Computer Recognition of Continuous Speech

by R. B. Purves* and W. J. Strong

Department of Physics and Astronomy, Brigham Young University, Provo, Utah 84602

Summary

A method of automatic speech recognition has been programmed on a small computer. The system accepts syntactic units of carefully spoken continuous speech from a single co-operative male speaker. The recognition parameters are low order cepstrum coefficients, zero crossing rate, slope change rate, cepstrum peak height and apparent place of articulation. The segmentation is performed using a "segmentation by recognition" method. Two phoneme choices are assigned to each segment. The utterance is identified by generating successive phoneme strings until one is found which satisfies the lexical and syntactic constraints. The lexical constraint requires the word string to consist only of phonemicon (phonemic dictionary) entries. The syntactic constraint requires the word string to satisfy a simplified English syntax. The phonemicon was built to contain about a thousand entries. Twenty utterances containing an average of 3.4 words were used to evaluate the system. Of these, 35 percent were correctly recognized without application of the syntactic constraint. Imposition of the syntactic constraint improved the recognition rate to 65 percent.

Computererkennung kontinuierlicher Sprache

Zusammenfassung

Auf einem Kleinrechner wurde eine Methode der automatischen Spracherkennung programmiert. Das System verarbeitet syntaktische Einheiten sorgfältig gesprochener kontinuierlicher Sprache eines gutwilligen männlichen Sprechers. Erkennungsparameter sind Cepstrumkoeffizienten niedriger Ordnung, Nulldurchgangsrate, Änderungsrate der Kurvensteigung, Höhe der Spitze und Lage gemäß der Artikulation im Cepstrum. Die Segmentierung wird mittels einer Methode der „Segmentierung durch Erkennung" durchgeführt. Jedem Segment sind zwei Phonemmöglichkeiten zugeordnet. Die Sprechweise wird identifiziert durch Erzeugung aufeinanderfolgender Phonemketten, bis eine gefunden wird, welche die lexikalischen und syntaktischen Beschränkungen erfüllt. Die lexikale Einschränkung erfordert, daß die Wortkette nur aus Einträgen in einem Phonemicon (phonemisches Wörterbuch) besteht. Die syntaktische Einschränkung erfordert, daß die Wortkette einer vereinfachten englischen Syntax genügt. Das Phonemicon war so aufgebaut, daß es etwa tausend Einträge aufwies. Es wurden zwanzig Sprechweisen aus im Mittel 3,4 Worten verwendet, um das System zu beurteilen. Ohne Anwendung der syntaktischen Einschränkung wurden davon 35 Prozent richtig erkannt. Durch Hinzunahme der syntaktischen Einschränkung wurde die Erkennungsrate auf 65 Prozent erhöht.

L'identification de la parole continue à l'ordinateur

Sommaire

On a programmé sur un petit ordinateur une méthode pour l'identification automatique de la parole. Le système admet à l'entrée des unités syntaxiques de discours continu soigneusement prononcé par un seul locuteur masculin collaborant à l'expérience. Les paramètres d'identification sont les coefficients d'ordre peu élevé du spectre, le taux de passages par zéro, les amplitudes de pointe du spectre et l'endroit apparent d'articulation; pour effectuer la segmentation, on a recours à une méthode de «segmentation par identification». A chaque segment correspond un choix possible entre deux phonèmes. On identifie l'émission en produisant des suites successives de phonèmes, jusqu'à ce qu'on en trouve une qui satisfasse aux conditions lexicales et syntaxiques. La condition lexicale est que la suite ne contienne que des phonèmes figurant au dictionnaire de phonèmes ou «phonémicon»; la condition syntaxique est que la suite de mots satisfasse à une syntaxe simplifiée de l'anglais. On a établi le phonémicon de manière qu'il contienne mille entrées. Pour apprécier le système, on a utilisé vingt phrases contenant en moyenne 3,4 mots: 35% d'entre elles furent reconnues correctement sans recourir à la condition syntaxique; à l'aide de celle-ci, le résultat était amélioré et atteignait 65%.

1. Introduction

This study represents a preliminary attempt to simulate a mechanical speech recognizer on a digital computer. The primary aim of the study was to design a system which would accept the carefully spoken speech of a single speaker and respond by correctly typing the utterance. The secondary aim was to examine the value of lexical and syntactic constraints in automatic speech recognition.

* Present address: The Boeing Company, Seattle, Washington.

Some (James [4], Pierce [7]) have argued that general speech recognition is feasible only if non-acoustic constraints are utilized. A possible classification of these constraints would be: (1) lexical, (2) syntactic, (3) semantic, and (4) statistical.

Once the language (or subset of a language) has been chosen, an automatic upper bound on the size and content of the vocabulary is implied. A lexical constraint will require that all identified words belong to this limited although perhaps large set. The possibility of recognizing nonsense words is naturally eliminated.

Only relatively few randomly generated word strings can be considered as structurally correct. This constraint is due to the grammar of the language. A general recognition system must require that any input word string satisfy the syntactic constraints on the language.

The semantic constraint requires not only that the word string make sense, but it must also impose a continuity of meaning. In a special purpose system, for example, a system designed to play chess from verbal instructions, the semantic constraint would require that the instructions correspond only to realizable moves (Reddy et al. [13]).

When one phoneme has been identified, the choice of the phoneme to follow is no longer arbitrary, but follows the statistics of phoneme sequence probabilities for the language (Fry and Denes [2]).

This machine recognition study was limited to the application of lexical and syntactic constraints and no attempt was made to apply semantic or statistical constraints.

2. A basis for speech recognition

One of the problems in designing a speech recognition system is in deciding on the fundamental unit to process (Hill [3]). If sentence identification is attempted, a comprehensive dictionary would need to contain about 10^{20} entries. If the word is chosen, 20000 entries might be adequate. It is important to note however, that in principle all of these words and sentences may be constructed from about forty perceptually identifiable speech sounds called phonemes. This economy of representation is one of the main reasons for the attractiveness of a speech recognition system based on identification of phonemes.

The above illustration was based on figures appropriate to a complete language; however, for a word recognizer of limited vocabulary, it may be more convenient to recognize the word directly. Typically, word recognizers have not relied upon phoneme recognition, but have generally used a matching of time-frequency patterns (or some other function in place of the frequency) for each word as a whole. This method has been used with some success (Denes and Mathews [1]), but for the reason mentioned above is not feasible for continuous speech or speech with a large vocabulary.

A phoneme recognition system faces the problem of segmenting the speech into phonemes. Some success in one method in which segmentation precedes identification has been demonstrated by Reddy [11], [12]. His method used the variation or stability of sound intensity levels to perform a primary segmentation with zero crossing rates as an aid in resolving ambiguities. Each segment was then classified as being either sustained or transitional. In a second method a sequence of phoneme estimates was made (e.g., Purves and Strong [10]) and the segment boundaries were set when the transition took place from one phoneme to another. This method is termed "segmentation by recognition".

Whatever the basic unit chosen, speech recognition eventually reduces to a problem in pattern matching. One must decide on a parameter set which adequately contains the required information of the speech signal. This basis set should be chosen so that perceptually similar speech sounds are represented by similar values of the parameters, and perceptually different speech sounds are represented by different values. From the point of view of practicality the parameter set should be of the smallest possible size.

It seems that a reasonable test on a parameter set would be the possibility of synthesis of intelligible speech from it. This criterion is neither necessary nor sufficient, although if the recognition parameters may be successfully used as control parameters for a speech synthesizer we are assured that all of the necessary information is included. The information may not, however, be in a useful form. For example, speech may be synthesized from a specification of the waveform, but this is not a useful representation for direct recognition. A parameter set should probably not be sensitive to phase.

Having chosen a parameter set, there remains the problem of what constitutes a "best fit". One much used method is to consider each entry of the set as a rectangular co-ordinate in a Euclidean space. In this case, each pattern corresponds to a point in the space and the "distance" between points can be used as a measure of closeness. This concept is useful because, by means of it, an utterance can be visualized as the trajectory of a particle in the space. A more complete view of closeness would use a metric for the space by appropriately weighting the most significant dimensions.

Another method of determing closeness uses a binary decision tree (e.g., Wiren and Stubbs [15]). This method has been applied in a vowel recognizer. The distinctive feature view of vowels was used and a decision made by sequential sub-classification. There is the danger in a system of this kind that one wrong decision will make the final decision incorrect.

A large number of feasible representations are available for the speech signal. A choice between them should be based on the criteria of efficiency of computation and efficiency and completeness of the representation.

The individual sounds of spoken English may be categorized in several different ways, but the most convenient primary classification for our purposes is a separation according to the parameters needed for identification. When considered in this way three main classes may be established: voiced continuants, fricative continuants and stop consonants. If separation into these broad subsets can be made without error then the intra-class identification becomes more efficient because of the smaller number of class members.

In this system, the voiced continuants were represented by the low order cepstrum coefficients. The fricatives were represented by slope change rate, zero crossing rate and cepstrum peak height. The stop consonants were represented by the apparent place of articulation, rise time and the presence or absence of frication immediately following the rise above silence. The choice of these parameter sets was largely a matter of convenience. In particular, use of the cepstral coefficients avoided the problems associated with formant tracking.

3. The recognition strategy

The utterances to be processed were first recorded in an anechoic chamber. The recording was then played through an amplifier, low pass filtered at 4.5 kHz, digitized at 10 kHz, and the digitized waveform written on the computer disc for further processing.

3.1. Acoustic analysis and parameter extraction

The object of the acoustic analysis was to extract a convenient parameter set to adequately represent the phonetic features of the acoustic signal.

The first part of the analysis program extracted information directly from the speech waveform: the slope change rate (SCR) defined as the number of slope changes per 100 samples of the digitized signal, the zero crossing rate (ZCR) defined as the number of zero crossings per 100 points, and the short time intensity level (STI) defined as the sum of the absolute values of 100 consecutive sample points.

A preliminary excitation decision was based on slope change rate and zero crossing rate.

In an attempt to reconcile the apparently conflicting criteria of good time and good frequency resolution, two different time windows were used in the program: for normal processing a 40 ms-window was used, but whenever the short time intensity rose above an arbitrary silence threshold a 10 ms-window was used. The purpose of the shorter window was to achieve better time resolution in the vicinity of stop consonants.

The cepstrum (Noll [5], [6]) was calculated in the usual way using a 512 point fast Fourier transform. The Hamming windowed speech was centred in the 512 point array and the remaining points set to zero. Using the thirty lowest frequency coefficients of the cepstrum, a smoothed logarithmic spectrum was calculated by means of an inverse DFT. The apparent place of articulation was then calculated by using a method outlined by Purves [8].

Almost all of the computation was performed in fixed point arithmetic and consequently a scaling factor was used throughout.

3.2. The phoneme set

The human vocal mechanism is capable of producing a very large number of distinguishable speech sounds. If we label each of these sounds and then describe an utterance in terms of them, the amount of information transmitted is overwhelming. A convenient concept in describing speech sounds is the phoneme. A phoneme is a distinguishable sound of speech and is the basic unit of which larger speech units are constructed. Spoken English recognizes about forty such phonemes.

The use of the word "phoneme" in the present context requires some liberalization. In the recognition system each voiced continuant phoneme was represented by a parameter set which, in turn, was considered as a vector in a multidimensional space. Thus, each speech sound was assigned a point in the space. The utterance (except for the complication of stop consonants and fricatives) corresponded to the trajectory of the "speech particle" through the space.

A phonetic description of the utterance could be considered as a gross representation of this trajectory in terms of a sequence of general regions through which the particle passed. When the problem is considered in these terms, there is no longer any need to express the phonemic spelling by means of a string of phonemes defined in the conventional way.

But rather, we may choose a "cloud" of points, which provides a satisfactory spread over the probable range of speech sounds, and describe the utterance as a sequence of nearest neighbors. Such a set of points could correspond to average values for the conventional phonemes, but this restriction is not necessary. This "cloud" may have a larger or smaller number of entries than the conventional set. In this system a set of 22 phonemes was used.

No attempt was made to separately accommodate diphthongs; in the present system these appeared as a sequence of the sounds through which the particle passed. The breakdown of phonemes by class is summarized in Table I. Internally in the system the phonemes were referred to by means of a phoneme index. The numerical sequence has not been completely followed because of a reduction in the number of phonemes from that used in the original version of the recognizer.

Table I.
Summary of the system phonemes.

Class	Number of members	Phoneme indices
Vowel	7	1–7
Nasal	3	8–10
Liquid	2	11, 12
Nasal-fricative	1	13
Fricative	4	14, 15, 18, 19
Stop	5	21–24, 26

3.3. Identification of continuants

The continuants are characterized by having a sustained quality. They may be subclassified as voiced or unvoiced. The identification of continuants proceeds on the assumption that these two subclasses may be distinguished from each other virtually without error.

The fricatives generally have a high slope change rate (greater than 60). The only non-fricative to possesss this property is /i/. Confusion between this phoneme and the fricatives may usually be eliminated by an examination of the 0.3 ms-coefficient of the cepstrum. In the case of fricatives, this coefficient does not often exceed a measured threshold, and in the case of /i/, it rarely falls below the threshold. Once the continuant class has been decided, further identification continues in each group independently.

Cepstrum matching was chosen as the recognition condition for quasi-periodic speech waveforms. This method had previously been used by Strasbourger [14] for speech recognition by using the low order coefficients.

In this work coefficients 1 through 10 were utilized. This corresponds approximately to the frequency range 0.1 to 1.0 ms. If these coefficients are considered as co-ordinates of a 10-dimensional hyper-space, the notion of closeness acquires a physical meaning. If we further allow that some of the components are more significant in the recognition process than others, a metric tensor of some kind may be used to reflect this relative importance. A very simple view is that those components which show the greatest intra-phoneme variation are less useful than the more stable ones.

For each non-fricative parameter set, two best fit phonemes were chosen. The best fit criterion actually used was minimization of the Euclidean distance. Provision was made in the program to use a non-unit metric tensor, but machine size limitations prevented its implementation.

Although it is difficult to visualize a 10-dimensional hyper-space, a projection of this space onto a 2-dimensional surface is easily obtained. Fig. 1 shows a projection of the space onto the C4-C5 plane. The numbers indicate the location of the corresponding standard phonemes. Also shown on the same figure is the trajectory of the speech particle, during utterance of the word "man". Table II lists the values of the standard voiced continuants and the approximate phonetic equivalents. It should be understood that the table values have not been chosen to correspond to the phonetic values indicated, but rather, each phonetic symbol has been chosen to give an approximate phonetic meaning

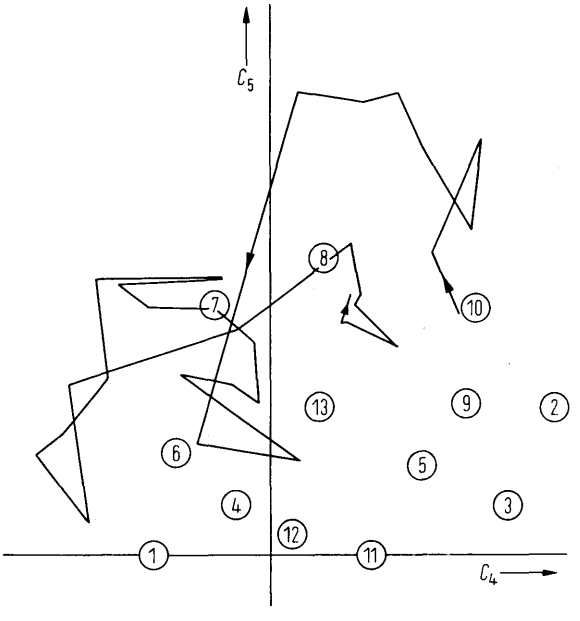

Fig. 1. Motion of the speech particle during the utterance of the word "man", shown as a projection onto the C_4-C_5 plane. The circled numbers show the projections of the voiced continuants onto the C_4-C_5 plane.

Table II.
Standard cepstral parameter phoneme values for continuants (phoneme number 13 has no single perceptual equivalent).

Suggest. equival.	Phoneme number	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
/a/	1	6800	1000	4000	−1200	0	− 800	−1100	−800	200	− 500
/i/	2	3000	−3000	3800	2800	1500	500	1000	−700	1500	500
/u/	3	9000	3000	3200	2400	500	1000	− 500	300	500	500
/Λ/	4	9000	−1000	3500	− 400	500	500	500	1000	−500	−1000
/I/	5	8000	2800	3800	1500	900	200	−1000	− 50	−900	− 100
/æ/	6	7500	−3000	3500	−1000	1000	−1000	500	1200	−900	0
/ɛ/	7	8000	−4000	3000	− 500	2500	200	−1500	500	−600	−1800
/m/	8	8000	− 500	2000	500	3000	1000	0	900	500	− 500
/n/	9	7500	−1500	2200	2000	1500	500	1500	−200	−700	− 400
/ŋ/	10	7000	250	1300	2000	2500	400	− 600	1000	−300	− 400
/l/	11	9500	4000	3500	1000	0	−1000	− 500	400	−500	− 500
/r/	12	10000	3000	2200	200	200	600	800	1000	800	0
/ʔ/	13	7000	1500	3000	500	1500	1000	2000	500	1000	0

to the particular parameter set. Phoneme 13 had no perceptual equivalent since several widely different sounds were sometimes represented by it. For example, the voiced fricative /ð/ is usually identified as phoneme 13; however, in some environments the nasals receive this identification. Because of this dual role, it has been classified as a nasal-fricative.

The parameters used for fricative identification were: the slope change rate (SCR), the slope change rate minus the zero crossing rate (SCR-ZCR), and the cepstrum peak height (CPH). The voiced fricatives differ from the unvoiced ones in that the superimposed periodicity lowers the ZCR and raises the CPH. The cepstrum of a pure fricative has no genuine peak and in this case the CPH refers to the maximum value encountered in the region of search. Table III summarizes the standard parameter values for the fricatives.

Table III.
Standard parameter values for the fricatives.

Suggested equivalent	Phoneme index	SCR	SCR − ZCR	CPH
/z/	14	75	35	300
/ʒ/	15	70	45	700
/s/	18	78	6	250
/ʃ/	19	65	13	700

A non-unit metric tensor was used in calculating distances in the fricative space. If we let u and v be two points in the fricative space, then the distance between them is defined by

$$D^2 = g_{ij}(u^i - v^i)(u^j - v^j)$$

where $g_{ij} = 0$ for $i \neq j$, and $g_{11} = 1$, $g_{22} = 1$, $g_{33} = 1/15$. Summation is implied over repeated indices. Presumably a detailed study would provide a more satisfactory g.

3.4. *Identification of stop consonants*

The stop consonants may be conveniently classified in terms of the place of articulation and the manner of excitation. An additonal identification cue may be obtained from the characteristic rise-time.

Each time the zeroth cepstral coefficient C_0 rose above the silence threshold the corresponding segment was labelled as a stop consonant. Frequently such a condition did not correspond to a stop at all, but merely the onset of speech. Consequently, the possibility of a "non-stop" was allowed in the identification routine. The minimization of the Euclidean distance was used as the best fit criterion. The following were the recognition parameters for the stop consonants:

1. the number of time frames in the following twelve judged as fricatives,
2. apparent place of articulation (on a scale of 1 to 11),
3. the number of time frames from the rise above the silence threshold to the first maximum of the zeroth cepstral coefficient C_0.

The five stops used in the system may be characterized as: (1) voiced front, (2) voiced back, (3) unvoiced front, (4) unvoiced back, (5) non-stop. Table IV lists the standard parameter values for the stop consonants.

Table IV.
Standard parameter values for the stop consonants.

Suggested equivalent	Phoneme index	Parameter 1	Parameter 2	Parameter 3
/b, d/	21	0	2	0
/d, g/	22	0	5	3
/p, t/	23	5	2	0
/t, k/	24	5	5	3
non-stop	26	12	5	9
non-stop	26	3	10	0

3.5. Segmentation

The segmentation operation had as its input a set of parameters for each time frame; these were:

1. Best fit standard phoneme index,
2. second best fit standard phoneme index,
3. distance from the best fit standard phoneme to the current time frame.

If the frame had been judged as silence its best fit phoneme number was zero.

Each of the time frames was given a numerical label according to the rules of Table V. This labelling system placed a "−1" at the end of each segment and a "+1" at the beginning. Any segment containing a "+3" was identified by the stop identification routine. The frames labelled with "+2" were the only frames used in the identification of the continuant segments. The frames labelled with "−2" were the only frames considered for segment boundaries in the subsegmentation routine. All frames which satisfied none of the labelling conditions were assigned a zero.

Table V.
Segmentation labels and conditions.

Condition	Label
Nasal preceding non-nasal	−1
Nasal following non-nasal	+1
Non-nasal following nasal	+1
Non-nasal preceding nasal	−1
Fricative preceding non-fricative	−1
Fricative following non-fricative	+1
Non-fricative preceding fricative	−1
Non-fricative following fricative	+1
4 frames before fall into silence	+1
5 frames before fall into silence	−1
1 frame after fall into silence	−1
2 frames after fall into silence	+1
Silence preceding non-silence	−1
1 frame later	+1
3 frames earlier	−1
2 frames earlier	+1
1 frame earlier	+3
Distance from best fit phoneme is smaller than that of previous frame	+2
Distance from best fit phoneme is a maximum	−2

Application of these rules provided a gross first attempt at segmentation. Improvement was made by elimination of intersecting segments, concatenation of short segments and subdivision of long ones. An example of the preliminary segmentation results are given in Table VI, with the approximate final segment boundaries shown by the horizontal lines.

The identification of each continuant segment required an averaging of the separate identifications of the individual frames. In this averaging process, only those frames were used which had the label "+2".

Table VI.
Illustration of the segmentation process. The utterance segment shown ("is in") is from "The shoe is in the room". The final segmentation boundaries are shown by the horizontal lines. The entries in the table are spaced at 10 ms intervals.

First choice phoneme index	Second choice phoneme index	First choice distance function	Segmentation label	Phoneme recognized
5	9	3295	1	
5	2	4111	0	
5	2	4584	−2	
5	2	4510	2	/I/
2	5	4119	2	
2	5	4469	0	
5	2	4845	0	
2	9	4551	−1	
9	8	5041	1	
9	8	4728	0	
8	9	4378	2	
10	13	4758	0	
13	10	5474	0	
15	14	4980	0	/z/
2	13	6708	0	
2	13	6759	−1	
15	14	4242	1	
15	14	7406	−2	
15	14	4889	2	
15	14	5176	0	
15	14	4516	−1	
2	9	5033	1	
2	9	4264	2	
9	2	4266	−1	
2	9	4595	1	
2	9	4280	2	/I/
2	5	4260	2	
2	5	4126	0	
5	2	3811	2	
5	9	3206	−1	
9	5	2960	1	
9	5	2992	0	
9	8	3665	−2	
9	10	2813	2	
9	10	2706	2	
9	5	2733	0	
9	10	1294	2	
9	5	1710	0	
9	5	2244	0	/n/
9	5	1971	2	
9	5	2602	0	
9	5	2049	2	
9	13	2472	0	
9	4	3690	0	
13	8	3877	0	
4	8	4691	0	
4	13	5420	−1	

3.6. The phonemicon

For the purposes of this study, a dictionary which contains words according to their phonemic rather than their orthographic spellings is called a "phonemicon". The phoneme recognition process converted the input utterance into a string of phonemes. The system then attempted to interpret the phoneme string as a word string by matching phoneme sequences with the entries of the phonemicon.

Often, a single word appeared as several slightly different phoneme strings on different occasions. As a result of this, it was useful to include these several versions of the word in the phonemicon. Each entry of the phonemicon contained the orthographic representation and some grammatical information. Homonyms were listed separately for each distinct meaning. The phonemic spellings were entered onto the system as they were estimated or computed from the calibration utterances.

3.7. Utterance extraction

Human recognition of speech is not based solely upon acoustic information; a listener must generally use both his knowledge of the language and of the topic of speech as aids in understanding. Sometimes allowance must even be made for the idiosyncracies of the speaker. A mechanical speech recognizer should also be able to gain in proficiency by utilizing extra-acoustic information.

Application of a general semantic constraint would require a considerable amount of stored information and complicated logic. A very severe semantic constraint, corresponding to a task-oriented recognition system, is much more realistic. A general syntactic analysis would be prohibitive, but it is possible to construct a syntax checking algorithm for a fairly general subset of English from a relatively small set of syntax rules.

The recognition system described in this study incorporated a specially developed syntax checking algorithm (Purves [8]). Although it would be very useful to have a syntax checking algorithm which operates infallibly with normal English, it is unlikely that such an algorithm will be constructed with only a small number of simple rules. A short list of grammatical rules always seems to have a long list of exceptions. The grammatical analysis scheme used in this system is not universally applicable to normal English. It rejects some grammatical utterances and accepts some ungrammatical ones. It does, however, serve the useful purpose of severely limiting the number of allowed word strings, thus substituting for a more general testing algorithm. Its most significant asset is that it is easily implemented as a FORTRAN subroutine containing about 300 statements (Purves [8], [9]).

The utterance extraction program attempted to convert the input phoneme string into a syntactically acceptable word string. If this could not be achieved with the original phonemes, the string was modified by replacement of phonemes from the reserve choices. Successive phonemes strings were generated from the original by substitution of reserve choice phonemes until a string was found which satisfied the lexical and syntactic constraints. The trial phoneme strings were generated by first replacing one phoneme at a time by its reserve choice, then two at a time, and so on. For example, if the original phoneme string had been (24, 4, 6, 9) and the backup set (26, 5, 3, 8), then the sequence of phoneme strings to be tested would have been (24, 4, 6, 9), (26, 4, 6, 9), (24, 5, 6, 9), (24, 4, 3, 9), (24, 4, 6, 8), (26, 5, 6, 9), (26, 4, 3, 9), (26, 4, 6, 8), and so on.

3.8. Practical considerations

The recognition system was implemented on a PDP-15 computer. The computer had a word length of 18 bits, but lacked a hardware floating-point arithmetic capability. The small core memory size (16K) gave rise to some special problems, making it necessary to write the recognizer as a sequence of programs. In some sense this was a natural arrangement, since some operations were prerequisites for others. In the case of the acoustic analysis and utterance extraction programs the machine size was scarcely adequate. These programs were so large that some CRT display handling programs could not be properly loaded simultaneously. Another problem in machine size arose in matching the test phoneme string with the phonemicon entries. Only a small part of the dictionary could be brought into core memory at a time (64 entries), making necessary a very large amount of disc manipulation.

The recognition system used three dictionaries: the spelling list, the grammatical features list and the phonemicon. The dictionaries were stored on disc and sections brought into core as needed. The sequence of entries provided a means of referring to entries. A mapping table used in conjunction with the phonemicon contained the location of the corresponding entries of the spelling list and the features list. Since a single word was allowed to have several entries in the phonemicon, it was necessary to allow all of these entries to point to the same spelling and the same features set.

In the phonemicon, three phonemes were stored in one computer word allowing 6 bits per phoneme.

The maximum allowed number of phonemes in a single word was 12, so that 4 computer words were required to store each phonemic spelling.

4. System evaluation

It is difficult to evaluate a speech recognition system and assign to it a single number indicating its proficiency. The fraction of correct identifications from a test set of utterances is certainly useful information, but hardly sufficient. Factors which would seriously influence such a single number rating include vocabulary size, the amount of effort expended in calibrating the system (its experience) and the number of word variants permitted. Constraints which limit the possible utterances also play a significant role. In this system a syntactic constraint was able to accept or reject postulated word strings. A contextual constraint could serve the purpose of further reducing the possible utterances. If the recognition system is explicitly task oriented with a reasonable number of different tasks, and no one will be injured by a wrong decision, the machine could be allowed to guess from the set of possible responses. A more responsible system would be unable to respond unless the decision were made with a high degree of confidence. It is apparent that an irresponsible system would have a better result rating than a responsible system.

The recognition system of this work had the choice of a very large possible number of utterances. The phonemicon contained phonemic spellings for about 800 different words. The syntax severely limited the number of acceptable utterances; nevertheless, this number was very large. For example, with an 800 word vocabulary and a string of three words, there would be of the order of 10^8 strings possible without the constraint and an estimated 10^7 strings allowed with the constraint.

4.1. Method of calibration

In the early stages of the system, words and word strings were processed and the computed phonemic spellings entered onto the phonemicon. In this way an early corpus of data was obtained. During this early stage the recognition logic and standard parameters were continually being changed so that a few of the phonemic spellings were rendered obsolete. When the system had reached a sufficiently advanced stage for evaluation, a large number (about 700) of estimated phonemic spellings were entered onto the phonemicon. These estimations were based on experience of how the system had behaved. The next stage was to calibrate the words to be used in the evaluation. This was accomplished by processing each word several times to compute the phonemic spellings.

Most words show some variation from one utterance to the next, so it was necessary to also estimate the probable variants from a consideration of the reserve choice phonemes. It may be argued that the inclusion of a large number of variants in the phonemicon makes the search time excessive. However, it is much faster to search in the phonemicon for a common variant than it is to generate that variant from the original phoneme string. In any case, a phonemo-numeric sequencing of phonemicon entries would have substantially reduced the search time.

The large number of estimated phonemic spellings were intended merely to act as decoys. The time required to calibrate such a large number of words would have been prohibitive. In attempting to match a phoneme string with phonemicon entries the program had no way of knowing which entries were genuinely calibrated and which merely guessed. In this way it was possible to estimate the effects of having a large vocabulary without paying the full cost. Since the estimated phonemic spellings occurred earlier in the phonemicon than the calibrated ones, their ability to act as decoys was increased.

4.2. A test on the system

A set of twenty short utterances was chosen to make an estimate of the proficiency of the recognition system. These utterances, together with a large part of the calibration materials were recorded at the same session. No special effort was made to pronounce the test utterances in exactly the same way as the calibration materials. The recordings were made using carefully spoken continuous speech from a single speaker. The words of the test strings were allowed to run together when it seemed natural to do so. Short pauses between words were permitted when the last phoneme of a word was the same as the first phoneme of the following word or when co-articulation effects seemed likely to seriously affect phoneme integrity. The system was invited to recognize the test set.

Of the twenty utterances used in the test, in fifteen cases recognition judgments were made. Of these, two were slightly incorrect. Table VII shows the phonemic spellings generated for the fourth test utterance, "man and machine". A general summary of the results obtained is given in Table VIII.

Of the fifteen utterances recognized, nine were recognized without the syntactic constraint, while six required application of the syntactic constraint.

Table VII.
Phoneme index string generated for the utterance "man and machine". The starred second choice phonemes were required for recognition. When two adjacent phoneme indices were the same they were concatenated. The first vowel sound of "machine" typically did not appear.

First choice index	Second choice index	Word
22	26*	man
8	10	
8	5	
6	4	
9	13	
9	13	
27	27	(pause)
24	26*	and (an')
6	7	
9	13	
26	22	machine
13	8	
19	18	
2	6	
13	9*	

Two of the five unrecognized utterances would probably have been recognized by letting the system run longer. Two utterances were not recognized because of inserted phonemes and one because of a missing phoneme.

5. Discussion and conclusions

This work represents an attempt to simulate an automatic speech recognizer for syntactic units of continuous speech. The following restrictions were made on the problem in order to keep it within acceptable size limits:

1. only one male talker was used,
2. the speech was carefully spoken,
3. the phonemicon contained a large number of dummy entries to act as decoys,
4. the calibration materials and test utterances were recorded in a quiet environment on relatively few occasions,
5. only about 50 of the 800 words on the system were used in the test utterances,
6. only lexical and syntactic constraints were used.

Although these simplifying qualifications were made, there is nothing fundamental in the system which prevents its extension to multiple speaker use and/or a larger vocabulary.

The choice of recognition parameters for this system was largely a matter of convenience. The principal advantage of using the cepstral coefficients was the avoidance of formant tracking. The cepstrum is a strictly formal representation and not subject to some of the vaguaries of a formant pick-

Table VIII.
Results of the recognition test.

	Input utterance	No. of phonemes in string	No. of reserves used	Comment	Syntactic constraint applied
1.	Sea smell	8	1	Recognized	Yes
2.	Peace and rest	10	3	Recognized	No
3.	She sees me	11	1	Recognized	Yes
4.	Man and machine	15	3	Recognized	Yes
5.	The shoe is in the room	19	0	Recognized	No
6.	I shall make some room	20	—	Recognition prevented by inserted phoneme	—
7.	Nine machines	11	0	Recognized	No
8.	A smooth shoe	10	2	Identified as "smooth shoe"	No
9.	I see the sea	10	1	Recognized	Yes
10.	The ill man is at rest	20	2	Recognized	No
11.	A small mass	11	3	Identified as "the small mass"	No
12.	On some animal	12	1	Recognized	No
13.	Send a mass to me	17	4	Unreas. amount of time needed	Yes
14.	Some men send peace	20	—	Recognition prevented by inserted phoneme	—
15.	Free money	8	1	Recognized	Yes
16.	Some loose money	11	1	Recognized	No
17.	Free machines	9	1	Recognized	Yes
18.	Does the short person see the smooth ice	29	3	Unreas. amount of time needed	Yes
19.	Still	5	1	Recognized	No
20.	Roads	5	—	Recognition prevented by omitted stop consonant	—

ing algorithm. However, some of the recent formant estimating algorithms may tip the balance in favor of formants as parameters in future systems.

Perhaps the most intuitive description of speech is the articulatory one. In this system recourse was made to the articulatory domain for identification of the stop consonants. The method allows only a single constriction in a uniform tube. If the place of articulation so calculated is used only in the vicinity of stop consonants, the constriction will dominate over any area expansion in the tract.

A question arises about whether a second generation version of this system should have a greater or smaller number of phonemes. The answer is not immediately obvious. If too few phonemes are used then different words may appear with the same phonemic spelling. If too many are used then the same word may have a large number of phonemic variants. The latter seems to be the preferable condition. In the second case the phonemicon size is large, but the information is available for correct identification. In the first case, if the confused words have the same grammatical features, no distinction between them is possible. The decision on the number of phonemes to use must be based on the particular vocabulary of the system and the confusions likely to arise.

A problem closely related to the number of system phonemes is whether the recognition procedure should be more or less sensitive. The acoustic analysis is capable of detecting very subtle and even imperceptible speech events. It is not always obvious which events should be taken seriously and which should be ignored. Some ignorable events last longer than significant ones. It leaves things in the quandary then, that sometimes more sensitivity is needed and sometimes less.

A logical extension to the present system would be a self calibrating facility. In its easiest form this would be applied directly to the phonemicon. Whenever a word is correctly recognized by the system, and reserve choice phonemes are necessary, the first choice phonemic spelling would be added to the phonemicon. Another self adjusting method would allow the standard phoneme parameters to be adjusted by a correct identification. In this way, the standard parameters might converge to an optimum set.

One of the more serious problems in speech recognition is that of omission and insertion of phonemes. The use of reserve choice phonemes goes part of the way toward solving this problem. For example, consider the word "meany." Suppose that the phonemicon entry is /m/i/n/i/. A possible phonemic spelling of the spoken word would be /m/i/i/i/, with reserve set /?/I/n/I/. Because of the repeated phoneme the original phonemic spelling will reduce to /m/i/. However, by substitution of a reserve choice phoneme the string /m/i/n/i/ may be generated. We have, in effect, been able to insert a phoneme into the original string. In a similar way, by considering the process in reverse, a spurious phoneme may be removed.

Word occurrences with missing phonemes may be considered as genuine variants if the system generates that phonemic spelling often. For example, in the case of the word "machine", /m/ʃ/i/n/ could be considered a genuine variant.

Insertion of phonemes poses a more serious problem, particularly if it takes place between words. As a partial solution the phoneme string variation could sequentially delete phonemes with large distance functions.

The sequence of variations used in the recognition system was obtained by starting the substitutions at the beginning of the utterance. The correct utterance could be found more rapidly, and perhaps more accurately, by beginning the substitution with those phonemes whose first choice distance functions are about the same as the corresponding reserve choice distance functions.

This system has no ability to recognize long sections of continuous speech. If the utterance to be recognized contains a large number of phonemes, the time taken to make even a few substitutions becomes prohibitive. Implementation of this system to long sections of continuous speech would require the use of prosodic features in some syntactic preprocessing. Presumably one could, in this way, segment the utterance into relatively short word groups whose syntactic function could be determined by pitch level and explicit recognition.

A further way of using prosodic features would be in the phonemicon search. Six bits were allowed for each phoneme in the phonemic spellings, but only five of these were actually used. An on-off stress marker could use this sixth bit. The extra information could be helpful in eliminating some confusions, especially with a large phonemicon. For example, the stress and pitch patterns make possible a distinction between "light-house keeper" and "light housekeeper".

The purpose of the syntactic constraint is to limit the choice of possible utterances, and so reduce the likelihood of an incorrect decision. The more severe the constraint the more liberties may be taken in computing phoneme string variations and in making deletions. The present syntax checking algorithm is unsatisfactory in the sense that some grammatically correct utterances are rejected by it.

The computer configuration used in this work was inadequate for the task. An additional 8K of core memory would have made a considerable reduction in the amount of data manipulation. Real time Fourier transform hardware would have reduced the real time ratio of approximately 300 to one very substantially. There is no doubt that parallel processing could be used to great advantage in further reducing the real-time ratio.

A system has been developed which attempts to recognize syntactic units of continuous speech when carefully spoken by a single male talker. This system, although not satisfactory for practical application, represents progress in the field of automatic speech recognition. The lexical constraint alone was adequate to allow correct recognition of 35 percent of the set of 20 test utterances. Imposition of the syntactic constraint raised the recognition rate to 65 percent. These results demonstrate that, with the phoneme recognition method used, non-acoustic information is essential to the design of a speech recognizer.

(Received June 12th, 1975.)

References

[1] Denes, P. and Mathews, M. V., Spoken digit recognition using time-frequency pattern matching. J. Acoust. Soc. Amer. **32** [1960], 914 (A).

[2] Fry, D. B. and Denes, P. L., The solution of some fundamental problems in mechanical speech recognition. Language and Speech **1** [1958], 35.

[3] Hill, D. R., Automatic speech regognition: a problem for machine intelligence. Machine Intelligence **1** [1967], 199.

[4] James, W., Talks to teachers on psychology and to students on some of life's ideals. Holt, New York 1899, p. 159.

[5] Noll, A. M., Short time spectrum and 'Cepstrum' techniques for vocal pitch detection. J. Acoust. Soc. Amer. **36** [1964], 296.

[6] Noll, A. M., Cepstrum pitch determination. J. Acoust. Soc. Amer. **41** [1967], 293.

[7] Pierce, J. R., Whither speech recognition? J. Acoust. Soc. Amer. **46** [1969], 1049.

[8] Purves, R. B., An automatic method for computer recognition of continuous speech. Ph. D. dissertation, Brigham Young University 1973.

[9] Purves, R. B., Toward machine recognition of spoken language. Proceedings of the Brigham Young University Language Research Center Linguistics Symposium, 1973.

[10] Purves, R. B. and Strong, W. J., An automatic method for computer recognition of continuous speech. J. Acoust. Soc. Amer. **53** [1973], 355 (A).

[11] Reddy, D. R., Segmentation of speech sounds. J. Acoust. Soc. Amer. **40** [1966], 307.

[12] Reddy, D. R., Phoneme grouping for speech recognition. J. Acoust. Soc. Amer. **41** [1967], 1295.

[13] Reddy, D. R., Erman, L. D., and Neely, R. B., A model and a system for machine recognition of speech. IEEE Trans. Audio Electroacoust. AU-**21** [1973], 229.

[14] Strasbourger, E., The role of the cepstrum in speech recognition. 1972 Conference on Speech Communication and Processing. Conference Report. Newton, Mass., paper H7.

[15] Wiren, J. and Stubbs, H. L., Electronic binary selection system for phoneme classification. J. Acoust. Soc. Amer. **28** [1956], 1082.