A Software Speech Recognition System Using a Phonetic Approach

L Robert H Everson

B Ing, Stellenbosch

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Engineering, in the Department of Electronic Engineering, University of Natal, South Africa.

Durban

April 1985

Preface

The author hereby declares that this thesis represents his own original and unaided work except where specific acknowledgement is made by name or in the form of a reference. The thesis has not been submitted to any other university for degree purposes.

Abstract

Computer speech recognition techniques were investigated. This investigation included a study of the hearing and speech process. An algorithm was developed that used nine features to identify the phonemes in speech signals.

Two of these features, the total energy and the number of zero crossings in a specific section of the speech signal, were obtained directly from the digitized speech signal. The other features, frequency energy bands and formant frequencies, were measured from a spectral analysis of the signal.

A Hewlett Packard mini-computer was used for the development of the necessary software in FORTRAN. For the testing of the algorithm ten words, "zero" through to "nine" were used.

Acknowledgements

The author especially wishes to thank Professor Anthony D Broadhurst of the Department of Electronic Engineering, University of Natal, under whose supervision this work was performed. Also, Mr D C Levy for his supervision of the project whilst Professor A D Broadhurst was away on subbatical leave.

The author would also wish to thank the following people and organisations:

a) The Council for Scientific and Industrial Research and the University of Natal for their financial support;

b) Messrs D S Peddie and R C S Peplow for their assistance in using the Department's computers;

c) Messrs D Beard and T Watts for their help in using the Department's equipment;

d) Mr P Strong for his advice and the use of one of his programs;

e) Mrs P Paul for her assistance with the grammatical presentation of the thesis;

f) his mother for her help with the drawings in, and the layout of the thesis;

g) and his family and friends for their encouragement and support.

Contents

5 RECOGNITION RATES OF VARIOUS SECTIONS OF THE SYSTEM

6 CONCLUSION AND RECOMMENDATIONS

APPENDICES

Symbols

A	reference pattern
a	LPC coefficient vectors of the reference pattern
a'	differential of a
B	test pattern
b	LPC coefficient vectors of the test pattern
b'	differential of b
BW	frequency range of the energy-bands in Hz
$C(A,B)$	correlation alignment function
$D(A,B)$	dynamic programming alignment function
$d(A,B)$	distance measurement
E_i	the i^{th} freqency energy-band
f_n	the n^{th} formant frequency
$h(n)$	window function
S	sampling rate
$SM2_j$	sum of the squares of the j^{th} features of a template
SM_j	sum of the j^{th} features of a template
T_k	the k^{th} training pattern
TE_i	total energy of the i^{th} frame
V'	matrix of the auto correlation coefficient
W_i	weight vector of the i^{th} reference template
$w(j)$	warping function
$X(k)$	power spectrum of $y(n)$
$x(n)$	input speech signal
$Y(k)$	Fourier representation of $y(n)$
$y(n)$	windowed speech signal
ZC_i	zero-crossings of the i^{th} frame

Abbreviations

ADC	analog to digital converter
CNT	counter
CSR	continuous speech recognition
CWR	connected word recognition
DC	direct current
dev	deviation
DFT	discrete Fourier transform
DMA	direct memory access
DP	dynamic programming
DTW	dynamic time warping
EL	error-limit
Eq.	equation
f	formant frequency
FFT	fast Fourier transform
Fig.	figure
HE	high-frequency energy-band
HME	high mid-frequency energy-band
IWR	isolated word recognition
LE	low-frequency energy-band
LHE	low high-frequency energy-band
LPC	linear predictive coding
max	maximum
ME	mid-frequency energy-band
min	minimum
ROM	read only memory

SI	silence
SM	sum
SM2	sum of the squares
TE	total energy
U	unvoiced sounds
UF	unvoiced fricatives
US	unvoiced stops
V	voiced sounds
VE	voiced-frequency energy-band
VF	voiced fricatives
VHE	very high-frequency energy-band
VL	vowel-like sounds
VO	vowels
VS	voiced stops
ZC	zero-crossings

1 Introduction to Speech Recognition Systems

1.1 Introduction

This thesis reviews some speech recognition techniques and describes the development of a speech recognition system. Speech recognition is the recognition of speech either by human beings or by machines. In order to understand the mechanisms used in speech recognition, it was necessary to study the speech process. Speech is a form of communication between human beings which involves the generation and reception of complex acoustical signals (Fig. 1.1), (1).

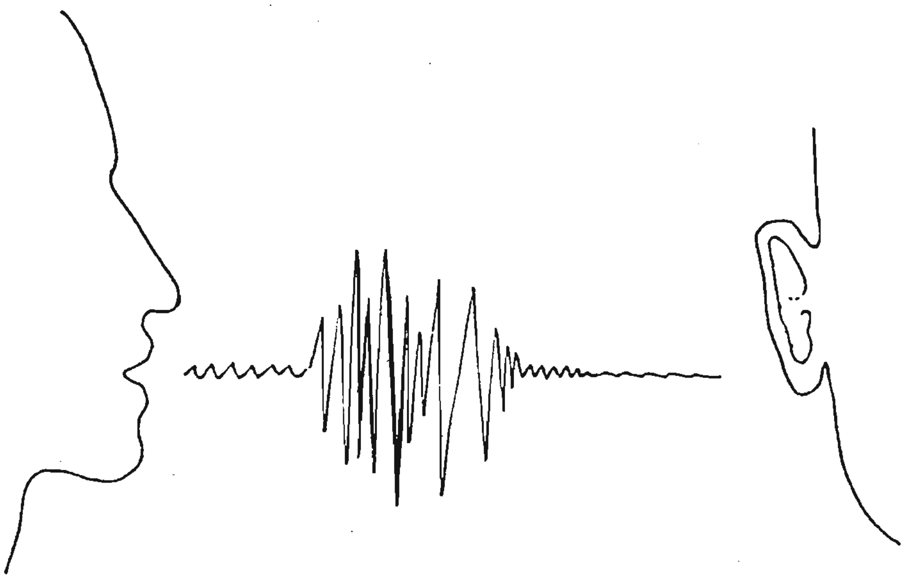


Fig. 1.1 The communication between two humans

If the generation of these speech signals is achieved by a machine, the speech sounds are called speech synthesized sounds. The recognition of speech sounds may be thought of as a decoding operation which takes place in the human ear and the brain. The decoding operation can be described as the identification of the different speech sounds and the understanding of the messages conveyed in the speech. Speech can also be recognized by limited extended computer systems and this is called Automatic Speech Recognition.

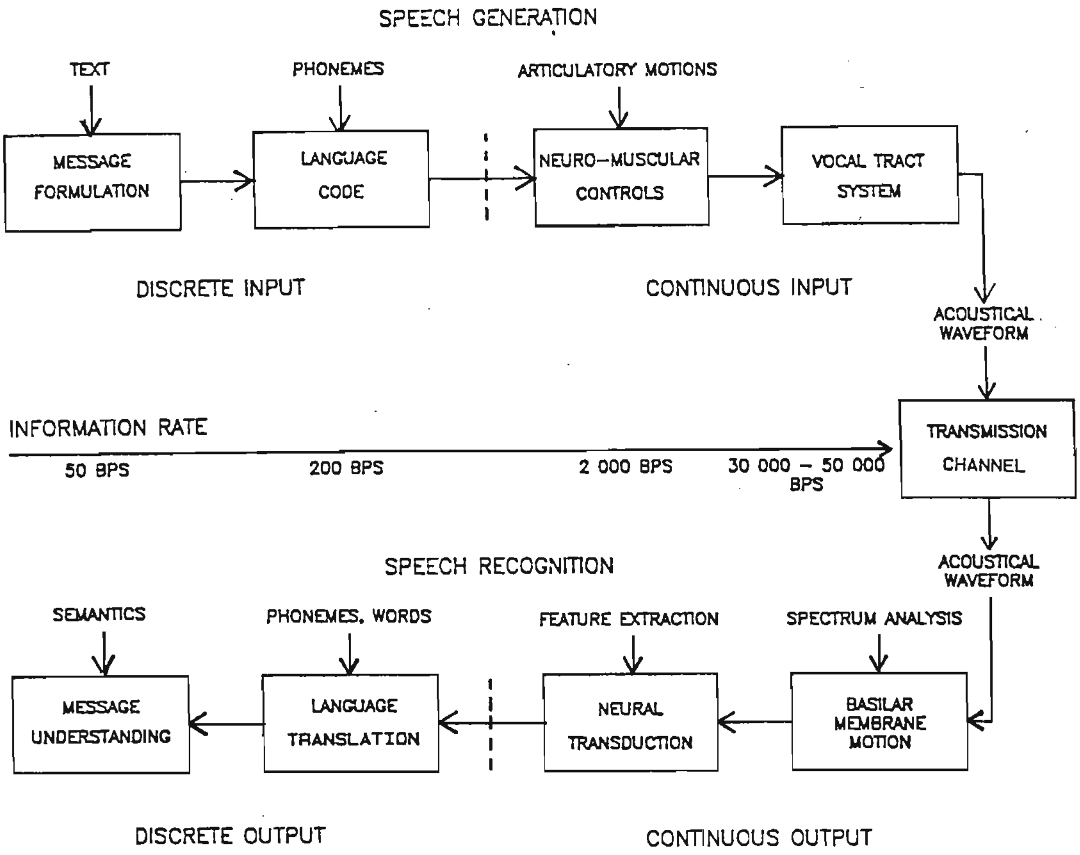


Fig. 1.2 Schematic representation of the speech chain (2)

Fig. 1.2 gives a schematic representation of human speech generation and recognition together with the relevant information rates in terms of bits per second. Simulating the two continuous output sections on a computer is not difficult to achieve, but the language translation section and the understanding of the message needs high intelligence as it takes place in the cerebral cortex of the brain. Accomplishing this with a computer requires recognition of phonemes, words, sentences and semantics, which is no easy task.

1.2 State of the Art

A number of speech recognition systems are already commercially available. They include speaker-dependent and independent systems. However the speaker independent systems can only be used with small vocabularies (Chapter 3).

a) Bell Laboratories has been very active in the researching of speech recognition systems. A system has been developed which can recognize isolated words, from sentences of a single speaker who pauses between words. This system, which is used for the retrieval of airline timetable information , can recognize 127 words, with a 98% sentence recognition rate (2). However, it is still under development and only a fast signal processor which is used as part of the system is available on the market (3).

b) Interstate Electronics is one of the bigger suppliers of isolated word recognition systems. This company offers two systems, a single chip, VRC008, 16-word vocabulary speaker-independent system and a two chip, VRC100-1, speaker-dependent 100-word vocabulary.

Both these systems are customized (made to the customer's specification) by Interstate Electronics. The vocabulary is stored in an external ROM. The word recognition accuracy is claimed to be better than 99% for the speaker-dependent system and 90% for the speaker-independent system. Interstate is also able to supply companies with single-board recognition modules which are compatible with DEC software and which can interface with any other system with its serial RS232 C interface (4).

c) The Nippon Electronic Company (NEC) is also engaged in research and development of speech recognition systems. Under development are a number of systems which include an unlimited Japanese vocabulary, speaker-dependent system and a 128-word vocabulary speaker-independent model. NEC has a number of special purpose signal processors on the market including a NMOS processor with a build in ROM, which does signal processing like FFT and LPC algorithms. These programs are stored in ROM (5), (6).

d) Texas Instruments has designed a voice recognition and speech-synthesis board aimed at personal computers. The board, SBSP3001, is built around the high-speed TMS320 single-chip signal processor. The system is a non-customized product and is trained by the speaker who is going to use the system. A maximum of 32 seconds of speech can be trained, which allows about 50 utterences. The recognition rate of the system is more than 98% (7).

e) The Verbex model 1800 is a telephone data entry and retrieval system. The 1800 accommodates eight users and can recognize the ten words "zero" through to "nine" and the words "yes" and "no" of nearly all the American dialects. The recognition vocabularies can be customized as required for individual applications. The vocabulary can be expanded up to 50 words (8).

Other systems available are the T-500/580 range, which are voice data entry terminals (speaker-dependent, isolated-words) systems from Threshold Technology. The V10 range from Voicetex is a range of speech recognition and voice output systems for personal computers and the V1000 form Votan which is a speaker-dependent isolated-word recognition system (4).

1.3 Future Use

Speech recognition systems are not yet being widely used in the business world. Only companies with money and patience are willing to install speech recognition systems which are 98% reliable, and can only be used by one speaker who in most cases is required to speak in a slow computer-like manner with pauses between each word. As soon as the available systems are speaker-independent and can understand continuous strings of words they will become essential tools in working environments.

For example it would be very convenient if one could simply speak into a telephone giving the name of the person one wished to talk to. The telephone speech recognition system would then find the number of the person and then dial the number required.

The telephone airline timetable system presently under development at the Bell Laboratories will be in use in the very near future. With this system it will be possible to ask any information in connection with the the airline timetable, in sentences with pauses between each word. The system will then recognize certain words in the sentences and answer back, with speech synthesized sentences stored in memory (9).

A speech typewriter or automated dictaphone which can convert any speech to text of any speaker, might not happen in this century. However, a speech typewriter with a limited speaker trained word vocabulary will probably be available in the not too distant future. Speech typewriters could also be used to help physically disabled people, enabling a blind person to "talk" to a deaf person, (Fig. 1.3).

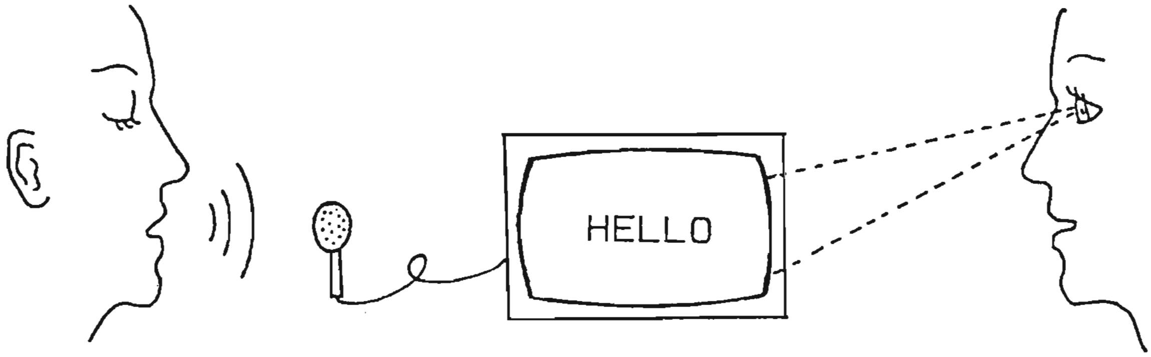


Fig. 1.3 Communication between two physically-disabled persons

The age old idea of human beings simply commanding a machine to do a task, seems to be possible and in some places is already in use. Although a machine which can hear, speak and understand as well as humans, will probably not be seen in the next ten years, it may become a reality before the turn of the century.

1.4 Purpose of the Project

The task of the author was to investigate speech recognition systems and to develop a system which might be used as a part of a speech typewriter. This investigation included a study of speech and hearing processes which helped to design a system which used minimum computer memory and minimum calculation time.

The system proposed by the author is a software speech recognition system which was developed on a mini-computer. A phonemic approach was used, which recognizes the phonemes in continuous phoneme strings (the different sounds in words). The author demonstrated that such a system is possible and although the recognition rate of the phonemes was much lower than 99% the word recognition rate of a limited vocubulary could easily be in the region of 80%. The author believes that this type of system, which uses a phonetic approach will in the end be the only method which can solve the problems of unlimited vocabularies systems.

The material in this thesis is divided into the following chapters.

Chapter 1 consists of the introduction.

Chapter 2 describes the speech and hearing processes.

Chapter 3 presents various different speech recognition methods.

In Chapter 4 the theory of Chapter 2 and Chapter 3 is used to develop a speaker-dependent, continuous phoneme recognition system.

Chapter 5 consists of the results obtained from tests done on the proposed system.

Finally the conclusion of this thesis is presented in Chapter 6.

2 The Speech and Hearing Processes

2.1 The Speech Process

Sounds are made by the expulsion of the air from the lungs. The air passes through the vocal system and the acoustic filtering behaviour of the system produces the speech sounds.

Speech consists of a sequence of different sounds. The transitions between these sounds serves as a symbolic representation of information. Phonetics is the study and classification of these sounds (symbols) (Apendix A), (10), (11), (12), (13).

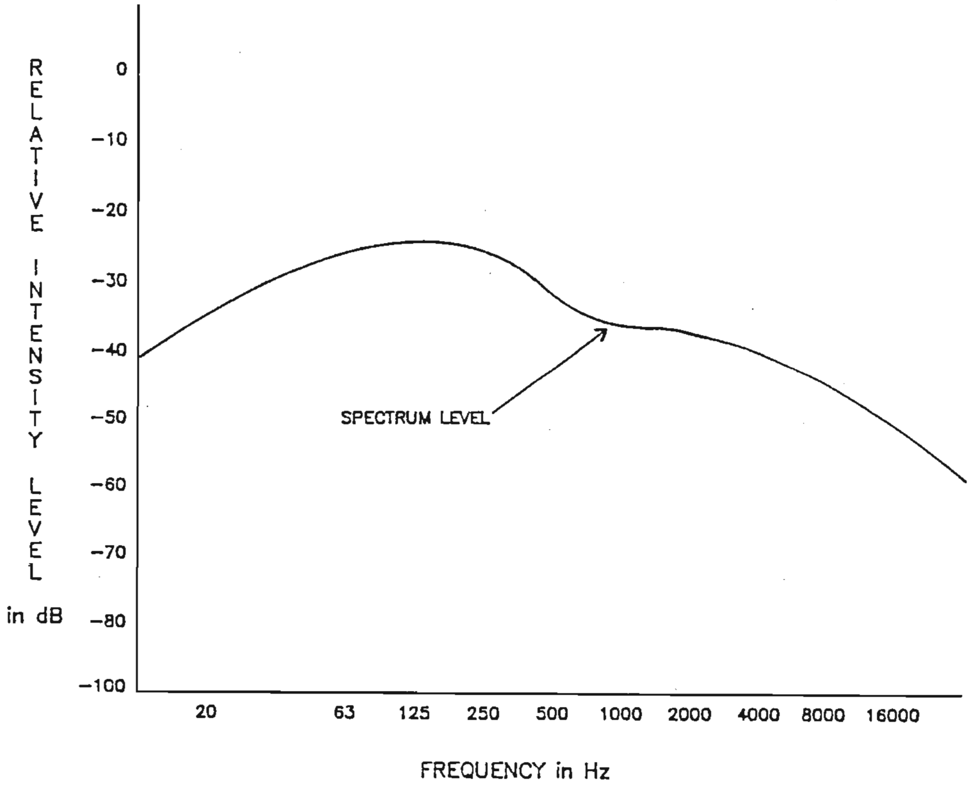


Fig. 2.1 Large time average spectrum of speech (10)

2.1.1 The Vocal System

An understanding of the acoustics of speech production can be indicated with the help of Fig. 2.2. The vocal organs are the lungs, the trachea (the windpipe), the larynx (containing the glottis, the vocal cords), the pharynx (the connection from the esophagus to the mouth, the throat), the nasal cavity (the nose), and the oral cavity (the mouth). The vocal cords are folds of ligament and are about 18 mm long. The space between the vocal cords is called the glottis and is typically 5 mm^2. Together, these organs form an intricately-shaped acoustic tube extending from the lungs to the lips.

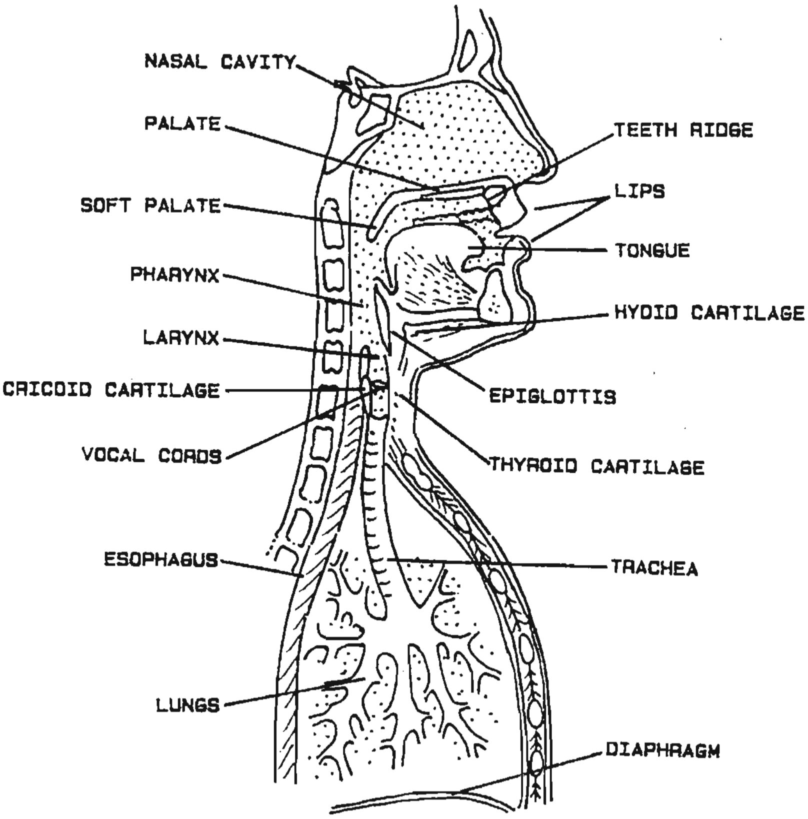


Fig. 2.2 The vocal organs

2.1.1.1 The Vocal Tract

The vocal tract is a nonuniform acoustic tube beginning at the opening between the vocal cords and ending at the lips. The total length of this acoustical tube is about 17 cm, in the average male and consists of the pharynx and the oral cavity. The oral cavity begins at the velums and terminates at the lips. The cross-sectional area of the tract varies from complete closure to about 20 cm^2, determined by the position of the tongue, lips, jaw and velum.

2.1.1.2 The Nasal Tract

The nasal tract begins at the velum and terminates at the nostrils. The tract is about 12 cm long and has a fixed volume of about 60 cm^3. The nasal cavity can be coupled to the vocal tract by the use of the velum (the soft palate).

2.1.2 The Speech Production

During speech, the diaphragm relaxes, and the degree of abdominal muscle contraction controls the extent to which the contents of the abdomen are pressed up against the diaphragm and carried into the chest cavity, where they squeeze the air out of the lungs.

Air from the lungs travels up the trachea, a tube consisting of rings of cartilage, and through the larynx towards the mouth and nose. The larynx acts as a valve between the lungs and the mouth.

2.1.2.1 Vocal Cords

The valve action of the larynx depends largely on the vocal cords which form a barrier. As the air pressure rises the air eventually blows the cords apart. Once apart the excess pressure is released, the cords return to their closed position, the pressure builds up again and the cycle is repeated. The vocal cords vibrate rhythmically, opening and closing as the air passes from the lungs to the mouth.

The frequency of the vibration is determined by how fast the cords are blown apart and the time taken to close again. The frequency is controlled by the size of the vocal cords, their tension and length. There is also the effect of low air pressure created in the glottis by air rushing through its narrow opening into the wider space above. Greater air pressure from the lungs enhances this effect and increases the frequency of vocal cord vibration. The range of these frequencies used in normal speech extends from 60 Hz to 350 Hz. Higher frequencies are occasionally used. Vocal cord vibration speech frequencies can cover about one and a half octaves.

The spectrum of the pressure waves created in the glottis due to its non sinusoidal and semiperiodic nature is rich in harmonics at the vocal cord frequency. The amplitude of these waves generally decreases as their frequency increases. In loud speech the higher harmonic amplitude increases and thus gives the sound a harsher quality.

The pressure waves generated by the vocal cords and the lungs are radiated by the mouth and nostrils as audible sound and the quality of this sound is changed by the configuration of the vocal tract and nasal tract. The shape and size of the pharynx changes during speech (Fig. 2.3). During non-nasal sounds the velum seals off the nasal tract and no sound is radiated from the nostrils.

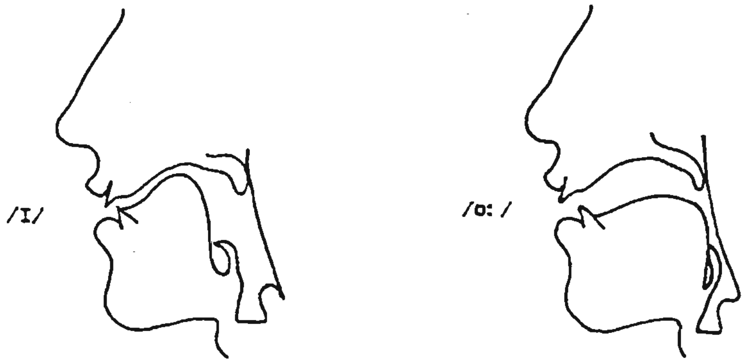


Fig. 2.3 Outlines of the vocal tract during the articulation of two vowels

2.1.2.2 The Mouth

The last and most important part of the vocal tract is the mouth. The shape and size of the mouth can be varied (more extensively than any other part of the vocal tract) by adjusting the relative positions of the palate, the tongue, the lips and the teeth.

The tongue is the most flexible, because its tip, edges and base can be moved independently. The tongue can move backwards, forwards and up and down. The lips affect both the length and the shape of the vocal tract and can be rounded or spread to various degrees. They can also be closed to stop the air flow altogether. The teeth also affect the vocal tract's shape. They can be used to restrict or stop the air flow by being placed close to the lips or the tip of the tongue.

2.1.2.3 Articulation of Speech Sounds

Speech sounds can be divided into four groups:

a) The voiced sounds are produced when the vocal cords vibrate because of the force of the air pressure from the lungs which produces periodic broad spectrum pulses. For example the "e" sound /E/ in "example".

b) Fricative sounds are generated by constricting the air flow in the oral tract (usually toward the mouth end) enough to produce an air turbulence. For example the "f" sound /f/ in "fricative".

c) Plosive sounds are made by blocking the air pressure somewhere in the oral tract and then abrubtly releasing the pressure. The airflow can be blocked by pressing the labial together or by pressing the tongue against the alveolar or the velar. For example the "p" sound /p/ in "plosive".

d) Click sounds are produced by blocking the vocal tract at two points and then sucking the air out between the two blocks and then re-opening the oral tract (14). These click sounds are not used in spoken English and therefore will not be discussed any further in this thesis.

It must be noted, however, that fricative and plosive sounds can either be voiced (V) or unvoiced sounds (U).

Thus by setting the shape of the vocal tract and its acoustic characteristics the vocal organs enable us to distinguish one speech sound (phoneme) from another.

2.1.3 Phonemes

Languages can be described in terms of a number of distinctive sounds called phonemes. There are more or less 47 phonemes in the English language. These phonemes include vowels, diphthongs, semivowels and other consonants (Appendix A), (14), (15).

2.1.3.1 Vowels

The enunciation of vowels (VO) can be described in terms of the tongue and lip position. All the vowels are voiced and the velum most of the time completely closes the nasal tract. The addition of a nasal quality to a vowel sound is not used to distinguish one English vowel from another.

Tongue positions used for making vowels are usually described by comparing them with the position used for making a number of references to cardinal vowels. The position of the tongue is described by specifying where the position of the highest part of the main body is. Cardinal vowels are a set of standard reference sounds whose quality is defined independently of language. There is no written definition of cardinal vowel quality possible, because the definition of quality is perceived only when listening to a trained phonetician making the sound.

A The Physical Description of vowels

There are four points to be considered in connection with the definition of vowel sounds.

i) Height of the tongue: The tongue may be in one of three positions in the mouth; high, mid or low. The higher the tongue the nearer does it approach to a consonant sound. Compare the "ee" sound /i:/ in "see" with the "y" sound /y/ in "yet". If the tongue is positioned for the vowel /i:/ and then raised slightly at the front until the sounds produces friction, the consonant /y/ is produced.

ii) Part of the tongue: In the formation of vowels, either the front or the back of the tongue may be used. In the first case the front of the tongue is raised towards the front of the palatum. The tongue forms a slope from the front to the back and the sounds made are called front vowels. When the tongue is retracted and is being raised at the back, the slope is from the back to front. The sounds made are called back vowels. There are also vowel sounds called flat vowels, because the tongue does not slope either way.

iii) Condition of the tongue: Vowels may be either tense or slack. In tense vowels the tongue is braced up so that there is a feeling of tension and it takes a somewhat round position. The slack vowels are produced when the tongue is relaxed and somewhat flattened.

iv) Condition of the lips: In standard English there are twice as many unrounded as rounded vowels. The lips do not seem to make much difference to the sound of a vowel although front vowels are usually made with spread lips and back vowels with rounded lips.

The schematic representation of the mouth and the position of the tongue for the English vowels are shown, and the eight cardinal vowels are shown by numeral, in Fig. 2.4 . Table 2.1 gives an example of each vowel (10), (15).

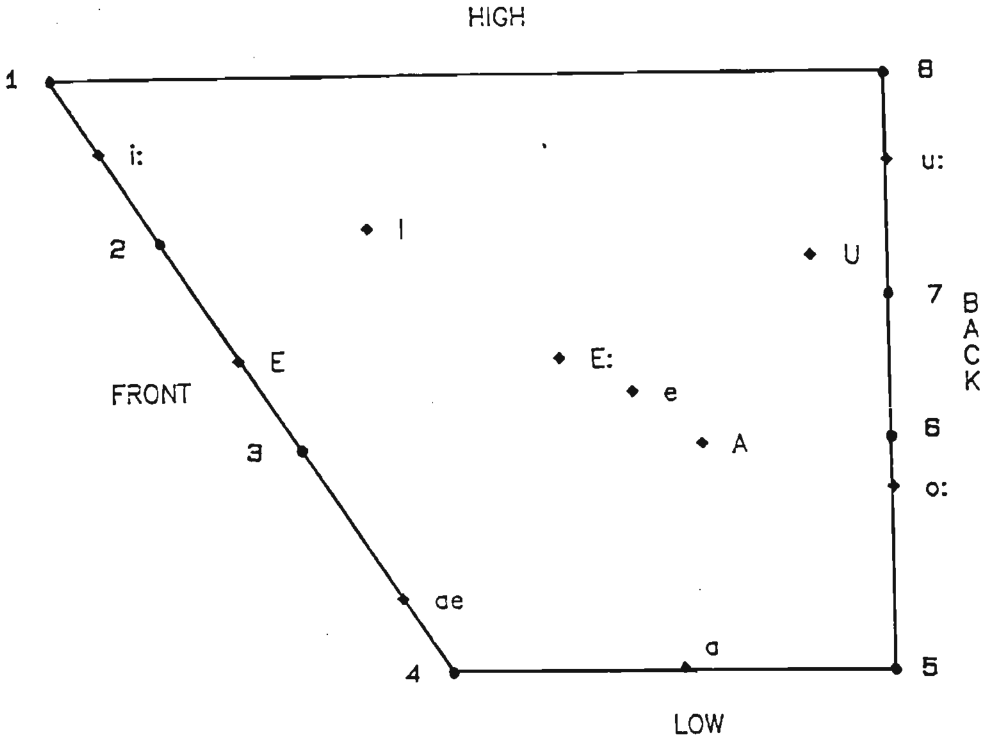


Fig. 2.4 Schematic representation of the position of the tongue for the vowels

PART	Front		Flat		Back	
CONDITION	Slack	Tense	Slack	Tense	Slack	Tense
POSITION						
	/I/	/i:/			/U/	/u:/
High	sit	seed			full	fool
	/E/		/e/	/E:/ /A/		
Mid	set		the	bird fun		
	/ae/		/a:/		/o/	/o:/
Low	sat		father		fog	fall

Table 2.1 Example of the part, condition and position of the tongue of each vowel (11), (15)

B The Frequence Spectrum of Vowels

The cross-sectional area of the vocal tract determines the resonant frequenies of the tract and thus the sound. Resonances of the vocal tract are called formants and their frequencies, the formant frequencies (f). Every tongue position in the vocal tract has its own set of formant frequencies. Thus, each vowel sound can be characterized by the vocal tract configuration and also by the formant frequencies. Peterson and Barney (13) have measured the formant frequencies of most vowels. Their results are shown in Fig. 2.5 and Table 2.2 .

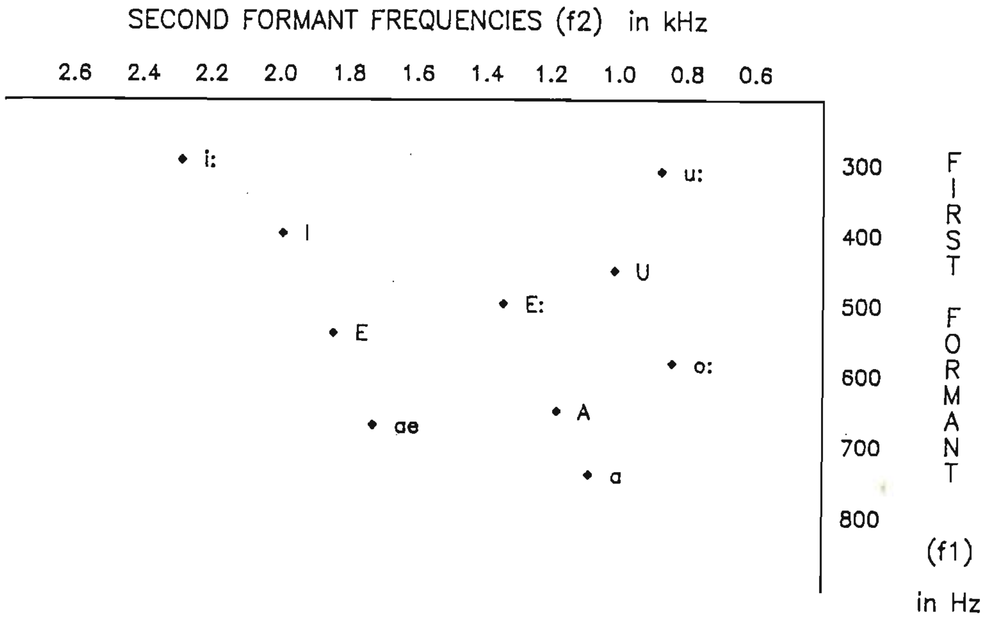


Fig. 2.5 Formant frequencies for vowels (13)

Phoneme	f1	f2	f3
/i:/	270	2290	3010
/I/	390	1990	2550
/E/	530	1840	2480
/ae/	660	1720	2410
/A/	640	1190	2390
/a:/	730	1090	2440
/o:/	570	840	2410
/U/	440	1020	2240
/u:/	300	870	2240
/E:/	490	1350	1690

Table 2.2 Average formant frequencies for vowels

If Fig. 2.4, which shows the tongue position, is compared with Fig. 2.5, which shows the first and second formant frequencies, it is very clear that the tongue position can be described in terms of the first two formant frequencies of the vowel pronouced.

2.1.3.2 Diphthongs

Diphthongs are produced by varying the vocal tract smoothly between two vowel configurations. There are about five diphthongs in English which may be classified into two classes. Table 2.3 shows all the diphthongs (11).

I-Diphthongs		U-Diphthongs	
/eI/	pay		
/aI/	buy	/aU/	out
/oI/	boy	/eU/	slow

Table 2.3 The two classes of diphtongs (11)

Diphthongs are voiced and non-continuant sounds. If the vocal tract configuration is time varying, the sound produced is a non-continuant sound.

2.1.3.3 Consonants

Consonants consist of voiced sounds, plosives and fricatives. In the production of consonant sounds, the oral tract is constricted, by either the tongue, teeth or the lips. A classification of all English consonants, according to place- and manner of articulation, is given in Table 2.4 .

PLACE of ARTICULATION	MANNER of ARTICILATION						
	Plosives		Fricatives		Glides	Liquids	Nasals
	V	U	V	U			
Labial	/p/	/b/			/w/		/m/
Labio-Dental			/f/	/v/			
Dental			/th/	/the/			
Aveolar	/t/	/d/	/s/	/z/	/y/	/l/ /r/	/n/
Palatal			/sh/	/zh/			
Velar	/k/	/g/					/ng/
Glottal			/h/				

V - Voiced U - Unvoiced

Table 2.4 Consonant and semi-vowel classification (10)

A Semivowels

This group of sounds is called semivowels because of their vowel-like (VL) nature. They are best described as transitional, vowel-like sounds and are similar in nature to the diphthongs. All the semivowels are voiced and can be divided into two classes.

i) The glides are produced by keeping the vocal tract briefly in a vowel-like position and then changing it rapidly to the position required for the the following vowel in the syllable.

ii) The liquids are made by putting the tip of the tongue against the gums and allowing the air to pass on either side of the tongue. In the case of the "r" sound /r/ the air pressure pushes the tongue slightly away from the gums.

B Nasals

Nasal consonants are also vowel-like sounds (VL), but the velum is lowered and thus lets sound radiate from the nostrils. The oral cavity is acoustically coupled to the nasal cavity and serves as a resonant cavity. The resonant frequencies of the oral cavity appear as anti-resonances (zeros). The nasal consonants are characterized by resonances which are spectrally broader than those for vowels. The reason for these broader resonances is that the nasal cavity has a relatively larger ratio of surface to cross-sectional area. Table 2.5 shows the half-power bandwidth in Hz of the formants for various intervocalic nasal consonants (16).

FORMANTS	/m/	/n/	/ng/
f1	60	40	80
f2	60	100	100
f3	100	110	230

Table 2.5 The half-power bandwith in Hz of the formants for various nasals

C Plosives

Plosives (stop consonants) can be classifed into two sections: voiced stop consonants (VS) and unvoiced stop consonants (US). These consonants are transient, non-continuent sounds which are produced by building up air pressure behind a total constriction somewhere in the oral tract and then suddenly releasing the pressure. The constriction can be achieved by pressing the tongue either against the lips, gums or the soft palate see Table 2.4 .

i) Voiced plosives: During the period when there is a total block in the tract no sound radiates from the lips. However, a small amount of low frequency energy radiates from the vocal cords through the walls of the throat. Voiced plosives are highly influenced by the vowel which follows the consonant, because of the dynamical nature of the voiced consonant.

ii) Unvoiced plosives are similar to their voiced counter-parts. Because the vocal cords do not vibrate, the period folllowing the closure consists of a brief interval of friction followed by a period of aspiration before voiced excitation begins.

The period of friction is due to the sudden turbulence of the escaping air and the period of aspiration is produced by the steady air flow from the glottis exciting the resonances of the vocal tract. The frequency content of the friction noise and aspiration varies greatly with the unvoiced plosives.

D Fricatives, Affricates and Whispers

i) Fricatives are produced by exciting the vocal tract with a steady airflow which became turbulent in the neighbourhood of a constriction in the oral tract. These sounds also have two classes: the voiced fricatives (VF) and the voiceless fricatives (UF). The voiced fricatives are produced by two excitation sources, because the vocal cords vibrate as well. Thus we can expect two distinct frequency components.

From the Table 2.4 it is clear that the place of articulation serves to determine which sound is produced. As in the case of nasal sounds, fricatives are influenced by two cavities. The tongue divides the oral cavity into two separate cavities. From the front cavity sound is radiated from the lips while the back cavity introduces an anti-resonance (17).

ii) Affricates. The consonants /tf/ (the "ch" sound in "chew") and /dz/ (the "j" sound in "judge") are not true fricatives. They are affricates and can be described as the linking of a stop and a fricative.

iii) Whispers. The phoneme /h/ (the "h" sound in "hear") is produced by a steady air flow which became turbulent in the region of the glottis without the vocal cords vibrating. In the production of this consonant the vocal tract assumes the characteristics of the following vowel. These types of excitation are also used for whispered speech.

Fricatives, affricates and whispers have a large amount of high frequency energy.

2.2 The Hearing Process

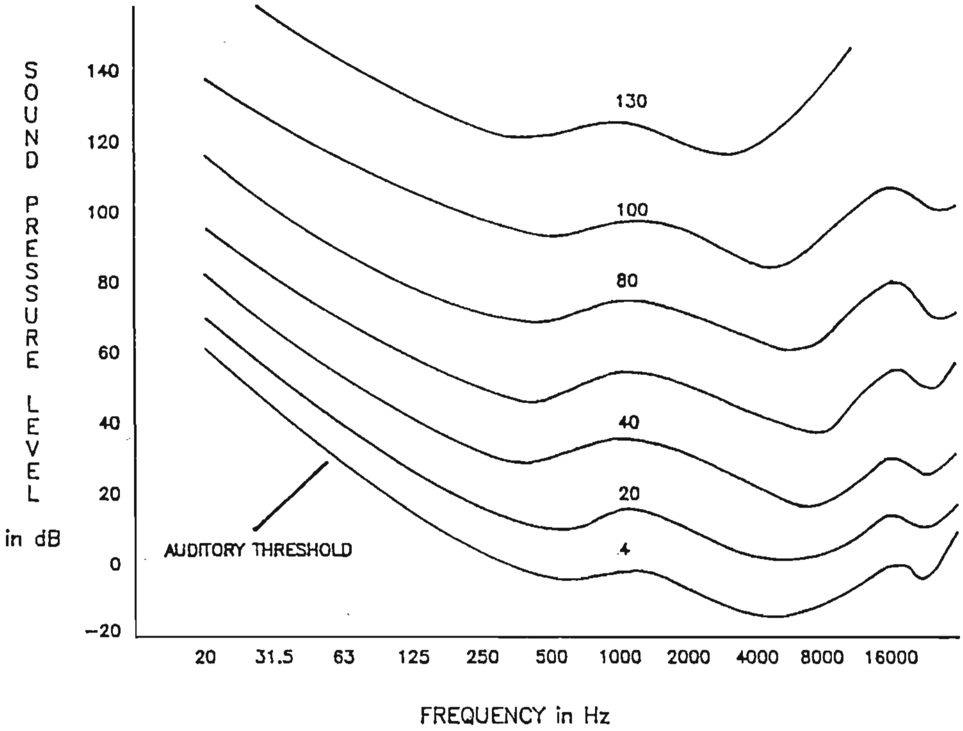


Fig. 2.6 Curves of equal loudness level

2.2.1 The Human Ear

The ear is conveniently divided into three sections: the outer ear, the middle ear and the inner ear (Fig. 2.7), (18), (19), (20), (21), (22).

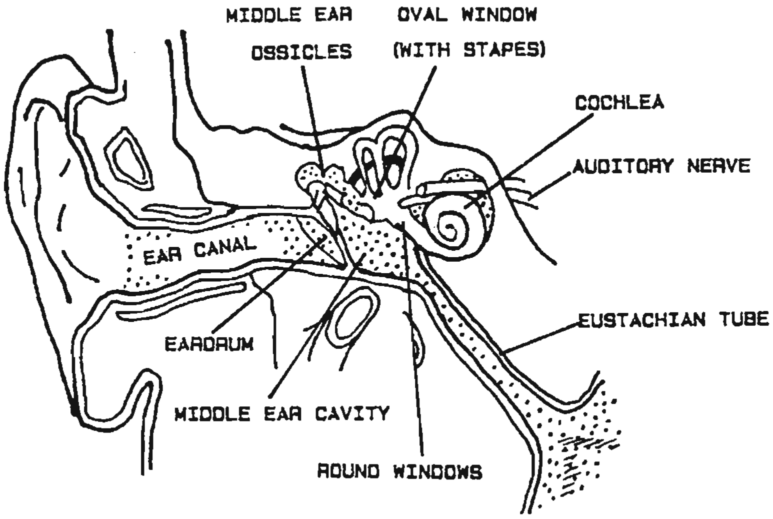


Fig. 2.7 The hearing organs

2.2.1.1 The Outer Ear

The outer ear plays a minor role in the hearing process and consists of the outer visible portion of the ear (the pinna) and the air-filled passageway (the auditory canal or meatus). The function of the outer ear is to increase the air pressure variations produced by any sound. Any sound wave reaching the ear is guided down into the head along the auditory canal to the eardrum, which then starts to vibrate.

The auditory canal is about 25 mm long and 8 mm in diameter. This canal is open at the pinna and terminates at the eardrum (the tympanic membrane). The meatus is an acoustic resonator and tends to amplify frequencies close to its resonant frequency. This resonant frequency is more or less between 3 000 and 4 000 Hz and amplifies sounds at these frequencies by about 12 dB.

2.2.1.2 The Middle Ear

The middle ear consists of three small bones: the malleas, the incus and the stapes. These bones are also called collectively the auditory ossicles. The auditory ossicles form a mechanical linkage between the tympanic membrane and the inner ear.

The malleus (hammer), which is connected to the tympanic membrane, transmits the vibrations from the eardrum to the inner ear with the help of the incus (anvil) and the stape (stirrip). The sound waves arrive in the form of amplified mechanical energy at the oval window, which is the entrance to the inner ear and is covered by the stapes.

The middle ear can be compared with an electronic pre-amplifier and thus performs two major functions.

a) It matches the acoustical impedance of air to that of the oval window. If a sound wave in air were to arrive at the oval window directly, almost all the incident energy would be reflected. The middle ear accomplishes amplification in two ways.

   i) The ossicles behave like a lever mechanism. The inner end (stapes footplate) of the ossicle lever moves through a shorter distance but exerts a greater force than the the outer end. These bones double the force of the vibrations of the eardrum.

   ii) It is the size of the oval window which produces an increase of amplification needed to match the impedance between sound waves in the air and the sound waves in the cochlea fluid. The area of the oval window is about thirty times smaller than the area of the eardrum. When force is transmitted from a greater to a smaller area the force is always increased proportionally.

   The combined effect of the two methods makes the pressure at the oval window about 60 times greater (35 dB) than it would be if the eardrum and the ossicles were not present.

b) The second function of the middle ear is to protect the inner ear from extremely loud sounds. Loud sounds trigger two sets of muscles. One pulls the stapes away from the oval window and the other tightens the eardrum and thus restricts its ablilty to vibrate. The muscle contractions also cause the ossicle to shift into a second mode of vibration in which the stapes axis of rotation changes to decrease the pressure variations transmitted to the inner ear.

2.2.1.3 The inner ear

The inner ear is perhaps the intricate section of the ear. This is the section where a spectral analysis is performed on the sound wave (Fig. 2.8).

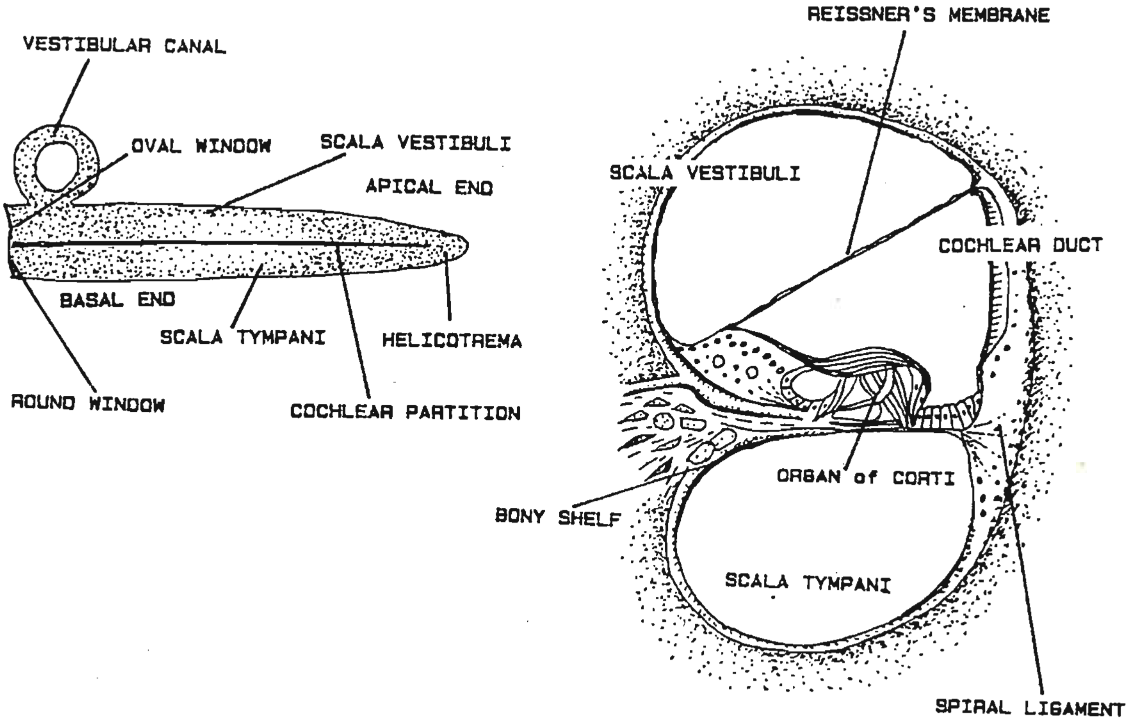


Fig. 2.8 The inner ear

The important transformation from mechanical vibrations to nerve impulses takes place in the cochlea. The cochlea is divided into two distinct regions by a membrane called the cochlea partition. The two regions are called the scala tympani and the scala vestibuli. The tympanic canal ends at the round window, a membrane-covered opening that leads back into the middle ear. This end is called the basal end. At the other end of the canal (the apical end) the two cavities are connected and a fluid (periymph) passes freely between the two canals. The vestibuli canal ends at the oval window at the basal end of the cochlea. The cochlea partition mentioned earlier consists of a hollow centre, the cochlea duct, which is filled with fluid called endolymph. The basilar membrane separates the tympanic canal from the duct. The scala vestibuli and the duct are separated by the Reissner's membrane. The basilar membrane is at its narrowest at the basal end, 0,04 mm, and about 0,5 mm wide at the apical end. This membrane is also more stiff and lighter near the oval window than at the helicotrema. When the oval window is set into vibration by the displacement of the stapes, the entire cochlea partition starts to vibrate. Sounds of various frequencies set the basilar membrane into varying degrees of vibration at different distances from the oval window (Fig. 2.9).

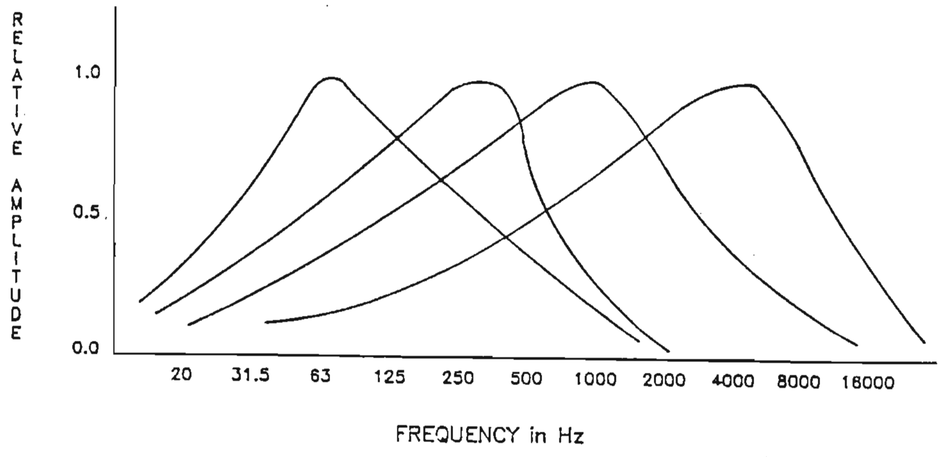


Fig. 2.9 Maximal amplitudes of the oscillation of the basilar membrane for sounds of different frequencies.

For higher frequencies the amplitude of the vibrations are highest at the stiffer and lighter section of the basilar membrane, while for lower frequencies, the point of maximum amplitude moves towards the broader and more elastic end of the basilar membrane. Thus the structure of the cochlea leads to a spatial separation of the maximum responses at different frequencies. Because the maxima are relatively broad the complete frequency analysis performed by the hearing mechanism, which is very selective analysis, cannot be accounted for by the formation of these maxima.

2.2.2 The Organ of Corti

The vibrations of the basilar membrane are converted into signals that are transmitted to the brain via the auditory nerve. This complicated process is perfomed by the organ of Corti (Fig. 2.10) .

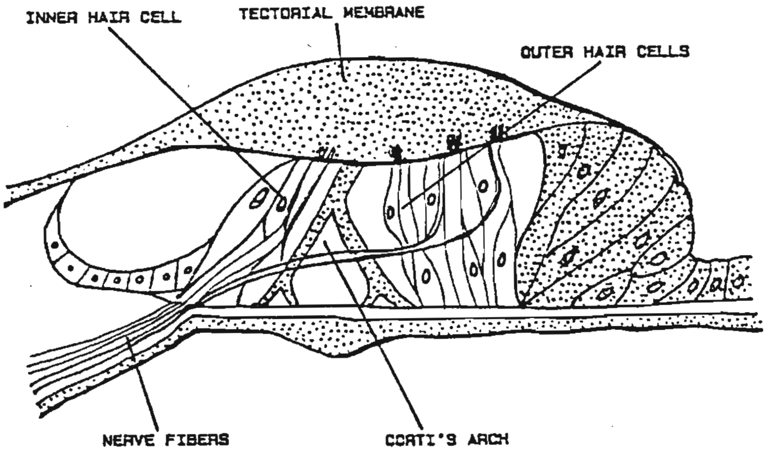


Fig. 2.10 The organ of Corti

The organ of Corti contains a number of small cells which are inside the cochlea duct and are supported by the basilar membrane. Fine hairs on these cells are the actual sensory receptors in the organ. These very fine hairs make contact with the tectorial membrane. There are two types of hair cells, the inner and the outer cells, and they are separated by the arch of Corti, which gives the organ its structural strength.

Movement of the the basilar membrane is transferred to the tectorial membrane. Because the hair cells are supported by the basilar membrane and the hairs rooted in the hair cells, the hair tops in contact with the tectarial membrane will be twisted, pulled and pushed as the two membranes vibrate. It is still not clearly understood how, but it is these physical stresses of the hairs in the hair cells which generate electrochemical pulses. Receptor cells, such as the hair cells, receive sensory information from their environment and help to code this information into the electrochemical pulses which are transmitted through the auditory nervous system to the auditory area of the cerebral cortex in the brain.

2.2.3 The Auditory Nervous System

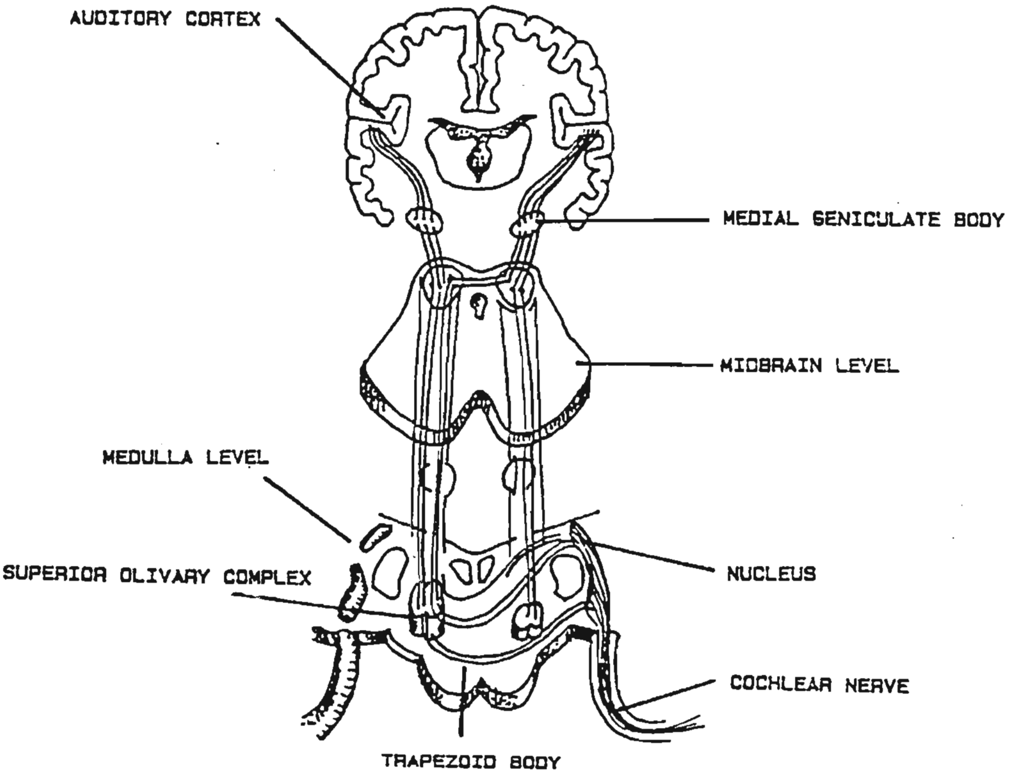


Fig. 2.11 The auditory nervous system

Nerve fibers extend all the way from the inner ear to the brain (Fig. 2.11). The auditory pathway starts at the auditory nerve. The auditory nerves are a broad bundle of about 30 000 individual fibers, which terminate in the cochlea, just above the basilar membrane. These fibers branch out into individual nerve cells (neurons) each one starting near a hair cell in the Organ of Corti.

Neurons are complex devices used for the detection and transmission of electrical signals. Each neuron picks up external stimuli, such as the hearing signal, from the hair cells. If such a stimulus is strong enough to activate the neuron, it produces an electrochemical pulse.

The auditory nerve enters the central nervous system in a complex interconnection called the cochlea nucleus. Each cell in the cochlea nucleus receives connections from many incoming fibers. From this nucleus, nerve fibers run in a bundle of nerves, known as the trapezoid body, to the superior olivary complex. The auditory pathway consists of a number of other bodies of cells which are used for the interconnections of the nerve fibers. The very last one of these bodies is called the medial geniculate body and from this point, the nerve fibers proceed directly to the auditory projection of the sensory cortex.

2.2.4 Information Processing in the Auditory pathway

According to W.D.Keidel, 1974 (23) some of the neurons in the higher levels of the auditory pathway activate on stimulus from sounds like vowels and consonants. The results of W.D.Keidel can be summarized as follows:

a) There exist neurons at geniculate level which respond to fundamental formant frequencies. Since a range of neurons is tuned in the same way to a fundamental frequency as vowels, these neurons could be considered to be vowel-detectors.

b) Another type of neuron fires when the stimulus frequency changes continuously from low to high frequencies and back. There are subgroups of neurons in this class which respond to a relatively slow frequency-modulation type of stimuli. These neurons can therefore correspond to phonemes which are transient speech sounds. Thus these neurons can be labelled as consonant-detectors or transient-detectors.

It is therefore believed that neurons at geniculate level can be used to detect and separate different phonemes in speech.

3 Computer Speech Recognition Techniques

3.1 Introduction

The complete successful recognition of spoken words in electronic speech signals will be one of mankind's greatest achievements. It is the subject of intensive research but nevertheless the goal has not yet been achieved. Perhaps the most successful results to date have been word recognition systems with systems that are typically reported to be able to recognize 1 000 different spoken words, (24). Many different speech recognition methods have been tried in the last two decades throughout the world (America, 1984 (24), England, 1975 (25), 1977 (26), India, 1982 (27), Japan, 1982 (28), Germany, 1983 (29), France, 1984 (30)).

Speech recognition may be classified into three major study fields:

a) isolated word recognition (IWR),
b) connected word recognition (CWR) and
c) continuous speech recognition (CSR).

The first method has been used for some time in commercial systems. Recently systems have been developed commercially which can recognize a string of words from a small vocabulary. However, continuous speech recognition systems are still the subject of intensive research, because of their enormous commercial potential. Each one of these fields can be subdivided into two categories according to the individuality of speakers, speaker-independent systems and speaker-trained systems, which are more commonly used in commercial systems.

3.2 Isolated Word Recognition Models

The main objective of any recognition system is to classify a pattern, in this case, the unknown digital speech waveform, into one of several possible classes (words). IWR (isolated word recognition) systems usually use words as the smallest recognition segment. Other systems use phonemes and sometimes syllables, for the recognition of words.

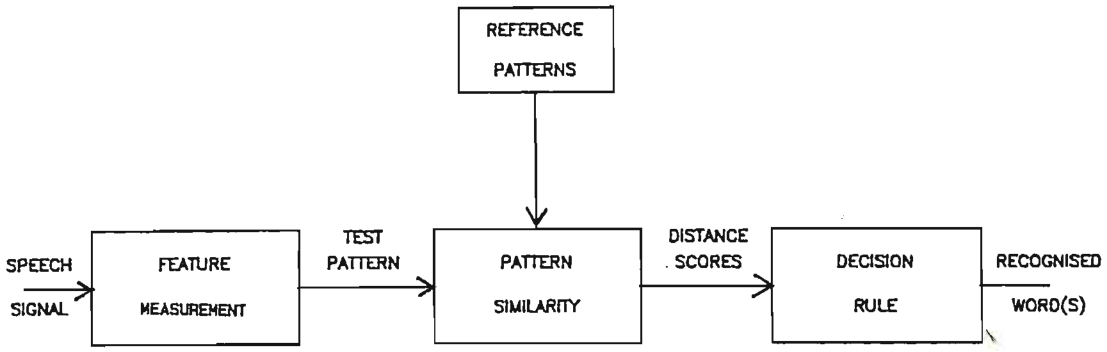


Fig. 3.1 Model of isolated word recognition

According to Levinson and Rabiner, (2) most of the isolated word recognition systems can be modelled with three basic building blocks (Fig. 3.1). First the input speech signal must be analysed and the necessary features extracted to obtain the test patterns. Secondly, these test patterns are matched to reference patterns to find pattern similarities. Finally, to recognize the actual word, distance scores, which are a measure of the pattern similarities, are used in the decision-making process of such systems. The distance scores are usually the difference between the numerical values of the test pattern and the values of the reference patterns (Section 3.2.2.2).

3.2.1 Feature Extraction

The key problem in speech recognition is the extraction of the best features. Mathematical and structural features are best suited for automatic computer speech recognition, although they may be quite different from features derived by the human recognition process, the hearing process (Section 2.2).

In selecting a certain set of features the following considerations must be optimized. The features must be computed in the shortest possible time, must use the smallest possible computer memory and must be easy to implement. Further, it must be noted that only a limited number of features are extracted for classification of the test pattern.

3.2.1.1 Frequency Spectrum Analysis

This type of feature extraction is probably the most common method because of its parallelism to the cochlea. The frequency spectrum of the input speech signal is discussed in this section.

a) A bank of analog filters,

b) or a series of digital filters covering the whole frequency range of the speech signal,

c) and a Fourier Transform can be used to obtain the power spectrum.

In the first two techniques, the analog and the digital methods, the speech signal is passed through a bank of bandpass filters which covers the frequency range of the speech signal, from just above 60 to 6 000 Hz (Fig. 2.1). The number of filters used in analysis can vary from 5 to as many as 35 filters (30). These filters can be spaced linearly in the speech frequency band, or the percentage bandwidth of the filters can be kept constant. In some cases the filters are spaced linearly up to more or less 1 000 Hz and from then on the percentage bandwidths are held constant. The output signal amplitude levels of the filters are usually derived or computed using non-linear operations and low pass filtering. Often the filtered signal levels are converted to a logarithmic scale to obtain the power spectrum level of the input speech signal.

The third method, which uses a descrete Fourier transform (DFT) to obtain the features, is also common in a number of recognition systems. Most of these systems use a radix-2 fast Fourier transform (FFT) algorithm (27), (31), which is a computationally efficient procedure, to calculate the DFT. The speech signal is digitized and then weighted over a certain duration with a window, commonly a Hamming or a Hanning window, to prevent leakage (32), (33). The durations typically vary from 10 ms to 25 ms . The FFT is normally performed on either 128 or 256 samples depending upon the sampling rate used. The logarithmic spectrum level in each of the frequency samples (64 or 128) is then computed to obtain the power spectrum of the input speech signal.

3.2.1.2 Linear Predictive Coding (LPC)

Linear predictive coding is a method of predicting the transfer function of a system which assumes that the output of the system is a linear function of the past outputs and the present and past inputs (34). In the case of speech signals, the transfer function of the vocal tract is computed and the linear predictor coefficients obtained are used as the recognition features of the speech signals. The speech recognition systems which use LPC, compare the measured speech signal with an artificially-generated speech computed from the output signal of an all pole digital filter whose input is then excited by either a random signal or a set of periodic inpulses (34). By minimising the sum of the squared differences (over a finite interval) between the actual samples and the linearly predicted samples, an individual set of predictor coefficients can be determined.

The excitation signal is assumed to be either quasiperiodic pulses for voiced sounds or random noise for unvoiced sounds (Fig. 3.2), (34), (35).

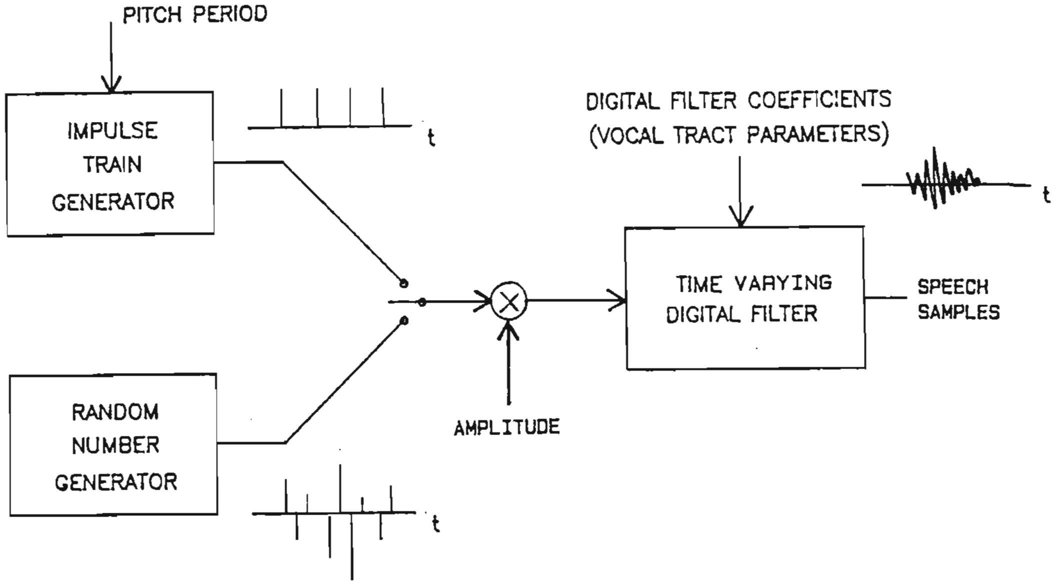


Fig. 3.2 The model of the vocal tract

As in the FFT analysis, a block processing model is used in which a frame consisting of a number of samples is processed at one time. The duration of the frame is typically 10 to 30 ms . To obtain the predictor coefficients the speech signal is first pre-emphasized to spectrally flatten the speech signal. Secondly the data is weighted with a window similar to the windows used for the FFT analysis. To find the predictor parameters, an autocorrelation analysis is performed on the window data (36).

3.2.1.3 Other Features

Sometimes in addition to above mentioned features zero-crossing rate (ZC) and total energy (TE) are calculated. Zero-crossing rate is defined as the number of zero crossings in a fixed frame length. Total energy is defined as the sum of the squared values of the speech waveform in a given frame, Table 3.1 (37).

	ZERO-CROSSINGS	TOTAL ENERGY (dB)
VOICED SOUNDS	12,8	51,5
UNVOICED SOUNDS	49,8	30,5
SILENCES	12,1	18,1

Table 3.1 Mean values of ZC and TE for room-quality speech (absolute silence is 0 dB and the frame length is 10 ms)

3.2.2 Pattern Similarity Determination

After the feature extraction is performed, a test pattern, consisting of a string of quantized numbers quantatively representing the extracted features, is computed for each frame. This test pattern, B, is then matched to a reference pattern, A, to determine the pattern similarity (Fig. 3.1). In the case of isolated word recognition the reference patterns and test patterns are usually the whole word rather than sections of the word like phonemes or syllables.

The speaking rates of words normally vary considerably and therefore pattern similarity involves time alignment. To find the minimum distance between the two patterns (words) an optimum time (warping) function, w(j), must be found (Fig. 3.3).

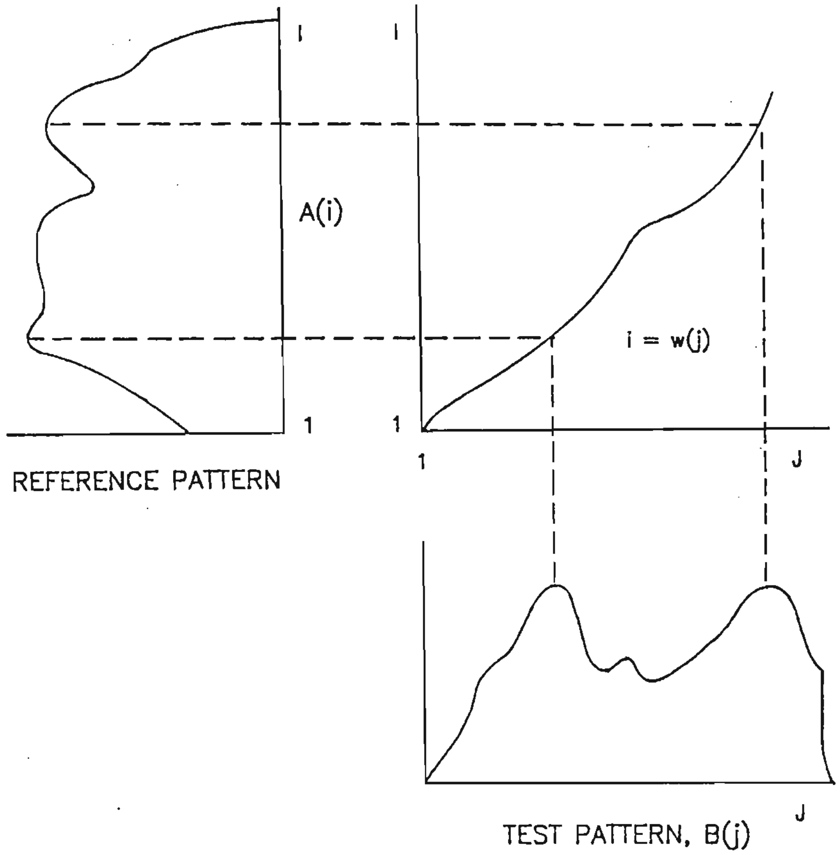


Fig. 3.3 Example of a warping function

3.2.2.1 The Warping Functions

The warping funtion, w(j), is a model of time-axis fluctuation in a speech pattern. Accordingly, it should approximate the properties of the actual time-axis fluctuation. In other words, function w(j), when viewed as a mapping from the time axis of pattern A onto that of pattern B, must preserve all the linguistically essential structures in pattern A time axis and vice versa (2).

A number of different techniques can be used to find the warping funtion.

a) Linear time alignment is used when the two patterns are linearly matched and the alignment function is a straight line:

$$i = w(j) = (j-1). \frac{(I-1)}{(J-1)} +1 \qquad 3.1$$

where j is the j^{th} sample of the test pattern B, i is the i^{th} sample of pattern A and J and I are respectively the maximum number of samples in pattern B and pattern A.

The other techniques involve a nonlinear time alignment.

b) Time event matching is the matching in time of significant events in the reference and test patterns. For each one of these events another warping function is computed which is typically a linear time alignment function.

c) Correlation maximisation is the warping function which is found when the correlation between pattern A and pattern B is maximized (Eq. 3.2):

$$C(A,B) = \max_{w(j)} \sum_{j=1}^{J} (A(j).B(j)) \qquad 3.2$$

d) Dynamic time warping (DTW), is the most important time alignment function for isolated word recognition and is sometimes called dynamic programming (DP), (40). Dynamic programming is the warping curve which is determined from the solution to the optimization problem:

$$D(A,B) = \min_{(k)} \left(\sum_{k=1}^{K} d(A(k),B(w(k))) \right) \qquad 3.3$$

where d(A(k),B(w(k))) is the distance between k^{th} frame of the test pattern B and $w(k)^{th}$ frame of the the reference pattern A, and w(k) is a function of i(k) and j(k) (Fig. 3.3).

3.2.2.2 Distance Measurements

A number of different distance measurements are available and depend heavily on the type of feature measured. In most recognition systems the distance calculation is one of the most time consuming computations of the system.

a) Euclidean distance is a very commonly used measurement and is of the form (27):

$$d^2(B,A) = \sum_{k=1}^{K} (B(k)-A(k))^2 \qquad 3.4$$

where A(k) and B(k) are k^{th} component of the reference and aligned test pattern respectively and K is the maximum number of features. This measurement requires K multiplications, subtractions and additions.

b) The Minimum Prediction Residual, Itakura (39) has proposed a remarkably useful distance measure for a feature set based on LPC parameters and is of the form:

$$d(A,B) = \log_{10} \frac{a.V'.a'}{b.V'.b'} \qquad 3.5$$

where a and b are the LPC coefficient vectors of the reference and test patterns respectively, a' and b' are respectively the time differential of a and b. V' is the matrix of the auto correlation coefficient (Section 3.2.1.2) of the test pattern. LPC distance measurements need a large number of additions and multiplication. However, the distance d can be rewritten in the form (39):

$$d(A,B) = d+\log_{10}((c.r)/(b.r)) \qquad 3.6$$

where r is the auto correlation vector $(v(i)/v(0))$, $(i=0, \ldots,p)$, $d=\log_{10}(b.b)$, and c is the vector $(1,c(1), \ldots,c(p))$, whose elements are defined by:

$$c(i) = 2. \sum_{i=0}^{p-i} a(j).a(j+i)/(a.a) \qquad 3.7$$

Using Eq. 3.6, the computation of the distance, reduce to only p+1 additions, multiplications and one log computation.

3.2.3 The Decision Rule for Isolated Word Recogniton

The final step in the recognition model (Fig. 3.1) is the decision rule operation. Here a reference pattern is selected which is the closest match to the unknown test pattern. There are a number of rules which may be used but the two most commonly employed are:

3.2.3.1 The Nearest Neighbour Rule

This rule can be explained as follows. Let there be N reference patterns, A_i ,i=1,2, ...N, and the distance score for each pattern calculated is D_i, from the pattern similarity algorithm. The test pattern, B, will be classified the same as the reference pattern, A_i, with the smallest distance score, D_i.

3.2.3.2 The K-nearest Neighbour Rule

The K-nearest neighbour rule is used when each isolated reference word is described by more than one reference pattern, as in the case where the words must be independent of the talker. In speaker-independent (Section 3.4) systems is it very common to have a template for a number of speakers and another template for other speakers although both templates are used for the same isolated reference word. Therefore there are M reference patterns for each of the N reference words. The j^{th} occurrence of the i^{th} reference pattern can be expressed as $A_{i,j}$, $1 \leq i \leq N$, $1 \leq i \leq M$, implying that the distance for the j^{th} occurrence of the i^{th} reference pattern is D_{ij}. The average distance of the K-nearst neighbour is then defined as:

$$d_i = (\sum_{k=1}^{K} D_{i,(k)})/K \tag{3.8}$$

In both cases, the nearest neighbour and the K-nearest neighbour, are the the test patterns classified the same as the reference patterns, A_i, with the smallest distance measurements to the test pattern, B.

3.3 Connected Word and Continuous Speech Recognition

Connected word recognition (CWR) and continuous speech recognition (CSR) systems are able to recognize a string of words instead of single isolated words. The model used to describe isolated word recognition (Section 3.2, Fig. 3.1), can also be used to model connected word and continuous speech recognition systems, with two modifications (Fig. 3.4): the incorporation of endpoint detection, and feedback frcm the decision rule secticn to any of the previcus sections of the model.

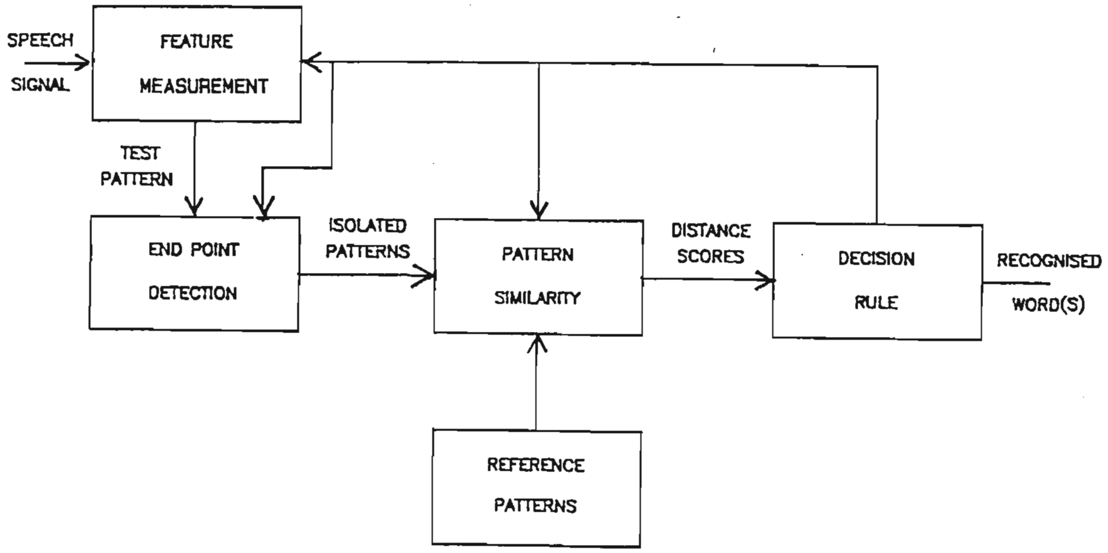


Fig. 3.4 Model of connected speech recognition

The operation of the connected speech recognition model can be described as follows.

a) The incoming string of words is analysed and a feature extraction performed.

b) The start and end of the words are located.

c) In the next step, as in isolated word recognition, these features are matched to reference patterns, previously obtained, to find the pattern similarities.

d) At this stage it is important to decide if the word can be matched to a word in the template. It is also possible to detect syntax and semantic errors in this section. Syntax is the branch of linguistics which deals with the grammatical arrangement of the words in sentences. Semantics is the branch of linguistics which deals with the study of the meaning and the principles that govern the relationship between sentences or words and their meanings (14).

e) Finally, if the decision rule decides that it is necessary to repeat any of the previous sections it will go back to that section and repeat the process, until the correct word is found.

3.3.1 Feature Extraction

Feature extractions in continuous speech recognition (CSR) systems is similar to the techniques used in in the isolated word recognition (IWR) systems (Section 3.2.1). However, most IWR systems extract their features from whole words, while CSR systems obtain their features from sections of the words (24), (25), (27), (41). This means that CSR systems use phonemes or syllables as building blocks for the recognition of the words and word strings. Hence, the period over which the features are extracted is shorter than in IWR systems. Most of the features are non-time-varying. The minimum length of a phoneme is about 30 ms, while the number of available samples for the extraction of the feature varies between 150 and 600 samples, depending on the sampling rate (between 5 000 and 20 000 Hz). This number of data samples is large enough to perform a frequency spectrum analysis.

It should be noted that CWR systems uses IWR reference patterns as reference patterns to recognize the test patterns. They therefore use exactly the same feature extraction techniques as the IWR systems.

3.3.2 Start and Endpoint Detection

Start and endpoints can be very easily detected in connected speech as the speaker pauses between each spoken word (9), (37), (38). These pauses can be detected as silences in the speech signal. The words between the pauses are recognized as isolated words which can be matched to a set of isolated word reference templates (36), (38), (42). This method is commonly used in CWR systems.

The CSR systems do not depend upon pauses in the speech, although the pauses do help in the decision rule section (Fig. 3.4). CSR systems first recognize all the phonemes in the speech frame, using the similarity section, and then pass this string of phonemes to the decision rule subdivision to analyse the phonemes and find the appropriate word (25), (41).

With other words the start and endpoints can be detected by detecting the pauses between the words of connected speech. In the case of an endpoint being wrongly detected, the error can be corrected by going back and trying to find other pauses in the connected speech. CWR systems need endpoint detection, although CSR systems can use it to improve the recognition rate.

3.3.3 Pattern Similarity Determination

It is again important to show the different methods used in CWR and CSR systems.

The CWR and IWR systems have similar problems, namely that the speaking rates of the words change from time to time. Therefore it is necessary to use a time alignment procedure (Section 3.2.2.1). Different types of dynamic time-warping algorithms have been proposed by a number of people and have been used with some success (9), (28), (29), (36), (38), (43). The determination of the distance between the input features and reference features is identical to the methods used in IWR systems (Section 3.2.2.2).

CSR systems extract features from very short time periods (Section 3.3.1), sometimes a fraction of the time it takes to say one single phoneme. These systems use these small time segments in their reference-templates as computer phonemes (24), (25), (27), (30), (41). Therefore the recognition is performed on phonemes and so the need for time alignment is eliminated, because most phoneme's features do not vary in time (11). It is therefore clear that CSR systems do not have the problem of time alignment, while the determination of the distance between the test pattern and the features of the reference pattern stay exactly the same for IWR, CWR and CSR systems (Section 3.2.2.2).

3.3.4 Decision Rule for CWR and CSR systems

This decision rule is much more involved than the decision rule for IWR systems. The first part of this section is more or less the same as the decision rule for IWR systems. The decision rule normally chooses the reference pattern closest to the unknown pattern (Section 3.2.3). The choice is a word in the case of CWR systems and a computer-phoneme in the case of CSR systems.

In most CSR systems the unknown pattern is first classifed into a certain broad acoustic category and then afterwards classified as a specific computer-phoneme in that broad acoustic category. For example, the system first recognizes the input pattern as a vowel and then afterwards classifies this vowel as a phoneme.

The output from the first part of the decision rule for CSR systems is a string of computer-phonemes, and it is at this stage that some time alignment must take place. The string of phonemes may contain several repeats of the same phoneme, because the speaking rates of the phonemes vary. This type of time alignment is very easy to handle as it is just a question of deleting all the extra phonems.

The next processing step for a CSR system is to convert the strings of phonemes into words. This can be done in two ways: namely, by constantly trying to match the string of phonemes to a dictionary of phonetic transcriptions, or to use some phoneme to text rules to obtain the correct words. Phoneme to text rules require considerable processing power (44). A string of phonemes cannot be directly converted into a word by using only simple spelling rules, because of phonological variations in natural speech. However, a number of the variations can be captured in general phonological rules, because some of these variations are governed by phonological environments. According to Oshika (44) one of the most apparent trends in continuous speech is the reduction of a vowel. The rule specifies that a vowel carrying reduced stress may be realized as /e/.

Example 3.1 is

spectrogram
/s p E k t r o g r ae m/ - /s p E k t r e g r ae m/

Another rule states that if a stressed syllable is followed by two syllables with less stress, then the vowel immediately following the stressed syllable is deleted

Example 3.2 is

chocolate
/c o k e l I t/ - /c o k l I t/

This method might need more processing than the matching of the phonemes to a dictionary of phoneme transcriptions, but will use less memory and there is the possibility of an unlimited vocabulary. This level of recognition is usually called the lexical level of recognition.

Fig. 3.5 shows the flow diagram of two typical recognition systems, a CWR and CSR system.

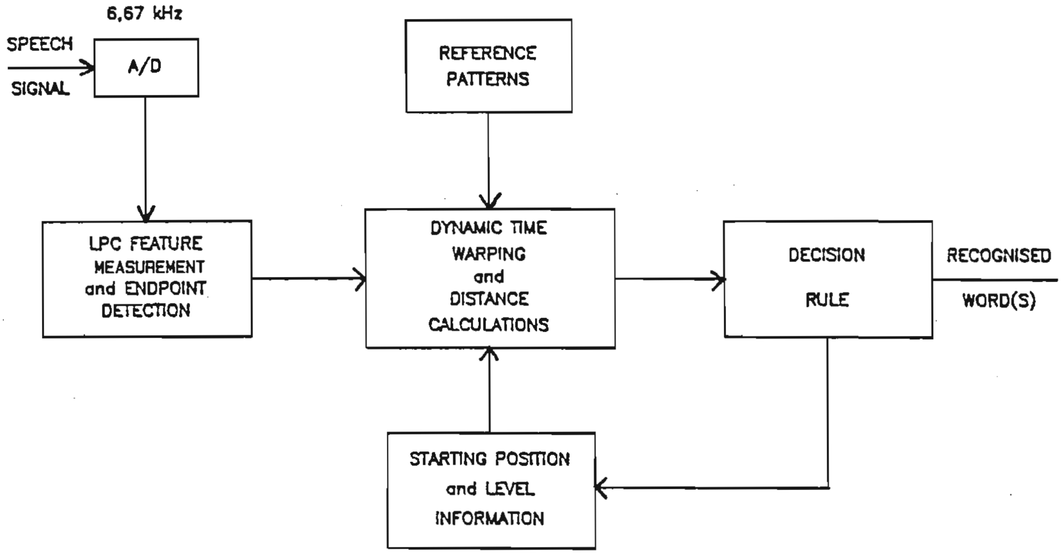

6,67 kHz
SPEECH
SIGNAL
A/D
REFERENCE PATTERNS
LPC FEATURE MEASUREMENT and ENDPOINT DETECTION
DYNAMIC TIME WARPING and DISTANCE CALCULATIONS
DECISION RULE
RECOGNISED WORD(S)
STARTING POSITION and LEVEL INFORMATION

Fig. 3.5 (a) Flow diagram of a CWR system, (2)

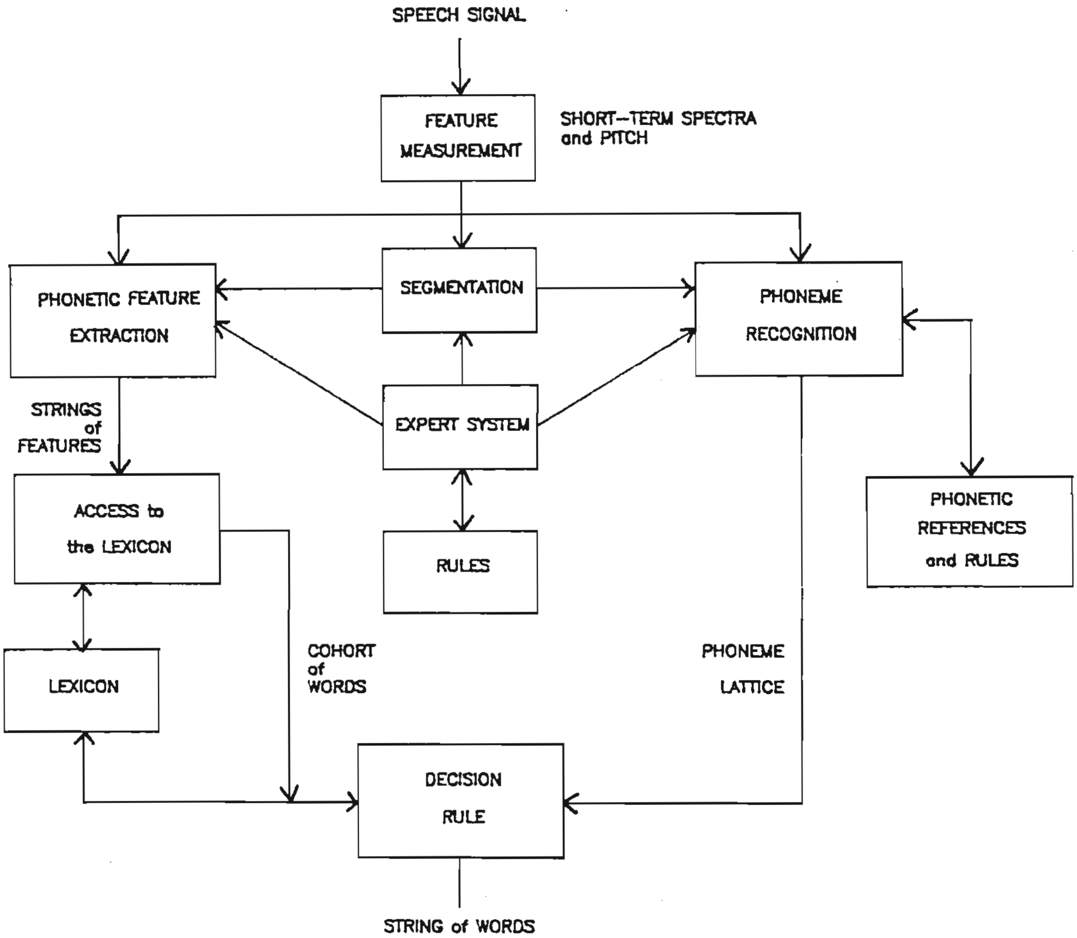


Fig. 3.5 (b) Flow diagram of a CSR system, (30)

Having recognized the word, the next step is to decide if this word is a possible word. It is therefore necessary to look at the syntax and the semantics of the words. If the syntax or the semantics of the word in a specific sentence is wrong the system must try to find a new word. This means there must be a feedback loop to help in correcting the error. The level at which the system decides if the syntax is correct is commonly known as the phrasal level. The semantics level is also called the conceptual level. The difference between syntax and semantics is clearly shown by the following example (45).

Example 3.3 is

The green colourless cloud pulled coldly
on the fast shirt.

According to any of the English syntax rules this sentence is formed correctly, but the sentence is not correct at all (see Section 3.3 for definitions of syntax and semantics). For this reason is it very useful to have some semantics rules in both the CWR and CSR systems.

A typical syntax rule is that prepositions should appear before the words they govern (46).

Example 3.4 is

We worked on the computer through the night.

If we apply this rule to Example 3.3, we will find that the preposition "on" is in correct position and the syntax is therefore correct. For the system to check these semantic errors, is it necessary however, to determine the subject. Looking at Example 3.4, the system must know the type of words that can follow the preposition "through". Nouns like "computer" and "cloud" will not fit and therefore the system must not only know the type (noun, verb, ...) of word which can follow another word, but also which words can follow a certain word.

3.3.5 The Feedback Loop

CWR systems recognize isolated words in connected speech from isolated word templates and therefore it is necessary for the feedback loop to return to the endpoint detection section of the connected speech recognition model (Fig. 3.4, Section 3.3.2) to obtain a word. This process must be repeated until the correct word is recognized. On the other hand, the CSR systems recognize small sections of words, normally phonemes, and therefore do not use endpoint detection. Phonological, syntax and semantics rules are applied to correct the phonemes and words. This means that the feedback loop must return to certain levels in the decision rule section. Fig. 3.6 shows the size and the complexity of a CSR system (after HEARSAY II, (47)).

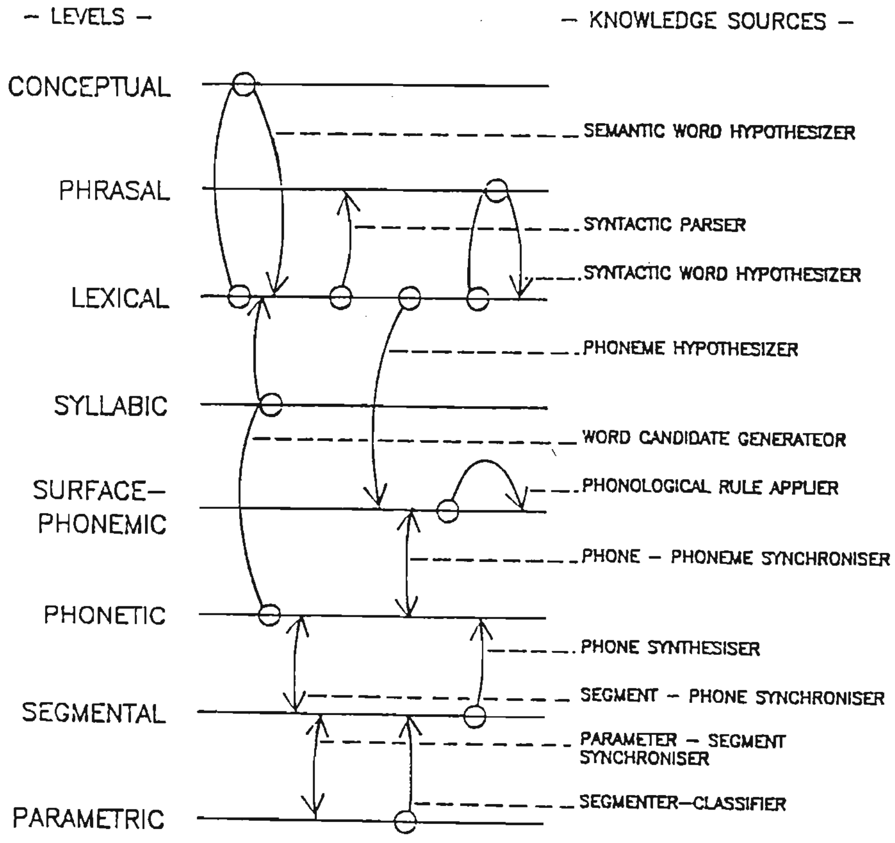

– LEVELS –
– KNOWLEDGE SOURCES –
CONCEPTUAL
PHRASAL
LEXICAL
SYLLABIC
SURFACE–PHONEMIC
PHONETIC
SEGMENTAL
PARAMETRIC
SEMANTIC WORD HYPOTHESIZER
SYNTACTIC PARSER
SYNTACTIC WORD HYPOTHESIZER
PHONEME HYPOTHESIZER
WORD CANDIDATE GENERATEOR
PHONOLOGICAL RULE APPLIER
PHONE – PHONEME SYNCHRONISER
PHONE SYNTHESISER
SEGMENT – PHONE SYNCHRONISER
PARAMETER – SEGMENT SYNCHRONISER
SEGMENTER–CLASSIFIER

Fig. 3.6 Decision rule and feedback section of a CSR system

3.4 Speaker Independent and Speaker-trained systems

Speaker-trained systems can only recognize words and phonemes spoken by the person who trained the recognition system. In the training operation a template is determined consisting of a set of features for a specific word or phoneme. This process is repeated until all the necessary templates are labelled (25), (38).

The speaker-independent systems usually have a number of labelled templates for each word, which are chosen in such a way that the system can recognize the words of any speaker's voice or at least a number of speakers' voices. Some of these systems use normalization factors to normalize the frequency and amplitude of the extracted features of the speakers' words. To find these normalising factors, the system asks the user to say a known word, the system then extracts the factors by calculating the difference between the features in the template of the known word and the features of the spoken word (28), (36).

4 Implementation of a Computer Speech Recognition System

4.1 Introduction

In this section the author gives a detailed description of the goal, restrictions and implementation of a computer speech recognition system.

4.1.1 The Goal

The original end goal of the system was to be able to recognize any letter of the alphabet in a continuous speech section. The idea behind this goal was that any speaker could could train the system to recognize his voice and then use the system as a dictaphone. The speaker will then speak into the computer rather than typing the text into the computer.

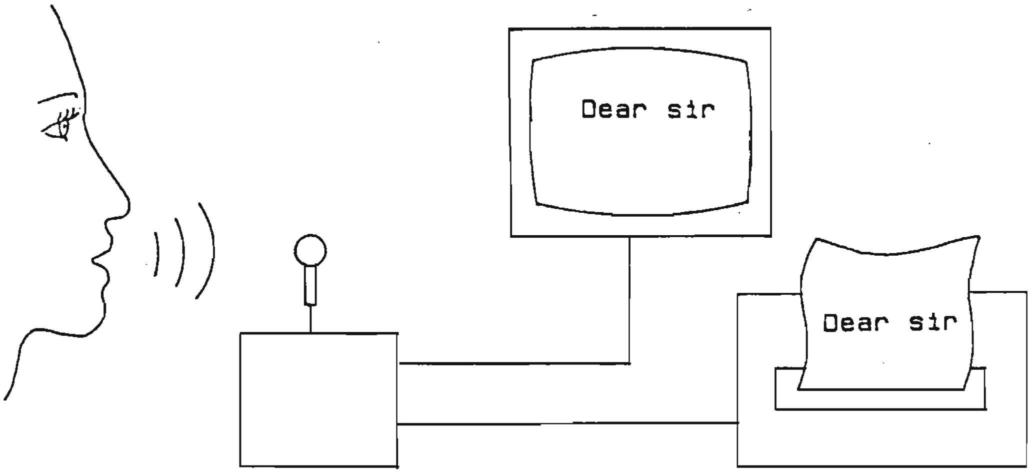


Fig. 4.1 The ultimate goal

This goal was set as the end product of the project, not to be accomplished in the amount of time given to the project. To achieve this goal it was necessary to look at all the systems presently available. Chapter 3 describes the three different study fields in speech recognition. See also Table 4.1.

STUDY FIELD	MODE OF SPEECH	TYPE OF PROCESSING	VOCABULARY SIZES
Isolated Word Recognition	Isolated words	recognize isolated words	10 - 1000 (24) words
Connected Word Recognition	Senteces with short pauses between the words	recognize isolated words	10 - 127 (38) words
Continuous Word Recognition	Continuous speech, no pauses	recognize the phonemes or syllables in the words	47 (30) phonemes

Table 4.1 Different speech recognition systems

In the studying of the goal, it was found that there are three factors which are the deciding factors in what type of recognition method must be used. First of all the system must be able to recognize any letter of the alphabet. Secondly, the speech will be continuous and finally the system can be speaker dependent.

By comparing these factors with the features of the different speech recognition systems (Table 4.1), the wanted system can be classified as a speaker dependent (speaker-trained), continuous speech recognition system.

To determine how to implement a set goal it is necessary to investigate the:

a) practical implication involving such a system,
b) availability of equipment,
c) cost of time and cost of computer memory.

4.1.2 A Practical Approach to the Goal

After a thorough investigation into continuous speech recognition systems and the equipment available it was time to re-evaluate the first goal and set a new goal. The new goal was set in such a manner that by constantly improving on the system, the initial goal can be accomplished. The new goal can be seen as the first stage of the initial goal.

The practical approach can be summarized as follows:

a) Develope the system on the mini-computer available for the project.

b) Digitize speech with the analog to digital converter and store the digitized speech samples in the mini-computer.

c) Recognize as many phonemes as possible.

d) Convert these phonemes to text.

4.1.3 Reasons for Using such an Approach

a) The use of a computer for the development of the system is of great value, because software programs can easily be changed with minimal cost, as opposed to a hardware system. Once the system works effectively, it can be analysed to see which sections of the system are the most time consumming sections and then the speed of these sections can be optimized, by integrating software and hardware. Futhermore a system developed on a computer can efficiently be changed to work on any other type of digital means, like other computers, hardware systems or even micro-computers. A HP 1000 F-series mini-computer (HP1000) was used for the development of the software system.

b) For convenience the speech signals are recorded on a tape-recorder beforehand, rather than talking straight into the computer each time the system must be trained or tested in the development process. These recordings can then be made with any type of background noise and are not limited to computer-room background noise.

The recordings are digitized with an analog to digital converter and are stored in the computer as files and thus save an immense amount of time, especially where one speech sample is continuously used to train the system and another sample is constantly used for the testing of the system. For the development of the system it was decided to train and test the system on the words "zero" through to "nine", which were spoken in a slow connected manner. These words were chosen, because they are very commonly used in speech and contain half of all the English phonemes. A system which can recognize the ten digits can be used in a number of practical applications, like a speech telephone dialling system and perhaps a calculator.

An UHER 4000 tape recorder is used for the analog recordings and a CAMAC 12 bit analog to digital converter (ADC) is used for the digital recordings which are stored in the mini-computer.

c) Why a phoneme based recognition system?
The answer to this is straightforward. The end goal of the the system is to use it as a dictaphone, and it must therefore be able to recognize any word used by the user (speaker).

There are two methods of achieving this. The first is to train the system with every word in the English language. Secondly, train the system on all the phonemes in the English language, which will take up much less computer memory and will take only a fraction of the time it will take to train with all the English words. To recognize a word from a 1000 word vocabulary will take much longer than to recognize a phoneme from a 47 phoneme template.

d) Once the phonemes are recognized, they need to be converted to ordinary English text, (Section 3.3.4 and Section 3.3.5). To convert these phonemes to words, two major study directions can be implemented. A string of phonemes can either be matched (looked up) in a phoneme transcriptions dictionary which contains all the phoneme transcriptions of all the English words (say about 10 000 words) or simple phoneme to text rules can be applied to convert the phonemes to text.

The later-mentioned method will need more intelligence, but will use less computer memory. Because the first method will need more and more memory as the dictionary enlarges, it was decided to implement the phoneme to text rules. The major advantage of such a system is that if this system is trained on the phonemes of the words "zero" to "nine", it can also convert other words to text (Example 4.1).

Example 4.1 is phonetics - /f e n E t I k s/

Every phoneme in the word "phonetics" can be found in the ten words mentioned above (Table 4.2).

To achieve the initial end goal from here, it is necessary to:

a) improve the phoneme recognition rate,
b) improve the phoneme to text program,
c) speed up the whole system, by using a very fast signal processor with an analog input (microphone) and digital output (monitor or printer).

4.2 System Implementation

In Fig. 4.2 a model of the proposed system is given and in Fig. 4.3 the recognition section is given in more detail. A detailed explanation of the recognition section can be found in Section 4.2.2.

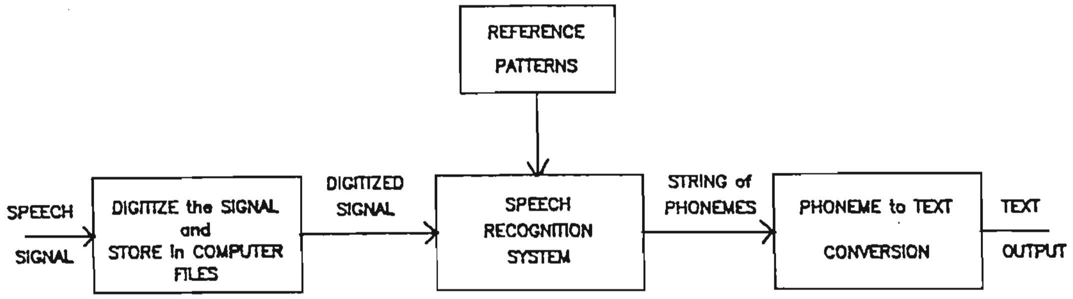

REFERENCE PATTERNS
SPEECH SIGNAL
DIGITIZE the SIGNAL and STORE In COMPUTER FILES
DIGITIZED SIGNAL
SPEECH RECOGNITION SYSTEM
STRING of PHONEMES
PHONEME to TEXT CONVERSION
TEXT OUTPUT

Fig. 4.2 A practical approach

4.2.1 Speech Data Acquisition

To obtain the necessary speech signals for the training and testing of the system, the speech was recorded on an UHER 4000 Reporter II, analog tape recorder, at a tape speed of 475 mm/s, with a M 514 UHER microphone. The recordings were done in a quiet room and thus the quality of the speech can be described as room-quality speech. Table 4.1 shows the words which were used in the testing and training of the system.

The digitizing of the speech signals were done with the department's HP1000-CAMAC system. A program written by one of the author's fellow-students was used, to perform the data transfer from the CAMAC's linear 12 bit analog to digital converter (ADC) to the mini-computer's (HP1000) memory, (48). The program uses a direct memory block access (DMA) technique to aquire the very fast sample rate require for speech signals. Because of the fact that the DMA drivers of the HP1000 can only access 32 000 words in which the actual program and the data must be stored, to save time, it was necessary to limit the maximum amount of samples digitized to 25 000 samples. At a sampling rate of 10 000 Hz the maximum length of of a continuous speech signal which can be recorded is 2,56 seconds.

For anti-aliasing, an adjustable 8th order Butterworth lowpass filter (ROCKLAND Model 1042F) was used and was set at a 3 dB cutoff frequency of 4,5 kHz. To eliminate the chances of adding a zero frequency (direct current, DC) noise or the main power supply frequency (50 Hz) noise, an 8'th order Butterworth highpass filter (ROCKLAND) was used at a 3 dB cutoff frequency of 60 Hz. This bandwidth (60 Hz to 4 500 Hz) is wide enough to contain all the important frequencies produced by the vocal system see Fig. 2.1 in Chapter 2.

The four sets of the words "zero" to "nine" (see Table 4.2 for the phonetic transcriptions of these words) were recorded and stored on disc in 20 files. Each file consists of 200 frames (2 560 ms of speech) each. All the records (also called frames) contain 128 speech samples (12,8 ms of speech), because a radix-2 FFT was used to obtain the frequency spectrum ($2^5 = 128$) and the frames must be short enough to contain features of any phoneme (phonemes can be as short as 30 ms). There are normally two words in each one of the files. For example the first and the fifth file both contains the words "zero" and "one".

WORDS	PHONETIC TRANSCRIPTIONS
zero	/z Ie r eU/
one	/w A n/
two	/t u:/
three	/th r i:/
four	/f o:/
five	/f aI v/
six	/s I k s/
seven	/s E v e n/
eight	/eI t/
nine	/n aI n/

Table 4.2 Phonetic Transcriptions of the Words "zero" to "nine"
Appendix A

4.2.2 A Phonetical Approach to Speech Recognition

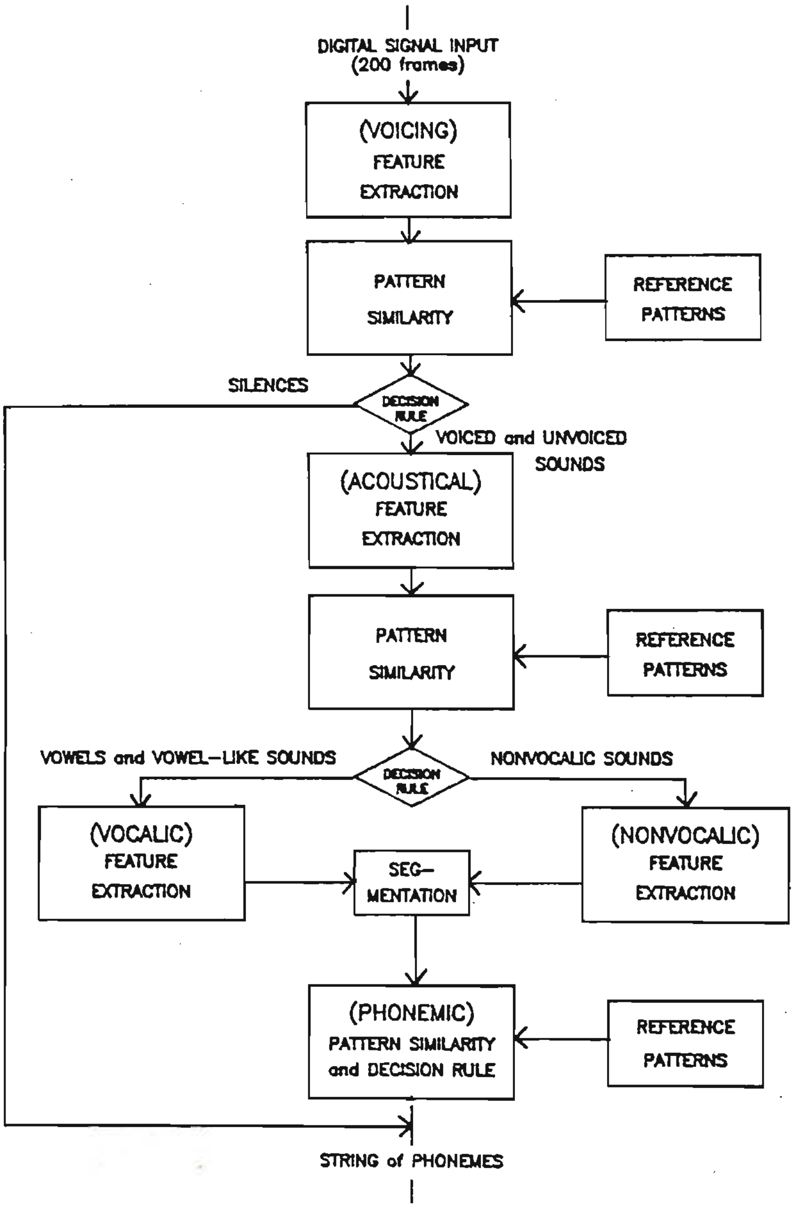

DIGITAL SIGNAL INPUT
(200 frames)
(VOICING) FEATURE EXTRACTION
PATTERN SIMILARITY
REFERENCE PATTERNS
SILENCES
DECISION RULE
VOICED and UNVOICED SOUNDS
(ACOUSTICAL) FEATURE EXTRACTION
PATTERN SIMILARITY
REFERENCE PATTERNS
VOWELS and VOWEL-LIKE SOUNDS
DECISION RULE
NONVOCALIC SOUNDS
(VOCALIC) FEATURE EXTRACTION
SEG-MENTATION
(NONVOCALIC) FEATURE EXTRACTION
(PHONEMIC) PATTERN SIMILARITY and DECISION RULE
REFERENCE PATTERNS
STRING of PHONEMES

Fig. 4.3 Flow diagram of the proposed speech recognition system

According to Fig 4.3 it is clear that the recognition process is done in stages. First of all some features are extracted from every 12,8 ms of the speech samples (each frame). These frames are then classified as either a silence, voiced sounds or unvoiced sounds. Then more features are extracted and the sound (frame) is now classified as a vowel and vowel-like sound or as a nonvocalic sound. The final features are then extracted to recognize the actual phoneme. This type of stage recognition system (segmental recognition) has been used in a number of systems (25), (26), (27), (30), (41), (47).

The reason for using this type of recognition is to save on execution time and to reduce the recognition error rate. For example once a frame is classified as a silence it is unnecessary to do any more processing on that frame. Also, once a frame is classified as a vocalic sound or nonvocalic it is not necessary to extract all the phonemic features but only the relevant features (Section 4.2.2.1 C). This method prevents the unnecessary matching of a vocalic sound to the nonvocalic templates and thus saves on exacution time. This also seems to be the method the human auditory system uses to recognize the phonemes (Section 2.2.4).

This approach is described in three sections:

a) feature extraction (Section 4.2.2.1),
b) pattern similarity (Section 4.2.2.2),
c) decision rule (Section 4.2.2.3).

Because the system is a segmental recognition system, it is divided into three recognition stages:

a) voicing (used for the classification of voiced, unvoiced sounds and silences, Sections 2.1.2.3)
b) acoustical (vowels, stops, fricatives, ... see Section 2.1.3)
c) phonemic (/A/, /t/, /f/, ... see Appendix A) recognition stages.

Therefore the Section 4.2.2.1 is also divided into three subsections (A, B and C). Section 4.2.2.2 and Section 4.2.2.3 describe the pattern similarity and decision rule respectively in general, because the same method is used in all three of the recognition stages, although in Section 4.2.2.3 the segmentation of the phonemes is also explained (Fig. 4.3).

4.2.2.1 Feature Extraction

This system extracts four different types of features for the classification of the input speech pattern.

A Voicing Features

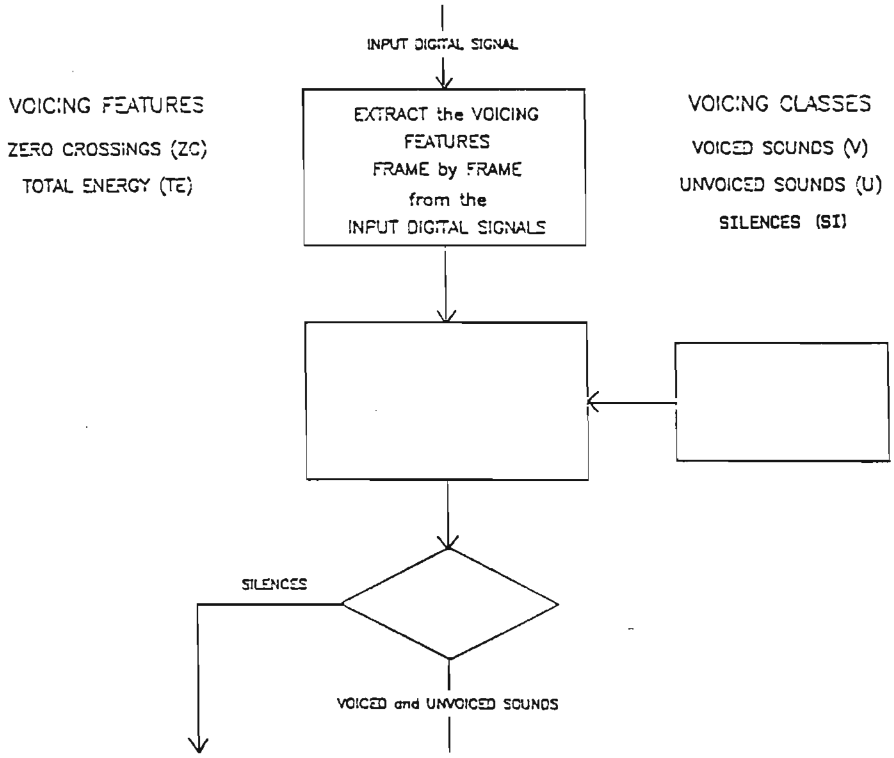


Fig. 4.4 Voicing features extraction

Each frame (record) of the speech samples is analysed and then classified. To obtain the voicing features, which will be used later to determine whether the frames are voiced, unvoiced or a silence, the total energy and the zero-crossings of the frames are calculated '(27), (37) (Section 3.2.1.3).

The total energy is calculated with the following formula:

$$TE_i = 10.\log_{10} (\sum_{n=1}^{N} x_i(n)^2) \qquad 4.1$$

where N is the maximum number of samples in a frame (N=128) and x(n), n = 1, 2, 3,...N-1, N is the n^{th} sample of the i^{th} record of the input speech pattern (file).

The zero-crossings are defined as the number of times the digital signal's digital value changes from a positive value to a negative value and thus is:

$$ZC_i = \sum_{n=2}^{N} (x_i(n-1) \geq 0 \text{ AND } x_i(n) < 0) \qquad 4.2$$

where N is the maximum number of samples in a frame (N=128) and x(n), n = 1, 2, 3,...N-1, N is the n^{th} sample of the i^{th} record of the 200 record digital speech file. The function in the round brackets is conditional formula which will have result of nil when the answer is false and a result of unity when the answer is true.

Once the voicing features are extracted they are matched to the features in the reference templates and then labelled as either a voice frame, unvoiced (voiceless) frame or a silence (Section 2.1.2.3). If the frame is a silence no further processing is done on that frame and this 12,8 ms of the signal is labelled a silence. On the other hand, if the frame is classified as a voiced or voiceless frame more processing is done on the frame.

B Acoustical Features

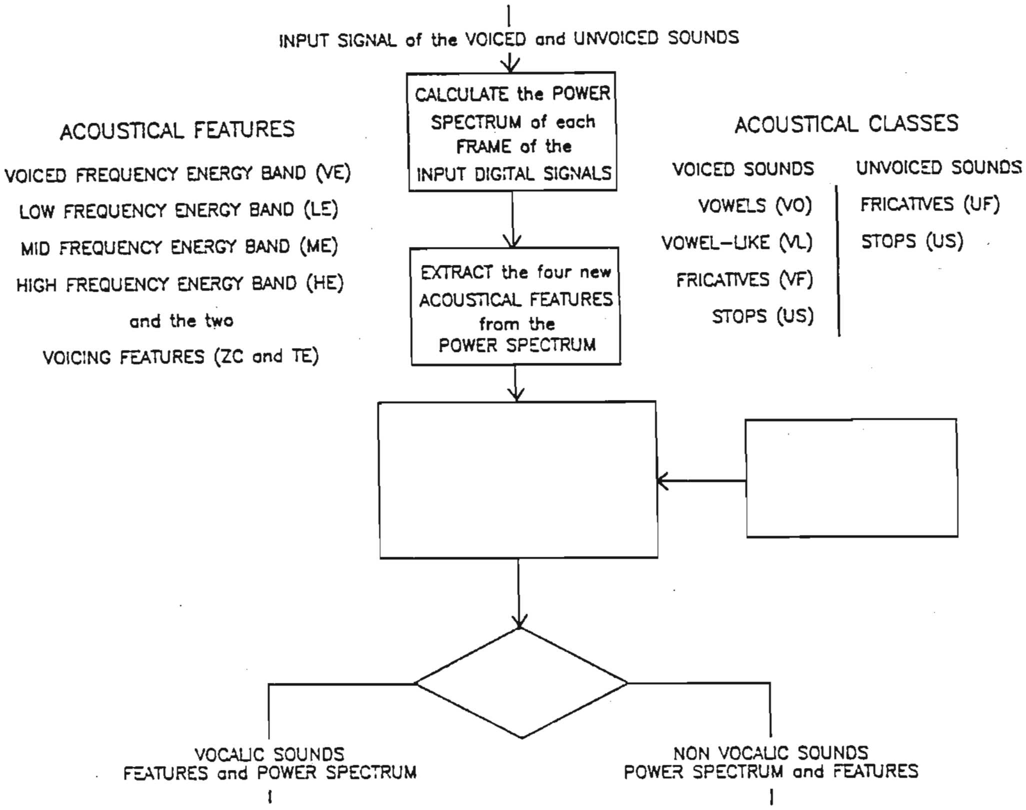

INPUT SIGNAL of the VOICED and UNVOICED SOUNDS
CALCULATE the POWER SPECTRUM of each FRAME of the INPUT DIGITAL SIGNALS
ACOUSTICAL FEATURES
VOICED FREQUENCY ENERGY BAND (VE)
LOW FREQUENCY ENERGY BAND (LE)
MID FREQUENCY ENERGY BAND (ME)
HIGH FREQUENCY ENERGY BAND (HE)
and the two
VOICING FEATURES (ZC and TE)
EXTRACT the four new ACOUSTICAL FEATURES from the POWER SPECTRUM
ACOUSTICAL CLASSES
VOICED SOUNDS
VOWELS (VO)
VOWEL-LIKE (VL)
FRICATIVES (VF)
STOPS (US)
UNVOICED SOUNDS
FRICATIVES (UF)
STOPS (US)
VOCALIC SOUNDS
FEATURES and POWER SPECTRUM
NON VOCALIC SOUNDS
POWER SPECTRUM and FEATURES

Fig. 4.5 Acoustical features extraction

The next set of features are extracted from the frames to find out in what broad acoustical category the specific frame is.

The categories are more or less the ones described in Section 2.1.3 see also Table 4.3 and Fig. 4.5.

VO	Vl	VS	VF	US	UF
Ie	m	d	v	p	f
eU	n	t	vh	t	th
A	ng	g	z	k	s
u:	r	dzh	zh	tsh	sh
I	l				
aI	w				
E					
e					
eI					

VO - Vowels and diphthongs

VL - Vowel-like

VS - Voiced stops

VF - Voiced fricatives

US - Unvoiced stops and affricates

UF - Unvoiced fricatives

Table 4.3 The acoustical categories are the phonemes (note, not all the vowels are given in the table above)

Six parameters are selected for the classification of the acoustical category of the frames, total energy (TE) and the zero crossings rate (ZC) as previously explained. A further four energy bands are calculated from the power spectrum of the frames.

The power spectrum is computed from a discrite fourier transform (DFT) which is calculated using the 128 samples in each frame. In order to minimize the leakage between two neighbouring frames each frame is weighted with a window prior to the calculation of the DFT. The other reason for doing so is to ensure a smooth estimation of the power spectrum (32), (Section 3.2.1.1). A 12,8 ms Hanning window is used to weight the digital speech signals, the window can be describes as follows (33):

$$h(n) = 0{,}5-0{,}5.\cos(\frac{2.\pi.n}{N-1}) \qquad 4.3$$

where h(n) is the n^{th} value of the window and N is the maximum number of samples in a frame. To obtain the weighted speech signal the signal is multiplied with the window. The windowed speech signal is defined as:

$$y(n) = x(n).h(n) \qquad 4.4$$

and y(n) is the n^{th} sample of the input signal, where n = 1, 2, 3,...128.

The DFT is the Fourier representation of a finite-duration sequence (like a speech signal) and is of the form:

$$Y(k) = \begin{cases} \sum_{n=0}^{N-1} y(n).W_N^{kn} & 0 \leq k \leq N-1 \\ 0 & \text{otherwise} \end{cases} \qquad 4.5$$

where $W_N = e^{-j(2.\pi/N)}$ and y(n) may be complex.

A Fortran program of a FFT algorithm which uses the decimation in time method to calculate the DFT of 2^i number of samples (pionts), where i is any positive integer is given in Appendix B. This algorithm computes the transform in $\log_2 N$ stages. Each stage has N complex multiplications and N complex additions, thus there will be a total of $N.\log_2 N$ complex multiplications and additions. This is a considerable drop in calçulations opposed to the original number of calculations needed for the calculation of the DFT.

In order to obtain the power spectrum from the FFT of the windowed signal y(n), n=1, 2, 3,...N:

$$X(k) = 20.\log_{10} (Y(k)) \qquad 0 \leq k \leq N \qquad 4.6$$

where N=128 for all the frames of the 200 frame files.

The four extra acoustical features are a voiced-frequency energy-band (VE, 80 - 300 Hz), a low-frequency energy-band (LE, 300 - 1 000 Hz), a mid-frequency energy-band (ME, 1 000 - 3 200 Hz) and a high-frequency energy-band (HE, 3 200 - 4 500 Hz) (Eq. 4.7). These parameters are found to provide adequate discrimination between the different acoustical categories and are based on work done by Paliwal and Roa (27). The Sperry Univac speech recognition system uses similar energy bands (49).

This calculation of the power spectrum and the classification of the frames into acoustical categories is closely related to what happens in the ear. The chochlea in the ear transforms the input signal to the power spectrum of signal (Section 2.2.1.3). It is believed that some kind of acoustical classification is done in the higher levels of the auditory pathway (Section 2.2.4).

C The Phonemic Features

From Fig. 4.3 it is obvious that depending on whether the segment is a vocalic or a nonvocalic sound a different set of features is extracted. In both cases it is necessary to calculate three more features. The first six features already explained in the previous sections are used with the new features. The system will thus use nine parameters to recognize a phoneme.

(I) Vowel and Vowel-like Sounds

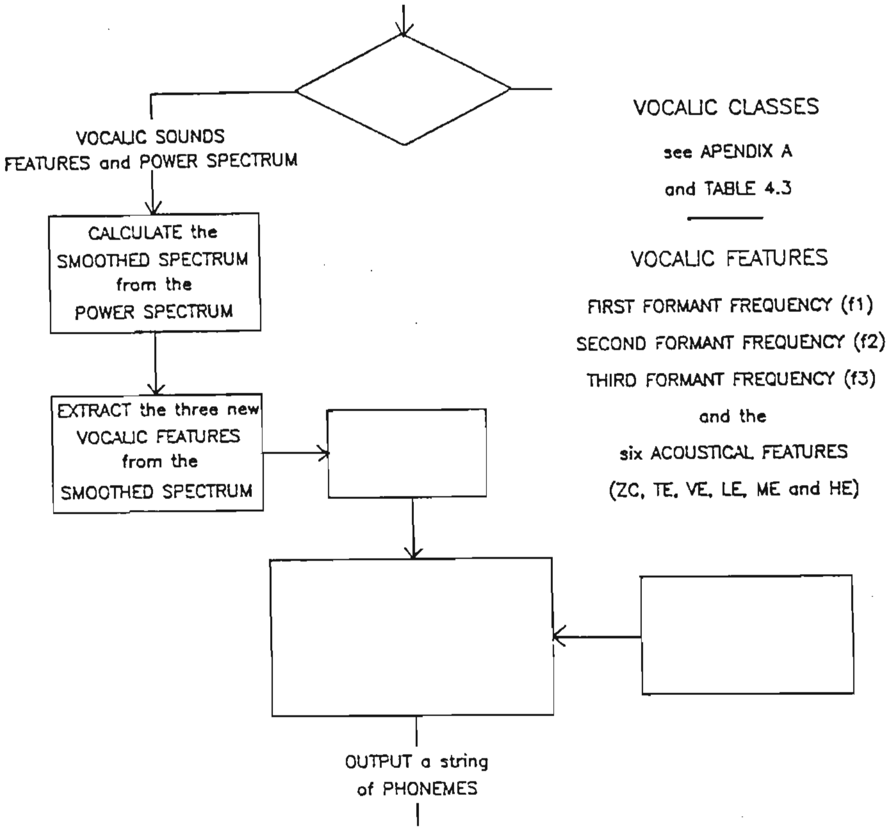

VOCALIC SOUNDS
FEATURES and POWER SPECTRUM
CALCULATE the SMOOTHED SPECTRUM from the POWER SPECTRUM
EXTRACT the three new VOCALIC FEATURES from the SMOOTHED SPECTRUM
OUTPUT a string of PHONEMES
VOCALIC CLASSES
see APENDIX A
and TABLE 4.3
VOCALIC FEATURES
FIRST FORMANT FREQUENCY (f1)
SECOND FORMANT FREQUENCY (f2)
THIRD FORMANT FREQUENCY (f3)
and the
six ACOUSTICAL FEATURES
(ZC, TE, VE, LE, ME and HE)

Fig. 4.6 Vocalic features

If these two acoustical categories are observed (Section 2.1.3.1 B and Section 2.1.3.3 A and B) it is clear that the formant frequencies of these categories are a very important feature (Fig. 4.6). The first two formants contain almost enough information for the recognition of any vowel (13). However, it was decided to use the first three formants, as the extra three features, to identify the vowel and vowel-like sounds. The use of these three parameters is very common in speech recognition systems (25), (27), (31), (49), (50), (51).

The formant frequencies of these sounds can be extracted from the power spectrum of each frame in the phonemic segment either by using a peak-picking procedure or by means of iterative methods (analysis-by-synthesis, (52)). Because the iterative methods use too much computation time, the peak-picking procedure is used for the extracting of the first three formants of all the frames.

The peak-picking method used is based on work done by Schafer and Rabiner (53) and is actually a double peak-picking procedure. The first peak-picking is done to smooth the spectra of the sound (Fig.4.7).

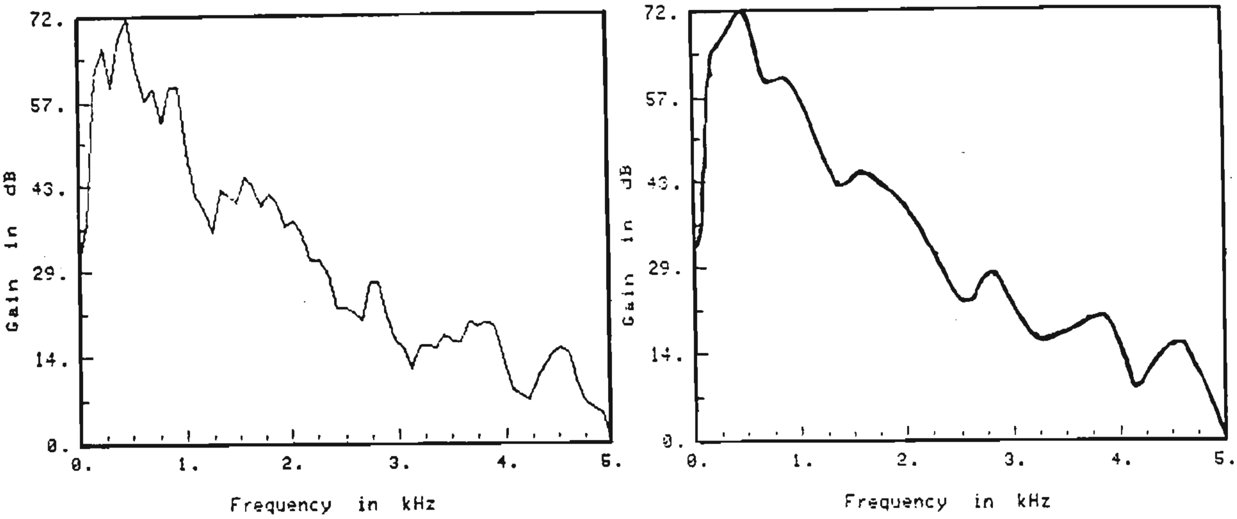


Fig. 4.7 Normal and smoothed spectra of the vowel /o:/ in word four

Although other systems use a selective linear predictive technique to smooth the spectra (27), (54), (55) the above-mentioned method is found to be an efficient method of smoothing the the power spectrum of the sound.

The second peak-picking procedure uses this smoothed spectra to obtain the first three formants. This procedure involves slightly more than just picking the first three highest peaks and labelling them as formants. From experimental data obtained from the speech recorded digitally and from results of previous analysis of the frequency regions of the different formant freqency, the following regions for the different formants arrives:

a) The First formant (f1) ranges from 230 to 1090 Hz,
b) the second formant (f2) ranges from 630 to 2890 Hz and
c) the third formant (f3) ranges from 1880 to 3670 Hz.

Analysis also showed that relative amplitudes of the formant peaks plays an important role in the estimation of the formants. For example if the difference between the amplitude of f2 and that of f3 is less than 17 dB, the chances are good that the formants f2 and f3 are incorrect and two new peaks must be found. (see Appendix C for the flow charts of the estimation of the formants).

(II) Nonvocalic Features

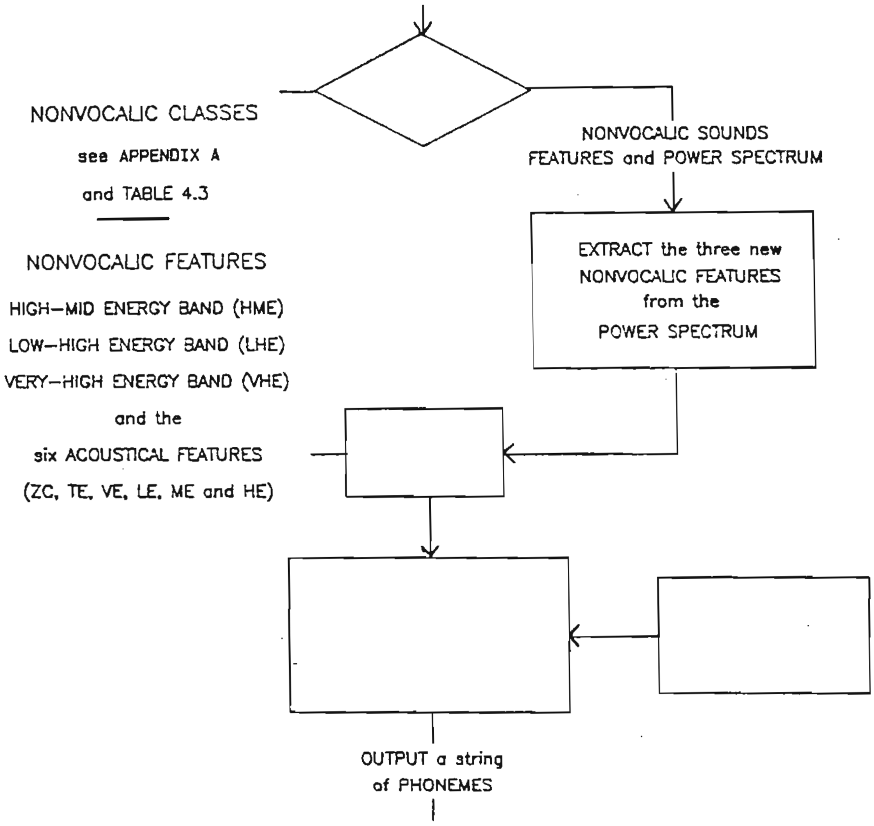


Fig. 4.8 Nonvocalic features extraction

The nonvocalic sounds usually contain more information in the higher frequencies than the vocalic sounds and therefore it is advisable to extract more features from the higher frequencies (25), (27), (50), (56), (Fig. 4.8 and Fig. 4.9).

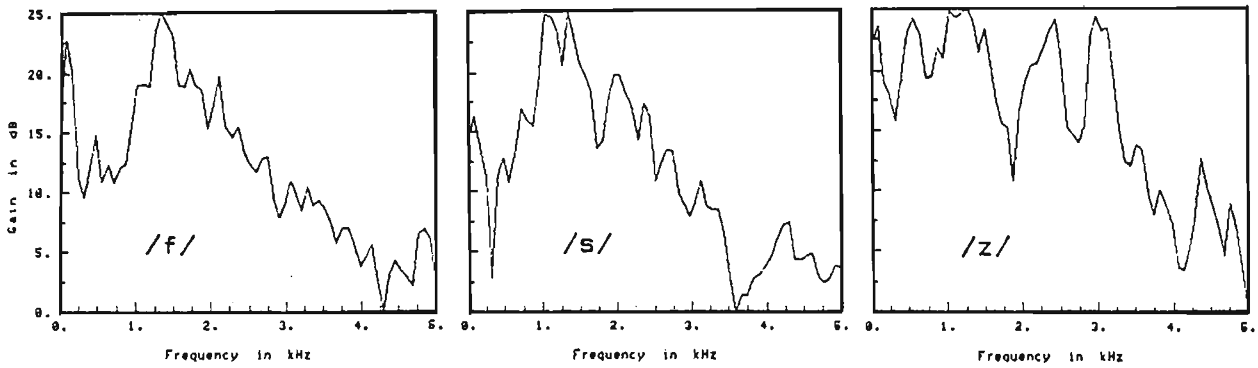


Fig. 4.9 Spectrograms of three Nonvocalic Sounds (compare with the spectra of the vocalic sound in Fig. 4.7)

Data obtained from analysing the nonvocalic sounds proved that the following three frequency bands will contain information which can be used to identify the different nonvocalic sounds,

a) The high mid-frequency energy-band (HME, 2000 to 2750 Hz),
b) the low high-frequency energy-band (LHE, 2750 to 3750 Hz) and
c) the very high-frequency energy-band (VHE, 3750 to 4500 Hz).

The frequency energy-bands are of the form:

$$E_i = 10.\log_{10}.(\sum_{k=n_{i1}}^{n_{i2}} X_i(k))/K_i \qquad 4.7$$

where E_i is the i^{th} frequency energy-band, E_i = VE, LE, ME, HE, HME, LHE, VHE, n_{i1} is the minimum and n_{i2} the maximum frequency of E_i, $K_i = n_{i2}-n_{i1}$, X(k) is the k^{th} value of the power spectrum and,

$$n = BW.(N/2+1)/(S/2) \qquad 4.8$$

N is the number of samples used for each power spectrum (N=128, Eq. 4.6), S is the sampling rate (S=10 kHz), BW is the frequency-range of the frequency energy-band in Hz.

Table 4.4 shows a summary of the features being used for the recognition of the two types of sounds.

Features	Vocalic sounds	nonvocalic sounds
Voicing	Total Energy (TE) Zero Crossings (ZC)	
acoustical	both above mentioned features and	
	Voiced-frequency Energy-band (VE, 80 - 300 Hz)	
	Low-frequency Energy-band (LE, 300 - 1000 Hz)	
	Mid-frequency Energy-band (ME, 1000 - 3200 Hz)	
	High-frequency Energy-band (HE, 3200 - 4500 Hz)	
Phonemic	all above mentioned features and	
	First Formant (f1)	High Mid-freqency Energy-band (HME, 2000 - 2750 Hz)
	Second Formant (f2)	Low High-freqency Energy-band (LHE, 2750 - 3750 Hz)
	Third Formant (f3)	Very High-freqency Energy-band (VHE, 3750 - 4500 Hz)

Table 4.4 Features of the different recognition stages

4.2.2.2 Pattern Similarity

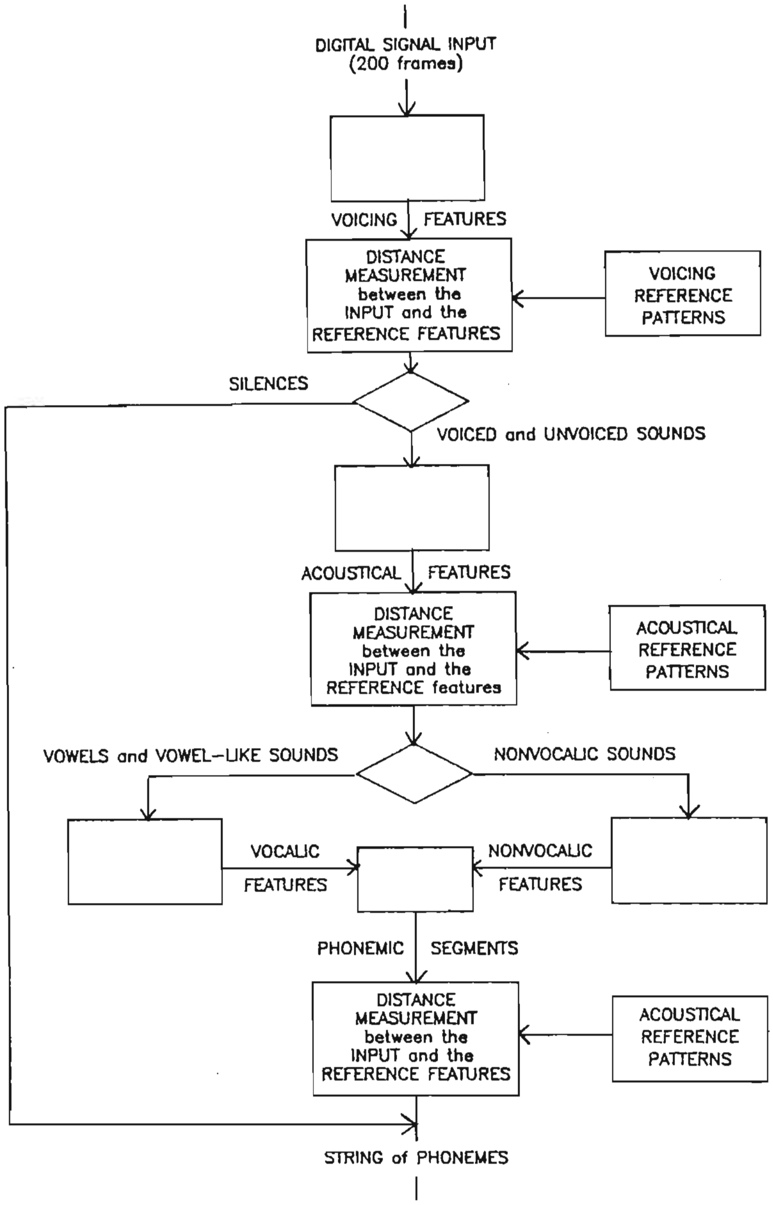

DIGITAL SIGNAL INPUT (200 frames)
VOICING FEATURES
DISTANCE MEASUREMENT between the INPUT and the REFERENCE FEATURES
VOICING REFERENCE PATTERNS
SILENCES
VOICED and UNVOICED SOUNDS
ACOUSTICAL FEATURES
DISTANCE MEASUREMENT between the INPUT and the REFERENCE features
ACOUSTICAL REFERENCE PATTERNS
VOWELS and VOWEL-LIKE SOUNDS
NONVOCALIC SOUNDS
VOCALIC FEATURES
NONVOCALIC FEATURES
PHONEMIC SEGMENTS
DISTANCE MEASUREMENT between the INPUT and the REFERENCE FEATURES
ACOUSTICAL REFERENCE PATTERNS
STRING of PHONEMES

Fig. 4.10 Pattern similarity

In the system proposed by the author two classifications are done before the actual phoneme can be recognized, (Section 4.2.2 and Fig. 4.10) therefore it is necessary to monitor the pattern similarity at three different points in the system. The voicing , the acoustical and the phonemic features of the input pattern need to be checked with the reference patterns for pattern similarity.

Each frame of the input pattern is matched with the necessary reference patterns (templates), to find the distance measurement between the input pattern, B, and reference patterns, A, (Section 3.3.3 and Section 3.2.2.2). The pattern similarities are obtained by calculating the distance between the features of the input frames and that of a specific set of reference patterns, which are of the same class (voicing or acoustical class) as the input frames (patterns).

For example, the word "four" /f o:/, (Fig. 4.13). First of all the voicing features are extracted (Section 4.2.2.1 A) from each one of the frames and the distance between the voicing reference templates and the frames is calculated (Fig 4.10) to classify the frame to an appropriate voicing class (a description of how the classification works can be found in Section 4.2.2.2 and Section 4.2.2.3). In this case the first number of frames will be classified as silences, the second lot as unvoiced sounds, the third couple of frames as voiced sounds and the final number of frames as silences again (that is if the word is pronounced in an isolated manner).

Now the acoustical features are extracted (Section 4.2.2.1 B) and the frames labelled as unvoiced sounds are matched to the unvoiced acoustical reference templates and the voiced frames are matched against the voiced acoustical reference patterns (Table 4.3) to find the pattern similarity between the test frames and the specific range of reference templates (Fig. 4.10). The system will now have segmented and classified (see Section 4.2.2.2 and Section 4.2.2.3 for explanation of classification of the frames) the two sets of frames into their acoustical group, the unvoiced frames as unvoiced fricatives (UF) and the voiced frames as vowels (VO) (Section 4.2.2.1 B).

The phonemic features are extracted (Fig. 4.6 and Fig 4.8) and the third pattern similarity takes place (Fig 4.10). In this case the distances between the features of the UF segments and UF phonemic reference templates has been calculated to find the correct fricative. The same is done with the features of the VO segments to recognize the correct vowel.

In the present system a weighted Euclidean distance measurement is in use (57) (Section 3.2.2 B) and is of the form:

$$d_i^2 = \sum_{j=1}^{N} (W_{ij}.(B_j-A_{ij}))^2 \qquad 4.9$$

where d_i is the distance of the i^{th} reference template and N the maximum number of features in the templates, A_{ij} and W_{ij} are the j^{th} feature of the reference vector and the weight vector, respectively, of the i^{th} reference template, B_j is the j^{th} feature of the input (test) pattern to be classified. Each feature of each template (reference template) has a weight W_{ij} which is proportional to the standard deviation of the respective features.

The reference vector can actually be described as a mean vector and the weight vector as a standard deviation vector:

$$A_{ij} = (\sum_{k=1}^{N_i} T_{ijk})/N_i \qquad 4.10$$

and

$$W_{ij}^2 = (\sum_{k=1}^{N_i} (T_{ijk}-A_{ij})^2)/(N_i-1) \qquad 4.11$$

where N_i is the maximum number of training patterns being used to obtain the i^{th} template (reference pattern) and T_{ijk} the j^{th} feature of the k^{th} training pattern of the i^{th} template.

This means that A_{ij} is the mean of a number of training patterns and W_{ij} is the standard deviation of these training patterns and thus from Eq. 4.10 and Eq. 4.11:

$$W_{ij}^{2} = \left(\sum_{k=1}^{N_i} T_{ijk}^{2} - \left(\sum_{k=1}^{N_i} T_{ijk} \right)^{2} / N_i \right) / N_i \qquad 4.12$$

The values for i and j change for each one of the three distance measurements (voicing, acoustical and phonemic distance measurement).

WEIGHT	REFERENCE PLATES		FEATURES
W_{ij}	minimum i		j
Voicing	3		2
Acoustical	3	V	6
	3	U	6
Phonemic	4	UF	9
	4	US	9
	4	VF	9
	4	VS	9
	6	VL	9
	9	VO	9

Table 4.5 Different values for i and j for Eq.4.12

The minimum number of templates (minimum i) is based on the phonemes in Table 4.3 (Section 4.2.2.1). The values of i are given as minimum values, because there might be times when the phonemic, the acoustical or the voicing classes need more than one template to identify a specific class, as in the cases where the pronunciations of a phoneme change so drastically, because of the following phoneme, that by averaging two pronunciations the recognition rate will worsen.

4.2.2.3 The Decision Rule

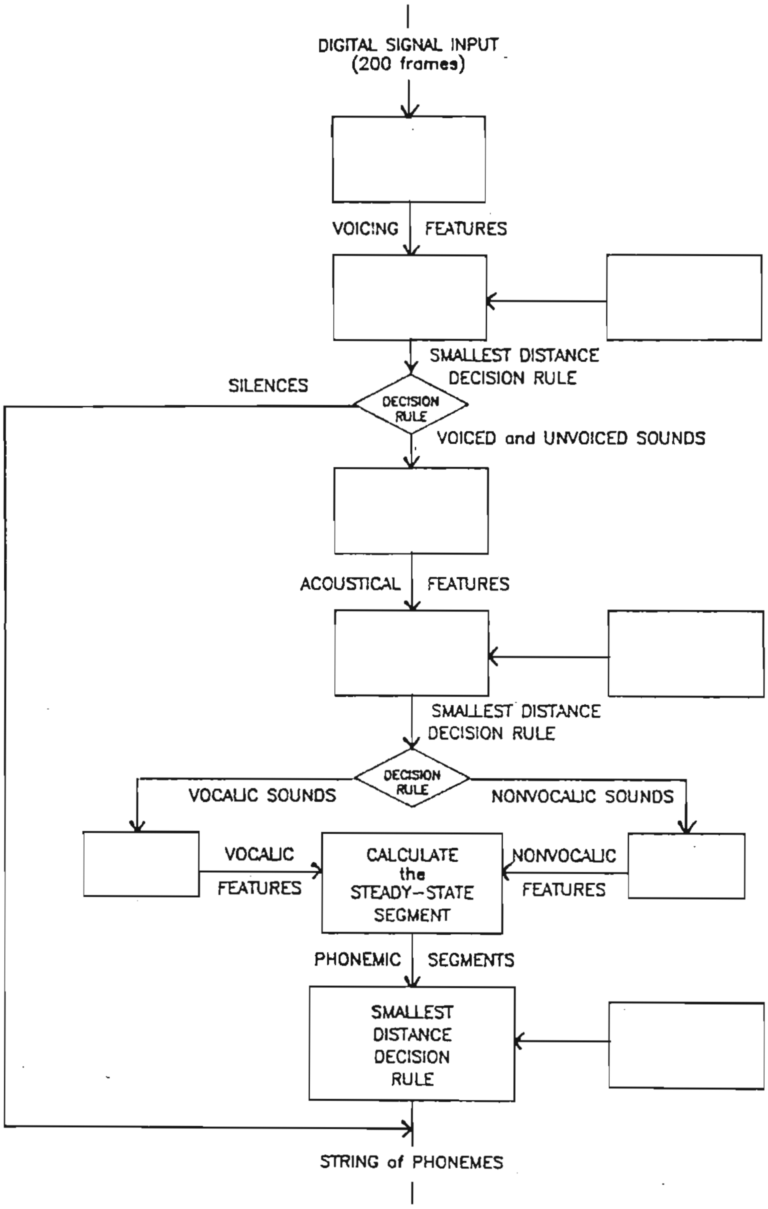


Fig. 4.11 Decision rule

Similar to the pattern similarity section (Section 4.2.2.2) the decision rule will take place at three points in the proposed system (Fig. 4.11). This decision rule takes place directly after the distance measurements have been calculated.

In all three cases the input patterns are labelled by the same name as the name of the template which gives the minimum distance according to Eq. 4.9, Eq. 4.10 and Eq. 4.11. This method is also called 'the nearest neighbour' rule (Section 3.2.3), (50), (57), (58).

However, to identify a phoneme some pre-processing needs to be done and this is to find the steady-state segment of the input pattern (frames) Section 4.2.2.1 B II explains how the phonemic segments are obtained. Once the six acoustical parameters (ZC, TE, VE, LE, ME and HE) are extracted from the input signal, and the frames are classified into the acoustical groups (Section 4.2.2.1 B and Table 4.3), these classified frames must now be segmented into phonemic segments (a group of frames which belong to the acoustical group). The actual phoneme will be recognized from these phonemic segments.

From previous studies (25), (27), (51), (59) it is acceptable to assume that the minimum duration of a phonemic segment will be 30 ms and therefore these phonemic segments will always be longer than three frames (three times 12,8 ms, 38,4 ms). Bearing this assumption in mind, whenever a frame differs from both its neighbours it is taken that this different frame is wrongly classified and it is then corrected, by classifying this frame to the same acoustical class as its neighbours.

Adjacent frames with the same acoustical label are grouped together to form a phonemic segment and segment boundaries are inserted wherever the present frame has a different label from the previous frames. In the cases where two contiguous phonemic segments belong to the same acoustical category (for example the "i"-sound /I/ and the "ou"-sound /e/ of the word "previous" /p r i: v I e s/, both are vowel-sounds), the segment boundary will be missed. However, these segments can be detected on the basis of their long duration. These long segments are divided into two equal segments and treated independently as two separate segments. The maximum duration of any segment is taken as 10 frames of 12,8 ms each. In other words, any segment longer than 128 ms is treated as two separate segments. In the case where the segment is longer than 256 ms, the system will first process the first 256 ms as two segments and then proceed to the rest of the original segment. The system will recognize these segments with the help of six more features, as one of the phonemes the system is trained on.

The steady-state segments are calculated in two steps. The middle 20% of the segment is considered as steady-state segment. Then for the vocalic sounds the average value of the second formant frequency is computed over the middle 20% of the segment. And the steady-state segment is then extended in both directions until the deviation of the formant exceeds 5% of the averaged formant value. In the cases where the steady-state segment exceeds 40% of the original segment, the middle 40% of the segment is considered as the steady-state segment. In the cases of nonvocalic phonemic segments, the middle 40% of the segment will be taken as the steady-state segments (Fig. 4.12).

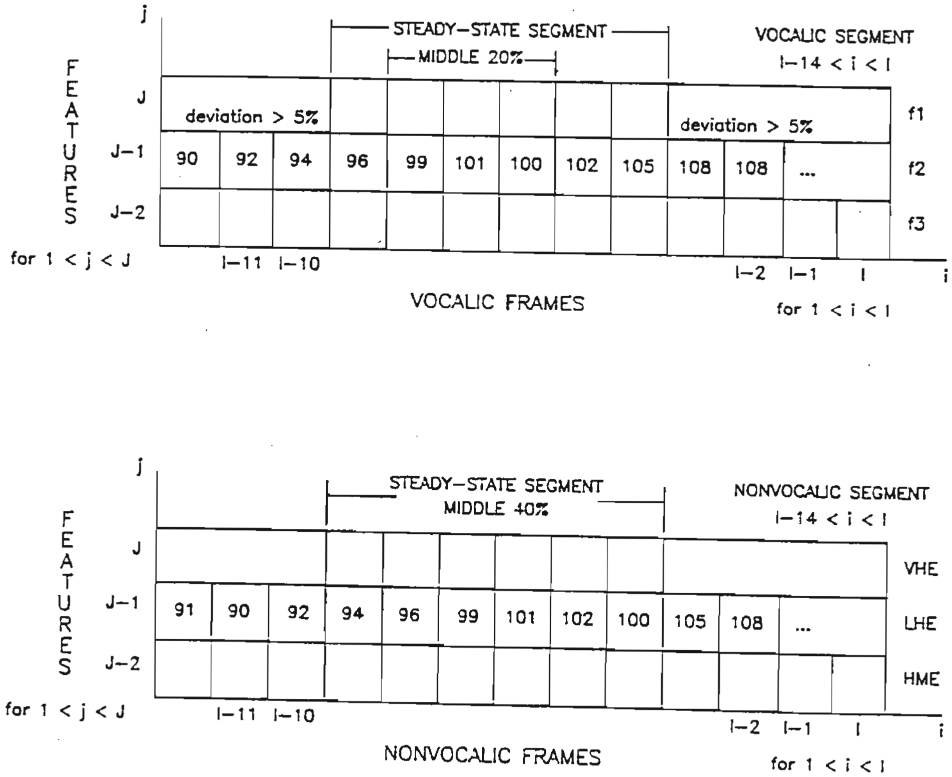


Fig. 4.12 Segmentation of the sounds

All the features of the steady-state segment are now averaged to obtain the averaged features of the original segment. These averaged parameters are used to calculate the distance measurement of the phonemic segment and are then used in the decision making process.

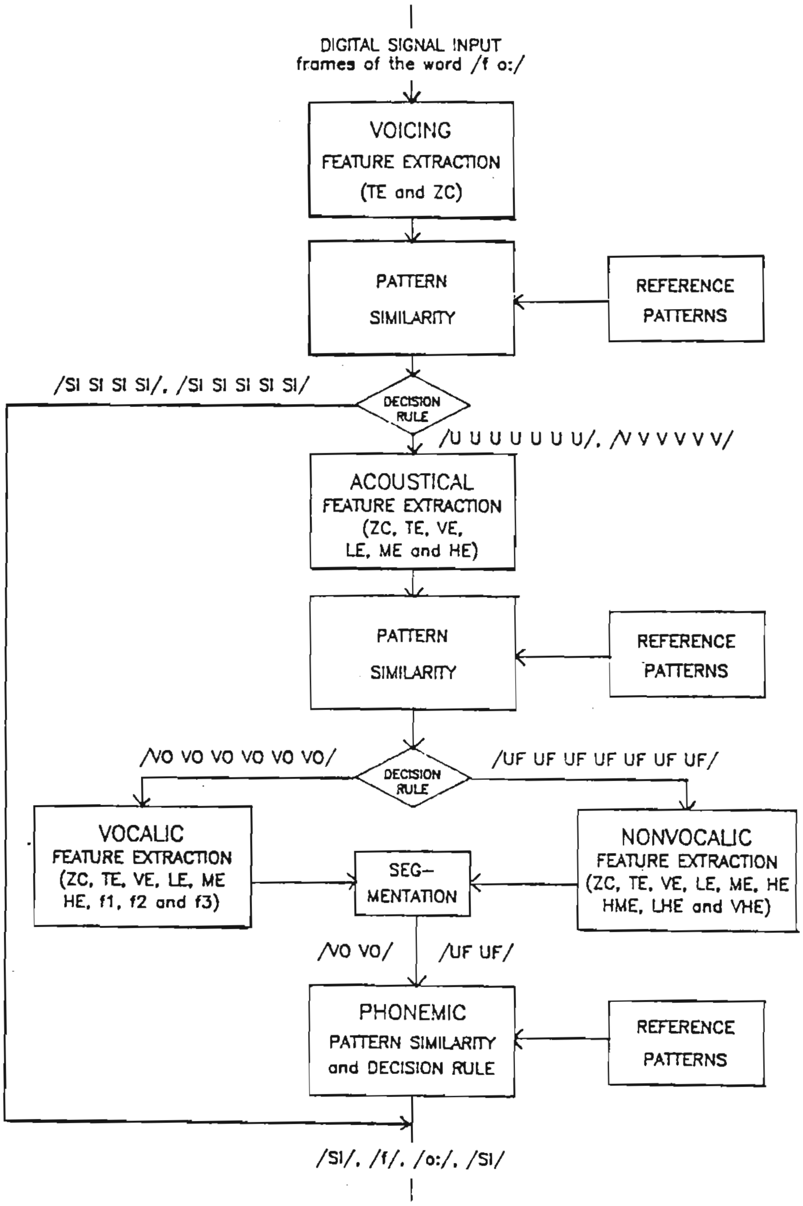


Fig. 4.13 An example of how the proposed algorithm uses three feature extractions, pattern similarities and decision rules to obtain the correct phoneme

4.2.3 Training Process

The training of the system is done with exactly the same program as the one which has been used for the recognition of the phonemes. The only difference is that the input pattern is no longer a test pattern, but a training pattern (Fig. 4.14).

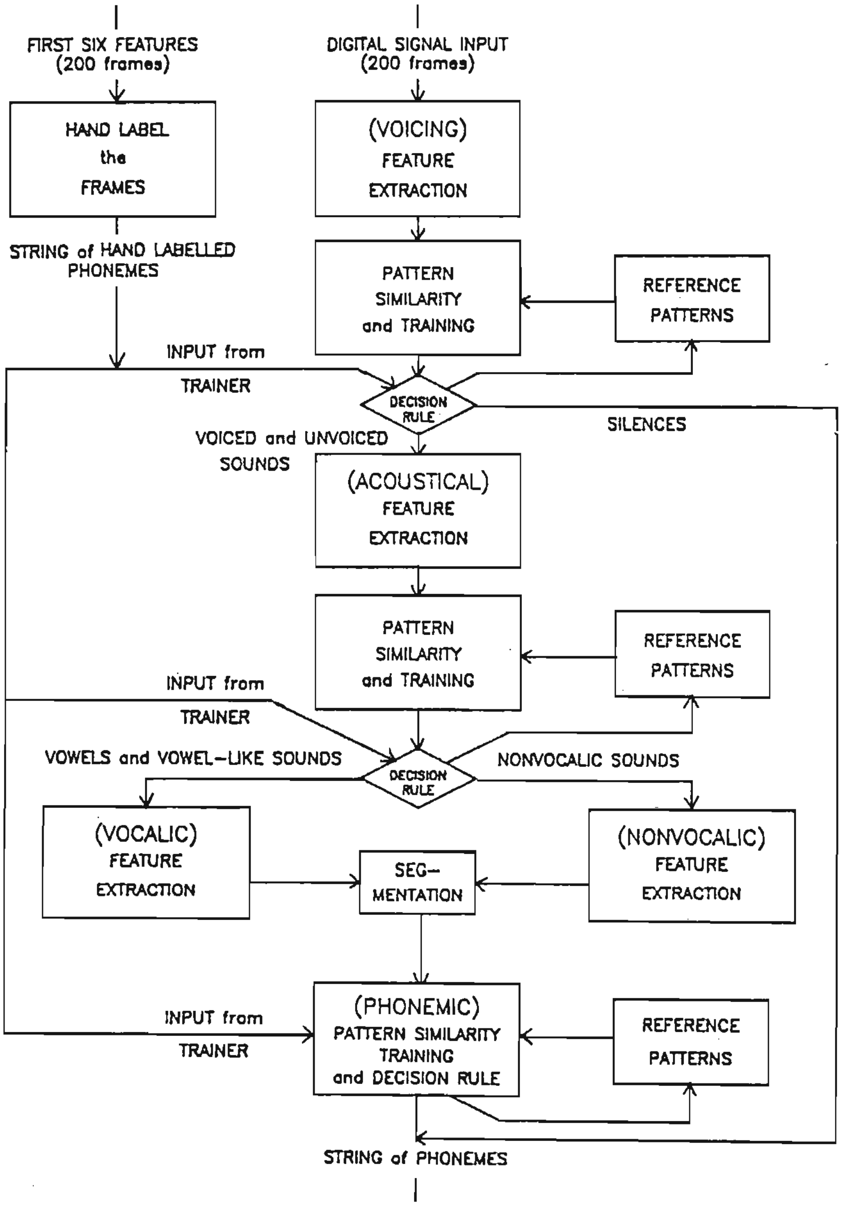

FIRST SIX FEATURES (200 frames)
DIGITAL SIGNAL INPUT (200 frames)
HAND LABEL the FRAMES
(VOICING) FEATURE EXTRACTION
STRING of HAND LABELLED PHONEMES
PATTERN SIMILARITY and TRAINING
REFERENCE PATTERNS
INPUT from TRAINER
DECISION RULE
SILENCES
VOICED and UNVOICED SOUNDS
(ACOUSTICAL) FEATURE EXTRACTION
PATTERN SIMILARITY and TRAINING
REFERENCE PATTERNS
INPUT from TRAINER
VOWELS and VOWEL-LIKE SOUNDS
DECISION RULE
NONVOCALIC SOUNDS
(VOCALIC) FEATURE EXTRACTION
SEG-MENTATION
(NONVOCALIC) FEATURE EXTRACTION
INPUT from TRAINER
(PHONEMIC) PATTERN SIMILARITY TRAINING and DECISION RULE
REFERENCE PATTERNS
STRING of PHONEMES

Fig. 4.14 The training process

In these types of systems (phoneme recognition systems (27), (49), (60), (61)) it is necessary to have a proper knowledge of each one of the training patterns. In other words every one of the 200 frames (records) of the training pattern (each training pattern like the test patterns is stored in files consisting of 200 records, Section 4.2.1) must be hand labelled before the system can be trained. This is done with the help of two programs: a graphics program and the program mentioned above.

The graphics program, written by the author, can display the actual input signal, draw a three dimensional spectro-gram of any section of the signal, average any section of the spectro display and draw it and can also display a number of the previous mentioned features the ZC, TE, VE, LE, ME and the HE (Section 4.2.2.1 A and B, Table 4.4).

Fig. 4.15 shows example graphs of the graphics program and demonstrates how the displays can help to identify first of all where the the words are and secondly where to find the actual phonemes.

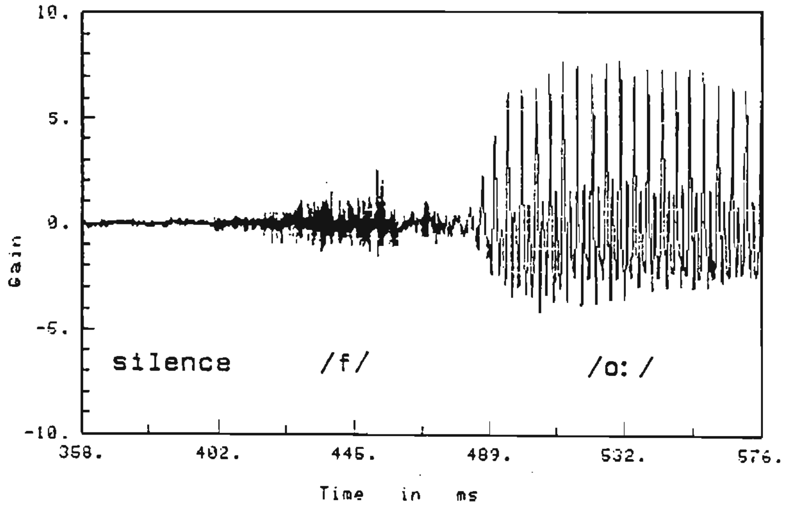


Fig. 4.15 (a) The time domain signal of the word 'four'

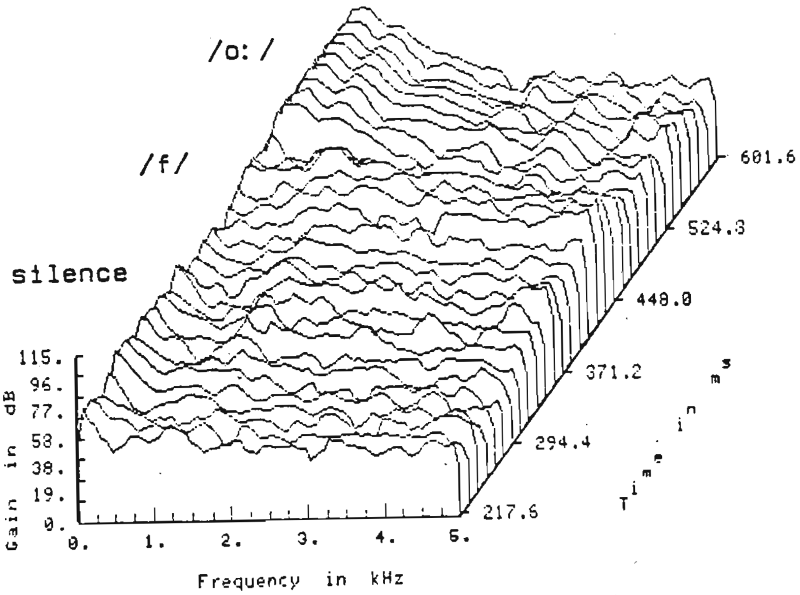


Fig. 4.15 (b) Spectra display of the word 'four'

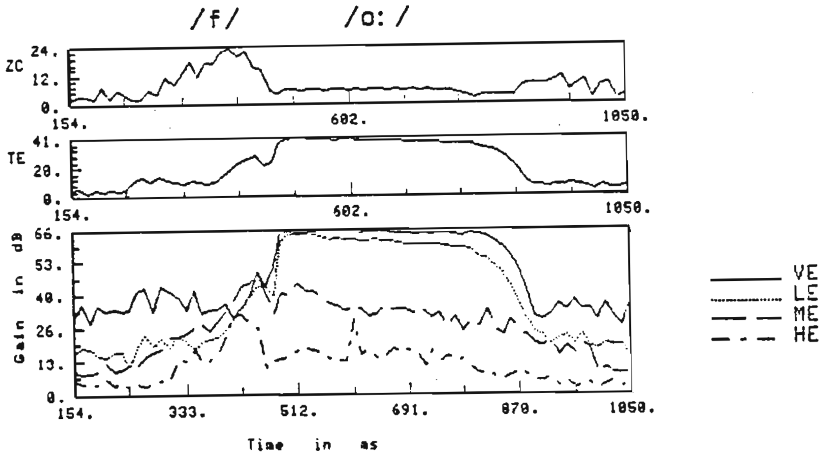


Fig. 4.15 (c) Feature display of the word 'four'

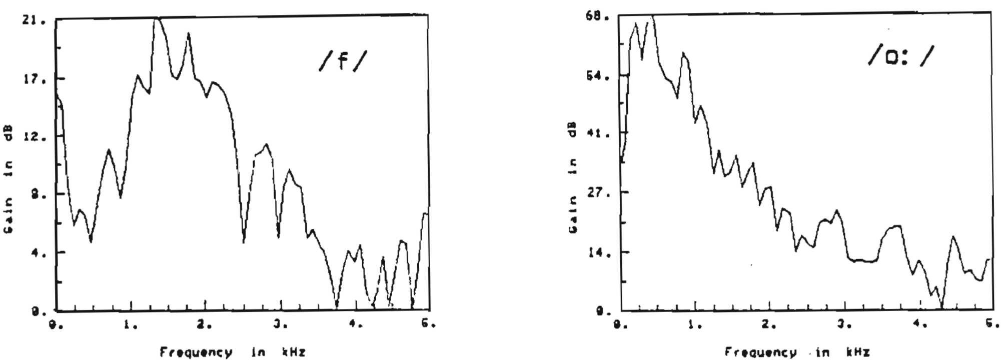


Fig. 4.15 (d) The frequency spectrums of the phonemes /f/ and /o:/ (see also Fig. 4.7 and Fig. 4.9)

After the position of the phonemes is established from the graphs, it is now necessary to label each frame. This is done by making a printout of the first six features (Table 4.6) with the frame's number next to it, with the help of the recognition program (system). The system is written in such a manner that the six features of all the frames can now be printed on either a computer terminal or a printer.

Once the printout of the first six features of the 200 frames is made, it is fairly easy to identify the phonemes (Table 4.6).

ZC	TE	VE	LE	ME	HE
2	14	48	26	19	11
2	15	49	31	22	10 Silence
4	16	50	29	26	13
13	12	41	28	32	20
19	13	49	26	34	26
18	11	41	29	33	21 /f/
22	22	34	37	45	39
23	28	46	46	52	36
14	24	50	50	47	17
4	41	73	68	48	22
6	43	72	71	52	25
5	43	72	71	48	25
6	42	73	70	45	22 /o:/
6	40	73	69	42	22
5	41	72	70	41	22
6	40	73	70	42	21
3	38	71	65	37	17
3	28	63	51	37	15
8	9	36	30	27	13 Silence
8	8	39	28	28	10

Table 4.6 Computer printout of the feature of the word "four" (these values are not the exact values, but rather proportional to the correct values)

The frames are now hand-labelled and the system is ready to be trained. Once again the procedure is done in three stages and in exactly the same way as the recognition process. Although there are two ways to achieve this, the author preferred the second method. The first method is to train all the voicing templates with numerous input patterns and for the time being to forget about the rest of the stages. Then train all the acoustical templates and finally train all the phoneme templates. The second method takes an input pattern and trains the system to recognize this input pattern in all its stages (voicing, acoustical and phoneme). In other words the system will be able to identify in which voicing and acoustical class each phoneme is and also to recognize the actual phonemes. This procedure is continued until the system can recognize and identify all the necessary voicing classes, acoustical classes and phonemes.

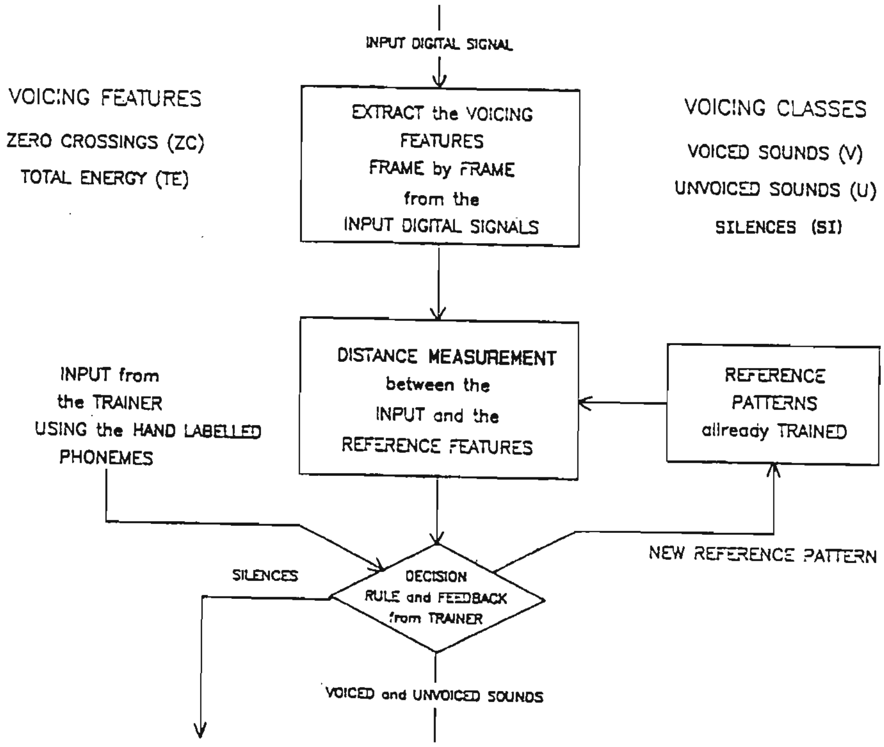

INPUT DIGITAL SIGNAL
VOICING FEATURES
ZERO CROSSINGS (ZC)
TOTAL ENERGY (TE)
EXTRACT the VOICING FEATURES FRAME by FRAME from the INPUT DIGITAL SIGNALS
VOICING CLASSES
VOICED SOUNDS (V)
UNVOICED SOUNDS (U)
SILENCES (SI)
DISTANCE MEASUREMENT between the INPUT and the REFERENCE FEATURES
INPUT from the TRAINER USING the HAND LABELLED PHONEMES
REFERENCE PATTERNS allready TRAINED
NEW REFERENCE PATTERN
SILENCES
DECISION RULE and FEEDBACK from TRAINER
VOICED and UNVOICED SOUNDS

Fig. 4.16 The training of the voicing templates

The training procedure from here onwards is very similar to that of the recognition procedure for the system, although the trainer has to do a number of extra tasks (Fig. 4.16). When the system program is run, the system is notified of the name of the training pattern and how many features must be displayed to the trainer's display unit. The system will now try to identify in which voicing category the first frame of the training pattern is, but because the templates contain zero features the system will come back and ask the trainer if these features of this frame are going to be a new template, or if they are already correctly classified. The computer always assumes that the features of the first frame are going to be a new template, for obvious reasons. The system will now ask the trainer what the name of the new template is. Once the trainer enters a name the system will then store two values (Eq. 4.13 and Eq. 4.14) for each feature of the frame as the parameters of the first template. The two values are used to obtain the weight vector and average features of the templates so that when the system is in recognition mode the weighted Euclidean distance between the templates and the input patterns can be calculated (Eq. 4.9). For each one of the features (j) the following three values are stored in the templates (i):

The sum of the squares of all features,

$$SM2_{ij} = \sum_{k=1}^{N_i} T_{ijk}^{2} \qquad 4.13$$

The sum of all the features,

$$SM_{ij} = \sum_{k=1}^{N_i} T_{ijk} \qquad 4.14$$

where N_i is the maximum number of training patterns being used to obtain the i^{th} template (reference pattern) and T_{ijk} is the j^{th} feature of the k^{th} training pattern of the i^{th} template.

Stored with these values are the name of template, the number of frames used to obtain these features and the error-limit.

a) The number of frames is the total number of frames or segments used for the training of the template is (CNT), N_i (Eq. 4.10. and Eq. 4.12).

b) The maximum weighted Euclidean distance (Eq. 4.9) calculated between the features of the training patterns and the averaged features already stored in the templates are called the error-limit (EL). When the first frame's features are stored in the template the EL is set to a preset small value, because it is impossible to calculate the Euclidean distance when the templates contain no features.

After the first frame has been trained on to a template the template is no longer empty and the Euclidean distance can be calculated (Eq. 4.9). Every time a new frame is trained onto a template the Euclidean distance between the features of this frame and the parameters of the current template is stored as the new EL (error-limit) if this distance is more than the previous EL. This is only done when the trainer told the system that the new frame is actually of the same type as the template it was matched to. In the case where the system thinks the the new frame is of the same type as the template, the system can be corrected by the trainer by either specifying the correct template's name, or telling the system that the frame is a completely new template. The system will then ask the trainer what the name of the new template is and then create a new template with the name the trainer specified in exactly the same way it was explained above.

For example, if the frame to be trained, has a minimum distance to the voiced template and this distance is less than the EL of the voiced template, the present frame will be recognized as a voiced frame and the system will not ask the trainer whether the frame is an old or new template. The system will accept that the frame is voiced (Fig. 4.19). On the other hand, if the minimum distance between the frame's features and the template's parameters is more than EL and the template is the correct one, the system will take the features of the frame and use it to update the parameters of the template by adding the squared features to SM2, add the features to SM, increment CNT and store this new minimum distance as the new EL. It must be remembered however, that each feature has its own three parameters which must be updated. But when the trainer corrects the system and tells it that the frame is actually an unvoiced sound, the system will create a new template.

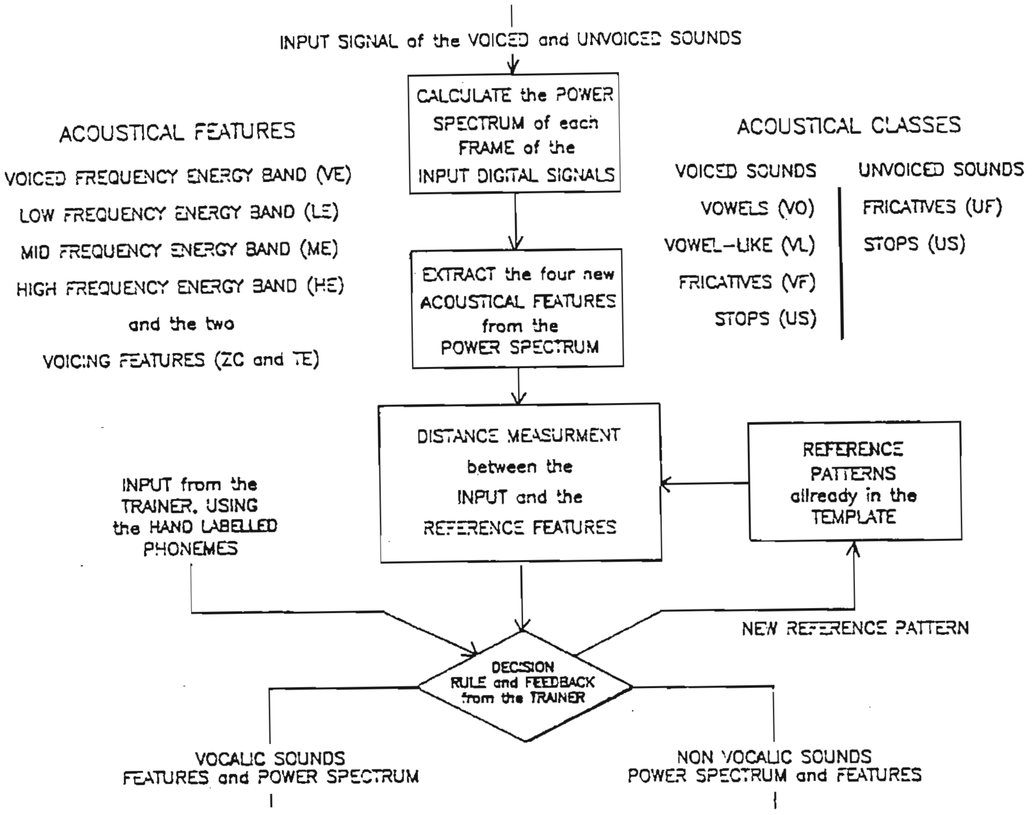

INPUT SIGNAL of the VOICED and UNVOICED SOUNDS
CALCULATE the POWER SPECTRUM of each FRAME of the INPUT DIGITAL SIGNALS
ACOUSTICAL FEATURES
VOICED FREQUENCY ENERGY BAND (VE)
LOW FREQUENCY ENERGY BAND (LE)
MID FREQUENCY ENERGY BAND (ME)
HIGH FREQUENCY ENERGY BAND (HE)
and the two
VOICING FEATURES (ZC and TE)
ACOUSTICAL CLASSES
VOICED SOUNDS
VOWELS (VO)
VOWEL-LIKE (VL)
FRICATIVES (VF)
STOPS (US)
UNVOICED SOUNDS
FRICATIVES (UF)
STOPS (US)
EXTRACT the four new ACOUSTICAL FEATURES from the POWER SPECTRUM
DISTANCE MEASURMENT between the INPUT and the REFERENCE FEATURES
REFERENCE PATTERNS allready in the TEMPLATE
INPUT from the TRAINER, USING the HAND LABELLED PHONEMES
NEW REFERENCE PATTERN
DECISION RULE and FEEDBACK from the TRAINER
VOCALIC SOUNDS FEATURES and POWER SPECTRUM
NON VOCALIC SOUNDS POWER SPECTRUM and FEATURES

Fig. 4.17 Training of the acoustical templates

The next step in the training procedure of the specific frame is to train the system to identify what type of acoustical class the frame is (Fig. 4.17). Exactly the same method is used as mentioned above. The only difference is that more parameters are used in each one of the acoustical templates, and of course there are more acoustical templates than voicing templates.

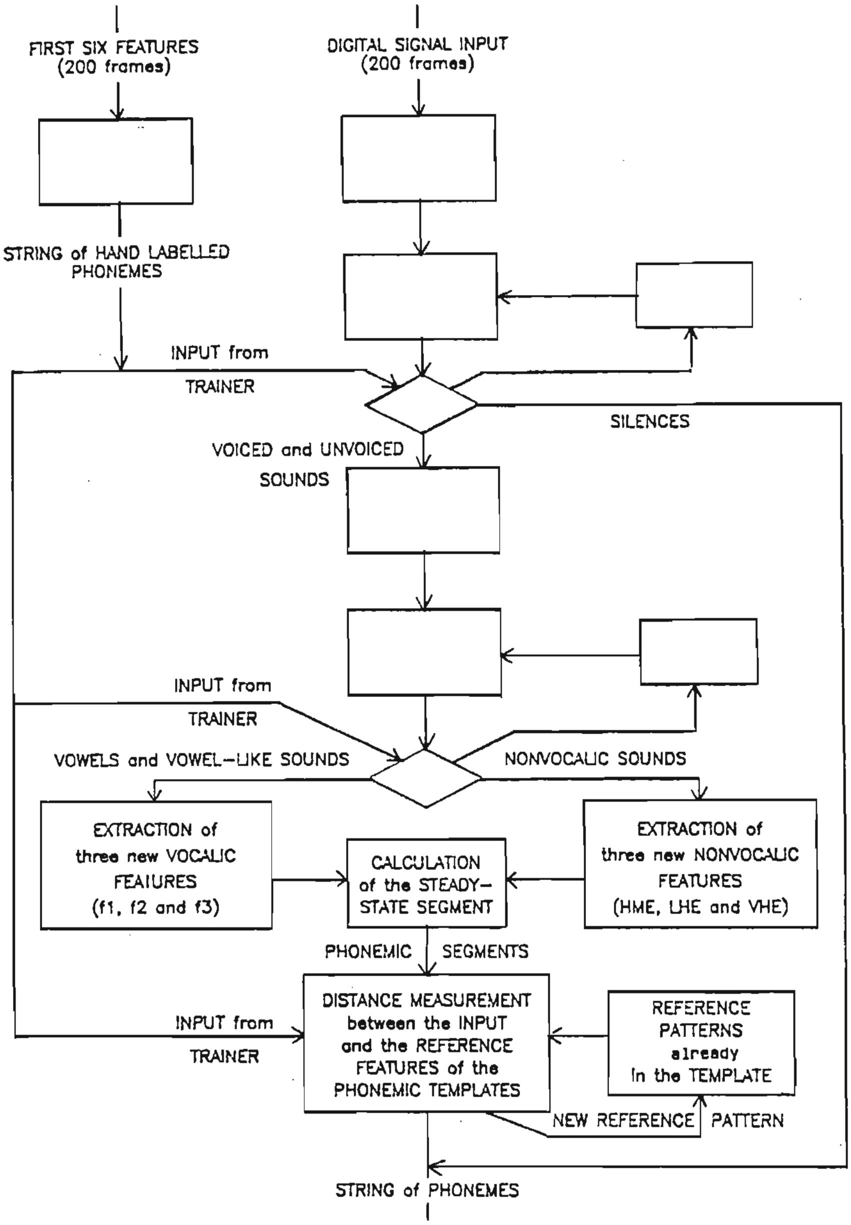

FIRST SIX FEATURES
(200 frames)
DIGITAL SIGNAL INPUT
(200 frames)
STRING of HAND LABELLED PHONEMES
INPUT from TRAINER
SILENCES
VOICED and UNVOICED SOUNDS
INPUT from TRAINER
VOWELS and VOWEL-LIKE SOUNDS
NONVOCALIC SOUNDS
EXTRACTION of three new VOCALIC FEATURES (f1, f2 and f3)
CALCULATION of the STEADY-STATE SEGMENT
EXTRACTION of three new NONVOCALIC FEATURES (HME, LHE and VHE)
PHONEMIC SEGMENTS
INPUT from TRAINER
DISTANCE MEASUREMENT between the INPUT and the REFERENCE FEATURES of the PHONEMIC TEMPLATES
REFERENCE PATTERNS already In the TEMPLATE
NEW REFERENCE PATTERN
STRING of PHONEMES

Fig. 4.18 Training of the phonemic templates

The final step is the training of the phoneme templates. Before the phoneme templates can be created, the steady-state segment of the phonemical segment must be calculated (Section 4.2.2.3 and Fig. 4.18). This steady-state segment consists of a number of frames of the same acoustical class and the features of these frames are then averaged to obtain a single set of features. The rest of the training procedure is again the same as the training of the voicing templates, the only differences are that once again more features have been used and the input pattern is more than just a single frame. It is an averaged frame.

Let us continue with the example of the frame which was labelled 'voiced' (Fig. 4.19). The system now extracts the four acoustical features and then tries to identify in which acoustical class the frame is. In the case where the acoustical templates are still empty, the system will ask the trainer what the acoustical name of the of the present frame is. The system will show on command the six features of the specific frame so that the trainer can more easily identify the frame. The system will now create the first acoustical template. In this case the frame could have been a vowel sound.

But when the acoustical templates are not empty and the system identifies the frame as a vowel sound, but the Euclidean distance between the template's parameters and frame's features is more than the EL of the template, the system will ask the trainer if this frame is correctly identified or if the frame belongs to a new acoustical group which has not yet been trained. On the answer that the frame is correctly identified, the system will update the parameters of the template (six SM2's, six SM's, the CNT, the EL and the name of the template). If the trainer specifies another template, the system will update the parameters of the other template with these new features. In the case where the trainer tells the system that the frame belongs to a new class, say vowel-like, the system will create a vowel-like-template with this frame's features.

With the addition of the three phonemic features (f1, f2 and f3) the average steady-state segment can be recognized as the phoneme /o:/. When the phonemic templates are still empty the system will ask the trainer what the phoneme name of the segment is and then create the /o:/-template. As in the previous cases, once the templates contain parameters, the system will only recognize a segment to a template when the Euclidean distance is less than EL. In the cases where this is not the case, the system will again ask the trainer what to do. The system will update the template if the segment was correctly recognized, otherwise the system will create a new template or update the correct template.

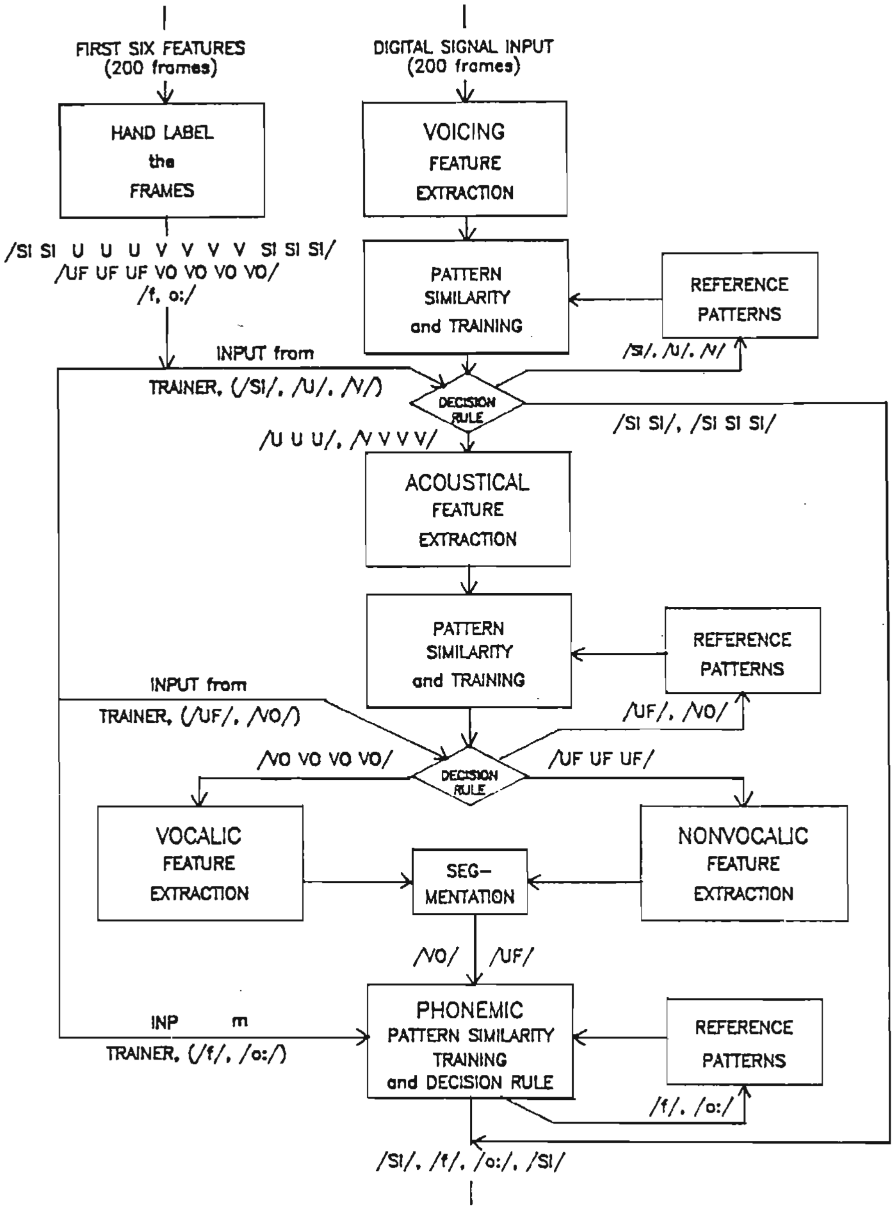


Fig. 4.19 Example of how the training procedure is done

4.2.4 Phoneme to Text Conversion

Once the recognition mode of the system recognizes the different phonemes of the test pattern, these phonemes are stored in a file. This file containing the phonemes is then read by a phoneme to text conversion program which will convert the phoneme strings to a more readable form. The program is actually part of the decision making process, but because the program is not part of the intial system it is dealt with in this section. Unlike the system which can be used to recognize any phoneme that is trained on the system, this conversion program can only be used to convert the testing words ("zero" to "nine") to text. To enlarge the capabilities of the program is not a difficult task. The conversion program works in such a mannar that the conversion of the phoneme is sometimes dependent on the previous phoneme.

Let us consider the three words which contain the "n"-sound /n/,

a) one, /w A n/,
b) seven, /s E v e n/ and
c) nine, /n aI n/ (Table 4.2).

From these three words it is clear that to convert the phoneme /n/ to text it is necessary to have two conversion rules. The first rule is to add an "e"-letter onto the "n"-letter when the phoneme before the /n/ is an /A/ or an /aI/. The second rule is that all the other /n/-phonemes are directly converted to "n"-letters.

The algorithm for the conversion of the /n/ phoneme to text can be seen in Fig. 4.20.

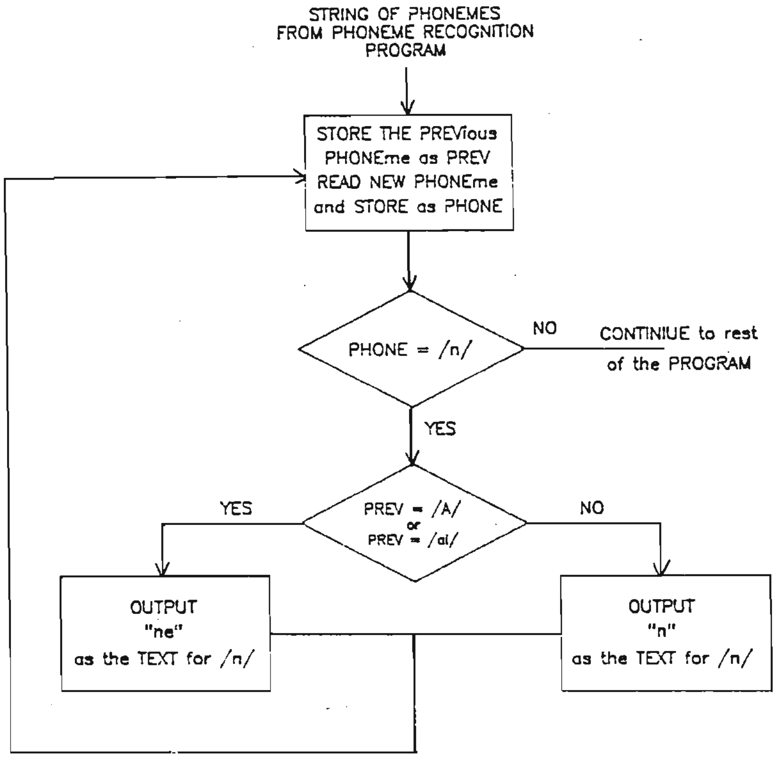


Fig. 4.20 Conversion of the /n/ phoneme

This program has other functions as well and one is to merge phonemes which follow similar phonemes into one phoneme. Another feature of the program is that silences are denoted by spaces between the words, except in the cases where the silence is followed by a plosive-sound, for example the /t/ phoneme in the word "eight" (Section 2.1.3.3 C).

Example 4.2 is:

input phoneme string to the conversion program,
/silence/, /eI/, /eI/, /eI/, /silence/, /t/,

after merging,
/eI/, /t/,

output of the program,
ei, ght (Table 4.7).

The output of this program is a more readable one than the phonemic transcriptions of the recognition system. It must be noted however, that the system up to this stage has no knowledge of the meaning or even the correct spelling of any word. The program only works according to the rules and is unable to detect phonemes which are deleted or even phonemes which are inserted.

Example 4.3 is:

If the input phoneme string of the program is the following,

/silence/, /f/, /f/, /o:/, /o:/, /u:/, /silence/,

and after the merging,

/f/, /o:/, /u:/,

the output will be,

f, our, wo (Table 4.7).

This method of converting sounds (phonemes) to text is a very common process which every one of us uses when we are writing a word down on paper, especially the words with a difficult spelling. We do not remember the spelling of the word, we merely use some simple spelling rules to convert the sound of the word to text.

PREVIOUS PHONEME	CURRENT PHONEME	TEXT OUTPUT
	/aI/ five	i
	/I/ six	i
	/E/ seven	e
	/eI/ eight	ei
	/i:/ three	ee
	/Ie/ zero	e
	/eu/ zero	o
	/o:/ four	our
	/u:/ two	wo
	/w/ one	nil
	/k/ six	nil
/w/	/A/ one	o
/k/	/s/ six	x
/A/	/n/ one	ne
/aI/	/n/ nine	ne
/aI/	/v/ five	ve
	/n/ nine	n
	/v/ seven	v
	/th/ three	th
/eI/	/t/ eight	ght
	/t/ two	t
	/r/ three	r
	/f/ four	f
	/s/ six	s

Table 4.7 Conversion rules for the phonemes of the words "zero" to "nine"

5 Recognition Rates of Various Sections of the System

5.1 Introduction

In order to obtain the recognition rates it is necessary to have a set of words to train and another set of words to test the system on (Section 4.2.1). Unfortunately it was impossible to capture and test the data in real- time using the HP CAMAC system and therefore the amount of training and testing of words was severely limited. At the time the system was developed and tested no ADC was readily available to capture the data. However, the HP CAMAC system proved to be reliable but rather tedious to use and speech samples were limited to 2,56 s. For this reason it was decided to digitize the words "zero" to "nine" (Section 4.2.1) and store them in 20 data files. Twenty of these words were used for the training procedure and all forty words for the testing of the system (Table 4.2).

5.2 Voicing Recognition Results

For the detection of the voicing class of the words two features were used, the zero-crossings (ZC) and the total energy (TE) (Section 3.2.1.3 and Section 4.2.2.1).

The averaged values of these two features of the three voicing classes can be seen in Fig. 5.1 and Table 5.1 . These values were used as the voicing templates (Table 5.1).

VOICING	FEATURES			
	ZC		TE	
	mean	deviation	mean	deviation
SI1	5	1	13	8
SI2	16	2	11	4
U	19	5	19	7
V	8	3	32	6

Table 5.1 Voicing features (TE is measured in dB)

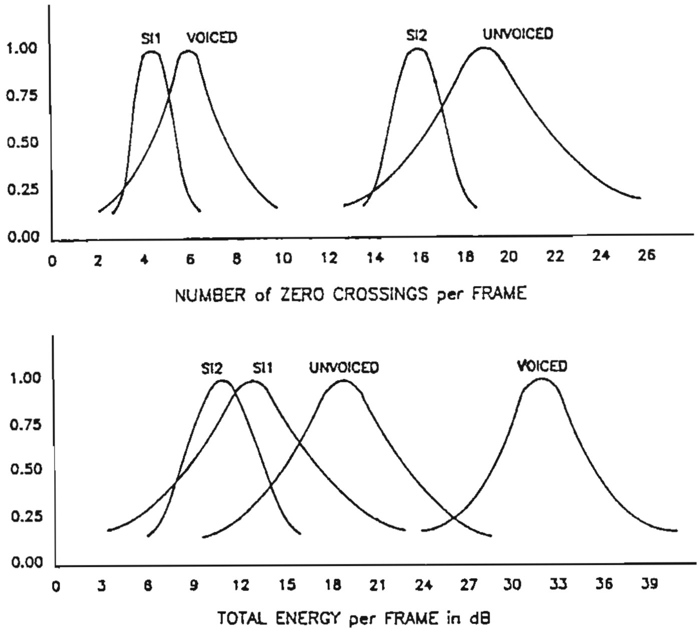


Fig. 5.1 The mean and standard deviation values of the features for the different classes.

Two types of silences are tabulated in Table 5.1. Silence (SI1) has a very low ZC rate of about 470 Hz. This frequency correlates very well with results of experiments done by L R Rabiner and M R Sambur (37) taking into account the fact that these experiments counted both the negative and positive zero crossings unlike this thesis (Section 4.2.2.1 A, Eq. 4.2).

The other silence (SI2) is a soundless whisper (Section 2.1.3.3 D 3). These whispers are usually present after a voiced sound (Fig. 5.2). It seems that the speaker's vocal cords stop vibrating (Section 2.1.2.1) but the speaker is still exhaling air and thus generating a whisper. Normally whispers consist of high frequencies, and therefore the ZC rates of silence (SI2) are higher than that of voiced sounds (Fig. 5.1).

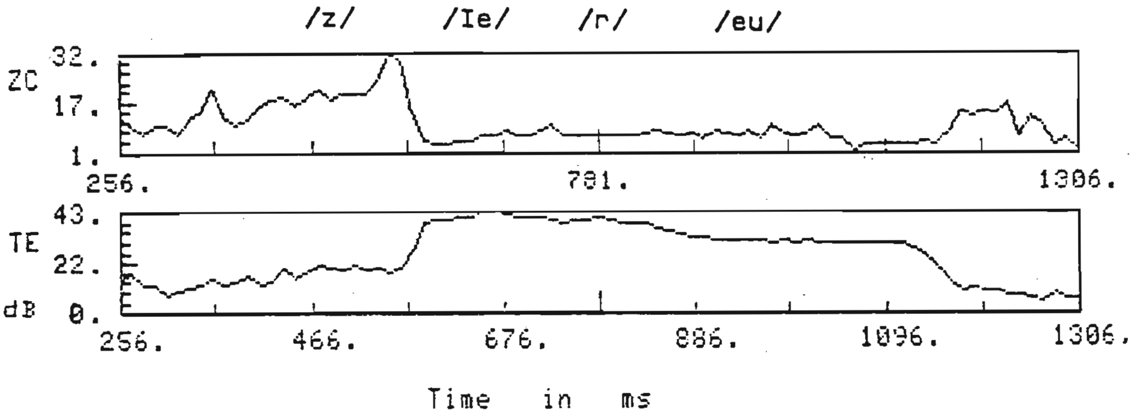


Fig. 5.2 ZC and TE of the word "zero", /z Ie r eu/

186 segments were tested and of these only 17 voicing segments were completely misrecognized giving a voicing recognition rate of 90,8%. This rate does not take account of extra silences which are detected in the decision rule section of the system (Section 3.3.4). Furthermore the /z/ was trained as a unvoiced sound, because the speaker's /z/ (ZC=19, TE=21 dB, Fig. 5.2) had similar features to the unvoiced sounds (ZC=19, TE=19 dB). To try and train the /z/ as a voiced sound would decrease the recognition rate.

HAND LABELLED	RECOGNIZED		
	SI	U	V
SI	42	-	1
U	7	34	-
V	6	3	93

Table 5.2 The voicing recognition rate

5.3 Acoustical Recognition Rates

Six features (TE, ZC, VE, LE, ME and HE) are used to recognize the six acoustical classes (Section 4.2.2.1 B and Table 4.3). In Table 5.3 the values of the acoustical templates were used to retrieve the acoustical classes of the frames. These values were calculated by the systems during the training process (Eq. 4.11 and Section 4.2.2.1).

	ZC		TE		VE		LE		ME		HE	
	mean	dev	mean	dev	mean	dev	mean	dev	mean	dev	mean	dev
VF1	5	3	31	9	63	10	54	11	40	4	33	4
VF2	22	1	20	1	45	4	42	6	42	1	32	1
UF	17	6	24	4	48	12	47	7	45	4	34	6
VL	5	1	33	6	66	5	55	9	41	7	20	5
VO	6	1	42	1	72	1	68	2	49	1	34	2

dev - deviation

Table 5.3 Values used in the acoustical templates (energy bands are measured in dB)

The fricative /v/ was classified as VF1 and fricative /z/ as VF2. The voiced fricative /z/ and the unvoiced stops ,/t/ and /k/, were trained as unvoiced fricatives, because the error distance (Section 4.2.2.2) between the unvoiced sounds and these sounds was too small for the sounds to be unique.

Fig. 5.3 shows the different features of some of the phonemes that were trained as unvoiced fricatives.

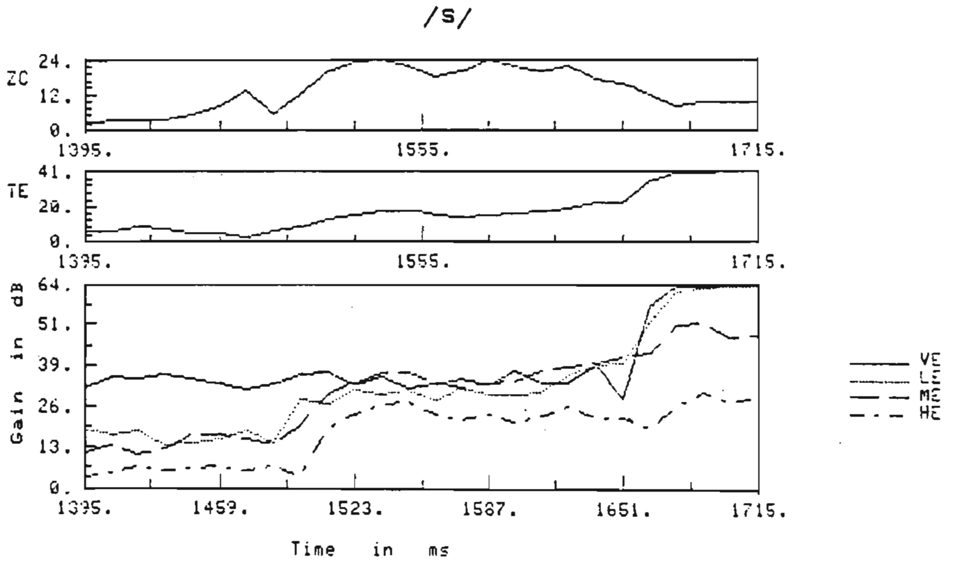


Fig. 5.3 (a) The features of the unvoiced phoneme /s/

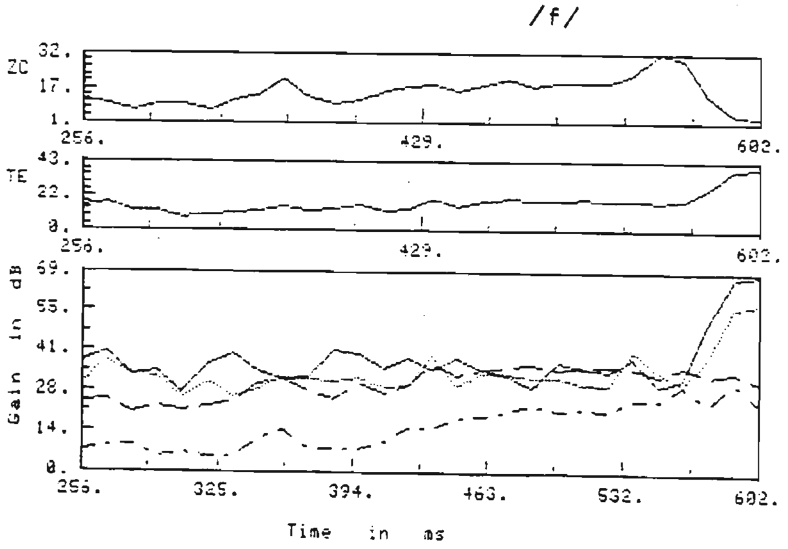


Fig. 5.3 (b) The features of the unvoiced phoneme /f/

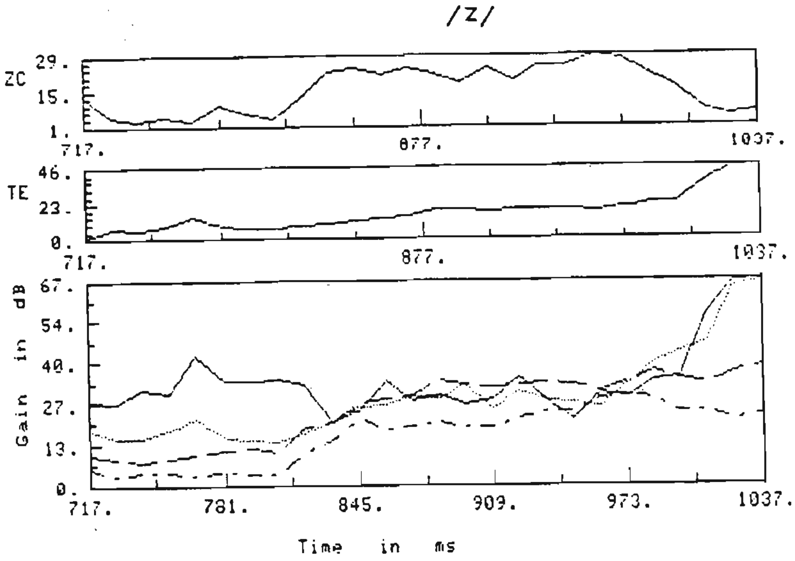


Fig. 5.3 (c) The features of the phoneme /z/

With this method of combining the unvoiced stops, the one voiced fricative (VF2) and the unvoiced frictives as one acoustical class, produced an acoustical recognition rate of 82% (Table 5.4). It must be noted that only the acoustical classes of the phonemes in the words "zero" to "nine" have been used in the training process (Compare Table 4.2 with Table 4.3).

HAND LABELLED	RECOGNIZED			
	VO	VL	VF	UF
VO	55	1	3	-
VL	10	10	8	-
VF	-	1	8	-
UF	-	-	-	34

Table 5.4 The confusion matrix of the acoustical classes

From the results in Table 5.4 it can be seen that the VL class segments have the lowest recognition rate of 64%. The reason for this is probably that the variations in the features of VL phonemes vary widely and therefore more distinct features must be used.

In Section 2.2.4 it is mentioned that the neurons at the genticulate level of the hearing process can detect two different acoustical classes. Therefore a possible solution might be to reduce the number of acoustical classes used in the system.

5.4 Recognition rates of the Phonemes

To improve the phoneme recognition rate the system was trained and tested with three different sets of templates, the vowels (Template1 with 11 vowels), the vowel-like phonemes and the voiced fricatives (Template2 with 5 phonemes) and also the unvoiced fricatives (Template3 with 5 phonemes).

5.4.1 Vowels (Template1)

A total of nine parameters (ZC, TE, VE, LE, ME, HE, and the first three formant frequencies f1, f2 and f3) were used to train and recognize a vowel, as previously mentioned in Section 4.2.2.1 C (II). Results obtained from the training process (Table 5.5 and Fig. 5.4) showed that the formant frequencies compare well with the results in Fig. 2.4 and Fig. 2.5 in Chapter 2.

	first formant	second formant
i:	301	2 690
I	461	2 306
E	538	1 845
o:	538	768
e	461	1 306
u:	307	922
A	691	1 076
aI	384	1 846
eI	691	2 229
Ie	304	1 306
eU	768	846

Table 5.5 The first two formant frequencies in Hz

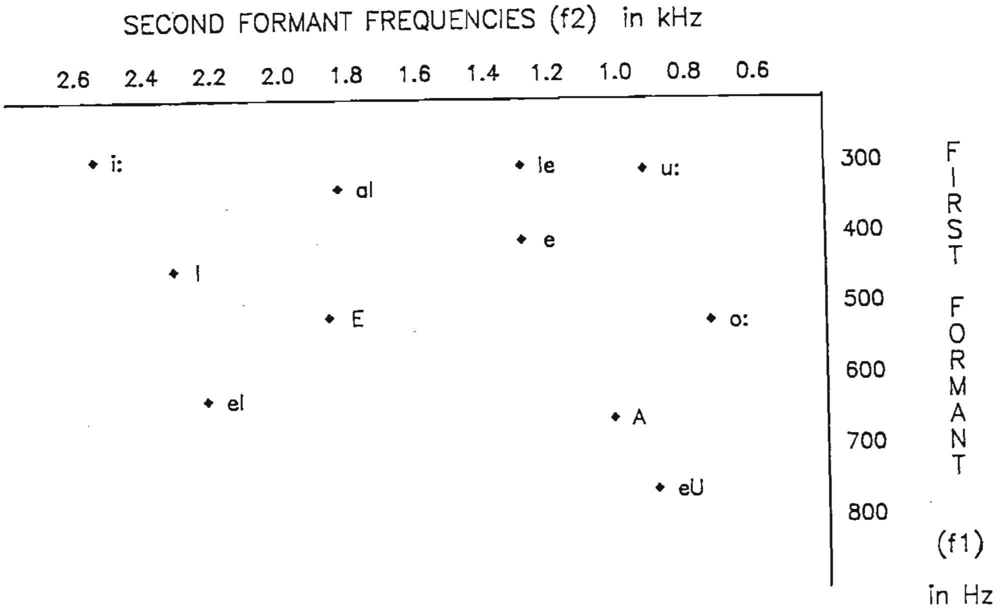


Fig. 5.4 Formant Frequencies (f1 against f2)

The confusion Table 5.6 shows a vowel recognition rate of 71,4%.

	RECOGNIZED									
HAND LABELLED	Ie	eU	A	u:	I	o:	aI	E	eI	e
Ie	1	-	-	-	-	-	-	-	-	1
eU	-	2	-	1	-	-	-	-	-	-
A	-	1	1	-	-	-	1	-	-	-
u:	-	-	-	3	1	-	1	-	-	-
I	-	-	-	-	5	-	-	-	1	-
o:	-	-	-	-	-	4	-	-	-	-
aI	-	-	-	-	-	-	5	-	-	-
E	-	-	-	-	-	-	-	1	2	-
eI	-	-	-	-	-	-	-	-	3	-
e	-	-	-	-	-	-	-	-	-	-

Table 5.6 Results of the vowel recognition rate

5.4.2 Vowel-like and Voiced Fricative Phonemes (Template2)

The vowel-like sounds use exactly the same features as the vowels (Section 5.4.1), but the voiced fricatives use three other features (HME, LHE and VHE) in place of the three formants (Section 4.2.2.1 C (II) and Table 4.4). Results obtained from tests showed that the VL and VF recognition rate was 75.6% and are tabulated in Table 5.7.

	RECOGNIZED				
HAND LABELLED	z	v	r	w	n
z	3	1	-	-	-
v	1	5	-	-	2
r	-	5	4	-	-
w	-	-	-	4	-
n	-	-	-	-	12

Table 5.7 Confusion matrix of the VL and the VF phonemes

The worst confusion was between the /r/ and the /v/. This was actually an error that repeated itself, the /r/ in the word "three" was constantly recognized as a /v/, because the /r/ in "three" follows the unvoiced /th/ phoneme which contains a great amount of high frequency energy (Fig. 4.9) and some of this high frequency energy continues into the /r/. On the other hand, the /r/ in the word "zero" follows a voiced sound /Ie/ which does not contain the high frequency energy and therefore is not recognized as a voiced fricative. Thus to improve the recognition rate the system can be trained with two different /r/ templates or the phoneme to text conversion program can correct the error.

5.4.3 Unvoiced Stops and Fricatives (Template3)

The features used for the recogintion of the US and UF are similar to those of the voiced fricatives (Section 5.4.2 and Table 4.4). It must be noted that because of the small error distance between the two acoustical classes (US and UF) all the phonemes were trained as UF. A recognition rate of 71,8% was calculated from the test results (Table 5.8).

	RECOGNIZED				
HAND LABELLED	th	f	s	t	k
th	5	-	-	-	-
f	-	4	6	-	-
s	-	1	13	-	-
t	1	-	4	3	-
k	-	1	-	-	3

Table 5.8 Test results of the UF phonemes

Two cases given in the above results have poor recognition rates, the /f/ in "four" and the /t/ in "two". In both cases they were recognized as a /s/ phoneme. These common errors can be eliminated in the phoneme to text conversion program (Section 5.5). The average phoneme recognition rate is in the region of 72,3% and with the help of the phoneme to text conversion program it can be improved.

5.5 Phoneme to Text Conversion

The recognized phonemes were entered into the conversion program and the results are given in Table 5.9 (Examples of twenty words are tabulated).

PHONEME INPUT		TEXT OUTPUT	
/z e r eU/,	/z Ie r eU/	zero,	zero
/w eU n/,	/w A n/	on,	one
/s u:/,	/s u:/	swo,	swo
/th v i:/,	/th v i:/	thvee,	thvee
/s o:/,	/s o:/	sour,	sour
/f aI v/,	/f aI v/	five,	five
/s I k f/,	/s I k s/	sikf,	six
/s E v n/,	/s eI v n/	sevn,	seivn
/eI t/,	/eI t/	eight,	eight
/n aI n/,	/n aI n/	nine,	nine

Table 5.9 The input and output of the conversion program

There is an error in ten of the words in Table 5.9 and this gives a very low word recognition rate of 50%. However, this is not a true indication of the system's performance, because in most cases there is only one letter wrong in the word. If the conversion program is now changed to cater for common errors, like the /t/ in "two", the /f/ in "four", the /r/ in "three" and the /e/ in "seven", the word recognition rate could be as high as 85% (Table 5.10).

FOUR COMMON ERRORS	CORRECT WORDS
swo	two
thvee	three
sour	four
sevn	seven

Table 5.10 The common errors

6 Conclusion and Recommendations

6.1 Conclusions

Speech recognition systems have been investigated and a continuous phoneme recognition algorithm has been developed. The system was developed and tested using a HP 1000 mini-computer.

The algorithm used a 128-point Fast Fourier Transform to calculate the frequency spectrum of the speech signals. A fixed frame length of 12,8 ms was used and the sample rate was 10 kHz. Two time domain features were calculated from the frames and the rest of the features from the frequency spectrum of the speech signal.

Before a phoneme could be recognized, the voicing and acoustical class of the phonemes had to be identified, by using different features, like zero-crossings and frequency energy bands. The recognition tests done with continuous phoneme strings showed that the voicing class recognition rate was 90,8% and the acoustical classes were 82,0% correctly classified. This acoustical recognition rate is a reasonable score compared with other phoneme recognition systems (83,9% (62) and 86,1% (63)).

Recognition tests using small reference templates produced an overall phoneme recognition rate of 72,3%. Although this score compares very well with scores of other systems (27), (50), the score is less than expected because of the fairly small templates. On the other hand it must be remembered that the amount of data used for training and testing purposes was limited and therefore the values in the templates are not a very good representation of the phoneme's features.

With the help of a phoneme to text conversion program it would be possible to have word recognition rate of about 80,0%. This score cannot be compared with isolated recognition systems, because they do not use continuous speech.

The system takes 35 seconds to recognize one data file which contains 2,56 seconds of speech. The system thus takes about 14 times real time to process the speech.

This proposed system demonstrates that phoneme recognition could be the answer to the speech recognition problem and isolated practical problem areas.

6.2 Recommendations

The most important recommendation is the need for a data acquisition system that is readily available and that can capture more than 2,56 seconds of speech at a time. Such a system is now available in the Department and this will help any future research in speech signal processing it the Department tremendously.

The voicing recognition score is relatively low. If this score can be improved to very near 100,0% the total recognition rate will also improve significantly. Other methods of obtaining the voicing class of the phonemes might be investigated.

The frame's features must be optimised so that the best frequency energy-bands are used and thus improve the acoustical recognition score. There must be an investigation into how the auditory system classifies these sounds and what features are the most important ones. The proposed recognition system tries to classify the sounds into acoustical classes like the human thinks the sounds should be classifed and not like the auditory system classifies the sounds. Once the recognition system can identify the phonemes' acoustical classes correctly, the phonemes will be more easily recognised.

In the recognition of the phonemes it might be a good idea to look at the previous phonemes. The pronounciation of a phoneme can change depending on its predecessor. The transition stage between the two phonemes can play a promising role in the recognition of the phonemes. Again it is necessary to find the most unique features for the recognition of the phonemes. The frequency energy bands and the sample rate of speech signal could be changed to notify their effect on the recognition rate. New methods of calculating the frequency spectrum and formant frequency could be developed to simulate the cochlea of the inner ear.

To summarise, it would be very useful to do more investigation on what features are used in the recognition process of the auditory system and how the phonemes are classified. The importance of the transition stage between two phonemes must also be investigated. The phoneme to text program can be extended to be able to convert more than just the phonemes of the words "zero" to "nine" to text.

Appendices and References

A

Pronunciation Key of the Phonemes

The symbols used in the pronunciation transcriptions and in the rest of the thesis are an orthographic representation based on the International Phonetic Alphabet. The following consonant symbols have their usual English values: b, d, f, h, k, l, m, n, p, r, s, t, v, w, z. The other symbols and their interpretation are listed in the table below (11), (14).

INTERNATIONAL	ORTHOGRAPHIC	INTERPRETATION	
PHONETIC ALPHABET	REPRESENTATION		
/a:/	/a:/	father	/f a: the e/
/æ /	/ae/	act	/ae k t/
/aI/	/aI/	dive	/d aI v/
/aIə/	/aIe/	fire	/f aIe/
/aʊ/	/aU/	out	/aU t/
/aʊə/	/aUe/	flour	/f l aUe/
/ɛ/	/E/	bet	/b E t/
/eI/	/eI/	paid	/p eI d/
/ɛə/	/Ee/	bear	/b Ee r/
/g/	/g/	get	/g E t/
/I/	/I/	pretty	/p r I t I/
/i:/	/i:/	see	/s i:/
/Iə/	/Ie/	fear	/f Ie/
/j/	/j/	yes	/j E s/
/ɒ/	/o/	pot	/p o t/
/əʊ/	/eU/	note	/n eU t/
/ɔ:/	/o:/	organ	/o: g e n/
/ɔI/	/oI/	void	/v oI d/
/ʊ/	/U/	pull	/p U l/
/u:/	/u:/	zoo	/z u:/
/ʊə/	/Ue/	poor	/p Ue/

INTERNATIONAL PHONETIC ALPHABET	ORTHOGRAPHIC REPRESENTATION	INTERPRETATION	
/ə/	/e/	potter	/p o t e r/
/ɜ:/	/E:/	fern	/f E: n/
/ʌ/	/A/	cut	/k A t/
/ʃ/	/sh/	ship	/sh I p/
/ʒ/	/zh/	closure	/k l eU zh e/
/tʃ/	/tsh/	chew	/tsh u:/
/dʒ/	/dzh/	jaw	/dzh o:/
/θ/	/th/	thin	/th I n/
/ð/	/the/	these	/the i: z/
/ŋ/	/ng/	sing	/s I ng/
/^{ə}/	/^{e}/	bundle	/b A n d ^{e} l/

B FORTRAN Program of a Fast Fourier Transform Algorithm

This program is based on an algorithm proposed by Oppenheim and Schafer (32). The digital input signal must be stored in a Common Complex aray x before the subroutine FFT is called, the output of the FFT subroutine will be placed in the same Complex aray x. A 128-point radix-2 FFT was used in this case.

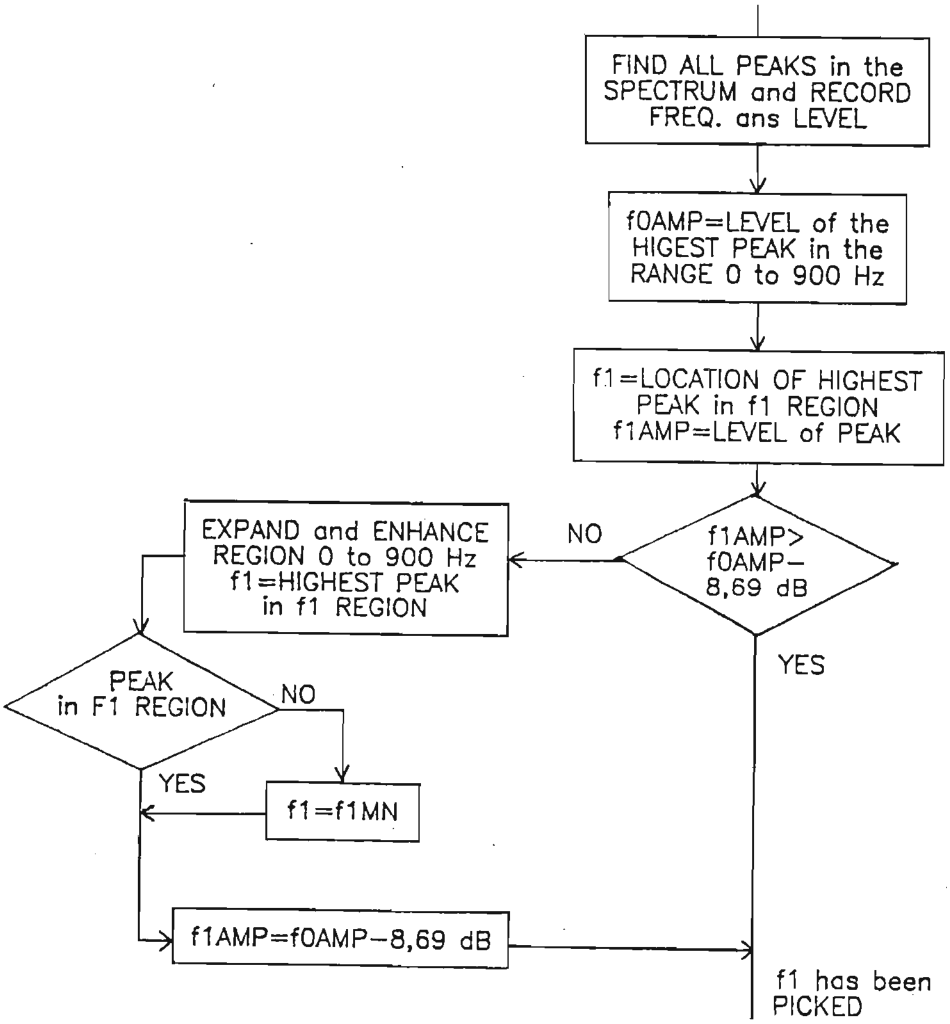
FORTRAN77

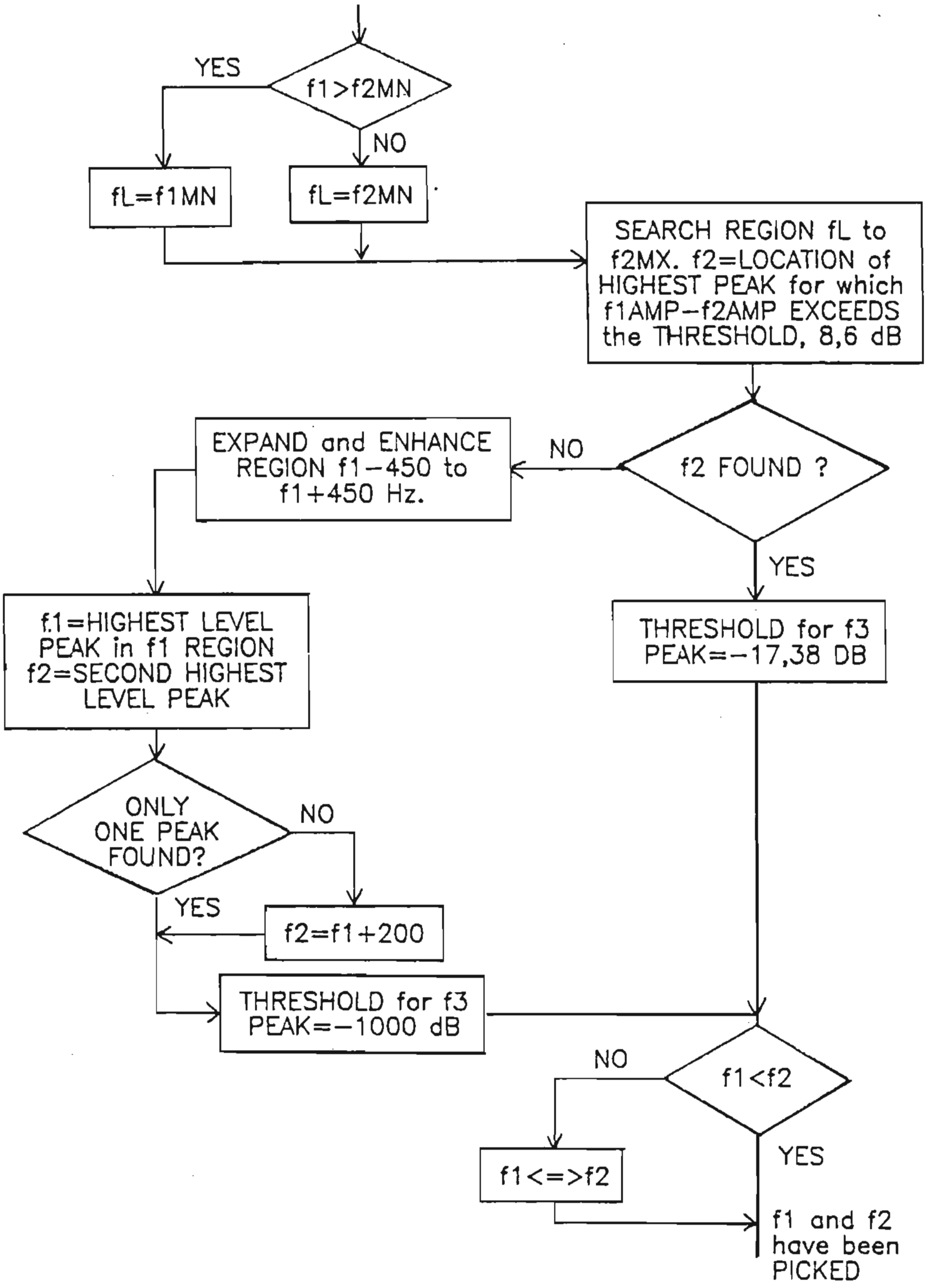
```
Subroutine FFT
Complex x, u, w, temp
Common  /fft/ x(128)
Data    pi    /3.14159265359/
n   = 128
m   =   7
nv2 = n/2
nm1 = n-1
j   = 1
Do i = 1, nm1
      If (i .lt. j) Then
            temp = x(j)
            x(j) = x(i)
            x(i) = temp
      End If
      k = nv2
      Do While (k .lt. j)
            j = j-k
            k = k/2
      End Do
      j = j+k
End Do
```

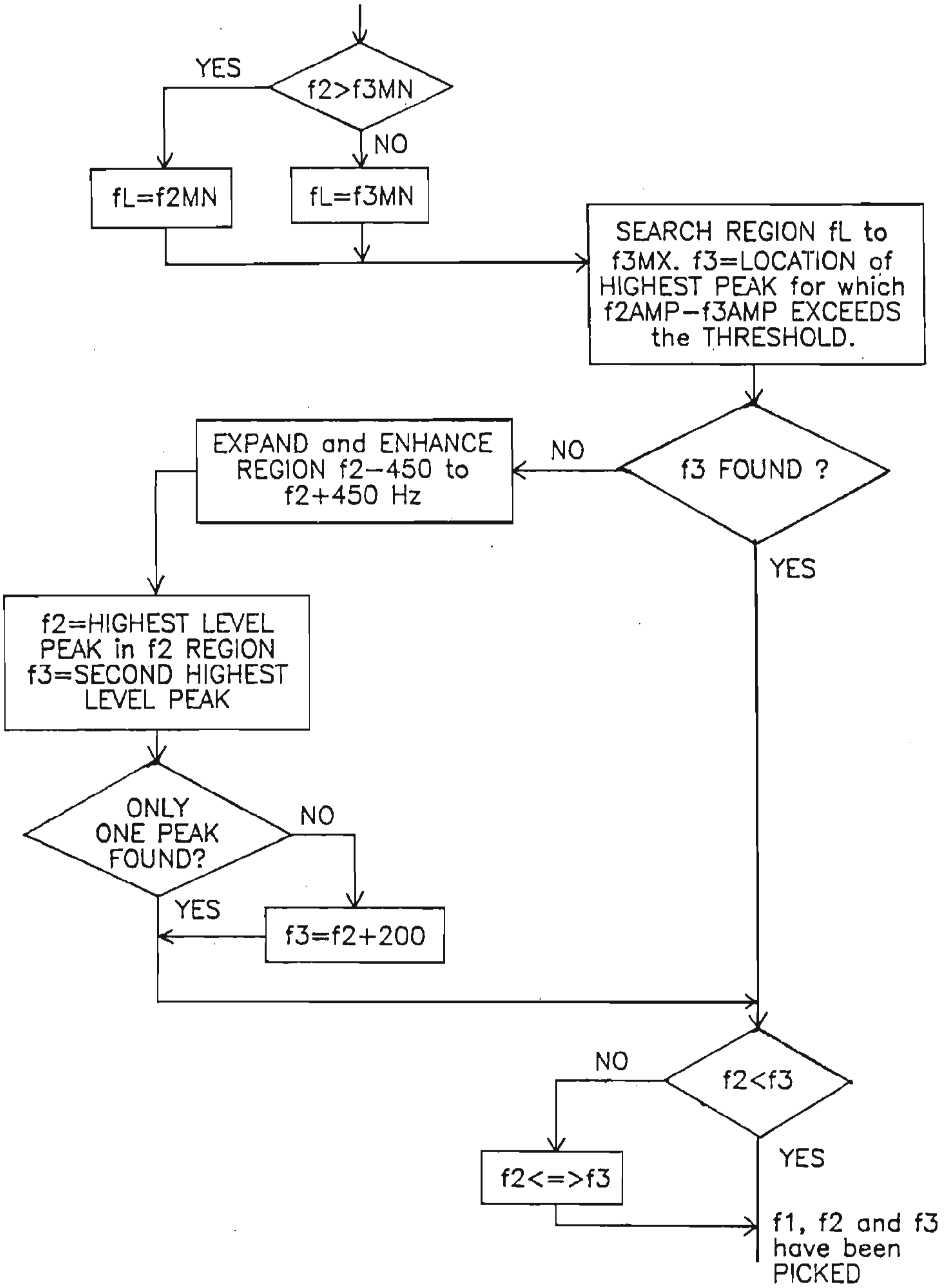
```
Do l = 1, m
      le  = 2**l
      le1 = le/2
      u = (1.0, 0.0)
      w = Cmplx(Cos(pi/Float(le1)), Sin(pi/Float(le1)))
      Do j = 1, le1
            Do i = j, n, le
                  ip    = i+2**(l-1)
                  t     = x(ip)*u
                  x(ip) = x(i)-t
                  x(i)  = x(i)+t
            End Do
            u = u*w
      End Do
End Do
Return
End
```

C Algorithm for the Estimation of the Formant Frequencies

This algorithm is based on work done by Schafer and Rabiner (53). The method uses a peak-picking method to obtain the formants, the peak-picking is done on the smoothed spectra of the input speech signal.


FIND ALL PEAKS in the SPECTRUM and RECORD FREQ. ans LEVEL
f0AMP=LEVEL of the HIGEST PEAK in the RANGE 0 to 900 Hz
f1=LOCATION OF HIGHEST PEAK in f1 REGION f1AMP=LEVEL of PEAK
f1AMP> f0AMP− 8,69 dB
NO
YES
EXPAND and ENHANCE REGION 0 to 900 Hz f1=HIGHEST PEAK in f1 REGION
PEAK in F1 REGION
NO
YES
f1=f1MN
f1AMP=f0AMP−8,69 dB
f1 has been PICKED


f1>f2MN
YES
NO
fL=f1MN
fL=f2MN
SEARCH REGION fL to f2MX. f2=LOCATION of HIGHEST PEAK for which f1AMP–f2AMP EXCEEDS the THRESHOLD, 8,6 dB
f2 FOUND ?
NO
YES
EXPAND and ENHANCE REGION f1–450 to f1+450 Hz.
THRESHOLD for f3 PEAK=–17,38 DB
f.1=HIGHEST LEVEL PEAK in f1 REGION f2=SECOND HIGHEST LEVEL PEAK
ONLY ONE PEAK FOUND?
NO
YES
f2=f1+200
THRESHOLD for f3 PEAK=–1000 dB
f1<f2
NO
YES
f1<=>f2
f1 and f2 have been PICKED


f2>f3MN
YES
NO
fL=f2MN
fL=f3MN
SEARCH REGION fL to f3MX. f3=LOCATION of HIGHEST PEAK for which f2AMP−f3AMP EXCEEDS the THRESHOLD.
f3 FOUND ?
NO
YES
EXPAND and ENHANCE REGION f2−450 to f2+450 Hz
f2=HIGHEST LEVEL PEAK in f2 REGION f3=SECOND HIGHEST LEVEL PEAK
ONLY ONE PEAK FOUND?
NO
YES
f3=f2+200
f2<f3
NO
YES
f2<=>f3
f1, f2 and f3 have been PICKED

References

(1) S R Hyde, "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature", Human Communication: A Unified View, E E David, Jr and P B Denes, Eds., 1972, pp. 339-438.

(2) L R Rabiner and S E Levinson "Isolated and Connected Word Recognition - Theory and Selected Applications", IEEE Trans. on Communications, vol. COM-29, no. 5, May 1981, pp. 621-659.

(3) N H E West, D J Burr and B D Ackland, "A Systolic Processing Element for Speech Recognition", 1982 International Solid-State Circuits Conference Digest, vol. 15, Session XIX, IEEE, 1982, pp. 274-275.

(4) Electronic Engineering, "Voice Input / Output Systems and Devices", Product Focus, Electronic Engineering, May 1982, pp. 76-191.

(5) (4), pp. 100.

(6) R J Godin, "Voice Input Output", Special Report, Electronics, 21 April 1983, pp. 126-143.

(7) (6), pp. 128-131.

(8) (4), pp. 101.

(9) S E Levinson and K L Shipley, "A Conversational-Model Airline Information and Reservation System Using Speech Input and Output", Bell System Technical Journal, vol. 59, no. 1, January 1980, pp. 119-137.

(10) P B Denes and E N Pinson, "The Speech Chain, The Physics and Biology of Spoken Language", Anchor Books, 1973.

(11) L R Rabiner and R W Schafer, "Digital Processing of Speech Signals", Prentice-Hall, 1978, pp. 38-108.

(12) J L Flanagan, "Voices of Men and Machines", J. Acoust. Soc. Am., vol. 51, no. 5 (part 1), May 1972, pp. 1375-1387.

(13) G E Peterson and H L Barney, "Control Methods Used in a Study of the Vowels", J. Acoust. Soc. Am., vol. 24, March 1952, pp. 175-184.

(14) P Hanks, T H Long and L Urdang, "Collins Dictionary of the English Language", Collins, 1980.

(15) W Davidson, J C Alock, "English Grammar and Analysis", Allman & Son, 1910, pp. 2-7.

(16) O Fujimura, "Analysis of Nasal Consonants", J. Acoust. Soc. Am., vol. 34, December 1962, pp. 1865-1875.

(17) J M Heinz and K N Stevens, "On the Properties of Voiceless Fricatives Consonats", J. Acoust. Soc. Am., vol. 33, May 1961, pp. 589-596.

(18) (10), pp. 79-147.

(19) R F Schmidt, "Fundamentals of Sensory Physiology", Springer-Verlag, 1981, pp. 180-203.

(20) E C Carterett and M P Friedman, "Handbook of Perception", Volume IV - Hearing, Academic Press, 1977, pp. 21-27.

(21) G A Briggs and J Moir, "About Your Hearing", Rank Wharfeldale, May 1967.

(22) S S Stevens and F Warshofsky, "Sound and Hearing", Time-Life Books, July 1970.

(23) W D Keidel, "Information Processing in the Higher Parts of the Auditory Pathway", Springer-Verlag, Facts and Models in Hearing, 1974, pp. 216-227.

(24) D Kaln, L R Rabiner and A E Rosenberg, "On Duration and Smoothing rules in a Demisyllable-based Isolated-Word Recognition System", J. Acoust. Soc. Am., vol. 75, no. 2, February 1984, pp. 590-598.

(25) C J Weinstein, S S McCandless, L F Mondshein and V W Zue, "A System for Acoustic-Phonetic Analysis of Continuous Speech", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-23, February 1975, pp. 54-67.

(26) J J Wolf and W A Woods, "The HWIM Speech Understanding System", 1977 IEEE Internat. Conf. Record on Acoust., Speech and Signal Process., May 1977, pp. 784-787.

(27) K K Paliwal and P V S Rao, "Synthesis-based Recognition of Continuous speech", J. Acoust. Soc. Am., vol. 71, no. 4, April 1982, pp. 1016-1024.

(28) N Ishii, Y Imai, R Nakatsu and M Ando, "Speaker-Independent Speech Recognition Unit Development for Telephone Line Use", Japan Telecommunications Review, July 1982, pp. 267-274.

(29) M H Kuhn and H H Tomaschewski, "Improvements in Isolated Word Recognition", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-31, no. 1, February 1983, pp. 157-167.

(30) J F Mari and J P Haton, "Some Experiments in Automatic Recognition of a Thousand Word Vocabulary", Unpublished, (Pattern Recognition and Artificial Intelligence Group, CRIN - University of Nancy 1, B.P. 239-54506, Vandoeuvre, France), 1984.

(31) A K Datta, N R Ganguli and S Ray, "Maximum Likelihood Methods in Vowel Recognition: A Comparative Study", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. PAMI-4, no. 6, November 1982, pp. 683-689.

(32) A V Oppenheim and R W Schafer, "Digital Signal Processing", Prentice-Hall, 1975, pp. 532-554.

(33) H Y-F Lam, "Analog and Digital Filters: Design and Realization", Prentice-Hall, 1979, pp. 569-572.

(34) J Makhoul, "Linear Prediction: A Tutorial Review", Proc. IEEE., vol. 63, April 1975, pp. 561-580.

(35) R W Schafer and L R Rabiner, "Digital Representation of Speech Signals", Proc. IEEE., vol. 63, April 1975, pp. 662-677.

(36) A E Rosenberg, L R Rabiner and J G Wilpon, "Recognition of Spoken Spelled Names for Directory Assistance Using Speaker-Independent Templates", Bell System Technical Journal, vol. 59, no. 4, April 1980, pp. 571-592.

(37) L R Rabiner and M R Sambur, "Some Preliminary Experiments in the Recognition of Connected Digits", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-24, April 1976, pp. 170-182.

(38) C S Myers and L R Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", IEEE Trans. Acoust., Speech and Signal Pocess., vol. ASSP-29, no. 2, April 1981, pp. 284-297.

(39) F Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-23, February 1975, pp. 67-72.

(40) G M White and R B Neely, "Speech Recognition Experiments with Linear Predication, Bandpass Filtering, and Dynamic Programming", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-24, April 1976, pp. 183-188.

(41) M J Hunt M Lennig and P Mermelstein, "Experiments in Syllable-based Recognition of Continuous Speech", IEEE, 1980, pp. 880-883.

(42) (2), pp. 639-644.

(43) H Sakoe and S Chibu, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-26, February 1978, pp. 43-49.

(44) B T Oshika, V W Zue, R V Weeks, H Neu and J Aurbach, "The Role of Phonological Rules in Speech Understanding Research", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-23, no. 1, February 1975, pp. 104-112.

(45) (2), pp. 645.

(46) (15), pp. 140-228.

(47) V R Lesser, R D Fennell, L D Erman and D R Reddy, "Organisation of the HEARSAY II Speech Understanding System", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-23, February 1975, pp. 11-24.

(48) P N Strong, "The Design and Implementation of a Linear Predictive Vocoder", Thesis for M Sc in Engineering, University of Natal, December 1983, pp. 18-19.

(49) M F Medress, T E Skinner, D R Kloker, T C Diller and W A Lea, "A System for the Recognition of Spoken Connected Word Sequences", IEEE Internat. Conf. Record on Acoust., Speech and Signal Process., April 1976, pp. 434-437.

(50) P Regel, "A Model for Acoustic-Phonetic Transcriptions of Fluently Spoken German Speech", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-30, no. 3, June 1982, pp. 440-450.

(51) K Tanaka, "A Parametric Representation and a Clustering Method for Phoneme Recognition-Application to Stops in CV Environment", IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-29, no. 6, December 1981, pp. 1117-1127.

(52) C G Bell, H Fujisaki, J M Heinz, K N Stevens and A S House, "Reduction of Speech Spectra by Analysis by Synthesis Technique", J. Acoust. Soc. Am., vol. 33, 1966, pp. 1725.

(53) R W Schafer and L R Rabiner, "System for Automatic Formant Analysis of Voiced Speech", J. Acoust. Soc. Am., vol. 47, February 1970, pp. 634-648.

(54) R D Kent and L L Forner, "Developmental Study of Vowel Formant Frequencies in an Imitation Task", J. Acoust. Soc. Am., vol. 65, no. 1, January 1979, pp. 208-217.

(55) H Kasuya and H Wakita, "Automatic Detection of Vowel Centers from Continuous Speech", IECE Trans. of Japan, vol. E 64, no. 10, October 1981, pp. 640-645.

(56) S E Blumstein, "Acoustic Invariance in Speech: Evidence from Measurements of Spectral Characteristics of Stop", J. Acoust. Soc. Am., vol. 66, no. 4, October 1979, pp. 1001-1016.

(57) (27), pp. 1017-1018.

(58) (2), pp. 630.

(59) (11), pp. 46-74.

(60) J E Paul, Jr and A S Rabinowitz, "An Acoustically Based Continuous Speech Recognition System", IEEE Symp. on Speech Recognition, April 1974, pp. 63-67.

(61) D R Reddy, L D Erman and R B Neely, "A model and a System for Machine Recognition of Speech", IEEE Trans. Audio Electroacoust., vol. AV-21, June 1973, pp. 229-238.

(62) (27), pp. 1018.

(63) (50), pp. 447.

Pronunciation Key of the Phonemes

The following consonant symbols have their usual English values: b, d, f, h, k, l, m, n, p, r, s, t, v, w, z. The other symbols and their interpretation are listed in the table below.

ORTHOGRAPHIC REPRESENTATION	INTERPRETATION	
/a:/	father	/f a: the e/
/ae/	act	/ae k t/
/aI/	dive	/d aI v/
/aIe/	fire	/f aIe/
/aU/	out	/aU t/
/aUe/	flour	/f l aUe/
/E/	bet	/b E t/
/eI/	paid	/p eI d/
/Ee/	bear	/b Ee r/
/g/	get	/g E t/
/I/	pretty	/p r I t I/
/i:/	see	/s i:/
/Ie/	fear	/f Ie/
/j/	yes	/j E s/
/o/	pot	/p o t/
/eU/	note	/n eU t/
/o:/	organ	/o: g e n/
/oI/	void	/v oI d/
/U/	pull	/p U l/
/u:/	zoo	/z u:/
/Ue/	poor	/p Ue/
/e/	potter	/p o t e r/
/E:/	fern	/f E: n/
/A/	cut	/k A t/
/sh/	ship	/sh I p/
/zh/	closure	/k l eU zh e/
/tsh/	chew	/tsh u:/
/dzh/	jaw	/dzh o:/
/th/	thin	/th I n/
/the/	these	/the i: z/
/ng/	sing	/s I ng/
/^{e}/	bundle	/b A n d ^{e} l/