INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE CENTRATOM TRIESTE

H4.SMR. 405/3

SECOND WORKSHOP ON TELEMATICS

6 - 24 November 1989

Automatic Speech Recognition

P.V.S. Rao
Tata Institute of Fundamental Research, Bombay, India

LECTURES 2 AND 3

AUTOMATIC SPEECH RECOGNITION

1. Introduction

1.1 Objective and scope

In April 1981, the Japanese government announced a comprehensive ten-year programme aimed at acquiring world supremacy in information technology. This plan was motivated by their assessment that the country's future economic viability depended on leadership in this field. The main thrust of this programme, undertaken as a joint venture by the Japanese Government, Industry and Academic Institutions, was the development of a new (fifth) generation of computers. The first comprehensive presentation of the philosophy behind this project was made at the International Conference on the Fifth-Generation Computer Systems held in Tokyo in October 1981. (Feigenbaum and McCorduck 1983).

The Japanese thrust in the direction of realisation of fifth generation computers triggered more or less similar exercise in several other countries: notably USA, France, USSR and UK. The Department of Electronics of the Government of India has also formulated plans for the initiation of activity of leading to the realisation of concepts techniques, technologies, and systems, relevant to the implementation of fifth generation computers, as a multi-institutional effort. Financial provisions have been made for this purpose in the seventh five-year plan.

The main features visualised by the Japanese researchers for their fifth generation computers were-the ability to provide assistance to the user at the level of an 'expert' in any given area of activity, the capability of accepting instructions from him in 'natural language' and the provision for interacting with him in the speech mode.

The last mentioned of these features, the ability to interact with the user in the speech mode, implies that it should be possible for the computer to accept information given to it as a spoken message as well as to give him information in the form of spoken messages: recognition and synthesis of speech. Work on automatic speech recognition (ASR) received a fresh impetus in the recent past as a consequence of its relevance to fifth generation computers.

Our aim here is to provide a brief overview of the state-of-the-art in ASR research. It is almost impossible to entirely cover the many advances that have taken place in the last few

years. In fact, some of these are not even fully reported in the literature for reasons of their commercial value. We have, therefore, tried to present here the current status of ASR, primarily in terms of those advances with which we are familiar enough to comment confidently upon. To supplement this, we have included at the end of the paper a selected bibliography for further reading, in addition to references.

1.2 Advantages of ASR

Work in the area of speech recognition has been pursued for the last three decades for various reasons. The main motivating factor has been the hope of utilising speech for communication between man and machine. Interaction in the speech mode would be much more convenient than other modalities (such a typing) because of its universality, convenience and speed. The best communication aids now available through modern technology notwithstanding, speech remains unrivalled as the fastest and most convenient means of interactive communication between man and man. The same advantages would be there in the case of speech interaction between man and machine.

There are also some disadvantages. Table 1 lists the advantages and disadvantages of man-machine interaction in the speech mode. It can be seen from this table that most of the disadvantages can be overcome. Even as early as ten years ago, it was evident on the basis of practical experience that the advantages of speech I/O far outweigh the disadvantages (Martin 1976).

1.3 Applications of ASR

There are, quite understandably, countless applications for ASR. Applications of individual systems would naturally depend on their capabilities and limitations. ASR systems which can recognise words spoken in isolation have been commercial use for the last ten years and have found a number of significant applications (Martin 1976). Typical among these are industrial robots, command and control environments or data entry situations, where the use of a key-board is not practical because the users hands are otherwise occupied.

As ASR technology advances, ASR systems are finding wider applications in civilian as well as in military tasks (Beek et al 1977; Woodard & Cupples 1983). In these applications, ASR systems are used either alone (e.g. in speech control, toys and games etc.) or in association with speech synthesis and speaker recognition systems. These applications are listed in table 2 (civilian tasks) and table 3 (military tasks).

Table 1. Advantages and disadvantages of speech I/O

Advantages

Engineering
* Can be faster than other modes of communication.
* Can be more accurate than other modes of communication.
* Compatible with existing communication systems, that is telephones.
* Can be more accurate in tasks currently performed by humans, that is, automatic speaker verification vs. identity verification by human visual inspection.
* Can reduce manpower requirement.
* Requires little panel space in cockpits.

Psychological
* Most natural form of human communication.
* Best for group or team problem solving.
* Universal (or nearly so) among humans.
* Can reduce visual information overload.
* Increase in value when the person is engaged in activities requiring highly complex cognitive processing.

Physiological
* Requires less effort and motor activity than other communication modes.
* Frees hands and eyes and does not require physical contact with a transducer.
* Permits multimodal operation.
* Is feasible in a darkened environment.
* Is omnidirectional and does not require direct line of sight.
* Permits considerable operator mobility.
* Contains information on identiy and emotional state of speaker.
* Contains information on physical state of the speaker.
* Simultaneous communication with machines and humans are possible.

Disadvantages

Engineering
* Competing acoustic signals may interfere with speech. These include noise, distortions, and competing talkers.
* Physical conditions can change the acoustic characteristics of speech, that is, vibration, g-forces, and physical orientation of the speaker.
* Unlike typing, there is no permanent record of speech (unless explicitly recorded).
* Microphones are required for speech input, and acoustic speakers are required for speech output.

Psychological
* Speech is not private and may be observed and recorded by others.
* Psychological changes (stress, for example) in the speaker may change his speech characteristics.
* Synthetic speech output may interfere with other aural indicators.

Physiological
* Fatigue can result from prolonged speaking, and this may change speech characteristics.
* Physical ailments such as colds may change speech characteristics.

(Source: Woodard & Cupples 1983)

Table 2. ASR applications in civilian tasks.

Application	User	Provider	Accetability parameters Quality	Price	Examples
Toys & games	Occasional	Industry	Immaterial	Low	Vice-controlled toys and games with speech output (spelling testers etc.)
Industry and commerce	Profes-sional	Business	High	Non-critical	Stock control, access conrol alarms systems CAD
Handicapped persons	Handi-capped persons	Govt. and ac-tions groups	High	critical or non-critical according to politi-cal cir-cumstances	Assistance in communication, access to in-formation reha-bilitation
Telecommu-nications	Occasional	Govt. P & T Dept.	Very high	Immaterial	Dialling, in-formation as sistance in call set-up.
Consumer goods	Occasional	Industry	Very high	Immaterial	Translations, data base ac-cess, inter-active servi-ces (goods ordering, tic-ket reserva-tion). auto-mobiles

Teaching	Govt.	Very high	Immaterial	Phonetic and programmed teaching.

(Source: Gagnoulet & Mercier 1981)

1.4 Problems in ASR

Though ASR research has been actively pursued over the past three decades, a successful recognition system for unrestricted continuous speech is yet to emerge. Why is continuous speech recognition so difficult ? The problems in developing an ASR system are really formidable and are described below.

Speech recognition is basically a decoding process: the inverse of the speech encoding process that takes place when one speaks. An understanding of the basis of this encoding process is therefore necessary to appreciate the problems of recognition.

Figure 1 shows the human speech production system along with its schematic representation. The lungs and the associated respiratory muscles constitute the source of power. This power is used to generate the quasi-periodic acoustic signal by means of the vibrating vocal cords for voiced sounds such as vowels. for fricative sounds (such as /f/ and /s/) it is converted into an aperiodic (noisy) signal due to the high velocity frictional flow of air through a narrow constriction formed in the mouth. For plosive sounds (such as /p/ and /t/) it is converted into short bursts of noise by the sudden release of pressure which is built up by completely closing the vocal tract for short durations. Thus, all of the above mechanisms convert the more or less steady pressure of the lungs (DC power) into an acoustic signal (AC power) which is used for exciting the vocal tract system to generate audible speech sounds.

Table 3. Speech I/O applications in military tasks

Security
* Speaker verification (authentication)
* Speaker identification (recognition)
* Determining emotional state of speaker (e.g. stress effects)
* Recognition of spoken codes
* Secure access voice identification, whether or not in combination with fingerprints, facial information, identity card, signature, etc.
* Surveillance of the communication channels

Command and control
* System control (ships, aircraft, fire control, situation displays etc.)
* Voice-operated computer input/output (each telephone a terminal)

* Data handling and record control
* Material handling (mail, baggage, publications, industrial applications)
* Remote control (dangerous material)
* Administrative record control

Data transmission and communication
* Speech synthesis
* Vocoder systems
* Bandwidth reduction or, more general, bit-rate reduction
* Ciphering/coding/scrambling

Processing distorted speech

* Diver speech
* Astronaut Communication
* Underwater telephone
* Oxygen mask speech
* High G force speech

(Source: Beek et al 1977)

In all these cases, the frequency-wise distribution of acoustic energy is achieved by the dynamically changing shape and size of the vocal tract. These changes are effected, and different sounds are produced, by the movement of the articulators: tongue, lips, jaws and velum. For nasal sounds, the velum moves to connect the nasal tract to the vocal tract.

The shape of the vocal tract uniquely determines the sound that is produced. The problem of speech recognition may be visualised as that of determining of surmising, from the information contained in the speech signal, the causative movements of the articulators, and from thence, the spoken message. This message can be seen to be composed of more or less discrete entities at various levels: sentences, phrases, words, syllables and so on. ASR therefore requires two operations: (1) segmentation or the process of dividing the running speech signal into discrete segments; and (2) classification of each segment or recognising it as one of the finite number of elements of the vocabulary e.g. words.

There is a trade-off between complexities in segmentation and classification, depending on the type of entity chosen as the basic element for recognition. Choice of a higher level of entity as the element may be expected to simplify the segmentation process in some sense but this complicates the task of classification because the number of classes (or different elements in the vocabulary) increases. For instances, segmentation becomes trivial if one chosen sentence length elements, but the number of classes (individual sentences possible) become infinitely many.

Segmentation and classification would both have to be performed on the basis of the acoustic properties of the speech signal. To do these operations quantitatively, it becomes necessary to select a small number of convenient parameters and to extract their values from the speech signal at various points of time. These parameters have to be collectively adequate for characterisising the dynamically changing configuration of the vocal tract as a function of time.

Spoken utterances can, for all practical purposes, be adequately expressed using a phonetic alphabet consisting of a small number of elements, called 'phonemes'. Each phoneme is defined, in articulatory terms, by the positions of the various articulators (and thus the shape of the vocal tract) that are necessary to produce it: the articulatory targets. During normal speech, the articulators are required to rapidly move from one articulatory target to another. The time constants of the movement of the articulators, and their consequent inability to assume configurations corresponding to individual phonemes abruptly, account for the continuity of the acoustic signal of speech (the speech signal). This also implies that there are portions in the speech signal (corresponding to movements from one target to another) that do not correspond to any single phoneme. Also, to achieve speed, individual articulators move to positions corresponding to a following phoneme while earlier phonemes are still being uttered. A consequence of this anticipatory coarticulation is that the acoustic properties of a given segment of speech do not depend merely on the identity of the corresponding phoneme, but on the phonemes that follow as well. A third complication is that, during rapid speech, the articulators do not necessarily assume the proper target positions for each phoneme.

As a consequence, one faces two main problems in recognising continuous speech by machine: 1) There is, in general, no practical criterion by which one can associate with each element of the phonemic string a well-defined segment of the corresponding speech signal in a clear-cut manner. 2) The acoustic characteristics of a given phonemic segment display enormous variability in different phonemic contexts.

Change of speakers adds another dimension to the problem of acoustic variability discussed above. The acoustic characteristics of a given phoneme vary widely when spoken by different persons. In addition, even the same person can pronounce a given sound differently from one rendition to the next. These differences become particularly pronounced when the speaking rates differ.

These two fundamental problems (the segmentation problem and the problem of acoustic variability resulting from different phonemic contexts, speaking rates and speakers) are illustrated

in figures 2 and 3, respectively. It has been observed that differences in acoustic characteristics occurring due to different speakers and speaking rates are, in many instances, more than the inter-phonemic differences.

There are, in addition, a number of engineering problems that the designer of an ASR system has to deal with. For example, a typical application of an ASR system may be insider the cockpit of an aeroplane; engine noise would be the main problem here. For such applications, special signal processing techniques have to be devised which can estimate the values of recognition parameters from noisy speech. Speech utterances may also be accompanied by speech-related noise such as lip smacks, tongue clicks, breath sounds, and inadvertantly spoken like 'uh's' and 'er's'. These cause problems in detecting end-points (beginnings and ends) of the speech utterances.

Because of these problems, it is impossible to achieve recognition of continuous speech using acoustic information alone. Humans, however, can understand spoken language even when the speech signal is corrupted by noise. For this they use not only acoustic information but also their broad knowledge of the world, which includes higher level sources of knowledge such as syntax of the language, and the semantics of the task environment. ASR systems can also make use of these higher level sources of knowledge to improve their recognition performance.

ASR systems which use such higher level sources of knowledge have been termed 'speech understanding systems' (Newell et al 1973). The goal of speech understanding systems is to understand what the speaker meant rather than recognising what he said; as long as the message is understood, it hardly matters whether each and every phoneme or word is recognised correctly or not. The problem here relates to codifying, storing and using all the knowledge of the world which human beings acquire through their experience of many years. In speech recognition systems, the emmphasis is on recognising every phoneme or word correctly.

1.5 Types of ASR systems

In the last subsection, we discussed the problems one encounters in developing an ASR system. To develop an ASR system that is sophisticated enough to recognise or understand continuous speech spoken by any speaker under any circumstances would be next to impossible. It is therefore necessary to simplify the speech recognition problem by specifying some constraints; the size of the vocabulary (number of phonemes or words), the type of speech (isolated words or connected speech), the number of speakers acceptable to the system, the task environment (airline flight time enquiries, telephone directory enquiries, numerical computation tasks, etc.), the acceptable noise level and speech quality (sound treated room, normal

office, telephone quality speech, etc.), and even the syntactic structure of sentences.

Depending on the type of constraints, ASR systems can be broadly classified as follows: 1) isolated word recognition systems, 2) connected word recognition systems, 3) phoneme recognition systems for continuous speech, and 4) speech understanding systems. Significant advances have been made in the last decade in each of these types. Isolated word recognition systems are already commercially available. One can buy, for less than US $ 100, toys and video games which respond to speech commands. A dictation machine which has unlimited syntax and a 5000 word vocabulary has been recently demonstrated by the IBM speech group. Table 4 gives a summary of the advances already made and the future outlook for different types of ASR systems. These advances in systems will be discussed in more detail in later sections of the paper.

Table 4. Speech recognition system milestones

Recognition capability	Isolated words, speaker-dependent	Connected words, speaker-dependent	Phonemic recognition of continuous speech speaker-dependent	Continuous speech, speaker-independent	Isolated words, speaker-independent	Isolated words, speaker-independent	Continuous speech speaker-independent
Syntax	Limited	Limited	Unlimited	Limited	Unlimited	Unlimited	Unlimited
Vocabulary, number of words	200	100	Unlimited	1000	5000	20,000	20,000
Processing speed used/required, megainstructions per second	1 to 10	1 to 10	1 to 10	100	300	1000	100,000
Technology used/required	Acoustic pattern matching of whole word Dynamic programming to solve problem of variations in duration of words (both achieved by Nippon Electric Co. commercial machines)	Same as for related words except for the dynamic programming algorithm modified for connected words (achieved by Nippon Electric Co. commercial machines)	Acoustic processing for segmentation and labeling Use of transition segments to compensate for acoustic variability due to different phonemic contexts and speaking rates (both achieved by TIFR system, but not in real time)	Beam-search strategy to narrow selection of words Better algorithms to determine word boundaries (both achieved by Harpy)	Probabilistic approach to determine words on basis of preceding words, for use in supplementing phonetics Faster searches using selection keyed to individual sounds (achieved by IBM experimental system)	Language constraints such as the fact that "vn" never begins English words – to narrow choices Relating acoustic signals to phonemes in the form of quantitative rules	Natural-language understanding Knowledge base to use context of speech to assist in recognition Learning from errors

(Source: Reddy & Zue 1983)

The acoustic processor divides the continuous input speech signal into small time-segments and extracts the relevant recognition parameters. It then performs a process of preliminary classification of the segments into some convenient groups or classes using a suitable pattern classification method. The selection of proper techniques for parametric representation, signal processing and pattern classification is important at this stage. In 2, we describe different signal processing techniques used for ASR. Section 3 describes different parametric representation used in ASR systems and their performance. In 4, we describe different pattern classification methods used in ASR and their relative advantages and disadvantages.

We present next an overview of different types of ASR systems and their present status. We deal with the isolated-word recognition systems in 5, connected-word recognition systems 6, phoneme recognition systems (for continuous speech) in 7 and speech understanding systems in 8.

We describe in 9 the IBM system which supports a large vocabulary and unlimited syntax. This system was demonstrated recently in March 1985, at the IEEE International Conference on Acoustics, Speech and Signal Processing held in Tampa, Florida, USA. In 10, we describe the present status of ASR research in India. We conclude with a discussion of the future outlook of ASR research in 11.

2. Signal processing techniques used in ASR systems

Sophisticated signal processing techniques are required in ASR systems to obtain accurate and reliable estimates of recognition parameters. Since the speech production systems generates sequences of phonemic sounds by changing the shape of the vocal tract, the parameters should characterize the time-varying shape of the vocal tract for speech recognition to be possible. As the sound sources and the vocal tract shapes are relatively independent, a reasonable approximation is to model them separately. This model is commonly known as the source-system model of speech production and is shown in figure 4a. In this model, a time-varying (digital) filter represents the vocal tract (system). This filter is excited by an appropriate (source) signal which is quasi-periodic for voiced sounds (figure 4b) and aperiodic for others (figure 4c). The output of this filter is the speech signal.

The speech signal is the result of the convolution of the source function and the vocal tract impulse response function. The short-time spectrum of this signal reflects the characteristics of both the source and the system. The periodicity of the source (in the case of voiced sounds) appears in the form of ripples in the spectrum, as shown in figure 4b, arising due to

the harmonics of the fundamental frequency (or pitch) of the vocal cords. The system characteristics are reflected in the overall shape of the smoothened power spectrum (spectral envelope).

The parameters of the filter (such as the impulse response function, transfer function or pole and zero frequencies) can be used to derive information about the vocal tract configuration: i.e. for speech recognition. The aim of acoustic analysis to estimate the parameters of this filter.

Since the parameters used in most ASR, systems are derived from the frequency domain representation of the speech signal, the main task of the signal processing technique is to compute the short-time power spectrum. In order to represent the speech signal as a sequence of short-time power spectra, one has to make the assumption that the signal remains stationary for the duration of the segment over which the analysis is made. The assumption is not valid for regions where there are sharp transitions, as when the articulators are moving fast from the target positions of one phoneme to those of another. For the stationarity assumption to be valid, it is necessary to choose as short an analysis segment as possible.

Two types of analysis procedures are possible for analysing speech signals: pitch-synchronous and pitch-asynchronous. In pitch-synchronous analysis (Pinson 1963), pitch pulses mark the beginnings of the analysis segments; the analysis segments can then be quite short (usually less than one pitch period). Thus, the stationarity assumption is quite easily satisfied for pitch-synchronous analysis.

However, it is not possible to reduce the analysis segment duration to that extent for pitch-asynchronous analysis. This is because arbitrary placement of the analysis segments (with respect to pitch pulses) can cause large errors in spectral estimation if the analysis segment is too short. A reasonable compromise for pitch-asynchronous analysis is to use a segment duration which is two to three times the pitch period.

The following three signal processing techniques have been used in ASR systems for obtaining the frequency domain representation of speech: 1) filter-bank analysis technique, 2) cepstral analysis technique and 3) linear prediction analysis technique.

Filter-bank analysis is the oldest spectral estimation technique but is still used even in the very recent ASR systems (Pols 1971; Dautrich et al 1983; Kuhn & Tomashchewrki 1983). Here, between 8 to 32 frequency bands are chosen covering the entire frequency range of interest. The amount of energy in each band is measured and averaged over the segment duration. The spacing and width of these frequency bands can either be uniform or vary in

some systematic way with frequency. Filter-bank anaylsis can be performed either by explicitly filtering the speech signal using a bank of filters or by processing the power spectrum computed through fast fourier transform (FFT) methods (Dautrich et al 1983 b). In some ASR systems, the frequency response characteristics of the ear have been used to assign filter parameter values (Zwicker et al 1979; Kates 1983). Though filter-bank analysis does not provide as parsimonious a representation of speech as the other two techniques, it has one major advantage. ASR systems using this technique have the most graceful performance degradation in noisy environments.

Filter-bank analysis provides gross information only about the composite spectrum which, as we noted earlier, is the result of the convolution of the source function and the vocal tract impulse response function. Cepstral analysis provides a technique for separating the two by computing the spectral envelope which truly reflects the system characteristics.

The Fourier transform of a speech signal is the product of the transforms for the source and the system. The ripples in the power spectrum therefore exhibit a periodicity $f = 1/T$, where T is the pitch period. If the logarithm of the power spectrum is taken and its Fourier transform is obtained, the resultant spectrum is the sum of the components corresponding to the source and system.

The power spectrum (square of the Fourier transform) of the logarithm of the power spectrum of the speech signal is defined as the 'Cepstrum' of the signal (Noll 1967). The independent variable of this function has the dimensions of the reciprocal of frequency (i.e., those of period) and is termed 'quefrency'.

The periodicity of the ripples in the power spectrum of the speech signal, which arises from the periodicity of the source function, appears in the cepstral domain as a sharp peak at a quefrency which is the same as the pitch period. The system characteristics manifest themselves as a broader peak at lower quefrencies. Since the cepstrum is the sum (and not the product) of the source and system components, separating the two is a straightforward process and can be accomplished by appropriate windowing. Going from the cepstral to the spectral doman is also quite straightforward.

The cepstral analysis technique (also known as homomorphic filtering) thus provides a procedure for computing the spectral envelope. This, in turn, can be used to extract the formant frequencies by the peak picking method (Schafer & Rabiner 1970). The procedure for computing the smoothed power spectrum can be outlined in the following four steps: 1) compute the log-power spectrum from the speech segment, 2) compute the cepstrum by taking the Fourier transform of the log-power spectrum, 3) window

the cepstrum for eliminating the pitch effects, and 4) compute the smoothed power spectrum by taking the inverse Fourier transform of the windowed cepstrum. Figure 5 shows the results obtained by using the cepstral analysis technique for computing the smoothed power spectrum. Figure 5a displays the power spectrum of vowel /a/ computed through FFT, (b) the cepstrum and (c) the smoothed power spectrum.

Apart from computing the smoothed spectrum, this technique provides an economic (but approximate) representation of spectral information for ASR in terms of the first few cepstral coefficients. Spectral pattern matching can be performed using these cepstral coefficients without having to compute the smoothed power spectrum explicitly (Gray & Markel 1976a).

The linear prediction (LP) analyais technique (Makhoul 1975b; Markel & Gray 1976) has been used more recently in ASR systems. Like the cepstral analysis technique, this technique too provides a procedure of estimating the short-time smoothed power spectrum. It also permits a more parsimonious representation of the smoothed spectrum than does the filter-bank analysis technique. It assumes an all-pole model for the speech signal shown in figure 6. According to this model, the speech signal is produced as the output of an all-pole (recursive) filter which is excited either by a periodic pulse train (for voiced speech) or by a random noise sequence (for unvoiced speech). The filter coefficients (also called linear predictor coefficients) are computed from the speech signal on the basis of a least-squares fit between the observed values and the values linearly predicted from preceding samples.

One basic difference between the LP and the cepstral analysis techniques is that while the LP analysis technique is parametric, the cepstral analysis techniques is not. The smoothed power spectrum can be exactly represented in terms of say p LP coefficients (where p is the order of the all-pole model used for LP analysis). In contrast, it is not possible to represent the smoothed spectrum exactly in terms of a few cepstral coefficients. Because of this, the LP analysis technique does a better job of spectral matching (without explicitly computing the smoothed power spectra) than the cepstral analysis technique.

Another advantage of LP analysis over cepstral analysis is with respect to formant extraction. The LP analysis technique computes the smoothed spectrum. The number of peaks in this spectrum can be controlled by prior specification of the order of the all-pole filter model. There are major peaks as well as 'kinks' in the smoothed spectrum computed through cepstral analysis. This can be seen from figure 7 which shows the smoothed spectra of a vowel segment computed through LP and cepstral analysis techniques.

Because of these advantages, LP analysis is the most popular signal processing technique used in ASR systems. As mentioned earlier, the LP coeffieicnts are obtained a solution of the mean-squared error minimization problem. There are three popular methods for estimating the LP coefficients which differ from each other in terms of solving the minimization problem (Markel & Gray 1976; Makhoul 1975a; Ullrych & Bishop 1975). These methods are: 1) autocorreletion method,) covariance method and 3) Burg method.

The covariance method does not ensure the stability of the estimated all-pole filter. The autocorrelation method guarantees this stability with floating-point computations. The Burg method ensures filter stability even with fixed point computations. In a recent study (Paliwal & Rao 1982a), the performance of these methods was found to be comparable for pitch-asynchronous analysis. For pitch-synchronous analysis, however, the autocorrelation and Burg methods do not perform as well as the covariance method.

A modified autocorrelation method (known as the cyclic autocorrelation methdo) has been proposed (Paliwal & Rao 1981) with performance comparable to that of the covariance method for pitch-synchronous analysis. Some modified versions of the Burg method have recently been proposed which bring the performance of the Burge method close to the covariance method for pitch-synchronous analysis (Paliwal 1984a).

More sophisticated techniques have to be used for ASR systems operating in noisy environments (as in airborne command posts or cockpits of fighter aircraft and helicopters). ASR systems operating in these environemtns use one of the following two strategies: 1) a preprocessing stage which uses speech enhancement techniques to eliminate noise (Lim & Oppenheim 1979; Hoy et al 1983; Paliwal '985c; Ephraim & Malah 1985); 2) robust signal processing techniques to estimate speech parameters even in the presence of noise (Lim & Oppenheim 1978; Chan & Langford 1982; Johnson et al 1983; Paliwal 1984c, 1985a, b; Jain & Atal 1985).

3. Parametric representations used in ASR systems

As mentioned earlier, the central component of all ASR systems is a pattern classifier which essentially recognises the input speech, segment by segment, by assigning a label to each segment (these labels are names of the elements of the vocabulary: permitted words in a word recognition system, phonemes in a phoneme-based system and so on). It does this by comparing the parameter vector associated with that segment with reference (prototype) parameter vectors representing different elements of the vocabulary. The test segment is recognised as the element which gives the best match.

Selection of an appropriate set of parameters and a proper classification technique are most important for correct classification. A number of parameters can be extracted - either directly from the speech signal or from its power spectrum. For classification, only those parameters should be chosen that are useful for discriminating between the classes efficiently.

For obvious reasons, the number of parameters must be kept small. It would therefore be desirable to choose uncorrelated parameters; alternatively, a feature selection method has to be used which takes advantage of statistical analysis techniques, such as principal component analysis and the analysis of variance, to reduce the dimensionality of the original space (Rao & Deodhar 1978; Paliwal et al 1978). It is easy to see that the following properties are desirable in these parameters: 1) they should convey information about class identity of the segment and should provide adequate separation between different classes 2) computing their values should be simple and easy, and most important, not prone to errors; 3) they should be stable over time and context, i.e., intra-class variations should be minimal; 4) their values should be insensitive to ambient noise; 5) ideally, their values should be speaker independent (at least for multi speaker ASR systems).

Formant frequencies (pole frequencies of the vocal tract transfer function) are the most frequently used parameters for ASR systems (Martin 1976; Reddy 1976; Paliwal & Rao 1982). This is for the following reasons : 1) formants have physical significance-they represent the vocal tract resonances. Formant trajectories manifest the dynamics of articulation fairly directly and to a fair degree of detail. For this reason, formant transition information is useful even for segmenting the speech signal (Broad 1972); 2) formant frequencies for given utterances by single speakers display remarkable inter-repetition stability (Peterson & Barney 1952); 3) formants provide a reasonable degree of the inter-class separation. It is possible to achieve a recognition accuracy of more than 80% in a speaker-dependent vowel recognition task using the frequencies of the first two formants alone as parameters (Forgie & Forgie 1959); 4) a large amount of data is available in a well-documented form from acoustic-phonetic studies aimed at characterising different speech sounds in terms of formants. This is a major incentive for using formants as the main parameters for ASR.

Many methods are available in the literature for automatic extraction of formants from the speech signal. Two recently proposed methods of formant extraction use heuristic methods for picking the peaks in the smoothed power spectrum (Schafer & Rabiner 1970; Markel 1972; McCandless 1974). The problem of automatic formant extraction is, however, far from solved. It is difficult to extract formants in the following situations: 1) two adjacent formants are so close that they merge into a single peak

in the smoothed spectrum (e.g. the second and third formants of the vowel /i/ and the first and second formants of vowel /u/); 2) one of the formants is either very weak or totally absent (e.g. the second formant for nasals /m/ and /n/, due to the presence of a zero in the same region of the spectrum); 3) spurious peaks appear in the smoothed spectrum.

There can be gross errors in automatic formant extraction in such situations. A vowel-recognition experiment was conducted (using continuous speech) to determine the severity of this problem and the extent to which it affects the performance of ASR systems (Paliwal & Rao 1980). Vowel recognition was first attempted using the automatic formant extraction method of Markel (1972). It was then repeated after manually identifying and correcting the gross errors in formant frequency estimates. Manual correction of gross errors improved the recognition performance by as much as 30%.

Because of such problems, LP parametric representation (which is not prone to such gross estimation errors) is slowly gaining ground over formant representation. In an experiment comparing the performance of formant and LP parametic representation (91.4%) were found to be better than for formant representation (84.4%). (Formant frequencies were corrected manually for gross errors in this experiment). While it may be possible to argue that such experiments, limited as they are in scope, are not necessarily conclusive, the trend is quite clear.

Several different LP parametric representations have been proposed in the literature (Viswanathan & Makhoul 1975; Gray and Markel 1976b) and are related to each other through nonlinear transformations. Though these representations provide equivalent information about the smoothed power spectrum, their recognition performance can be and is different. Table 5 summarises the results of an experiment aimed at comparing different LP parametric representations as to their performance in a vowel recognition task using a Euclidean distance measure (Paliwal & Rao 1982c; Paliwal 1982c). It can be seen from this table that the cepstral coefficients derived through LP analysis yield the best recognition performance (91.4%). A similar experiment was reported earlier regarding some of the LP parametric representations in an isolated word recognition task (Itchikawa et al 1973). Results from this experiment are also listed in table 5. These results confirm the superiority of the cepstral coefficients representation over the other LP representations.

When these cepstral coefficients were multiplied by their respective quefrencies, vowel recognition performance improved further (Paliwal 1982a). In a recent study (Davis & Mermelstein 1980) LP cepstral coefficients, linear-frequency cepstral coefficients and mel-frequency cepstral coefficients were compared in a speaker-dependent word recognition task; the mel-frequency

cepstral coefficients representation was found to give the best performance.

Various other parameters have been reported in the literature for ASR. For example, energies and zero crossing rates in different frequency bands have been used by Reddy et al (1973) for ASR. Zero crossing parameters have the problem that they sensitive to the DC level of the signal and to noise in the signal. Some ASR systems, such as the IBM system (Dixon & Silverman 1976), used as many as 80 parameters covering the whole log-power spectrum.

Table 5. Speech recognition performance of different LP parametric representations.

LP parametric representation	Performance (%) in Vowel recognition task	Isolated-word recognition task
Predictor coefficients $[a_n]$	70.3	77.0
Impulse response $[h_n]$ of the all-poll filter	87.4	
Autocorrelation coefficients of $[a_n]$	60.7	
Autocorrelation coefficients of $[h_n]$	80.0	92.0
Cepstral coefficients	91.4	100.0
Area coefficients	55.4	
Reflection coefficients $[k_n]$	82.6	98.0
Log area ratios	85.1	
Log error ratios	76.7	
Inverse sine of $[k_n]$	83.2	
Poles of the all-poll filter	60.6	

(Sources: Paliwal & Rao 1982c; Paliwal 1982c; Ichikawa et al 1973)

4. Pattern classification techniques used in ASR systems

As mentioned in 3, it is important to choose the right type of pattern classification method for the front-end acoustic processor used in the ASR systems. A pattern classification method provides an effective method of combining the contributions of a number of speech parameters which individually may not be adequate to discriminate between the classes. A large number of pattern classification techniques are available; these differe from each other with respect to the technique used (statistical, syntactic or based on fuzzy logic), type of training required (supervised or unsupervised) and the type of density function used (parametric or non-parametric). Though syntactic- and fuzzy logic-based pattern classification techniques have been used in some ASR systems, the majority take recourse to statistical pattern classification techniques. Choice of a parametric technique would be appropriate if there is adequate confidence in the parametric model assumed for the class-conditional probability density.

Most of the speech parameters normally used in speaker-dependent speech recognition systems follow a Gaussian distribution (Atal & Rabiner 1976; Paliwal 1978). A speaker-dependent vowel recognition experiment using the first three formant frequencies as parameters (Paliwal & Rao 1980) confirmed that classifier performance improves as the amount of information used increases). The experiment assumed a multivariate normal distribution and studied the performance of three pattern classification methods: 1) Bayesian classifier, 2) maximum likelihood classifier and 3) minimum distance classifier with Mahalanobis distance measure. Their vowel recognition performance scores were 77.8%, 77% and 75.9%.

Results of another vowel recognition experiment (Paliwal & Rao 1982c; Paliwal 1982a) using the LP cepstral coefficients representation are listed in table 6. This compares the performance of minimum distance classifer with the following four distance measures: 1) Euclidean distancfe measure, 2) correlation distance measure, 3) Mahalonobis distance measure and 4) Itakura's log-likelihood distance measure. It can be seen that the performance of the Mahalanobis distance measure, which requires second-order statistics about the different vowel classes, is the best. However, in many situations where the corpus of speech available for training is very limited, it is not possible to use a distance measure which requires second-order statistics. It can be seen from table 6 that among the three distance measures which require first-order statistics, Itakura's log-likelihood distance measure gives the best performance. This is because this distance measure is specially suited for comparing two speech segments represented in terms of LP coefficients. The other distance measures do not use any speech-specific property. Thus, it is advantageous to use distance measures, which use speech specific properties, for ASR systems. Recently, Nocerino et al (1985) compared several such distance measures in an isolated word recognition task. These include the Itakura-Saito, the log-likelihood ratio, the likelihood ratio, the cepstral, the weighted likelihood ratio and the weighted slope metric distance measures. They found that the log-likelihood and the weighted slope metric distortion measures give the highest recognition accuracy.

Table 6. Vowel recognition performance of different distance measures.

Distance measure	Recognition performance (%)
Euclidean	91.4
Correlation	90.8
Mahalanobis	96.0
Log-likelihood	94.2

(Source : Paliwal & Rao 1982c)

Though most of the better known ASR systems use the parametric pattern classification techniques discussed so far, some use nonparametric techniques. For example, some speaker-independent isolated-word recognition systems use multiple reference patterns to represent a single word (Rabiner 1978; Gupta et al 1978). Clustering techniques are employed in such systems to create these multiple reference patterns from the utterances of a large number of speakers. In order to employ multiple reference patterns per word, these systems use a non-parametric pattern classification technique using the k-nearest neighbour decision rule. Though the performance of the k-nearest neighbour classifer depends on various factors, the most important factor is the type of distance measure used (Paliwal & Rao 1983a).

5. Isolated word recognition systems

As mentioned in 1, isolated word recognition systems are the only type of ASR systems to have achieved commercial success so far. The last decade has witnessed a large number of research efforts which led to the development of speaker-dependent, small-vocabulary, isolated-word recognition systems. In these systems, the whole word is treated as the recognition unit; words are uttered either in isolation or in connected speech but with pauses between the words. There is thus no segmentation problem. Also, there is no problem of acoustic variability resulting from coarticulation effects.

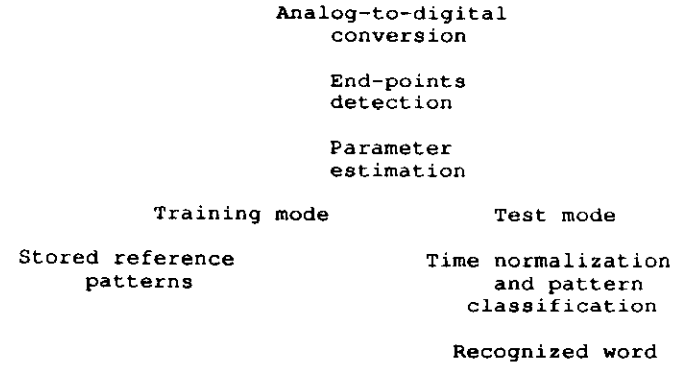
The only problem which these systems face is the acoustic variability resulting from fluctuations in speaking rates. Two utterances of the same word may be dissimilar, even if spoken by the same person, if the speaking rates are different. These utterances exhibit warping along the time axis as shown in figure

8a. It therefore becomes necessary to perform nonlinear time normalisation. A dynamic programming algorithm for performing this operation was first applied in isolated-word recognition systems by Velichko & Zagoruyko (1970) and Sakoe & Chiba (1971) and is commonly referred to as the dynamic time warping (DTW) algorithm. It effectively eliminates the nonlinear mismatch between two speech utterances by warping the time axis of one to get the best alignment with the other. This is illustrated in figure 8b where the second utterance is time-warped using DTW algorithm and plotted with the first utterance. Linear time normalisation works reasonably well for the recognition of monosyllabic words (Paliwal et al 1982c; White & Neely 1976), but the DTW algorithm becomes necessary for multisyllabic words (Itakura 1975; White & Neely 1976). This technique is therefore of major significance for isolated word recognition.

A typical isolated-word recognition system is shown in figure 9. Here, the spoken word to be recognised is digitized and its end points are detected. The values of recognition parameters are computed from the speech signal. The pattern so formed is time normalised using the DTW algorithm and compared with stored reference patterns for all the words present in the vocabulary of the system. The input pattern is recognised as the word whose reference pattern is most similar to the input pattern. It might be noted here that the DTW algorithm accomplishes both nonlinear time alignment and pattern matching in one step.

Speaker-dependent, isolated-word recognition systems perform with a recognition accuracy of 99% or more and have achieved remarkable success in a number of real life applications. However, they are of limited use in more general practical situations because of the following main reasons: 1) the requirement that pauses have to be deliberately inserted between words makes them unusable with natural speech; 2) high speaking rates are not possible because pauses with durations of 100 ms or more are required. Because of this, individuals using voice input systems could achieve average speaking rates of 30 to 70 words per minute only in factory environments (Martin 1976); 3) accurate detection of end points of the utterance is another problem. Though there are in the literature some algorithms for end point detection (Rabinder & Sambur 1975; Wilpon et al 1984), a satisfactory solution to this problem is yet to be found; 4) the number of words that can be recognised (the vocabulary of the system) has, of necessity, to be quite small.

So far, we discussed speaker-dependent isolated-word recognition systems. Attempts have recently been made to make these systems speaker-independent. Isolated-word recognition systems developed at Ell Laboratories have used clustering algorithms to find multiple reference patterns for a single word from a population of speakers (Rabiner 1978; Rabiner et al 1979). Rabiner et



al (1979) have found that upto 12 different reference patterns may be needed to represent the range of English pronunciations for single words. Though this approach gave good results for small groups of speakers, it could not be extended to a speaker whose speech characteristics were significantly different from those of subjects used for deriving the referecne patterns. In addition, this approach needs several reference patterns per word, i.e., more storage and computational power. Consequently, the size of the vocabulary has to be rather small.

Paliwal & Ainsworth (1985) proposed another approach for speaker-independent isolated-word recognition which requires only one reference pattern per word. Motivated by the success of the DTW algorithm for time normalisation, they tried dynamic frequency warping for normalising the differences between speakers. Though this algorithm worked well for speaker adaptation, it did not perform satisfactorily for speaker-independent isolated-word recognition.

6. Connected-word recognition systems

As we already saw, the isolated-word recognition systems described in the preceding section have the drawback that the speaker has to deliberately insert pauses between words. Connected-word recognition systems do not have this limitation. Though these systems treat the whole-word as a recognition unit, they allow the input speech to be in the natural form of connected words. Of course, even such systems permit only limited vocabularies.

The whole-word pattern matching technique used for isolated-word recognition systems has been extended for use in connected-word recognition systems. This is done by trying out various

sequences of words as possible matches for the input speech and finding the sequence which best matches the test pattern. The following three different algorithms, all based on dynamic programming, have been proposed for this 1) the two-level algorithm (Sakoe & Chiba 1979), 2) the level building algorithm (Myers & Rabiner 1981), and 3) the one-stage algorithm (Bridle & Brown 1979; Bridle et al 1982; Ney 1984). All the three algorithms perform equally well (about 98% for a 10 digit vocabulary, Sakoe & Chiba 1979), but differ in terms of their computational and storage requirements (see table 7). It can be seen from this table that the one-stage algorithm is most efficient in terms of computational load and storage requirements. In addition, unlike the other algorithms, there is no restriction regarding the maximum number of words in the input string.

The three operations of word-boundary detection, nonlinear time alignment and recognition are performed simultaneously by the DTW algorithm in connected-word recognition systems. Because of this, these systems do not face the end point detection problem. In addition, pauses are not required between words; input speech can therefore be more natural and there is no limitation on speaking rate. However, there is the problem of coarticulation across word boundaries which may modify the beginnings and the ends of words. This is illustrated in figure 10. Since the reference patterns used for pattern matching are concatenations of isolated words, they do not incorporate word-junction effects. This causes problems in recognition. However, if speaking rates are not very high and the speech is not deliberately slurred, these systems work reasonably well (Bridle et al 1982).

Connected-word recognition systems also can operate only with finite and small vocabularies, and hence, cannot be used for general purpose speech recognition.

Table 7. Computational comparison of dynamic programming algorithms for connected-word recognition and typical computational requirements for voice-dialing (i.e., 12 digits in a string).

	Two-level algorithm	Level building algorithm	One-stage algorithm
Number of basic time warps	K.N	K.M.	K
Size of time warps	J.(2R + 1)	J.N/3	J.N
Total computation	K.N.J.(2R+1)	K.M.J.N/3	K.J.N.
Storage	2.N.(2R + 1)	3.N.M.	2.(N + K.J)
Number of basic time warps	3600	120	10
Size of time warp	875	4200	12,600
Total computation	3,150,000	504,000	126,000
Storage	18,000	12,0001	1420

where J = 35 = average length of template;
K = 10 = number of templates;
M = 12 = maximum number of words in the input string, e.g., for voice-dialing (5 + 7) digits;
N = 360 = length of the input string;
R = 12 = range parameter for time warping

(Source : Ney 1984)

7. Phoneme recognition systems for continuous speech

Word-based recognition systems, described in 5 and 6, can operate on speech composed from a small and prespecified vocabulary. it is not possible to extend these systems for larger vocabularies because recognition scores fall as the size of the vocabulary increases. Computational and storage requirements also increase for a larger vocabulary. Using the phoneme as the recognition unit overcomes these limitations. Most languages have only about 40 phonemes; this is not a very large number. A phoneme-based ASR can operate on unlimited vocabularies even though it has reference patterns for only forty or so phonemes. In this sense, phoneme-based systems can be considered to be the ultimate in generality, conceptually. They are therefore likely to outlive other types of systems, eventually.

In phoneme recognition systems, speech recognition is performed in two stages: 1) the segmentation stage, and 2) the labelling stage. In the segmentation stage, the speech signal is divided into acoustic segments of phonemic length. This segmentation of continuous speech is a nontrivial problem because of the acoustic variability in the speech signal. This acoustic variability, as we saw earlier, occurs due to three factors 1) different speakers, 2) different speaking rates, and 3) different phonemic contexts.

Since speaker-independent speech recognition is too difficult a problem, most phoneme recognition systems reported in the literature are of the speaker-dependent type. A majority of these systems attempt segmentation and labelling without worrying about problems arising from different inter-phonemic context effects and speaking rates. However, there is growing realisation among speech researchers (Fant 1974; Paliwal 1978; zue 1983) that though the acoustic variability due to different phonemic contexts and speaking rates is large, it is systematic in nature and, hence, that some explicit rules can be applied to compensate for this variability. in order to provide a state-fo-the-art account on phonemic recognition of continuous speech, we briefly describe three phoneme recognition systems developed at: 1) the University of Erlangen, West Germany (Regel 1982), 2) the IBM

Research Laboratories, USA (Dixon & Silverman 1977), and 3) the Tata Institute of Fundamental Research (TIFR), Bombay (Paliwal & Rao 1982b). A brief description of these systems is given in table 8.

The Erlangen system (Regel 1982) does not provide any compensation for inter-phonemic context variations and speaking rate variations. The IBM system (Dixon & Silverman 1977) compensates for inter-phonemic context variations, but not for speaking rate variations. The TIFR system (Thosar & Rao 1971, 1976; Paliwal & Rao 1982b) compensates for both inter-phonemic context and speaking rate variations. It uses formant transitions for improved recognition scores. Also, it is synthesis based: it employs transition segments synthesised on the spot rather than prestored templates for comparison with the test sample. In a clever way, this prevents variations in the steady-state segment properties from interfering with recognition of the transition segments, and yields a significant improvement in the recognition scores. These systems are compared in table 8. The important point to be noted is that compensation for phonemic context variations alone improves the performance of the recognition system by 5.4% (Dixon & Silverman 1977), while compensation for both phonemic context and speaking rate variations improves the recognition performance by 9.7% (Paliwal & Rao 1982b). These differences, though small, indicate that if explicit rules are applied to compensate for systematic variations in acoustic characteristics occurring due to different phonemic contexts and speaking rates, the performance of speaker-dependent phoneme recognition systems can be improved significantly.

Recently, some experiments involving human reading of spectrograms have been reported in the literature (Cole et al 1980). In these experiments, subjects were taught a set of explicit rules for reading spectrograms. Subjects so trained could correctly identify upto 90% of the phonemes from spectrograms (independent of the speaker). This performance level in some sense sets the goal to be achieved by phoneme based ASR systems. Thus, there is still a long way to go in this direction.

8. Speech understanding systems

Due to the substantial financial support provided by the Advanced Research Projects Agency (ARPA), many research groups in USA have worked towards the development of systems capable of understanding the meaning or intent of naturally spoken utterances. Such systems should therefore be able to surmise the intended meaning even if they are not able to correctly recognise every phoneme or word. This requires "knowledge" of the world. Speech understanding systems use a number of knowledge sources operating at various levels: the characteristics of speech sounds (acoustic phonetics), variability in pronunciation (phonology), stress and intonation patterns of speech (prosodics), sound patterns of

words (lexicon), grammatical structure of the language (syntax), meanings of words and sentences (semantics) and context of the conversation (pragmatics) (Reddy 1976). Such sources of knowledge were used in combination in these systems to achieve higher performance than would be possible by using only some of these.

Speech understanding implies visualising, considering and choosing from a number of possible alternative ways of interpreting the signal at every stage (e.g., a phonemic segment may be one phoneme or another) and eventually choosing the correct sequence. It might appear that it would be safer to retain all alternatives til the end and then choose the most appropriate one from among these. This, however, would be impractical because of the combinatorial explosion that would arise in such a case. For example, if ther are 20 phonemes in a string and each of these could be any one of three phonemes, the number of possible combinations would be 3^{20}. It is therefore important to reject at every stage, alternatives which, for some reason or the other, are inapproapriate. It is, however, most important to avoid hasty rejection of wrong choices because 'back-tracking' would become necessary if a rejected choice has to be resurrected and this is quite messy.

A number of speech understanding systems have been developed under the ARPA project. These systems have been described in an excellent review article written by Klatt (1977). Klatt's article also describes the achievements of the ARPA programme and suggests some guidelines for future development of speech understanding systems. With the completion of the ARPA project in 1976, research on speech understanding in USA has been on a somewhat low key. A few speech understanding systems have been reported recently from Europe (Gillet et al 1982; Mariani 1982); none of them, however, is as ambitious as the ARPA speech understanding systems.

In order to present the state-of-the art in the speech understanding area, three ARPA speech understanding systems developed at Carnegie-Mellon University are selected here for a brief description. These are the Hearsay-1, Dragon and Harpy systems.

Hearsay-1 (Erman 1974) was historically the first speech understanding system to be demonstrated live. Three sources of knowledge (acoustics, syntax, and semantics) were used in the system as cooperating independent parallel processes. The processes used the hypothesize-and-test paradigm. The system worked in a particular task environment: voice-chess. The task was to recognize a spoken "move" in a given board position. The chess moves were expressible with a 31-word vocabulary and a finite context-free grammar of 18 production rules which was capable of generating about five million sentences. The system followed the best-first search strategy. it was tested on 79 sen-

Table 8. Comparison of three speaker-dependent phoneme recognition systems for continuous speech

Features	Erlanger system	IBM system	TIFR system
Compensation for phonemic context variations	No	Yes	Yes
Compensation for speaking rate variations	No	No	Yes
System structure	Recognition using steady-state segments	Recognition using steady-state and transition segments	Recognition using steady-state and transition segments
Recording environment	Unprepared room	Heavily sound-treated room	Ordinary office room
Sampling	10 kHz, 12 bit	20 kHz, 15 bit	20 kHz, 12 bit
Frame window	20 ms Hamming window	20 ms	12.2 ms Hamming window
Frame shift	12 ms	10 ms	10 ms
Parametric representation	18 parameters (energies in different bands, formants, LP error, autocorrelation coefficients)	40-point spectrum	9 parameters (energies in different bands, formant frequences, zero crossing rate)
Signal processing methods used	LP analysis	FFT	FFT and LP analysis
Classifier	Two-stage Bayes classifier	Mean-corrected minimum distance classifier	Two-stage minimum distance classifier
Segmentation results			
missed segments	6%	6.2%–6.9%	5.4%
extra segments	22%	6.1%–10.5%	8.3%
Labelling results			
without transeme classification	51%	57.9%	52.3%
with transeme classification	—	63.3%	62%
Computer used	PDP 11/34 (FORTRAN)	IBM 360/91 (PL1)	DEC 10 (FORTRAN)
Real time factor	400:1 (execution time)	9:1 (CPU time) without transeme classification	35:1 (CPU time) with transeme classification

(Sources: Regal 1982, Dixon & Silverman 1977, Paliwal 1978, Paliwal & Rao 1982b)

tences comprising 352 words. A word recognition accuracy of 79% was obtained.

In the Dragon system (Baker 1975), the knowledge sources were modelled as probabilistic functions of Markov processes. The system used a dynamic programming scheme of searching all possible paths in parallel to find the most optimal path. In other words, the system searched all possible sentences in the grammar, all possible pronunciations of each sentence and all possible dynamic time warpings of each such phonetic string to fit it best to the acoustic observations. In comparison the the Hearsay-1 system, the Dragon system provided much higher accuracy, but was found to be slower by a factor of 5 to 10.

The Harpy system (Lowrre 1976) used the best features of the Hearsay-1 and Dragon systems and additional heuristics to achieve higher speed and accuracy. This system used a beam-search strategy in which a restricted beam of alternatives around the best scoring path was considered, thus reducing the search time significantly without requiring backtracking. The Harpy system represented the syntactic, lexical and juncture knowledge in the form of a unified network of 15,000 states. Phonetic classification was achieved by a set of speaker-dependent acoustic-phonetic templates based on LP parameters which represented the acoustic realizations of the phonemes in the lexical portion of the network. The Harpy system displayed the best performance among the systems demonstrated as part of the ARPA speech understanding project (Klatt 1977). It satisfied most of the design goals that were specified at the beginning of the ARPA project (Newell et al 1973) and achieved a recognition score of 95% of the naturally spoken test sentences composed from a 1011-word lexicon. it ran in 6.8 MIPSS (millions of instructions per second of speech).

Though the ARPA sponsored speech understanding systems were fairly successful and (some of them) fulfilled the goals specified for them, these systems were not adequately general, being constrained in terms of the syntax of the language, the specificity of the task environment and so on. It would therefore not be correct to term them as general purpose ASR systems.

9. IBM speech recognition system for dictation transcription

The system developed by the Speech Recognition Group at IBM, Yorktown Heights, USA, is a significant landmark in the development of ASR systems (Jelinek 1985). This group implemented an experimental, real-time, speaker-dependent, isolated-word, speech recognition system with a large vocabulary and unconstrained syntax. This sytem can deal with a 5000-word vocabulary and can be used for dictating office correspondence. Figure 11 provides a brief description of its functioning.

This system has been tested on five speakers (four male and one female) and achieved remarkably high recognition accuracies (98% for prerecorded speech, 96.9% for read speech and 94.3% for spontaneous speech).

While it is a major achievement, the system still has the following limitations: 1) it uses speaker-dependent reference patterns for acoustic processing; 2) it is an isolated-word recognition system, i.e. it requires pauses between words; 3) its vocabulary at 5000 words is fairly large, but still limited.

10. ASR research in India

A number of research groups in India (mainly at research and educational institutions) are engaged in speech research. Unlike in the west where speech research is prompted by potential for practical applications, the interest in India has been primarily academic. Some of the better known among these groups are: 1) Tata Institute of Fundamental Research (TIFR), Bombay 2) Indian Statistical Institute (ISI), Calcutta 3) Indian Institute of Technology (IIT), Madras, 4) Indian Institute of Science (IISc), Bangalore, 5) Central Electronics Engineering Research Institute (CEERI), New Delhi and 6) Aligarh Muslim University (AMU), Aligarh. Out of these six, the first three have been interested in ASR research.

The TIFR group has been engaged in speech related activities for the last two decades. In preparation for work on ASR, this group started its activities with studies in speech synthesis (Rao & Thosar 1968, 1974) and human speech perception (Menon et al 1974; Rao 1974). Subsequently, it extended its interest, besides speech recognition, to speech coding (Paliwal & Ramasubramanian 1985; Paliwal & Krishnan 1985), speech enhancement (Paliwal 1984b, 1985c), speech parameter reduction (Deodhar et al 1974; Rao & Deodhar 1978; Paliwal et al 1978) and other related areas. The TIFR Group's work on various aspects of ASR included contributions to synthesis-based recognition (Thosar & Rao 1971, 1976), pitch extraction (Sreeniwas & Rao 1979a, 1981; Paliwal & Rao 1983b), signal processing techniques (Sreenivas & Rao 1979b, 1980; Paliwal 1981, 1982a; Paliwal 1984a), parametric representations of speech (Paliwal & Rao 1982c; Paliwal 1982a), pattern recognition techniques (Paliwal & Rao 1982c, 1983a) isolated word recognition system for Hindi digits (Paliwal et al 1982), and phoneme recognition system for continuous speech (Paliwal 1978; Paliwal & Rao 1982b). These contributions have already been discussed in earlier sections.

The main interest of the ISI group has been the various pattern recognition techniques used in the ASR systems. Datta et al (1980) used statistical pattern classification techniques to study the effectiveness of different formant parameters for the recognition of unaspirated plosives. Pal & Majumdar (1977) introduced the use of fuzzy logic-based pattern classification techniques for speech recognition.

The IIT group made a number of contributions in the speech recognition field. Yegnanarayana et al (1984b) proposed a signal processing method which reconstructs the signal from spectral magnitude or phase using group delay functions. This method has been extended to process noisy speech (Thomas et al 1985). Recently, this group proposed the use of signal-dependent parameter estimation and pattern matching for isolated-word recognition (Yegnanarayana & Sreekumar 1984; Raman et al 1984).

The IISc group worked mainly on the speaker recognition problem for which it developed several pattern recognition techniques (Dante & Sarma 1979; Sarma & Venugopal 1977). These techniques are equally applicable for ASR. In addition, this group developed a formant extraction procedure from the LP phase spectrum (Yegnanarayana 1978). Dattatreya & Sarma (1980) used an LP distance measure for vowel recognition and Bharathi Devi & Sarma (1980) have applied fuzzy set concepts in vowel recognition.
CEERI group and the AMU group worked mainly on human perception of speech (Ahmed & Agrawal 1969; Gupta et al 1969; Agrawal & Pavte 1980). These groups studied the importance of different acoustic cues in the perception of speech with the hope that these cues will be equally useful for ASR.

11. Future outlook and conclusions

Our aim is to present an overview of the different aspects of speech recognition by machine. Admittedly, this is a difficult and risky venture for an area like ASR which is very much open and where the last word is far from having been said. The magnitude and depth of the problem is evident from the fact that despite three decades of effort, a truly successful ASR system capable of dealing with continuous speech of any arbitrarily chosen speaker is far from realisation. To say this is not to be little the significant advances achieved in ASR technology during the last thirty years. Speaker-dependent, limited-vocabulary, isolated-word recognition systems have attained commercial success and are practically in everday use. Though it is speaker-dependent, the r ecently demonstrated IBM system has real-time isolated-word recognition capability for a 5000-word vocabulary and unconstrained syntax - no mean achievement. This system can be used for automatic dictation applications.

A brief but adequate summary of the current status of ASR research given in table 4. Columns 2 to 6 cover the milestones reached in ASR technology so far. Columns 7 and 8 indicate future trends.

One of the stated objectives of the Japanese fifth generation computer systems project is to develop a voice-operated

typewriter, which can recognise words from a 10000-word vocabulary and can work for hundreds of speakers. Considering the problems associated with large vocabularies and speaker-independent systems, it would seem unrealistic to seek to achieve this objective by as e arly as 1990. That the Japanese goal is not unachievable is clear from human spectrogram reading experiments where, merely with the help of some acoustic phonetic rules, human subjects could identify phonemes with 90% accuracy (independent of the speakers). An advantage of the Japanese language is the lack of ambiguity in spelling and pronunciation as compared to English.

Indian languages have the advantage of being phonetic and are therefore well placed for speech recognition. The fifth generation computer system programme initiated by the Government of India will provide a welcome incentive to speech research activities in the country and a practical motivation for ASR research in the country.

References :

Agrawal S S, Pavte K D 1980 Proc. Int. Symp. on speech processing (ed.) P V S Rao (Bombay:TIFR) pp. 801-834

Ahmed R, Agrawal S S 1969 J. Acoust. Soc. Am. 45: 758-763

Atal B S, Rabiner L R 1976 IEEE Trans. Acoust., Speech, Signal Process. 24:201-212

Baker J K 1975 Stochastic modelling as a means of automatic speech recognition Ph.D. thesis, Carnegie Mellon University, Pittsburg, USA

Beek B, Neuberg E P, Hodge D C 1977 IEEE Trans. Acoust., Speech, Signal Process. 25:310-322

Bharati Devi B, Sarma V V S 1980 Acoust. Lett. 4: 44-48

Bridle J S, Brown M D 1979 Proc. Institute of Acoustics Autumn Conference pp. 25-28

Bridle J S, Brown M D, Chamberlain R M 1982 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 899-902

Broad D J 1972 Int. J. Man-Mach. Stud. 4: 411-424

Chan Y T, Langford R P 1982 IEEE Trans. Acoust., Speech, Signal Process. 38: 689-698

Cole R A, Rudnicky A I, Zue V W, Reddy D R 1980 Perception and production of fluent speech (ed.) R A Cole (Hillsdale, New Jersey; Lawrence Erlbaum) pp. 3-50

Dante H M, Sarma V V S 1979 IEEE Trans. Acoust., Speech, Signal Process. 27: 255-263

Datta A K, Ganguli N R, Ray S 1980 IEEE Trans. Acoust.,Speech, Signal Process. 28: 85-91

Dattatreya G R, Sarma V V S 1980 J. Inst. Electron. Telecommun. Eng. 26: 77-81

Dautrich B A, Rabiner L R, Martin T B 1983a Bell Syst.Tech. J. 62:1311-1336

Dautrich B A, Rabiner L R, Martin T B 1983b IEEE Trans. Acoust., Speech, Signal Process. 31: 793-807

Davis S B, Mermelstein P 1980 IEEE Trans. Acoust., Speech, Signal PRocess. 28: 357-366

Deodhar M R 1975 Principal component analysis for machine recognition of speech Ph.D. thesis, Tata Institute of Fundamental Research, Bombay

Deodhar M R, Rao P V S, Rao V 1974 Proc. 8th Int. Congr. Acoust., London p. 252

Dixon N R, Silverman H F 1976a IEEE Trans. Acoust., Speech, Signal Process. 24: 137-162

Dixon N R, Silverman H F 197b IEEE Trans. Acoust., Speech, Signal Process. 24: 289-295

Dixon N R, Silverman H F 1977 IEEE Trans. Acoust, Speech, Signal Process. 25: 367-379

Ephraim Y, Malah D 1985 IEEE Trans. Acoust., Speech, Signal Process. 33: 443-445

Erman L D 1974 An environment and system for machine understanding of connected speech Ph.D. thesis, Carnegie Mellon university, Pittsburg, USA

Fant G 1974 Speech Recognition (ed.) D R Reddy (New York: Academic Press) pp. ix-x

Feigenbaum E A, McCorduck P 1983 The fifth generation artificial intelligence and Japan's computer challenge to the world (Reading, MA: Addison-Wesley)

Forgie J W, Forgie C D 1959 J. Acoust. Soc. Am. 31: 1480-1489

Gagnoulet C, Mercier G 1981 L'echo des Recherches (English issue) 35-46

Gillet D, Nouhenbellex A, Siroux J, Quinton P 1982 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 1633-1636

Gray A H, Markel J D 1976a IEEE Trnas. Acoust., Speech, Signal Process. 24: 380-391

Gray A H, Markel J D, 1976b IEEE Trans. Acoust., Speech Signal Process. 24: 459-473

Gupta J P, Agrawal S S, Ahmed R 1969 J. Acoust. Soc. Am. 45: 770-773

Gupta V N, Byran J K, Gowdy J N 1978 IEEE Trans. Acoust., Speech, Signal Process. 26: 27-33

Hoy L, Burns B, Solden D, Yarlagadda R 1983 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 3: 1133-1136

Ichikawa A, Nakano Y, Nakota K 1973 IEEE Trans. Audio Electroacoust. 21: 202-209

Itakuara F 1975 IEEE Trans. Acoust., Speech, Signal Process. 23: 67-72

Jain V K, Atal B S 1985 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2: 473-476

Jelinek F 1985 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 2: 858-861

Johnson R, Shore J E, Buck J, Burton D 1983 Proc. IEEE Int. Conf. Acoust, Speech, Signal Process. 3: 1129-1132

Kates J M 1983 IEEE Trans. Acoust., Speech, Signal, Process. 31: 148-156

Kay S M 1980 IEEE Trans. Acoust., Speech, Signal Process. 28: 292-303

Klatt D H 1977 J. Acoust. Soc. Am. 62: 1345-1366
Kuhn M H, Tomashchewrki H H 1983 IEEE Trans. Acoust. Speech, Signal Process. 31: 157-167
Levinson S E, Liberman M Y 1981 Sci. Am. 2444): 56-68
Lim J S, Oppenheim A V 1978 IEEE Trans. Acoust., Speech, Signal Process 26: 197-210
Lim J.S Oppenheim A V 1979 Proc. IEEE 67: 1586-1604
Lowrre B T 1976 The Harpy speech recognition system Ph.D. thesis, Carnegic Mellon university, Pittsburg, USA
Makhoul J 1975a Proc. IEEE 63: 561-580
Makhoul J 1975b Speech recognition (ed.) D R Reddy (New York: Academic Press) pp. 183-220
Mariani J J 1982 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process 1637-1640
Markel J D 1972 IEEE Trans. Autom. Control 20: 129-137
Markel J D, Gray A H 1976 Linear prediction of speech (Berlin: Springer-Verlag)
Martin T B 1976 Proc. IEEE 64: 487-501
McCandless S S 1974 IEEE Trans. A coust., Speech, Signal Process. 22: 135-141
Menon K M N, Rao P V S, Thosar R B 1974 Lang. Speech 17: 27-45
Myres C S, Rabiner L R 1981 IEEE Trans. Acoust., Speech, Signal Process. 29: 281-297
Ney H 1984 IEEE Trans. Acoust., Speech, Signal Process. 31: 263-271
Newell A, Barnett J, Forgie J W, Green C, Klatt d, Licklider J C R, Munson J, Reddy D R, Woods W A 1973 Speech Understanding systems: Final report of study group (Amsterdam: North Holland)
Nocerion N, Soony F K, Rabiner L R, Klatt D H 1985 Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 1: 25-28
Noll A M 1967 J. Acoust.Soc. Am. 41: 293-309
Pal S K, Majumdar D D 1977 IEEE Trans. Syst., Man, Cybern. 7: 625-629
Paliwal K K 1978 Computer recognition of continuous speech Ph.D. thesis, Tata Institute of Fundamental Research, Bombay
Paliwal K K 1982a Speech Commun. 1: 151-154
Paliwal K K 1982b Comparison of formant and LP parametric representations for vowel recognition, Technical Report, SDS Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K 1982c Comparison of different nonlinear transformations of reflection coefficients for vowel recognition, Technical Report, SDS Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K 1984a Speech Commun. 3: 221-231
Paliwal K K 1984b Speech enhancement using multipulse excited linear prediction system, Technical Report, CSC Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K 1984c Two new R spectrum estimation algorithms using extended Yule-Walker equations for noisy signals, Technical Report, CSC Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K 1984d Speech Commun. 3: 101-106
Paliwal K K 1985a An interative method of robust LP analysis of noisy speech, Technical Report, CSC Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K 1985b A robust LP analysis method based on pitch information for noisy speech, CSC Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K 1985c Linear phase FIR filter design for speech enhancement, Technical Report, CSC Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K, Ainsworth W A 1985 J. Phon. 13: 123-134
Paliwal K K, Agarwal A, Sinha S S 1982a Indian J. Tech. 20: 10-14
Paliwal K K, Agarwal A, Sinha S S 1982b Signal Process. 4: 329-333
Paliwal K K, Agarwal A, Sinha S S 1982c Advances in information science and technology (ed.) DD Majumdar (Calcutta: ISI) pp. 245-250
Paliwal K K, Krishnan S, 1985 A fully vector quantized adaptive transform coding system for speech, Technical Report, CSC Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K, Rao P V S 1977 Proc. 9th Int. Conf. Acoust. Madrid, Spain, 1:40
Paliwal K K, Rao P V S 1980 Indian J. Tech. 18: 285-289
Paliwal K K, Rao P V S 1981 Signal Process. 3: 181-185
Paliwal K K, Rao P V S 1982a Signal Process. 4: 59-63
Paliwal K K, Rao P V S 1982b J. Acoust. Soc. m. 71: 1016-1024
Paliwal K K, Rao P V S 1982c Signal Process. 4: 323-327
Paliwal K K, Rao P V S 1983a IEEE Trans. Pattern Anal. Mach. Intell. 5: 229-231
Paliwal K K, Rao P V S 1983b Speech Commun. 2: 37-45
Paliwal K K, Ramasubramanian V 1985a A study of vector quantization of speech waveforms, Technical Report, CSC Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K, Ramasubramanian V 1985b Adaptive vector quantization of speech, Technical Report, CSC group, Tata Institute of Fundamental Research, Bombay
Paliwal K K, Ramasubramanian V 1985c Distribution-free adaptive quantization of speech using clustering techniques, Technical Report, CSC Group, Tata Institute of Fundamental Research, Bombay
Paliwal K K, Sinha S S, Agrawal A 1982d J. Inst. Electron. Telecommun. Eng. 29: 18-22
Paliwal K K, Sreenivas T V, Rao P V S 1978 J. Acoust. Soc. India 6: 132-137
Peterson G E, Barney H L 1952 J. Acoust. Soc. Am. 24: 175-184
Pinson E N 1963 J. Acoust. Soc. Am. 35: 1264-1273
Pols L C W 1971 IEEE Trans. Commun. 20: 972-978
Rabiner L R, Levinson S E, Rosenberg A E, Wilpon J G 1979 IEEE Trans. Acoust.,Speech Signal Process. 27: 336-349
Rabiner L R, Sambur M R 1975 Bell Syst. Tech. J. 54: 297-315
Raman R, Yegnarayana B, Sundar R, Chandrasekaran V V 1984 Proc. int. Conf. computers, systems and signal processing, Bangalore, India, pp. 896-900

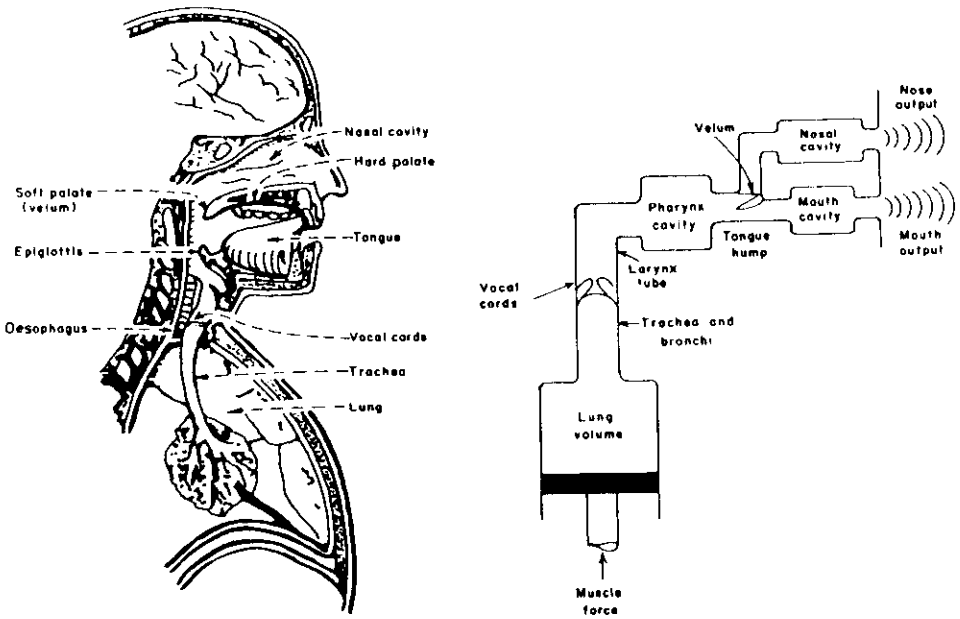


Figure 1. Speech production system and its schematic representation

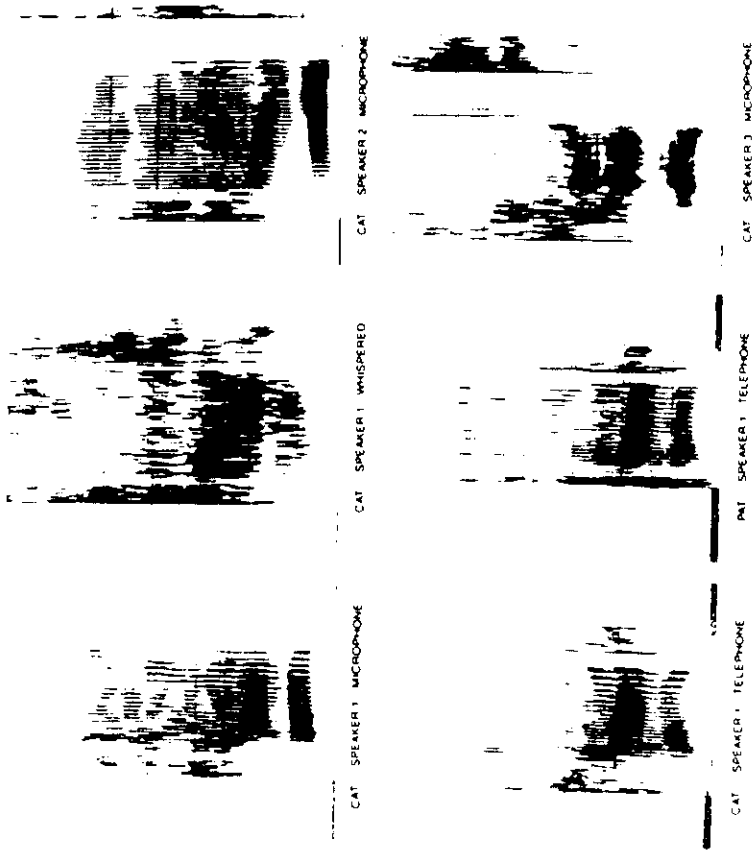


Figure 3. Illustration of the acoustic variability problem. Spectrograms of distinct but acoustically similar words may be more alike than the spectrograms of the same word pronounced under various conditions by different speakers. (Source Levinson & Liberman 1981)

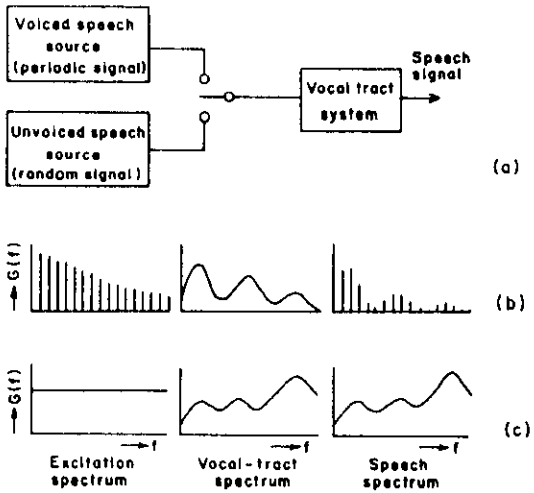


Figure 4. Source-system model of speech production and spectra of voiced and unvoiced speech. **a.** Source system model, **b.** voiced speech, **c.** unvoiced speech.

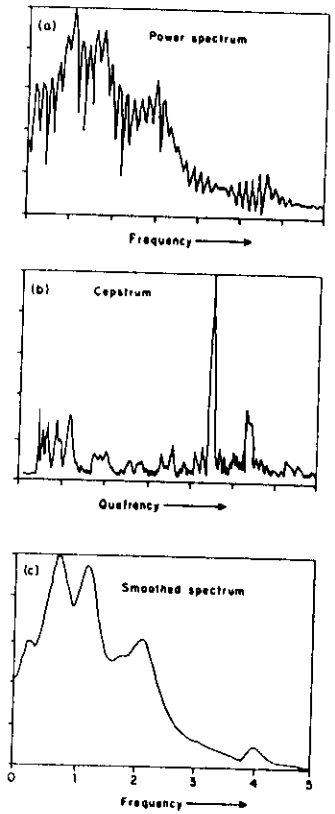


Figure 5. Illustration of cepstral analysis method for computing the smoothed power spectrum of vowel /a/. **(a)** unsmoothed power spectrum computed through FFT, **(b)** cepstrum, and **(c)** smoothed power spectrum.

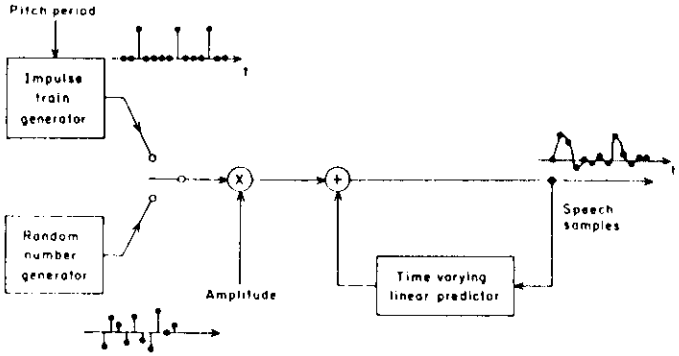


Figure 6. All-pole model of speech production used in linear prediction analysis.

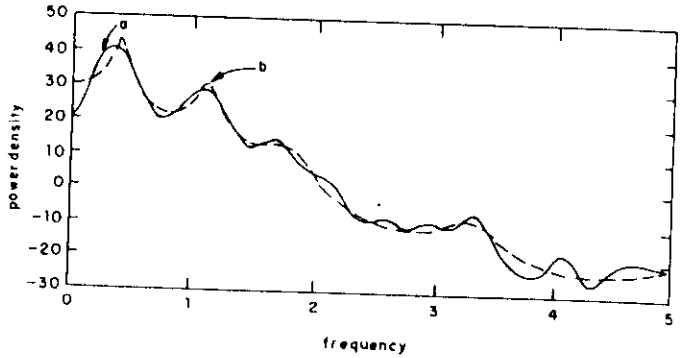


Figure 7. Comparison of cepstral and linear prediction analysis techniques for computing smooth power spectrum (units: power density - dB, frequency - kHz).

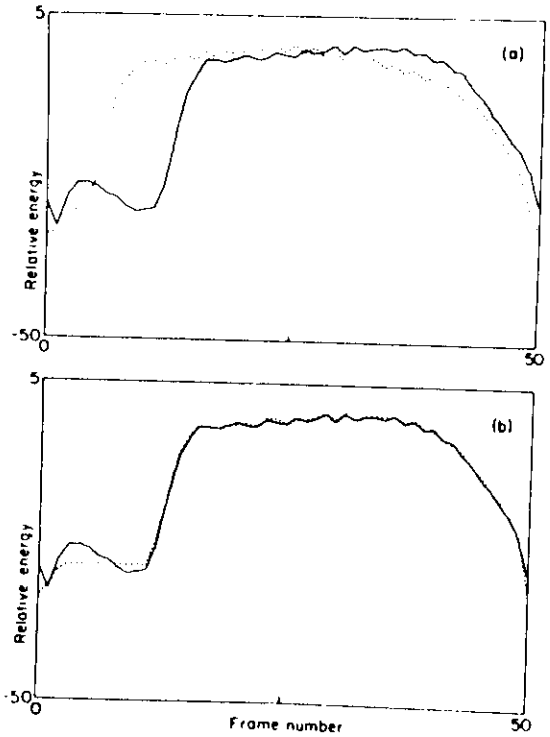


Figure 8. Time alignment of two utterances of the Hindi digit 2 (/do/) using (a) linear time warping, and (b) nonlinear dynamic time warping. Energy contours of the first and second utterances are shown by the solid and dotted lines, respectively (relative energy - dB). (Source: Paliwal *et al.* 1982a.)

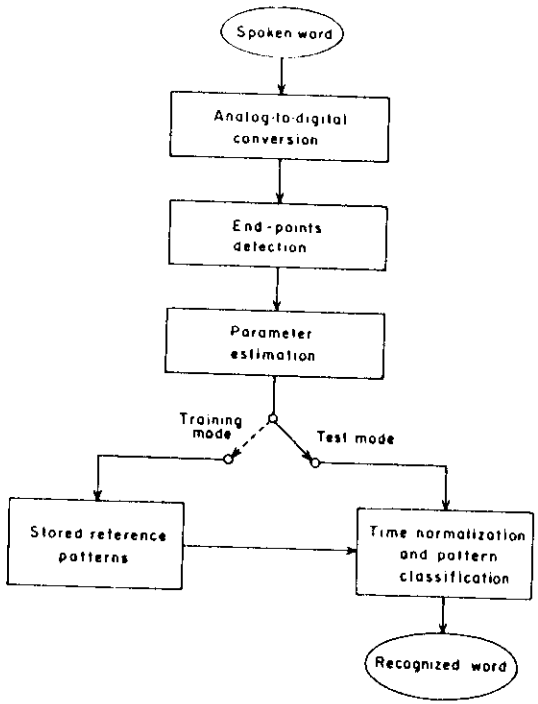


Figure 9. Block diagram of isolated-word recognition system.

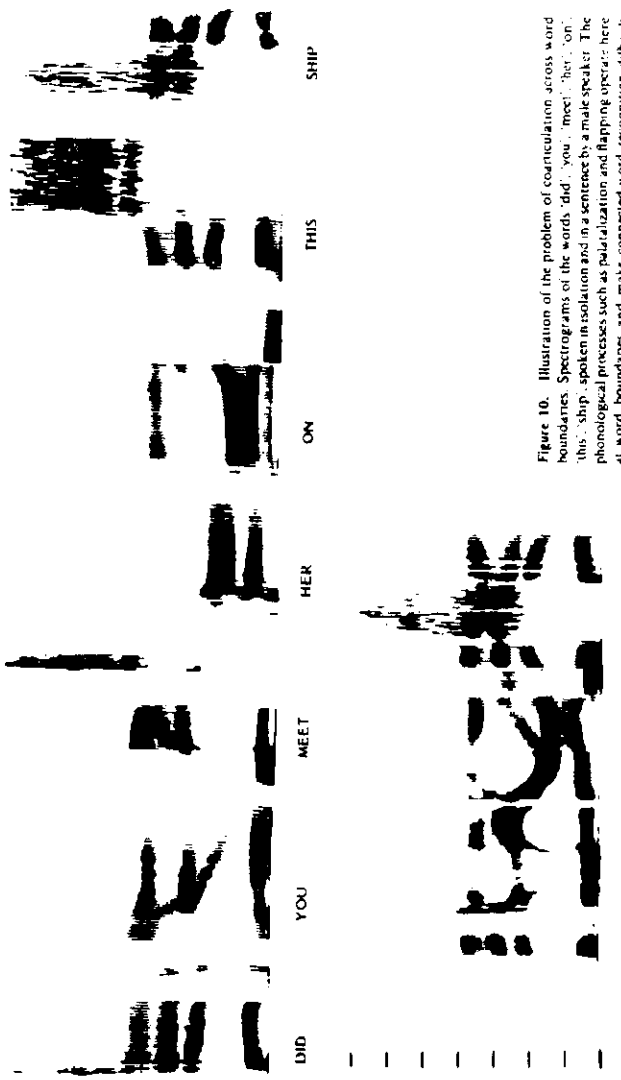


Figure 10. Illustration of the problem of coarticulation across word boundaries. Spectrograms of the words 'did', 'you', 'meet', 'her', 'on', 'this', 'ship', spoken in isolation and in a sentence by a male speaker. The phonological processes such as palatalization and flapping operate here at word boundaries and make connected-word recognition difficult (Source: Zue 1983).

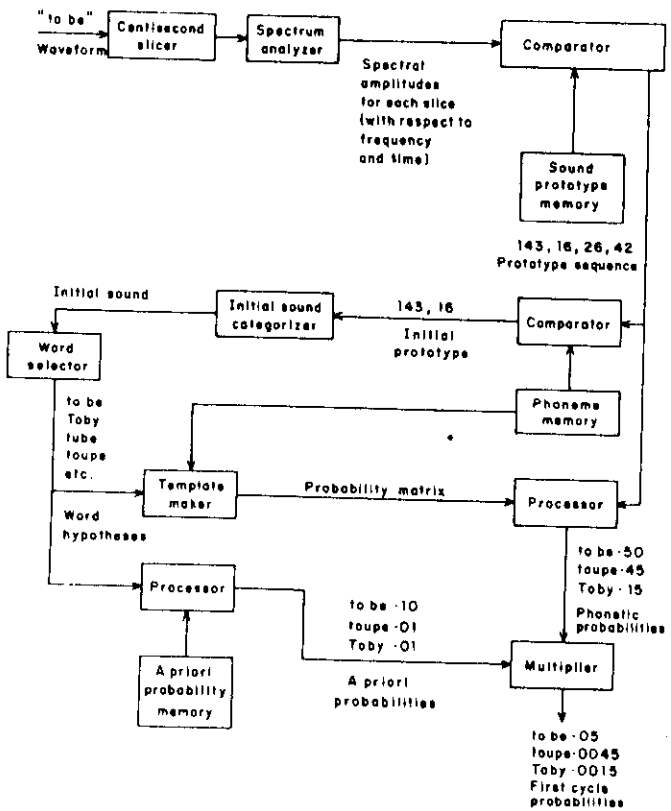


Figure 11. IBM speech recognition system. Incoming digitized signals are broken into centisecond slices and spectrally analyzed by the system. Each slice is compared with a collection of sound prototypes and the prototype closest to each slice is entered into a sequence. The prototype sequence is then used to roughly categorize the initial sound of the word, which in turn is used to produce word hypotheses. Each word is then tested by creating a probability matrix that determines the probability that a given prototype sequence is actually that word. From this matrix and the actual prototype sequence input, a phonetic probability for each word is arrived at. Simultaneously, a *a priori* probability for the word existing in a given part of a sentence following previously hypothesized words, is arrived at. The two probabilities are multiplied for each word hypothesis to give a ranking of overall probabilities. The cycle then repeats for the next word until an entire sentence is identified. (Source: Reddy & Zue 1983.)