STANFORD ARTIFICIAL INTELLIGENCE PROJECT October 7, 1964
Memo No. 26

EXPERIMENTS ON AUTOMATIC SPEECH RECOGNITION BY A DIGITAL COMPUTER

by R. Reddy

Abstract: Speech sounds have in the past been investigated with the aid of spectrographs, vo-coders and other analog devices. With the availability of digital computers with improved i-o devices such as Cathode Ray tubes and analog to digital converters it has recently become practicable to employ this powerful tool in the analysis of speech sounds.

Some papers have appeared in the recent literature[5] reporting the use of computers in the determination of the fundamental frequency and for vowel recognition.

This paper discusses the details and results of a preliminary investigation conducted at Stanford. It includes various aspects of speech sounds such as waveforms of vowels and constants; determination of a fundamental of the wave; Fourier (Spectral) analysis of the sound waves formant determination, simple vowel recognition algorithm and synthesis of sounds. All were obtained by the use of a digital computer.

The research reported here was supported in part by the Advanced Research Project Agency of the Office of the Secretary of Defense (SD-183)

ACKNOWLEDGEMENTS

This project was undertaken under the supervision of Professor John McCarthy. The author would like to express his gratitude to Professor McCarthy for his interest, encouragement and suggestions which were the main source of inspiration.

The author is also indebted to Mr. Dow Brian, Mr. Mark Finkelstein, and Mr. Harry Ratchford for the valuable discussions he had with them. An exploratory investigation undertaken together with Mr. Brian laid the foundation for the present investigation. Mr. Ratchford has been mainly responsible for the additions and modifications to the computer system.

The author would like to thank Mrs. Fran Thomson for her excellent typing and many other friends who volunteered to be experimental subjects.

EXPERIMENTS ON AUTOMATIC SPEECH RECOGNITION BY A DIGITAL COMPUTER

by R. Reddy

Introduction:

Speech sounds have been a subject of extensive research in the post-war period by various groups like communication engineers, phoneticists and linguists. Hearing, like other sensory mechanisms like seeing, smelling, tasting, and feeling, etc., can be looked upon as input to a highly complex pattern recognizer, the human mind. It is natural, therefore, for people working on artificial intelligence to take active interest in speech sounds. It is our hope that such an overall view could lead to more fruitful results.

A sound is any disturbance of the air that causes a displacement of the ear drum which, after transmission by the bone chain, affects the liquid in the inner ear in such a way that the auditory nerves are stimulated. It is known, by the experiments conducted by Drs. Wever and Bray on cats, that sounds are communicated to the brain in the form of electrical energy which is essentially the same as that of a microphone. Both can be amplified and played-back to produce the same sound. It appears that very little classifying of sounds is done before they reach the human mind. Thus the human mind is required to analyze the attributes of each sound and associate a concept with it. In so doing it not only makes use of acoustic information of speech but also linguistic information (contextual and grammatical). Any automatic speech recognition system will also have to take these aspects into consideration to be successful. In this paper we shall mainly concern ourselves with some aspects of the acoustic information of speech.

Until now spectrographs and vo-coders have been extensively utilized in the analysis of speech sounds to determine the energy of a speech sound at various frequency levels.[1,2,3] These energy patterns usually have sufficiently distinguishing characteristics which have formed the basis of formant theory.

Vocal cords, set to vibrating by the escaping air[4], are blown apart at regular intervals. The escaping air, at a higher pressure, acts as an impulse which sets the column of air from glottis to the lips and nostrils vibrating at its own natural frequencies known as the formant frequencies. This column of air will tend to resonate at the above formant frequencies irrespective of the fundamental frequency of the vocal cords (the so called pitch of the sound) provided the shape of the vocal organs remains unaltered. This forms the basis of formant theory. However, our investigations have shown (see Section 5), that introduction of the pitch of the voice as a parameter helps overcome the problem of overlapping formants in vowel recognition. The above problem may be due to the fact that certain muscles of the vocal organs become tense at a higher pitch thus altering the shape of the resonating chamber.

In between impulses from the vocal cords, the amplitude of the formant frequencies will tend to decrease due to the dampening effect of the resonating chamber. Thus one should, perhaps attempt to express the fundamental wave as a combination pure sin-cos waves of decreasing amplitudes rather than a combination of those of constant amplitudes.

In order to develop an Automatic Speech Recognition system at Stanford, preliminary investigations were started this summer to determine the various requirements of such a system. The equipment used consisted of a PDP-1 computer with 20 k core memory. A Cathode Ray tube (for visual display) and a microphone (connected to the computer through an A-D converter) were part of the input-output devices attached to the computer. Speech sounds were digitized to form numbers between 0 and 63, the no-noise level being 32. Signals were sampled at about 50 microsecond intervals.

Segmentation of sounds, fundamental determination, Fourier analysis, Formant determination and Vowel Recognition were all performed by the computer using the digital information obtained above. Pictures of various waves and analyses, displayed on the scope, were taken by using a polaroid camera. Some of them are included in this report. The details of the investigation are given in the following sections.

Determination of the Fundamental Frequency of a Speech Sample:

Fundamental frequency is the frequency at which a given sound is perceived. It is also called the pitch of the sound. It is well known that this frequency is equal to the greatest common divisor of all the frequencies present in the given sound. Thus given a fundamental wave we can, by means of Fourier analysis or otherwise, determine the various frequencies (and their amplitudes) which are combined together to make up the fundamental. Hence, it is desirable to be able to select a fundamental of a given sound.

Fundamental waves can be perceived visually when the speech sound is displayed on the scope. The same pattern can be seen to repeat itself. However, such a definition is vague and does not suggest a means of determining the fundamental. Using the notions of set theory we shall attempt to define precisely the concept of the fundamental wave of a speech sample.

Definition 1. Let P be a set of patterns and let Q be a proper subset of P. A characteristic of Q is defined as that which contributes information to distinguish Q from P-Q, the compliment Q relative to P.

Local maxima and minima and their relative positions in time domain, number of zero crossings and their relative positions, slope of the wave at the zero crossings, relative positions of the maxima and minima of the envelope of the wave are some of the characteristics which may contribute information for distinguishing a speech wave. Relative width and height, vertical lines, horizontal lines, inclined lines, curves with change of direction of 90°, etc. can be regarded as some of the characteristics which contribute information in distinguishing the characters of the alphabet.

Definition 2. Let $p \epsilon P$ and $q \epsilon P$. Let C be a given characteristic mapping function. Let P be the domain and E^n, n-dimensional Euclidean space, be the range of C. p is said to be equivalent to q under C if $C(p) = C(q)$.

C is a vector-valued function and the dimension of range space will be dependent upon the number of characteristic rules specified by C. Thus two fundamental waves shall be said to be equivalent if they have the same relative positions of maxima and minima or the same number of zero crossing etc. depending on the rules specified by C. It is important to note that p and q are equivalent under C does not imply that p and q are equivalent under another mapping C_1: eg. A, *A* may have the same number of branches at each node but they will not remain equivalent when orientation of the character is also added as a characteristic.

Lemma 1. Characteristic mapping function C partitions P into a collection of disjoint sub-sets each of which is mapped to a single point in E^n.

Proof. Let T be a subset of E^n such that $C(P) = T$. The lemma states that $\forall_t : t \epsilon T, \exists$ a unique inverse image.

$$Q = C^{-1}(t) = \{p \mid t \epsilon T, p \epsilon P, C(p) = t\}$$

The union of all the inverse images is P and their intersection is the null set. The fact that they are disjoint follows from definition 1.

From now on P shall be considered as a set of sound waves. Each element of P, a fundamental wave, can itself be considered as a function of time. Thus P, the collection of all such waves, can also be considered as a function of time. Using a computer P will be sampled at discrete time intervals, resulting in an ordered sequence defined on the time domain. This sequence when suitably grouped together will give us P, the set of sound waves we desire. Now we are in a position to define the fundamental.

Definition 3. Let P be a set of waves. Let C, the characteristic mapping function, partition P into disjoint subsets Q_1, Q_2, etc. Then any element of the largest subset Q_n shall be defined to be a fundamental of P. By largest subset we mean that subset which has the largest number of points in the time domain.

Thus if a given sound consists of 50 equivalent waves at 5000 cps, 10 waves at 500 cps and 3 waves at 100 cps then one of the waves with 100 cps will be selected as the fundamental because 3 waves at 100 cps will have the largest number of points in the time domain (50 waves at 5000 cps take 1/100th of a second; 10 at 500 cps take 2/100th of a second and 3 at 100 cps take 3/100th of a second). If Q_1, Q_2 ... are singletons i.e. if C is one-to-one on P then we shall say that there exists no fundamental in P. This would happen if each wave had different characteristics. The algorithm used for forming the set P and the characteristic mapping function C used in the investigation are given below.

Several sets of P, the possible fundamental waves are formed by the use of modified positive Zero-Crossing method. By a positive Zero-Crossing is meant a point at which the no-noise level line intersects with the given speech sample, the slope of the wave at the point of intersection being positive. Usually they intersect at a number of points. Even within a given fundamental they may intersect a number of times. However, between fundamentals the intersections tend to occur at approximately the same intervals and the same number of times. This fact is used in selecting candidates for fundamentals.

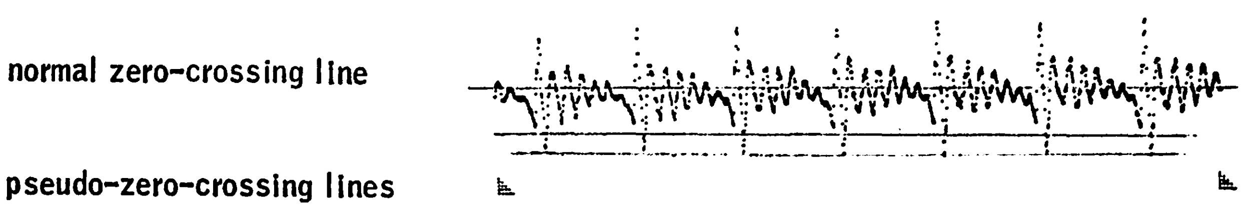


Fig. 2.1

For most of the speech sounds, the no-noise line intersects with the fundamental several times, slowing the process of recognition. Instead of the center, we can use a pseudo-no-noise line towards the lower (or higher) part of the wave. Then there will usually be fewer Zero-Crossings to deal with. Further the no-noise level line fails when presented with a sound sample representing low random noise (such as that which occurs in a computer room) because out of the numerous random-crossings some may have a semblance of regularity. To cater for the possibility of fundamental waves with gradually increasing or decreasing amplitude, several zero-crossing lines at different levels may be used (three were used in this case). If the first fails the second and then the third set of P's can be tested for likely fundamentals.

So much attention is necessary because sounds vary a great deal in pitch, amplitude and general pattern. Use of local minima and other approaches were tried but not with the same degree of success.

Having formed the successive differences (of location) between positive zero crossings these are grouped together in groups of 1,2, ... 10 each representing a likely fundamental wave. This gives us the set of waves P from which we attempt to choose a fundamental.

The characteristic mapping function C was chosen to be the following. Two members p and q of set P will belong to the same subset Q if

a) Lengths of the waves do not differ by more than 1% of their average length.

b) Each occurs adjacent to at least one other member of the subset Q in time domain.

c) The frequency of the wave shall be within normal range of human voice eg. 90 to 300 cycles per second.

The last has the advantage of increasing the probability of success since some irregular variations in a speech sample can cause the selection of two fundamentals as one. However, it has the disadvantage of not being useful for higher frequency sounds like whistling and also for some low frequency sounds.

sample fundamentals selected

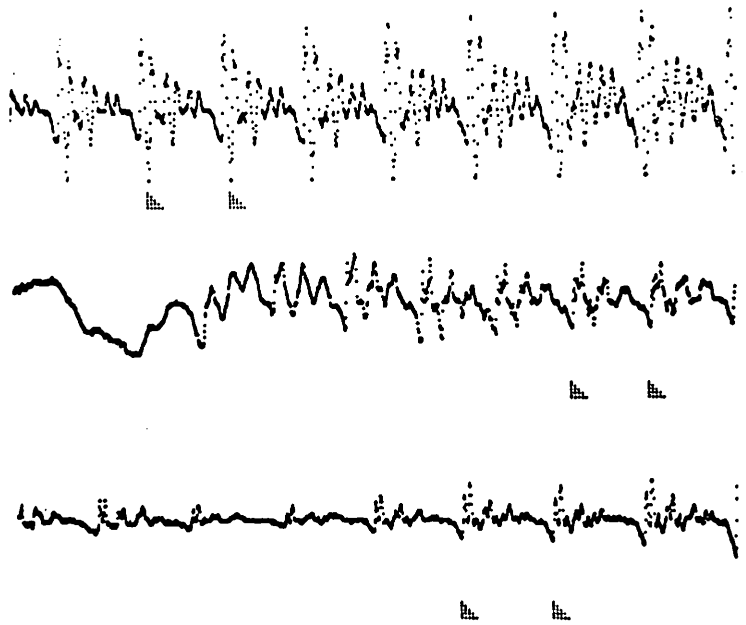


Fig. 2.2

Some of the fundamentals chosen by the use of the above algorithm are illustrated above. The wave between the markers represents the fundamental chosen in each case.

3. Fourier Analysis of a Fundamental

Periodic functions [6], such as a fundamental of a speech sample can be expressed in the form of a series

$$f(t) = a_o + a_1 \cos t + a_2 \cos 2t + \ldots + a_n \cos nt + \ldots.$$

$$+ b_1 \sin t + b_2 \sin 2t + \ldots + b_n \sin nt + \ldots.$$

where

$$a_n = \frac{1}{p} \int_d^{d+2p} f(t) \cos \frac{n\pi t}{p} \, dt$$

$$b_n = \frac{1}{p} \int_d^{d+2p} f(t) \sin \frac{n\pi t}{p} \, dt$$

where (d,d+2p) is the interval in which the function is to be expanded.

In a computer the continuous function f(t) is sampled and stored at discrete time intervals. Using this interval as 'dt' and assuming that f(t) is the ordinate in the middle of the interval, a_n and b_n can be expressed as follows.

$$a_n = \frac{dt}{p} \sum_{t=0}^{2p-1} f(t) \cos \frac{n\pi t}{p}$$

$$b_n = \frac{dt}{p} \sum_{t=0}^{2p-1} f(t) \sin \frac{n\pi t}{p}$$

In the analysis of speech we are mainly interested in the relative values a_n and b_n. These represent the amplitudes of sine and cosine waves with periodicity of 2p/n (the nth harmonic), all of which when added together result in the complex wave that is being analyzed. Since $\frac{dt}{p}$ is constant for all n, we may re-define a_n and b_n to be

$$a_n = \sum_{t=0}^{2p-1} f(t) \cos \frac{n\pi t}{p}$$

$$b_n = \sum_{t=0}^{2p-1} f(t) \sin \frac{n\pi t}{p}$$

The number of coefficients to be calculated is determined by the least amount of information that is required for further analyses. For most speech sounds energy distributions are required up-to at least 3000 cps. The pitch of the speech sound is normally over 90 cps. So if we determine a_n and b_n for n up to 32, the desired result would be achieved.

The power of the spectrum at the n^{th} harmonic is defined to be

$$c_n = a_n^2 + b_n^2$$

c_n represents the amplitude of the phase independent nth harmonic. Let c = max ($c_1 \ldots c_{32}$). Since only the relative power at various harmonics is required c_n shall be defined to be

$$c_n = \frac{a_n^2 + b_n^2}{c}$$

These c_n shall be called normalized power coefficients. Some sound waves and their normalized power coefficients are shown below.

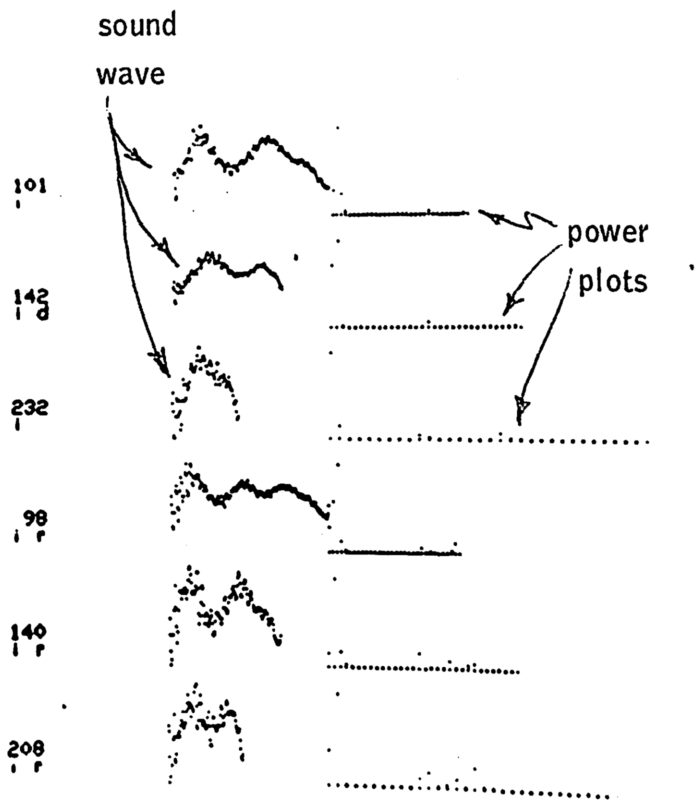


Fig. 3.1

Phoneme 'i' as spoken by two people is illustrated in Figure 3.1. Each person was asked to say 'i' at three different frequencies. Frequency of each utterance is given to the left along with identification of the phoneme and the originator. The fundamental selected is plotted as a function of time and the power as a function of absolute frequency. The increment between two adjacent points on the power plot is equal to the frequency of the wave.

4. Formant Analysis

The spectral distribution of a sound wave is an indication of the resonance characteristics of the air column above the vocal cords which depends on the shape of the vocal organs at the time of utterance. Every frequency at which there exists a local maximum shall be called a formant frequency. To estimate the locations of the local maxima, we consider the spectral distribution as a continuous function. Values of this function are known at discrete points. Estimates of the local maxima can be obtained by interpolation or simple curve fitting.

If we assume that the resonance characteristic curve of a simple resonator is normally distributed (as most authors appear to do) and that the vocal air column is composed of several simple resonators then we can regard the spectral distribution as a combination of several normally distributed curves, some of which might be overlapped. The mean of a normally distributed curve is also the point of the maximum ordinate. Thus a maximum likelihood estimator of the mean would give us a formant frequency.

The spectral distribution is divided into one or more normally distributed curves. Since we only have values at discrete points, a curve shall be defined to start at a local minima and end at the succeeding local minima. Having separated the curves, an estimate of the mean is calculated for each curve, giving the expected values of the Formant frequencies.

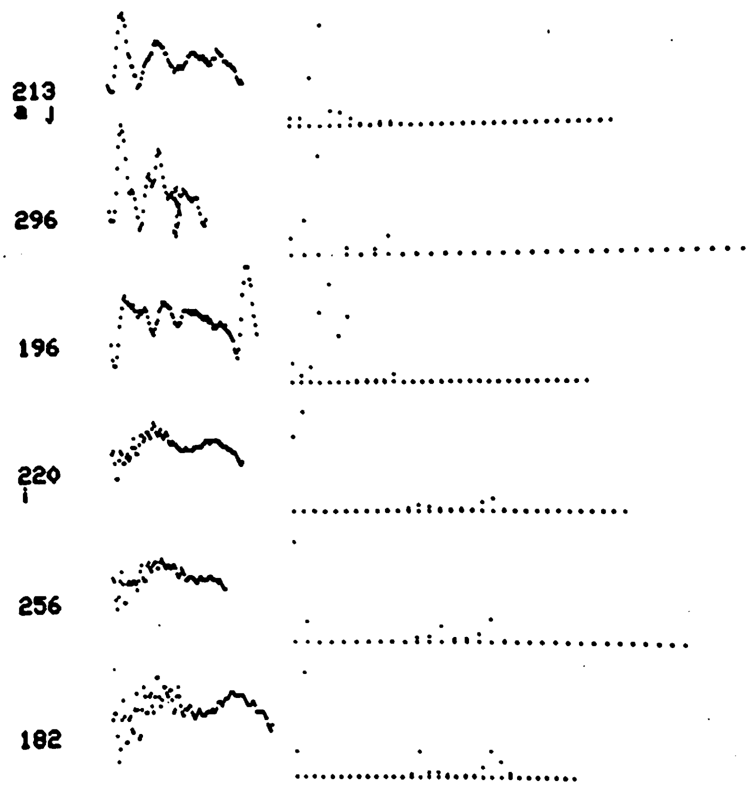


Fig. 4.1

The formant frequencies estimated in the above manner have three sources of error. 1. Error due to the assumption of normality; 2. errors in a_n and b_n due to truncation; 3. error in curve segmentation. Looking at Fig. 4.1, it can be seen that normality assumption is at best an approximation. The vocal air column is a complex resonator and the assumption that it be considered as a combination of several simple resonators is not entirely valid. Curve segmentation can be improved by the use of the slope information.

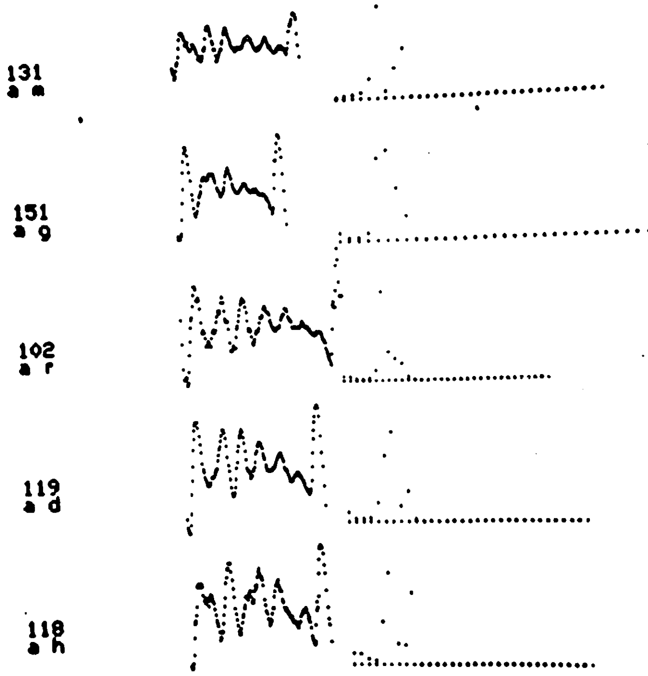


Fig. 4.2

Although it might be desirable to modify these assumptions later the accuracy obtained was satisfactory for the present investigation.

The formant frequencies obtained above were, in general, in agreement with the published results obtained by the use of the spectrograph. However, there were other inexplicable formants before the first and between the first and the second. The power at these formants was generally low but on occasions the power of some formants before the first was larger than that of the second. Therefore, one cannot neglect the local maxima below a certain power without the risk of losing some of the required ones. Further the level of significance varies from person to person. On occasions the second formant was totally absent. This was particularly noticeable in the case of 'u' and 'o'. This is possibly due to the errors 2 and 3 mentioned above. These characteristics may be observed in some of the spectral distributions shown in figures 4.1, 4.2, 5.1, and 5.2. Even with such variability, the results were encouraging enought to be used to recognize some vowels described below.

5. A Simple Vowel Recognition Algorithm:

Five phonemes i,e,a,o, and u were chosen for recognition out of the eleven or so vowels in the phonemes of the English language. These were chosen for their ease of generation as continuous vowel sounds compared to the others.

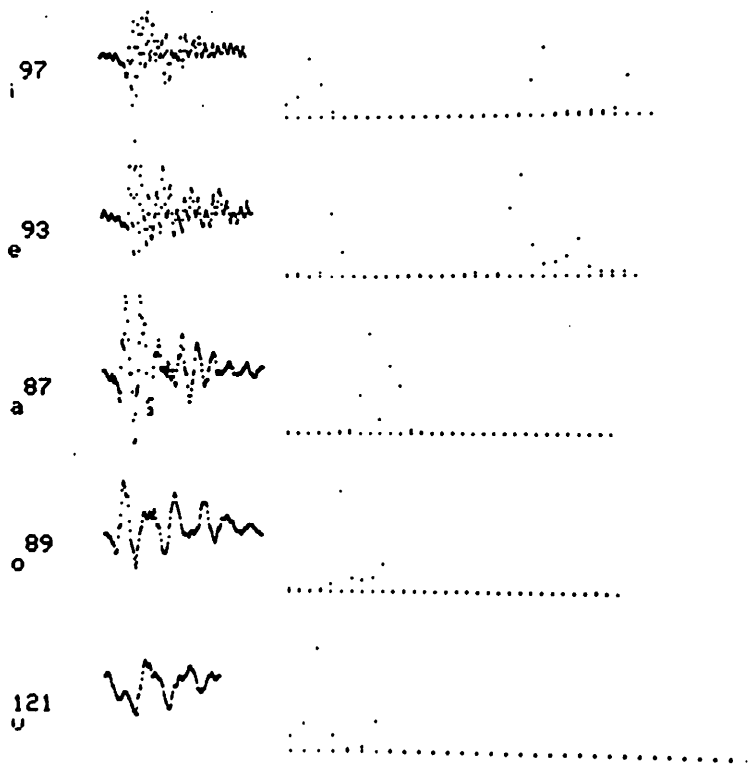


Fig. 5.1

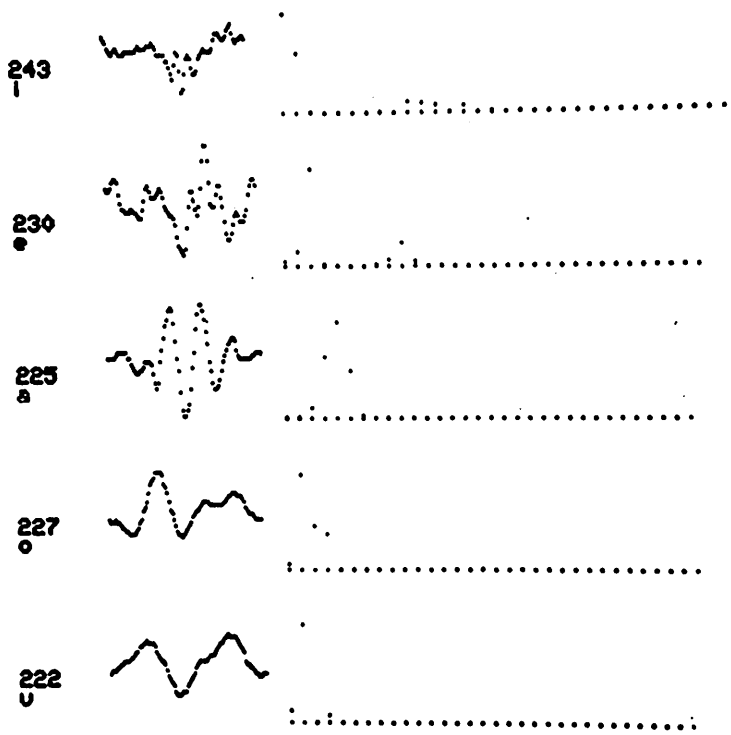


Fig. 5.2

When the intensity of loudness if over a pre-determined level the computer starts recording the speech sound for 1/10 of a second. Using this speech sample it selects a fundamental which is then used to determine the spectral distribution of the wave. Figures 5.1 and 5.2 are typical spectral distributions of i,e,a,o, and u when uttered by a male and female subject respectively. Using the spectral distribution, the formants of the given sound wave are determined.

Be examining several sets of formant frequencies of the vowels by different speakers the following relationships were derived.

Let p be the pitch of the sound. Then the spectral distribution can be divided into three regions.

1. First formant region. Frequency interval $[0,p + 1000]$
2. Second formant region. Frequency interval $(p + 1000,3300]$
3. No interest region. Frequency interval $(3300,\infty)$

Let f_1 be defined as the formant frequency with the greatest power in $[0,p + 1000]$. Let f_2 be defined as the formant frequency with the greatest power in $(p + 1000,3300]$. Then vowels i, e, a, o, and u may be distinguished from each other by the following algorithm.

If $f_1 > p/2 + 670$ then 'a' else if $f_2 > 1660$ then

(if $f_1 > p/4 + 375$ then 'e' else 'i') else if $f_1 > 3p/8 + 425$

then 'o' else 'u'.

The main characteristic of the above algorithm is it uses the pitch of the sound as a parameter and it is simple to use. It has been used with a high degree of success among men, women and children.

It is highly doubtful that the recognition algorithm would remain so simple as the number of phonemes increase. We believe that the recognition effort required will increase exponentially with the number of phonemes.

6, Synthesis of Vowel Sounds:

In an attempt to understand the composition of the sounds better, waves of different shapes were created and played back. The following approaches were used to create the sound waves. 1. Free-hand sketches of a desired sound wave were generated on the scope using a light-pen-tracker program. 2. The wave was specified by the use of a polynomial of any desired degree. Different polynomials could be used to generate different parts of the wave. 3. The wave was specified as a combination of sin and cosine waves of given frequency and amplitude.

The results of the first two approaches were disappointing. They needed a lot of effort to create and when played back usually bore no resemblence to any known sound. The results of the third were more encouraging. Using the power spectrum of the vowel sounds generated during the analysis and neglecting powers of low magnitude both phase-dependent and phase-independent sound waves were created. These and the original fundamental that was used to create these when played back showed little noticeable difference. Although all of these were recognizable, the clarity of a normal speech sample was lacking. This was attributed to the fact that an isolated wave when played back repeatedly does not have the minor variations of a normal speech sample which perhaps adds to its clarity.

7. A Proposed Automatic Speech Recognition System

Although there are said to be about 40 phonemes in the English language, it is by no means true that each has associated with it a unique sound (acoustically speaking). Different samples of the same sound vary a great deal in pitch, duration and spectral energy between individuals (and to a lesser extent for the same person).

The human mind classifies these sounds to fit one of the patterns it is aware of. When it comes across a phoneme that is foreign to its vocabulary it merely associates with it a phoneme that is close (e.g. german ü is usually classified as u.) Further even when certain words are omitted from a spoken sentence, the mind is capable of arriving at the intended meaning. It can do all this because of the highly complex pattern recognition system it possesses. Time taken for recognition of a familiar sound is less than that for an unfamiliar sound. That is why people have difficulty in understanding a southern drawl or a foreigner speaking English and many times the meaning is grasped belatedly.

In its effort to classify sounds, the human mind draws upon the resources it has built up over the years, not only through hearing but also through other sensory mechanisms. Each concept it forms has associated with it a sound pattern, visual pattern (of characters) and others (like hot, sweet, smooth etc.); what we might call a property list.

Experiments show that human recognition drops markedly when presented with a string of nonsense syllables as against well-constructed sentences. Certain sound patterns produce a sensation of joy (perhaps because of the composition and regularity). Other sounds cause distress (as in the case of noise) perhaps because the mind gets overworked trying to associate a concept with it and fails.

Therefore it is obvious that a good Automatic Speech Recognition System should make use of linguistic and contexual information[7] to cope with the inconsistencies of the acoustic properties of the speech sounds.

Using a spoken sentence as an input, such a system would have to split it into an ordered set of phonemes; combine such phonemes into an ordered sequence of words within its vocabulary and use linguistic information to associate a concept with it. It would be further desirable to have it make audio (or visual) responses in reply.

A spoken sentence of normal length usually takes less than 5 seconds of time. Assuming that audio signals are sampled at 50 microsecond intervals we would need a high speed store of 50,000 - 18 bit words packing 2 samples per word. While the sampling is proceeding the whole sentence can be scanned and segmented to form a number of distinct sounds. This segmentation will be complicated by the fact that different phonemes have different starting and ending characteristics depending on whether they are voiced or unvoiced, plosive, ficative etc.

Each such segment is then analyzed to determine which phoneme has similar characteristics to the segment in question. Time normalization of the segments may or may not be required in such an analysis. Determination of the distinguishing attributes of various phonemes would constitute a major portion of the research. Fourier analysis may be used to determine the energy at various frequencies. Zero crossing counts, number of local maxima and minima, and pitch of the sound may also carry additional information. Means and variances of various measures should be collected on individual basis which can then be used to identify various phonemes uttered by that speaker. These can be consolidated later to form general measures of various attributes. Duration of utterances and effect of the adjacent phonemes on the acoustic characteristics of a phoneme should also be investigated. The vocabulary of the system should consist of 200 or so most frequently used English words with their phonetic equivalents and grammatical information. Too many words would make the system very slow and too few would make it less versatile. The phonemes should be grouped to form an ordered sequence of most likely words using grammatical information and phrase structure of the language. Any string of phonemes that cannot be grouped together may then be grouped into one or more phonetic words. This would be the first phase of the system. The output can then be displayed on the scope or typed out or both.

The second phase would deal with the semantics of such an ordered set of words, with a view to associating a concept and generating a reply in response. Such a function of the system may well be taken over by the Advice Taker formalism now being developed at Stanford. Such a response can be in the form of audio messages formed by synthesis of speech sounds by the computer.

However, the Advice Taker assumes that the input to the system is error-free. It does not have any mechanism like that of the human mind which can fill in the blanks when the information presented is incomplete or unintelligible by the use of the linguistic information available to it. Some form of error detection and correction should also form part of the second phase.

It is estimated that the total time required for the analysis will be over 20 times the time taken for recording the sentence. Perhaps the segmentation of the sentence into phonemes can be performed simultaneous to recording. In the absence of large high speed storage the segmented utterances will have to be stored on a disk involving a lot of bookkeeping and delaying the process of recognition.

Such a project as outlined above would require a computer with large storage and cycle time in the sub-microsecond range for real-time operation. Of course most of the preliminary ground work can be done before we are in possession of such a computer.

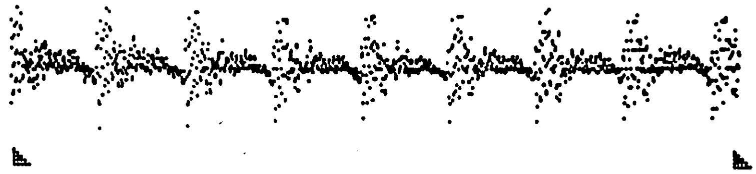
An excellent collection of articles on various aspects of the above topic may be found in "Automatic Speech Recognition" put out by the University of Michigan for a summer conference under the direction of Dr. G. E. Peterson.

7. Conclusions

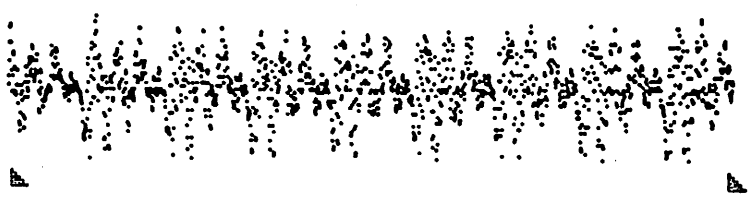
This investigation shows that most of the speech research that was conducted in the past by the use of special equipment can now be done by the use of a digital computer just as well if not better. This paper illustrates the use of a digital computer in the determination of the fundamental, spectral analysis of the fundamental, determination of formant frequencies, for vowel recognition and for synthesis of sounds. The paper concludes with some comments on the future of research in this field.

APPENDIX: Wave Forms of Some Speech Sounds:

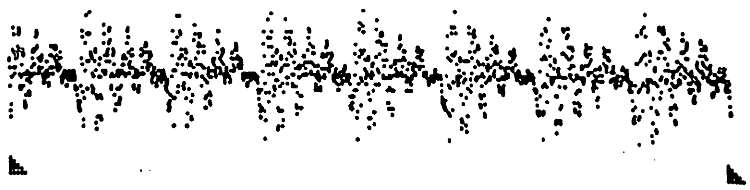
The following pictures illustrate the wave form of several vowels and some consonants. Each wave has the identification to the right of it.



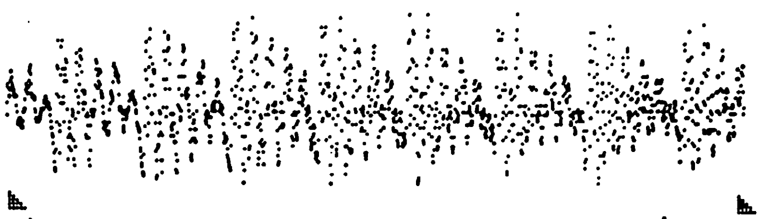
i as in beat



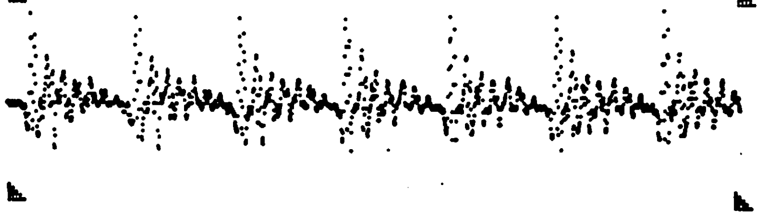
I as in bit



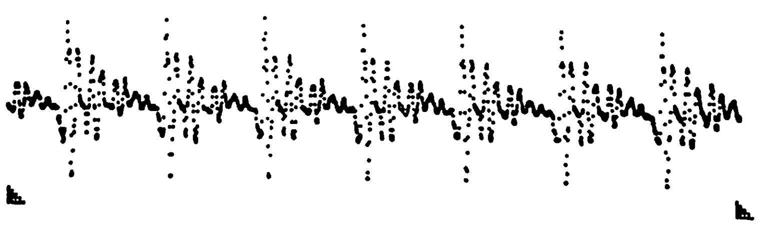
e as in bate



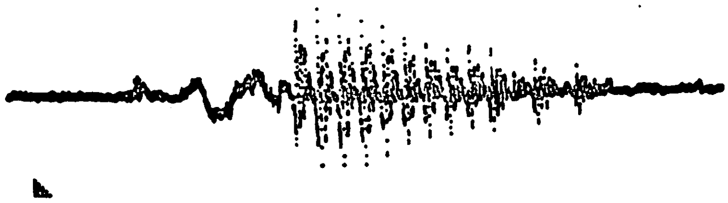
ϵ as in bet



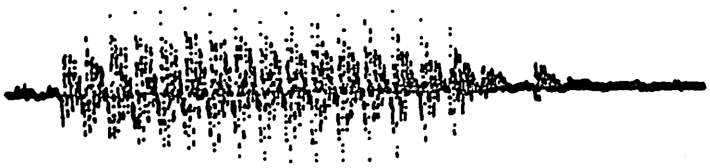
æ as in bat



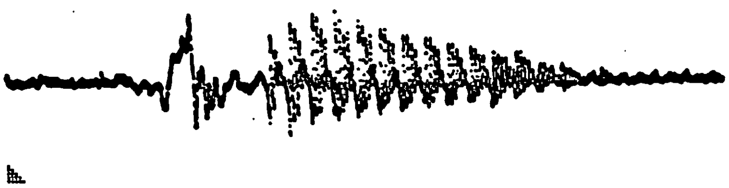
a as in father



ʌ preceded by unvoiced stop consonant t

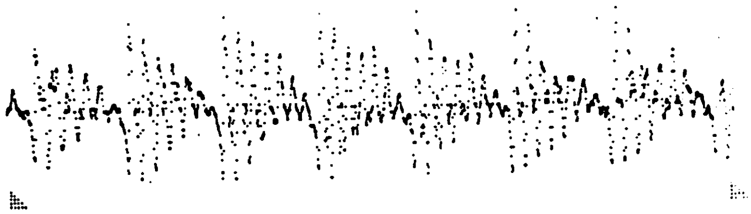


ʌ preceded by voiced stop consonant d

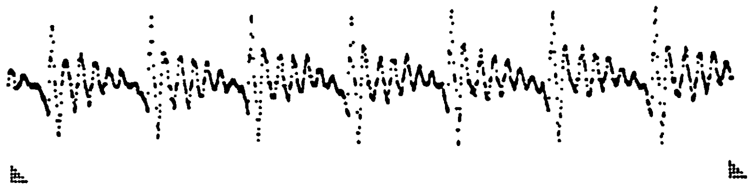


ʌ preceded by unvoiced p

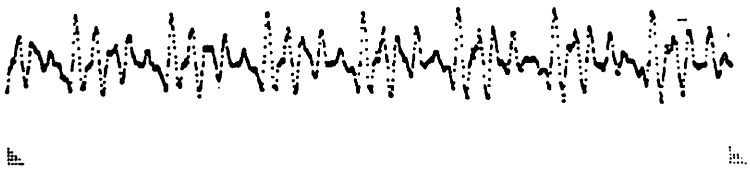
ʌ preceded by voiced b



ʌ as in but

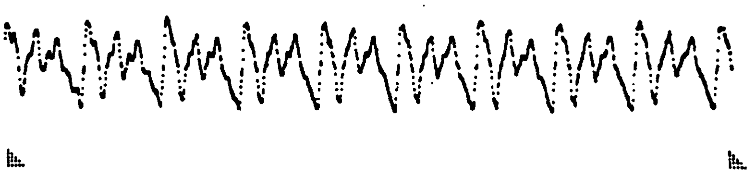


ɔ as in ought



o as in boat

U as in Buddha



u as in boot

a preceded by unvoiced fricative s (as in saw)

a preceded by unvoiced fricative ʃ (as in shaw)

a preceded by voiced fricative ω

a preceded by voiced fricative υ

REFERENCES

1. Chapter 3, Psycholinguistics by Saporta; Holt, Rinehart and Winston

2. Visible Speech by Potter, Kopp and Green; Van Nostrand

3. Introduction to Spectrography by Pulgram; Mouton and Co.

4. Elements of Acoustic Phonetics by Ladeford, The University of Chicago Press.

5. Automatic Speech Recognition, collected papers; University of Michigan Engineering Summer Conference Series

6. Advanced Engineering Mathematics by C. R. Wylie; McGraw-Hill

7. See the articles by Prof. Fry in reference 5 on this topic.