3

THE RECOGNITION OF SPEECH BY MACHINE

GEORGE W. HUGHES

TECHNICAL REPORT 395

MAY 1, 1961

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
RESEARCH LABORATORY OF ELECTRONICS
CAMBRIDGE, MASSACHUSETTS

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, jointly by the U.S. Army (Signal Corps), the U.S. Navy (Office of Naval Research), and the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA 36-039-sc-78018, Department of the Army Task 3-99-20-001 and Project 3-99-00-000,

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 395 May 1, 1961

THE RECOGNITION OF SPEECH BY MACHINE

George W. Hughes

This report is based on a thesis submitted to the Department of Electrical Engineering, M. I. T., September 1, 1959, in partial fulfillment of the requirements for the degree of Doctor of Science.

Abstract

The problem of engineering a mechanical (automatic) speech recognition system is discussed in both its theoretical and practical aspects. Performance of such a system is judged in terms of its ability to act as a parallel channel to human speech recognition. The linguistic framework of phonemes as the atomical units of speech, together with their distinctive feature description, provides the necessary unification of abstract representation and acoustic manifestation. A partial solution to phoneme recognition, based on acoustic feature tracking, is derived, implemented, and tested. Results appear to justify the fundamental assumption that there exist several acoustic features that are stable over a wide range of voice inputs and phonetic environments.

TABLE OF CONTENTS

I. SOME THEORETICAL ASPECTS OF SPEECH ANALYSIS

1.1 THE SPEECH COMMUNICATION CHANNEL

The faculty of speech, unique to human beings, has long been the subject of intensive study and investigation. Man is able to use organs intended for the intake of oxygen and food to produce an information-bearing acoustical signal. He has the concomitant ability to extract from this complex signal, even in the presence of much noise or interference, enough information to allow effective communication. Direct study of these phonemena in terms of the human nervous or auditory systems is at best extremely difficult at this time. However, much may be learned about speech production and perception by postulating various models consistent with observed data on human behavior and implementing them in circuitry or logical procedures. Of interest is a comparison between the response of such a device and human subjects, both faced with identical stimuli.

In performing speech recognition, the human organism is capable of selecting from a given ensemble one or a sequence of symbols which represents the acoustical signal. The aim of mechanical speech recognition studies, then, is to develop measurement techniques capable of duplicating this function, that is, extracting the information-bearing elements present in the speech signal.

1.2 DISCRETENESS IN SPEECH

Communications systems serving to transmit information generally fall into two categories:

(a) Those for which the input can not be described by a fixed set of discrete values. The range of the input is a continuum, and the output is (through some transformation) the best possible imitation of the input. Although the range of values assumed by the input function may be limited, an attempt must be made to produce a unique output for every value of the input in that range. Thus, the number of values possible at the output approaches infinity as the quality of transmission increases. Examples of systems of this sort are high-fidelity amplifiers, tape recorders, and radio transmitters.

(b) Those for which the input can be expressed in terms of a fixed set of discrete (and usually finite) values. The input is often considered to be encoded in terms of members of the set. Here the output may only be a representation or repetition of the input rather than a (transformed) imitation. Again we require an output symbol or value for every value of input. If the input range is bounded, however, only a finite number of output values will serve to distinguish among all possible inputs. Examples of systems of this sort are pulse code modulation, digital voltmeters, and the Henry system of fingerprint classification.

Even assuming only the mildest restrictions, that is, bounded inputs and the nonexistence of noiseless or distortionless links, it is evident that systems composed of many links of type 2 will perform quite differently than those involving links of type 1. No matter how small the imperfections in the individual links of type 1, a sufficient number in cascade will produce a system in which there is no measurable correlation between output and input. If, on the other hand, the imperfections in the links of type 2 are only small enough not to cause a given input to produce more than one of the discrete outputs, compound systems will perform perfectly regardless of the number of links.

In any information processing system in which repeatability is to be possible the set in terms of which all messages are represented must be discrete (denumerable). The input may be said to be encoded in terms of the units of the output set. In information-processing systems where no encoding or decoding takes place, an attempt usually is made only to reproduce or amplify a signal.

One important characteristic of the speech process is the preservation of repeatability as a message is communicated from one speaker to another. It may be noted, however, that in no sense is the speech signal representing a given message reproduced as it is passed from speaker to speaker. Since successful speech communication depends neither on the total absence of noise nor on an ability to imitate perfectly an auditory stimulus, we may expect to find a code or discrete and finite set of units common to all speakers of a language which will suffice to represent any speech event recognizable as part of that language.

Alphabetic writing provides further evidence of discreteness in language. It is always possible to transcribe a long or complex speech utterance as a time sequence of smaller, discrete units. Simply stated, the problem of mechanical speech recognition is to do this automatically.

1.3 THE CODE

Before formulating any identification scheme it is necessary to define the set of output symbols in terms of which all possible inputs must be described. The units of the set may in general be quite arbitrary in nature; that is, the laws of information theory which describe the operation of information-processing systems do not impose any restrictions on the choice of symbols. However, in devising realizable mechanical recognition procedures, the choice of the output set is crucial and should be governed by at least the following criteria:

(a) Each member must be in some way measurably distinct from all others.

(b) The set must be of sufficient generality so that only one combination of its constituent units will form each more complex utterance of interest.

(c) The economy of a description of all possible complex input utterances in terms of units of the set must be considered. In general this means the size of the set is minimum although details of implementation may dictate otherwise.

For many purposes of identification, sets satisfying only criterion 1 are both convenient and sufficient. For example, in specifying a lost article such as a light blue, four-door, 1951 Ford, or a medium-build, blond, brown-eyed, mustached husband, one tacitly defines the set in terms of immediately recognizable features. No attempt is made to specify the item in terms of sufficient generality to identify all cars or people respectively.

In many areas, including speech analysis, considerations of generality and economy dictate the nature of the set of output symbols. If a general set is found whose units will describe every meaningful utterance, then, of course, any solution to the problem of finding measurements on the speech waveform that will physically specify that set is, by definition, a complete solution. The price paid for this guarantee of completeness is the difficulty of discovering and instrumenting the physical measurements necessary to separate the units of a fixed, linguistically-derived set.

The aim of speech recognition devices is to distinguish among all utterances that are not repetitions of each other. Thus, for example, the English words "bet" and "pet" are to be classified as "different" (although perhaps physically quite similar), and all the utterances of "bet" by sopranos, baritones, and so forth, are to be classified as "same" (although physically quite different). In other words, we must discover those properties of speech signal that are invariant under the multitude of transformations that have little effect on our ability to specify what was said. For example, if listeners were asked to judge which of two different words were spoken (even in isolation), their response would be relatively independent of the talker's voice quality or rapidity of speech. Also, listeners would have no trouble in recognizing that the "same" vowel (/u/) occurs in the words moon, suit, pool, loose, two, and so forth. Such differences in phonetic environment and/or the speaker's voice quality generally have serious acoustical consequences which are difficult for mechanical recognition procedures to ignore. Most devices constructed to perform speech recognition show inordinate sensitivity to the many features of the speech signal which are readily measurable but have no linguistic significance. That is, a change in several properties of the input will cause the device to register a different output but will not cause a panel of listeners to significantly modify their responses. Success of a given mechanical speech recognition scheme, therefore, may be judged in terms of how closely its output corresponds to that of human subjects presented with the same input stimuli. For example, an adequate identification scheme for a given set of words would at least make identical judgments of "same" or "different" when applied to pairs of the words which speakers of the language would make. The set of all phonetically different utterances in a language defines a possible set of complete generality if the number of units, n, is made large enough. Assuming, however, that measurements could be found to separate each member (utterance) from all others, as many as $\frac{n(n-1)}{2}$ such measurable differences might have to be taken into account. Of course if such a procedure were adopted, many of the measurements would overlap or actually be identical, so that $\frac{n(n-1)}{2}$ represents only an upper bound. However, the number of measurements that would have to be defined would still be very large for sets of even moderate size. Furthermore, for a natural language, it is impossible even to specify an upper bound on n (the number of words for example). It is apparent that a solution based on a set of the phonetically distinct utterances themselves is not only uneconomical, but unfeasible.

The problem of identifying physical phenomena belonging to an unbounded set is known to other disciplines, cf. analytical chemistry. The solution lies in regarding complex phenomena as configurations of simpler entities whose number is limited. In the case of speech, all utterances may be arranged in a kind of linguistic hierarchy roughly as follows: (a) phrases, sentences, and more complex units, (b) words, (c) morphemes and syllables, and (d) phonemes.

The number of units in each set of utterances of complexity greater than the syllable is infinite – that is, no procedure can be given which will guarantee an exhaustive catalog.

The phoneme is the smallest unit in terms of which all speech utterances may be described. If one phoneme of an utterance is changed, the utterance will be recognized as different by speakers of the language. (For a more complete definition of the phoneme and discussions of its role in linguistics, see Jakobson and Halle (14), and Jones (15).) Each language possesses its own set of phonemes. No two known languages have identical sets of phonemes nor entirely different sets. In English, the number has been given by various linguists to be between 30 and 40 (see Table I).

Thus, the phonemes of a language will provide a set of output symbols which meet the requirements of generality and economy – their number is limited, yet there exists no speech utterance which cannot be adequately represented by a phoneme sequence.

1.4 TWO APPROACHES TO RELATING A CODE TO AN INPUT SIGNAL

There remains the question of relating this set to measurable properties of the speech waveform. This problem has often been minimized as "a detail of instrumentation" or "to be worked out experimentally." However, upon closer examination there appear at least two fundamentally different approaches to discovering this relationship.

The first is to choose carefully a set of measurements and then define the members of the output set in terms of the expected or observed results of applying these to speech events. Various sets of measurements have been chosen and optimized under such diverse considerations as equipment available, and experimental results with filtered, clipped, or otherwise distorted speech. An output set derived in this fashion will in general be larger, more flexible, and of limited generality. Devices instrumented on this basis may include rules for transforming the complex output set into a simpler set whose members may have more linguistic significance. This approach is the basis of many recognition devices reported in literature. (For examples see Davis, Biddulph, and Belashek (3), Fry and Denes (6), and Olson and Belar (16).)

Many of these devices illustrate an approach to mechanical speech recognition often termed "pattern matching." This term is somewhat misleading, because in a sense any speech recognition device will, at some stage near the final determination of an output symbol, make a match or correlation between a measured set and a fixed set of parameters. However, in the pattern-matching schemes the measurements themselves are taken to form patterns which are to be matched against a set of stored standards. The

Table I. Phonemes of English and their distinctive feature composition.

DISTINCTIVE FEATURES	PHONEMES																																		
	i	ɪ	u	ʊ	ɛ	e	æ	o	ɔ	ɑ	ʌ	r	l	w	j	m	n	ŋ	f	s	ʃ	θ	ð	v	z	ʒ	dʒ	tʃ	p	t	k	b	d	g	h
1. SONORANT/NON-SONORANT	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
2. CONSONANT/NON-CONSONANT	−	−	−	−	−	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−
3. COMPACT/NON-COMPACT						−	+	−	+	+						−	−	+	−	−	+	−	−	−	−	+	+	+	−	−	+	−	−	+	
4. DIFFUSE/NON-DIFFUSE	+	+	+	+	−	−	−	−	−	−	−																								
5. GRAVE/ACUTE	−	−	+	+	−	−	−	+	+	+	+	−	+	+	−	+	−		+	−		+	+	+	−				+	−		+	−		
6. TENSE/LAX	+	−	+	−	−	+	+	+	+	+	−								+	+	+	+	−	−	−	−	−	+	+	+	+	−	−	−	
7. FLAT/PLAIN									+	−		+	−	+	−																				
8. NASAL/ORAL												−	−	−	−	+	+	+																	
9. CONTINUANT/INTERRUPTED																			+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	
10. STRIDENT/MELLOW																			+			−	−	+			+	+			−			−	

implementation of this approach usually consists of detecting a maximum of common properties between an unknown input and a set of measurement values taken to define an output. Many of the different acoustical patterns that a speech recognizer must deal with should result in identical outputs, and many should be disregarded altogether. Therefore for each single output symbol (a word, for example), a very large number of patterns or templates must be stored if account is to be taken of differences among speakers or even in the speech of one individual. There is a fantastically large limit to the complexity of analysis and equipment needed to instrument pattern matching if an attempt is made to include a large number of speech events. For these reasons it is not surprising that the pattern-matching approach fails to solve the problem of invariance under transformations which are linguistically insignificant.

A second approach proceeds from a fixed output set to a system of linguistically derived parameters which serve to distinguish among members of the set (which usually have been stated in terms of articulatory positions of the vocal tract). The relationship between these parameters, their acoustical correlates, and possible measurements procedures is treated as a separate problem which has no influence either on the selection of the set or on the parameters in terms of which the set is described. As a result of efforts made in the past few years to include linguistic principles in the design of speech recognition systems, more careful attention has been paid to the criteria by which a set of output symbols is chosen, the detailed nature of this set, and the theory upon which is based a procedure for relating the mechanical selection of an output symbol to measurable properties of the speech waveform.

This approach is epitomized in the development of the distinctive-feature description of phonemes by Jakobson, Fant and Halle (12). (For a more complete discussion of distinctive features, their applications, and implications than follows here, see Cherry (1), Cherry, Halle and Jakobson (2), Halle (7), and Jakobson and Halle (14).) Here the phonemes are subjected to a binary classification scheme based on linguistic observations and for the most part utilizing the terminology of phonetics. It is to be noted that the authors applied the same structure of reasoning from phonemes to distinctive features as had previously been applied to reasoning from words to phonemes. Their work shows that identification of linguistic units should proceed from the definition of a set of distinctive differences among the members of the output set and the expression of these differences in terms of measurable properties of the input signal.

Several speech recognition devices have been built with an attempt to incorporate this principle in their design. (See Hughes and Halle (11) and Wiren and Stubbs (18).)

1.5 A PARTICULAR DISTINCTIVE-FEATURE REPRESENTATION OF SPEECH

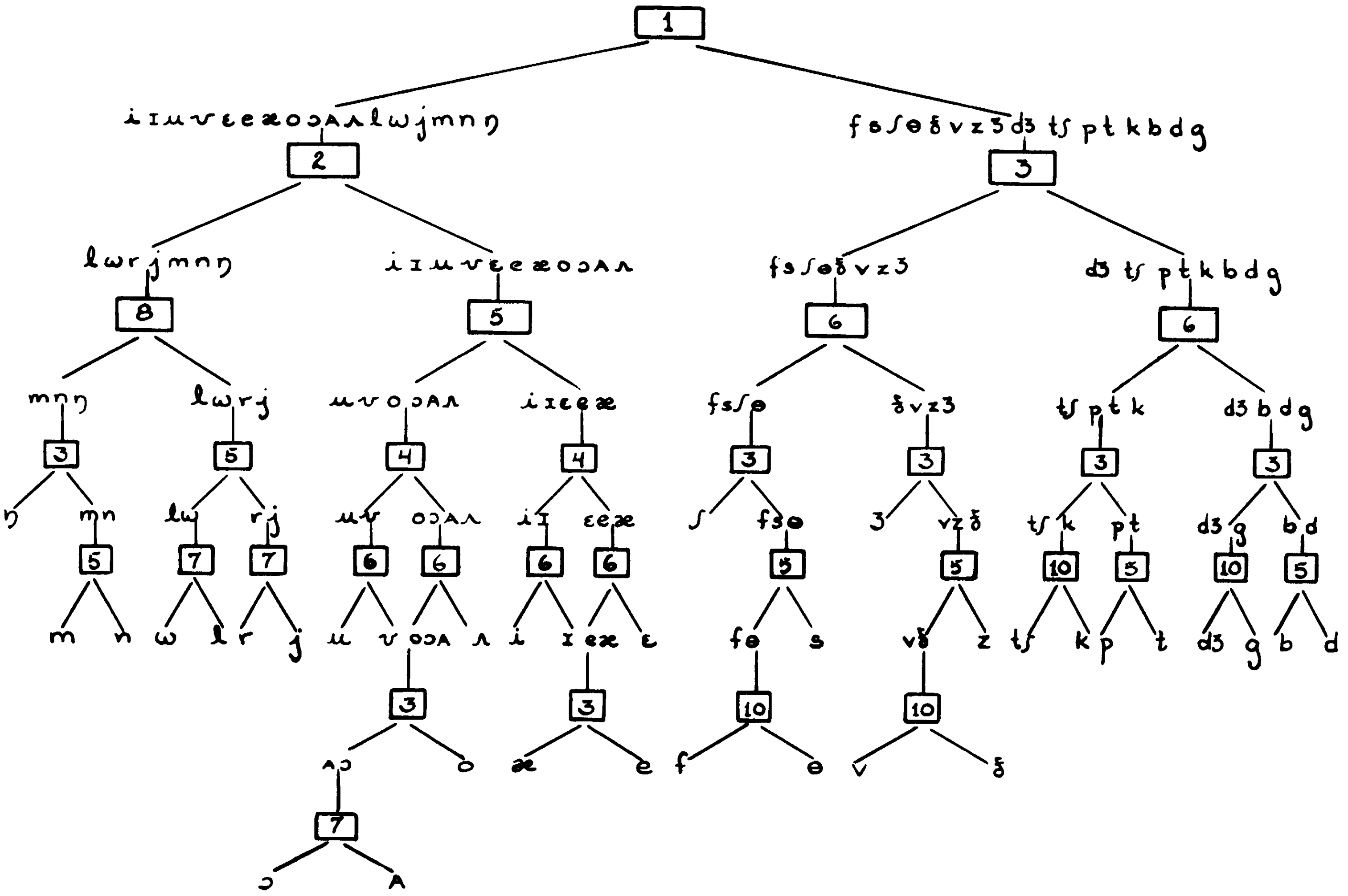
The analysis of the phonemes of a given language in terms of distinctive features is dictated largely by the facts of articulation, questions of the economy of the description, and the degree of consistency with most probable explanations of phenomena occurring when phonemes are connected into sequences (grammar, morphology, and so forth). A distinctive-feature analysis of the phonemes of English is given in Table I. (The original distinctive-feature table for English was made by Jakobson, Fant and Halle (13). Certain modifications and changes have been made by the author in the preparation of Table I, notably in the features 1, 8, and 10.) Table II shows a possible tree structure for phoneme identification based on this analysis.

The high degree of correlation between such a set of distinctive features and observable linguistic behavior is a strong argument in favor of attempting to base a mechanical identification procedure directly on the detection of the presence or absence of these features in the speech waveform. The generality of such a solution and its elimination of redundant features such as voice quality are obvious. In addition, the economy of successive classifications versus, for instance, determining the $\frac{n(n-1)}{2}$ individual differences between n phonemes, is evident. A third and perhaps most important advantage of this approach from a practical standpoint is the usefulness of schemes based only on detection of a few of the features. For example, if only features 3-6 in Table I were instrumented, the device would be capable of classifying all vowels, confusing only /a/ and /ɔ/. In any case, the confusions made by a partial scheme would be completely predictable. Whether or not the particular set of distinctive features shown in Table I is taken to describe the phonemes, the generality and economy of the theoretical framework they illustrate is maintained if the principle of successive classification is preserved. A mechanical procedure built on this framework need only track those acoustical features of the speech input necessary to distinguish among the selected classes which define a representation of the input.

1.6 RELATIONSHIP BETWEEN ABSTRACT CLASSES AND PHYSICAL MEASUREMENTS

Difficulties arise in connection with relating a set of distinctive features to a set of measurable properties of the speech waveform. Although this is the final step in the general solution to phoneme recognition, it is also the least well understood at present. The science of linguistics which furnished an output (the phonemes) of great generality does not provide a complete, mechanically realizable separation procedure. In many cases the phonemes and phoneme classes are well described in terms of articulation; however, this knowledge itself does not indicate the appropriate measurement procedures. Although the acoustical output is theoretically calculable from a given vocal tract configuration, the converse is not true, that is, proceeding from sound to source involves a many-to-one relationship. Also, little is known in detail about human perception of the many acoustical consequences of each state of the vocal organs. In the search for invariant acoustical correlates to a set of distinctive features we are thus forced to rely almost wholly on experimental results. The essential contribution of the distinctive-feature approach is to point out what type of experiments to make and to outline clearly what constitutes success and failure in the application of a given measurements scheme. In other words, projected measurements procedures do not determine

Table II. Tree diagram for the identification of phonemes.



the end results, but postulated end results do determine the measurements procedures. This is not to say, however, that no reasoning from immediately available measurements techniques to a modified output set (less general) is warranted, if some sort of partial identification scheme is wanted.

The concepts implied by the terms "distinctive feature," "acoustic feature," and "physical measurement" as used hereafter should be made clearly separate at this point. The abstract set of distinctive features is such as that given in Table I. The set of acoustic features performs an analogous function (that is, separating classes of events) but is defined solely from measurable parameters. A particular acoustic feature may serve to adequately define a particular distinctive feature, or several in combination may be the acoustical correlates of a distinctive feature. However, there is in general no one-to-one correspondence between these two sets. The set of acoustical features is defined in terms of a set of physical measurements together with threshold values and procedures for processing the resulting data. Thus, the distinctive feature "diffuse/non-diffuse" may have the acoustical feature "first formant low/not low" as its correlate that in turn is derived by tracking the lowest vocal resonance above or below 400 cps.

The final solution to the problem of mechanical speech recognition will map out an orderly procedure for making the transformation from measurements to distinctive features. That this transformation will be complex and not a one-to-one relationship can be seen from the following known facts of speech production.

(a) Absolute thresholds on many single measurements are not valid. For example, in languages in which vowel length distinguishes phonetically different utterances, a short vowel spoken slowly may be longer in absolute time duration than a long vowel spoken rapidly. A correct interpretation of the duration measurement would necessitate the inclusion of contextual factors such as rate of speaking.

(b) The inertia of the vocal organs often results in mutual interaction between features of adjacent phonemes -- so-called co-articulation.

(c) The presence of a feature may modify the physical manifestation of another present at the same time. For example, the acoustical correlate of the feature "voiced-unvoiced" is normally the presence or absence of a strong low-frequency periodic component in the spectrum. However, in the case of stop or fricative sounds this vocal cord vibration may or may not be measurably apparent in the speech waveform, the distinction "voiced-unvoiced" being manifest by the feature "tense-lax."

(d) Some special rules of phoneme combination may change the feature composition of certain segments. For example in many dialects of American English the distinction between the words "writing" and "riding" lies not in the segment corresponding to the stop consonant, as would be indicated in the abstract phonemic representation, but in the preceding accented vowel.

The many complex interrelationships among features and the dependence of phoneme feature content on environment will eventually play a dominant role in speech analysis. However, work in these areas will no doubt be based on a thorough knowledge of those properties of the speech signal that are stable throughout a language community. The work reported here is directed towards qualitatively determining some of the invariant physical properties of speech.

II. DESIGN OF EXPERIMENTS IN COMPUTER RECOGNITION OF SPEECH

2.1 PARAMETERS USED TO SPECIFY ACOUSTICAL MEASUREMENTS OF SPEECH

In order to realize a workable mechanical speech recognition system, we must attempt to bridge the gap between phonemes on the abstract level and measurement procedures on the physical or acoustical level. Traditionally the phonemes and their variants have been described in great detail in articulatory terms. One may view these descriptions as a set of instructions to the talker on how best to duplicate a certain sound. The distinctive-feature approach to phoneme specification, as developed by Jakobson, Fant, and Halle (12), replaces detailed description by a system of class separation and indicates that a small number of crucial parameters enable listeners to distinguish among utterances. If a set of distinctive features such as that given in Table I remains on the abstract level, it is because we lack further knowledge about the correlation between acoustic parameters and linguistic classification of sounds by listeners.

In particular, for a machine designed to duplicate the function of a listener, physical reality means only measurement procedures and measurement results. By means of measurements the designer of speech automata seeks to give a definite physical reality to abstractions such as phonemes and distinctive features.

Some features, known to be of importance in separating general classes of speech sounds, have been discovered and described in terms of articulation, for example, the front or back placement of the tongue, opening or closing of the velum, and so forth. Observed or calculated acoustical consequences of these articulatory positions may be of help in indicating appropriate measurement techniques. Much data on articulation has been gathered by linguists and physiologists using x-rays. The assumption of simple models of the vocal tract such as tubes, Helmholtz resonators, transmission lines, and RLC networks often allows calculation of the appropriate acoustic consequences of a given mode of articulation. These results have been useful in suggesting possible acoustical correlates, but oversimplification in modeling the vocal tract, together with large physical differences among speakers, limits their application towards deriving specific measurement procedures.

Experiments on human perception of certain types of sounds have been conducted. Although not enough is known about the auditory system to be of much help in proposing mechanical identification procedures, bounds on sensitivity and range of measurements needed can often be deduced. At least the acoustic parameters chosen may be checked by observing the response of listeners to artificial stimuli in which these parameters are singly varied. For example, experiments by Flanagan (5) on the perception of artificially produced vowel-like sounds set a limit on the precision needed in measuring formant frequencies.

The choice of a set of acoustic parameters upon which to base speech analysis procedures has often been confused with the problem of explaining results of experiments in perception of distorted or transformed speech or with the problem of maintaining equipment simplicity and elegance. This has led many to search for a single transformation (perhaps a very complex one) which will extract, or at least make apparent, the information-bearing properties of the speech waveform. For example, it has been shown that amplitude compression of speech, even to the extent of preserving only the axis crossings (infinite peak clipping), does not destroy intelligibility. However, no successful attempts to correlate axis-crossing density with phoneme classification have been reported. The discovery of waveform transformations that have proven useful in other fields seems to lead inevitably to their application to speech, whether or not results could reasonably be expected. Autocorrelation, various combinations of fixed filter bands whose outputs are arranged in a multidimensional display, and oscilloscope displays of the speech waveform versus its first derivative are examples of transformations that may put one or more important acoustic features in evidence, but cannot by themselves hope to produce a physical description of a significant number of any type of linguistic unit.

The answer, of course, lies in finding physical parameters on which to base a complex system of individually simple transformations rather than a simple set (one or two members) of complex transformations.

Past studies of speech production and perception make it possible to list here certain acoustic features of speech which are known to play important roles from the listener's point of view.

(a) The presence of vocal-cord vibration evidenced by periodic excitation of the vocal tract.

(b) The frequency of this periodic vocal-tract excitation.

(c) The presence of turbulent noise excitation of the vocal tract as evidenced by a random, noise-like component of the speech wave.

(d) The presence of silence or only very low frequency energy (signaling complete vocal tract closure).

(c) The resonances or natural frequencies (poles and zeros) of the vocal tract and their motion during an utterance.

(f) General spectral characteristics other than resonances such as predominance of high, low, or mid-frequency regions.

(g) Relative energy level of various time segments of an utterance.

(h) Shape of the envelope of the speech waveform (rapid or gradual changes, for example).

Although this is not a complete catalog of the important acoustical cues in speech, a system of class distinctions based upon only these would be able to separate a great many general categories of linguistic significance. The problem is to detect as many of these features as possible by appropriate measurements and then, on this basis, to design logical operations which lead to segment-by-segment classification.

2.2 THE SONAGRAPH TRANSFORMATION

The Sonagraph (sound spectragraph) performs a transformation on an audio waveform which puts in evidence many of the important acoustic features of speech. Sonagrams are a three-dimensional display of energy (darkness of line), frequency (ordinate), and time (abscissa). As such, any organization of a signal in the time or frequency domains is visually apparent. In particular, resonances, type of vocal tract excitation, and abrupt changes in level or spectrum are readily discernible. Since results and procedures described in succeeding chapters are conveniently illustrated or discussed in terms of a sonagraphic display, an example is given here.

Figure 1 shows a sonagram of the word "faced" which includes many of the acoustic parameters listed above. The frequency scale has been altered from the conventional linear function to one in which frequency is approximately proportional to the square of the ordinate. This modification allowed more accurate determination of the position of low-frequency resonances. Note that general spectral characteristics such as vowel resonances and the predominance of high-frequency energy in the fricative are obvious. Also the vocal frequency is easily computed (by counting the number of glottal pulses per unit time) to be approximately 130 cps. Certain temporal characteristics are also evident, such as the duration of the various segments and the abruptness of onset for the stop burst. Since the dynamic range (black to white) is small (only about 20 db), and high-frequency pre-emphasis is incorporated in the Sonagraph circuitry, the envelope shape, frequency-band energy-level ratios, and general over-all level characteristics can only be crudely estimated.

Although sonagrams display several important acoustic cues, particularly for vowels, attempts to read speech from sonagrams have been largely unsuccessful. For consonants so much information is lost or is not easily extracted from the sonagram that complete distinctions are difficult or impossible. The principal values of sonagraphic speech studies are to provide a qualitative indication of what kind of measurements might prove fruitful and to provide gross quantitative data on resonant frequencies, time duration, and so forth. From no other single transformation can as many important acoustic parameters of speech be viewed simultaneously.

2.3 OBJECTIVES OF THE EXPERIMENTAL WORK

In order to test the feasibility of a feature-tracking approach to automatic speech recognition, I undertook an experimental program utilizing the versatile system simulation capabilities of a large scale digital computer. The objective was not to instrument a complete solution, since this would assume knowledge of how to track all the distinctive features in speech. Rather the aim was the more limited one of developing tracking and classification procedures based on features of the speech waveform made evident by the sonagraphic presentation. Such a partial solution was then evaluated from the results.

The experiments actually performed were designed to yield statistically significant

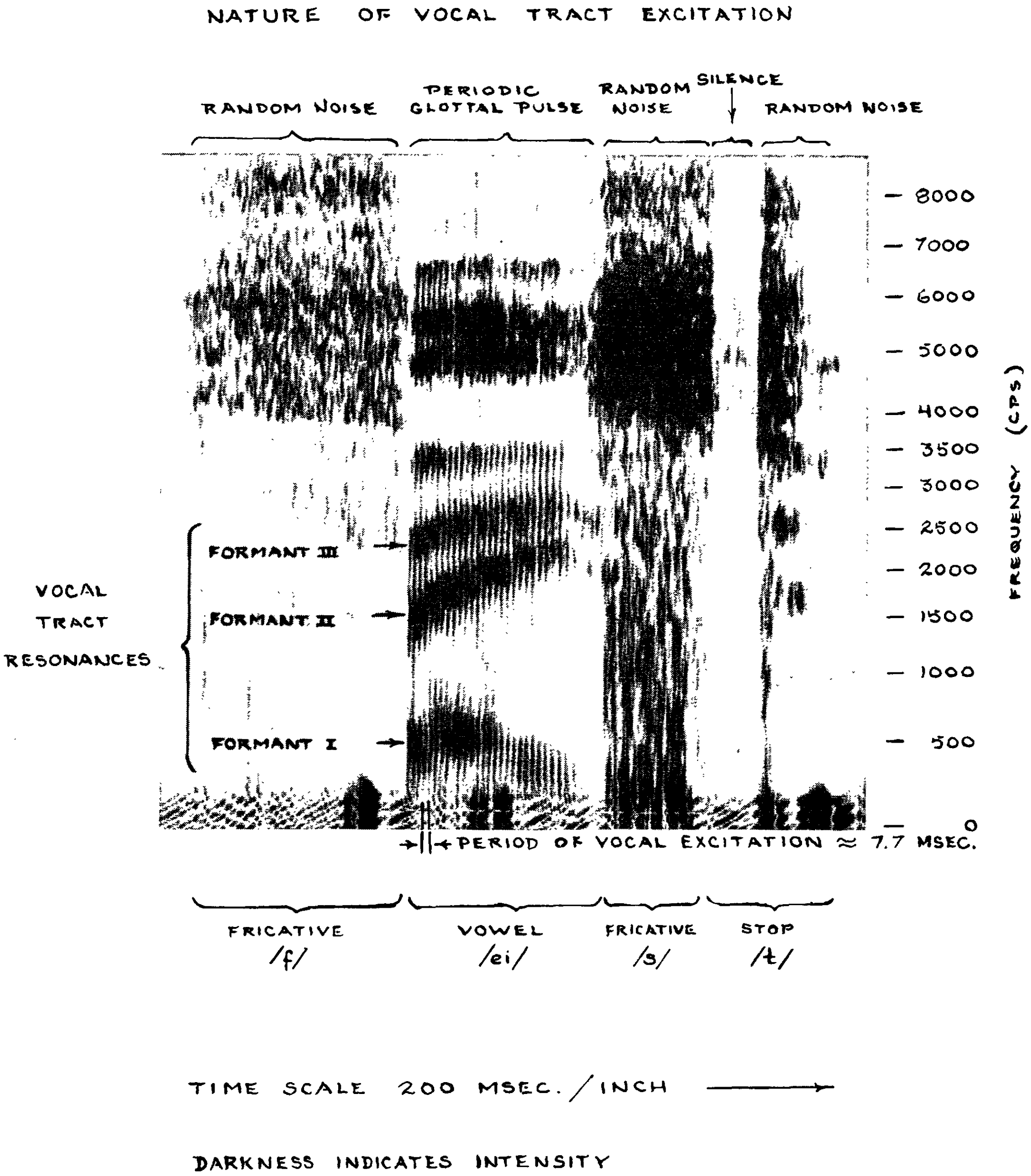


Fig. 1. Sonagram of the word "faced" showing various acoustic features.

information about the following questions:

(a) The stability of the relationship between the readily available short-time spectra of speech sounds and parameters such as formant-frequency positions, types of fricative spectral envelopes, and so forth, upon which classification procedures were based.

(b) The extent and nature of the dependence of tracking and classification procedures on arbitrary or ad hoc threshold constants.

(c) The possible improvement in over-all reliability of classification by increasing the interdependence of the various measurement and decision criteria.

(d) The stability, relative to variation in context and talker, of the relationship between intended (spoken) linguistic unit or class and output symbol chosen by fixed classification procedures.

(e) The usefulness of a partial over-all classification scheme as evidenced by the performance of an actual procedure tried on a reasonably large variety of input stimuli.

Of the acoustical features of speech put in evidence by the sonagraph transformation, five were chosen to form the basis of a partial identification scheme. These are presented below, together with a discussion of the procedure whereby each was related to the incoming raw spectral data. A flow diagram summarizing the operation of the analysis program is shown in Fig. 2.

Results of previously published studies by diverse investigators led to the choice of this particular set of acoustical features. These investigations show that the set chosen possesses two characteristics consistent with the aim of the present work, that is, to ascertain the strengths and weaknesses of the basic approach to combining individual feature tracking into an over-all identification scheme.

(a) Changes in these features, or in the parameters they represent, are directly perceived by listeners as a change in the utterance and, in most cases, as a change in the linguistic classification of the utterance. In other words, these acoustical features are known to be important in speech perception. There was no need for further psychoacoustical evidence to justify the present study.

(b) Relatively straightforward measurement procedures have been postulated which relate most of these features to the type of spectral data available as a computer input. The main research effort was, therefore, placed on developing these procedures into a workable partial identification scheme rather than generating measurement techniques.

2.4 DESCRIPTION OF THE ACOUSTICAL FEATURES TRACKED

a. Formant Frequencies. It has been shown that the frequency locations of the two lowest vocal tract resonances play the central role in separating vowel classes. (If the vowel sound in the standard American pronunciation of the word "her" is included, the location of the third resonance must also be taken into account.) In particular, the distinctive features (see Table I) Compact/Non-compact, Diffuse/Non-diffuse, and Grave/Acute are directly correlated with the position of the lowest frequency resonance (first formant or F1) and the difference in cps between the two lowest resonances (F2-F1).

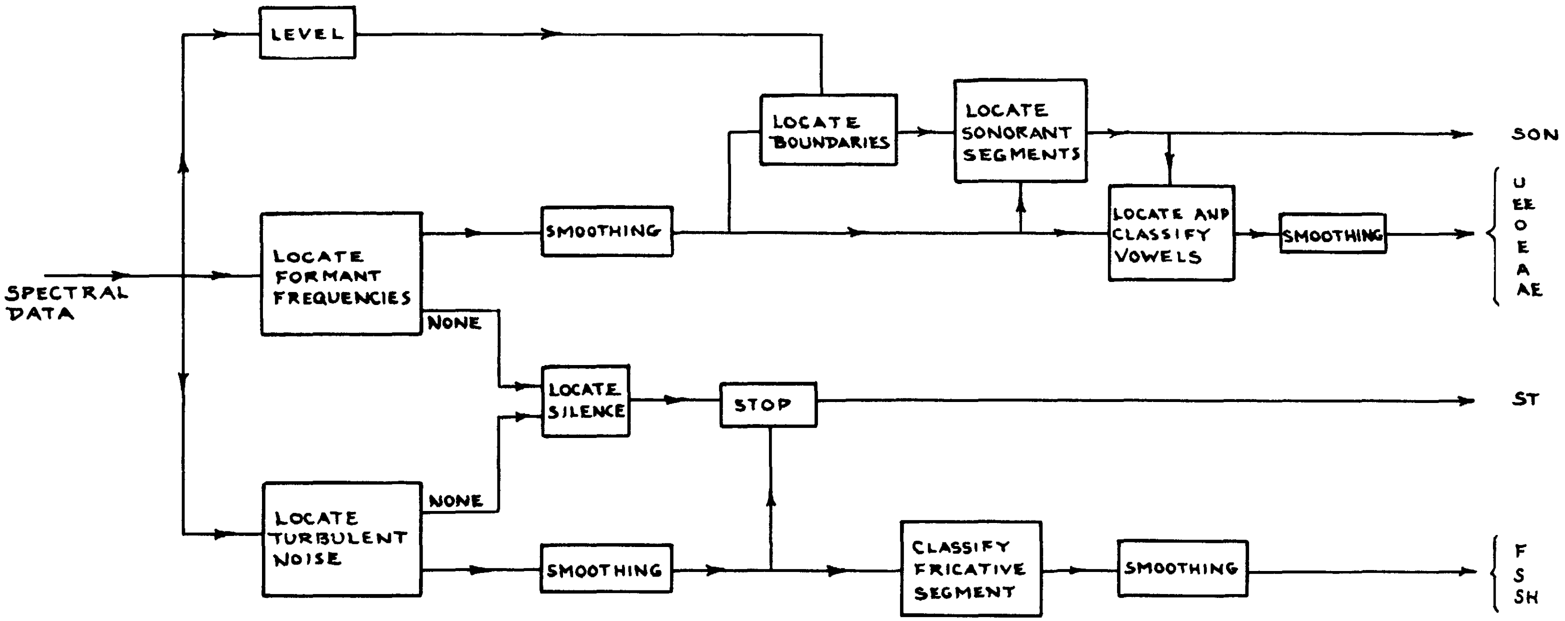


Fig. 2. Flow diagram showing procedures of the analysis program.

The acoustical correlates of the remaining distinctive features pertaining to vowel identification (Tense/Lax, Flat/Plain) are less well understood and are apparently more complex than the others. However, it has been suggested by Jakobson, Fant, and Halle (12) that the acoustical description of these features will also be based in large part on formant positions.

The motion of the formant positions at the beginning or end of a vowel segment is often an important cue to the identity of the preceding or following consonant. Vocal resonant frequencies change during the production of diphthongs, glides, and so forth, and often change discontinuously at vowel-sonorant boundaries.

It is apparent that any speech recognition system (man or machine) must rely heavily on tracking the frequency positions of at least the first two formants during vowel and sonorant portions of an utterance.

The formant-tracking computer program developed for this study assumes as a basis that a very simple and direct relationship exists between formant position and the short-time spectral input to the computer, that is, that the filter channel whose center frequency is closest to the actual formant frequency at the time the filter output is sampled will be clearly maximum relative to the other filters in the formant region. This approach was chosen because a set of fixed band-pass filters was used to obtain the spectral input data to the computer. An important invariant characteristic was revealed by a comparison of vowel spectra developed using this set of filters with spectra of the same vowel segments developed using a continuously variable spectrum analyzer (Hewlett-Packard Wave Analyzer modified to have a bandwidth of 150 cps). Although much loss of definition occurred when the fixed filters were used (particularly near the "valleys" or spectral minima), spectral "peaks" or local maxima were always well correlated with those found using the more laborious variable single-filter technique.

As a complete formant-tracking scheme, implementation of simple peak picking would encounter the serious difficulty that the frequency region in which it is possible to observe the second formant overlaps both the first and third formant regions. Data collected using adult male and female speakers of English show that Fl may occur in the region from approximately 250-1200 cps, F2 from 600-3000 cps and F3 from 1700-4000 cps, although occurrences in the extremes of these regions are rare. In addition to formant region overlap, a strong first or second harmonic of a high-pitched voicing frequency may enter the Fl region. Vocal frequencies of 150-300 cps are common with female speakers. Thus, it is impossible to define two fixed, mutually exclusive sets of filters in which the maxima will be related to the first and second formants with any degree of certainty.

In the spectra of most vowel utterances the amplitude of the first formant exceeds that of the second which in turn exceeds that of the third, and so forth; that is, vowel spectra characteristically fall with frequency. The exceptions to this rule in the case of closely spaced formants, together with the vagaries of glottal spectra, make relative amplitude alone an unreliable criterion for distinguishing among formants. Short-time glottal disturbances may cause no distinct maxima at all to appear in one or more of the formant regions.

In order to pursue the peak-picking technique, so well adaptable for use with data from a fixed filter set, the attempt was made to overcome the above difficulties by making the following additions to the basic scheme:

(i) In both the F1 and F2 range, provision was made to store not only the filter channel number whose output was the maximum during the sampling period but also those channel numbers (termed here "ties") whose output was within a fixed amplitude threshold of the maximum. Thus, small measurement errors could be more accurately corrected later by imposing formant continuity restraints, and so forth, on several possible alternatives.

(ii) The F1 range was fixed to be smaller than that observed in normal speech. In order to reduce confusions between F1 and F2 or a voicing component, only those filters covering the range from 290-845 cps were examined for F1 maxima. (No significance should be attached to the odd values of frequency limits reported here. They are the result of design criteria applied when the filter set used for these experiments was built. See Table XIII.) These confusions, if present, always caused more serious errors than those introduced by abbreviating the allowable range. For purposes of vowel classification, at least, it makes little difference whether the first formant is located at 800 cps or 1000 cps. Because of the finite bandwidth of the vocal-tract resonances, one of the two channels at the extremes of the allowed F1 range would exhibit maximum output even if the actual formant were outside the fixed limits.

(iii) For each time segment the F1 maximum was located first and the result used to help determine the allowed F2 range. Like the F1 range, the nominal F2 range was abbreviated and included 680-2600 cps. To help prevent confusion between F1 and F2, an additional constraint was imposed which limited the lower boundary of the F2 range to one-half octave above the previously located F1 maxima.

(iv) Some spectrum shaping (additional to that inherent in the increase of filter bandwidth with frequency) was programmed. Frequencies in the lower F1 range were attenuated to reduce confusions (particularly for female speakers) between F1 and voicing, and the frequencies in the F2 middle range were boosted to help reduce F2 → F1 errors.

(v) Continuity of formant position with time was imposed in two ways:

(a) If the ties among spectral maxima were spread over a large frequency range, the closest in frequency to the formant location for the previous or following time segment was selected to represent the formant position.

(b) As the last step in the formant-tracking procedure, certain types of jumps in F1 and/or F2 were smoothed.

Results of the formant-tracking portion of the computer program were not only of interest in themselves but were also the most important source of information for much of the rest of the analysis program. For this reason, the first and second formant frequencies determined as above were printed out for most of the speech utterances fed into the computer before proceeding with the extraction of other features and segment classification. These data were plotted directly on the sonagrams of each utterance for performance evaluation.

b. Presence of Turbulent Noise (Fricatives). The first judgment made on the short-time spectrum of every 11-msec time interval was whether or not sufficient constriction in the vocal tract was present to cause the generation of turbulent noise. If this noise was not present the formant program was entered; if the noise was present the interval was classified "fricative."

Two assumptions underlie the programmed procedure for identifying the presence of a fricative:

(i) All fricatives are manifest by the presence of a strong random noise component in the acoustical wave. Oscilloscope and sonagraphic displays of speech show this to be true for the large majority of speech segments articulated with narrow constriction but not complete tongue closure. Exceptions sometimes occur in unstressed syllables of rapid or careless speech, for example the /v/ in "seven."

(ii) A concomitant feature of the presence of random noise is the predominance of high-frequency energy. Since the resonances of oral cavities behind the point of stricture play little or no role in shaping the output spectrum, and in English the point of fricative articulation is always far enough forward along the vocal tract, no resonances appear below about 1500 cps. Resonances above 3-4 kc do appear in fricative spectra, however, resulting in a high ratio of energy above the normal F1-F2 frequency range to energy in that range. For languages which possess fricatives articulated farther back in the vocal tract (such as /x/ in the German "Buch") this assumption would be questionable unless very careful attention were paid to defining the "high" and "low" frequency ranges.

These assumptions led to equating the existence of more energy above the 3 kc than below to the presence of a fricative. The program to perform this initial judgment simply subtracted the linear sum of the outputs of the filter channels in the frequency range 315-990 cps from the sum of those covering the range 3350-10,000 cps. If this difference exceeded a threshold, the 11-msec speech segment in question was classified a fricative. The value of the threshold, although small and not critical, was nonzero to prevent silence or system noise from being marked as part of a fricative. No distinction was attempted between voiced and unvoiced fricatives; the lower limit of 315 cps was chosen to exclude most of the effects of vocal cord vibration, if present, during fricative production.

One smoothing operation was instituted on the fricative vs nonfricative classification. Isolated single segments judged as fricative (or nonfricative) were arbitrarily made to agree with their environment. This procedure reduced, but did not completely eliminate, momentary dropouts occurring in both classes because of quirks in the acoustical signal or in the input system.

c. Spectral Shape of the Segments Judged as Fricative. A previous study (10) of the correlation between fricative spectra and fricative classification proposed a set of energy-band ratio measurements to distinguish among three classes of fricative sounds in English. Since the same set of filters used in deriving this set of measurements was employed in the computer input system for the present study, it was decided to program the given measurements without further experimentation or changes. Figure 3 summarizes the measurements performed on each 11-msec segment previously classified as a fricative.

Measurement I reflects the relative prominence of high-frequency energy. Absence of appreciable energy below 4 kc is characteristic of /s/, and spectral peaks below 4 kc are characteristic of /ʃ/. The spectral characteristic of /f/ however, varies in one respect from speaker to speaker. Although consistently flat below 6 kc, many, but not all, speakers produce a sharp resonance in the vicinity of 8 kc when articulating /f/. Thus, /f/ appears on both sides of Measurement I.

Measurement II separates /f/ from /s/ by distinguishing a spectrum rising with frequency from 720-6500 cps from one generally flat in this region.

Measurement III detected the presence of sharp resonances in the frequency region in or above the F2 range, a feature characteristic of palatal consonants such as /ʃ/.

As expected, the numbers representing the results of the three measurements, as defined for digital computation, differed slightly from those determined from the analog procedures of Hughes and Halle (10). Therefore, a redetermination of threshold values that define the interpretations "large" or "small" was made. This was done experimentally by observing for what values of these constants maximum separation occurred among a group of about 100 fricative utterances.

A fourth class of English fricative, /θ/ and /ð/, was not included in the identification procedure since relatively little is known about its distinguishing physical characteristic(s). Most of the /θ/ and /ð/ sounds included in the spoken input data were identified as /f/.

Most fricative sounds in English speech last about 50-150 msec. This means that the analysis program made /f/-/s/-/ʃ/ judgments on from 5-15 segments per fricative. Since it was rare for all these judgments to agree, some smoothing procedure was necessary. One of the simplest rules possible was adopted, that is, whichever class was present most often during any given fricative was taken to represent that fricative. Ties were settled arbitrarily.

d. Discontinuities in Level and Formant Position. One of the most difficult measurement problems in speech analysis has been to find a set of parameters which will distinguish sounds in the vowel category from those termed non-vowel sonorants. In English we have seven such phonemes /r/, /l/, /w/, /j/, /m/, /n/, and /ŋ/, hereafter termed simply "sonorants." All have general spectral characteristics very similar to vowels. In particular, the sonorant spectra exhibit strong resonances below about 3 kc which are virtually indistinguishable from vowel resonances or formants. In short, there have yet to be found measurable parameters of the spectrum which will separate isolated

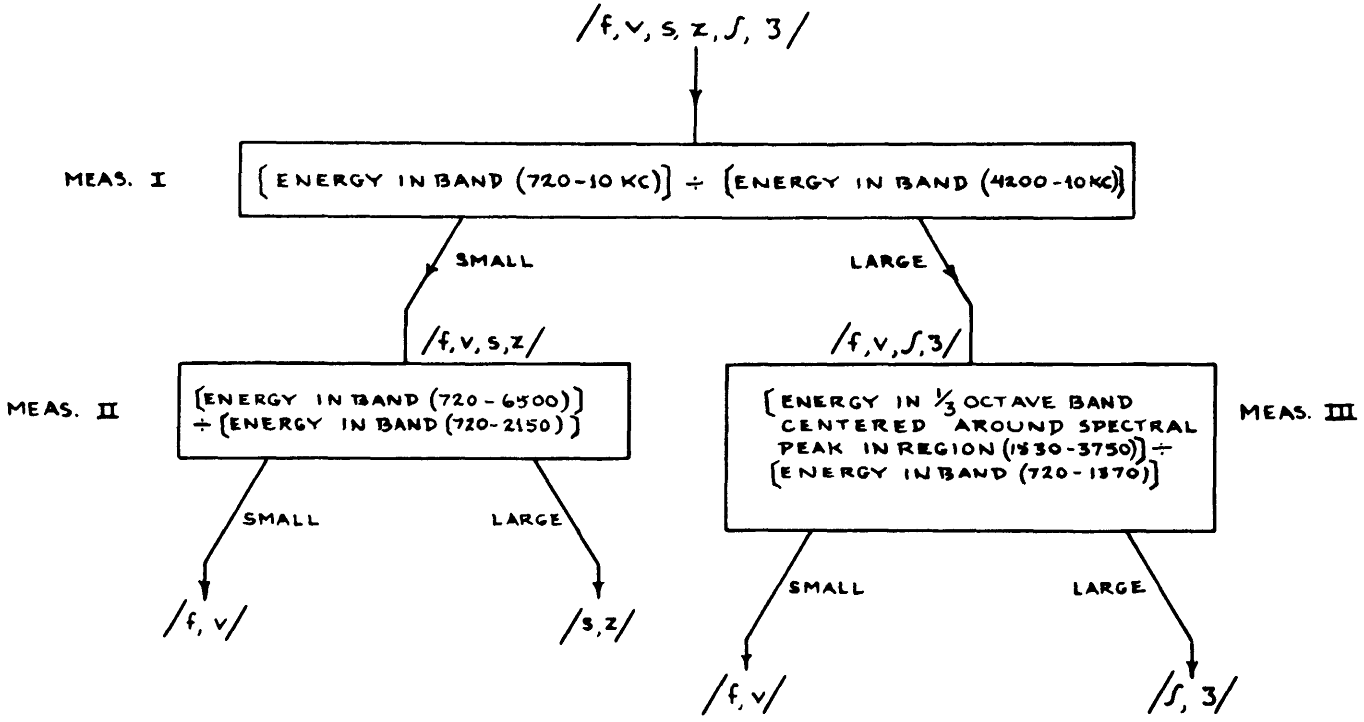

/f, v, s, z, ʃ, ʒ/
MEAS. I
[ENERGY IN BAND (720-10 KC)] ÷ [ENERGY IN BAND (4200-10KC)]
SMALL
LARGE
/f, v, s, z/
/f, v, ʃ, ʒ/
MEAS. II
[ENERGY IN BAND (720 - 6500)]
÷ [ENERGY IN BAND (720-2150)]
[ENERGY IN ⅓ OCTAVE BAND
CENTERED AROUND SPECTRAL
PEAK IN REGION (1530-3750)] ÷
[ENERGY IN BAND (720 - 1370)]
MEAS. III
SMALL
LARGE
SMALL
LARGE
/f, v/
/s, z/
/f, v/
/ʃ, ʒ/

Fig. 3. Classification of fricative spectra.

segments of these sonorants from the general class of vowels. Stevens and House (8) have shown, for example, that a definite correlation exists between the perception of nasality (applied here to /m/, /n/, and /ŋ/) and a slight broadening of formant bandwidth together with introduction of zeros in the spectrum. Attempts to utilize these results in a speech analysis scheme would meet with the formidable problem of instrumenting procedures that are able to detect the presence of these or similar features in the actual waveforms of spoken utterances.

The articulatory correlate of sonorant production is partial or complete constriction of the vocal tract. If the oral passage is completely stopped, the nasal passage is opened by dropping the velum (in normal English speech, nasal passage opening is distinctive only when accompanied by oral closure). Although greater than that present during vowel production, the degree of vocal-tract constriction for sonorants is not sufficient to produce turbulent noise.

A partial solution to the problem of identifying sonorant sounds was attempted in this study. It was based on detecting two of the acoustic correlates of vocal-tract constriction in continuous speech:

(i) Rapid formant or spectral maxima shift in frequency as the constriction goes from open (vowel) to closed (sonorant) or vice versa. This phenomenon is a sufficient, but not necessary, cue for vowel-sonorant boundaries since the place of articulation differs only slightly for some vowel-sonorant pairs.

(ii) Rapid decrease in sound level as vocal-tract constriction occurs and the converse. Constriction simply reduces the efficacy of the vocal tract as a sound radiator.

Assuming flawless detection of these acoustic parameters and their perfect correlation with constriction, the following limitations would nonetheless hold:

(i) At least one vowel-sonorant boundary must be present in an utterance to allow detection of the presence of a sonorant segment. Sonorants isolated from vowels (rare in English) would be indistinguishable from vowels. If one vowel-sonorant boundary were present, the sonorant segment would be closed by a silence, fricative, or second vowel, allowing specification of the presence, relative location, and duration of the sonorant. It may also be noted that there is no redundancy to provide a possible recovery from failure to detect a vowel-sonorant boundary.

(ii) No further breakdown of sonorant segments into subclasses is possible unless other parameters, relating to place of articulation, and so forth, are included.

(iii) Sonorant clusters are treated as a single sonorant.

In the boundary-detection program the parameters of level and formant change were derived from previously computed data that made available the frequencies of F1 and F2 and the level for each 11-msec interval of the utterance. Prior to the development of this program, three types of level were investigated to check correlation of change with vowel-sonorant boundaries:

(i) A simple linear sum of the outputs of the filters.

(ii) Log_2 of the sum of the squares of the filter outputs.

(iii) A number representing the height of the speech envelope. This was not calculated but read into the computer along with the spectral data. The speech waveform was highpass filtered (115 cps), rectified, smoothed, and fed to one of the rotary switch positions.

As expected, all three levels showed reasonably appropriate variation at boundaries. The behavior of the envelope height tended to be sluggish or erratic, depending on the value of smoothing capacitance. The linear sum exhibited the largest dynamic range and was the simplest to compute, hence it was chosen to represent level. A few of the low-frequency channels were omitted in forming the level sum because energy level in the voicing and lower first formant range is the least affected by vocal tract constriction.

Over each span of about 40 msec the magnitude of the net change in formant frequency and the ratio of final level to initial level were calculated. Four criteria for the existence of a sonorant-vowel boundary were applied:

(i) Magnitude of formant change > T_1.

(ii) Level ratio > T_2.

(iii) Formant change > $T_3 < T_1$ and level ratio > $T_4 < T_5$.

(iv) Level ratio > $T_5 < T_2$ and formant change > $T_6 < T_3$.

If any of the above four criteria were met, a boundary was said to exist at that point in the utterance. Each boundary was further classified as vowel-sonorant vs sonorant-vowel by noting in which direction the level changed. Vowels were assumed to always have the greater level. The six threshold constants $T_1 \ldots T_6$ were determined by a detailed examination of formant and level data for about fifty utterances.

Each boundary in the spoken utterance resulted initially in a set of several contiguous boundaries as determined by the computer program. Only one boundary mark was allowed to remain for each such set. The average time location of the set specified the place in the utterance at which a vowel-sonorant boundary was finally inserted.

Because of large level changes, the above process also marked boundaries between vowels or sonorants and fricatives or silence. These redundant boundaries had to be removed so that those remaining would be unambiguously separating vowels and sonorants. An erasure program arbitrarily deleted all boundary marks within four 11-msec segments of a fricative or silence.

e. Silence. Silence, caused by complete closure of the vocal tract and nasal passage, is an important cue for the perception of the distinctive feature continuant vs interrupted that separates stop sounds from fricatives. The period just preceding the explosion of voiced stops may contain very low frequency energy because of vocal cord vibration. However, if "silence" is generalized to mean lack of energy in the first formant frequency region and above, then silence may be taken as a necessary, but not sufficient, cue for the presence of a stop. Two other cues for stops are a short burst of noise following the silence, and rapid vowel formant transitions adjacent to the stop.

With the accuracy of formant tracking thus far achieved, transitional stop cues would be difficult to extract. However, the detection of silence was inherent in the formant and fricative analysis programs. If a segment was classified nonfricative and thus referred to the formant location section of the program which in turn found no formants, it was termed silence by default. The later, overall classification program then used this information to pair a silence followed by a short fricative into the category "stop."

Since only one utterance at a time was read into the computer for this work, initial and final silences (unless voiced) are not detectable. Another difficulty is the separation of true silences from dropouts during fricatives and/or vowels. As a first-order solution to this problem, silences were required to last at least 30 msec before being recognized.

f. Generation of a Set of Output Symbols. Based on the foregoing five acoustic features of the speech input, a program was developed to transform this information into a limited set of output symbols describing the utterance in linguistic terminology. The purpose was threefold:

(i) The usefulness and limitations of this method of feature tracking and classification are readily apparent if the results are displayed in a form resembling linguistic categories. For example, it will be possible to predict, given a list of spoken words, whether or not these mechanical procedures will be able to separate them.

(ii) A gross check may be made on the performance of the individual feature-tracking portions of the program by comparing the output with phonetic transcriptions of the input. Suspected malfunctions are more readily localized prior to more detailed error analysis.

(iii) Data were needed on the feasibility and complexity of the rules necessary to perform the transformation: Feature tracked → Linguistic categories.

Six vowel and five consonant symbols formed the output for each utterance. They correspond, in general, to the following linguistic interpretations:

(i) U. Diffuse, grave, tense vowel as in "who."

(ii) EE. Diffuse, acute, tense vowel as in "he."

(iii) O. Non-compact, non-diffuse, grave vowel as in "hoe."

(iv) E. Non-compact, non-diffuse, acute vowel as in "hay."

(v) A. Compact, grave vowel as in "ah."

(vi) AE. Compact, acute vowel as in "hat."

(vii) SON. Any one of the consonantal sonorants, that is, liquids, nasals, and glides.

(viii) F. Grave, non-compact fricative as in "feel" and "veal."

(ix) S. Acute, non-compact fricative as in "seal" and "zeal."

(x) SH. Compact fricative as in "she'll."

(xi) ST. Interrupted, consonantal, non-sonorant, that is, a stop or affricate, as in "pet," "binge," "deck," and so forth.

The rules for arriving at each symbol from the acoustic features are given in detail in the Appendix. From them the following additional comments may be deduced:

(i) The vowel /I/ (pill) will be classified as E in most cases but as EE in many.

(ii) The vowel /ʊ/ (pull) will be classified as O in most cases but as U in some.

(iii) The vowel /ɛ/ (pet) will be classified as E.

(iv) The vowel /ɔ/ (Paul) will be classified as A for most dialects, as O for some.

(v) The vowel /ʌ/ (putt) will be classified as A in most cases, as O in some.

(vi) Final, unexploded stops will not be detected.

(vii) Initial stops will either be missed altogether or identified as fricatives.

(viii) The identification of affricates as ST, F, S, or SH will be marginal since no information as to envelope rise time is included.

III. EXPERIMENTAL PROGRAM AND RESULTS

3.1 PROCEDURE

The analysis procedures described in Section II and the Appendix were applied to a large corpus of speech in the form of isolated words. Results of computer processing of the real-time input data were made available for analysis in four forms that may be listed in order of complexity of the analysis preceding output presentation:

(a) Final transformation of each word into a sequence of symbols representing 6 vowel classes, 3 fricative classes, sonorant or stop.

(b) Same as (a) less the final combination and deletion procedures. Data in this form indicate how the individual 11-msec intervals of the speech waveform were classified. Short-time variations in classification (less than 30 msec) were deleted, however.

(c) The frequency positions of the lowest two vocal resonances (formants) for each 11-msec real-time interval of the input.

(d) Printouts of the contents of certain computer registers which showed the state of the analysis at various stages in the procedure.

Data of the last type above represented at least 90 per cent of the total collected during the course of this investigation. However, none will be given here, since it was used primarily to improve analysis procedures, locate errors, set threshold constants, and so forth. The achievements and shortcomings of the program as finally developed are best illustrated, rather, in terms of a comparison between the sequence of the output symbols produced by the programmed analysis procedures and the features as deduced from a sonagram or phonetic transcription of the utterance.

3.2 RESULTS OF OVERALL CLASSIFICATION

The sequences of symbols (for all the words and speakers listed in the Appendix) that represented the final computer output for each utterance, after all the feature extraction and classification procedures had been applied, were studied in great detail. (Only tabular condensations of this data are presented here. For a complete presentation of all the output sequences of symbols obtained experimentally, see Hughes (9).) In general, it became apparent that the best results were obtained on words containing a single stressed vowel and the poorest results on polysyllabic utterances containing one or more unstressed syllables. Words such as "verse," "mile," "men," "woo pa," and "oozy" produced outputs in good agreement with their phonetic content. However, the utterances "maul over," "energy," and "ahead," for example, were virtually unrecognizable from the output because of almost complete failure to identify anything but stressed vowels and surrounding consonants. Most of the data fell in the category of one or two mistakes per word in feature tracking or segment classification.

It should be stated here that words 1-50 were chosen with the objective of placing all the vowels of English in as many consonantal contexts as possible. This resulted in words that provided very useful stimuli for developing and evaluating a feature-tracking system, but that would require a more sophisticated analysis procedure than was attempted here to produce outputs well correlated with phonetic input. Words 51-100 are, for the most part, of the consonant-vowel-consonant (CVC) type and lie more within the direct capabilities of such a partial identification procedure as was under study.

A summary of all the data is given in Tables III and IV. These were obtained by correlating phonetic transcriptions of the input utterances with the set of output symbols.

Each entry in Table III is the number of times (expressed in per cent of the number in the second column) a given phoneme (left column) present in the input resulted in the appearance of the output symbol or symbols in the top row. The second column gives the number of occurrences of each phoneme in the entire collection of utterances used as input data. Columns 14 through 21 represent several commonly occurring combinations of single output symbols. The shaded boxes locate where the maximum percentages would be expected to fall, given the programmed criteria for selecting the various output symbols. For example, assuming that the acoustical correlates of all /ʌ/ phonemes are high first-formant frequency position and low difference in cps between the first and second formant positions, and knowing that these were exactly the specification for the output symbol "A", we would expect all /ʌ/'s to be classified as "A".

Table IV conversely gives the percentage of occurrence of the expected phonemes that are given the appearance of an output symbol.

3.3 FEATURE TRACKING

The results may be further broken down to help evaluate the performance of the individual acoustic feature-tracking schemes. Numbers in Tables V through X are left as counted rather than re-expressed in per cent.

Table V gives a measure of the success in transforming the presence of any of the phonemes /f/, /s/, /ʃ/, /θ/, /p/, /t/, /k/, /tʃ/, /h/ into any combination of the symbols F, S, SH, and ST. This is termed here "detecting the presence of a fricative," under the assumption that all these phonemes will be manifest by the appearance of some high-frequency random noise. Failure to detect this noise occurred most often in the case of the stops and /h/. Other failures resulted from the low intensity of noise present during articulation of many /f/ and /θ/ phonemes. The appearance of fricative judgments in some segments of highly diffuse vowels (very low first formant) indicates that F1 was sometimes located in the low-frequency voicing range that was arbitrarily excluded from the measurement giving energy above 3 kc relative to that in the normal F1 range. (For illustrations of these errors and those described below, together with the sonagrams of the actual utterances, see Hughes (9).)

Data given in Table VI was determined under the assumption that all appearances of the phonemes /p/, /t/, /k/, and /tʃ/ (with the exception of initial stops and final unexploded stops whose preceding silence is indistinguishable from silences preceding and following all isolated words) would cause the symbol ST to occur at the output. The circumstances causing errors were mainly dropout occurring at the beginning of an

Table III. Computer output symbol vs phonetic transcription of input (expressed in per cent).

		EE	E	AE	A	O	U	SON	F	S	SH	STOP	ST+F	ST+S	ST+SH	A·O (U)	A·AE E·EE	O(A) E·EE	E·EE	EE·E O·U	#
i	88	91	4.5							1											3.5
ɪ	70	21.5	67																		11.5
e	13		100																		
ɛ	67		91	3	3	1.5															1.5
æ	57		40	55	3.5												2				
ɑ	63		1		86	13															
ʌ	59		1.5		63	34	1.5														
ɔ	64		1.5		66	31															1.5
o	59			1.5	8.5	81	2									7					
ʊ	26	4	23			69	4														
u	100	2	1		1	10	80	1													5
ɜ	44	20	36		2	38															4
ə	24	17	9	8	17	8	8														33
aʊ	50				22	14										58					6
aɪ	51		2		5.5												77	15.5			
oɪ	31				3	3											3	88			3
eɪ	56	2	16														5.5		73		3.5
ju	18	56																		44	
l_1	52					13.5	7.5	71	2												6
l_2	109				6	82		3.5								5					3.5
ω	73					1	8	80													11
r	66	1.5	3			9	6	70				1.5									9
j	18	50	11					6		6	11	11									5
m	156	7					2	73													18
n	168	6.5					0.5	86													7
ŋ	52	13				2	4	77													4
f·v	108							2	69	8.5	1	10	5.5								4
s·z	115								2.5	94	1	1		1.5							
ʃ·ʒ	63								11	26	62		1								
θ·ð	42							9.5	47	12		4.5	3	3							21
p·b	60							6.5	8.5		1	57									27
t·d	80							16	8.5	1.5	8.5	45	2.5	1.5	1.5						15
k·g	81							9.5	6	5	3.5	40		3.5	7.5						25
tʃ·dʒ	55								2	29	24	3	3	22	15						2
h	18			18				28	4		4	18									28
#								87	2			11									

SHADING INDICATES EXPECTED OUTPUT SYMBOL

= NO PHONEME OR SYMBOL PRESENT

Table IV. Probability that a given output symbol represents a particular phoneme or group of phonemes.

EE	i	46%
E	e, ɛ, ɝ, ɪ	69
AE	æ	48
A	ɑ, ʌ, ɔ, ə	75
O	o, ʊ, ɚ, ℓ_2	62
U	u	74
SON	ℓ_1, w, r, j, m, n, ŋ	86
F	f, θ, p_1	78
S	s, t_1	68
SH	ʃ, k_1	58
STOP	p, t, k, tʃ	81
ST+F	f, θ	58
ST+S	s	11
ST+Ṣ	tʃ	63
ST+SH	tʃ	53
A·O·U	aʊ	77
A·AE·E·EE	aɪ	89
O·A·E·EE	oɪ	78
E·EE	eɪ	100
EE·E·O·U	ju	100

Table V. Presence of fricative.

		MEASURED	
		F	FNP
INTENDED	FRICATIVE PRESENT	511	93
	FRICATIVE NOT PRESENT	19	1628

Table VI. Presence of silence.

		MEASURED	
		S	NS
INTENDED	SILENCE	139	70
	NO SILENCE	36	1942

Table VII. Presence of discontinuity.

		MEASURED	
		D	ND
INTENDED	DISCONTINUITIES	439	138
	NO DISCONTINUITIES	33	937

otherwise strongly articulated fricative, deletion of very short silences, and stop bursts or fricatives too short or too weak to be identified as such.

Discontinuities in formant position and/or level caused a sonorant to be indicated. Thus, Table VII was obtained from the data by noting the distribution of the SON symbol. In counting the number of times false discontinuities appeared, those caused by fricatives or stops were not included. No separation of errors into those caused by improper formant-tracking vs those caused by level measurement was attempted, although the voluminous quantity of data available might have made this possible.

Tabel VIII summarizes the results of the classification of fricative segments into F, S, or SH. Note that /θ/ is generally classified as F, with no detection at all (#, ST, or SON) a close second. The affricate /tʃ/ was ambiguously identified as S or SH, apparently depending on the influence of the initial /s/- or /t/-like portion.

Tables IX and X pertain to the classification of vowel segments. The two acoustic features of vowels used in this study were the tripartite division of first-formant frequency position and the binary classification of the distance in cps between F1 and F2. Extraction of these features depended both on correct formant tracking and on the values of the threshold constants (in cps) that defined "high," "low," and "medium." In preparing these charts, assumptions were made as before concerning the acoustic feature composition of the phonemes comprising the input. These are given in Tables IX and X by the coupling of a feature and the phonemes beside it in parentheses. For example, the entries in the row labeled Medium (First Formant) were obtained by noting the distribution of the output symbols for input sounds which were given the phonetic transcriptions /e/, /ε/, /o/ vs /i/, /u/, and /æ/, and /a/. Previously published studies indicate that not all English vowels will fall into the rigid classes set up here. The lower part of Table IX separates the results for some of the vowels included in the input stimuli that (quite predictably) do not fit as well into the same fixed F1 classes as those above.

Table X shows that very good results were obtained in separating the vowel classes "grave" and "acute" on the basis of F2-F1. The only phoneme for which the fixed threshold produced serious ambiguity was /ə́/.

3.4 ACOUSTIC PARAMETERS AND CLASSIFICATION PROCEDURES

Of the acoustic parameters used in this work to specify acoustic features, the most important and most readily correlated with the audio signal (by means of the Sonagraph) is formant position. The formant position during the vowel and sonorant portions of every utterance used in this study were separately printed out. Many of these were plotted directly on a sonagram of the utterance as a performance check. It is difficult to give an exact quantitative evaluation of the formant-tracking procedure because of the enormous quantity of data involved and the lack of distinct formant indication on some sonagrams. However, the following general observations were made:

(a) The frequency positions of the stressed or long vowel portions of an utterance were accurately determined in almost all cases.

Table VIII. Classification of fricative spectra.

	MEASURED		
INTENDED	F	S	SH
F	80	9	1
S	3	110	1
ʃ	8	16	39
θ	21	7	0
tʃ	3	28	21

Table IX. Position of first formant.

	F 1 MEASURED		
F 1 INTENDED	L	M	H
LOW (i, u)	162	22	2
MED. (e, ɛ, o)	2	122	27
HIGH (æ, a)	0	38	88
MED. (ɪ, ʊ, ɝ)	21	48	1
HIGH (ʌ, ə)	1	55	87

Table X. Separation of first and second formants.

	F2 - F1 MEASURED	
F2 - F1 INTENDED	H	L
HIGH (i, ɪ, e, æ, ɛ)	301	8
LOW (a, ʌ, ɔ, u, o, ʊ)	22	373

(b) During less intense portions of an utterance, extraneous effects of glottal spectrum peculiarities, noise, and so forth, sometimes caused uncertainties (jumps and dropouts) in the location of the formant frequencies.

(c) The smoothing procedures necessary to help overcome false indications introduced some errors, especially during short vowels actually having different formant positions than their surroundings.

(d) Closely spaced formants and formants located at the extremes of their normal frequency ranges were the most difficult to track accurately.

(e) Independent determination of F2 (aside from frequency range adjustment) and setting a limit on energy level below which no formant was allowed provided enough flexibility to justify the number of errors introduced.

The classification procedures given in the Appendix are not given as a set of rigorously determined rules based on fundamental principles. They were included in a final step in the analysis program both to obtain some experimental evidence of the consequences of imposing relatively _ad hoc_ procedures on data derived from physical measurements and to aid in evaluating the feature-tracking schemes.

3.5 CONCLUSIONS

We have sought to demonstrate that the problem of automatic speech recognition is best attacked from the theoretical basis of phoneme identification by means of distinctive-feature tracking. The many advantages of this approach, such as simplicity, guaranteed completeness of final solution, and consistency with observed human behavior in perceiving speech, furnish strong enough arguments for basing analysis schemes on this theory. However, as often happens, the price paid for theoretical elegance is practical difficulty. Up to the time the present study was undertaken there existed little experimental evidence that implementation of a complete feature-tracking system (with the speech of any member of a linguistic community as the input and a fixed discrete set of symbols as the output) was feasible. The most important general conclusion that may be drawn from the data summarized above is that procedures for implementing at least a partial solution are indeed feasible and practical with presently available equipment. The data are relevant to the problem of organizing what might be termed a "partial-complete" solution; partial in the sense that an insufficient number of features were acoustically defined and tracked to specify all phonemes, and complete in the sense that the input was speech and the output was linguistic classification.

The first step in the experimental procedure was to postulate a set of acoustical features. The fundamental assumption is that there are some acoustical features of speech that are stable, that may be defined and measured in such a way as to remain invariant over long periods of time or from speaker to speaker. Most of the features and classification rules formulated for the present study were thought to have this property. (Results obtained show more directly than in previous studies for what areas this assumption appears to be valid.) In some cases these features are more directly correlated with the set of distinctive features postulated for English than in others. In all cases the acoustical features chosen were those made evident by the sonagraphic transformation of speech. Measurement procedures are more easily developed for acoustical features derived from study of sonagrams; and it was of vital importance to have a ready method of checking the results of experiments in automatic tracking of the features.

Tracking procedures were implemented using a general-purpose digital computer. For each feature the procedure was developed sufficiently to permit detailed specification of the causes of error and to yield results good enough to justify further development of an overall system.

The information supplied by the results of individual feature tracking was put together to enable linguistic classification of the various segments of the input utterance. This segmental classification was the final output of the system herein described.

Application of the above procedures to a large corpus of speech has resulted in data from which several detailed conclusions may be drawn. The most important of these relates to the existence of fixed, well-defined acoustic features of speech (whether actual tracking procedures have been found or not). If we require that for a feature to exist and be useful some way must be found to define it (and thus track it) with vanishingly small probability of error for all speakers, then it follows that there will be a vanishingly small set of such features. No such criteria need be imposed. What must be found is a set of acoustic features (not necessarily complete) which can be defined and tracked with sufficient accuracy to be termed stable over the possible range of inputs. As a working definition of "stable" we will take "trackable by fixed procedures to an accuracy (averaged over many speakers) of from 75-85 per cent or better."

Any complete set of features (sufficient to identify all utterances) will contain many which may be termed variable, that is, no procedure fixed in advance can be specified which will track these features for all speakers with any degree of accuracy. These features are nonetheless necessary; the information needed to determine the details of how they are to be extracted from a particular utterance will come from the results of tracking the stable features.

As an example, consider the problem of identifying which of the thirteen or so English vowels is contained in an utterance. Details of vowel spectra, duration, intensity, and so forth, may vary from speaker to speaker. The procedure most likely to succeed would begin with the extraction of the most stable features and end with the determination of those features whose exact specification depends most on the individual characteristics of the speaker. Such a procedure might be:

(a) Locate the portion of the utterance produced with least oral constriction, eliminating fricatives, nasals, and so forth.

(b) Eliminate from consideration those portions of the vowel segment likely to have been influenced by a neighboring consonant.

(c) Divide the vowel segment into individual sections if large changes in formant position or level indicate that two or more adjacent vowels or vowel-like sounds were produced.

(d) Classify the spectrum of each section very grossly, such as F1 very high or very low, F2 high or low.

(e) Make finer distinctions among possible positions of F1 and F2 based on the information from (d).

(f) Add the information from stable features concerning high or low voicing frequency, class of neighboring segments, and so forth.

(g) From all of the above, set thresholds on very fine F1 or F2 positions, energy-band ratios, bandwidth of resonances, and so forth.

The data as summarized in Tables V through X give an indication as to what some of the stable acoustic features of speech are. Tables XI and XII below rephrase these data on feature-tracking performance in terms of probabilities of correct tracking. For some features the word lists used in the experimental work do not give a balanced distribution for the dichotomy "present-absent," for example "silence-no silence," Table VI, and "fricative-no fricative," Table V. For this reason the ratios of a posteriori to a priori probabilities are also shown in Tables XI and XII to give a measure of the change from random distribution of a feature, ceteris paribus.

Table XI gives the probability that a feature is tracked correctly as evidenced by the output symbol, given that this feature is present in the corresponding segment of the speech waveform as shown by the broad phonetic transcription of what the speaker intended to say. The most significant results, those which indicate directly how much of the time a feature was tracked correctly, are starred. This table suggests listing the various features involved in the order of per cent of judgments correct to obtain an indication of relative feature stability over the words and speakers comprising the input. From Table XI this order is:

(a) F2-F1 high, (b) /s/-type fricative spectral shape, (c) F2-F1 low, (d) /f/-type fricative spectral shape, (e) F1 low, (f) fricative present, (g) F1 medium, (h) sonorant (discontinuity) present, (i) F1 high, (j) silence (+fricative) present, and (k) /ʃ/-type fricative spectra.

The ratios $P(Y|X)/P(Y)$ tabulated in Table XI give a measure of the effect a given input sound (feature) had on the appearance of each of the mechanically extracted features. It will be noticed that the three features having the lowest percentage of correct judgments had relatively high values of this ratio.

Table XII gives the probability that the appearance of a feature in the output indicates that this feature was indeed present in the input. Results most informative as to the reliability of individual feature-tracking are starred. From Table XII the features, as evidenced by the output, are listed in order of the confidence that may be placed on their correctness:

(a) F1 low, (b) F1-F2 low, (c) fricative present, (d) /ʃ/-type fricative spectral shape, (e) F2-F1 high, (f) sonorant (discontinuity) present, (g) /f/-type fricative spectral shape, (h) silence (+fricative) present, (i) F1 high, (j) F1 medium, and (k) /s/-type

Table XI. Probability that a feature present in an utterance will be correctly tracked.

	P(Y|X)**	P(Y)	$\frac{P(Y\|X)}{P(Y)}$
P{ no F, S, SH | no fricative present}	99%	74%	1.3
P{ no ST | no silence present}	98	92	1.1
P{ no SON | no sonorant present}	97	70	1.4
*P{ F2-F1 high | /i/, /I/, /e/, /æ/, /ɛ/}	97	46	2.1
*P{ S | /s/}	96	49	2.0
*P{ F2-F1 low | /a/, /ʌ/, /ɔ/, /u/, /o/, /ʊ/}	94	54	1.7
P{ S or SH | /tʃ/}	94	67	1.4
*P{ F | /f/}	89	33	2.7
*P{ F1 low | /u/, /i/}	87	27	3.2
*P{ F, S, or SH | fricative present}	85	26	3.3
*P{ F1 medium | /e/, /ɛ/, /o/}	81	39	2.1
*P{ SON | sonorant}	76	30	2.5
P{ F | /θ/}	75	33	2.3
*P{ F1 high | /æ/, /a/}	70	25	2.8
P{ F1 medium | /e/, /ɛ/, /o/, /I/, /ʊ/, /ə/}	70	42	1.7
*P{ ST | silence + fricative}	67	8	8.4
*P{ SH | /ʃ/}	63	18	3.5
P{ SH | /ʃ/, /tʃ/}	52	18	2.9
P{ F1 high | /ʌ/, /ɔ/, /æ/, /a/}	44	30	1.5
P{ SH | /tʃ/}	41	18	2.3

Table XII. Probability that a given feature-tracking result is correct.

	P(X|Y)**	P(X)	$\frac{P(X\|Y)}{P(X)}$
*P{ /i/, /u/ | F1 low} /I/, /ʊ/, /ə/, /ʌ/, /ɔ/ excluded	99%	40%	2.5
*P{ /a/, /ʌ/, /ɔ/, /u/, /o/, /ʊ/ | F2-F1 low}	96	56	1.8
*P{ fricative present | F, S, or SH}	97	27	3.6
P{ no silence | no ST}	97	90	1.1
*P{ /ʃ/, /tʃ/ | SH}	97	33	2.9
P{ no fricative | no F, S, or SH}	95	73	1.3
*P{ /i/, /I/, /e/, /æ/, /ɛ/ | F2-F1 high}	93	44	2.1
*P{ sonorant | SON}	93	37	2.5
*P{ /f/, /θ/ | F}	88	34	2.6
P{ /i/, /u/ | F1 low} /I/, /ʊ/, /ə/, /ʌ/, /ɔ/ included	87	28	3.1
P{ no sonorant present | no SON}	87	63	1.4
P{ /ʌ/, /ɔ/, /æ/, /a/ | F1 high} /I/, /ʊ/, /ə/ included	85	40	2.1
P{/s/, /tʃ/, | S}	81	48	1.7
*P {silence | ST}	79	10	7.9
*P {/æ/, /a/ | F1 high} /I/, /ʊ/, /ə/, /ʌ/, /ɔ/ excluded	75	27	2.8
*P {/e/, /ɛ/, /o/ | F1 medium} /I/, /ʊ/, /ə/ /ʌ/, /ɔ/ excluded	66	33	2.0
*P {/s/ | S}	65	33	2.0
P {/e/, /ɛ/, /o/, /I/, /ʊ/, /ə/ | F1 medium} /ʌ/, /ɔ/ included	60	22	2.7

fricative spectral shape.

The ratios $P(X|Y)/P(X)$ tabulated in Table XII give a measure of the contribution of the mechanical tracking procedures towards specification of the feature content of the input. Note that $P(X|Y)/P(X)$ is not equal to $P(Y|X)/P(Y)$ since the input (acoustic feature content) is not in one-to-one correspondence with the output symbols (interpreted in terms of features). In a complete solution this one-to-one correspondence would be present.

The data produced by the tracking procedures developed for this study support the assertion that at least some acoustic features can be defined that are stable. To summarize:

(a) The difference in cps between the first and second formant frequencies (F2-Fl) exceeding a fixed threshold, the frequency of Fl falling below a threshold, the presence of high-frequency (random) energy, and the general ternary classification of fricative spectral shapes were the most stable acoustic features tracked.

(b) The classification of Fl as high or medium in frequency, the presence of discontinuities, the presence of silence, and detailed aspects of the fricative spectral shape classification also showed evidence of stability on the basis of these data.

The above conclusions on feasibility and stability of distinctive-feature solutions to the problem of mechanical speech recognition are based on the performance of an implemented partial solution having inherent limitations. A critical examination of these limitations indicates definite areas in which further research effort will yield the most progress towards a complete solution.

Improvement in the accuracy of extraction of the acoustic features is, of course, a problem of fundamental importance. A study of the mistakes made by the individual feature-tracking routines shows that errors may be attributed to two main sources:

(a) The system of measurements made on each segment of speech input as represented by the short-time spectrum of that segment.

(b) The correction and smoothing procedures instituted on this preliminary data processing to achieve smooth tracking of the features.

Experience has shown that to within certain limits the complexity of analysis procedures may be divided arbitrarily between (a) and (b) above. The more dependable the initial data processing, the less sophistication is required in smoothing and correction. There is, however, a need for both first-order feature extraction and smoothing in any system no matter what degree of reliability is achieved in either. No segment-by-segment procedure can be postulated which will achieve significant results as long as the input is as erratic and variable as is common with ordinary speech. Also, it is impossible to imagine procedures, no matter how complex, that will smooth and correct totally unreliable preliminary data processing.

For the analysis system which produced the data for the present study, the chief limitation of type (a) above was formant tracking. It is difficult to obtain a number that will adequately describe the accuracy that was achieved in locating the lowest two formants during an utterance, since only the relatively crude checking procedure of comparison of computer output to sonagram was available. Early in the research program some 5000 computed formant locations for some 60 words were carefully evaluated. Results showed that well over 90 per cent of the computed points fell on the formants as indicated by dark bars on the sonagram. Yet this is still not adequate reliability to allow full use to be made of the information carried by formant-frequency position. (Some improvement was achieved thereafter and further correction was entrusted to later smoothing procedures.) Since the formant-tracking errors were for the most part random in size and location, the smoothing procedures imposed later to eliminate them were necessarily ad hoc and arbitrary. The main consequences of these errors (other than the occasional appearance of an incorrect vowel identification) are: first, the effect on other features such as discontinuities which depend on formant tracking, and second, the effect of the necessary increase in severity of later smoothing and correction procedures on feature tracking itself. (Poza (17) reports a system for formant tracking wherein an attempt is made to include more of what is known about the restrictions imposed on the spectra of vowels by the physical nature of the source (vocal tract). Following the model suggested by Fant (4), the three lowest resonant frequencies are varied, the overall vowel spectrum calculated for each pole configuration, and a best-match type of comparison made with the incoming speech-sample spectrum. From the data made available so far, this method of matching the calculated consequences of postulated source configurations to the input measurements seems to make as many or more errors than the one presented here. However, the errors appear to be more systematic. Eventually the best scheme will probably evolve as a combination of the two.)

The first consequence is, of course, quite predictable and is mentioned because a majority of the features of speech depend at least partly on formant information, for example, the distinctive features given in Table I. The second consequence was manifest in the inability of various feature-tracking procedures to follow short-time departures from a steady state. The correction procedures were designed to eliminate as much as possible the random (incorrect) departures from constant or smoothly changing formant positions caused by voice irregularities or initial tracking errors. Consequently many unstressed vowels, short consonants, and even short silences were corrected and went unmarked in the final output.

Overcorrection was also present to a lesser degree in tracking the other acoustic features. Many errors in the final result may be traced to placing too much of the burden of reliability on smoothing and correction procedures and not enough on initial measurement procedures.

Some reduction in the number of errors of type (a) could no doubt be achieved by introducing additional data into the initial stages of the system. Although the short-time spectrum was the only input employed, there exists ample evidence that the inclusion of voice pitch and envelope information will eventually be desirable.

For many of the measurements, calculations, and decisions made by the analysis scheme, fixed threshold constants had to be determined. These constants were determined experimentally by adjusting values suggested by published studies or theoretical calculations to obtain best results. Further improvement of feature tracking by redetermination of the threshold values would be insignificant.

Aside from the overcorrection just discussed, errors (or better, "inadequacies," since the sins were usually those of omission) of type (b) above stem from a fundamental limitation imposed on the correction and smoothing procedures. The time span over which these procedures were allowed to operate was either zero or very small. In other words, the account taken of mutual influence between even adjacent time segments of the input was almost nil. Smoothing of formant positions operated at most over time segments of about 44 msec. In the case of vowel classification and the fricative present feature, about 33 msec was the extent of time dependence. The only mutual constraints imposed between features were the requirements that the second formant be located at least one-half octave above the first, and that boundaries in vowels near previously determined fricative or silence segments be erased. A final classification of a large segment of the input (based on what features were present) in no way influenced the final classification of any other segment. It is well known, however, that both mechanical and statistical a priori relationships do exist among the final feature content of various time segments of an utterance. One of the next logical steps towards developing improved speech recognition will be the inclusion of at least approximations to these constraints.

No experimental evidence appeared that such factors as (a) the choice of bandwidth and type of filters used in the input system, (b) the type of spectrum input (linear vs log amplitude, contiguous vs disjoint frequency scale, and so forth), (c) dynamic range of input, and (d) computer limitations played any significant role in the outcome of the data. Therefore, the performance of the system developed by this study, as evidenced by the output data, may be taken to approximate that which is obtainable under the limitations of moderately good formant tracking and no inclusion of intersegmental mutual influence.

It is interesting to consider the practical utility of such a partial solution to speech identification as has been presented here. Assuming that all the features herein described are tracked throughout an utterance, we have the following classifications available for each segment:

(a) Six classes of vowel sounds, (b) Three classes of fricative sounds, (c) Sonorant, non-vowel sounds, and (d) Exploded stops in non-initial position.

If a restricted vocabulary of monosyllabic words of the CVC type formed the input, the system would be capable of responding differently to each word on a list as long as 120. Furthermore, the number of such lists that could be drawn up of words or nonsense syllables of English is fantastically large. The only requirement is that each word on a list differ from all others on that list by at least one of the features tracked. If the device accepts an utterance as long as CVCVC then the maximum number of separable words on a list becomes 3600, and so on.

Evidently as more features are tracked by such a system these lists become longer.

The most important point is that as long as the acoustic features tracked have a well-defined relationship (not necessarily one-to-one) to linguistically significant (distinctive) features, the performance of the identification scheme applied to any speech input will be completely predictable. If the tracking procedures on one or more of the features is imperfect (as will certainly be the case in the foreseeable future), the exact nature of the distribution of errors in the output will be predictable. For example, given the feature composition of a set of output symbols, output data such as represented by Table III could be derived knowing only the word lists used as an input.

This immediate usefulness of even incomplete and not wholly reliable systems based on the distinctive-feature approach to speech analysis furnishes perhaps the best argument for further energetic study of the formidable practical difficulties.

APPENDIX

1. GENERAL DESCRIPTION OF DATA-PROCESSING PROCEDURES

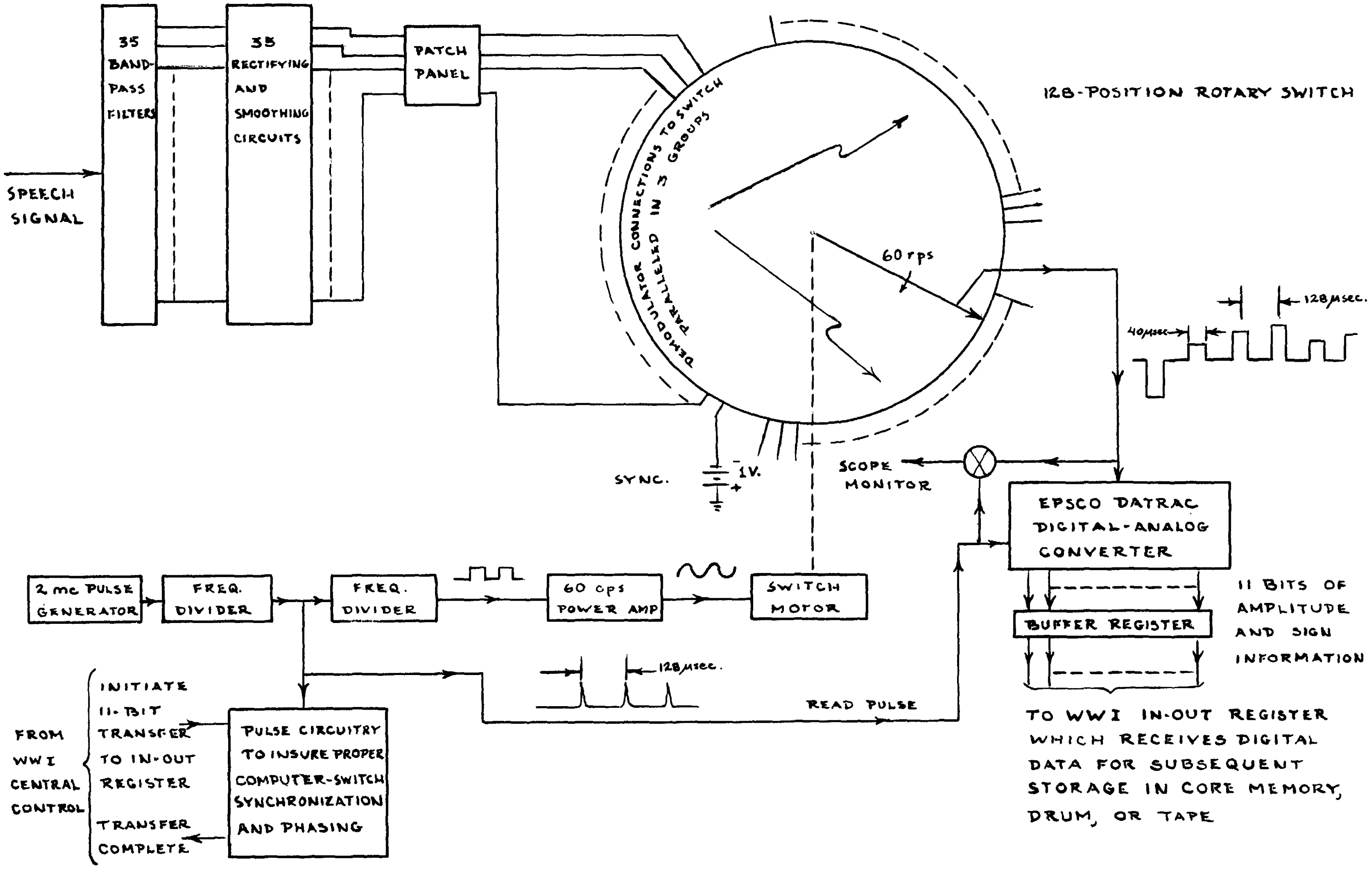
The purpose of this appendix is to outline briefly those details of the experimental procedure which may have affected the results. In addition, salient features of the identification logic are presented in schematic form.

Speech utterances subjected to analysis were first recorded on magnetic tape with the talker sitting in an anechoic chamber about one foot from the microphone. (For a more complete description of the informants, analog input equipment, and digital computer (Whirlwind I) employed for the analysis, see Hughes (9).) The speech data actually stored in the computer were entirely of a short-time spectral nature. Outputs from a set of 35 band-pass filters were fullwave rectified, smoothed with a time constant of 5 msec, and sampled by means of a mercury-jet rotary switch. The time between successive samples of a given filter output was 5. 5 msec. Level quantization was achieved by use of an Epsco Datrac analog-to-digital converter whose output was thus a ten-bit binary representation of the already time- and frequency-quantized speech spectrum. During read-in, raw spectral data were fed to the computer at a rate of approximately 8000 bits per second. The band limits for each of the filters are given in Table XIII. A summary of the real-time input system is shown by the diagram of Fig. 4.

The data were first stored in the core memory of the computer. After examining a scope display of the speech waveform envelope (also stored during read-in) in order to verify that the correct time segment had been sampled, the data were stored in digital form on magnetic tape for future analysis. Because of the limited size of the rapid-access memory, 16,000 msec was the longest duration of a signal that could be processed per read-in operation.

Figure 5 shows the over-all sequence of operations performed on each utterance from the time the digital spectral data were called in from magnetic tape storage until a set of symbols was printed on the high-speed (Analex) printer. After the data were transferred from magnetic tape to drum storage, two groups of 35 registers (each group representing the spectrum of approximately 5. 5 msec of speech) were called into core memory, averaged together, and frequency weighting applied. This averaged, weighted group of 35 registers may be considered as one (11 msec) look at the speech spectrum. Weighting factors for each frequency are given in Table XIII. Channels 13 and below were subjected to a de-emphasis of approximately 3 dv per octave. This was done chiefly to help avoid confusion in the formant tracking portion of the program between the first (lowest frequency) formant and the first or second harmonic of the voicing frequency. The lower order components of voicing are particularly prominent in the data from female speakers. Channels in the vicinity of 1800 cps were pre-emphasized to lower confusion between formant 2 and formants 1 or 3.

The first step in the analysis program was to extract as much information as possible from each averaged spectral sample, that is, make judgments as to formant positions,



SPEECH SIGNAL
35 BAND-PASS FILTERS
35 RECTIFYING AND SMOOTHING CIRCUITS
PATCH PANEL
DEMODULATOR CONNECTIONS TO SWITCH PARALLELED IN 3 GROUPS
128-POSITION ROTARY SWITCH
60 rps
SYNC.
1V.
SCOPE MONITOR
40μsec.
128μsec.
EPSCO DATRAC DIGITAL-ANALOG CONVERTER
BUFFER REGISTER
11 BITS OF AMPLITUDE AND SIGN INFORMATION
TO WWI IN-OUT REGISTER WHICH RECEIVES DIGITAL DATA FOR SUBSEQUENT STORAGE IN CORE MEMORY, DRUM, OR TAPE
2 mc PULSE GENERATOR
FREQ. DIVIDER
FREQ. DIVIDER
60 cps POWER AMP
SWITCH MOTOR
128μsec.
READ PULSE
FROM WWI CENTRAL CONTROL
INITIATE 11-BIT TRANSFER TO IN-OUT REGISTER
TRANSFER COMPLETE
PULSE CIRCUITRY TO INSURE PROPER COMPUTER-SWITCH SYNCHRONIZATION AND PHASING

Fig. 4. Analog-to-digital system.

Table XIII. Characteristics of bandpass filters.

Channel Number	Band Limits (cps)*	Center Frequency (cps)	Bandwidth (cps)	Weighting
1	115-165	140	50	0. 386
2	165-215	190	50	0. 450
3	215-265	240	50	0. 505
4	265-315	290	50	0. 555
5	315-365	340	50	0. 602
6	365-415	390	50	0. 644
7	415-465	440	50	0. 685
8	465-520	493	58	0. 725
9	520-580	550	60	0. 765
10	580-645	613	65	0. 808
11	645-720	683	75	0. 853
12	720-800	760	80	0. 900
13	800-890	845	90	0. 950
14	890-990	940	100	1. 000
15	990-1100	1045	110	1. 000
16	1100-1230	1165	130	1. 100
17	1230-1370	1300	140	1. 200
18	1370-1530	1450	160	1. 300
19	1530-1700	1615	170	1. 400
20	1700-1900	1800	200	1. 500
21	1900-2150	2025	250	1. 400
22	2150-2400	2275	250	1. 300
23	2400-2700	2550	300	1. 200
24	2700-3000	2850	300	1. 100
25	3000-3350	3175	350	1. 000
26	3350-3750	3550	400	1. 000
27	3750-4200	3975	450	1. 000
28	4200-4700	4450	500	1. 000
29	4700-5250	4975	550	1. 000
30	5250-5850	5550	600	1. 000
31	5850-6500	6175	650	1. 000
32	6500-7250	6875	750	1. 000
33	7250-8100	7675	850	1. 000
34	8100-9000	8550	900	1. 000
35	9000-10, 000	9500	1000	1. 000

*The frequency at which attenuation relative to passband is 1 db.

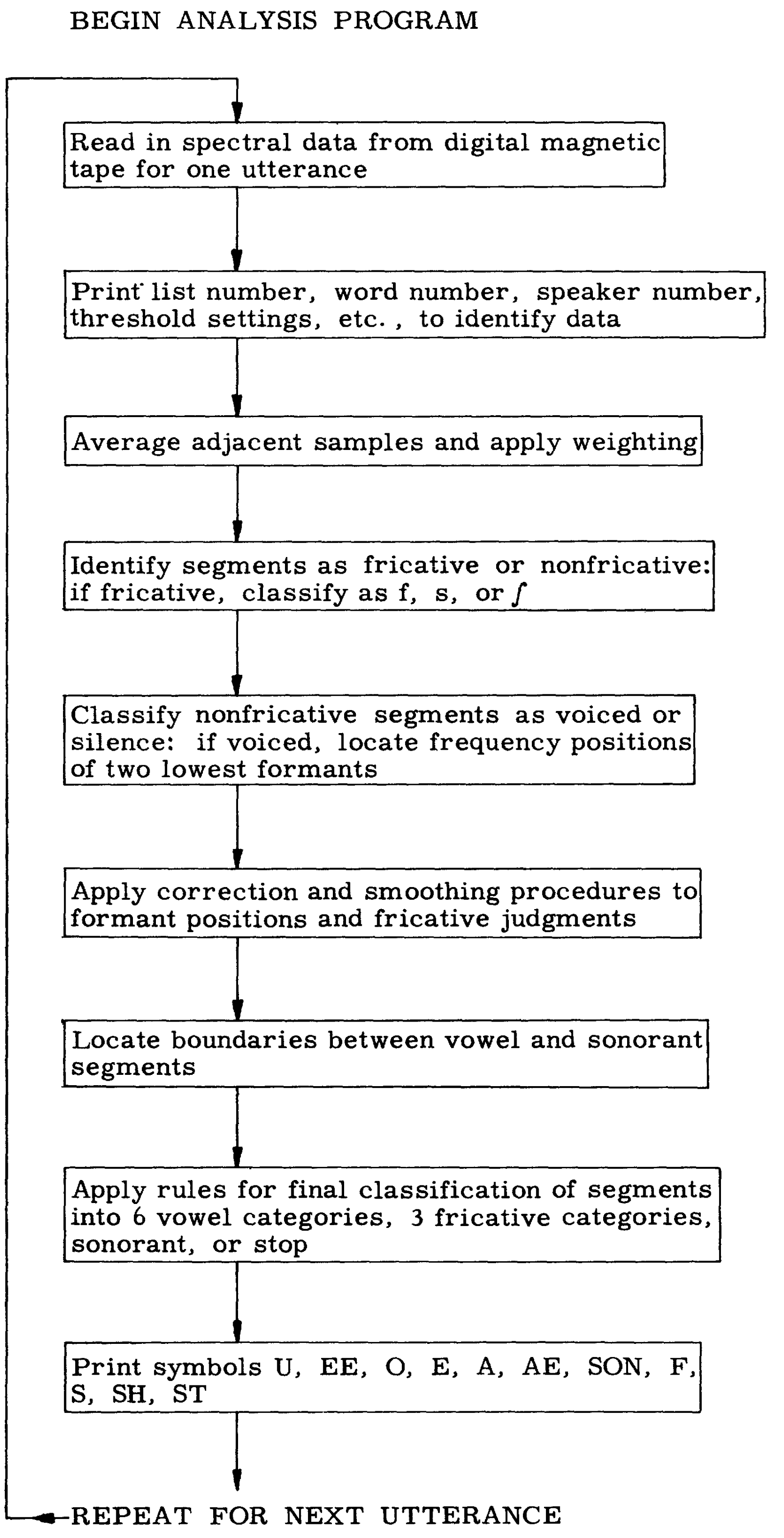

BEGIN ANALYSIS PROGRAM
Read in spectral data from digital magnetic tape for one utterance
Print list number, word number, speaker number, threshold settings, etc. , to identify data
Average adjacent samples and apply weighting
Identify segments as fricative or nonfricative: if fricative, classify as f, s, or ʃ
Classify nonfricative segments as voiced or silence: if voiced, locate frequency positions of two lowest formants
Apply correction and smoothing procedures to formant positions and fricative judgments
Locate boundaries between vowel and sonorant segments
Apply rules for final classification of segments into 6 vowel categories, 3 fricative categories, sonorant, or stop
Print symbols U, EE, O, E, A, AE, SON, F, S, SH, ST
REPEAT FOR NEXT UTTERANCE

Fig. 5. General sequence of operations in analysis program.

fricative class, level, and so forth, for each 11.1-msec interval of real time. This reduced form of the data was stored in various groups of registers for subsequent use and/or transformation. These groups were each 100 registers in number (to accommodate the longest possible utterance), each register in a group corresponding to a given real-time segment. It will be convenient to refer to these groups as Group A, B, C, and so forth. The rest of the programmed procedure may then be considered to be the application of corrections, smoothing, and criteria for the identification of the various segments with the proper printed linguistic symbols.

2. PORTIONS OF THE ANALYSIS PROGRAM OPERATING DIRECTLY ON SPECTRAL DATA

The raw spectral data for an utterance were brought into the computer core memory only once. Measurements were performed on these data on a look-by-look basis. Only after measurements were performed on each look separately and independently was any attempt made to extend the influence of results on one time segment to those on nearby segments. Figure 6 shows procedures imposed on each block of 35 registers containing numbers (representing the weighted spectrum of an 11-msec interval of speech signal) immediately after they were brought into core memory.

After storing the sum of channels 9-35 in register D_i (used later in the boundary program), the first step was to classify the look as fricative or nonfricative. This was done by subtracting the sum of 10 channels in the F1 region from the 10 highest frequency channels and comparing the result with a threshold.

Those looks classified as fricative were further subclassified as /f/, /s/, or /ʃ/. The three measurements used to make this tripartite division were as follows:

<u>Measurement I</u>. A number a is calculated as the difference of two numbers b and c. Numbers b and c are such that:

$\frac{D}{2} \leqslant$ [sum of the squares of channels 26-35] $2^b < D$

and

$\frac{D}{2} \leqslant$ [sum of the squares of channels 12-35] $2^c < D$ where D is the largest number stored in a WWI register. If $a < 2$ proceed to measurement II. If $a \geqslant 2$ proceed to measurement III.

<u>Measurement II</u>. A number a' is calculated as the difference of two numbers b' and c'. Numbers b' and c' are such that:

$\frac{D}{2} \leqslant$ [sum of the squares of channels 12-31] $2^{b'} < D$

and

$\frac{D}{2} \leqslant$ [sum of the squares of channels 12-21] $2^{c'} < D$. If $a' < 4$, look is classified as /f/. If $a' \geqslant 5$, look is classified as /s/.

<u>Measurement III</u>. Channels 19-26 are examined to locate the spectral maximum in this region. A number a"is calculated as the difference of two numbers b" and c". Numbers b" and c" are such that:

$\frac{D}{2} \leqslant$ [sum of the squares of channels 12-17] $2^{b''} < D$

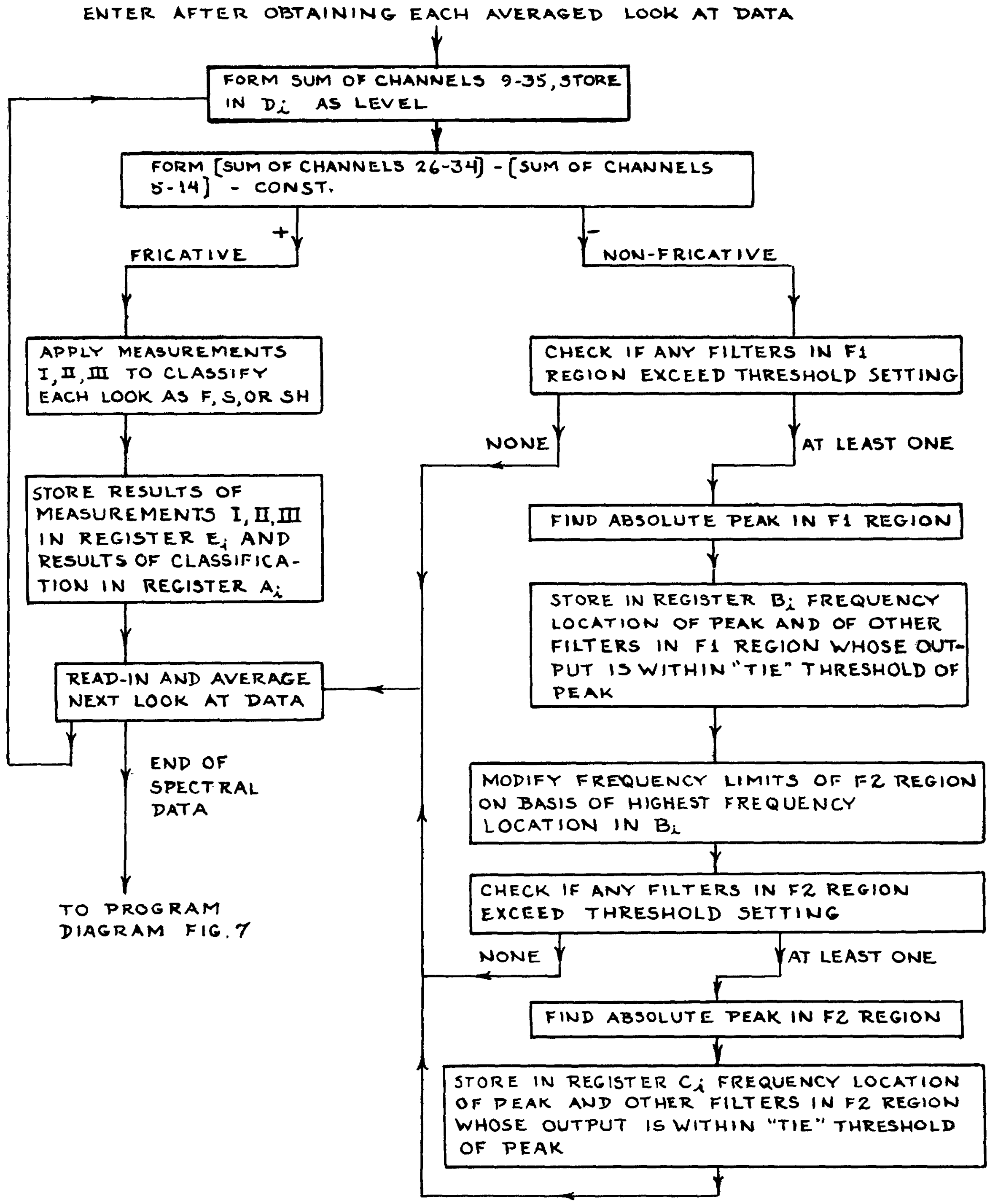

ENTER AFTER OBTAINING EACH AVERAGED LOOK AT DATA
FORM SUM OF CHANNELS 9-35, STORE IN Di AS LEVEL
FORM [SUM OF CHANNELS 26-34] - [SUM OF CHANNELS 5-14] - CONST.
+
FRICATIVE
-
NON-FRICATIVE
APPLY MEASUREMENTS I, II, III TO CLASSIFY EACH LOOK AS F, S, OR SH
STORE RESULTS OF MEASUREMENTS I, II, III IN REGISTER Ei AND RESULTS OF CLASSIFICATION IN REGISTER Ai
READ-IN AND AVERAGE NEXT LOOK AT DATA
END OF SPECTRAL DATA
TO PROGRAM DIAGRAM FIG. 7
CHECK IF ANY FILTERS IN F1 REGION EXCEED THRESHOLD SETTING
NONE
AT LEAST ONE
FIND ABSOLUTE PEAK IN F1 REGION
STORE IN REGISTER Bi FREQUENCY LOCATION OF PEAK AND OF OTHER FILTERS IN F1 REGION WHOSE OUTPUT IS WITHIN "TIE" THRESHOLD OF PEAK
MODIFY FREQUENCY LIMITS OF F2 REGION ON BASIS OF HIGHEST FREQUENCY LOCATION IN Bi
CHECK IF ANY FILTERS IN F2 REGION EXCEED THRESHOLD SETTING
NONE
AT LEAST ONE
FIND ABSOLUTE PEAK IN F2 REGION
STORE IN REGISTER Ci FREQUENCY LOCATION OF PEAK AND OTHER FILTERS IN F2 REGION WHOSE OUTPUT IS WITHIN "TIE" THRESHOLD OF PEAK

Fig. 6. Portion of program operating directly on spectral data.

and
$\frac{D}{2} \leqslant$ [sum of squares of 3 channels surrounding spectral maximum] $2^{c''} < D$. If $a'' < 4$, look is classified as /f/. If $a'' \geqslant 4$, look is classified as /ʃ/.

The results of measurements performed (a, a', a") were stored in group E to facilitate error evaluation and the classification stored in group A for later use.

If a look was classified nonfricative, information was extracted concerning the frequency location of peaks or relatively large concentrations of energy. Channels 4-13 were taken to encompass the first formant (F1) range for male speakers. For female speakers channels 6-14 defined the F1 range. If no filter output in the F1 frequency range exceeded a fixed (existence) threshold, the examination of the spectral data representing that particular time segment was terminated. The lack of energy in either the high or low frequency regions was taken to indicate the presence of a look of silence. The threshold of minimum filter output was approximately 4 per cent of the maximum filter output reached during the utterance.

If the F1 threshold was exceeded, the channel number of the filter in the F1 region whose output was the largest, together with the channel numbers of any other filters whose outputs were within the tie threshold of the absolute peak, was stored in group B. The tie threshold was (arbitrarily) fixed at approximately 2 per cent of maximum filter output during the utterance. It was found that more accurate smoothing and correction of formant positions in subsequent sections of the over-all formant-tracking procedure was possible if information as to possible alternatives to the absolute peak was furnished.

The channels 12-23 for male speakers and channels 14-24 for female speakers were taken to nominally encompass the second formant (F2) region. However, for a given look, the lowest-frequency filter from which the F2 peak and ties were selected was not allowed to be less than 4 channels above the highest frequency F1 tie previously stored. This precaution is necessary since the absolute limits of the F1 and F2 frequency ranges must be allowed to overlap, and for almost all vowel spectra the amplitude of F1 is greater than that of F2. The existence and tie thresholds for F2 were reduced to approximately 2 per cent and 1 per cent of maximum filter output respectively since vowel spectra generally fall in amplitude with frequency. Except for these changes, the procedure for locating the absolute peak and ties was identical to that followed for F1.

The extraction of level, fricative, and formant information from successive averaged looks was repeated until the end of the spectral data for one utterance was reached. The remaining analysis on the utterance consisted of interpreting the numbers derived from this process which were now stored in register groups A, B, C, and D.

3. SMOOTHING OPERATIONS

Following the operations described above which used the spectral data directly, a sequence of smoothing and correction procedures was initiated. These procedures operate on the numbers stored in register groups A, B, and C, leaving the final judgments for class of fricative segment, presence of silence, or frequency position of formants 1

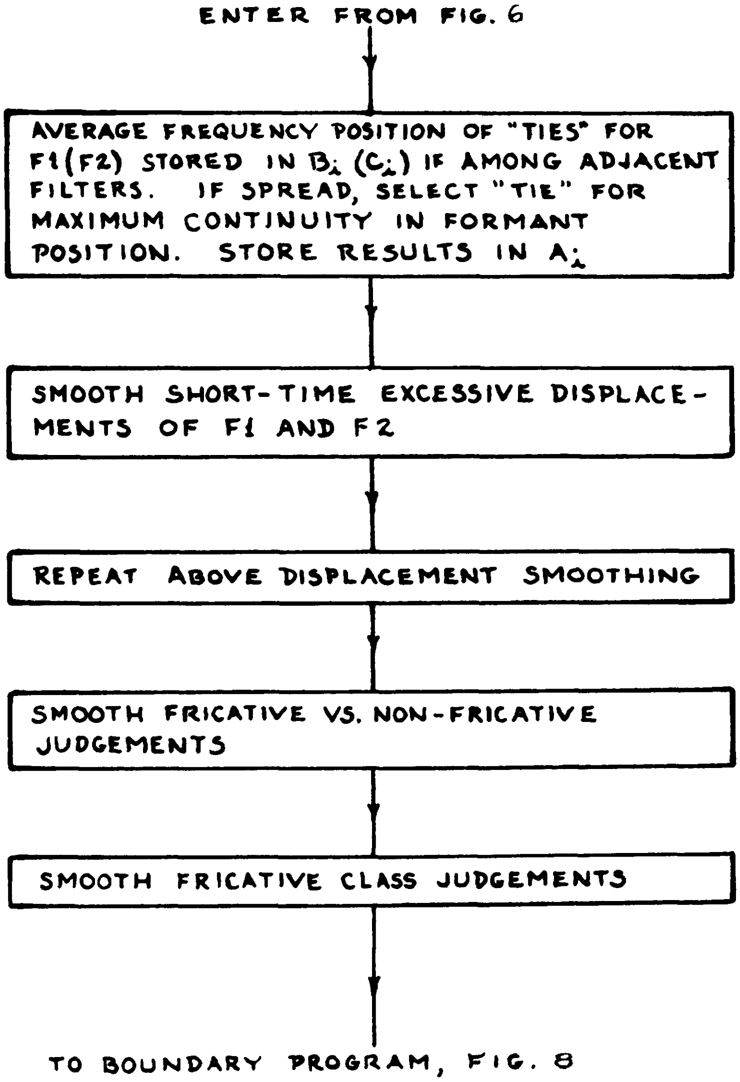

ENTER FROM FIG. 6
AVERAGE FREQUENCY POSITION OF "TIES" FOR F1(F2) STORED IN B_i (C_i) IF AMONG ADJACENT FILTERS. IF SPREAD, SELECT "TIE" FOR MAXIMUM CONTINUITY IN FORMANT POSITION. STORE RESULTS IN A_i
SMOOTH SHORT-TIME EXCESSIVE DISPLACEMENTS OF F1 AND F2
REPEAT ABOVE DISPLACEMENT SMOOTHING
SMOOTH FRICATIVE VS. NON-FRICATIVE JUDGEMENTS
SMOOTH FRICATIVE CLASS JUDGEMENTS
TO BOUNDARY PROGRAM, FIG. 8

Fig. 7. Smoothing of fricative and formant judgments.

and 2 in group A. Refer to the block diagram of Fig. 7.

The first step in smoothing procedure was to search groups B and C for the presence of ties, that is, locate looks at the spectrum showing no single filter output clearly maximum in either or both the F1 and F2 frequency ranges. These ties, when present, were eliminated by choosing a channel number to represent the formant position in one of two ways:

(a) If the ties were among adjacent channels, the average location was calculated and stored in the group A register corresponding to that look. The number of possible (discrete) formant-frequency positions thus was 2n-1, where n was the number of channels examined.

(b) If the ties were not among adjacent channels, the one whose frequency location was closest to that selected to represent the formant position for the previous (or following) look was selected.

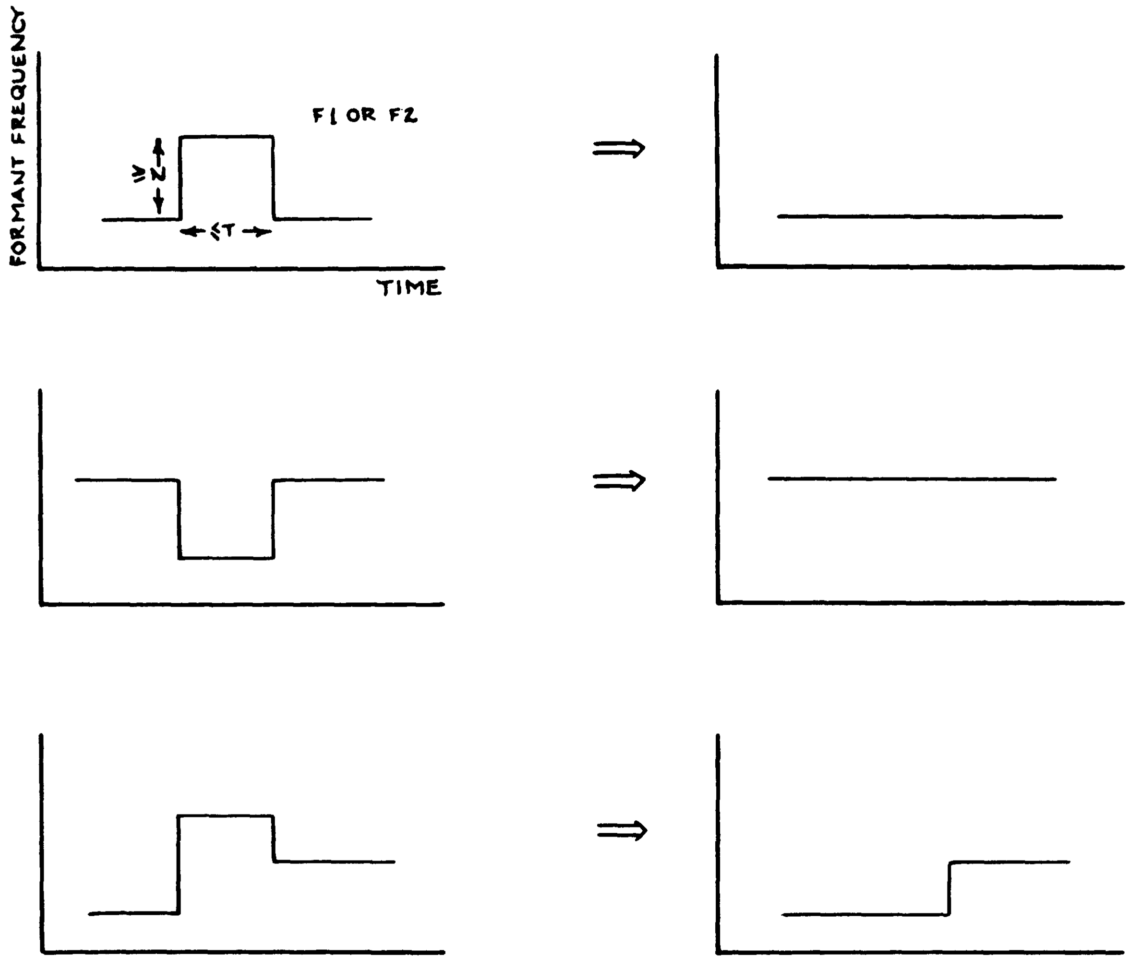
The formant positions (numbers stored in group A) were next subjected to correction or smoothing of large jumps of opposite sign spaced closely in group A, that is, short-time disturbances in formant-position continuity. This procedure was an attempt to include to some extent the constraint on formant motion imposed by the inertia of the vocal organs.

It was very important to include some formant-smoothing procedure in the program because of the random departures of the raw spectral data from the idealized vowel spectrum curves. Two effects were particularly noticeable: (a) A formant might suddenly weaken considerably in intensity for a short period (probably because of changes in the glottal spectrum), causing the simple peak-picking operation to find a second order maximum in the formant region or perhaps track an adjacent formant until the true formant returned to normal amplitude. (b) During vowels exhibiting closely spaced formants such as /a/ or /i/, the lower-frequency formant does not always have the higher intensity. Again a simple peak-picking operation may jump randomly between adjacent formants, depending on how the allowed formant–frequency ranges are chosen. Jump correction is not entirely adequate to deal with this difficulty. This is discussed in Section III. However, it often corrected many of the confusions between adjacent formants if they did not persist too long in time.

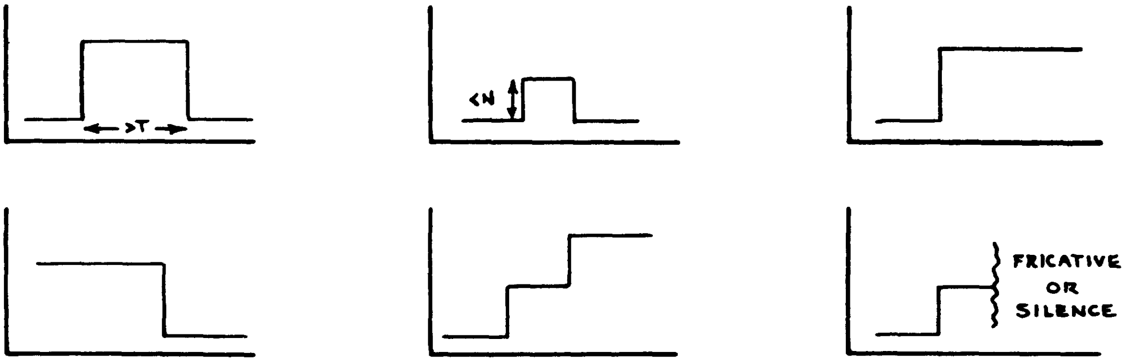
Although sonagrams (even of rapid speech) never exhibit large discontinuities in formant position closely spaced in time, numerical specification of this fact for purposes of programming is difficult. For this reason, formant tracking data were compiled before introducing this constraint into the program to determine what sort of false formant jumps were caused by imperfections in the previous stages of the system. Analyses of this data resulted in the formulation of correction procedures summarized in Table XIV and the specification of the constants N = 3 filters (about 1/3 octave in frequency) and T = 44 msec (4 looks at the spectrum).

It was found experimentally that complex patterns of (false) F1 or F2 discontinuities were not completely smoothed by this procedure. In many such cases a simple repetition

Table XIV. Jump corrections.



etc.



etc.

of this part of the program improved the results. Since only rarely did the smoothing "correct" true variations in formant positions, the double correction procedure was incorporated as a permanent feature of the formant tracking program.

Two smoothing procedures related to looks judged as fricative were instituted. The first was designed to eliminate isolated nonfricative judgments during fricative portions of a sound as well as isolated fricative judgments during nonfricative portions. During the previous look-by-look analysis, the fricative vs nonfricative classification of each look, together with that of both the preceding and following looks, was stored in register A_i. Final judgment was then formulated in accordance with the rules given in Table XV.

Table XV. Fricative vs nonfricative smoothing.

A_{i-1}	A_i	A_{i+1}	Smoothed Judgment Made on A_i
F	F	F	F
NF	F	F	F
F	F	NF	F
F	NF	F	F
F	NF	NF	F
NF	F	NF	NF
NF	NF	F	NF
NF	NF	NF	NF

The second smoothing procedure classified entire segments of speech as either /f/, /s/, or /ʃ/. Only rarely did all the individual classifications of each look in a segment agree. A segment was given the same classification as had been given the largest number of individual looks within that segment. If two or more classifications were distributed equally, arbitrary preference was given to the classes /ʃ/, /s/, /f/, in that order.

4. DETERMINATION OF BOUNDARIES BETWEEN SONORANTS AND VOWELS

Boundaries between sonorant and vowel segments of an input speech signal were determined from two parameters:

(a) Changes in level (linear sum of channels 9-35) over a span of 5 looks or 55 msec of time.

(b) Shift in both F1 and F2 frequency positions over a span of 3 looks or 33 msec of time.

The boundary program logic is outlined in Fig. 8. For every 5.5-msec spectral look at the input, two numbers were formed to represent the change in level and formant

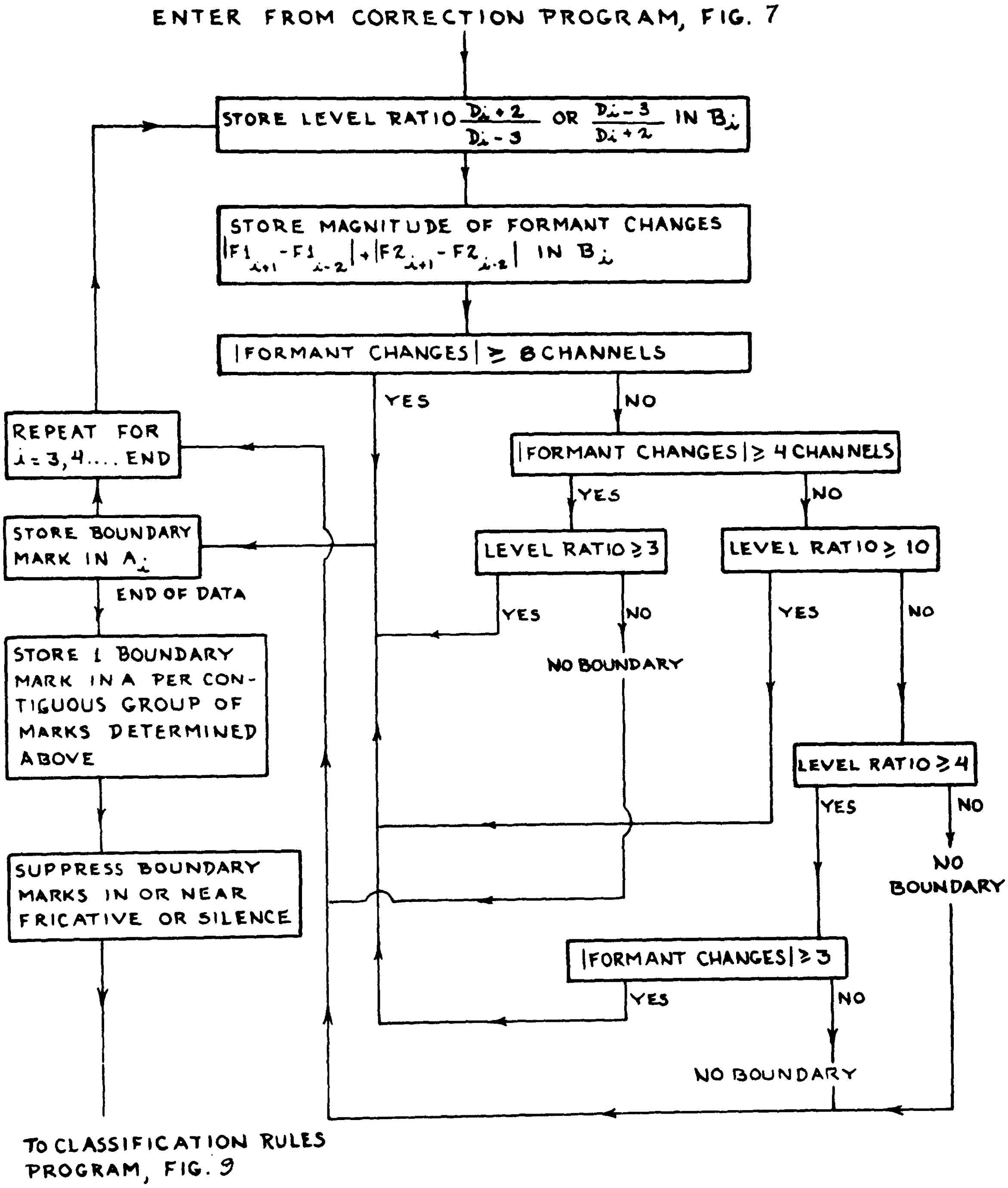

ENTER FROM CORRECTION PROGRAM, FIG. 7
STORE LEVEL RATIO $\frac{D_{i+2}}{D_{i-3}}$ OR $\frac{D_{i-3}}{D_{i+2}}$ IN B_i
STORE MAGNITUDE OF FORMANT CHANGES $|F1_{i+1} - F1_{i-2}| + |F2_{i+1} - F2_{i-2}|$ IN B_i
|FORMANT CHANGES| ≥ 8 CHANNELS
YES
NO
REPEAT FOR $i = 3, 4 \ldots$ END
|FORMANT CHANGES| ≥ 4 CHANNELS
YES
NO
STORE BOUNDARY MARK IN A_i
LEVEL RATIO ≥ 3
LEVEL RATIO ≥ 10
END OF DATA
YES
NO
YES
NO
NO BOUNDARY
STORE 1 BOUNDARY MARK IN A PER CONTIGUOUS GROUP OF MARKS DETERMINED ABOVE
LEVEL RATIO ≥ 4
YES
NO
NO BOUNDARY
SUPPRESS BOUNDARY MARKS IN OR NEAR FRICATIVE OR SILENCE
|FORMANT CHANGES| ≥ 3
YES
NO
NO BOUNDARY
TO CLASSIFICATION RULES PROGRAM, FIG. 9

Fig. 8. Determination of boundaries between sonorants and vowels.

frequency shift respectively. The first was the ratio of level (stored in group D) two looks ahead to that of three looks behind (or the inverse such that this number was < 1). The second was calculated from the smoothed formant channel numbers (stored in group A) and was the sum of the magnitude of the change in F1 and the magnitude of the change in F2 from two looks behind to one look ahead. After these numbers were calculated for the entire utterance and stored in group B, the boundary existence criteria were applied. The information as to whether the level change was upward or downward with time was also stored and used to classify a boundary as sonorant-vowel or vowel-sonorant. Channels 1-8 were omitted from the original level calculation to improve the detection of boundaries between final nasal consonants and preceding diffuse vowels; both of the latter have high concentrations of energy in the very low frequencies but may differ markedly in the amount of energy present above 500 cps.

It will be noted from Fig. 8 that there were 4 possible paths to the indication of a boundary.

(a) Large formant shift. The formant change over 3 looks was ⩾ 8 channels (approximately 1 octave).

(b) Large level change. The ratio of levels was ⩾ 10. Due to the method used to formulate the so-called level used here, it would be difficult to interpret this ratio in terms of standard units such as the db.

(c) Marginal formant shift, marginal level change. The formant change was ⩾ 4 channels and the level ratio ⩾ 3.

(d) Marginal level change, marginal formant shift. The level ratio was ⩾ 4 and the formant shift ⩾ 3 channels.

If one of the above 4 conditions was met, a boundary mark was stored in register A_i. A single boundary between vowel and sonorant segments of the speech input usually resulted in a group of from 2-8 contiguous boundary marks in group A. It was found that very satisfactory location of the single boundary could be made by simply averaging the positions of each such group.

The final operation of this part of the program was the erasure of all boundary marks within 44 msec of a segment established as fricative or silence. These boundaries were already signaled by virtue of the change in classification, thus the concomitant variations in level or formant shifts were redundant.

5. CLASSIFICATION OF SEGMENTS

The final classification of segments of the input speech signal into linguistic categories was based on the information assembled in register group A. In summary this was:

(a) The filter channel numbers representing the positions of the first and second formant of vowel segments.

(b) The presence of a fricative and its classification.

(c) The presence of a sonorant-vowel (CV) or vowel-sonorant (VC) boundary.

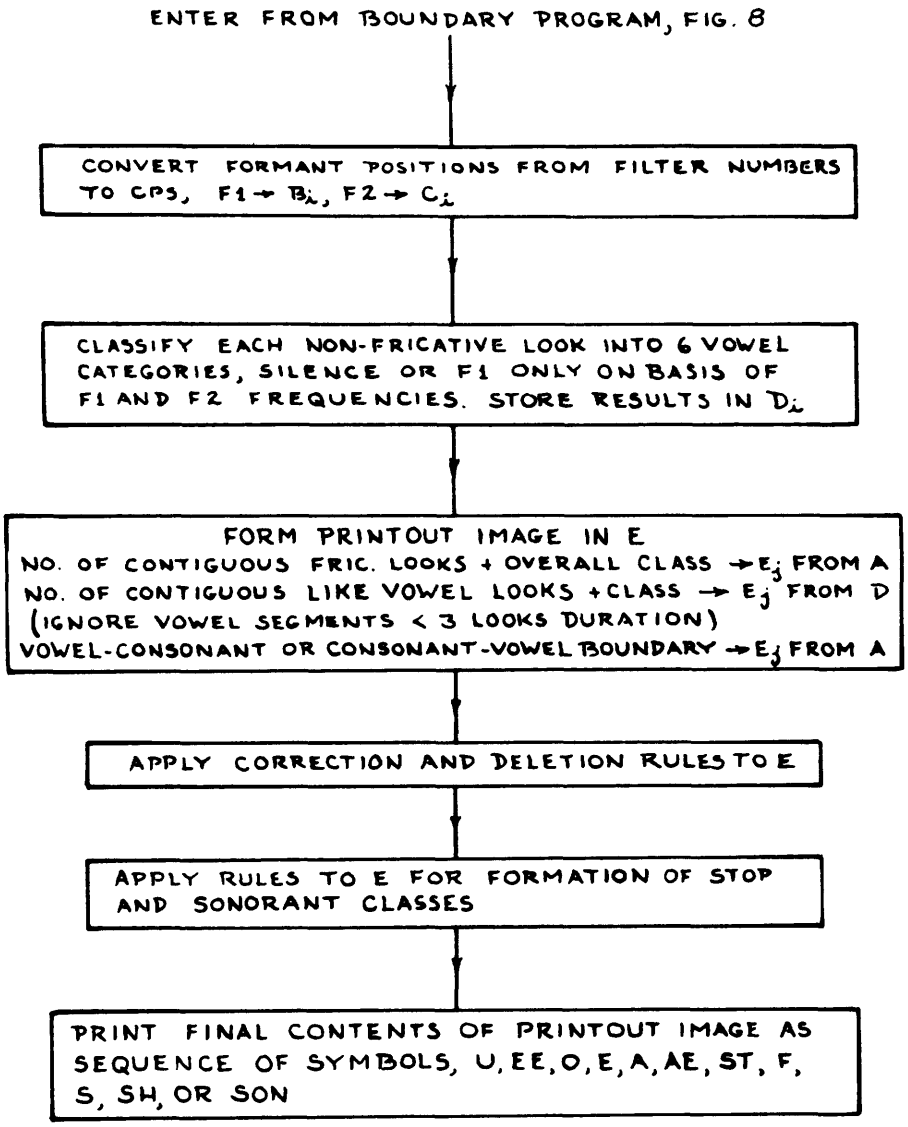

ENTER FROM BOUNDARY PROGRAM, FIG. 8
CONVERT FORMANT POSITIONS FROM FILTER NUMBERS TO CPS, F1 → B_i, F2 → C_i
CLASSIFY EACH NON-FRICATIVE LOOK INTO 6 VOWEL CATEGORIES, SILENCE OR F1 ONLY ON BASIS OF F1 AND F2 FREQUENCIES. STORE RESULTS IN D_i
FORM PRINTOUT IMAGE IN E
NO. OF CONTIGUOUS FRIC. LOOKS + OVERALL CLASS → E_j FROM A
NO. OF CONTIGUOUS LIKE VOWEL LOOKS + CLASS → E_j FROM D
(IGNORE VOWEL SEGMENTS < 3 LOOKS DURATION)
VOWEL-CONSONANT OR CONSONANT-VOWEL BOUNDARY → E_j FROM A
APPLY CORRECTION AND DELETION RULES TO E
APPLY RULES TO E FOR FORMATION OF STOP AND SONORANT CLASSES
PRINT FINAL CONTENTS OF PRINTOUT IMAGE AS SEQUENCE OF SYMBOLS, U, EE, O, E, A, AE, ST, F, S, SH, OR SON

Fig. 9. Final classification of segments into vowel categories, fricative, sonorant or stops.

Figure 9 outlines the general procedures for deriving single categories that describe gross segments of the input data.

After converting the formant positions from channel numbers to frequency in cps, each nonfricative look was classified into one of eight vowel categories. This classification was based on the absolute frequency of the first formant and the difference in cps between the position of the second and first formants. Table XVI summarizes the position of the second and first formants. Table XVI summarizes the rules employed.

Table XVI. Vowel clasification rules.

Vowel Class		Fl (cps)	F2-Fl (cps)
0.	V0 (silence)	Not present	Not present
1.	Vl (no F2)	Anywhere in Fl range	Not present
2.	U	< 450	> 850
3.	EE	< 450	< 850
4.	O	Between 450 and 650	> 850
5.	E	Between 450 and 650	< 850
6.	A	> 650	> 850
7.	AE	> 650	< 850

This was the last operation performed on any spectral data (derived or direct) on a look-by-look basis. All following procedures were aimed at reducing the data step-by-step until one symbol could be printed to represent the classification of each segment of the original utterance. For this purpose a group of registers in the computer termed "printout image" (group E) was set up. All rules for formation of classes, correction, deletion, and so forth, were performed on the sequence of numbers in this group. Finally, these numbers were translated into orders to print alphabetic symbols on the high-speed Analex printer that was available.

Formulation of the printout image began by placing in successive registers one of the following pieces of information, numerical order of the registers corresponding to order of occurrence in the utterance.

(a) Class of fricative and number of contiguous looks duration.

(b) Class of vowel and number of contiguous looks duration. Vowel segments of any class determined to be less than three looks duration were ignored to help reduce the appearance of random variations or transitions of short duration.

(c) Presence of VC or CV boundary.

The.following thirteen rules were programmed to operate on the contents of the printout image:

(a) Coalesce adjacent registers that indicated identical classes, that is, AE (4 looks), AE (3 looks) ⇒ AE (7 looks), VC, VC ⇒ VC, and so forth.

(b) Of the vowel-class segments enclosed by two boundaries, delete those adjacent to the boundaries that are of less duration than one-fifth the total number of enclosed looks. For the purpose of this rule, "boundaries" were defined as the beginning or end of a word, fricative, VO, VC or CV.

(c) Delete CV if separated from preceding fricative by less than four vowel-class looks. Delete VC if separated from following fricative by less than four vowel-class looks.

The following four rules complete the specification of the sonorant segments:

(d) Delete final V0 segments. Final VC ⇒sonorant. Initial CV ⇒ sonorant. Initial V0 and/or V1 followed by CV⇒ sonorant.

(e) VC-CV⇒ sonorant. VC-vowel Class-CV⇒ sonorant.

(f) If vowel segments between preceding boundary (fricative, sonorant, beginning or end of word, V0, or V1) and CV changed by less than two classes (see Table XVI), delete vowels. Follow same procedure for VC and following boundary. Finally, all CV and VC⇒ sonorant.

(g) Initial V1 of greater duration than five looks and followed by vowel⇒ sonorant, otherwise delete initial V1.

The following five rules completed the specification of the fricative and stop segments:

(h) Initial V0⇒ fricative (/f/) if of greater than five looks duration and followed by /f/ or vowel.

(i) Initial fricative⇒ stop if of less than five looks duration.

(j) Delete all fricative of duration less than two looks not immediately followed by a vowel.

(k) V0 of n_o looks duration followed by fricative of n_f looks duration: ⇒ stop if $n_o \geqslant n_f \leqslant 6$, or⇒ stop + fricative if $n_o < n_f$.

(l) Non-final V0⇒ stop if followed by a vowel.

(m) Delete all remaining segments of class V0 and V1.

After all thirteen rules had been applied to the printout image, each register in turn was interpreted as a single output symbol, the printing of which completed the computer analysis of the utterance. Thus, the final vocabulary of output symbols in terms of which all data was described was: U, EE, O, E, A, AE, SON (sonorant), ST (stop).

6. WORDS AND SPEAKERS USED

Two different word lists, read by a total of seven speakers, constituted the input data used in this study. The speakers were instructed to read each word distinctly at normal conversational speed but with a 5-second pause between words and no drop in level or pitch at the end. These instructions proved to be difficult ones, as almost all of the speakers became nervous and slurred their speech somewhat. Also it is almost impossible to overcome the natural tendency to drop pitch and level while reading items from a list.

Word List No. 1 (51-100) contains monosyllables, and List No. 2 (1-50), polysyllables.

Word List No. 1

51	SASH	sæʃ	76	LOOM	lum
52	ZOO	zu	77	KNOWN	noʊn
53	FOOT	fʊt	78	RANG	ræŋ
54	SAW	sɔ	79	MOON	mun
55	FEZ	fɛz	80	WAN	wɑn
56	AH	ɑ	81	MEN	mɛn
57	VERSE	vɝs	82	LAWN	lɔn
58	SHOVE	ʃʌv	83	KNELL	nɛl
59	FEE	fi	84	MAN	mæn
60	SHOWS	ʃoʊz	85	NUMB	nʌm
61	IF	ɪf	86	MAUL	mɔl
62	NOW	naʊ	87	WELL	wɛl
63	I	aɪ	88	MOLE	mol
64	JOY	dʒoɪ	89	NIM	nɪm
65	A	eɪ	90	MOLL	mɑl
66	RING	rɪŋ	91	MILL	mɪl
67	LOIN	loɪn	92	RAIL	reɪl
68	MAIN	meɪn	93	OWL	aʊl
69	WRONG	rɔŋ	94	WOUND	wund
70	MILE	maɪl	95	KNURL	nɝl
71	RUNG	rʌŋ	96	WOOL	wʊl
72	LIME	laɪm	97	MULL	mʌl
73	ROAM	roʊm	98	WEEMS	wimz
74	WORM	wɝm	99	WON	wʌn
75	MEAN	min	100	KNEEL	nil

Word List No. 2

1 COUGHDROP	kɔfdrap	26 IN AWE	ɪnɔ
2 DOVE TAIL	dʌvteɪl	27 KEY GANG	kigæŋ
3 EARTHQUAKE	ɚθkweɪk	28 NOW BABE	naʊbeɪb
4 ROOK GOOCH	rʊkgutʃ	29 YOU THOUGH	juðoʊ
5 MUTE DOLL	mjutdal	30 BUY FOAM	baɪfom
6 ANDES	æ ndiz	31 SHOE BOOTH	ʃu buθ
7 GHANA	gana	32 H. I.	eɪ tʃ aɪ
8 JIM OZ	dʒ ɪmaz	33 THEY KNEW	ðeɪ nu
9 OWL EGG	aʊleɪg	34 OH JOY	oʊ dʒ oɪ
10 WOO PA	wupa	35 MAUL OVER	mɔlovɚ
11 UNCLE	ʌŋkəl	36 VOO DOO	vudu
12 KOW TOW	kaʊtaʊ	37 LONG AGO	lɔŋəgo
13 COQUETTE	koʊkɛt	38 HAT CHECK	hættʃɛk
14 AZURE	æʒɚ	39 ICE EDGE	aɪsɛdʒ
15 POOPING	pupɪŋ	40 SHIP OUT	ʃɪpaʊt
16 AUGER	ɔgɚ	41 ENERGY	ɛnɚdʒi
17 THE VEEP	ðəvip	42 TO PUFF	tupʌf
18 OOZY	uzi	43 LIGHT ON	laɪtɔn
19 OIL ASH	oɪlæʃ	44 FEW FISH	fjufɪʃ
20 EAT EVE	itiv	45 AHEAD	əhɛd
21 WASH ROOM	waʃrum	46 SOY SAUCE	sɔɪsɔs
22 YE SHOULD	jiʃʊd	47 PAY DAY	peɪdeɪ
23 CHOW FUZZ	tʃaʊfʌz	48 BEE HIVE	bihaɪv
24 AH DEATH	adɛθ	49 THESIS	θisɪs
25 WEE YAM	wijæm	50 RESCUE	reskju

These words, together with the broad phonetic transcription (in I. P. A. symbols), were used to formulate the results of Section III.

A brief description of the speakers follows. All speakers read both lists except No. 5, who read only words 51-100.

Speaker 1	Male	Low vocal pitch
Speaker 2	Male	Medium vocal pitch
Speaker 3	Female	High vocal pitch
Speaker 4	Male	Very low vocal pitch
Speaker 5	Male	Low vocal pitch
Speaker 6	Male	Low vocal pitch
Speaker 7	Female	Medium vocal pitch

All speakers employed an Eastern United States dialect.

Acknowledgment

The author wishes to express his gratitude for the able assistance and continued inspiration given throughout this research program by Professor Morris Halle, Department of Modern Languages, and Research Laboratory of Electronics, M. I. T. Also, the helpful comments of Professor Peter Elias, Professor Robert M. Fano, and Professor Kenneth L. Stevens are very much appreciated.

References

1. E. C. Cherry, Roman Jakobson's distinctive features as the normal coordinates of language, For Roman Jakobson (Mouton and Company, The Hague, 1956), pp. 60-64.

2. E. C. Cherry, M. Halle, R. Jakobson, Toward the logical description of languages in their phonetic aspect, Language 29, 24-46 (1953).

3. K. H. Davis, R. Biddulph, S. Belashek, Automatic recognition of spoken digits, Communication Theory, Willis Jackson, editor (Butterworth's Scientific Publications, London, 1953), pp. 433-441.

4. C. G. M. Fant, On the predictability of formant levels and spectrum envelopes from formant frequencies, For Roman Jakobson (Mouton and Company, The Hague, 1956), pp. 109-121.

5. J. L. Flanagan, A difference limen for vowel formant frequency, J. Acoust. Soc. Am. 27, 613-617 (1955).

6. D. B. Fry, P. Denes, Mechanical speech recognition, Communication Theory, Willis Jackson, editor (Butterworth's Scientific Publications, London, 1953), pp. 426-432.

7. M. Halle, The strategy of phonemics, Word, 10, 197-209 (1954).

8. A. S. House, K. N. Stevens, Analog studies of the nasalization of vowels, J. Speech and Hearing Disord. 21, 218-232 (1956).

9. G. W. Hughes, On the Recognition of Speech by Machine, Sc. D. Thesis, Department of Electrical Engineering, M. I. T., September 1, 1959.

10. G. W. Hughes, M. Halle, Spectral properties of fricative consonants, J. Acoust. Soc. Am. 28, 303-310 (1956).

11. G. W. Hughes, M. Halle, Vowel identifier, Quarterly Progress Report No. 42, Research Laboratory of Electronics, M. I. T., October 1956, pp. 109-117.

12. R. Jakobson, C. G. M. Fant, M. Halle, Preliminaries to Speech Analysis, Technical Report No. 13, Research Laboratory of Electronics, M. I. T., 1952.

13. Ibid. p. 43.

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

15. D. Jones, The Phoneme, It's Nature and Use (W. Heffer and Sons, Cambridge, 1950).

16. H. F. Olson, H. Belar, Phonetic typewriter, J. Acous. Soc. Am. 28, 1072-1081 (1956).

17. F. Poza, Formant Tracking by Digital Computation, M. S. Thesis, Department of Electrical Engineering, M. I. T., September 1959.

18. J. Wiren, H. Stubbs, Electronic Binary Selection System for Phoneme Classification, J. Acous. Soc. Am. 28, 1082-1091 (1956).