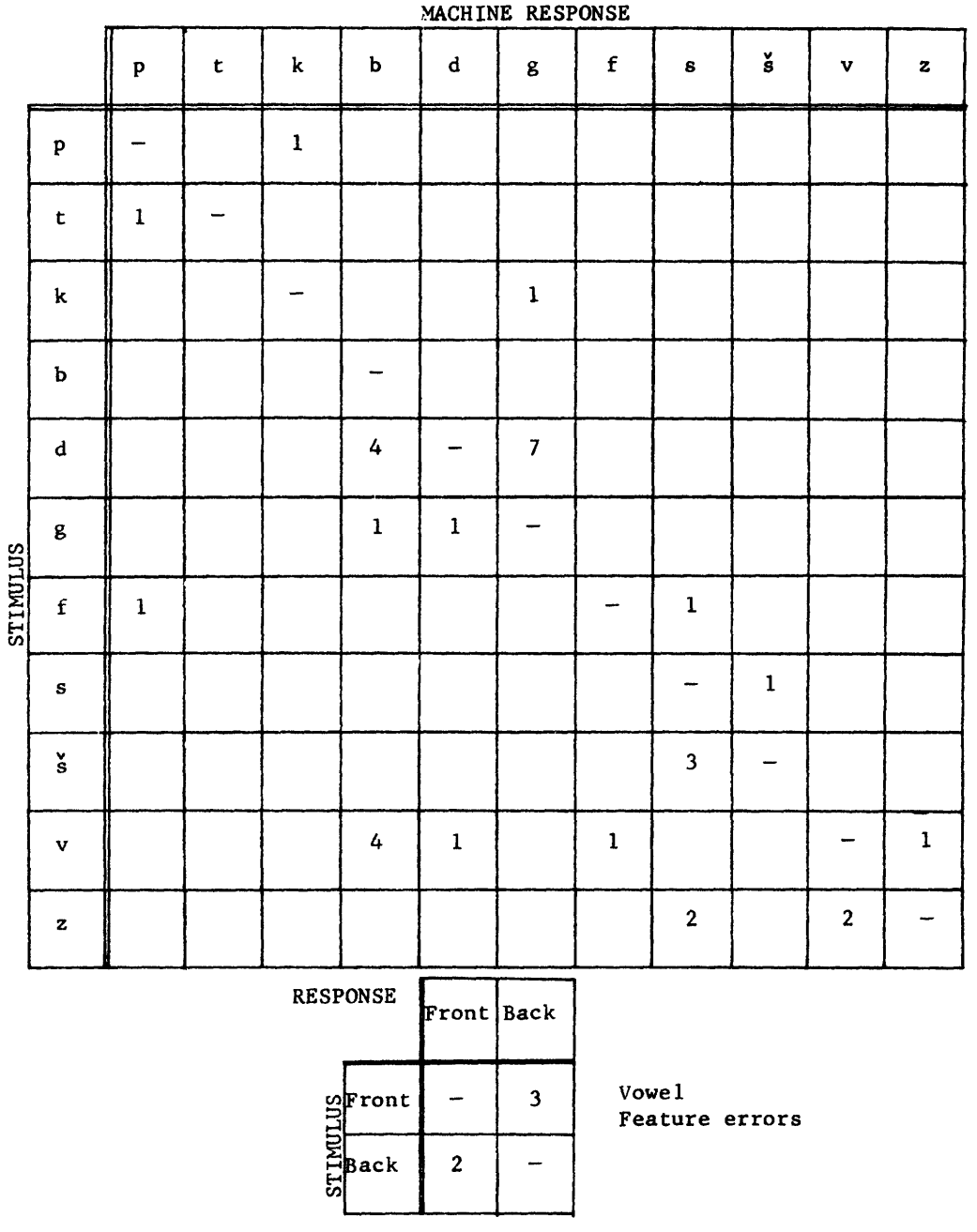


Fig. 47. Consonant and vowel errors in the reduced error list (Exp. 2B). 6 male and 6 female speakers.

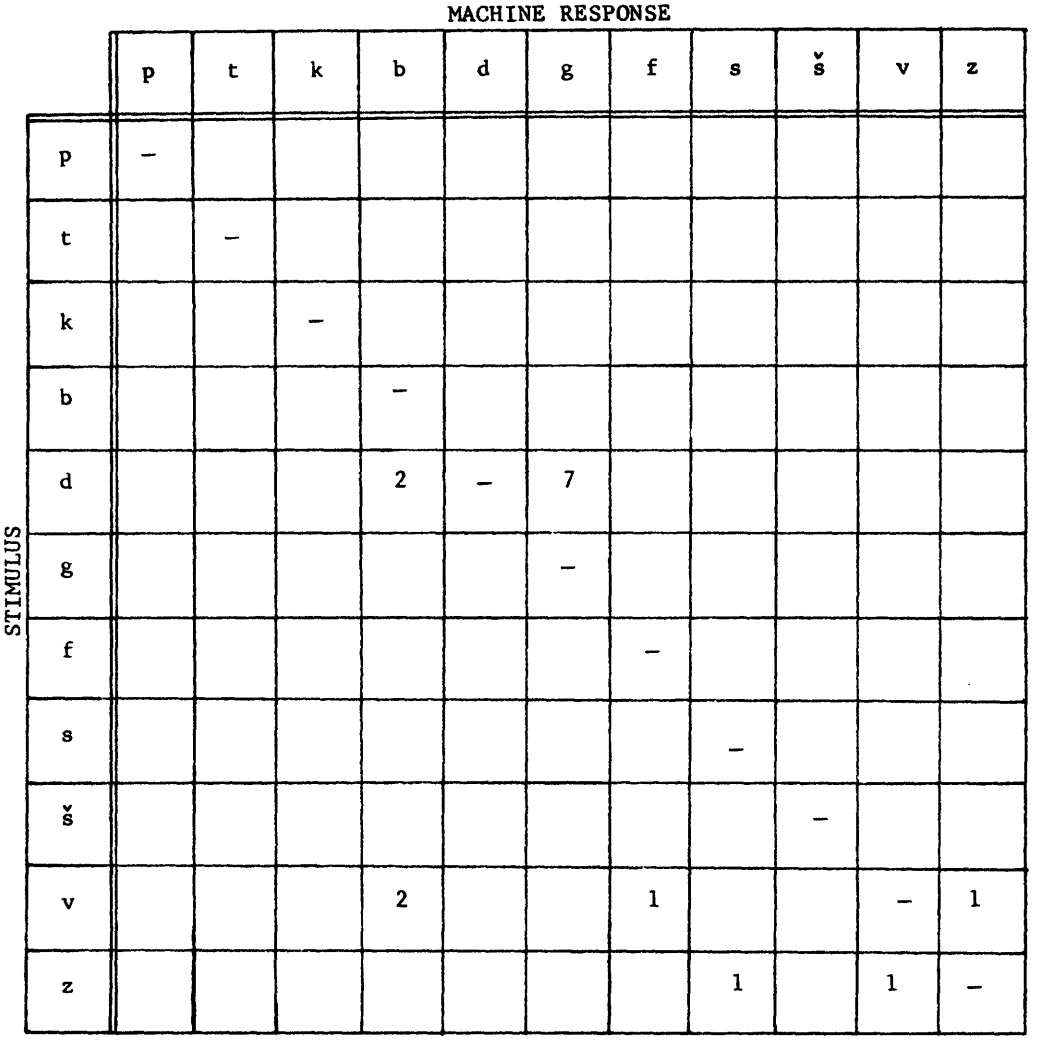


Fig. 48. Consonant errors in the final error list (Exp. 2C).
6 male and 6 female speakers.
No vowel errors in the final error list.

average error rate would decrease to less than 2.3%. Judging from the kind of errors in the final error list and from my knowledge of the recognition algorithm, however, it seems that the 2.3% error rate can be most effectively reduced by appropriate changes in the recognition algorithm. Whether the 2.3% average error rate could actually be attained or not can be determined only by further experimentation. It is this experimenter's opinion that such an error rate is indeed possible to attain after a few learning sessions, provided the definition of a recognition error is taken as follows: If an utterance is incorrectly recognized, it is considered as a recognition error only if the speaker is unable to effect correct recognition upon immediately repeating the utterance. This definition is consistent with the definition of successful adaptation and learning of an utterance given previously and upon which the 2.3% figure was obtained.

For a detailed look at how the different speakers performed relative to the average given in Fig. 44, Fig. 45 shows the progression of error rates from Experiment 1A, 2A to 2B to 2C, for each of the 12 participating subjects. It is worth noting that Speaker 3 was always very self-conscious, and Speaker 5 suffered from a slight speech impediment which was very evident during continuous speech. Often his [t] sounded like [k], but he was able to correct it with some effort. It was more difficult for him to articulate [v] properly; it almost always sounded like [b]. Speaker 12 (a 17-year old freshman) did very well in Experiment 1A, but not as well in Experiment 2. Actually, she was very interested in the experiment, so much so that she was trying too hard, which resulted in some unnatural pronunciations that led to incorrect recognition. I believe that the final 5.5% error rate could easily be reduced with one or two more sessions.

It is instructive to examine the types of errors remaining in the reduced and final error lists. Figure 46 shows the consonant confusion matrix for the second run (Experiment 2A). The vowel confusion matrix was shown in Fig. 40. The distribution of the errors in the reduced error list (Experiment 2B) is shown in Fig. 47. Note that the only vowel errors remaining were due to 3 front/back errors by Speaker 10 and two back/front errors by two other speakers; all 9 high-mid-low confusions were rectified. The errors in the final error list (Experiment 2C) are shown in Fig. 48. There were no vowel errors. The consonant errors in Fig. 48 occurred with [d] before front vowels and with the voiced fricatives [v] and [z]. These errors account for the 2.3% error rates shown in Fig. 44. It is interesting that they reflect the main weaknesses of the recognition algorithm which were discussed in section 6.5.1.

VII. FINAL REMARKS

We have examined the feasibility and limitations of speaker adaptation in improving the performance of a fixed (speaker-independent) automatic speech recognition system. The vocabulary was composed of 55 [əCVd] utterances, where C is one of 11 stops and fricatives, and V is one of 5 tense vowels. The results of the experiment on speaker adaptation, performed with 6 male and 6 female adult speakers show that speakers can learn to change their articulations to improve recognition scores. The initial average error rate was 18.3%; as a result of the experiment it was predicted that the average error rate could possibly decrease to 2.3%. Further experimentation is needed, however, to confirm the actual magnitude of the decrease in the average error rate caused by speaker adaptation. The preliminary results also indicate that most of the necessary adaptation can be achieved in a relatively short time, provided that the speakers are instructed how to change their articulations to produce the desired effects. Because of the experimental nature of this investigation, the experimenter's participation in supplying the necessary instructions to the speakers was necessary. In a system designed for actual use as a speech recognizer, the speaker would perhaps interact solely with the machine. Instructions could be programmed and presented to the subject, either visually or orally. Even under such a system, it might be necessary to provide initial training with the help of a human experimenter. After the initial training period, the instructions from the machine would be sufficient.

As a general rule of thumb, speakers adapted their articulations in a consistent manner only when they were able to consciously correlate the changes in articulation with the corresponding acoustic output. Several methods of changes in articulation were found useful. Rounding and protruding of the lips and diphthongization were used with vowel errors; deliberate efforts at voicing and/or frication were used with consonant errors. Place-of-articulation errors with stop consonants were corrected by a proper production of the stop burst; this was successful mainly with labial-nonlabial errors. Errors between nonlabials were the most difficult to correct by changes in articulation. The main reason for this is that the speaker was required to consciously manipulate the position of the tongue body, and all untrained speakers were unable to do this. There were cases such as with nonlabial voiced stops before front vowels in which even when the speaker was able to consciously manipulate the tongue position the resulting acoustic changes were not sufficient to result in correct recognition. In such cases, one should concentrate on improving the recognition algorithm.

The results of this study demonstrate the feasibility of a very limited speaker-independent speech-recognition system. The search for reliable, largely speaker-independent speech parameters proved to be fruitful for the limited vocabulary of 55 [əCVd] syllables, a fact that adds weight to the arguments supporting the use of the multispeaker approach to speech recognition. Even for an adaptive speech-recognition system, the search for speaker-independent parameters is bound to lead

to reliable and very useful parameters. That, of course, does not mean that important parameters will not be speaker-dependent. A parameter such as the burst length in stops is certainly speaker-dependent, but for any particular speaker it could be used very reliably in manner, as well as in place-of-articulation recognition of stops. It should not be construed from this discussion that any recognition system with limited vocabulary (e.g., up to 50 words) could be made speaker-independent without speaker normalization. The limited success of systems that recognize the 10 digits is a living testimony to the difficulties involved in speaker-independent speech recognition. The key issue in a limited speaker-independent recognition system is the vocabulary for which the system is designed. It so happens that the 10 digits in English are mostly one-syllable words with many similar features, and hence the difficulty in discrimination. On the other hand, if one were allowed to choose the vocabulary freely, then it would be easy to design a successful 50-word speaker-independent recognition system. In practice, the problem is that the vocabulary is often dictated by the specific application, such as ZIP Code or space-flight applications. In such cases, some form of adaptation by the recognition system is almost certainly necessary for successful recognition.

The performance of the recognition system was in general quite good, for females as well as for males, considering that the system was fixed and no speaker normalization was employed. The major errors for which speaker adaptation was not completely successful lie in three categories: (a) the detection of frication, (b) the detection of voicing in fricatives, and (c) the recognition of voiced stops before front vowels. In each of these three cases the recognition was based on one test based on only a single measurement. The lack of sufficient measurements and tests undoubtedly contributed to the poor results. In case (c), the lack of testing was mainly due to a lack of knowledge about how to design an algorithm to recognize stops before front vowels. In cases (a) and (b), the reasons for insufficient measurements were twofold: first and foremost, using the same processing techniques, such measurements would have increased the recognition processing time to a degree intolerable to a subject undergoing a process of adaptation, and second, it was believed (and hoped) that the subjects would be able to control frication and voicing of fricatives with little effort. This was indeed the case for most subjects, but not all. There is no doubt, however, that the recognition algorithm could be so improved that frication and voicing would cease to be a major source of error. It must be emphasized that the main aim of this study was to investigate the problem of speaker adaptation. The speech-recognition system that was designed served as a tool for the investigation. The system had to perform well enough for this application with its time limitations, and there was no attempt to optimize the performance of the system as an end in itself.

It is rather difficult to compare the performance of this recognition system with that of other nonadaptive systems that have been reported. This is due to differences in the

vocabularies and the conditions under which the results were obtained. For example, the inclusion of female, as well as male, speakers with different accents, and on-line testing with a signal-to-noise ratio of only 20-25 dB, are conditions that have been rarely present in other recognition systems. Also, a vocabulary of [əCVd] utterances with emphasis on place, as well as manner-of-articulation, recognition is not a common occurrence among other systems. The recognition system of Martin and his co-researchers[22] at RCA comes closest to this system, at least in terms of the vocabulary. The RCA system had a vocabulary of [CVd] utterances (with the consonants always in initial position), where V is one of 10 English vowels. The consonants were divided into three classes: stops, fricatives, and semivowels. The stops and fricatives were identical with those in Fig. 1 except for an additional [h], which was considered to be one of the fricatives. The recognition was performed only within the classes; there was no cross-class recognition (e.g., stop-fricative recognition was not available). The vowels were explicitly recognized only after semivowels. Therefore, we can compare the results for stop and fricative recognition in both systems. The results in the RCA report were based on recordings of the [CVd] words by 6 male speakers, with each word repeated twice. Keeping in mind the different conditions under which the RCA experiment was performed, a comparison of the recognition results of the RCA system and the results reported in Fig. 41 for Experiments 1A and 2A (with no speaker adaptation) is given in Table 3. (The stop/fricative and fricative/stop confusions are not included in these results. Also, the [h] confusions have been eliminated from the RCA results.)

Table 3. Comparison of error rates between the RCA recognition system and the one used in this study.

	RCA System	Present System
Stops before front vowels	28.6%	19.9%
Stops before back vowels	16.2%	3.1%
Average for all stops	21.8%	9.8%
Fricatives	3.8%	9.6%

In Table 3, the recognition results for stops are much better in the present system. On the other hand, the results for fricatives are better in the RCA system. Almost half the fricative errors in the present system were voicing errors, which have already been discussed. Furthermore, all error rates decreased appreciably with speaker adaptation.

It is interesting to speculate about how a human would perform, given the same task as in the present recognition system. There is no doubt that the human performance would be better than that reported in this study. The real question is, given that human listeners would make some errors, what kinds of errors would they be? Because of the

nature of the five vowels in this vocabulary, there probably would be no vowel confusions; most of the errors would be consonant errors. The study by Miller and Nicely[33] of perceptual confusions among some English consonants gives a clue as to the kinds of general errors to be expected. The largest number of consonant errors would be place-of-articulation errors, followed by frication errors, and then voicing errors, which should be minimal. Unfortunately, the study by Miller and Nicely was not comprehensive enough for our purposes. The vocabulary comprised [Ca] syllables only. "Five female subjects served as talkers and listening crew; when one talked, the other four listened" (Miller and Nicely[33]). This meant that the listeners were always able to adapt to the speaker. What is needed is another study that would include front, as well as back, vowels in the vocabulary, and have a randomized list of utterances spoken by several male and female speakers presented to a large group of listeners. My prediction is that stop consonants followed by front vowels would cause the largest number of confusions, as was the case in the present recognition system.

The processing techniques that were employed in the recognition system included storing the sampled time waveform and performing spectral analysis on it. The availability of the time waveform at all times provided a flexibility that is desirable in designing an experimental recognition system. In this system it was particularly useful for locating the stop burst accurately. Spectral analysis employed a bank of 36 bandpass filters and the Fast Fourier transform. This aspect of the system is discussed further in Appendix B. Fourier spectral analysis is not the only way by which speech could be analyzed. Researchers have experimented with several techniques, one of which I wish to mention briefly. Dolansky[64] described a method for expressing voiced speech sounds in a pitch-synchronous manner by means of complex-exponential base functions. The idea is very attractive but has the obvious drawback that the choice of the particular set of base functions to be used is a major undertaking that may not prove to be optimum for a large number of sounds and speakers. What might be a compromise between Fourier analysis and Dolansky's method is the evaluation of the spectrum on a line in the s-plane other than the $j\omega$-axis. A method for computing the z-transform of a sequence of samples at points in the z-plane that lie on spiral contours (which correspond to straight lines in the s-plane) has been described by Rabiner, Schafer, and Rader[65]; it is called the "chirp z-transform algorithm" (CZT). One of the applications of the CZT algorithm is the enhancement of poles for use in spectral analysis, which is very useful in the accurate location of formants. More recently, attention has been given to analyzing speech in terms of Walsh functions because of their attractiveness for computing purposes. I am not aware of any concrete results in that direction. (For other applications of Walsh functions see Harmuth.[66]) A similarly attractive set of functions is the Haar functions (see Collatz[67] for a definition).

My limited experience with the Haar functions did not show them to be promising for speech-recognition purposes. Whatever the method used in the processing of speech the crucial problem in speech recognition today is not that, but rather what one does after the initial processing.

The recognition scheme itself was based on the extraction of several acoustic features from the speech signal. This was accomplished by a hierarchy of decisions made on carefully selected parameters that were computed from a spectral description of the speech signal by means of a set of energoids, each energoid representing the center of energy concentration in a particular energy band. The positions of the energoids and their movements in time proved to be very valuable tools in developing parameters that were to a large degree speaker-independent. The resolution of energoid movements largely depended on the frequency resolution of the spectrum. When recognition depended on small shifts in energy concentration, as in the [CV] transitions in [gi] and [di], the energoid movements were unreliable, because of the coarse resolution of the spectral description at high frequencies, and hence were not used. The energoid is simply one method of describing energy concentrations and their movements. It has worked well in general, but more should be done in finding other means for the description of dynamic energy movements in frequency and time. Indeed, I wish to hypothesize that the human ear, although relatively insensitive to the exact positions of specific energy concentrations (e.g., formants), is highly sensitive to dynamic shifts in the positions of those energy concentrations, and that such information is essential for recognition. In particular, the human recognition of [gi] from [di] most probably depends on such information. In the absence of more accurate descriptions of energy movements, the designer of automatic speech-recognition systems would do well not to include words in the vocabulary that differ only in the place of articulation of a voiced stop when followed by a front vowel.

Finally, the phonemes employed in the vocabulary were restricted to stop and fricative consonants and tense vowels. Future experiments in speaker adaptation should include other phonemes in different contexts. Research in speaker adaptation should go on simultaneously with experiments in machine adaptation.

APPENDIX A

A Note on Distinctive Features

Much of the groundwork and further development of the concept of distinctive features may be found in the following works: Jakobson's Kindersprache (1941), which has been translated into English as Child Language, Aphasia and Phonological Universals[68]; Preliminaries to Speech Analysis by Jakobson, Fant and Halle[17]; and The Sound Pattern of English by Chomsky and Halle.[18] The vowels and consonants in Fig. 1 will be specified in terms of the distinctive feature system of Chomsky and Halle and that of a different system which will be proposed. The addition of a new feature [labial] is suggested and discussed.

According to the Chomsky and Halle distinctive feature system all vowels are classified as $\begin{bmatrix}+\text{vocalic}\\-\text{consonantal}\end{bmatrix}$ and all stops and fricatives as $\begin{bmatrix}-\text{vocalic}\\+\text{consonantal}\end{bmatrix}$. The five vowels shown in Fig. 1 are classified in terms of distinctive features in Fig. A-1.

	i	e	a	o	u
back	-	-	+	+	+
high	+	-	-	-	+
low	-	-	+	-	-

Fig. A-1. Distinctive feature compositions of the five vowels in Fig. 1.

The features shown in Fig. 2a are directly related to those in Fig. A-1. The feature front is equivalent to [-back] in Fig. A-1. The feature mid is equivalent to $\begin{bmatrix}-\text{high}\\-\text{low}\end{bmatrix}$.

Figure A-2a shows the distinctive feature composition of the eleven consonants shown in Figs. 1 and 2b according to the Chomsky and Halle system. Figure A-2b shows a different composition, with the feature [coronal] replaced by the feature [labial]. The Chomsky and Halle system can be mapped into the system in Fig. A-2b by the following substitutions:

$$\begin{bmatrix}-\text{coronal}\\+\text{anterior}\end{bmatrix} \longrightarrow [+\text{labial}]$$

all other segments $\longrightarrow$ [-labial].

Labial consonants are produced by forming a constriction using the lips; nonlabial consonants are produced by making constrictions that do not employ the lips.

The features in Fig. 2b are related to those in Fig. A-2b in the following manner:

stop $\longrightarrow$ [-continuant]; fricative $\longrightarrow$ [+continuant]

voiced $\longrightarrow$ [+voice]; unvoiced $\longrightarrow$ [-voice]

labial ⟶ [+labial]

dental ⟶ $\begin{bmatrix}-\text{labial}\\+\text{anterior}\end{bmatrix}$

palato-alveolar ⟶ $\begin{bmatrix}-\text{labial}\\-\text{anterior}\end{bmatrix}$

velar ⟶ $\begin{bmatrix}-\text{labial}\\+\text{back}\end{bmatrix}$

(Note the [+labial] is redundantly $\begin{bmatrix}+\text{anterior}\\-\text{back}\end{bmatrix}$ and hence need not be explicitly specified. Also, [+anterior] implies [-back], and [+back] implies [-anterior].) The features of Fig. 2b could, of course, also be specified in terms of the distinctive features in Fig. A-2a.

The remaining discussion focuses on the features [labial] and [coronal] without being confined to the eleven consonants in Fig. 1.

	p	t	k	b	d	g	f	s	š	v	z
continuant	-	-	-	-	-	-	+	+	+	+	+
voice	-	-	-	+	+	+	-	-	-	+	+
coronal	-	+	-	-	+	-	-	+	+	-	-
anterior	+	+	-	+	+	-	+	+	-	+	+
back	-	-	+	-	-	+	-	-	-	-	-

(a)

	p	t	k	b	d	g	f	s	š	v	z
continuant	-	-	-	-	-	-	+	+	+	+	+
voice	-	-	-	+	+	+	-	-	-	+	+
labial	+	-	-	+	-	-	+	-	-	+	-
anterior	+	+	-	+	+	-	+	+	-	+	+
back	-	-	+	-	-	+	-	-	-	-	-

(b)

Fig. A-2. Distinctive feature composition of the eleven consonants according to (a) the Chomsky and Halle system, (b) a different system where the feature [labial] is introduced.

Reasons for considering the feature [labial] are threefold. First, the lips are articulators physically distinct from other articulators. They are involved in two types of sound-generation activity: (a) production of labial consonants; and (b) rounding or labialization of consonants. The feature [rounded] is used to specify rounding. According to Chomsky and Halle, "All classes of consonants, including labials, may be rounded."[69] This shows that the feature [rounded] is independent of all features that specify the different consonant classes. The independence of the feature [rounded] from other consonant features indicates that the motor control for the lips may be quite independent from that of other articulators. Since the production of labial consonants involves mainly movement of the lips, it seems reasonable to postulate the existence of a feature [labial] which is also independent of other features.

Second, labials form an important class of sounds during the first stage of language acquisition by children. Jakobson writes, ". . . the acquisition of vowels is launched with a wide vowel, and, at the same time, the acquisition of consonants by a forward articulated stop. An [a] emerges as the first vowel, and generally a labial as the first consonant, in child language. The first consonantal opposition is that of nasal and oral stop (e.g., mama-papa), which is followed by the opposition of labials and dentals (e. g., papa-tata and mama-nana)."[70] He also writes, "After the splitting of the consonants into two series of phonemes –labials and nonlabials – the latter are normally realized as dentals, which the child occasionally replaces with palatals."[71] Therefore, the labial-nonlabial opposition is very real and basic in language acquisition. Although "the dentals become the natural foundation of the consonantal system in its subsequent autonomous development (Jakobson[72]), it does not necessarily make the labial-nonlabial distinction any less real.

Third, labials have acoustic features that are different from those of nonlabials. Some of those features have been discussed in section 4. 2. 2.

Let us now consider the feature [coronal]. As defined by Chomsky and Halle "Coronal sounds are produced with the blade of the tongue raised from its neutral position; noncoronal sounds are produced with the blade of the tongue in the neutral position."[73] The dental, alveolar, palato-alveolar consonants and liquids are coronal. Among the noncoronals are the back consonants (e. g., velars) and labials. Phonologically, coronals very often act as a single class in opposition to noncoronals. (It was very interesting for me to discover, for example, that in Lebanese Arabic all of the "sun letters" are coronal and all of the "moon letters" are noncoronal.) Now, let us consider a different approach to articulations employing the tongue.

For physiological reasons, it is more likely that, in the anterior portion of the oral cavity, sounds be produced with the blade of the tongue rather than with the body of the tongue. Similarly, articulations in the posterior portion of the oral cavity are more naturally produced by the body of the tongue. Somewhere in between the two positions it is likely that, at approximately the same point of articulation, distinct sounds could be produced by raising either the blade or the body of the tongue. In this case it would be very natural to differentiate between such sounds by the feature [coronal]. English

does not have such sounds, but they do exist in other languages. For example, in German, [s̬] in [lis̬t], "extinguishes," is coronal, and [ç] in [liçt], "light," is noncoronal; both [s̬] and [ç] have approximately the same point of articulation (Halle[74]).

According to the Chomsky and Halle system, the only case in English in which the feature [coronal] is distinctive is in separating labials from dentals. Labials are $\begin{bmatrix} -\text{coronal} \\ +\text{anterior} \end{bmatrix}$ and dentals are $\begin{bmatrix} +\text{coronal} \\ +\text{anterior} \end{bmatrix}$. Therefore, the feature [coronal] could be completely replaced by another feature [labial], as was proposed at the beginning of this discussion. The feature [coronal] would still be necessary in other languages but as a secondary feature.

Acoustically, what seems to be more important is not whether a constriction is formed by some articulator or another but rather by the shape of the constriction and the point of articulation. Whether a consonant is articulated using the lips, the tongue blade or the tongue body is in itself not acoustically significant; what is significant is the fact that the lips happen to be at one end of the vocal tract, that the tongue blade forms constrictions in the anterior portion of the oral cavity, and that the tongue body forms constrictions in the posterior portion of the oral cavity; such articulations divide the vocal tract in different ways and hence different acoustic outputs result. The other significant factor is the type of orifice or constriction that is formed. For example, bilabial and labiodental continuants, which exist in Ewe (Chomsky and Halle[75]), have the same point of articulation; what differentiates them is the manner in which they are produced, i. e., the type of constriction that is formed. The fact that [l] and [r] in English are produced by the blade of the tongue, in itself, sheds no light on the acoustical properties of these sounds. As a matter of fact, [l] and [r] share no major acoustic features with other coronals. In cases in which raising the blade or the body of the tongue is distinctive, what could be significant is the kind of constriction that is formed and not what is forming it. In those cases, we could perhaps dispense with the feature [coronal] and look for other features that describe the manner of the constriction, such as the features [distributed] and [strident]. The feature [strident] is, however, an acoustic feature; in fact, it is the only acoustic feature in the Chomsky and Halle system. The whole problem of describing the manner of consonant production in terms of distinctive features is still under study.

The discussion above advocates the introduction of a new distinctive feature [labial]. The feature [coronal] would either remain as a secondary feature or, as suggested, could be replaced by features that describe different types of constrictions in consonant production. More research is necessary before any definitive statements could be made in support of this proposal. Furthermore, the linguistic significance of such a proposal should be investigated.

APPENDIX B

FFT vs Filter Bank

Some of the characteristics of the FFT that was implemented in this work will be outlined and followed by a comparison between FFT and filter-bank spectra.

The FFT algorithm used in the recognition scheme is of the frequency decimation type (see Cochran et al.[37]). Integer arithmetic with scaling is employed; the machine word length is 18 bits, and the accuracy was more than sufficient for our purposes. The fact that the time waveform is a real (noncomplex) function of time was exploited to save computation time. The time samples are multiplied by a window function before the FFT is computed. The energy spectrum is obtained by taking the square of the absolute value of the transform of the time samples.

B. 1 FFT OF A REAL SEQUENCE

Let $x_o, x_1, \ldots, x_{N-1}$ be a sequence of N real time samples whose discrete Fourier transform (DFT) is desired. The DFT is defined by

$$X_k = \sum_{n=0}^{N-1} x_n W^{kn}, \qquad k = 0, 1, \ldots, N-1, \tag{B. 1}$$

where X_k is a sequence of N frequency samples, and

$$W = e^{-j\frac{2\pi}{N}}.$$

If the time samples x_n are separated by τ seconds, then the frequency samples X_k occur at multiples of $f_o = 1/N\tau$. Since there are N frequency samples to be computed, the total frequency range is equal to $Nf_o = \frac{1}{\tau}$, which is equal to the sampling frequency. Since the sequence x_n is real, it can be shown in Eq. B. 1 that the sequence X_k is Hermitian symmetric; that is,

$$X_k = X^*_{N-k}, \tag{B. 2}$$

where the star denotes complex conjugate.

The energy spectrum is given by

$$E_k = X_k X_k^* = |X_k|^2. \tag{B. 3}$$

From Eq. B. 2 it is clear that

$$E_{N-k} = E_k \tag{B. 4}$$

and the sequence E_k is, therefore, real and even. So, it is sufficient to compute

E_k for $k = 0, 1, \ldots \frac{N}{2}$, which is the physically meaningful portion that represents the spectrum. It has a frequency range or bandwidth of $1/2\tau$ which is half the sampling frequency. If the sampling frequency is equal to or greater than the Nyquist rate, then the frequency range $1/2\tau$ covers the relevant range for the spectrum.

Let the sequence of N samples x_n be divided into two equal sequences y_n and z_n, where y_n is the collection of even-numbered samples $(x_o, x_2, x_4, \ldots)$, and z_n is the collection of odd-numbered samples $(x_1, x_3, x_5, \ldots)$. This can be written formally

$$\left.\begin{aligned} y_n &= x_{2n} \\ z_n &= x_{2n+1} \end{aligned}\right\} \qquad n = 0, 1, \ldots, \frac{N}{2} - 1. \tag{B.5}$$

Equation B. 1 can be rewritten in terms of y_n and z_n as follows:

$$X_k = \sum_{n=0}^{\frac{N}{2}-1} \left[y_n e^{-j\frac{2\pi k}{N} 2n} + z_n e^{-j\frac{2\pi k}{N}(2n+1)} \right],$$

$$X_k = \sum_{n=0}^{\frac{N}{2}-1} \left[y_n e^{-j\frac{2\pi}{N/2} nk} \right] + e^{-j\frac{2\pi k}{N}} \sum_{n=0}^{\frac{N}{2}-1} \left[z_n e^{-j\frac{2\pi}{N/2} nk} \right]$$

$$X_k = Y_k + W^k Z_k; \qquad k = 0, 1, \ldots, N-1, \tag{B.6}$$

where Y_k and z_k are the $\frac{N}{2}$-point DFT's of y_n and z_n, respectively. Now we shall make use of the fact that y_n and z_n are real sequences.

Define another sequence c_n of N/2 complex samples as

$$c_n = y_n + jz_n, \qquad n = 0, 1, \ldots, \frac{N}{2} - 1. \tag{B.7}$$

Then

$$C_k = Y_k + jZ_k, \qquad k = 0, 1, \ldots, \frac{N}{2} - 1, \tag{B.8}$$

where C_k is the DFT of c_n. Since y_n and z_n are real samples, their DFT's are Hermitian symmetric.

$$Y_k = Y^*_{\frac{N}{2}-k}, \qquad Z_k = Z^*_{\frac{N}{2}-k}. \tag{B.9}$$

Now, from Eq. B. 8 we have

$$C^*_{\frac{N}{2}-k} = Y^*_{\frac{N}{2}-k} - jZ^*_{\frac{N}{2}-k}. \qquad \text{(B. 10)}$$

Substituting Eq. B. 9 in Eq. B. 10, we obtain

$$C^*_{\frac{N}{2}-k} = Y_k - jZ_k. \qquad \text{(B. 11)}$$

Equations B. 8 and B. 11 can be solved for Y_k and Z_k. Substituting the result in Eq. B.6, we obtain

$$X_k = \frac{1}{2}\left[C_k + C^*_{\frac{N}{2}-k} - jW^k\left(C_k - C^*_{\frac{N}{2}-k}\right)\right], \qquad k = 0, 1, \ldots, \frac{N}{2}. \qquad \text{(B. 12)}$$

(X_k need be computed only for the given range because x_n is a real sequence.) The sequence X_k can be computed by first computing the $\frac{N}{2}$-point DFT of the sequence c_n using the FFT algorithm, and then applying Eq. B. 12 to the result. This method of computing the DFT of a real sequence results in a saving of approximately half the time from an N-point FFT for a large N. The flow graph for computing the FFT has been well documented (see Cochran et al.[37]). There remain the computations in Eq. B. 12; they can be computed using the flow graph in Fig. B-1. The plus and minus signs indicate

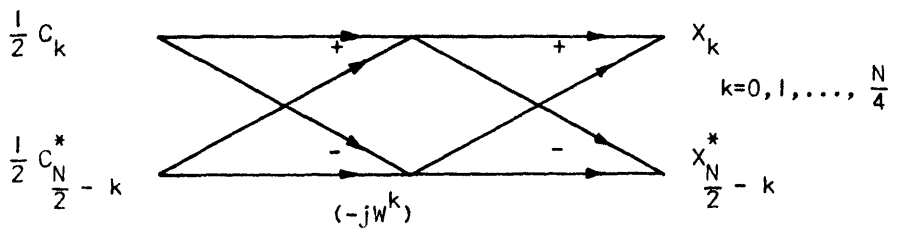


Fig. B-1. Supplementary flow graph in computing the FFT of a sequence of real samples.

addition and subtraction of the elements from the place where the arrows emanate. The term $(-jW^k)$ is multiplied at the node after subtraction. Note that W is still defined by $\exp\left(-j\frac{2\pi}{N}\right)$. If the flow graph of Fig. B. 1 is to be implemented in a computer program, it is best to compute the values for k = 0 separately from the others. Since $C_{N/2} = C_o$, it can be shown that X_o and $X_{N/2}$ are both real and are given by the following equations;

$$X_o = \text{Re}\,(C_o) + \text{Im}\,(C_o)$$
$$X_{N/2} = \text{Re}\,(C_o) - \text{Im}\,(C_o), \qquad \text{(B. 13)}$$

where Re and Im represent real and imaginary parts, respectively.

B. 2 WINDOW FUNCTIONS

The short-time complex spectrum $X(\omega, t)$ of a continuous time signal $x(t)$ can be viewed as the Fourier transform of $x(t)$ as seen through a weighting window function $h(t)$ at a given instant of time. This can be expressed formally as

$$X(\omega, t) = \int_{-\infty}^{\infty} x(\gamma)\, h(t-\gamma)\, e^{-j\omega\gamma}\, d\gamma, \tag{B. 14}$$

where $X(\omega, t)$ can be computed at different instants of time by appropriately shifting the window function. Presumably, $h(t)$ is effective only within a "short" time interval. Equation B. 14 can be rewritten

$$\begin{aligned} X(\omega, t) &= e^{-j\omega t} \int_{-\infty}^{\infty} x(t-\gamma)\, h(\gamma)\, e^{j\omega\gamma}\, d\gamma \\ &= e^{-j\omega t} \left[x(t) \circledast h(t)\, e^{j\omega t}\right] \end{aligned} \tag{B. 15}$$

where $\circledast$ represents convolution. The short-time amplitude spectrum $|X(\omega, t)|$ is then given by the two equivalent formulations:

$$|X(\omega, t)| = \left|x(t)\, e^{-j\omega t} \circledast h(t)\right| \tag{B. 16}$$

$$|X(\omega, t)| = \left|x(t) \circledast h(t)\, e^{j\omega t}\right|. \tag{B. 17}$$

Equation B. 17 can be interpreted as the absolute output of a filter whose input is $x(t)$ and whose impulse response is $h(t)\, e^{j\omega t}$. For a specific frequency $\omega = \omega_o$, the transfer function corresponding to $h(t)\, e^{j\omega_o t}$ is equal to $H(\omega-\omega_o)$, which is $H(\omega)$ centered around ω_o. If $H(\omega)$ is a narrow bandpass filter, then $|X(\omega_o, t)|^2$ gives the approximate energy content of $x(t)$ in the neighborhood of ω_o. The exact form of $h(t)$ is important and will now be discussed.

Since we are interested in computing the short-time spectrum using a computer, $h(t)$ must necessarily be time-limited. Let us also assume that $h(t)$ is an even function of time. The last two conditions can be written

$$\begin{aligned} &h(t) = 0, \quad |t| > T \\ &h(t) = h(-t), \end{aligned} \tag{B. 18}$$

where T is some constant, and $2T$ is the window size. Then, Eq. B. 14 takes the form

$$X(\omega, t) = \int_{t-T}^{t+T} x(\gamma)\, h(\gamma-t)\, e^{-j\omega\gamma}\, d\gamma. \tag{B. 19}$$

For any particular instant of time $t = t_o$, the time coordinate can be redefined so that $t_o = 0$. Keeping this in mind, we can substitute $t = 0$ in Eq. B.19. A change in the integration variable gives the result

$$X(\omega) = \int_{-T}^{T} x(t)\, h(t)\, e^{-j\omega t}\, dt. \tag{B.20}$$

Although the time-dependence has been removed from $X(\omega, t)$ in Eq. B.20, it is still implicitly existent. With proper sampling of x(t) and h(t) the Fourier integral can be converted into a DFT of two real sequences x_n and h_n:

$$X_k = \sum_{n=0}^{N-1} x_n h_n W^{kn} \tag{B.21}$$

The computation in Eq. B.21 can be performed on the computer by using the method for real sequences that was described in the previous section.

Recall that the derivation of Eqs. B.20 and B.21 was based on the assumption that h(t) is an even and time-limited function (see Eq. B.18). It is obvious that beyond this assumption the actual shape of h(t) can markedly affect the resultant $X(\omega)$ or X_k. A discussion of some window functions follows.

The simplest (perhaps trivial) window function that could be used is the rectangular window shown in Fig. B-2a, and defined by

$$h_1(t) = \frac{1}{2T}\, u_{-1}\left(1 - \frac{|t|}{T}\right) \tag{B.22}$$

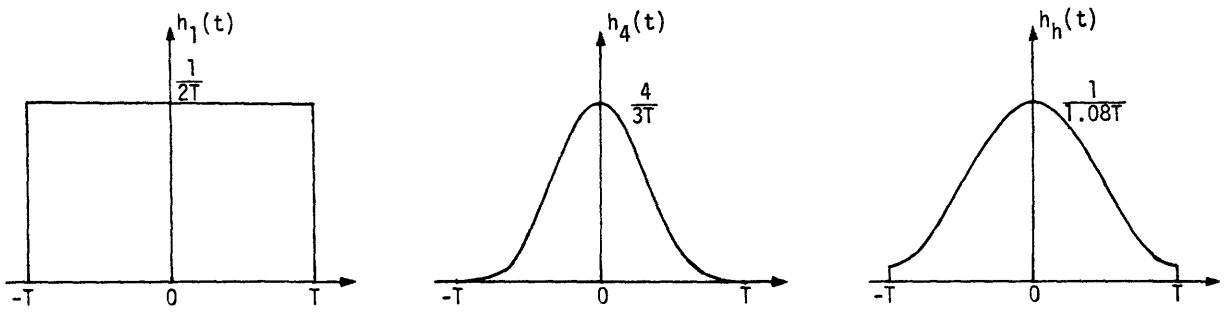
where $u_{-1}(t)$ is the unit-step function. $H_1(\omega)$ is given by

$$H_1(\omega) = \frac{\sin T\omega}{T\omega}. \tag{B.23}$$

The first sidelobe of $H_1(\omega)$ is only -13.3 dB from the major lobe, and the high-frequency roll-off is only -6 dB/octave. Therefore, $H_1(\omega)$ is a poor bandpass filter; it suffers from the "leakage" problem. The result is a marked lack of frequency resolution of the computed spectrum.

One method of improving on $H_1(\omega)$ is to raise it to some integer power n. The corresponding impulse response is an (n-1)-fold convolution of $h_1(t)$, which means that the resultant h(t) has the property $h(t) = 0$, $|t| > nT$. In order to keep the window width equal to 2T instead of 2nT, the frequency ω must be divided by n. The result is

$$H_n(\omega) = \left[\frac{\sin (T\omega/n)}{T\omega/n}\right]^n \tag{B.24}$$



$$h_1(t) = \frac{1}{2T}\, u_{-1}\left(1 - \frac{|t|}{T}\right)$$

$$h_4(t) = \frac{8}{3T}\left[\left(1 - \frac{|t|}{T}\right)^3 u_{-1}\left(1 - \frac{|t|}{T}\right) - 4\left(\frac{1}{2} - \frac{|t|}{T}\right)^3 u_{-1}\left(\frac{1}{2} - \frac{|t|}{T}\right)\right]$$

$$h_h(t) = \frac{1}{1.08T}\left(0.54 + 0.46 \cos\frac{\pi t}{T}\right) u_{-1}\left(1 - \frac{|t|}{T}\right)$$

$$H_1(f) = \frac{\sin 2\pi fT}{2\pi fT} \quad ; \quad H_4(f) = \left[\frac{\sin(2\pi fT/4)}{2\pi fT/4}\right]^4$$

$$H_h(f) = H_1(f) + \frac{0.23}{0.54}\left[H_1(f+f_0) + H_1(f-f_0)\right], \text{ where } f_0 = \frac{1}{2T}$$

Fig. B-2. Three window functions. (a) Rectangular window. (b) Parzen window. (c) Hamming window.

$H_n(\omega)$ has the property that the first sidelobe level is -13.3 n dB below the major lobe, and the high-frequency roll-off is -6 n dB/octave. It would seem then that as n increases, the result is less leakage and better frequency resolution. As n increases, however, so does the bandwidth of $H_n(\omega)$. A compromise has to be reached depending on the specific application. $H_2(\omega)$, where n = 2, corresponds to the well-known triangular or Barlett window, which has been used at times. A much superior window for speech applications is $H_4(\omega)$, which is known as the "Parzen window." $h_n(t)$, the Fourier transform of $H_n(\omega)$ in Eq. B.24, can be shown to be equal (Makhoul[76]) to

$$h_n(t) = \frac{(n/2)^n}{T(n-1)!} \sum_{k=0}^{\left[\frac{n-1}{2}\right]} (-1)^k \binom{n}{k} \left(1 - \frac{2k}{n} - \frac{|t|}{T}\right)^{n-1} u_{-1}\left(1 - \frac{2k}{n} - \frac{|t|}{T}\right) \qquad \text{(B.25)}$$

where n is any positive integer,

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

and

$$\left[\frac{n-1}{2}\right] \equiv \text{integer portion of } \frac{n-1}{2}.$$

The Parzen window, $h_4(t)$, is shown in Fig. B-2b.

A different type of window function which has been quite popular is the Hamming window. It is shown in Fig. B-2c.

All three window functions shown in Fig. B-2 are normalized so that

$$\int_{-T}^{T} h(t)\, dt = H(0) = 1. \tag{B.26}$$

Some spectral characteristics are shown in Fig. B-3. Figure B-3a is a linear plot of the major lobe for all three windows. Other characteristics are shown in Fig. B-3b. For a logarithmic plot of $H_h(\omega)$, the Hamming window transform, see Blackman and Tukey,[77] where other similar window functions may be found. All three windows were available in the computer program that computed the FFT in the recognition algorithm. The rectangular window was never actually used because of its undesirable characteristics. The Parzen and Hamming windows were both used in the analysis. Experimentation with sinusoids and speech sounds showed that the Parzen window gave somewhat better results than the Hamming window. The requisite cosines for the computation of the Hamming window were already available in the computer memory because they were also used in the FFT computation. The result was that the Hamming-window computations took less time than the Parzen-window computations. Because of time limitations the Hamming window was finally used in the recognition scheme. Had there been no memory space limitations, the window coefficients could have been stored directly in the computer memory. In that case, I would have chosen the Parzen window.

B.3 FFT vs FILTER BANK

We may now conclude that the FFT of the product of the time waveform and a window function h(t) of width 2T is equivalent to a filtering operation with N filters whose center frequencies are multiples of $f_o = 1/2T$ but otherwise have identical spectral characteristics described by $H(\omega)$, the Fourier transform of the window h(t). The nominal averaging time for these filters is equal to 2T. The "effective" averaging time, however, is a function of the specific window; it is inversely proportional to the filter bandwidth. For any specific window the averaging time for each of the N filters is the same. The FFT is reminiscent of the spectrograph in the analogue domain where the characteristics of the filter and the time averaging, as performed by a separate lowpass filter, are fixed for all frequencies.

In order to facilitate comparison between the FFT and filter-bank characteristics, a formal definition of averaging time will be given. Intuitively, it seems reasonable to assume that the averaging time for a rectangular window should equal the window width 2T. By setting that as a standard, the averaging time for any bandpass filter is given by

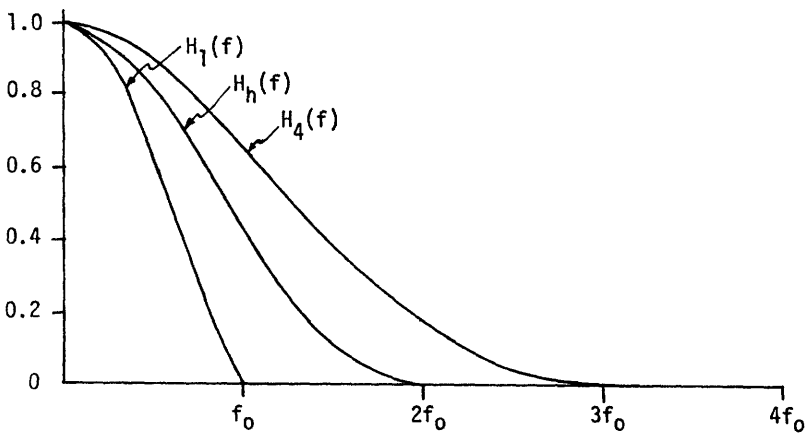
$$T_A = 2T \frac{W_1}{W}, \tag{B.27}$$

where W_1 is the bandwidth of the rectangular window filter, and W is the bandwidth of the filter whose averaging time T_A is desired. From Fig. B-3b, $W_1 = 0.9/2T$. Substituting in Eq. B. 27, we obtain

$$T_A = \frac{0.9}{W}, \tag{B.28}$$

where W is measured in Hz and T_A in sec.

In the final recognition algorithm the FFT employed N = 512 samples, sampled at intervals of 50 μs (sampling rate = 20 kHz), resulting in a window width 2T = 25.6 ms. The spectrum is computed at 256 points in frequency, ranging from 0-10 kHz and separated by $f_1 = 10^3/25.6 = 39.1$ Hz. The Hamming window filter has a bandwidth = 52 Hz (see Fig. B-3b) and the averaging time, as computed by Eq. B. 28, is equal to 17.3 ms.



(a) FREQUENCY f - ($f_0 = \frac{1}{2T}$)

WINDOW	3-dB BANDWIDTH	LARGEST SIDELOBE RELATIVE TO MAIN LOBE	HIGH-FREQUENCY ROLL-OFF
RECTANGULAR	0.9 f_0	-13.3 dB (1st sidelobe)	-6 dB/oct
PARZEN	1.8 f_0	-53.1 dB (1st sidelobe)	-24 dB/oct
HAMMING	1.33 f_0	-43 dB (4th sidelobe)	-6 dB/oct

(b)

Fig. B-3. Spectral characteristics of the three window functions in Fig. B-2. (Only the main lobes are shown.)

The filter bank has been described in section 3.3. The filters are linearly spaced in the range 150-1650 Hz, with a constant bandwidth of 100 Hz, and logarithmically spaced in the range 1650-7025 Hz, with a linearly increasing bandwidth (see Fig. 5). Each filter is followed by a rectifier and a single-pole RC lowpass filter with a time constant of 10 ms. Considering the lowpass filter as a bandpass filter, it has a bandwidth of $\frac{2*10^3}{2\pi*10 \text{ ms}} = 31.8$ Hz. The corresponding averaging time, by Eq. B.28, is 28.3 ms, roughly three times the time constant. (Note that the averaging time for the lowpass filter is much larger than and, hence dominates the effective averaging time of all bandpass filters of the filter band.) Compare 28.3 ms, the averaging time for the filter bank, with 17.3 ms, the averaging time for the Hamming window filter. It is not surprising that the FFT-computed spectrum was much more sensitive to variations in time than the filter-bank spectrum. A major difference between the filter bank and the FFT is that the averaging time for the filter bank is determined by a circuit that is completely independent of the filters, while the averaging time for the FFT is determined by the window function, which also specifies the filter characteristics. The center frequencies and bandwidths of the filters in the filter bank are designed quite independently, while the FFT window function and its width completely specify the filter characteristics and the averaging time. There is an independence of design in the filter bank which does not exist in the FFT. Nevertheless, this difference, to a large extent, could be rectified. The FFT is ordinarily computed at a relatively large number of frequencies. So, the computed spectral energies E_k (Eq. B.3) at adjacent frequencies could be averaged to simulate a bank of filters, where each is an equivalent filter with one center frequency and one equivalent bandwidth. This method was actually used in the recognition system in order to have identical representations of the filter bank and FFT spectra. The average energy E_A from n adjacent FFT filters, whose center frequencies are separated by f_o and whose bandwidth is W Hz each, is defined by

$$E_A = \frac{1}{n}(E_{i+1} + E_{i+2} + \ldots + E_{i+n}). \tag{B.29}$$

The equivalent filter has a center frequency

$$f_c = \left(i + \frac{n}{2}\right) f_o, \tag{B.30}$$

and an equivalent bandwidth

$$W_e = (n-1)f_o + W. \tag{B.31}$$

By this method, the FFT is rendered quite flexible, and different groupings of filters could be investigated in order to obtain the desired representation. As an example, see the filter-bank spectrum in Fig. 6 and the FFT-computed spectrum (after appropriate regroupings of energies to simulate a bank of 36 filters with the

center frequencies approximately equal to those of the hardware filter bank) in Fig. 7 of [z] in [əzid]. In this manner it was possible to critically compare both spectra.

A similar method could be used to change the averaging times after the FFT spectrum is computed. There is, of course, a minimum averaging time that is dictated by the window function size. Beyond that, one could perform a further computation that would average the energy at any frequency over a longer period of time by introducing a digital lowpass filter.

Therefore, potentially, the FFT-derived spectrum is quite flexible. What is required is to compute the spectrum at relatively many frequencies in the desired frequency range, and to employ a window function whose averaging time is the minimum desired. The computed spectrum could then be frequency- or time-averaged to obtain the desired representation. A major problem, or course, is the extra computation time involved.

Acknowledgment

I wish to thank Professor Murray Eden for supervising my thesis, and for his criticism and encouragement. I am grateful to Professor Kenneth N. Stevens for many enlightening discussions and valuable suggestions throughout the course of this study, and also to Professor Dennis H. Klatt for many valuable comments before and during the preparation of this manuscript. I am indebted to Professor Samuel J. Mason for suggesting this problem as a thesis topic. I also wish to thank Professor Morris Halle for the discussions that we had, especially on distinctive features, and for agreeing to be a reader at the last minute.

I would especially like to thank the Speech Communications Group of the Research Laboratory of Electronics for allowing me the use of its computer and other facilities, without which this study could not have been made. The members of the Speech Communications Group were always generous with their time; in particular, I benefited from many discussions on speech recognition with my colleague Mark F. Medress.

I would also like to thank Joan M. Maling, of the Linguistics Group of the Research Laboratory of Electronics, for her help in the preparation of this manuscript.

References

1. S. J. Keyser and M. Halle, "What We Do When We Speak," in Recognizing Patterns, P. A. Kolers and M. Eden (eds.) (The M. I. T. Press, Cambridge, Mass., 1968), pp. 64-80.
2. M. Halle, and K. Stevens, "Speech Recognition: A Model and a Program for Research," IRE Trans. on Information Theory, Vol. IT-8, No. 2, 155-159, February 1962.
3. A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the Speech Code," Psychol. Rev., Vol. 74, No. 6, pp. 431-461, November 1967.
4. K. N. Stevens and A. S. House, "Speech Perception," in Foundations of Modern Auditory Theory, J. Tobias and E. Schubert (eds.) (in press).
5. N. Lindgren, "Automatic Speech Recognition," IEEE Spectrum, pp. 114-136, March 1965.
6. N. Lindgren, "Theoretical Models of Speech Perception and Language," IEEE Spectrum, pp. 44-59, April 1965.
7. J. L. Flanagan, Speech Analysis Synthesis and Perception (Academic Press, Inc., New York, 1965), see pp. 158-164.
8. S. R. Hyde, "Automatic Speech Recognition Literature Survey and Discussion," Research Report No. 45, Post Office Research Department, Dollis Hill, London, 30 September 1968.
9. D. G. Bobrow and D. H. Klatt, "A Limited Speech Recognition System," BBN Report No. 1667, Final Report, Contract No. NAS 12-138, NASA, Cambridge, Mass., 15 May 1968; also Proc. FJCC 1968, pp. 305-317.
10. M. F. Medress, "Computer Recognition of Single-Syllable English Words," Ph. D. Thesis, Massachusetts Institute of Technology, September 1969.
11. K. N. Stevens, "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," in Human Communication: A Unified View, E. G. Davis, Jr. and P. B. Denes (eds.) (in press).
12. P. A. Kolers and M. Eden (eds.), Recognizing Patterns (The M. I. T. Press, Cambridge, Mass., 1968), Preface, p. ix.
13. G. Nagy, "State of the Art in Pattern Recognition," Proc. IEEE 56, 836-862 (1968).
14. C. Cherry, On Human Communication (The M. I. T. Press, Cambridge, Mass., 2d Edition, 1966); see p. 299.
15. J. L. Flanagan, op. cit., p. 7.
16. J. E. Karlin and S. N. Alexander, "Communication between Man and Machine," Proc. IRE 50, 1124-1128 (1962).
17. R. Jakobson, C. G. M. Fant, and M. Halle, Preliminaries to Speech Analysis (The M. I. T. Press, Cambridge, Mass., 1952, revised edition, 1963).
18. N. Chomsky and M. Halle, The Sound Pattern of English (Harper and Row, New York, 1968).
19. W. A. Wickelgren, "Distinctive Features and Errors in Short-Term Memory for English Consonants," J. Acoust. Soc. Am. 39, pp. 388-398 (1966).
20. J. Wiren and H. L. Stubbs, "Electronic Binary Selection System for Phoneme Classification," J. Acoust. Soc. Am. 28, 1082-1091 (1956).
21. T. B. Martin, A. L. Nelson, and H. J. Zadell, "Speech Recognition by Feature Abstraction Techniques," Wright-Patterson AFB Avionics Laboratories Report AL-TDR 64-176, AD-604 526, August 1964.

22. T. B. Martin, A. L. Nelson, and H. J. Zadell, "Speech Recognition by Feature Abstraction Techniques," Report AFAL-TR-65-317, AD-476317, December 1965.

23. G. W. Hughes and J. F. Hemdal, "Speech Analysis," Final Report, AF 19(628)-305, TR-EE 65-9, AFCRL-65-681, 1965. Prepared for AFCRL, Office of Aerospace Research, U. S. A. F., Bedford, Mass.

24. B. Gold, "Word-Recognition Computer Program," Technical Report 452, Research Laboratory of Electronics, M. I. T., Cambridge, Mass., June 15, 1966.

25. D. R. Reddy, "Computer Recognition of Connected Speech," J. Acoust. Soc. Am. 42, 329-347 (1967).

26. D. G. Bobrow, A. K. Hartley, and D. H. Klatt, "A Limited Speech Recognition System II," BBN Report No. 1819, Final Report, Contract No. NAS 12-138, NASA, Cambridge, Mass., 1 April 1969.

27. N. Chomsky and M. Halle, op. cit., p. 176.

28. S. E. G. Öhman, "Coarticulation in VCV Utterances: Spectrographic Measurements," J. Acoust. Soc. Am. 39, 151-168 (1966).

29. B. Lindblom, "Spectrographic Study of Vowel Reduction," J. Acoust. Soc. Am. 35, 1773-1781 (1963).

30. B. E. F. Lindblom and M. Studdert-Kennedy, "On the Role of Formant Transitions in Vowel Recognition," J. Acoust. Soc. Am. 42, 830-843 (1967).

31. K. N. Stevens and A. S. House, "Perturbation of Vowel Articulations by Consonantal Context: An Acoustical Study," J. Speech Hearing Res., Vol. 6, No. 2, pp. 111-128, June 1963.

32. K. N. Stevens, A. S. House, and A. P. Paul, "Acoustical Description of Syllabic Nuclei: An Interpretation in Terms of a Dynamic Model of Articulation," J. Acoust. Soc. Am. 40, 123-132 (1966).

33. G. A. Miller and P. E. Nicely, "An Analysis of Perceptual Confusions among English Consonants," J. Acoust. Soc. Am. 27, 338-352 (1955).

34. J. W. Glenn and N. Kleiner, "Speaker Identification Based on Nasal Phonation," J. Acoust. Soc. Am. 43, 368-372 (1968).

35. J. J. Wolf, "Acoustic Measurements for Speaker Recognition," Ph. D. Thesis, Massachusetts Institute of Technology, September 1969.

36. W. L. Henke, "Speech Computer Facility," Quarterly Progress Report No. 90, Research Laboratory of Electronics, M. I. T., Cambridge, Mass., July 15, 1968, pp. 217-219.

37. W. T. Cochran et al. (G-AE Subcommittee on Measurement Concepts), "What Is the Fast Fourier Transform?," IEEE Trans. on Audio and Electroacoustics, Vol. Au-15, No. 2, pp. 45-55, June 1967.

38. G. E. Peterson and H. L. Barney, "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am. 24, 175-184 (1952).

39. L. J. Gerstman, "Classification of Self-Normalized Vowels," Conference on Speech Communication and Processing, 6-8 November 1967, Conference Reprints, pp. 97-100.

40. H. Suzuki, H. Kasuya, and K. Kido, "The Acoustic Parameters for Vowel Recognition without Distinction of Speakers," Conference on Speech Communication and Processing, 6-8 November 1967, Conference Reprints, pp. 92-96.

41. J. L. Flanagan, op. cit., pp. 139-140.

42. P. Ladefoged, Three Areas of Experimental Phonetics (Oxford University Press, London, 1967), see pp. 79-91.

43. R. K. Potter, G. A. Kopp, and H. C. Green, Visible Speech (D. Van Nostrand Co., Inc., Princeton, N. J., 1947).

44. G. Fant, "On the Predictability of Formant Levels and Spectrum Envelopes from Formant Frequencies," in M. Halle (ed.), For Roman Jakobson (The Hague: Mouton and Co., 1956), pp. 109-120.

45. G. Fant, Acoustic Theory of Speech Production (Mouton and Co., 's-Gravenhage, The Netherlands, 1960).

46. O. Fujimura, "On the Second Spectral Peak of Front Vowels: A Perceptual Study of the Role of the Second and Third Formants," Language and Speech, Vol. 10, Part 3, pp. 181-193, 1967.

47. D. H. Klatt, Private communication, 1970.

48. G. W. Hughes and M. Halle, "Spectral Properties of Fricative Consonants," J. Acoust. Soc. Am. 28, 303-310 (1956).

49. P. Strevens, "Spectra of Fricative Noise in Human Speech," Language and Speech, Vol. 3, pp. 32-49, 1960.

50. J. M. Heinz and K. N. Stevens, "On the Properties of Voiceless Fricative Consonants," J. Acoust. Soc. Am. 33, 589-596 (1961).

51. K. N. Stevens, "Acoustic Correlates of Place of Articulation for Stop and Fricative Consonants," Quarterly Progress Report No. 89, Research Laboratory of Electronics, M. I. T., Cambridge, Mass., April 15, 1968, pp. 199-205.

52. E. Fischer-Jorgensen, "Acoustic Analysis of Stop Consonants," Miscellanea Phonetica 11, 42-59 (1954).

53. M. Halle, G. W. Hughes, and J. Radley, "Acoustic Properties of Stop Consonants," J. Acoust. Soc. Am. 29, 107-116 (1957).

54. F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some Experiments on the Perception of Synthetic Speech Sounds," J. Acoust. Soc. Am. 24, 597-606 (1952).

55. A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, "The Role of Consonant-Vowel Transitions in the Stop and Nasal Consonants," Psychol. Monographs 68, No. 379, 1954.

56. K. S. Harris, H. S. Hoffman, A. M. Liberman, P. C. Delattre, and F. S. Cooper, "Effect of Third-Formant Transitions on the Perception of the Voiced Stop Consonants," J. Acoust. Soc. Am. 30, 122-126 (1958).

57. H. S. Hoffman, "Study of Some Cues in the Perception of the Voiced Stop Consonants," J. Acoust. Soc. Am. 30, 1035-1041 (1958).

58. T. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic Loci and Transitional Cues for Consonants," J. Acoust. Soc. Am. 27, 769-773 (1955).

59. K. N. Stevens, "Acoustic Correlates of Certain Consonantal Features," Conference on Speech Communication and Processing, 6-8 November 1967, Conference Reprints, pp. 177-184.

60. R. Jakobson and M. Halle, Fundamentals of Language (Mouton and Co., The Hague, 1956).

61. C. D. Schatz, "The Role of Context in the Perception of Stops," Language, Vol. 30, No. 1, pp. 47-56, 1954.

62. K. S. Harris, "Cues for the Discrimination of American English Fricatives in Spoken Syllables," Language and Speech, Vol. 1, No. 1, pp. 1-7, 1958.

63. C. G. M. Fant, "Descriptive Analysis of the Acoustic Aspects of Speech," LOGOS, Vol. 5, No. 1, pp. 3-17, April 1962.

64. L. Dolansky, "Choice of Base Signals in Speech Signal Analysis," IRE Trans. on Audio, Vol. AU-8, pp. 221-229, November-December 1960.

65 L. R. Rabiner, R. W. Schafer, and C. H. Rader, "The Chirp z-Transform Algorithm," IEEE Trans. Audio and Electroacoustics, Vol. AU-17, No. 2, pp. 86-92, June 1969.

66. H. F. Harmuth, "Applications of Walsh Functions in Communications," IEEE Spectrum, pp. 82-91, November 1969.

67. L. Collatz, Functional Analysis and Numerical Mathematics (Academic Press, Inc., New York, 1966), pp. 73-74.

68. R. Jakobson, Child Language, Aphasia and Phonological Universals (Mouton: The Hague, Paris, 1968).

69. N. Chomsky and M. Halle, op. cit., p. 307.

70. R. Jacobson, op. cit., pp. 47-48.

71. Ibid., p. 30.

72. Ibid., p. 87.

73. N. Chomsky and M. Halle, op. cit., p. 304.

74. M. Halle, Private communication, 1970.

75. N. Chomsky and M. Halle, op. cit., p. 329.

76. J. I. Makhoul, "SINC^n – A Family of Window Functions," Quarterly Progress Report No. 97, Research Laboratory of Electronics, M. I. T., Cambridge, Mass., April 15, 1970, pp. 145-150.

77. R. B. Blackman and J. W. Tukey, The Measurement of Power Spectra (Dover Publications, Inc., New York, 1958). p. 97.

JOINT SERVICES ELECTRONICS PROGRAM
REPORTS DISTRIBUTION LIST

Department of Defense

Assistant Director (Research)
Office of Director of Defense Research
& Engineering
Pentagon, Rm 3C128
Washington, D. C. 20301

Technical Library
DDR&E
Room 3C-122, The Pentagon
Washington, D.C. 20301

Director For Materials Sciences
Advanced Research Projects Agency
Room 3D179, Pentagon
Washington, D.C. 20301

Chief, R&D Division (340)
Defense Communications Agency
Washington, D.C. 20305

Defense Documentation Center
Attn: DDC-TCA
Cameron Station
Alexandria, Virginia 22314

Dr. Alvin D. Schnitzler
Institute For Defense Analyses
Science and Technology Division
400 Army-Navy Drive
Arlington, Virginia 22202

Central Intelligence Agency
Attn: CRS/ADD/PUBLICATIONS
Washington, D.C. 20505

M. A. Rothenberg (STEPD-SC(S))
Scientific Director
Deseret Test Center
Bldg 100, Soldiers' Circle
Fort Douglas, Utah 84113

Department of the Air Force

Hq USAF (AFRDDD)
The Pentagon
Washington, D.C. 20330

Hq USAF (AFRDDG)
The Pentagon
Washington, D.C. 20330

Hq USAF (AFRDSD)
The Pentagon
Washington, D.C. 20330
Attn: LTC C. M. Waespy

Colonel E. P. Gaines, Jr.
ESD (MCD)
L. G. Hanscom Field
Bedford, Massachusetts 01730

Dr L. A. Wood, Director
Electronic and Solid State Sciences
Air Force Office of Scientific Research
1400 Wilson Boulevard
Arlington, Virginia 22209

Dr Harvey E. Savely, Director
Life Sciences
Air Force Office of Scientific Research
1400 Wilson Boulevard
Arlington, Virginia 22209

Mr I. R. Mirman
Hq AFSC (SGGP)
Andrews Air Force Base,
Washington, D.C. 20331

Rome Air Development Center
Attn: Documents Library (EMTLD)
Griffiss Air Force Base, New York 13440

Mr H. E. Webb, Jr (EMBIS)
Rome Air Development Center
Griffiss Air Force Base, New York 13440

Dr L. M. Hollingsworth
AFCRL (CRN)
L. G. Hanscom Field
Bedford, Massachusetts 01730

Hq ESD (ESTI)
L. G. Hanscom Field
Bedford, Massachusetts 01730

Professor R. E. Fontana, Head
Dept of Electrical Engineering
Air Force Institute of Technology
Wright-Patterson Air Force Base,
Ohio 45433

AFAL (AVT) Dr H. V. Noble, Chief
Electronics Technology Division
Air Force Avionics Laboratory
Wright-Patterson Air Force Base,
Ohio 45433

Director
Air Force Avionics Laboratory
Wright-Patterson Air Force Base,
Ohio 45433

AFAL (AVTA/R. D. Larson)
Wright-Patterson Air Force Base,
Ohio 45433

Director of Faculty Research
Department of the Air Force
U.S. Air Force Academy
Colorado 80840

Mr Jules I. Wittebort
Chief, Electronics Branch
Manufacturing Technology Division
AFAL/LTE
Wright-Patterson Air Force Base,
Ohio 45433

Academy Library (DFSLB)
USAF Academy, Colorado 80840

Director of Aerospace Mechanics Sciences
Frank J. Seiler Research Laboratory (OAR)
USAF Academy, Colorado 80840

Major Richard J. Gowen
Tenure Associate Professor
Dept of Electrical Engineering
USAF Academy, Colorado 80840

Director, USAF PROJECT RAND
Via: Air Force Liaison Office
The RAND Corporation
Attn: Library D
1700 Main Street
Santa Monica, California 90406

Hq SAMSO (SMTAE/Lt Belate)
Air Force Unit Post Office
Los Angeles, California 90045

AUL3T-9663
Maxwell Air Force Base, Alabama 36112

AFETR Technical Library
(ETV, MU-135)
Patrick Air Force Base, Florida 32925

ADTC (ADBPS-12)
Eglin Air Force Base, Florida 32542

Mr B. R. Locke
Technical Adviser, Requirements
USAF Security Service
Kelly Air Force Base, Texas 78241

Hq AMD (AMR)
Brooks Air Force Base, Texas 78235

USAFSAM (SMKOR)
Brooks Air Force Base, Texas 78235

Commanding General
Attn: STEWS-RE-L, Technical Library
White Sands Missile Range,
New Mexico 88002

Hq AEDC (AETS)
Arnold Air Force Station, Tennessee 37389

European Office of Aerospace Research
Technical Information Office
Box 14, FPO New York 09510

Electromagnetic Compatibility Analysis
Center (ECAC) ATTN: ACOAT
North Severn
Annapolis, Maryland 21402

VELA Seismological Center
312 Montgomery Street
Alexandria, Virginia 22314

Capt C. E. Baum
AFWL (WLRE)
Kirtland Air Force Base, New Mexico 87117

Dr Billy Welch
USAFSAM (SMC)
Brooks Air Force Base, Texas 78235

Department of the Army

Director
Physical & Engineering Sciences Division
3045 Columbia Pike
Arlington, Virginia 22204

Commanding General
U.S. Army Security Agency
Attn: IARD-T
Arlington Hall Station
Arlington, Virginia 22212

Commanding General
U.S. Army Materiel Command
Attn: AMCRD-TP
Washington, D.C. 20315

Director
U.S. Army Advanced Materiel
Concepts Agency
2461 Eisenhower Avenue
Alexandria, Virginia 22314

Commanding General
USACDC Institute of Land Combat
Attn: Technical Library, Rm 636
2461 Eisenhower Avenue
Alexandria, Virginia 22314

Mr H. T. Darracott (AMXAM-FT)
U.S. Army Advanced Materiel Concepts Agency
2461 Eisenhower Avenue
Alexandria, Virginia 22314

Commanding Officer
Harry Diamond Laboratories
Attn: Dr Berthold Altman (AMXDO-TI)
Connecticut Avenue and
Van Ness Street N. W.
Washington, D.C. 20438

Commanding Officer (AMXRO-BAT)
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground
Aberdeen, Maryland 21005

Technical Director
U.S. Army Land Warfare Laboratory
Aberdeen Proving Ground
Aberdeen, Maryland 21005

U.S. Army Munitions Command
Attn: Science & Technology Information
Branch, Bldg 59
Picatinny Arsenal, SMUPA-RT-S
Dover, New Jersey 07801

U.S. Army Mobility Equipment Research
and Development Center
Attn: Technical Documents Center, Bldg 315
Fort Belvoir, Virginia 22060

Commanding Officer
U.S. Army Engineer Topographic
Laboratories
Attn: STINFO Center
Fort Belvoir, Virginia 22060

Dr Herman Robl
Deputy Chief Scientist
U.S. Army Research Office (Durham)
Box CM, Duke Station
Durham, North Carolina 27706

Richard O. Ulsh (CRDARD-IP)
U.S. Army Research Office (Durham)
Box CM, Duke Station
Durham, North Carolina 27706

Technical Director (SMUFA-A2000-107-1)
Frankford Arsenal
Philadelphia, Pennsylvania 19137

Redstone Scientific Information Center
Attn: Chief, Document Section
U.S. Army Missile Command
Redstone Arsenal, Alabama 35809

Commanding General
U.S. Army Missile Command
Attn: AMSMI-RR
Redstone Arsenal, Alabama 35809

Commanding General
U.S. Army Strategic Communications
Command
Attn: SCC-ATS (Mr Peter B. Pichetto)
Fort Huachuca, Arizona 85613

Commanding Officer
Army Materials and Mechanics
Research Center
Attn: Dr H. Priest
Watertown Arsenal
Watertown, Massachusetts 02172

Commandant
U.S. Army Air Defense School
Attn: Missile Science Division, C&S Dept
P. O. Box 9390
Fort Bliss, Texas 79916

Commandant
U.S. Army Command and General
Staff College
Attn: Acquisitions, Lib Div
Fort Leavenworth, Kansas 66027

Dr H. K. Ziegler, Chief Scientist
Army Member TAC/JSEP (AMSEL-SC)
U.S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Mr I. A. Balton, AMSEL-XL-D
Executive Secretary, TAC/JSEP
U.S. Army Electronics Command
Fort Monmouth, New Jersey 07703

Director (NV-D)
Night Vision Laboratory, USAECOM
Fort Belvoir, Virginia 22060

Commanding Officer
Atmospheric Sciences Laboratory
U.S. Army Electronics Command
White Sands Missile Range,
New Mexico 88002

Commanding Officer (AMSEL-BL-WS-R)
Atmospheric Sciences Laboratory
U.S. Army Electronics Command
White Sands Missile Range,
New Mexico 88002

Chief
Missile Electronic Warfare Tech
Area (AMSEL-WL-M)
Electronic Warfare Laboratory, USAECOM
White Sands Missile Range,
New Mexico 88002

Product Manager NAVCON
Attn: AMCPM-NS-TM, Bldg 439
(H. H. Bahr)
Fort Monmouth, New Jersey 07703

Mr A. D. Bedrosian, Rm 26-131
U.S. Army Scientific Liaison Office
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, Massachusetts 02139

Commanding General
U.S. Army Electronics Command
Fort Monmouth, New Jersey 07703
Attn: AMSEL-SC
DL
GG-DD
XL-D
XL-DT
XL-G (Dr S. Kronenberg)
XL-H (Dr R. G. Buser)
BL-FM-P
CT-D
CT-R
CT-S
CT-L (Dr W. S. McAfee)
CT-O
CT-I
CT-A
NL-D (Dr H. Bennett)
NL-A
NL-C
NL-P
NL-P-2
NL-R
NL-S
KL-D
KL-I
KL-E
KL-S
KL-SM
KL-T
VL-D
VL-F
WL-D
RD-PB (Miss F. Morris)

Department of the Navy

Director, Electronics Programs
Attn: Code 427
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Mr Gordon D. Goldstein, Code 437
Information Systems Program
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217

Commander
Naval Security Group Command
Naval Security Group Headquarters
Attn: Technical Library (G43)
3801 Nebraska Avenue, N. W.
Washington, D.C. 20390

Director
Naval Research Laboratory
Washington, D.C. 20390
Attn: Code 2027
Dr W. C. Hall, Code 7000
Mr A. Brodzinsky, Supt,
Electronics Div

Code 8050
Maury Center Library
Naval Research Laboratory
Washington, D.C. 20390

Dr G. M. R. Winkler
Director, Time Service Division
U.S. Naval Observatory
Washington, D.C. 20390

Naval Air Systems Command
AIR 03
Washington, D.C. 20360

Naval Ship Systems Command
Ship 031
Washington, D.C. 20360

Naval Ship Systems Command
Ship 035
Washington, D.C. 20360

U.S. Naval Weapons Laboratory
Dahlgren, Virginia 22448

Naval Electronic Systems Command
ELEX 03, Rm 2534 Main Navy Bldg
Department of the Navy
Washington, D.C. 20360

Commander
U.S. Naval Ordnance Laboratory
Attn: Librarian
White Oak, Maryland 20910

Director
Office of Naval Research
Boston Branch
495 Summer Street
Boston, Massachusetts 02210

Commander (ADL)
Naval Air Development Center
Attn: NADC Library
Johnsville, Warminster,
Pennsylvania 18974

Commander (Code 753)
Naval Weapons Center
Attn: Technical Library
China Lake, California 93555

Commanding Officer
Naval Weapons Center
Corona Laboratories
Attn: Library
Corona, California 91720

Commanding Officer (56322)
Naval Missile Center
Point Mugu, California 93041

W. A. Eberspacher, Associate Head
Systems Integration Division, Code 5340A
U.S. Naval Missile Center
Point Mugu, California 93041

Commander
Naval Electronics Laboratory Center
Attn: Library
San Diego, California 92152

Deputy Director and Chief Scientist
Office of Naval Research Branch Office
1031 East Green Street
Pasadena, California 91101

Library (Code 2124)
Technical Report Section
Naval Postgraduate School
Monterey, California 93940

Glen A. Myers (Code 52Mv)
Assoc Professor of Electrical Engineering
Naval Postgraduate School
Monterey, California 93940

Commanding Officer (Code 2064)
Navy Underwater Sound Laboratory
Fort Trumbull
New London, Connecticut 06320

Commanding Officer
Naval Avionics Facility
Indianapolis, Indiana 46241

Director
Naval Research Laboratory
Attn: Library, Code 2039 (ONRL)
Washington, D.C. 20390

Commanding Officer
Naval Training Device Center
Orlando, Florida 32813

U.S. Naval Oceanographic Office
Attn: M. Rogofsky, Librarian (Code 1640)
Washington, D.C. 20390

Other Government Agencies

Dr H. Harrison, Code RRE
Chief, Electrophysics Branch
National Aeronautics and
Space Administration
Washington, D.C. 20546

NASA Lewis Research Center
Attn: Library
21000 Brookpark Road
Cleveland, Ohio 44135

Los Alamos Scientific Laboratory
Attn: Reports Library
P. O. Box 1663
Los Alamos, New Mexico 87544

Mr M. Zane Thornton, Chief
Network Engineering, Communications
and Operations Branch
Lister Hill National Center for
Biomedical Communications
8600 Rockville Pike
Bethesda, Maryland 20014

U.S. Post Office Department
Library - Room 6012
12th & Pennsylvania Ave., N. W.
Washington, D.C. 20260

Non-Government Agencies

Director
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Mr Jerome Fox, Research Coordinator
Polytechnic Institute of Brooklyn
333 Jay Street
Brooklyn, New York 11201

Director
Columbia Radiation Laboratory
Columbia University
538 West 120th Street
New York, New York 10027

Director
Coordinate Science Laboratory
University of Illinois
Urbana, Illinois 61801

Director
Stanford Electronics Laboratory
Stanford University
Stanford, California 94305

Director
Microwave Laboratory
Stanford University
Stanford, California 94305

Director
Electronics Research Laboratory
University of California
Berkeley, California 94720

Director
Electronics Sciences Laboratory
University of Southern California
Los Angeles, California 90007

Director
Electronics Research Center
The University of Texas at Austin
Engineering-Science Bldg 110
Austin, Texas 78712

Division of Engineering and
Applied Physics
210 Pierce Hall
Harvard University
Cambridge, Massachusetts 02138

Dr G. J. Murphy
The Technological Institute
Northwestern University
Evanston, Illinois 60201

Dr John C. Hancock, Head
School of Electrical Engineering
Purdue University
Lafayette, Indiana 47907

Dept of Electrical Engineering
Texas Technological University
Lubbock, Texas 79409

Aerospace Corporation
P. O. Box 95085
Attn: Library Acquisitions Group
Los Angeles, California 90045

Airborne Instruments Laboratory
Deerpark, New York 11729

The University of Arizona
Department of Electrical Engineering
Tucson, Arizona 85721

Chairman, Electrical Engineering
Arizona State University
Tempe, Arizona 85281

Engineering and Mathematical
Sciences Library
University of California at Los Angeles
405 Hilgred Avenue
Los Angeles, California 90024

Sciences-Engineering Library
University of California
Santa Barbara, California 93106

Professor Nicholas George
California Institute of Technology
Pasadena, California 91109

Aeronautics Library
Graduate Aeronautical Laboratories
California Institute of Technology
1201 E. California Boulevard
Pasadena, California 91109

Hunt Library
Carnegie-Mellon University
Schenley Park
Pittsburgh, Pennsylvania 15213

Dr A. G. Jordan
Head of Dept of Electrical Engineering
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Case Western Reserve University
Engineering Division
University Circle
Cleveland, Ohio 44106

Hollander Associates
Attn: Librarian
P. O. Box 2276
Fullerton, California 92633

Dr Sheldon J. Welles
Electronic Properties Information Center
Mail Station E-175
Hughes Aircraft Company
Culver City, California 90230

Illinois Institute of Technology
Department of Electrical Engineering
Chicago, Illinois 60616

Government Documents Department
University of Iowa Libraries
Iowa City, Iowa 52240

The Johns Hopkins University
Applied Physics Laboratory
Attn: Document Librarian
8621 Georgia Avenue
Silver Spring, Maryland 20910

Lehigh University
Department of Electrical Engineering
Bethlehem, Pennsylvania 18015

Mr E. K. Peterson
Lenkurt Electric Co. Inc.
1105 County Road
San Carlos, California 94070

MIT Lincoln Laboratory
Attn: Library A-082
P. O. Box 73
Lexington, Massachusetts 02173

Miss R. Joyce Harman
Project MAC, Room 810
545 Main Street
Cambridge, Massachusetts 02139

Professor R. H. Rediker
Electrical Engineering, Professor
Massachusetts Institute of Technology
Building 13-3050
Cambridge, Massachusetts 02139

Professor Joseph E. Rowe
Chairman, Dept of Electrical Engineering
The University of Michigan
Ann Arbor, Michigan 48104

New York University
Engineering Library
Bronx, New York 10453

Professor James A. Cadzow
Department of Electrical Engineering
State University of New York at Buffalo
Buffalo, New York 14214

Department of Electrical Engineering
Clippinger Laboratory
Ohio University
Athens, Ohio 45701

Raytheon Company
Research Division Library
28 Seyon Street
Waltham, Massachusetts 02154

Rice University
Department of Electrical Engineering
Houston, Texas 77001

Dr Leo Young, Program Manager
Stanford Research Institute
Menlo Park, California 94025

Sylvania Electronic Systems
Applied Research Laboratory
Attn: Documents Librarian
40 Sylvan Road
Waltham, Massachusetts 02154

Dr W. R. LePage, Chairman
Department of Electrical Engineering
Syracuse University
Syracuse, New York 13210

Dr F. R. Charvat
Union Carbide Corporation
Materials Systems Division
Crystal Products Department
8888 Balboa Avenue
P. O. Box 23017
San Diego, California 92123

Utah State University
Department of Electrical Engineering
Logan, Utah 84321

Research Laboratories for the
Engineering Sciences
School of Engineering and Applied Science
University of Virginia
Charlottesville, Virginia 22903

Department of Engineering and
Applied Science
Yale University
New Haven, Connecticut 06520

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) Research Laboratory of Electronics Massachusetts Institute of Technology Cambridge, Massachusetts 02139	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP None

3. REPORT TITLE
Speaker-Machine Interaction in Automatic Speech Recognition

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical report

5. AUTHOR(S) (First name, middle initial, last name)
John I. Makhoul

6. REPORT DATE December 15, 1970	7a. TOTAL NO. OF PAGES 126	7b. NO. OF REFS 77
8a. CONTRACT OR GRANT NO. DA 28-043-AMC-02536(E) b. PROJECT NO. 20061102B31F c. NIH Grant 5PO1 GM15006-03 d. NIH Grant 5 PO1 GM14940-04	9a. ORIGINATOR'S REPORT NUMBER(S) Technical Report 480 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None	

10. DISTRIBUTION STATEMENT
THIS DOCUMENT HAS BEEN APPROVED FOR PUBLIC RELEASE AND SALE; ITS DISTRIBUTION IS UNLIMITED.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Joint Services Electronics Program Through U. S. Army Electronics Command

13. ABSTRACT

This study examines the feasibility and limitations of speaker adaptation in improving the performance of a fixed (speaker-independent) automatic speech recognition system. A fixed vocabulary of 55 [əCVd] syllables is used in the recognition system, where C is one of eleven stops and fricatives, and V is one of five tense vowels. The results of the experiment on speaker adaptation, performed with 6 male and 6 female adult speakers, show that speakers can learn to change their articulations to improve recognition scores. The initial average error rate was 18.3%. As a result of the experiment it was predicted that the average error rate could decrease to 2.3%. The preliminary results obtained also indicate that most of the necessary adaptation can be achieved in a relatively short time, provided that the speakers are instructed in how to change their articulations to produce the desired effects.

The recognition scheme is based on the extraction of several acoustic features from the speech signal. This is accomplished by a hierarchy of decisions made on carefully selected parameters that are computed from a spectral description of the speech signal by means of a set of energoids (energy centroids), each energoid representing the center of energy concentration in a particular spectral energy band. Short-time spectra were obtained either from a bank of 36 bandpass filters covering the range 150-7025 Hz, or by directly computing the Fast Fourier Transform of portions of the sampled speech signal.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Fast Fourier Transform						
Feature Extraction						
Linguistic Distinctive Features						
Man-Machine Interaction						
Signal Processing						
Speaker Adaptation						
Speech Analysis						
Speech Recognition						