

SPEECH RECOGNITION :
STUDY OF FUNDAMENTAL FREQUENCY
AS A VARIABLE

Register No : M 8701
Ashok Kumar

A Dissertation submitted as part fulfilment for
Final year M.Sc. [Speech and Hearing]
to the University of Mysore

ALL INDIA INSTITUTE OF SPEECH & HEARING
MYSORE - 570 006
MAY - 1989

To,
PAPA - MUMMY

CERTIFICATE

This is to certify that the dissertation entitled

**SPEECH RECOGNITION : STUDY OF FUNDAMENTAL
FREQUENCY AS A VARIABLE**

*is the bonafide work in part fulfilment for the degree of Master of
Science [Speech & Hearing], of the student with Register No. M 8701*

N. R. Patil
12/1/89

Director

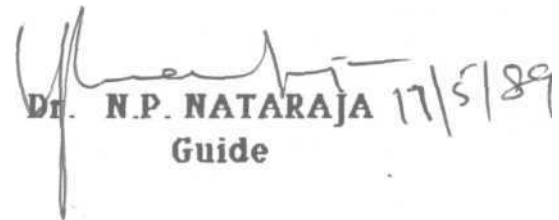
**All India Institute of Speech & Hearing
Mysore - 570 006**

CERTIFICATE

This is to certify that this dissertation entitled

**SPEECH RECOGNITION :
STUDY OF FUNDAMENTAL FREQUENCY AS A VARIABLE**

has been prepared under my Supervision and guidance.


Dr. N.P. NATARAJA 17/5/89
Guide

DECLARATION

This dissertation is the result of my own study undertaken under the guidance of Dr. N.P. Nataraja, Reader in Speech Science, All India Institute of Speech and Hearing, Mysore, and has not been submitted earlier at any University for any other Diploma or Degree.

**Mysore
May 1989**

Reg No. M 8701

ACKNOWLEDGEMENT

I express my deep gratitude to My teacher & guide Dr. N.P.Nataraja, , without whose unceasing inspiration & constant interest this dissertation could not have been achieved.

i am grateful to Dr. Rathna, Director, A.I.I.S.H, Mysore for giving me the opportunity to carry out this dissertation.

My must acknowledge my deep sense of gratitude to Mr (Venkatesh.C.S. Lecturer, Dept.Of Speech Sciences, A.I.I.S.H, Mysore who gave me quite a bit of his valuable time during various stages of this research work.

i would like to thank to my parents , brothers & sister for their encouragement at every stage of my educational life.

I thank, all my classmates, Mr.Ani & Mr.Kittu for their encouragement and help.

I am grateful to Madhu and Ruchi for everlasting inspiration & constant encouragement

I am thankful to Mr.dabu for neat and timely typing .

Above all; I am grateful to all the subjects and those who helped me in completing, this work.

Last, but not least, my sincere thanks to Creative Computer Centre for taking such great care & typing out my dissertation.

CONTENTS

Chapter	Page No.
I INTRODUCTION	1 - 6
II REVIEW OF LITERATURE	7 - 72
III METHODOLOGY	73-78
IV RESULTS AND DISCUSSION	79-82
V SUMMARY AND CONCLUSION	83-85
BIBLIOGRAPHY	

INTRODUCTION

Man's primary method of communication is speech. He is unique in his ability to transmit information with his voice. Of the myriad varieties life sharing our world, only man has developed the vocal means for coding and conveying information beyond a rudimentary stage. It is more of his credit that he has developed the facility from apparatus designed to subserve others, more vital purposes.

As early as the eighteenth century, attempts were made to model the human speech mechanism. Without the aid of electricity these were physical analogues of the apparatus - the lungs, vocal cords, vocal tract, tongue, teeth and lips - but were accurate enough to yield sounds recognisable as human utterances.

In 1939 a device called the 'voder' (voice operation demonstrator) was produced at the Bell Telephone Labs in New Jersey which could be operated by trained operators to emit speech sound.

Following the second world war, much more advanced speech synthesizers were produced using developments of the same technique. Algorithms which were suitable for digital implementation were then produced in Laboratories and Universities around the world as soon as mini-computers became readily available in 1960's.

Recognition task:

Even the simplest attempt at speech recognition requires electronic equipment, in order to capture utterances with a microphone and to analyse them. In 1947 this process was demonstrated by an important device known as the spectrogram which produced a graphical representation of the continuous spectrum of the speech.

From 1950 onwards many experiments began in which the parameter extracted from speech utterances by some form of electronic filtering were used to make automatic 'decisions' about the speech itself.

In 1952 a team at Bell Labs developed the first device that could be properly called an automatic speech recognizer. Since 1952, many techniques have been developed to enhance the performance of speech recognizers. However, the basic principle of stored information representing various options for comparison is always used, together with some form of parameter extraction and pre processing that takes place before the decision - making process.

The first requirement in the design of a speech recognition system, then, is to define clearly what particular aspects of speech are to be recognized. Some of the best used categories in this respect, in order of increasing complexity and expenses are; recognition of isolated words, recognition of discrete words in connected speech recognition of strings of words in connected speech.

Automatic recognition of speech, with its promise of voice -operated type writers and other machine marvels, has been one of the more glamorous technological dreams for several decades. With the emergence of ever faster computers the voice - operated millennium seemed at hand.

Automatic speech recognition and speaker verification are among the most challenging problems of modern man-machine interaction.

Among their numerous useful applications are a future "checkless" society in which all financial transactions are executed over the telephone and "signed" by voice. Access to confidential data can be made secure by speaker certification. Other applications include voice information and reservation systems covering a wide spectrum of human activities from travel and study to purchasing and partner matching. In these applications, spoken requests (over the telephone, say) are understood by machines and answered by synthesized voice. Voice control of computers and spacecraft (and machines in general whose operators have limited use of their hands) is an aspiration of long standing.

Activation by voice could be particularly beneficial for the severely handicapped who have lost one or several limbs.

Application of speech recognition in the field of speech therapy has been recognized as useful. But it has not been

worked upon particularly in India. Before using these programmes for clinical use it was felt necessary to know the efficiency of minimum prediction residual method, cash measure and linear prediction coefficients (euclidean distance measure); which were available.

The review of literature has shown that duration of utterance and fundamental frequency as important variables among other variables in the process of speech recognition.

A program of speech recognition written using basic language based on description provided by Gray and Markel (1976); was available to investigator. This program provides measurement of distances between the stored group data and test data using minimum prediction Residual method, cash measure and Linear Predication Coefficient (Euclidean distance measurements)

The present study was limited to find out the effect of variations in fundamental frequency on speech recognition with the program available to the investigator.

It was decided to use digits to start with, to make the problem simpler and also as others had used digits for recognition

HYPOTHESIS:

1. There is no difference in terms of recognition of digits when the pitch is varied from the habitual pitch, i.e.,

1(a) There is no difference in terms of recognition of digits when the digits are uttered at high pitch with respect to habitual pitch.

1(b) There is no difference in terms of recognition of digits when the digits are uttered at low pitch with respect to habitual pitch

1(c) There is no difference in terms of recognition of digits when the digits are uttered at low pitch with respect to high pitch.

2. There is no difference between the three methods of speech recognition in terms of recognition of digits as used in present study.

2(a) There is no difference between the methods Cosh Measures and Linear Prediction coefficient in recognizing the digits as used in this study.

2(b) There is no difference between the methods Minimum prediction residual and Cosh Measure in recognizing the digits as used in this study.

2(c) There is no difference between the methods Minimum Prediction Residual and Linear Prediction Coefficient in recognizing the digits as used in this study.

Limitations of this study:

1. Only five subjects were used.
2. The Speech Recognition Program that was available to the investigator was limited.
3. Only digits have been considered as speech sample
4. Only frequency has been considered as a variable here.

Implication of this Study:

1. The study has indicted that this particular program available to the Investigator had limited use.
2. It has shown that the fundamental frequency can be a variable in speech recognition.
3. The duration of the stimuli must be constant for the use of this particular program.
4. Attempts can be made to use this for consonant recognition in Articulation Testing and therapy.

REVIEW OF LITERATURE

Speech Communication is the transfer of information from one person to another via speech, which consists of variations in pressure coming from the mouth of a speaker. Such pressure variations propagate as waves through the air and reach the ears of listeners, who decipher the waves into a received message. The chain of events from conception of a message in the speaker's brain to their arrival of the message in the listener's brain is called speech chain. The chain consists of a speech production mechanisms located in the speaker, transmission through a medium such as air, and a speech perception process in the ears and brain of a listener.

In many applications of speech processing, a part of the chain is implemented by a simulation device. Automatic synthesis or generation of speech by algorithm (as in a computer) can take the speaker's role, except for generation of the original message, which is usually in the form of a text. In automatic speech of speaker recognition, an algorithm takes the listener's role in deciphering speech waves into either the underlying textual message or a hypothesis concerning the speaker's identity.

Speech coders allow replacing the analog transmission medium (such as air or telephone lines) with a digital

version, modifying the representation of the signal; in this way, speech can be efficiently stored and transmitted, often without noise problems and with enhanced security.

Vocal communication between people and computers includes the synthesis of speech from text and automatic speech recognition (ASR), or speech-to-text conversion. The design of automatic computer algorithms to perform these two tasks has been more successful for synthesis than for recognition because of a symmetries in producing and interpreting speech.

(The main difficulty of Automatic speech recognition, are the problems of segmentation and adaptation. For both synthesis and recognition, the input is often divided up for efficient processing, typically into segments of some linguistic relevance. In text-to-speech synthesis, the input text is easily divided into words and letters. Whereas the speech signal that serves as input to automatic speech recognition provides (at best) only indications of phonetic segment boundaries. Sudden large changes in speech spectrum or amplitude are often used to estimate segment boundaries, which are nevertheless unreliable due to coarticulation. Boundaries corresponding to words are very difficult to locate except when the speaker pauses. Most commercial recognizers require speakers to pause briefly after each word to facilitate segmentation.

According to Stark (1981); the segmentation problem can be partially overcome through compensation in speaking style. Three styles of speech (in order of increasing recognition difficulty) can be distinguished: Isolated-word or discrete utterance speech, connected-word speech, and continuous speech. Continuous speech recognition (CSR) allows natural conversational speech, with little or no adaptation of speaking style imposed on system users. Continuous speech allows the most rapid input (e.g., 150-250 words/min.), but it is the most difficult class to recognize. Requiring the speaker to pause for at least 100-250 ms after each word in isolated word recognition (IWR) is unnatural for speakers and slows the rate at which speech can be processed (e.g., to about 20-100 words/min.), but it alleviates the problem of isolating words in the input speech signal. Connected-word speech represents a compromise between the two extremes, the speaker need not pause but must pronounce and stress each word clearly.

The other major difference between synthesis and recognition concerns adaptation. Human listeners modify their expectations when hearing synthetic speech and usually accept it as they do speech from a strange dialect or with a foreign accent. In automatic speech recognition, however, it is the computer that must adapt to the different voices used as input. It is much easier to produce one synthetic voice

to which human listeners adapt than to design a recognition algorithm that can cope with the myriad ways different speakers pronounce the same sentence or indeed to interpret the variations that a single speaker uses in pronouncing the same sentence at different times - Human listeners are more flexible in adapting to a machine's accent than a computer is in deciphering human accents.

Current systems require speaker to modify their speech e.g., by pausing after each word or by speaking clearly and slowly. No commercial systems have yet been developed that accept truly natural, continuous speech. Most systems are speaker dependent, demonstrating good performance only for speakers who have previously trained the system. These system "adapt" to new users by requiring them to enter their speech patterns into recognizer memory. Since memory and training time in such systems grow linearly with the number of speakers, less accurate speaker - independent recognizers are useful if a large population must be served. These systems are trained by many speakers, not necessarily by those who later use the systems. Some systems of both types truly adapt in time via learning procedures as speaker enter speech. Correctly recognized speech modifies the patterns stored in memory, keeping up-to-date on new speakers and on evolutions in speaking style)

Components and Continuum:

An intonation 'contour' implies a continuous functions running through an utterance. Although, at an abstract level, description may be in terms of discrete components such as a 'fall' or a 'level' or a 'fall-rise' which are then concatenated, perceptual impression suggests, that pitch movement in an utterance is continuous. This is essentially a correct impression - the fundamental frequency of the speech signal, except for unavoidable breaks caused by voiceless segments, rises and falls throughout utterance.

The most basic problem of speech recognition is how to reconstruct the discrete linguistic component from the overlapping and interwoven cues in the acoustic signal.

Variation:-

In considering the variation encountered in speech it is useful to separate out kinds of variation which can be described in terms of the linguistic system from those kinds which stem from personal characteristics of the speaker. Under both headings there are aspects which remains relatively constant for a given speaker, and aspects which varies across utterances made by the same individual.

Linguistic Variation:-

As yet it is not well understood how far informality, rate, and redundancy interact in their control of phonetic

processes. In designing speech recognizers there should be pay off between the difficulty of incorporating increasingly sophisticated knowledge of different styles of speaking, and the advantage of reducing the constraint on speakers to adhere closely (and perhaps unnaturally) to a single speaking style.

Personal variation:

According, to Bristow (1986) The physique of the vocal apparatus determines the range within which acoustic parameters can vary, rather than fixed values. For instances the smaller size and mass of female vocal cords compared to male determines (other things being equal) a higher fundamental frequency range; but there is usually some overlap between the normal ranges of a man and a woman, and a man can override his natural range by producing falsetto.

Unfortunately (from the point of view of speech recognizers aiming to cope with different speakers) the relation between the acoustic o/p of different vocal tracts is not straight forward. It is possible to say, for instance, that female formant frequencies are on average about 20% higher than those of male speakers, a mean value such as that disguises considerable non-uniformity in the effects - resulting partly from the fact that vocal tracts differ more in pharynx length than in the length of the mouth

cavity. A procedure to equate male and female formant frequencies has to be specific to different types of vocoid and to each formant.

Most of the work in this area has concentrated on male female scaling of vocoids; but similar scaling problems exist between any two speakers even of the same sex, and across the whole range of sounds. Considerable progress needs to be made before an automatic speech recognition system can be achieved which will adapt to any new speaker without a lengthy "training" phase - that is, one which will replicate what human listeners do without difficulty on encountering a new voice.

In such a recognizer adaptation may well have to be a continuing process, since the personal characteristics of a person's voice change even in the short term. Even assuming a constant linguistic style, some of the properties of a voice will change as the speaker become tired, stressed and irritable, louder (perhaps as a result of increased background noise) and so on.

In the longer term, changes correlate diurnal rhythm, health (such as colds and ailments of the larynx) and (probably of least consequence to speech recognition) the ageing process.

ELEMENT OF SPEECH RECOGNITION

"Thirty years of history of speech recognition is speckled with limited successes and repetitive rediscoveries of old ideas, and yet with a growing ability to successfully handle small vocabularies of words spoken in isolation. Recent trends have added successes in recognition of continuous speech such as strings of digits and spoken sentences related to a restricted task domain, and the technology is currently expanding rapidly. Some important gaps still remain and future work will have to overcome some challenging problems" (Lea; 1980).

(Speech signals convey information about who spoke what message in what manner and in what environment. The task of a speech recognizer is to automatically determine the message (i.e., what was said), regardless of (or perhaps with some help from knowledge of) the variabilities introduced by speaker identity, manner of speaking, and environment conditions.

Speech recognition can be generally defined as the process of transforming the continuous acoustic speech signal into discrete representations which may be assigned proper meanings, and which, when comprehended, may be used to affect responsive behavior.

The ultimate goal is to understand the input sufficiently to select and produce an appropriate response.)

In actual fact, the effectiveness of every recognizer (even small devices for in dated word recognition) is determined by the appropriateness of its response. Input utterances may be considered equivalent if they have the same intended response, or different if they should yield different responses. This produces "equivalence classes" of spoken inputs, regardless of any irrelevant signal changes due to changes in talker, the speaking rate, the recording environment or variabilities in details of pronunciation.

A recognizer's decision is correct if the input assigned to the right equivalence class and incorrect if the wrong response would be produced (regardless of how many "phonemes", words, phrases, or aspects of signal are properly classified). The recognizer must recover the original intended message.

Performance evaluation of a recognizer is concerned in part with establishing what percentage of the spoken utterances produce a correct machine response. There is, of course, some uncertainty associated with the selection of correct responses, so decisions are made with the hope of minimizing the number of errors, or equivalently, reducing the probability of an error. However, the ultimate

evaluation of the recognizer is concerned with the correctness of its final responses, not its intermediate results.

According to Lea (1980); Another related aspect of system performance concerns how severely a system is affected by deficiencies of usual information, or deprivation such as the removal or malfunction of a system component. It is useful to know how much of the action of the recognizer is attributable to each knowledge source or system component, and to design systems so as to permit leeway or errors in one aspect, by permitting another component (such as syntactic analysis) to recover from errors in an earlier part of the system (such as the component that identifies small units). "Gradual degradation" is desired so that minor changes in either the input channel conditions or the system structure will have catastrophic effects on system accuracy.

Other aspects of the evaluation of speech recognizers are concerned with generality and enhanceability of the system. Systems are obviously of more general utility, if they permit a large number of speakers (of both sexes and various dialects) to speak in a natural manner (usually, involving large vocabularies, continuous, uninterrupted speech and loosely contained message structures), even in the environment of some noise and signal distortions such as a telephone channel might introduce.

The speed with which machines make recognition decisions, the required memory size and processing power, and the cost are other factors that must be considered. Also of some importance is the ease with which an available limited recognition capability can be enhanced to handle somewhat more difficult tasks such as new vocabularies, new structures in the spoken commands, new talkers, etc.

"According to Lea (1980) One of the primary reasons given for the use of more complex linguistically oriented recognition schemes instead of simpler mathematical techniques is the expected ease of enhancement and ultimate generality of linguistic approaches. However, for immediate success on limited problems, the more mathematically oriented approaches have repeatedly proven to give better results in shorter times"

Some designers of recognizers focus only (or primarily) on mathematical representations of the recognizers i/p - o/p characteristics, asserting that each i/p must merely be composed with previously stored representatives or templates of each equivalence class of inputs, and the nearest higher or minimally different representative must be selected as the identity (or equivalence class) of the current i/p signals (or, if no template is near enough, an error message might be given). This is the basis of generalized i/p - o/p functions (Newell, 1975), linear discriminant analysis, other

statistical models, and general pattern recognition and signal processing schemes. Such recognition models could apply as well to signals other than speech, and indeed their technology is highly developed because of such other applications. They do not consider how the signal was produced by the speaker, nor how it is normally perceived by the human listener. They of course, do not require that the recognizer operate internally in any way similar to the human's perception processes.

According to Lea, (1980) This generalized i/p - o/p or signal processing approach represents the first of the following four basic viewpoints about how to be guided towards the design of successful speech recognizers:

1. The Acoustical signal viewpoint:

It asserts that since the speech signal is just another waveform (or vector of numbers), simple general signal analysis techniques (Fourier Frequency spectrum analysis, principal component analysis, statistical decision procedures and other mathematical schemes), can be applied, to establish the identity (or representative "nearest neighbour") of the input.

2. The Speech Production Viewpoint:

It suggests that the communicative "source" of the speech signal is understood by individual and also individual

can capture essential aspects of the way in which speech was produced by the human vocal system (e.g., vocal tract resonances, rate of vibration of vocal cords, manner and place of articulation; coarticulatory movements etc.)

3. The Sensory Reception Viewpoint:

It suggests that duplicating the human auditory reception process, by extracting parameters and classifying patterns as is done in the ear, auditory nerves, and sensory feature detectors.

4. The Speech Perception Viewpoint:

It suggests that features are extracted and categorically distinguished that are experimentally established as being important to human perception of speech (e.g., voice onset times and formant transitions as cues to state of consonant voicing, "single equivalent formants" as vowel distinguishers, perceptual "feature detractors" etc.).

Combination of these viewpoints can also be devised. The four viewpoints reflect different ways in which the linguistic message being communicated is encoded at various stages in the production and reception of speech. The sensory reception and speech perception viewpoints have had much less effect on actual recognition systems than the speech production and acoustic signal viewpoints.

The speech production, sensory reception and speech perception viewpoints may be characterized as what Newell (1975) called "knowledge - source driven representation", which assume that recognition can be based on available knowledge of processes and speech encoding or decoding.

These three viewpoints assert that knowledge of the acoustic speech signal alone is not enough to fully determine the message (or intended machine response); other sources of knowledge must be brought to bear on the recognition problem. These viewpoints also acknowledge that, while a machine need not operate internally in the same manner as the human. The human speech processing abilities can serve as a successful prototype system" for guiding the development of machine algorithms for speech recognition. The conversion from acoustic signal to machine response, without the intervention of help of the human, involves the machine functionally duplicating the overall. (i/p - o/p) function of a human perceiver. It need not structurally duplicate the human ear brain system but the knowledge - source - driven viewpoints would suggest that recognizers may glean guidelines for effective recognition from study of human speech processing techniques.

Computers can currently do some analyses better than humans, and some others less adequately; so a controversy continues between mathematical (statistical, information-

theoretic, signal processing, or pattern-classifying) methods and human-oriented (phonetic, linguistic, perceptual or neurological) approaches.

Distance Measures and Template Matching:

A critical task in speech recognition is to extract all (and only) those parts that convey the message. No matter how many parts or times. The speech signal is divided into, to study minutely of all its timing data, an even finer analysis is always possible. Similarly, no matter how exactly the pressure (or voltage) of a speech wave is measured at an instant, a finer grain analysis is conceivable. Thus, speech is a two dimensional non - denumerable continuum. It is possible to analyse it without arbitrary or linguistically dictated divisions into time segments and no quantizations into significant changes in signal levels. Point by point differences in incoming and stored signals can be calculated, without regard to the possibility that some signal differences are more important than others.

To determine the identity of the incoming speech and effect the correct machine response, a recognizer can determine the difference between the incoming signal and the expected signal for each message. 'Expected' signals, or templates, can be actual stored training samples that were previously declared by the human to be associated with each appropriate machine response, or they can be averages or

other composite signals obtained from many such training samples. For each equivalence class of utterances (associated with a specific correct response), a template (or perhaps several templates representing allowable variations within the equivalence class) must be specified, and a distance measure determined such as calculation of a 'Euclidean distance' formed from the squares of the point-by-point differences in signals. The selection of an appropriate distance measure is thus one of the important concerns in this signal matching approach to recognition.

According to Lea (1980) ; this (in its simplest form) is basically an ignorance model of the significant aspect of the incoming signal. Each deviation from previously stored signals is assigned equal significance, and no particular features of the speech are considered more important than others. This ignores or disregards whatever is learnt about the important physiological and linguistic regularities of both the source of the speech and the intended intelligent receiver of spoken messages. It requires accumulation of representative training signals for each possible message. For large message sets (large vocabularies of isolated words, or large number of alternative sentences that might be continuously spoken), obtaining training templates is a time consuming and costly, process and results in extensive storage requirements in the machine.

Users are required to invest the time in training the machine to their voices or speaker - independent templates must be devised by the developers, based on averages or alternative variant of word pronunciations.

In addition, it is difficult to obtain truly 'representative' training signals, that are quite distinct from message to message, and that are likely to be closely approximated by all repetitions by all speakers in all environments. To account for speaker differences, variations due to environmental noise and channel distortions, several alternative templates usually have to be stored for each message and each set of conditions.

This further increases the storage requirements and makes recognition costly and unwieldy. However, with rapidly advancing speeds and storage capacities in low cost computers, this storage need is less of a problem than it might have seemed in earlier years.

The template approach also avoids the danger of inadvertently 'throwing away' important information while focussing on only certain parameters that are expected to be important.

The variety of template matching procedures is extensive, since a slightly different technique can appear to emerge from each new set of possible parameter patterns to

extract, each new method of data alignment, each new distance ensure and each decision or word matching method. Vector quantization (cf. Burton and Shore, 1986; Tsab and Gray, 1986) is an example of a template matching process that at the same time may seem very different from, and yet essentially a terminological equivalent of, previous methods. Advantage is the ability to define automatically the set of alternative spectral templates for categorizing portions of speech.

Traditional phonetic Recognition Procedure:

Here, the justifications of discrete representation of speech and the traditional view of segmentation of speech into phonological units of various sizes and kinds are considered.

The Discrete Representation of Speech:

There is extensive justification for characterizing speech by discrete units in time and as a finite set of simultaneous features.

Engineering and information - theoretic considerations suggest discrete representations. Speech must be converted into discrete voltages and switching states in digital computers, so that a discrete representation will be required for speech recognition using practical digital computers. Also, the sampling theorem (Pierce, 1961; Rabiner and Gold, 1975) asserts that only any limited signal (such as speech

may be considered to be) can be completely represented by $2 F_u$ samples per unit time, where F_u is the upper bound on the frequency content of the signal (which can certainly be taken to be at most 10 KHz for speech).

The fidelity criterion of communication theory (Chomsky and Miller; 1963) acknowledges that certain sound changes (e.g., those due to emotion, fatigue, speaker idiosyncrasies, et.) are irrelevant to the receiver, while others are significant, so that the infinity of possible speech waves of a limited length can be partitioned into a finite set of discrete, mutually exclusive 'equivalence classes', which are basic to speech recognition.

Halle (1954) has also noted that "if a discrete view be adopted, correction of errors begin upon receipt of each discrete unit (quantum)", so there is no need to wait until the entire continuous signal is completed to correct errors. This may prove significant in the real time recognition of spoken sentences. Finally, most messages for machine i/p would have identical intended responses if they had been written (or typewritten), rather than spoken, so the fact that the written code is discrete suggests the sufficiency of a discrete representation of speech.

No representation could be fully discrete unless it breaks up each of the two acoustic continua (continuum of

time and continuum of pressure levels into a countable in practice, finite) number of discrete parts. A non-denumerable infinity of possible speech wave forms of a fixed length must be identified in the same discretely representable class. Likewise, an infinity of possible lengths must be classified as 'equivalent'. Thus, a discrete representation of the time domain requires a segmentation of the continuous speech waveform into some sort of units or 'segments'. While the segments or units need not be separated with strict 'boundaries' between them, and could actually overlap each other or be spaced apart from each other, the usual method is to segment speech into juxtapose units of non-aero length, which classify as equivalent any of an infinity of 'insignificantly different' wave shapes spanning about the same stretch of time.

According to Bristow (1986) speech is segmented at fixed intervals. Commonly, such segments are selected to be short enough to allow to proper features extractions, such as proper spectral averaging or detections of periodicities in the wave form. This smallest unit of time (such as a short 10 ms unit) can be sensibly categorized into one of a finite number of sound categories, such as the most similar of a set of stored training templates obtained from previous processing of training data (Baler, 1975; Klatt, 1980) Alternatively, the segment may be assigned to a phonetic

category based on the spectral content and other distinguishing feature of the wave form within that unit.

On the other hand, the total utterance (which may be a word, for isolated word recognizers, or a sentence or discourse for continuous speech recognizers) can be interpreted as an undivided entity which determines the appropriate machine response.

Segmentation of speech into phonological units:

The most controversial aspect of segmentation has concerned the relative values of intermediate size units, such as 'phones'; phoneme-to-phoneme transitions or 'diphones'; syllabic subunits like 'syllabic onsets', 'syllabic nuclei', and 'coda' 'syllables'; 'words'; and 'phrases'. There is a vast literature on the 'psychological reality' and the linguistic utility of segments like phonemes and syllables (Sapir, 1938; Bloomfield, 1933; Pike, 1945; Wells, 1947; Trager and Smith, 1951; Chomsky, 1957; Chomsky and Miller, 1963; Chomsky and Halle, 1968; Hockett, 1972).

A. typical speech Recognition system:

The speech signals of representative training samples are processed to detect utterance boundaries. Points where energy rises above a threshold level are declared beginnings of words, unless the energy peak is so short in duration that it is clearly not speech, but rather a noise impulse.

Endings occur where energy drops below the threshold and stays down for at least some reasonable time. Short noise bursts have to be excluded from the regions called words, by using their short durations and perhaps their spectral character to distinguish them from speech sounds. Breath noise, lip smacks, stop bursts, weak fricative and lack of clear gaps between words within connected speech make word boundary location a difficult problem, and a primary reason for errors in recognisers. Changes of sound structure due to the sound structure of adjacent words also create word boundary-induced problems for connected speech recognizers.

Between utterance boundaries, various features are extracted, in each short time frame, to yield a matrix of $F_i N_i$ numbers where F_i is the number of features monitored, and N_i is the number of time segments in which such features are extracted. Each word is that is spoken during training is accompanied by a user-specified (possibly computer-prompted) identification of what that utterance meaning was (i.e. what word sequence was spoken, or what response is expected). A lexicon of expected pronunciation for all the words is thus obtained.

Later, when an unknown utterance is spoken, its matrix of numbers is compared with all the stored matrices, and the one lexical entry that is 'closest' to i/p pattern is the selected identity of the word. Hypothesized words may have

to be 'warped', or normalised, in time and other parameters, to match the i/p. Other normalisations may adjust signal amplitudes or other data values to aid proper alignment, comparison and scoring of closeness of fit.

After words have been hypothesized they may be subjected to syntactic, semantic and pragmatic constraints to select the most likely actual word utterance. Prosodic cues can aid syntactic analyses, by locating phrase boundaries, stresses, regions of phonetic reliability that yield highly reliable word hypotheses, other structural cues.

Acoustic Parameterization and Normalisations:

The parameters can be mainly divided into time-domain parameters and spectral parameters.

Time-domain parameters include peak amplitudes and peak-to-peak measures. One can monitor maximum within the whole utterances, or within moderate size regions like syllables, or within short segments, such as the maximum within each single cycle of the periodic voiced speech regions. Large waveform peaks are produced by the periodic excitations of the human vocal tract by puffs of air from the vocal cords, and the time interval between successive excitation peaks provides the pitch period. The pitch period is thus detectable from the time interval between large waveform peaks, or the time interval before the waveform basically

repeats itself. A common method for deriving pitch is autocorrelation method (Sondhi, 1968), which assumes that a signal will clearly correlate with itself at displacements of one (or any integer multiple of one) pitch period. The reciprocal of the pitch period is the fundamental frequency of the voice, which is useful in prosodic analysis. Within each pitch period, there are successions of progressively smaller peaks, indicating the resonance characteristic of the human vocal tract as a resonating tube. The number of peaks per pitch period can be a crude indicator of whether the signal is rapidly varying, as in fricatives, or slowly varying, as in vowel-like sounds. For vowel-like sounds, the peak count in each pitch cycle can be a cue to the frequency of the dominant formant or perceptually prominent spectral resonance of the speaker's vocal tract. Crude 'Single-equivalent formant' tracking, which estimates the perceptually prominent formant, has been done by simply extracting the duration of the first half-cycle of the waveform after the glottal excitation (Focht, 1967).

Counting the number of times the signal goes through any specific signal level (in either the positive or negative direction; or both) can be a measure of repetitiveness. For example, many research efforts and a few commercial products have explored, the use of the number of zero crossings per unit time. The reciprocal of the time interval between two successive zero crossings has been called the instantaneous

frequency (Baker, 1975). Zero crossing counts and instantaneous frequencies are high for noise like sounds like fricative low for vowel-like periodic sounds.

Another measure of the speech signal, frequently used in recognisers, is the intensity, or energy of the wave, which can be computed as the sum of the squares of the values of the wave at each point in time, within some window of time.

Most speech analyses has been done in the frequency domain, with techniques such as filter banks, further analysis (especially the Fast Fourier Transform), linear predictive coefficient (LPC) analyses, and cepstrum analysis. In the frequency domain, voice fundamental frequency can be found from the spacing between harmonics, or from cepstral analysis, which separates the harmonic activity from the more gradual changes in spectra due to vocal tract resonances.

Among the other possible spectral parameters, band limited energy contours, such as the 'sonorant energy' in the frequencies from 60 to 3000 Hz, or the 'voicing energy' in the low frequencies from 60 to 450 Hz, or the high frequency 'sibilant' energy from 3000 to 5000 Hz, can be useful. Tracking natural resonances of the vocal tract, or formants, is difficult to do reliably from the complex FFT spectrum or the output levels from a bank of narrowband filters distributed across the spectrum. However, the smoothed frequency spectrum that results from LPC analysis permits

formant tracking to be done with some reliability, using simple peak picking on the LPC spectrum, or pole tracking on the actual LPC model. It is generally acknowledged that if one can derive accurate formant frequency tracks vs. time, they can be valuable for the phonetic content of speech.

A few studies (e.g., Davis and Mermelstein, 1978) suggest that 'mel scale cepstral coefficients' can be at least as effective as LPC coefficients or other spectral parameters in determining the phonetic content of speech. Experiments have shown fairly comparable performance in word matching with either LPC coefficients, filter bank outputs, or FFT outputs (c.f. White and Neely, 1976; Wholford, Smith, and Sambur, 1980; Doutrich et al., 1983). Less effective were zero crossings counts, formant amplitudes and formant bandwidths.

A survey of experts with an average of ten years experience in speech recognition (Lea and Shoup, 1979, 1980) indicated that among the most preferred acoustic parameters were the formants (derived from LPC spectra, for example), fundamental frequency, LPC coefficients, energy measures, and poles of the LPC spectrum. This author favours LPC -derived formants, fundamental frequency (usually derived from auto-correlation analysis), sonorant (60-3000 Hz) energy contour, very low frequency (60-450 Hz) energy, high frequency (3000-5000 Hz) energy, a two-pole (Primary spectral peak) analyser,

a spectral derivative (monitoring how much the spectrum changes from one 10 ms frame to the next), detection of wide band nasal resonances, and mel scale cepstral coefficients.

To answer the question regarding which acoustic parameters to use in a speech recognition system is to try out all conceivable parameters, and mathematically determine which parameters account for the largest portions of the variance in large samples of speech. Several studies of principal component analysis, or eigenvector analysis, have demonstrated that the energy in the signal, the balance of energy between high vs. low frequencies, and the energies and resonances in the second, first and third formant regions are among the consistently important parameters.

Developers of new recognition systems will need good microphones, tape recorders, audio cables, earphone, speakers audio amplifiers, and other acoustic signal processing and computer equipment. An excellent way to get started quickly on parameter extraction is to purchase standard data acquisition equipment (such as the Digital sound cooperation A/D conversion system or the Data Translation or Analog Devices A/D boards), and standard software packages, such as the interactive Laboratory system (ILS) from signal Technology incorporated, which includes previously programmed procedures for deriving many of the parameters and which also permits principal component analyses and easy experimental comparisons of alternative parameters for specific tasks.

Robust Categorical features:

Voicing can be detected either by (1) observing that the very low frequency energy exceeds a threshold; (2) noting a high value of ratio LF/HF between the low frequency (60-900 Hz) energy and the high frequency (3000-5000 Hz) energy; (3) detecting small numbers of zero crossings per unit time; or (4) noting whether an Fo value is detected for the local-region.

Syllabic nuclei:

Syllabic nuclei are detectable from peaks in the sonorant energy from bounded by significant dips (such as 4 or 5 dB dips) (Lea, 1974, 1976, 1980, 1986 d). The boundaries of each nucleus may be fairly reliably to be at the points where energy drops down to half of the total amount of dip at that syllable boundary. A slight refinement could involve replacing the sonorant energy function by a "perceived loudness function", which is spectrally, weighted to be large in vowels and other parts of the syllabic nucleus.

Algorithms exist for accurate detections of syllabic nuclei (Lea, 1974; Mermelstein, 1975), and their performances indicate that this is one of the most reliable decisions that can be made in recognition. When syllables are bounded by sonorant consonants, not obstruents, the energy dips are not sufficient to be always reliably detected, but Lea (1976) still obtained over 90% correct syllable nucleus detection in the difficult case of all sonorant sentences.

Sibilant detection is possible with either the high frequency (3000 to 5000 Hz), energy function, or preferably, the ratio of the low frequency (60-900 Hz) to the high frequency (3000-5000 Hz) energy. If this ratio is below a threshold value, the spectrum is dominated by high frequencies, so that a nosy, intense fricative (specifically, sibilant) is present. (Many axis crossings per unit time is also a cue to sibilants).

Sibilants are another one of the most reliably detected categories of speech sounds, with well over 90% usually correctly detected (Medress, 1980).

Retroflexive detection is another reliable analysis procedure. Retroflexive (/r/ & /ʒ/) are detectable from a low value of the third formant F3 (below a threshold of about 1750 Hz, and possible as low as 1600 Hz), plus a pattern of F2 and F3 being very close (i.e. $F3 - F2$ is small).

Nasal stops are associated with broad form out bandwidths, very low (250Hz) formant F1 and dominant very low frequency energy, and abrupt spectral changes in formant patterns, as new formant positions occur due to the nasal tract resonances. Nasals are weaker than vowels, and they show antiresonances, or spectral dips. F2 is weak or absent. The low energy around 800 Hz (due to an antiresonance) and weaken high frequencies help distinguish nasals from /l/s.

Vowel detection is also reliably accomplished, based on the high (sonorant) energy region of each syllabic nucleus and the presence of voicing, and procedures for stripping away the non-vowel parts of the nucleus. Other primary cues to vowels are the prominent formant structures and the major energy concentration at low frequencies.

Algorithms can be designed to search directly for a match to the specific vowel sound patterns, such as the formant structures for /i/, /a/, /u/, etc. Throughout the history of speech recognition, vowel identification has been generally successful, particularly for simple syllable structures.

Stop consonants are another general class of sounds which have often been included in preliminary classification procedures, with considerable success. Stops are evidenced by gaps, or clear steady states of low energy (either silences or low energy voicing bars); large values of the spectral derivative, or extensive frequencies spectrum change from one time frame to the next, at the opening of closure; bursts of noisy, broadband energy of short duration and aspiration, or frictional sound following the gap and burst, of moderate duration (50 ms), for unvoiced stops.

These general categorical decisions leave one or more 'left over categories' to account for weak fricatives /f,,v,á/, /l/,glides /w,y/,and possible flaps latter parts

of diphthongs, transitional areas, and other areas that cannot be readily classified as sibilants, retroflexives, nasals, vowels or stops.

Detailed Phonetic Decisions:

Given preliminary decisions, a recogniser can attempt to narrow down the alternatives by attempting vowel identification, specific determination of diphthongs, detection of laterals, glides, affricates, weak fricatives, stop identification, and nasal identification. Effects of context can be taken into account, and phonological rules applied to compile a pronunciation that can be matched to expected pronunciations of words in the vocabulary.

Vowel identification is a rather well developed aspect of recognition. Vowels can be identified by their formant positions. A simple identifier could match incoming formant values stored expectations about formant values for various vowels, where stored templates were obtained from standard values.

A few other parameters help establish vowel identities. The 'spectral balance' is a general shape variable, indicating whether more energy is at low frequencies (under 1000 Hz), which is true for back vowels, or if higher frequencies (above 1200 Hz or more, as for front vowels) have a higher energy than for neutral vowels. 'Roundedness' of

vowels (as for /u, o, /) , is characterized by lower than usual values of the sum $F1 + F2 + F3$ (overall effect is lower frequency concentration).

The prosodic features of vowel energy and duration (and even fundamental frequency, F_0) also provide confirming cues to vowel identity. High vowels have low energy, short duration and high F_0 ; low vowels are the opposite.

Diphthong detection represent another refined categorical decision that might be attempted. Diphthongs are characterized by long durations, and may be confused with vowel plus glide, or vowel plus liquid, combinations. A prominent characteristic is the smooth changing formant pattern which is strictly dictated by the diphthong identity, laterals are fairly difficult to detect. They have $F1$ and $F2$ low ($F1$ lower than for vowels, usually) and are /o/ -like in spectrographic appearance, but they are weaker than /o/ usually and they appear on the edges of detected syllable nuclei. They have an extra formant at high frequencies, and they show discontinuity with neighbouring vowels.

Glides are also difficult to detect. The glide /y/ is /i/ - like, but shows more rapid transitions. A little dip in $F3$ often is evident in $F3$ contours during /y/s. The glide /w/ is characterised by a low $F2$, and major transitions into or out of low $F2$ condition.

Affricates are combinations of stop characteristics followed by fricative characteristics. They can be confused either with stop plus fricative combinations or with aspirated stops.

Fricatives, in general, have a broad band spectrum for which the high frequency energy is as much as more than the low frequency energy. They are noisy, and can be detected by from many zero crossings, also. In general, it is the overall spectrum of noise for a fricative that suggest its identity.

Stop identification involves voicing detection and a decision about place of articulation. In addition to usual energy based voicing decisions throughout the stop, other stop voicing cues include whether there is aspiration, and whether there is a delay after opening of the consonant closure before the formant structure becomes apparent (voice onset time).

The alveolar flap [r] is like a [t] or [d], but reduced in intensity, short in duration and without aspiration.

Nasals [m]; [n] and [ŋ] are distinguishable by their spectral peaks during closure, with [m]'s peak at about 1300 Hz, while the peak for [n] is at about 1800 Hz, and that for [ŋ] about 2000 Hz. Transitions into and out of nasals also show the spectral characteristics of the corresponding labial, alveolar and velar oral stops.

Phonological Analysis:

The various acoustic parameter extractors and phonetic category detectors provide the information needed to specify the sound structures of utterances, but procedures are needed for combining all these decisions into a specification of the phonetic sequence of the utterance. A strategy is needed for combining all the information so it is suitable for comparing with expected pronunciations, to decide what message was said.

Following are a few distinctive strategies for phonologic sequence matching;

1) Centisecond Labelling:- Phonetically classify each 10 ms time frame of the speech, based on its own internal features and how they match those expected for each phoneme, then use that long sequence of classifications to match to expected pronunciations.

2) Separate segmentation and- labelling: Segment at major acoustic boundaries, defined by major variations in robust important features, and then select the best phonetic label for each of those acoustic segments.

3. Phonetic detection without boundaries: Detect the presence of phonetic categories and specific phone identities based on the previously discussed manner and place characteristics, and thus specify central positions of those

apparent occurrences as detected phonetic units, but do not specify segment boundaries (and then attempt to match the order of detected units with the expected order of phonetic units for various words).

4) Phonetic Lattice: Detect, and find the beginning and ending (boundaries) of phonetic units, based on the manner and place characteristics, and allow alternative choices as to the identify of various regions as well as alternatives as to where units begin and end, to yield a phonetic 'lattice' of overlapping alternative sound sequences which can be matched to expected sequences.

5) Strict phonetic segmentation and labelling: Find (i.e., detect and specify boundaries of) a strict sequence of phonetic units (either sub-phonemic segments, or phonemes, or tranemes, etc.) that completely cover the utterance without overlapping.

The usual goal sought in recognition has been either strict phonetic segmentation and labelling or else separate acoustic segmentation and subsequent first choice phonetic labelling. Studies, as in the ARPA SUR project, have allowed several alternative choices for labelling each segment, where labels have been based on the most likely or 'closest' phonetic categories. Thus, probability vectors or preference lists are composed for each segment, indicating the first

choice label for the segment, the second choice label, the third etc. order of choices is based on the closeness to the acoustic characteristics of the analyzed segment.

Segmentation and labelling must include scoring procedures, for assessing how sure the analysis is about the various claimed segments and their identities. A measure of likelihood of correctness or error might be used for scoring.

Recognizers must consider whether to retain knowledge of only the best scoring phonetic unit for each position of speech, or whether to retain a vector of scores of various phonetic labels, and thus to give a priority list of most likely labels for each of the segments. The same issue arises at the word level; only the most likely (highest scoring) word should be chosen or a list of possible words and their relative scores should be retained for each region of the incoming speech. These are typical problems in the search mechanisms of AI systems, regarding the merit of 'best-first' analysis, vs. 'breadth-first' analysis, vs a compromise like the 'best-few-first' analysis and Lowerence, 1980; Wolf and Woods, 1980).

One other type of phonological information might help in identifying words in speech. That is the so-called 'phonotactic information' or language dictated constraints on allowable sound sequences of the language.

Lexical Constraints:

Zue and his colleagues (1982), showed that recognition with a large (20000 word) vocabulary can be reduced to a selection among just a few confusable words, once the major phonetic categories are established.

Waibel (1982), showed that prosodic features could also drastically reduce the set of candidate words from a large vocabulary. The stress pattern of the word, the duration of the syllables, and the portions of each syllable that is voiced can cut the set of candidate words down to about 1.6% of the vocabulary. Coupling such prosodic information with crude phonetic decisions can leave only a few words as candidates, on the average.

Prosodic Aids:

Prosodic information provides acoustic cues to more than just the wording of an utterance. Indeed, the primary contributions from prosodies may prove to be in aiding syntactic parsing and guiding phonetic analyses.

Based on linguistic and psychological arguments that syntax is used in the early stages of speech perception, Lea (1973, 1974, 1980 C, 1986d) has suggested novel theory of speech recognition, in which early use is made of prosodic cues to syntactic structures, and within that structure, analysis is focused on important (stressed) words and islands of phonetic reliability.

An extensive series of experiments showed other benefits offered by prosodic features. Lea showed that interstress intervals were the best indicator of rate of speaking, and could be used to select suitable (e.g., fast speech is slow speech) phonological rules. Unusually long interstress intervals were cues to major phonological phrase boundaries. Intonation contours and stress patterns could indicate sentence type, subordination of phrases under other, and special grammatical structures, like conjuncts with word repetition. Detailed acoustic phonetic analyses could be more efficiently and accurately done when guided by prosodic cues such as syllabic nucleus locations and stress determinations (Lea and Clermont, 1984).

SYNTACTIC CONSTRAINTS:

Most recognition systems have focused entirely on using the grammaticality constraints to 'weed out' alternative word sequences, without offering any abilities in phrasal grouping, labelling and definition of grammatical relations. This filtering of word sequences to establish which ones satisfy grammatical rules can be a major factor in assuring correct sentence understanding and reducing computations for alternative word sequence.

There are many types of grammars, of varying powers to generate complex languages (Lea, 1966; Chomsky and Miller,

1963) but perhaps the most important to speech recognition are finite state, context free, context sensitive and augmented transition network grammars.

Finite state grammars are at the forefront of current capabilities in highly reliable speech recognition. Harpy (Reddy and Lawrence, 1980), the IBM system (Jelnek, 1982), all commercial recognizers and dynamic programming systems have endeavoured to recognise with this restrictive form of grammar, and systems like the BBN HWIN system (Wolf and Woods, 1980) that have tried to go beyond the limitations of finite state grammars have had comparatively limited success. A finite state grammar is equivalent to a 'finite state automation' or 'Markov model'. In which generation (or recognition) of the next word in a sentence is determined by a fixed memory of the previous n words (where n is frequently only one, so the immediately previous word restricts the allowable next word).

Linguists (e.g., Chomsky, 1957) have shown that finite state grammars cannot properly characterise major subsets of English sentences if no fixed limit is placed on the complexity of sentences. Thus, finite state grammars cannot generate (or recognise) all such English sentences and only the acceptable sentences. Context free grammars have been devised to permit more generative power, in which sentences

need not be generated a single word at a time, but large units can be divided into phrasal sub-units, which in turn get expanded until the smallest units are represented by words of the acceptable vocabulary.

However, even such context free grammars cannot capture some of the contextual constraints that seem to be involved in aspects of the English language, again assuming no fixed limit on sentence complexity. Transformational grammars (Chomsky, 1957, 1965) were devised to account systematically for complex contextual effects and total derivational histories of sentences type (such as passive vs. active sentences, etc.).

However, transformational grammars have proven difficult to use in recognition procedures, so the 'augmented transition network' (or ATN) grammar has been devised as a practical substitute, of equally general power.

Thus, there is a hierarchy of ever more powerful grammars, ranging from finite-state grammars to upto ATN grammars. The more powerful the grammar, the versatile the language that can be characterised. More importantly, however, for the current uses of syntax in recognition, the more restrictive the grammar the better it is for strictly limiting the acceptable word sequence.

As important issue in syntactic analysis for speech recognition concerns the assessment of the complexity of a language for speech interaction with machines. Goodman (1976) developed the idea of an average 'branching factor', which indicates the average number of words that can appear next in a sentence of the voice i/p language. The higher the branching factor, the more difficult the recognition task, though this is hardly a fully adequate measure.

Semantic and Pragmatic Constraints:

Semantic networks can be used to show semantic relations between words, objects that can be 'contained in' other objects may be connected in a semantic network. In early works on speech understanding system, semantic networks were expected to play an important independent role in determining the correct word sequences to hypothesize in a system, and which hypothesisable word sequence should be ruled out due to their semantic anomalies (Nash-Webber, 1975).

Pragmatic information may be used in speech recognisers to verify or rule out hypothesized word combinations by establishing their agreement with prior discourse or their applicability to the task being undertaken during the human-machine interaction. Knowledge of previous discourse can help and can permit the full expansion of elliptical (truncated) utterances that follow similar utterances.

COMPUTATIONAL TECHNIQUES

The Principles of Speech Pattern Matching:

According to Moore (1985) It is widely acknowledged that it may be many years before the techniques of automatic speech recognition (ASR) are able to challenge the accuracy and reliability of normal human speech perception; the development of a machine with the ability to transcribe accurately any spoken message from a wide range of talkers under less than optimal environmental conditions.

However, since the early 1970's an approach to ASR has been evolving which, although rather superficial in appearance, is nevertheless achieving a modest amount of success, both from a scientific and from a commercial point of view. This approach has become known as speech pattern matching (SPM).

Like most other approaches to automatic speech recognition, SPM is based on the premise that for a machine to be able to recognise speech, it must have access to knowledge about speech and about how words and sounds manifest themselves in acoustic signals. It also requires that this knowledge should be structured and manipulated in appropriate ways. However, SPM differs from other approaches in that it attempts to minimise the amount of heuristic a

prior information about these structures and manipulations by capitalising on the fact that a prime source of potentiality reliable speech knowledge is the information contained in actual speech patterns.

In SPM, first a speech signal undergoes some kind of pre-processing which transforms the acoustic waveform into a sequence of analysis vectors. Then, during an initial training phase, example patterns are used to generate suitable models which are subsequently stored in the model store. The complexity of these models varies; in the simplest case a model might just be an example of a particular word, on the other hand more advanced schemes use relatively sophisticated statistical models. Finally, in the recognition phase, an unknown utterance is compared with the models in the model store and is assigned to the class of the model with which it is (in same sense) most similar.

Pre-Processing:

There are two main reasons why it is necessary to pre-process speech signals in advance of the pattern matching stages. First, it is desirable to transform the audio waveform into a domain where the patterning of speech is more explicit. Second, the data-rate in the transformed signal may be too high for the subsequent stages. Hence a large range of signal processing techniques are applicable to speech signals, both for achieving a suitable signal representation and for reducing the data-rate.

A. SIGNAL REPRESENTATION:

1. Fourier Analysis:

The most common and perhaps most informative way to analyse a speech signal is to estimate its short-time power spectrum using the Fourier transform. With a suitable choice of analysis window, the harmonic structure of the excitation function (voice pitch) may be ignored and the resulting wide-band envelope spectrum contains information which derives mainly from the shape of the vocal tract.

This process may be done using discrete Fourier transform (DFT) or, more easily, using a bank of analogue band-pass filters. The latter has the advantage that the distribution of frequency bands (channels) may be readily modelled on the critical bands of the human ear.

2. Cepstral Analysis:

Direct Fourier analysis requires a short data window in order to ignore the harmonic structure of speech signals. However, an alternative approach which is can use a wider time window is homomorphic or cepstral processing. Essentially, a narrow band spectrum (which contains the harmonic structure) is further transformed, using Fourier analysis, into the cepstral domain (the spectrum of the spectrum) where the components due to the pitch of the value may be filtered out, and then transformed back to obtain a smooth envelope spectrum.

3. Linear Predictive Analysis (LPC):

Linear predictive coding is a speech analysis technique which is particularly attractive from the computational point of view. In this scheme the autocorrelation characteristics of the speech waveform are exploited by estimating the value of the current sample using a linear combination of the previous n samples. The result is an analysis which is based on an all-pole model of the vocal tract. This means that LPC is particularly good at estimating the positions of spectral peaks during vowel sounds. However, during speech sounds which do not conform to an all - pole model (nasals and many consonants) LPC tends to over estimate the bandwidths of the peaks.

B. DATA REDUCTION:

The result of the initial transformation of a speech waveform is thus typically a regular sequence of analysis vectors, where each vector describes the distribution of energy at different frequencies (or the configuration of the vocal tract) at different times during an utterance. Further processing is then able to reduce the data-rate by capitalising on the redundancy in the transformed signal.

1. Vector Quantisation:

It is possible to reduce the data-rate of a speech signal by vector quantisation (VQ). VQ is a technique whereby

each frame of spectral data is coded by comparing it with a pre-stored set of reference frames, each of which is associated with a different output symbol.

The set of reference frames (or codebook) is normally generated by a clustering procedure which minimises the average distortion resulting from coding a suitably long sequence of vectors. However, care is necessary in the use of VQ in automatic speech recognition, since the set of reference frames must be large enough to preserve the smallest distinction required by the subsequent pattern matching algorithms.

Data adaptive Coding:

As VQ reduces the amount of information associated with a single frame of speech data, variable rate coding techniques may be used to exploit sequential properties. Such schemes are based on the observation that the changing pattern of sound in a speech signal gives rise to analysis vectors (e.g. spectra) which are often fairly similar over several frames. In these circumstances it is possible to re-sample the analysis vectors at a rate dictated by the amount of change in the signal, thereby reducing the overall frame-rate.

For a continuous signal, the simple scheme is to set a threshold such that an analysis vector is produced only if

the similarity between it and the previous vector exceeds the threshold; the higher the threshold, the fewer samples will be taken in the more stationary regions of the signal. In the special case where the first and last frames of data are known, it is possible to set the threshold such that a fixed number of frames are retained; this particular method is known as trace segmentations.

Modelling Speech Patterns:

The principle of speech pattern matching is that a prior knowledge (i.e., in sight and assumptions) about the structure of speech (e.g., words) is supplemented with analytical knowledge (examples of actual speech patterns) in order to construct models against which unknown patterns may be compared for recognition.

The quality of both a prior and analytical knowledge is obviously of paramount importance; assumptions about suitable model structures must not be too far from reality, and the example speech patterns must be reasonably representative. The key to successful modelling is to maximise the use of the actual measured data by sharing information (pooling) where possible, whilst at the same time minimising the loss of information by inappropriate pooling. However, by far the most important concept is that, no matter how well the derived data structures model the example patterns, it must be possible to generalise that information in appropriate

ways in order to accomodate the inherent variability of speech and to be able to correctly recognise speech patterns that have not been observed previously. This means that the models can be regarded as having the capability of generating (synthesizing) a range of patterns conditioned by the structure of the model and by the examples that have actually been observed.

Models:

Different types of models are used for speech recognition. They are; simple whole-word models; Stochastic models; markov models; Hidden markov models; Semi-Markov models; Sub-word models.

Other types of recognition procedures based on distance measures for speech processing are: The root mean square (rms) log spectral distance, cepstral distance, likelihood ratio (minimum residual principle or delta coding (DELCO) algorithm and a cash measure (based upon two nonsymmetrical likelihood ratios).

The properties and interrelationships among four measures of distance in speech processing are theoretically and experimentally studied by Gray and Markel (1976). It has been shown that the cepstral measures bounds the rms log spectral measures from below, while the Cosh measures bounds it from above. A simple non-linear transformation of the likelihood ration has been shown to be highly correlated with rms log spectral measures over expected ranges.

According to Gray and Markel (1976); for the purpose of knowing their interrelationships the rms log spectral distance was taken as a reference. It has been found that Cosh measure satisfy all of a specified set of distance measure criteria. Based upon the heavier weighting of large differences, it has been seen that Cosh measure is a better choice where large differences are expected.

Recently time-domain speech analyses based on linear predictability of signal waveform has been successfully adopted for efficient coding of a redundant speech signal. Several efforts have been made toward application of the linear predictor coefficients (LPC) for speech recognition. (Itakura and Saito; 1968, 1970). Here, in this procedure, a reference pattern for each word to be recognized is stored as a time pattern of linear prediction Coefficients (LPC). The total log prediction residual of an input signal is minimized by optimally registering the reference onto the input auto correlation coefficients using the dynamic programming algorithm (DP). The input signal is recognized as the reference word which produces the minimum prediction residual. Sequential decision procedure is used to reduce the amount of computation in DP. A frequency normalization with respect to the log-time spectral distribution is used to reduce effects of variations in the frequency response of telephone connection.

The system has been implemented on a DDP-516 computer for the word recognition - experiment. The recognition rate for a designated male talker is 97.3 percent for telephone input, and the recognition time is about 22 times real time (Itakura; 1975).

STUDIES ON AUTOMATIC SPEECH RECOGNITION

Earlier Studies:

The first truly successful recognizer was reported in 1952 by Davis, Biddulph, and Balashek of Bell Laboratories. This device could recognize ten digits, spoken over telephone by a single talker, with an accuracy of 100%. On the speech of a different talker, however, the accuracy could be as low as 50%. The talker was expected to speak clearly and to pause between digits. The recognition method used was to assign the spoken word to the most probable digit category on the basis of appropriate F1/F2 measurements during vowel sounds.

Wiren and Stubbs (1956) produced a device which partially implemented the distinctive feature binary oppositions. Phoneme classification was based on voice/unvoiced, turbulent/nonturbulent, stop/fricative, and acute/grave determinations. This gave 94% correct recognition of the vowels in short words by 21 talkers. Also in 1956, Olsen and Belar of RCA reported a machine with a vocabulary of ten monosyllabic words with which a syllable

recognition accuracy of 98% could be obtained in complete sentences by a single talker. "Careful" pronunciation was required with a pause between each syllable. Eight frequency bands and five time intervals per syllable were used to define an eight-cell by five-cell matrix. In each cell a "1" or a "0" was stored depending on whether the signal energy appropriate to that cell was above or below a threshold level. The decoded matrix pattern was used to operate a typewriter key and so a typed transcript of the sentence was obtained.

A later version of the Olson and Belar syllable recognizer was reported in 1961. Again this made use of eight frequency bands and five time samples, but these time samples were only taken when a significant change occurred in the spectral power distribution. The machine's vocabulary was increased from 10 to 100 syllables, but no recognition performance figures were given.

Following suggestion by Fry, Denes (1959) produced a speech recognizer. This consisted of a spectrum analyzer, a spectral pattern -matching system and stored probabilistic information concerning the probability of any phoneme recognized by the machine following another in spoken English. The phoneme-recognition set comprised four vowels and nine consonants. The overall performance was not

particularly good, but use of the linguistic data (phoneme pair probabilities) improved the word recognition accuracy from 24 to 44%.

Forgie and Forgies (1959), used a 35 - channel filter bank whose outputs were envelope-detected, sampled, and fed to a computer. The computer program made use mainly of the frequency positions of F1 and F2 and of fundamental voice frequency measurements. It could recognize ten English vowels in isolated words of the form /b/ -vowel/t/. For 21 male and female talkers, with no adjustment for the talker, the vowel recognition accuracy was 93%.

This work demonstrated the value of the digital computer in recognition studies.

The availability of powerful, high-speed computers in recent years has brought about a significant change in research in this field.

One common use for a computer has been in place of the hardware concerned with classifying the utterance on the basis of output signals from a hardware spectrum analyzer. The spectrum analyzer is typically of the vocoder type, having a bank of filters followed by rectifying and smoothing circuits. Spectral analyses of speech in this manner is well established and many such analysers exists in speech - research centers.

Where the spectral analysts section is also subject to investigation there are considerable advantages in simulating the whole system by computer. However, where real-time operation is required, or where large amount of speech material are to be processed, it has been common to fix the design during an initial period of complete simulation and then to build hardware units for those sections requiring long computer time. Now even this restriction has been greatly released with the development of the 'fast Fourier Transform' algorithm of Cooley and Tukey (1972). Thus in many present day laboratory investigations valuable research can be conducted with no further equipment than a tape recorder, an analog-to-digital converter, and a general-purpose computer.

The most interesting use to which computers have been put in some recently reported investigations has been in on-line computer studies using man-machine communication peripherals. These highly flexible systems allow for rapid manipulation of the speech signal in intensive experimental sessions, which permit progress at a high speed.

Hardware studies:

Suzuki and Nakata (1961) studied vowel recognition using a 26 channel spectrum analyzer and separate wideband, formant - related channels. The channel outputs were separately grouped and connected to individual vowel - decision

circuits. The speech was segmented into voiced and unvoiced segments, and envelope-intensity and fundamental voice-frequency measurements were also used. To reduce the errors due to formant movement during vowel sounds, separate recognition decisions were made at intervals throughout voiced segments. The final classification was then made by observing the phoneme most frequently recognized. Sakai and Doshita (1962) reported a most comprehensive recognizer. This used a separate circuits for segmenting the speech into vowels and consonants and for classifying the segmented phonemes. Zero-crossing analysis was combined with measurements of variation of energy in various frequency regions. The segmentation operation made use of measures of the "stability" of, and the "distance" between the digital patterns generated. 90% correct recognition was obtained on vowels and 70% on consonants.

A hardware study of massive proportions has been the subject of periodic - reports by Martin et al (74). This system makes use of Analog Threshold Logic (ATL) elements which are based on a model of the biological neuron. The ATL element is a transistor circuit having excitatory and inhibitory inputs which provides an output only when the linear sum of input currents (inhibitory inputs are subtracted) exceeds a Pre-set threshold. This output is then proportional to the sum of inputs until saturation is

reached. When reported in 1964 the system contained well over 500 of these circuits. The recognizer employed a 19 - channel spectrum analyzer, ATL elements, AND gates, and monostable multivibrators. Various features of the speech signal were used for recognition of individual phoneme pattern, such as steady-rate and transitory spectral-energy patterns and intensity ratios.

Falter (1965), measured machine's performance on CVC utterances by six male talkers. Recognition accuracies ranging from 82 to 99 percent were obtained for 22 different phonemes including vowels, stops, and vowel-like consonants.

Gazdag (1966), used a 12 channel spectrum analyzer (highest frequency 3 KHz), and the measurement space defined by its outputs was partitioned by six hyperplanes, implemented by summing amplifiers and trigger circuits, each trigger circuit changing state as its amplifier 's output passed through zero. The output pattern was thus in the form of six binary quantities whose time variation was characteristic of the spoken word. The outputs were recorded on a multichannel pen recorder and transcribed by human operators in sequences of six-digit binary numbers.

The machine was tested on 10 digits spoken by four talkers. Detailed results were not given but the performance was described as "promising". It was also claimed that the system was invariant for time variations but no proof of this was given.

Gilli and Meo (1967) described a system for recognizing Italian numbers - A 17 - channel spectrum analyzer was used which was followed by threshold detectors which reduced the range of outputs from each channel to binary statements. These signals were sampled and connected to separate circuits for the recognition of each digit. The decisions were mainly based on the sequential occurrences of vowel and consonant patterns and patterns due to certain transitions.

For the 10 numerals spoken by 10 talkers, which were used to provide data on which the machine was designed, no errors occurred. For new utterances by the same talkers, correct recognition was obtained for 90% of the utterances. The authors claimed that the results proved that the crude patterns obtained were sufficient for recognition of the digits by a simple machine and that a redesign could have eliminated most of the errors.

Ross (1970) used a four-channel spectrum analyzer and a digital disk storage system. A 20-bit binary pattern was generated for each word by sampling five two-state signals derived from four frequency channels. Four of these signals were obtained by calculating the ratio of the mean energy level in each frequency channel to the mean energy in the overall signal. The fifth signal resulted from a comparison of the energy in the highest frequency channel with that in the two lowest. Each of these signals were compared with a

threshold level to produce five binary outputs. The resulting patterns were classified by means of a "nearest neighbor" comparison with patterns stored in the machine. The permissible "distance" between patterns giving the same output response was under the control of the experiments. The training procedure was to present a number of utterances representing a single word in sequence. If the incoming pattern was not sufficiently close to any of the patterns already stored, the new pattern was added to the stored patterns. Thus a set of different patterns were retained for each word in the machine's vocabulary.

After training a new input pattern was given the same label as that of the nearest stored pattern if it was within the specified distance of one of these. Otherwise it was rejected. The machine was tested on spoken digits.

When the permitted distance was set to one bit (i.e., two patterns were considered the same if they differed by only one bit out of the 20 pattern bits) 86 stores patterns were required to correctly identify the members of the training set. For ten new samples of each spoken digit, 56 gave the correct response, 8 gave incorrect responses, and 36 gave no response. When the permitted distance was zero (requiring an exact match), with 218 patterns stored, 46 out of 100 new digits gave the correct response, 2 gave incorrect responses and 52 gave no response.

COMPUTER STUDIES

Denes and Mathews (1960) reported a system for recognizing the 10 digits from complete word patterns. This used a 17-channel spectrum analyzer, the outputs of which were sampled and recorded on magnetic tape. This signal then formed the input to a computer. Time-frequency patterns for a number of utterances of each word were averaged and stored as reference patterns. Patterns from unknown utterances to be classified were compared, by a cross-correlation process, with each stored pattern in turn. The classification of the reference pattern giving the best match was chosen as that of the unknown utterance. A comparison was made of the results with and without time normalization of the patterns.

One female and six male talkers were used and the only restriction imposed on their pronunciation of the digits was that they should pause between each. The error rate averaged over all talkers was 6% with time normalization and 12% without formation of reference patterns by averaging over several talkers and the use of time normalization were both valuable techniques.

Sebestyen (1960) devised a digit recognizer which had an 18 channel vocoder analyser and computer. The computer implemented linear transformations of the analyzer output signals to obtain maximal clustering of the data in the 361-

dimensional space formed by time -sampling the channel outputs. Time normalization was used and the normalization factor was included as one of the space dimensions.

Four hundred utterances of the 10 numerals, by 10 talkers were used. it was claimed that, if atleast seven examples of each numerals were used to optimize the clustering transformations, no recognition errors were made. When the talkers used in testing the recognizer were different from those used in training it.

Gold (1966); reported a word - recognition system. This used a 16-channel spectrum analyzer (maximum frequency 3 KHz), a pitch extractor, and a voicing detector. The computer program segmented the words into approximate phonemic units on the basis of voicing, magnitude, and spectral information. A further 15 spectral, duration, and intensity measurements were then made on a group of five segments centered on one segment which was defined as stressed. A scoring method based on the similarity with previous measurements on known words was used to classify the word. For 50 isolated words by 10 talkers, 86% were correctly identified and 96% received either the highest or second-highest score.

A word recogniser using a low-data-rate spectral-matching process has been developed by Shearme and Heach. The output of a 20-channel vocoder analyzer is sampled and

fed to a computer. Each time sample is considered to define a point in 20 dimensional space. The variation of this point traces out a path in the space which is characteristic of the word and of the talker. Twenty different "proto type spectra" are chosen which represent certain fixed points in the space. The path taken by the signal during an utterance is specified by noting the nearest prototype point at each sampling instant. The result is a series of numbers in the range 1 to 20. It is this sequence which forms the pattern to be recognized in the matching process. Separate number sequences due to the utterance of words of known classification are first stored in the computer as "template patterns". Unknown-word patterns are then compared with each prototype pattern and the classification of that giving the highest score is chosen.

For 32 isolated words by 10 talkers and using nine template patterns for each vocabulary word, an accuracy of 90% correct identification was obtained. Purton (1970) has produced system which makes use of an autocorrelation technique. The signal is first split into two bands approximating the ranges of F1 and F2. The two waveforms are then infinitely clipped to produce zero-crossing signals. These signals are separately auto correlated and time sampled to produce a 36 X 30 matrix for each utterance. This pattern is compared with master patterns and the best match

determined by a scoring method. One master patterns once formed from several utterances of known classification by one or by several talkers.

In single-talker experiments using a machine vocabulary of 10 words, 696 utterances gave individual accuracies ranging from 78 to 99%. The average accuracy was 89%.

Louis (1971); proposed a new real time word recognition system that uses only a small computer (8K memory) and a few analog peripherals. The essentials of the procedure are as follows. During the pronunciation of a word, a spectral analysis is carried out by a bank of $17\frac{1}{3}$ - octave bandpass filters. The outputs of the filters are logarithmically amplified and the maximal amplitude of the envelope is determined and sampled every 15 ms. In this way a word is characterized by a sequence of sample points in a 17 dimensional space. Then a principal components analysis is performed, reducing the original 17 dimensions of the space to 3. After a linear time normalization, the 3 - dimensional trace of the spoken word is compared with 20 reference traces, representing the 20 possible utterances (the digits, plus 10 computer commands). The machine responds by naming the best fitting trace. With the 20 speakers of the design set, the machine is correct 98.8% of the time.

Fumitada Itakura (1975); devised a computer system in which isolated words, spoken by a designated talker, are recognized through calculation of a minimum prediction residual. A reference pattern for each word to be recognized is stored as a time pattern of linear prediction coefficients (LPC). The total log prediction residual of an input signal is minimized by optimally registering the reference LPC onto the input autocorrelation coefficients using the dynamic programming algorithm (DP). The input signal is recognized as the reference word which produces the minimum prediction residual. A sequential decision procedure is used to reduce the amount of computation in DP. A frequency normalization with reference to the long-time spectral distribution is used to reduce the effects of variations in the frequency response of telephone connections.

The system has been implemented on a DDP - 516 computer for the 200 word recognition experiment. The recognition rate for a designated male talker is 97.3% for telephone input and the recognition time is about 22 times real time.

Sambur and Rabiner (1974); studied speaker independent digit recognition the digit classification scheme was based on segmenting the unknown word into three regions and then categorical judgements were made as to which of six broad acoustic classes each segment belongs to. The measurements made on speech wave form include energy, zero crossings.

two -pole linear predictive coding analyses and normalized error of the linear predictive coding analyses. A formal evaluation of the systems showed an error rate of 2.7% for a carefully controlled recording environment and a 5.6% error rate for on line recordings in a noisy computer room.

The experimental test of the digit recognizer was conducted in two parts. The first part consisted of 10 speakers (5 female and 5 male) each of whom made 10 complete recordings of the 10 digits. The recording sessions were spaced over a five - week period to include the effect of time variation in the testing . The recordings were made in a quiet room with a high - quality microphone. The decision algorithm was not designed for the characteristics of each particular speaker, so as to give a true test of the speakers - independent nature of the scheme.

A confusion matrix for each of the 100 tests of each digit 100 was presented. The confusion matrix indicated that all occurrence of initial friction were correctly detected by the decision algorithm. In only 6 out of 200 examples of the digits 1 and 9 was the initial nasal like consonant incorrectly determined. The confusion matrix also showed that most errors were made in the final detailed decision.

Rabiner and Sambur (1976); have done some experiment in the recognition of connected digits. The overall recognition system consists of two separate but interrelated parts. The

function of the first part of the system is to segment the digit string into the individual digits which comprise the string , the second part of the system then recognizes the individual digits based on the results of the segmentation.

The segmentation of the digits is based on a voiced or unvoiced analysis of the digit string, as well as information about the location and amplitude of minima in the energy contour of the utterance. The digit recognition strategy is similar to the algorithm used by Sambur and Rabiner (1974) : for isolated digits, but with several important modifications due to the impreciseness with which the exact digit boundaries can be located. To evaluate, the accuracy of the system in segmenting and recognizing digit strings a series of experiments was conducted. Using high -quality recordings from a sound proof booth both the segmentation accuracy was found to be about 99% and the recognition accuracy was about 91% across 10 speakers (5 male and 5 female) with recordings made in a noisy computer room the segmentation accuracy was about 87% across another group of ten speakers (5 males and 5 females).

Zuicker and Paulus (1979) ; studied speech recognition using psychoacoustic models. In which it follows the concept of 1) preprocessing in term of auditory parameters. 2) subsequent classification and recognition. The preprocessing system has been realized in analog hardware, while

recognition is carried out on a digital computer. In the processing system, the essential psychoacoustic principles of the perception of loudness, pitch, roughness and subjective duration were implemented with some approximation. The system essentially consist of 24 bandpass fitters, non linear transformation of each filter output into specific loudness, and specific roughness and final transformation of these parameters into total loudness, total roughness and three spectral momenta. As a means to further reduce the information flow, continuous selection of dominant parameters was also considered the basis of psychoacoustic data. The subsequent recognition process is mainly characterized by 1) discrimination between speech and silent periods. 2) detection of syllable peaks and classification of syllable nuclei, and 3) assumption of syllable boundaries and classification of consonant clusters. Result of the study indicated the concept provided a systematic and promising way towards automatic speech recognition of continuous speech.

Significant advances in speech recognition are likely to come not from researches into signal analysis, adaptive pattern matching, or computer implementation (although these fields have valuble techniques to offer the speech researches), but from the studies of speech perception and generation, phonetics and linguistics a much greater understanding of the whole speech process is required before

an automatic recognizer can be built whose performance will approach human ability.

As the review of literature shows that several attempts have been made to use computer for speech recognition several methods have been proposed and tried with varying success. Most commonly found methods are: Markov models; Hidden Markov models; Semi-Markov models; Cosh Measure; Minimum Prediction Residual Method and Linear Prediction Coefficients (Euclidean Distance Measures).

Thus it can be concluded that computer can be used for speech recognition at least to a limited extent. To the present investigator a program based on distance measure (Cosh Measure, Minimum Prediction Residual Method and Linear Prediction Coefficients) developed on the lines suggested by Gray and Markel (1976) was available. Since speech recognition has several applications in the field of speech diagnosis and therapy; it was felt necessary to explore program and to identify the variables related to speech recognition with the program available.

Due to several limitations this study was restricted to only Fundamental frequency as a variable.

METHODOLOGY

Speech recognition has been considered as important from various points of views. Several methods and models have been proposed and used for this purpose, for example: Simple whole word models ; Markov models ; Hidden Markov models and several distance measures are ; Cosh Measures ; minimum prediction residual principle ; linear predictive coefficient, Euclidean distance measures etc.

According to Gray and Marcel (1976) :- Minimum prediction residual method, Cosh measure and euclidean distance (LPC coefficient) measurement have been considered useful methods for speech recognition.

Application of speech recognition in the field of speech therapy has been recognizes as useful. But it has not been worked upon particularly in India. Before using these program for clinical use it was felt necessary to know the efficiency of minimum prediction residual method, Cosh measure and linear -prediction coefficients (Euclidean distance measure); which were available with this program.

The review of literature has shown that duration of utterance and fundamental frequency as important variables among the variable in the process of speech recognition.

A program* of speech recognition written, using basic language based on a description provided by Gray and Markel (1976) was available to the investigator. This program provided distances between the stored group data and test data using minimum prediction residual method, Cosh measure and LPC (Euclidean distance) measure.

The present study was limited to find out the effect of variations in fundamental frequency on speech recognition with the program that was available to the investigator.

A preliminary investigation had shown that the program ("speech recognition") was not capable of recognizing either digits or words when the duration of utterance varied.

In order to find the effect of variation in fundamental frequency on speech recognition with the present program, it was decided to make the subject utter the digits, (which were already recorded and analyzed) at habitual pitch** and pitches higher and lower than habitual pitch of the subjects. Therefore the present study was undertaken.

*The program that was used was a commercially available one. No modification or changes in the program were made by the investigator. It is beyond the limits of the present investigation to consider any of the aspects of the program except for its possible usefulness in speech recognition for clinical purposes.

**Habitual pitch is the pitch which is most frequently used by the subject. Usually for adult males it is around 125 Hz

Subjects :- Five subjects (all males) in the age range of 17 to 25 Years with a mean age of 20 Years took part in this study. The selection of subjects was done on a random basis and also on the basis of following criteria :- No history of hearing loss or vocal pathology at the time of recording.

The experiment was conducted in two phases. In the first phase the subjects were asked to utter ten digits (0 to 9) using their habitual pitch. Loudness was normal with the microphone at a distance of approximately five cms., from the mouth. The loudness of the counting was monitored so that the amplifier indicator of high voltage (about +5 volts) was not on.

Phase I Procedure

Step (i) (a) Instruments Used :- for recording of the digits from 0 to 9 for each subjects ; an external mic was used, which was connected to speech interface unit (voice and speech system) and finally the input was given to the computer (PC-XT).

Step (i) (b) Recording Material :- The digits from 0 to 9, written on flash cards (one digit written on each flash card) were used as recording material. The cards were presented randomly to the subjects and were asked to repeat them as soon as it was flashed.

Step (i) (c) Recording Environment :- The recording was carried out in a sound treated room. The ambient noise level, present in the test room was become the maximum permissible (ISO - 1964) level.

Step (iii) Instruction :- The subjects were instructed to utter the digit as soon as the cards were presented. For each subject each digit was recorded five times. the data out -put for each digit was heard through speaker connected to speech interface unit .

The instruments were arranged as shown in fig-1.

The data acquisition for each of these digits (five times each) at a sampling frequency of 8000 Hz was done. Then each digit was segmented using the program DSEGF and the segmented digital data were stored on the floppy - disk.

Each data file was normalised. All the digitized data was submitted for LPC analysis. the program ANALP was used for LPC (linear prediction coefficient) analysis.

After LPC analysis for each digit - the program RECGN (Recognition) was used. Parameter estimation of these LPC files for each subject was done using the mean mode.

Parameter estimation of each subject was followed by making group vocabulary for each subject. And this group vocabulary was taken as the reference point for recognition.

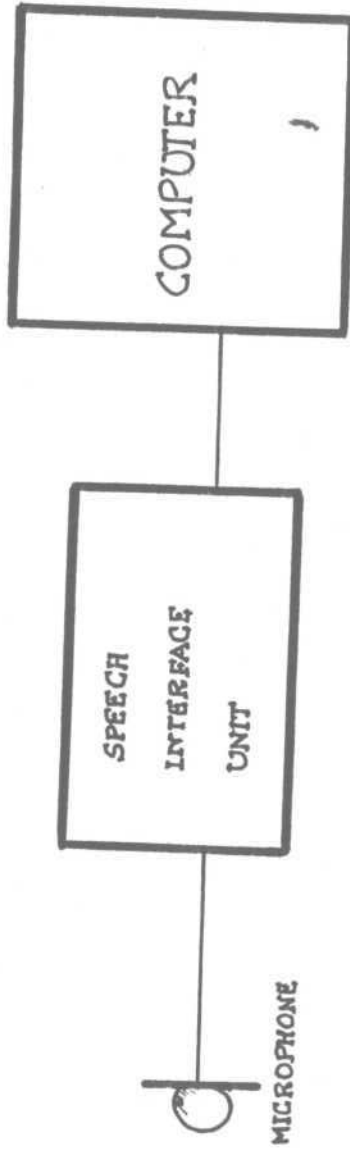


FIG:1 BLOCK DIAGRAM OF INSTRUMENT SET-UP .

Phase II

The subjects who had participated in phase I were taken this part of study also.

Step (i) (a) - same as in Phase I

Step (i) (b) - same as in Phase I

Step (i) (c) - same as in Phase I

The purpose of this phase of experiment was to find out variables which would affect the recognition of speech uttered by different speakers.

step (ii) Instruction

The subjects were instructed to utter the digits shown to them using the flash card at habitual pitch, and then at pitches higher and lower than the habitual pitch. The order of presentation of the flash cards were randomized.

Each subject was made to utter each of the ten digits five times each at three different pitch levels, as previously determined. Thus for each subject 150 recordings were made. Out of five recordings of each digit at each pitch level one which had a duration similar to the one used for group vocabulary ; for example the utterance of subject 1 -of digit say one had 255, 250, 255, 245 and 255 msec duration at habitual pitch. the utterance "two" which had 250 msec duration was similar to the duration of the utterance "two" in the group vocabulary of subject 1(as recorded in phase I).

Thus for each subject ten recordings out of 50 recordings at each pitch level were selected covering all the ten digits. i.e., a total of 30 utterance of each subject were selected.

Using this procedure 30 utterances, ten at habitual and ten each at higher and lower pitches for each of the five subjects were selected. thus a total of 150 utterances were used as test data to find out efficiency of the program to recognize at different pitch levels.

After data acquisition each of these digits were normalized using the program NORM. Again LPC analysis was done using the program ANALP as described in phase -I.

In the final stage of the experiment each digit recorded in phase II was compared (i.e., being recognized) with their respective group vocabulary using the program (i) Cosh measure (ii) minimum prediction residual method and (iii) Linear prediction coefficient (euclidean distance measure).

Each of the methods used provided the distance values for each digit with the respective group vocabulary. The nearest distance was taken as the recognized digit.

The correctly recognized digits were tabulated and the percentage of correct recognition (using each method, for three different pitch levels) were tabulated.

RESULTS AND DISCUSSION

The purpose of the study was two folds, i.e.,

- (1) To Review the literature concerned with speech recognition and
- (2) To find out the effect of change of fundamental frequency on Speech (digit) recognition task.

An extensive review of literature has been made in Chapter II.

As seen in the methodology adopted here; this study was conducted in two phases. Parameter estimation was done and group vocabulary was made for the data collected in the first phase of experiment for each subject. Finally, each digit recorded and analyzed in the second phase of the experiment were compared (i.e., were fed to be recognized) with the data obtained from the first phase of experiment.

Results obtained from the experiment have been tabulated and percentage of correctly recognized digit from each method have been calculated.

Table-I shows the number of correctly recognized digit at different pitch levels using three different recognition methods i.e., Cosh Measure, Minimum Prediction Residual Method and Linear Prediction Coefficient.

The criteria for considering a digit being recognized was that at least three out of five utterances of the subjects were recognized correctly.

TABLE - I

CORRECTLY RECOGNISED DIGITS DENOTED BY 'R'

METHODS USED	COSH MEASURE			MINI HUH PREDICTION-RESIDUAL METHOD			LINEAR PREDICTION COEFFICIENTS!		
	N	H	L	N	H	L	N	H	L
ONE	R	R	R	R	-	R	R	R	-
TWO	R	R	R	R	R	R	R	-	R
THREE	R	R	-	-	-	-	-	-	-
FOUR	R	-	R	R	R	R	R	R	R
FIVE	R	R	R	R	R	R	R	R	-
SIX	R	R	-	R	R	-	R	-R	-
SEVEN	-	R	-	-	-	-	-R	-R	R
EIGHT	R	-	-	R	-	-	R	-	R
NINE	R	-	-	-	-	-	R	-	-
ZERO	-	R	-	-	R	-	R	R	-

Table-II shows the percentage of correctly recognized digits at three different pitches and percentage of correctly recognized digits using each method of recognition used in this study.

As the Table II indicates, there is difference between the three methods of speech recognition. Hence the hypotheses (2) is rejected. Table III shows the difference of percentage of correctly recognized digit using the methods adopted in this study for speech recognition. As per Table III there is slight difference, (about 10%) between the methods; Cosh Measure and Linear Prediction Coefficient. Hence the hypothesis-2a is rejected. Using the methods Minimum Prediction Residual Method and Cosh Measure; the difference is seen more (about 20% at Normal and High pitch) that means there is difference between the two methods of speech recognition. It rejects the hypothesis-2b.

Still more differences are obtained by using the method Minimum Prediction Residual method and Linear Prediction Coefficient, for digits recognized at habitual pitch - i.e., there is difference between the two methods. It rejects the hypothesis-2c.

It is very clear from Table II that at habitual pitch both Cosh Measure and Linear Prediction Coefficient yields good results. The digits are recognized 80% to 90% by these

TABLE - 11

PERCENTAGE OF SPEAKER - DEPENDENT CORECTLY RECOGNIZED DIGITS FOR 5 SPEAKERS.

DIFFERENT RECOGNITION PROGRAM USED	COSH	MEASURE	MINIMUM PREDICTION RESIDUAL METHOD	LINEAR	PREDICTION COEFFICIENT
THREE DIFFERENT PITCH LEVELS	N	H L	N H L	N	H L
Percentage of correctly recognised digits at three different pitches	80%	70% 40%	60% 50% 40%	90%	60% 40%

TABLE - 11

**DIFFERENCE IN PERCENTAGE OF
CORRECTLY
RECOGNIZED DIGITS AT DIFFERENT
PITCH
USING DIFFERENT PROGRAM.**

THREE DIFFERENT PITCH LEVELS	N	H	L
DIFFERENCE OF PERCENTAGE OF CORRECTLY			
RECOGNIZED DIGIT USING METHODS :-			
1 COSH MEASURE AND LINEAR PREDICTION COEFFICIENT	10%	10%	0%
2 MINIMUM PREDICTION RESIDUAL METHOD AND COSH MEASURE	20%	20%	0%
3 MINIMUM PREDICTION RESIDUAL METHOD AND LINEAR PREDICTION COEFFICIENT	30%	10%	0%

two methods at habitual pitch. But as the pitch is varied the percentage of recognized digit goes down. That means recognition of digits is affected by varying the pitch. This rejects the hypothesis-1.

Table IV shows the difference in percentage of recognition of digits at different pitch Level. It is very clear from Table IV that there is difference in terms of recognition of digits when the digits are uttered at - a) high pitch with respect to habitual pitch; b) low pitch with respect to habitual pitch and c) low pitch with respect to high pitch. This finding rejects the hypothesis-1a, 1b and 1c respectively.

At habitual pitch levels the digits are recognized satisfactorily using the methods Cosh Measure and Linear Prediction Coefficient (Euclidean Distance Measurement). The finding that Cosh Measure is a reliable method of speech recognition is also supported by Gray and Markel (1976). And the finding that Linear Predictive Coding Analysis is a reliable method of recognition of digits is supported by Sambur and Rabiner (1974).

The methods Cosh Measure and Linear Prediction Coefficient (Euclidean Distance Measurement) yields the same percentage of results i.e., 63.33%. That means 63.33% of digits are correctly recognized using these two methods. But the method, Cosh Measure is more time consuming when compared

TABLE - IV

**DIFFERENCE IN PERCENTAGE OF
RECOGNITION OF DIGITS AT DIFFERENT PITCH
LEVELS**

THREE DIFFERENT PITCH LEVELS	HABITUAL & HIGH PITCH	HABITUAL & LOW PITCH	HIGH & LOW PITCH
DIFFERENCE OF PERCENTAGE OF CORRECTLY			
RECOGNIZED DIGIT USING METHODS :-			
1 COSH MEASURE	10%	40%	30%
2 MINIMUM PREDICTION RESIDUAL METHOD	10%	20%	10%
3 LINEAR PREDICTION COEFFICIENT	30%	50%	20%

to Linear Prediction Coefficients (which is least time consuming among all the three methods used).

The most atypical finding was that the digits uttered at low pitch level when compared to habitual pitch yielded the same percentage (40%) using all the three methods.

Studies done on speaker recognition have (Abberton, 1979; Lariviere, 1975, Atal, 1972) suggested that fundamental frequency is an important cue in speaker recognition tests. It is not all obvious what measures of fundamental frequency are likely to be the most appropriate. But it can be concluded that change in fundamental frequency may affects speaker recognition.

SUMMARY AND CONCLUSION

Computer has invaded all walks of life of human beings. Speech Pathology is no exception to this. Computer has been used extensively for Speech analysis and synthesis for the diagnosis and treatment of speech disorders. However, speech recognition has not been used well for the purpose of speech and language therapy, even though it, has lot of potentials in terms of providing reinforcement to the cases, motivation and drill to the cases. Presently some of the attempts have been made to use this for articulation testing and therapy. In this context it was necessary to review the literature regarding speech recognition and also study possible variables affecting speech recognition. Therefore, this is part of a proposed extensive study on application of speech recognition in speech therapy and diagnosis.

The present study was aimed at reviewing the relevant literature and answering the following questions:-

1. Is there any difference in terms of recognition of digits when the pitch is varied from the habitual pitch?

2. Is there any difference in terms of three methods of speech recognition (Cosh Measure, MPR & LPC i.e., Euclidean Distance Measure) in terms of recognition of digits?

It was decided to use digits to start with, to make the problem simpler and also as others had used digits for recognition.

A sample comprising of five subjects (all males) with no history of hearing loss and vocal pathology were taken for this study.

The experiment was conducted in two phases. In the first phase the subjects were asked to utter the digits (0 to 9) using their habitual pitch at normal loudness. In the second phase of the experiment the subjects were asked to utter the same digits (0 to 9), in three different pitches i.e., habitual, high and low pitch

The recorded speech samples were analyzed with the help of a computer (PC-XT). For recognition of digits the methods, Cosh Measure, Minimum Prediction Residual and Linear Prediction Coefficients were used.

The presented study has revealed that the percentage of correctly recognized digits is varied with respect to habitual pitch. The digits uttered at low pitch were very

poorly recognized (about 40 %) . Some what better result was seen at higher pitch level with respect to low pitch level. But when compared to habitual pitch the digits were less recognized at high pitch.

A comparision was also made between the three methods of speech recognition used in this study (Cosh Measure, MPR and Linear Prediction Coefficient). On comparison, it was found that there was difference between the three methods of recognition. But the methods Cosh Measure and Linear Prediction Coefficient (Euclidean Distance Measure), yielded almost same result; where as the Minimum Prediction Residual method yielded very poor result.

Above all it was found that Cosh Measure is reliable method of speech recognition. Similar results have been reported by Gray and Markel (1976).

Based on the above results it may be inferred that the variation in pitch affects the recognition of digits.

This finding however, is restricted to small sample. Further studies need to be carried out before generalizing these results.

Recommendations for future research :

1. Similar experiments can be carried out by with different speech samples.
2. This study can be done on larger population.
3. Other recognition programmes may be, used with similar stimuli and conditions.

BIBLIOGRAPHY

- Ann, K. Syrdal and Gopal, H.S. (1986): "A perceptual model of vowel Recognition based on the auditory representation of American English Vowels", J.Acoust.Soc.Amer., 79(4), 1086-1101.
- Atal, B. (1972): "Automatic speaker Recognition based on pitch contours", J.Acoust.Soc.Amer., 52, 1687-1697.
- Atal, B. (1974): "Effectiveness of Linear Prediction characteristics of the speech wave for automatic speaker identification and verification", J.Acoust.Soc.Amer., 55, 1304-1312.
- Bristow, G. (1986): "Electronic Speech Recognition, Techniques, Technology and Applications", Collins Professional and Technical Books; London.
- Chaslav, V.P., (1987): "Derivation of Primary Parameters and Procedures for use in Speech Intelligibility Predictions", J. Acoust.Soc.Amer., 82, 413-422.
- Chen, F.R., and Zue, V.W., (1983): "Exploring Allphonic and lexical constraints in a continuous recognition system", J.Acoust.Soc.Amer.Suppl. 174, S15.
- David, G. and Joan, S. (1983): "Speaking fundamental frequency: Some physical and social correlates, language and speech 26, 351-365.
- David, B. Pisoni (1985): "Speech Perception: Some new directions in Research and Theory", J.Acoust. Soc.Amer. 78(1), 381-387.
- Flanagan, J.L. (1972): "Speech Anlysis Synthesis and Perception", 2nd Edn., Heidelberg, New York.
- Gray.A.H. Jr., and Markel.J.D. (1976): "Distance measures for speech Processing" IEEE, Trans. ASSP, ASSP -24, 380-390.
- Gupta, V and Mermelstein, P. (1982): "Effects of speaker accent on the performance of a speaker independent isolated word recognizer", J.Acoust.Soc.Amer., 71, 1581-1587.

- Itakura, F. (1975): "Minimum Prediction Residual Principle applied to Speech Recognition", IEEE Trans. ASSP, ASSP -23, 67-72.
- Jakimik, J. and Hunnicutt.S (1981) "Organizing the lexicon for recognition", J.Acoust.Soc.Amer.Suppl.1.69, S41
- Klatt (1977) "Review of ARPA Speech Understanding Project", J,Acoust.Soc.Amer., 62, 1345-1366. .
- Lea, W, 91980): "Speech Recognition Pat, Present and Future" 39-98, in "Trends in Speech Recognition" Prentice-Hall: Englewood Cliffs, NJ).
- Paliwal, K and Rao, P. (1982): "Synthesis - based recognition of continuous speech", J.Acoust.Soc. Amer., 71, 1016-1024.
- Rabiner, L and Wilpon, J. (1980): " A simplified, robust training procedure for speaker trained, isolated word recognition systems", J.Acoust.Soc.Amer., 68, 1271-1276.
- Rabiner, L.R. and Sambur, M.R. (1976): "Some preliminary experiments in the Recognition of connected digits", 244-254 in N.R.Dixon and T.B.Martin; ed. Automatic Speech and Speaker Recognition, IEEE Press, Inc. New York.
- Reddy, D.R. (1976): "Speech Recognition by Machine: A review" Proc. IEEE 64, 501-523.
- Sarabur, M.R. and Rabiner, L.R.(1975): "A speaker -independent Digit -Recognition system" 151-172; in N.R.Dixon and T.B.Martin¹, Ed. Automatic Speech and Speaker Recognition, IEEE Press, Inc. New York.
- Schroeder, M.R. (1968): "Similarity measure for Automatic Speech and Speaker Recognition", J.Acoust.Soc.Amer., 43, 375-377.
- Schroeder, M.R. (1985): "Speech and Speaker Recognition", KARGER, New York.
- Subrata, K.D. (1982): "Some experiments in discrete utterance recognition", IEEE Trans. ASSP, ASSP - 30, 766-770.
- Tosi Oscar and others (1972): "Experiment on voice identification", J. Acoust. Soc. Amer., 51, 2030-2043.

- Keller, T.G. (1971): "On-line Recognition system for spoken digits", J.Acoust.Soc.Amer., 49, 1288-1296.
- J. (1972): "Efficient acoustic parameters for speaker recognition". J.Acoust.Soc.Amer., 51, 2044-2056.
- G.M. (1976): "Speech Recognition: A tutorial overview; 87-100, in Dixon N.R. and Martin, T.B.; ed. Automatic Speech and Speaker Recognition, IEEE Press, Inc. New York.
- .e, G.M. and Neely, R.B. (1976): "Speech Recognition Experiments with Linear prediction, Bandpass filtering, and dynamic programming; 172-187, in N.R.Dixon and T.B.Martin; ed. Automatic Speech and Speaker Recognition, IEEE Press, Inc. New York.
- Yegnanarayana, B., Chandran.S. and Agarwal, A. (1984): "On improvement of performance of isolated word recognition for degraded speech", Signal Processing 7, 175-183.
- Zwicker, E., Terhardt, E., and Paulus, E. (1979) "Automatic Speech Recognition using Psychoacoustic models". J.Acoust.Soc.Amer. 65, 487-498.
- Zue, V. (1985): "The use of Speech Knowledge in automatic speech recognition", Proc. IEEE 73, 1602-1615.