A SPEECH PROCESSOR FOR SINGLE CHANNEL AUDITORY PROSTHESIS

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Master of Technology

by

R.M. SAPRE

(90307001)

Guide: Dr. P.C. Pandey

Department of Electrical Egineering

Indian Institute of Technology, Bombay

January 1992

DISSERTATION APPROVAL SHEET

Dissertation entitled "A SPEECH PROCESSOR FOR SINGLE CHANNEL AUDITORY PROSTHESIS" is approved for the award of the degree for Master of Technology in Electrical Engineering.

Guide: _____

Internal Examiner: _____

External Examiner: _____

Chairman: _____

21/2/92

R.M. Sapre: A Speech Processor for Single-Channel Auditory Prosthesis, M. Tech. dissertation, Department of Electrical Engineering, I I T Bombay, 1992.

ABSTRACT

Single channel auditory prosthesis is a practical solution to enhance the lipreading skill used by the deaf. The objective of the project was to come up with a speech processing scheme for single channel auditory prosthesis, after critically examining the existing schemes, in the light of phonetic features of Indian languages.

A speech processing scheme that presents some information about unvoiced segments of speech, along with intonation and rhythm information during voiced sections is developed. This scheme is based on a scheme for single channel cochlear prosthesis proposed in the literature. A method for mapping random pulses from one band to those confined to a lower band was designed, tested, and incorporated in the off-line implementation of the scheme.

Two sets of twelve vowel-consonant-vowel syllables spoken by a male and a female speaker were processed and used for informal listening tests. The discriminating features between each pair of sounds within the sets were noted down by author to prepare qualitative confusion matrices. The results indicate that the features like manner of articulation, voicing, aspiration along with some spectral information about unvoiced fricatives can be presented.

Acknowledgements

I take an opportunity to express the sense of gratitude towards my guide Dr P.C. Pandey for his guidance and encouragement throughout the project work. I am thankful to him for providing the technical facilities as well as the neatly organized literature and reference material within the Standards Lab.

I specially acknowledge Miss Darshana Kulkarni for her suggestions through informal discussions and for the utility programs that made my task much easier. I am grateful to Prof T.G. Thomas and Mr Amitabh Sharma for patiently checking the manuscript of this dissertation.

I am thankful to my friends from the department and hostel who helped me in several ways during the course of the project.

My sincere thanks to Mr Anil Vartak and Mr A.D. Apte from Standards Lab for their timely help and cooperation.

I specially mention the help and support extended by my brother Nilesh, during the final phase of the project work.

Rajendra M Sapre

List of Symbols & Abbreviations

A/D	analog to digital
D/A	digital to analog
F0	fundamental frequency of the voice pitch
F1	first formant frequency
F2	second formant frequency
F3	third formant frequency
fc	cut off frequency
HF	high frequency
LF	low frequency
LP	linear prediction
PRR	pulse repetition rate
RAM	random access memory
RMS	root mean square
SD	standard deviation
VCV	vowel-consonant-vowel context
ZCR	zero-crossing rate

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW OF THE PROBLEM

Profoundly deaf persons face a lot of problems as they cannot converse easily. For prelingually deaf persons the problem is more difficult as they cannot acquire speaking skill due to lack of auditory feedback. The speech quality may deteriorate if a person becomes deaf postlingually.

Obtaining cues from facial expressions or articulatory movements, i.e. lipreading, has been a natural way to compensate for deafness. Various sensory aids are developed to enhance the lipreading skill. These aids extract the important cues like voicing, rhythm, pitch, or formants from speech and present them in a recoded form through residual hearing or alternate sensory modalities such as touch, electrical stimulation of auditory nerve, foveal or peripheral vision, etc. The prosthesis are single channel or multichannel. The single channel aids are simple in mechanism, consume less power, and are more practical at present.

Extraction of essential speech parameters, and a suitable recoding to match the characteristics of user's sensory organ is one of the most challenging aspects of these sensory aids.

Digital signal processing (DSP) is used to advantage for this purpose. The DSP microprocessors have made the technique feasible for such real-time applications.

1.2 OBJECTIVES OF THE PROJECT

A speech processing scheme for single channel cochlear prosthesis was proposed by Pandey et al (1987). The scheme includes presentation of prosodic features like intonation, and stress along with some high frequency information about consonants. These cues are coded by mapping schemes to match the small dynamic range and small bandwidth available.

Indian languages have a rich set of consonants due to the combination of features like manner of articulation, voicing, and aspiration for the same place of articulation. The objectives of the project are to study the speech processing schemes for various modalities and the scheme mentioned above in particular, considering the phonetic features of Indian languages and to develop and evaluate a scheme for single channel auditory prosthesis.

1.3 OUTLINE OF THE REPORT

Chapter 2 overviews the speech production mechanism, phonetics, hearing mechanism and disorders. An introduction to auditory prostheses is given towards the end. Understanding of lipreading, its limitations, and cues supplementary to it is

essential for designing a speech processing scheme. Chapter 3 covers these topics. The single channel sensory aids are reviewed along with the speech processing scheme for cochlear prosthesis.

The design and implementation details of the single-channel speech processing scheme are given in Chapter 4. The testing of pitch estimators, noise mapping method, and the scheme as a whole is dealt with in Chapter 5. The informal testing and results of the tests are reported in Chapter 6. The last chapter summarizes the work done. The suggestions for future work are also given. The appendix provides details of experimental set-up. The utility programs used are briefly described.

The listing of the programs developed during the project work is available in seperate volume (Sapre, 1992).

CHAPTER 2

SPEECH, HEARING, AND HEARING DISORDERS

2.1 INTRODUCTION

This chapter is an overview of speech, auditory system, and hearing disorders. Speech production, phonetics, hearing mechanism and hearing disorders are discussed. The various types of auditory prostheses are overviewed at the end.

2.2 SPEECH PRODUCTION & ACOUSTIC PHONETICS

Human speech production mechanism (Flanagan, 1972; Ledfoged, 1982; O'Shaughnessy, 1987) consists of three main parts: lungs, larynx, and vocal tract as shown in Fig. 2.1. Lungs serve as a power source. Air is forced from lungs through larynx into vocal tract. The larynx is a frame of cartilages and connects the lungs and vocal tract through the wind pipe or trachea. Within the larynx are the vocal cords or vocal folds, which is a pair of elastic structures of muscles, mucous membrane and tendons. The epiglottis is a cover for larynx and avoids passage of food in windpipe. The vocal tract consists of oral cavity and nasal cavity. The shape of oral cavity can be changed by the movement of articulators such as velum, tongue, teeth, lips and jaws.

The voiced sounds can be produced by passing the air from

lungs through the vocal folds under tension.The vocal cords vibrate at fundamental frequency (F0) which depends on mass and tension of the cords. This gives quasi-periodic pulses as output. The vocal tract shape gives filtering effect with resonant frequencies known as "formant frequencies" or "formants".

The vocal tract can be completely blocked followed by sudden release to give stops. Or it can be constrained somewhere to produce frication. Aspiration is a result of high air flow through unconstrained vocal cords.

Phoneme is the smallest distinct sound in a language. Phonemes can be classified as voiced and unvoiced depending on glottal excitation being present or absent respectively. Continuants are produced with fixed vocal tract configuration thus giving fixed formants. Noncontinuants are produced by varying vocal tract shape from one configuration to other thus show varying formants. Continuants include vowels, fricatives, and nasals. Noncontinuants include diphthongs, semivowels, stops, and affricates. The English phonemes (IPA symbols), keywords and their features are shown in Table 2.1. Similar chracterization for Hindi phonemes is given in Table 2.2.

Vowels are produced by fixed unconstrained vocal tract configuration with quasi-periodic glottal pulses as source. They are characterized by formants in F1 - F2 plane (first and second formants). The average positions of vowels are shown in

Fig. 2.2. Vowels can be classified on the basis of tongue position, i.e., location of the arch of the tongue and its height. Lax vowels are short duration vowels ,e.g., /I/ and tense vowels are long duration, e.g., /i/.

Diphthongs are produced by smooth variation of vocal tract from one position to other. Semivowels like /w j/ are produced by a gliding change in vocal tracts shape between adjacent phonemes. Their acoustic characteristics strongly depend on the context.

Nasals are produced by voiced excitation with oral cavity constricted at some point and nasal cavity coupled to it by lowering the velum. The oral cavity gives antiresonances whereas nasal cavity gives resonances.

Voiced stops /b d g/ are produced by completely closing the vocal tract at a point followed by sudden release with vocal cords vibrating. Unvoiced stops /p t k/ are produced similarly but with vocal cords released. Voiced fricatives /v z/ are produced by passing glottal excitations through vocal tract constricted at some point thus giving turbulence or high frequency components. Unvoiced fricatives are produced similarly but the voicing is lacking. Affricates like /tʃ dʒ/ are combinations of a stop followed by frication. The phoneme /h/ is produced by forcing air through relaxed vocal cords thus giving turbulence at glottis instead of vibration. The characteristics of /h/ are those of the vowel it follows.

In Indian languages, aspiration plays important role. Many consonant pairs (e.g., /b b^h/, /k k^h/) differ only on the basis of presence or absence of aspiration and are additionally classified as aspirated or unaspirated (Ladefoged, 1982).

The segmental features refer to characteristics of individual phonemes whereas suprasegmental features refer to the way the phrases or sentences are spoken as a whole (Levitt, 1980b). Some suprasegmental characteristics vary the meaning and others are related to voice quality which is helpful in speaker identification.

Intonation is the variation in pitch (FO) in a sentence. In English, a question ends in a rising pitch. Also, pitch rises at the word to be stressed. The stress is also marked by varying intensity or duration of the phonemes. Rhythm is created by a combination of stressed and unstressed phonemes. Phrasing refers to grouping of words in a sentence. The phrasing is either due to language syntax or occurs naturally for breathing. For a natural speech, there should be a neat combination of segmental as well as suprasegmental features.

2.3 AUDITORY SYSTEM:

The major components of ear as shown in Fig. 2.3, are the outer ear, middle ear, and inner ear. The anatomy and function of each section are discussed below (Flanagan, 1972; Gelfand,

1981, Pickles, 1982; O'Shaughnessy, 1987). The outer ear consists of external visible part called pinna and a cavity open at pinna and closed at other end by eardrum. The pinna serves as a protection for ear. Also it boosts the sound coming from front compared to those coming from back side. This helps in sound localization. The ear canal serves as a quarter wavelength resonator. Typically it amplifies energies between 3 kHz to 5 kHz by up to 12-15 dB. The resonance is broad due to the flexibility of canal.

The middle ear is an air filled cavity which has the ossicular chain of three tiny rigid bones; malleus, incus and stapes. The ossicular chain connects the eardrum to the beginning of inner ear called oval window. The cavity is connected to throat (nasopharynx) via Eustachian tube. The middle ear bones match the acoustic impedance of inner ear fluid (about 400 times that of air) with that of air. A gain of about 30 dB is obtained through lever mechanism and (mainly) because of the ratio of vibrating area at eardrum to that of stapes area. The middle ear attenuates frequencies above 14 Hz by about 15 dB/octave.

The inner ear consist of cochlea, an approximately 35mm long spiral tube, with an area at the base of about 4 mm^2 and shrinking to 1 mm^2 at the apex, as shown in Fig. 2.4. The cochlea is divided into three chambers due to two membranes, viz., upper Reissner's membrane and lower basilar membrane. The outer two chambers, scala

vestibuli and scala tympani, are filled with fluid similar to extracellular fluid, whereas middle chamber, scalar media, is filled with intracellular fluid. The two outer chambers are connected to each other at the apex via a small opening called helicotrema.

Since the fluids are incompressible, vibrations at oval window due to stapes set the flexible membranes to vibration. On the basilar membrane lie the organs of Corti, which contain about 30 thousand sensory hair cells. There is a single row at inside and a several rows of outer hair cells along the length of cochlea. Due to vibration, a relative motion is created between these cells and the tectorial membrane. The deformations cause neural firings at the nerve ending, from the auditory nerve fibers. The neural information is carried to brain after several recodings.

The basilar membrane is stiff, and thin at the base and thick and compliant at the apex. Each location on the membrane responds maximally to a particular characteristic frequency (CF). The nature of a section of membrane is like a bandpass filter with almost constant Q (ratio of center frequency to bandwidth) everywhere. The distance of maximal response point from apex is approximately proportional to the logarithm of the frequency of incident sound wave. At low frequencies the individual nerve fibers are phase locked to the stimulating frequency. This is

known as temporal coding. This is possible only up to 0.5 kHz. Since different points on basilar membrane are sensitive to different frequencies, the frequencies can be analyzed by "place coding". This explains the response to high frequencies as well.

2.4 HEARING DISORDERS

There are four main causes for hearing impairment: conductive loss, sensori-neural, central, and functional deafness.

The conductive loss is due to defects in ear canal, ear drum, or middle ear. Fixation of stapes in oval window, i.e., otosclerosis or heavy conductive loss in middle ear cavity filled with fluid, or puss due to infection (otitis media) are possible reasons.

Sensori-neural loss is due to damage of sensory hair cells (cochlear) or due to damage to auditory nerve (retro-cochlear). The causes could be certain diseases, old age, exposure to high noise, or effect of ototoxic chemicals. The high frequency response is very much affected and frequency resolution is reduced. So fricatives cannot be properly discriminated. Also the dynamic range is reduced, i.e., threshold of detection is elevated and a small increase in sound above this level causes uncomfortable sensation.

Central deafness is characterized by inability to discriminate complex sounds like speech. The reasons for this deafness

are, damage to auditory cortex by cerebral hemorrhage, meningitis, skull trauma, or congenial deafness. Sometimes, the auditory system looks normal and the deafness cannot be attributed to above three categories and it is called functional deafness. The reasons could be psychological in some cases.

The deafness due to old age is called presbycusis. The reasons are degeneration of hair cells mostly from basal end, change in cochlear fluids, loss of neurons in the ascending pathway and cells in auditory cortex, mostly resulting high frequency loss. Diplacusis is perception of two different pitches for the same sound in two ears. It could be caused by local irritation, fatigue, or mild injury to organs of corti. Tinnitus or ringing in ears is believed to be due to spontaneous discharge of hair cells or nerve fibers. The reasons could be many including drugs or high intensity sound. A severe form may affect speech intelligibility.

In some cases the deafness could be due to combination of some of the disorders discussed above. So the perfect diagnosis becomes quite involved.

2.5 SENSORY AIDS FOR THE DEAF

The sensory aids for the deaf can be classified in several ways. Functionally they can be classified as speech training aids and speech perception aids. The speech training aids are specially designed to improve the speech by the deaf. In some cases the

speech perception aids can also serve as speech training aids. The alternate channels for auditory prosthesis are, residual hearing, foveal or peripheral vision, electrical stimulation of cochlea, (cochlear implant) and the sense of touch (tactile aids).

Based on speech processing techniques they can be classified as non-speech, speech specific, feature extraction, speech recognition (Levitt, 1988). Non-speech processing aids treat speech or any other sound as one and the same. The speech specific aids match the spectral or temporal characteristics of speech with the characteristics of sensory organ. Feature extraction is detection of articulatory or phonetic features of speech. Finally, the speech could be directly recognized and prese-nted to the sensory organ.

The multichannel aids present stimuli to more than one location on the sensory organ whereas single-channel aids use a single location.

Table 2.1(a). Classification of English vowels alongwith keywords.

Phoneme IPA (keyword)	Features tongue height, tongue position, lax/tense, lip-rounding
i (beet)	high, front, tense
I (bit)	high, front, lax
e (bet)	mid, front, tense
(bat)	low, front, tense
ʌ (but)	mid, back, lax
a (father)	low, back, lax
ɑ (cot)	low, back, tense
U (foot)	high, back, lax, rounded
u (boot)	high, back, tense, rounded
o (coat)	mid, back, tense, rounded
(bird)	mid, central, tense
(ado)	mid, central, lax
(all)	mid, back, lax

Table 2.1(b) Classification of English coansonants alongwith keywords

Phoneme IPA (keyword)	Features manner, voicing, aspiration, place
p (pop)	stop, unvoiced, aspirated, bilabial
b (bib)	stop, voiced, unaspirated, bilabial
t (tot)	stop, unvoiced, aspirated, alveolar
d (did)	stop, voiced, unaspirated, alveolar
k (kick)	stop, unvoiced, unaspirated, velar
g (gig)	stop, voiced, unaspirated, velar
f (fluff)	fricative, unvoiced, unaspirated, labiodental
v (valve)	fricative, voiced, unaspirated, labiodental
θ (thin)	fricative, unvoiced, unaspirated,
ð (then)	fricative, voiced, unaspirated, dental
s (sun)	fricative, unvoiced, unaspirated, alveolar
z (zoo)	fricative, voiced, unaspirated, alveolar
ʃ (shoe)	fricative, unvoiced, unaspirated, palatal
ʒ (measure)	fricative, voiced, unaspirated, palatal
h (he)	fricative, unvoiced, unaspirated, glottal
tʃ (church)	affricate, unvoiced, unaspirated, alveopalatal
dʒ (judge)	affricate, voiced, unaspirated, alveopalatal
m (me)	nasal, voiced, unaspirated, labial
n (none)	nasal, voiced, unaspirated, alveolar
ŋ (bang)	nasal, voiced, unaspirated, velar

Table 2.2(a) Classification of Hindi vowels

Phoneme IPA (keyword)	Features tongue height, tongue position, lax/tense, lip-rounding
अ (^)	mid, back, lax
आ (a)	mid, back, lax
इ (I)	high, front, tense
ई (i)	high, front, lax
उ (U)	high, back, tense, rounded
ऊ (u)	high, back, lax, rounded
ए (e)	mid, front, tense
ओ (o)	mid, front, tense, rounded

Table 2.2 (b). Classification of Hindi consonants.

Phoneme Hindi (IPA)	Features manner, voicing, aspiration, place
क् (k) क	stop, unvoiced, unaspirated, velar
ख् (k^h)	stop, unvoiced, aspirated, velar
ग् (g)	stop, voiced, unaspirated, velar
घ (g^h)	stop, voiced, aspirated, velar
ङ (ŋ)	nasal, voiced, unaspirated, velar
च (tʃ)	affricate, unvoiced, unaspirated, alveopalatal
छ ($tʃ^h$)	affricate, unvoiced, aspirated, alveopalatal
ज (dʒ)	affricate, voiced, unaspirated, alveopalatal
झ ($dʒ^h$)	affricate, voiced, aspirated, alveopalatal
ञ (n)	nasal, voiced, unaspirated, alveopalatal
ट (t)	stop, unvoiced, unaspirated, alveolar
ठ (t^h)	stop, unvoiced, aspirated, alveolar
ड (d)	stop, voiced, unaspirated, alveolar
ढ (d^h)	stop, voiced, aspirated, alveolar
ण (n)	nasal, voiced, Ustop, alveolar
त (t)	stop, unvoiced, unaspirated, dental
थ (t^h)	stop, unvoiced, aspirated, dental
द (d)	stop, voiced, unaspirated, dental
ध (d^h)	stop, voiced, aspirated, dental
न (n)	nasal, voiced, unaspirated, dental

Table 2.2 (b) (continued). Last six phonemes are used to represent phonemes used in foreign words.

Phoneme Hindi (IPA)	Features manner, voicing, aspiration, place
प (p)	stop, unvoiced, unaspirated, bilabial
फ (p^h)	stop, unvoiced, aspirated, bilabial
ब (b)	stop, voiced, aspirated, bilabial
भ (b^h)	stop, aspirated, bilabial
म (m)	nasal, voiced, unaspirated, bilabial
य (j)	glide, voiced, unaspirated
र (r)	liquid, voiced, unaspirated, retroflex
ल (l)	liquid, voiced, unaspirated, lateral
व (w)	glide, voiced, unaspirated
श (ʃ)	fricative, unvoiced, unaspirated, palatal
ष (ʂ)	fricative, unvoiced, unaspirated, alveolar
स (s)	fricative, unvoiced, unaspirated, alveolar
ह (h)	fricative, unvoiced, aspirated, glottal
ऴ (ɵ)	fricative, unvoiced, unaspirated, dental
फ़ (f)	fricative, unvoiced, unaspirated, labiodental
ज़ (z)	fricative, voiced, unaspirated, alveolar
ऱ़ (ʒ)	fricative, voiced, unaspirated, palatal
ऴ़ (ð)	fricative, voiced, unaspirated, dental
व़ (v)	fricative, voiced, unaspirated, labiodental

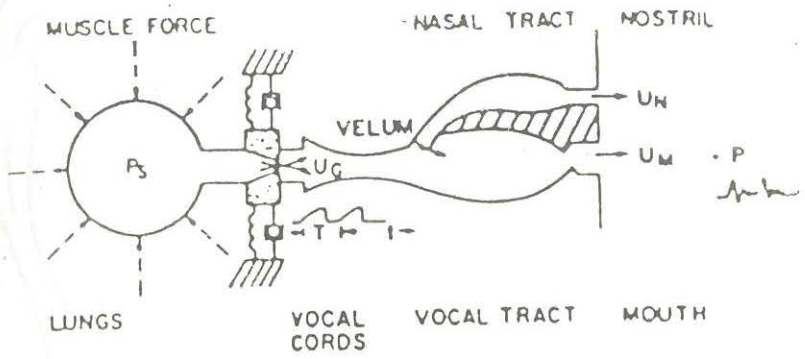


Fig. 2.1. A schematic representation of vocal apparatus. Adapted from Rabinar & Schafer (1978). Fig. 3.2.

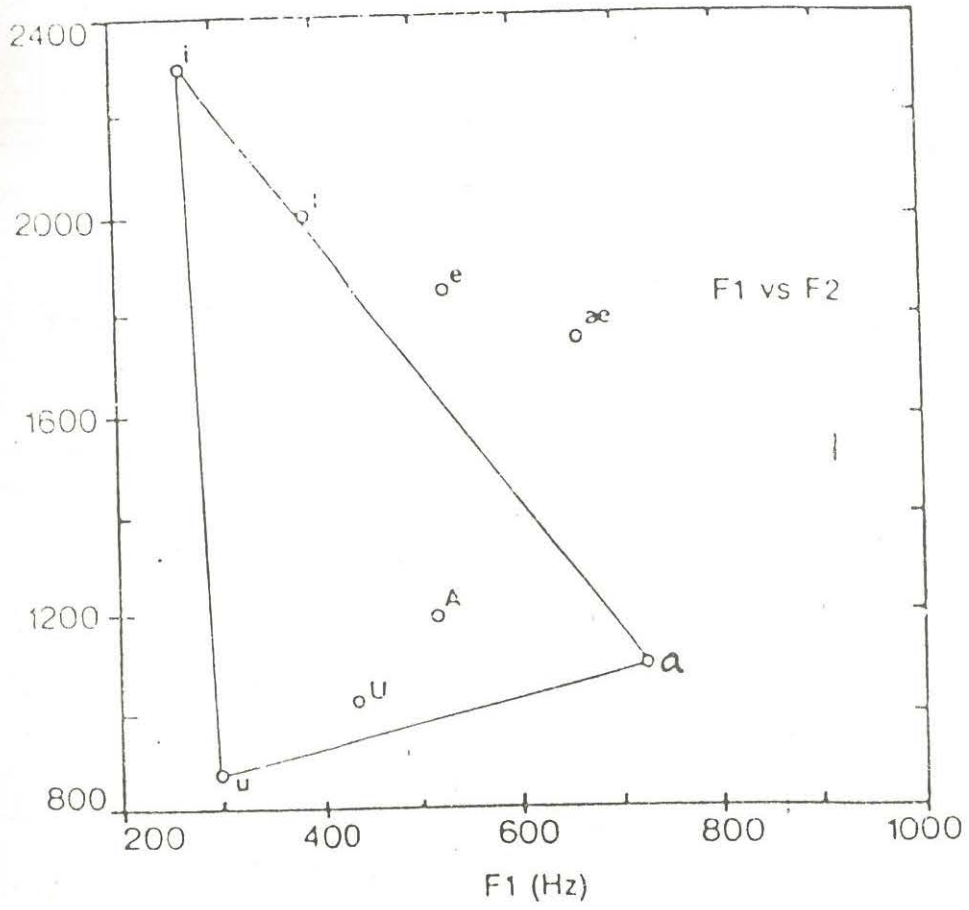


Fig 2.2 Average values of first and second formant frequencies of vowels for adult male speakers. Source: Pandey (1987), Fig 2.2.

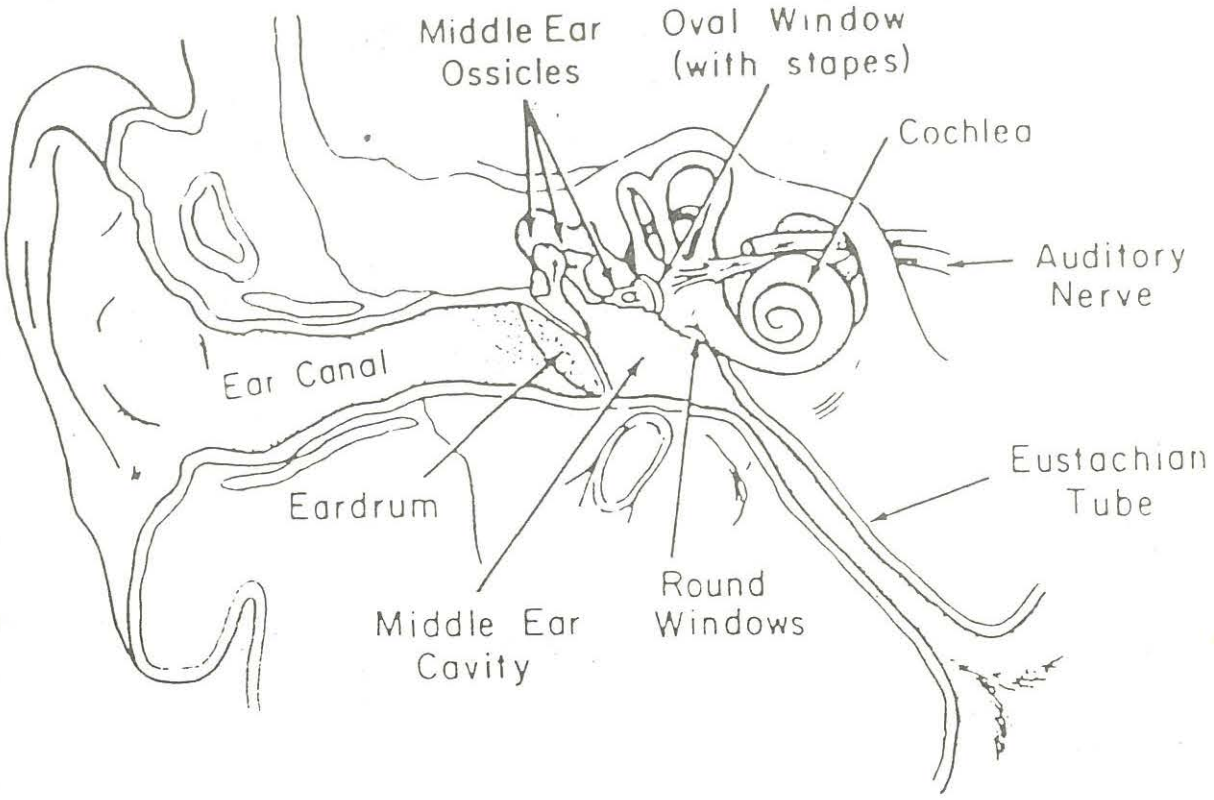


Fig. 2.3. Anatomy of the ear

Source: Levitt (1988b) Fig. 1.

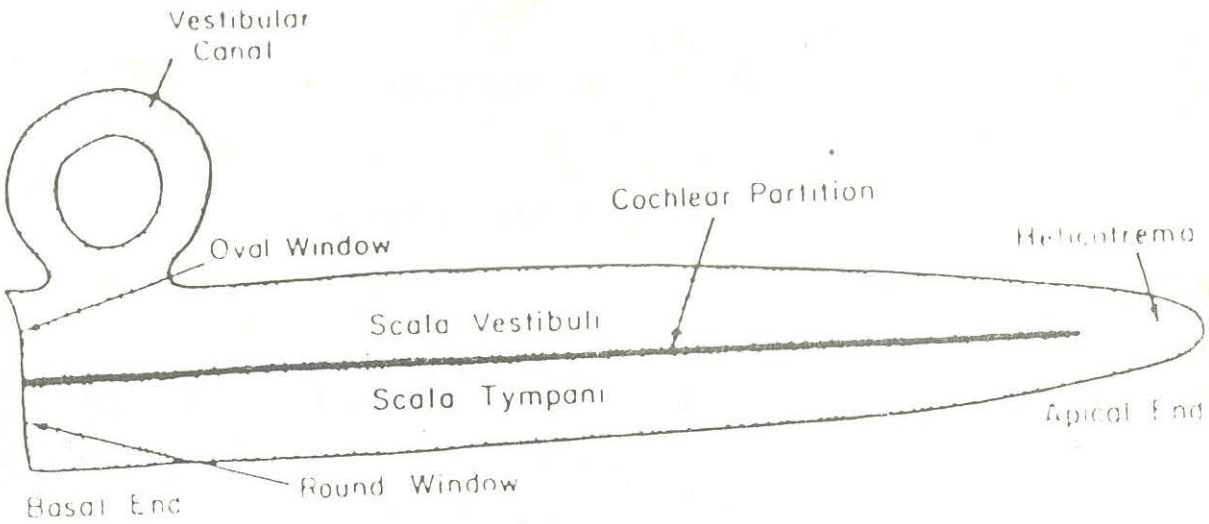
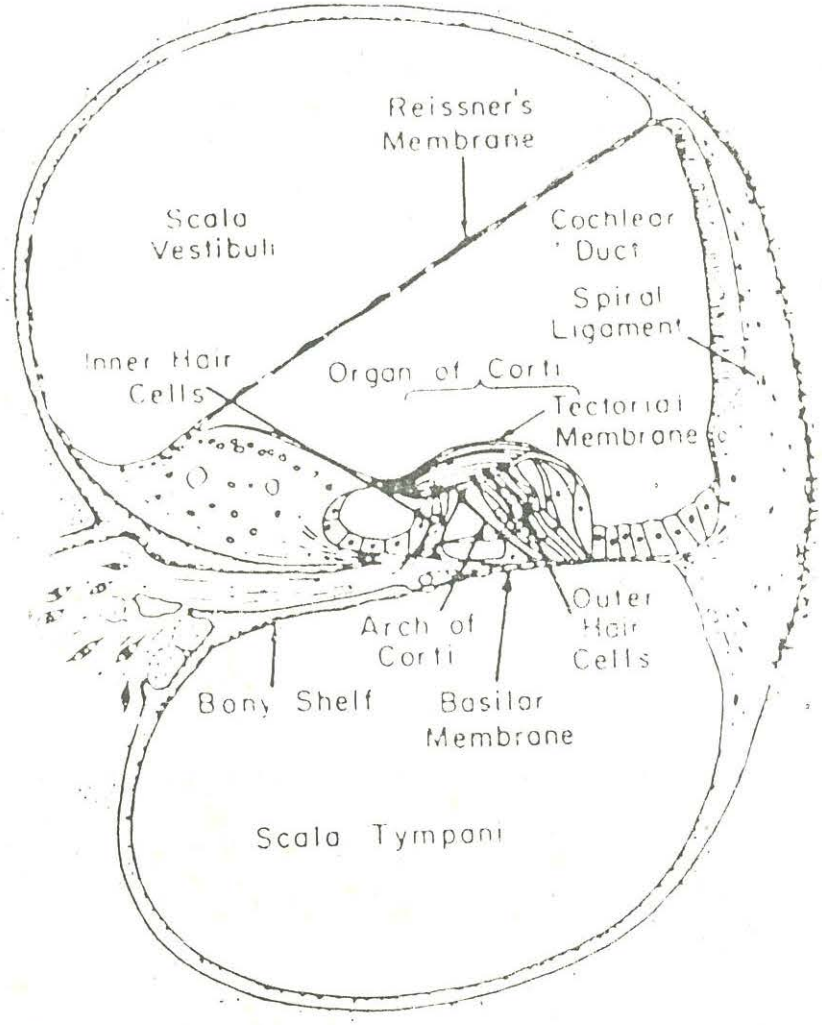


Fig. 2.4 (a) Longitudinal section of uncoiled cochlea.
Source: Levitt (1980b), Fig. 2(a)-(b)



(b) Transverse section of cochlea.

CHAPTER 3

SINGLE CHANNEL AIDS FOR THE DEAF

3.1 INTRODUCTION

The deaf persons use lipreading to perceive speech during face to face conversation. The single channel aids should provide necessary cues to supplement lipreading and design of the aids requires a study of cues for supplementing lipreading and their relative importance.

In this chapter, the information supplementary to lip-reading is discussed. Various types of single channel aids for the three modalities are reviewed. Finally a scheme for single channel cochlear prosthesis, proposed by Pandey et al (1987) will be discussed.

3.2 LIPREADING

Lipreading or speechreading is a skill of perceiving speech from facial expressions and articulatory movements. The information about consonants is obtained by observing the place of articulation, while some vowels are distinguished from lip configuration. The consonants may be classified from visual observability point of view into bilabial (/p b m/), rounded labial (/w/), and nonlabial (rest of the consonants) groups. Vowels

are observable as: /i/ (horizontal lip extension), /u/ (liprounding and protrusion), and /a/ (no liprounding and neutral horizontal extension).

Thus we see that the place of articulation of many consonants is not at all visible. Also voicing and nasalization are not sensed. The phonemes which are visually indistinguishable are called "homophones" sounds (e.g., /p b m/).

The articulatory movements are affected by the neighboring phonemes. This is called "coarticulation" (O'Shaughnessy, 1987). This can affect visibility of important sounds like /t n s/ Thus lipreading requires continuous attention, and contextual guess work. Sensory aids can relieve the burden to some extent.

3.3 INFORMATION SUPPLEMENTARY TO SPEECHREADING

The speech signal is very complex if viewed in time domain. A lot of research work has been devoted to investigate the mechanism of speech perception. Many models for speech perception have been suggested (O'Shaughnessy, 1987). These indicate certain speech features like intonation, rhythm, formant frequencies etc. as important cues for perceiving various categories of phonemes, as well as for continuous speech.

3.3.1 Voicing and Low Frequency Energy

The phonemes are classified as voiced, or unvoiced based on the presence or absence of pseudo periodic glottal excitation. Vowels show strongest low frequency energy followed by semivowels and glides. Voiced fricatives show a combination of voicing with frication. Information about voicing is missing in speechreading. Nasals show a concentration of low frequency energy in the band of 100 to 400 Hz, due to nasal coupling.

For voiced stops, (e.g., /g d/), the excitation starts immediately after the high frequency burst. For unvoiced stops, (e.g., /k t/), voicing starts after a delay thus showing a large delay in voice onset time (VOT). The delay is less for labial stop and increases towards palatal stop. Voiced fricatives also show early development of low frequency energy compared to unvoiced ones.

3.3.2 Prosody

Prosody or suprasegmentals deal with the duration, amplitude, and fundamental frequency of phoneme sequences, i.e., words and sentences. Normally one or more characters in a word are stressed by rule of the language. The stress is expressed by varying pitch, duration, or amplitude of the phoneme. A question ends with rising pitch. The rise is more if answer expected is of yes/no type. Words expressing new concept are stressed to seek

attention from listeners.

Rosen et al (1981), and Risberg (1974) found pitch as important cue to enhance lipreading in subjects with normal hearing.

Duration of phonemes is also an important clue according to Klatt (1976). Lax and tense vowels differ mainly in duration. The voiced fricatives are shorter in duration, by about 40 ms than corresponding unvoiced ones.

Grant et al (1985) studied the speechreading performance for connected speech, in normally hearing subjects, when supplemented with voicing duration, pitch, amplitude envelope stress and pitch, lowpass filtered speech, and natural speech. The performance improved in the following order: amplitude envelope, voicing duration, pitch, amplitude and pitch, lowpass filtered speech.

Similar experiments were performed with profoundly deaf subjects by Grant, (1985). Several frequency transformations were applied to pitch. Three out of five subjects could integrate the speechreading information with mapped pitch, while other two could judge intonation and stress when not speechreading.

3.3.3 Formants and High Frequency Energy

Formants are peaks in the vocal tract frequency response and result from different configurations of oral cavity. The

formants depend on the place of articulation. Vowels are characterised by first three formants. Stevens (1983), showed that the place of articulation is related to position and variation of second formant with respect to first and third.

Breewer and Plomp (1984), presented energies in 3 bands with center frequencies at 0.5, 1.6, 3.2 kHz modulated over sinusoids of respective frequencies to normal hearing subjects. Two of the 3 bands were selected at a time. They found that the combination 0.5 kHz and 3.2 kHz bands resulted in highest consonant discrimination scores. First and second formant were presented as sinusoids and were found useful for discriminating the voiced consonants, but the results were superior in the case of information about energies. This was attributed to the rhythm present in the lower band (Breeuwer and Plomp, 1985).

3.3.4 Relative Importance of Acoustic Parameters

The auditory prosthesis have limited information capacity. Thus a limited amount of information can be provided. This necessitates the study of relative importance of cues that can be presented. Thus redundant cues should be avoided; also more important cues should be assigned larger part of information capacity. Statistical study of a language is made to find out most useful cues. The relative importance or "functional load" can be determined by statistical study of a language. According to Denese

(1963), manner of consonant articulation was used more often than place of articulation. Many cues together can indicate a manner of articulation; but a single cue out of them may not be of significance. For example, the transient energy, aspiration, duration of stop gap, voice onset time etc, together could be most important for voiceless plosives; but they may not be equally important as individual cues. Thus the above mentioned cues correlate with manner of articulation (Risberg, 1969).

The relative importance of cues depends on various conditions like presence or absence of lipreading. Functional load on cues related with place of articulation decreases when lipreading is used (Woodland, 1960; reviewed by Risberg, 1969).

3.4 FREQUENCY LOWERING AIDS

The sensorineural deafness is characterized by high frequency hearing loss and many deaf persons have almost no hearing hearing left above 1 kHz. For such persons the amplifying type aids can hardly provide any information about consonants. Also deaf subjects are reported to prefer voice of adult male speakers to female or child speakers because there is more concentration of energy at the higher frequencies in the speech of females and children. These observations led to the development of frequency lowering aids. These devices attempt to recode high frequency information into the available band, typically upto 1 kHz.

Four major techniques are found in the literature, viz., modulation, vocoding, distortion, and frequency division (Risberg, 1979). Typical examples of each of the technique are shown in Fig. 3.1.

Johansson (1966), proposed the modulation technique, and has reported experiments with deaf children. The device called transposer has two channels. Normal speech is passed through amplitude compressor in channel-1. The channel-2 transposes the frequencies above 4.0 kHz to 0-1.5 kHz range. A modulator with carrier frequency around 5 kHz is used. The frequencies above 4 kHz are selected to minimize the effect on voiced phonemes. Johansson concluded that the device had some potential for speech perception but it was mainly useful for speech training. Ling (1969), compared the linear amplifier aid to modulator and a vocoder scheme. Experiments with deaf children showed that these transposing strategies were not significantly better than linear amplifier. However, two subjects with more training showed better consonant discrimination, but the performance was not attributed to any particular type of consonants. Subjects with residual hearing up to 4 kHz were also benefited by using this device.

In the "distortion" system, nonlinearity is used to generate low frequency noise from frication noise. In Johansson's system (Risberg, 1969), signals in 4-5 kHz range are passed through distortion network. The energy of low frequency noise depends on

that of frication noise but characteristic spectra for different fricatives are not generated. Risberg and Spens (Risberg, 1969) had tried a 3 channel distortion system for discriminating Swedish fricatives.

Vocoder systems divide the speech frequency range into a number of bands and the energies in these bands are modulated over pure tones or noise bands in the hearing range of subjects. The vocoder type seems to be most sophisticated. Two to ten channels have been used. Some systems use separate bands for low frequency periodicity and transposed noise bands while some do not. According to Risberg (1969), the number of channels in the vocoder should depend on the number of fricatives in the language, e.g., for four fricatives / f ʃ tʃ s/ there should be 3 or 4 channels.

Lippman (1980) reported a 10% improvement in consonant discrimination. The aid was evaluated by Posen et al (1984). They found degradation in the perception of some nasals and semivowels. So the level of modulated noise was reduced when low frequency component dominated the speech signal. This improvement nullified the negative effects maintaining the processing advantage for stops and fricatives.

Guttanman and Nelson (1968) used frequency division method. The aid was designed to make clear distinction between fricatives /s/ and /ʃ/. The spectrum of high frequency components is modified

by spectrum shaper to enhance the spectral difference between the two fricatives. The zero-crossing pulses of these signals are passed to a divide-by-N counter if the energy in this band is high compared to that in a low frequency band. The low frequency pulses are added to normal speech. From the test results, they concluded that the aid is better suited as a speech training aid.

3.5 COCHLEAR IMPLANT

For the deaf with intact auditory nerve, the nerve fibers in cochlea can be electrically stimulated. The stimulation has a dynamic range of 2-15 db (Moore, 1985). At low frequencies up to 500 Hz, the nerve fibers phase lock with input stimuli and the variation in stimulation is felt like change in pitch. The perception hardly depends on waveform of stimulation.
The stimulation electrode is placed on the round window in the case of extracochlear stimulation and it is inserted in scala tympani in intracochlear prosthesis. In bipolar stimulation, the reference electrode is placed in the same tissue, while in monopolar stimulation, the reference electrode is located elsewhere. The stimulation used is either broadband analog waveform obtained from speech, or low frequency pulses corresponding to pitch.

The aids like 3-M House or 3M-Vienna (Moore, 1985) use broadband analog wave forms. The subjects are known to use them for detecting environmental sounds and the aids are somewhat

useful, for enhancing lipreading. But the way in which the information is perceived is unknown and only trial-and-error method is possible to improve the speech coding procedure (Moore, 1985)

Another way is to extract some speech feature like pitch, amplitude variation, etc, and code it in terms of low frequency pulses (Moore, 1985). These aids use some mapping function to match the pitch with the characteristics of the user's cochlea. These indicate that pitch and intonation can be successfully presented through single channel aid. An attempt has been made to provide some high frequency information by mapping the information into the available pulse repetition rate (PRR). The pitch variation is mapped into lower PRR range while the high frequency information corresponding to some fricatives, stops etc., is mapped as noise bursts in high PRR range. The results from a processor, based on a simplified version of the scheme, (Pandey et al, 1987) indicate that such information can be presented without affecting the low frequency information. Thus there is a scope to enhance the processing scheme.

In a scheme put forth by Ifkube (1988), second formant and pitch information are presented simultaneously. An additional pulse with delay proportional to second format is inserted between pitch pulses. Further details are not available.

3.6 TACTILE AIDS:

The tactile sense is characterized by small dynamic range up to 20 dB and frequency discrimination up to 200 Hz. The skin is most sensitive to touch in the range of 30-200 Hz (Kirman, 1974). In spite of these limitations, the tactile aids have proven successful for distinguishing environmental sounds and as an aid to lipreading.

In the early attempts, the audio input was directly amplified and provided to vibrator. By knowing the properties of skin, the aids like Tactaid-1 were designed (Goldstien et al, 1985). These aids present the amplitude envelope modulated on 250 Hz frequency at which the skin is most sensitive. AGC is used to overcome the wide range of input sound. These aids help detection of environmental sounds, present rhythm and stress in speech (Sherrick, 1984; Wiesenberger et al, 1987). The aids have been tried on children of the age 6 weeks to 5 years and reported benefits include, attention to speaker, early linguistic development (Goldstein et al, 1985; Proctor et al, 1988).

The aids like "MiniFonator" modulate amplitude envelope over a carrier frequency of, up to 1000 Hz, obtained from environmental sound, in order to provide some spectral information. But the results obtained are not superior to those from simple aid providing amplitude envelope only (Wiesenberger et al, 1987).

The voice pitch F0 information has been used along with

rhythm (Goldstein et al, 1985).

Rosenberg & Molitor (1979) have shown that pitch variation can be sensed by tactile aids. In their experiments subjects could identify small sentences from a list based on intonation patterns only (no lipreading) with 50% accuracy.

Plant & Dillon (1985), provided pitch information in 30 to 100 Hz range modulated with amplitude envelope. The aid extracts pitch and presents it as either fixed (110 Hz) or variable (30-100 Hz) frequency. The intensity can be constant or proportional to that in 0.5-1 kHz band. If frequencies above 5 kHz are present, 300 Hz tone is triggered. Two subjects were trained for fixed frequency of variable amplitude and two for variable frequency of constant amplitude. Both methods were found useful in discrimination of consonants. Mainly due to improved perception of consonant voicing and enhanced identification of stop, sibilant, and nasal consonants. The aid was useful in speech training for syllable duration and sibilant production.

The electrostatic aids directly stimulate the nerves electrically. The dynamic range is up to 12 dB whereas pulse rates up to 150 pps can be discriminated with varying limens (Blamey, 1985). The power required is some what lower than that for vibrotactile aids but the sensations can be irritating some times at some locations. This modality is rarely used for single channel aids but multi channel aids like Tickletalker are available in

wearable form (Cowan et al, 1988).

3.7 THE SPEECH PROCESSING SCHEME

The survey of the various sensory aids indicates some similarities among the three modalities. The available frequency and dynamic ranges are very limited compared to those of normal ear. The information provided for discrimination of fricatives should not affect perception of voiced sounds.

A scheme that takes into account these points, was proposed by Pandey et al (1987). The scheme was designed to provide some information about certain high frequency fricatives and stops like /s ʃ z t/ along low frequency periodicity and rhythm information in the limited PRR and dynamic range available with single channel cochlear prosthesis.

The scheme is shown in Fig. 3.2. The acoustic input is processed in two parallel channels. Channel-1 provides low frequency information, and channel-2 provides high frequency information. The input signal is bandpass filtered by filter BPF-1. The information about rhythm is obtained by the envelope detector. Pitch pulses is obtained from pitch extractor or by center-clipping and peak-limiting the signal. The available pulse repetition rate (PRR) range is divided into two bands. The expected pitch range is mapped onto the lower PRR band by a mapping scheme. The amplitude envelope is compressed to suit the available dynamic

range and is superimposed on the mapped pulses to give channel-1 output.

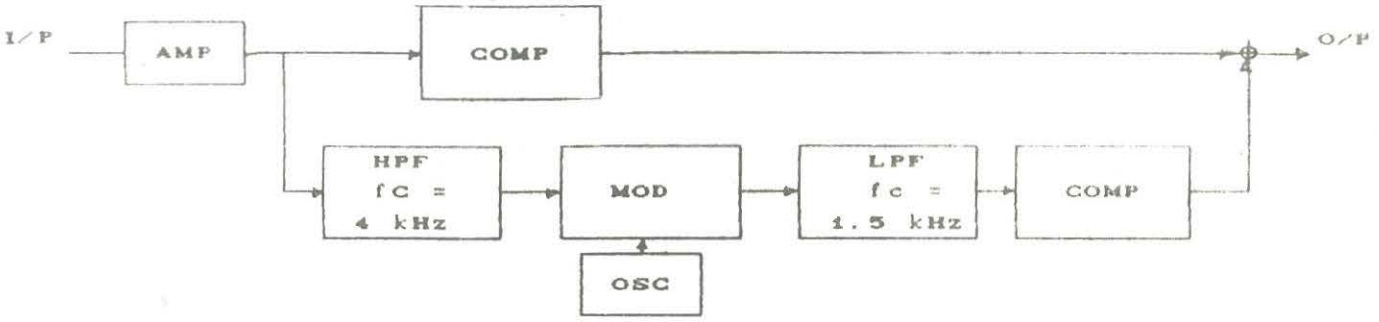
The high frequency contents are separated by filter BPF-2. (3-6 kHz for typical male speaker). The signal is center-clipped and infinite-peak-limited to obtain random pulses. The pulses are mapped into random pulses confined to the upper PRR band through a mapping scheme. The compressed amplitude envelope is superimposed on these random pulses to give output of channel-2. The relative strength of signals in the two bands is judged by a comparator, which switches the output from the channel with stronger signal.

Since the parameters like pitch vary over a wide range for adult male, adult female, and child voices, separate mapping schemes could be provided for each range. The dynamic range and PRR range is characteristic to a person. So the mapping should be user specific. The two types of information may mask each other. To avoid masking, the presentation is kept distinct in texture (periodic vs random) and in PRR range; and only one information is presented at a time.

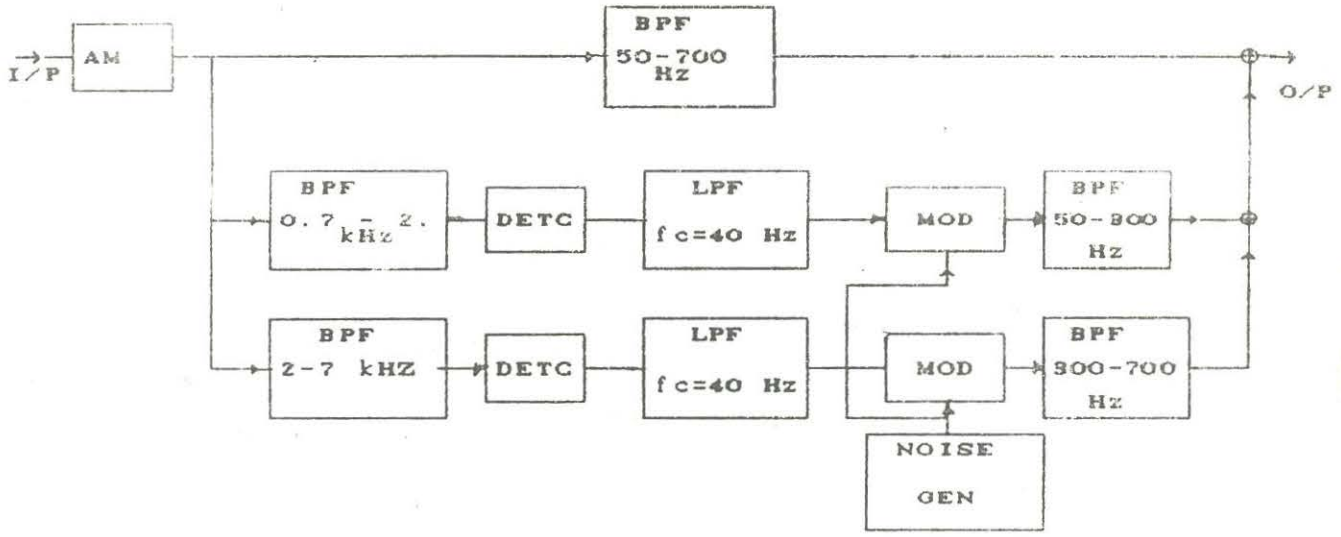
A speech processor shown in Fig.3.3, illustrating the principle of the scheme, was built in hardware, and tested for extracochlear implant by Pandey et al (1987). The results indicated that high frequency information is perceived distinctly low frequency pulses of periodicity. The low frequency information is

not masked by the presentation of high frequency information.

The scheme could be used for other modalities like residual hearing or tactile aids.

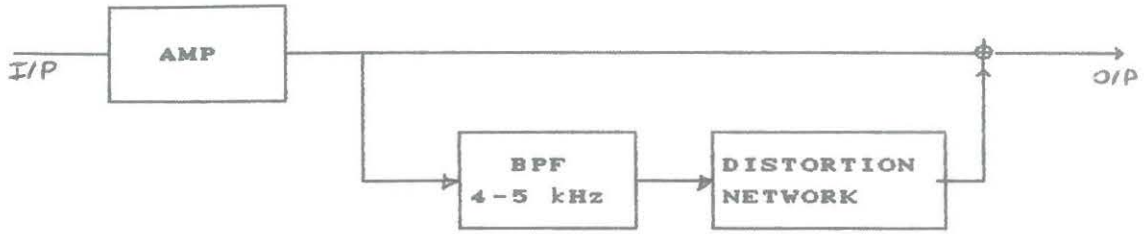


(a) Modulation: 'Transposer' by Johansson (1966)

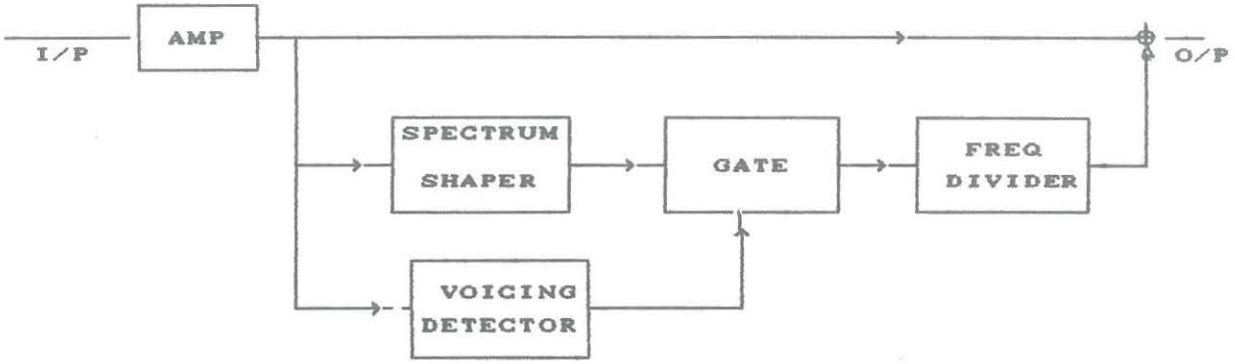


(b) Vocoder technique by Piminow.

Fig. 3.1. Frequency lowering techniques. AMP: amplifier, BPF: band pass filter, COMP:amplitude compression, DETC: detector, GEN: generator, HPF: Highpass filter, LPF: lowpass filter, MOD: modulator, OSC: oscillator.



(c) Dostortion type aid by Johansson (Risberg, 1969).



(d) Frequency divider by Guttaman and Nelson (1968).

Fig. 3.1 contd. Frequency lowering techniques. AMP: amplifier, BPF: band pass filter, COMP:amplitude compression, DETC: detector, GEN: generator, HPF: Highpass filter, LPF: lowpass filter, MOD: modulator, OSC: oscillator.

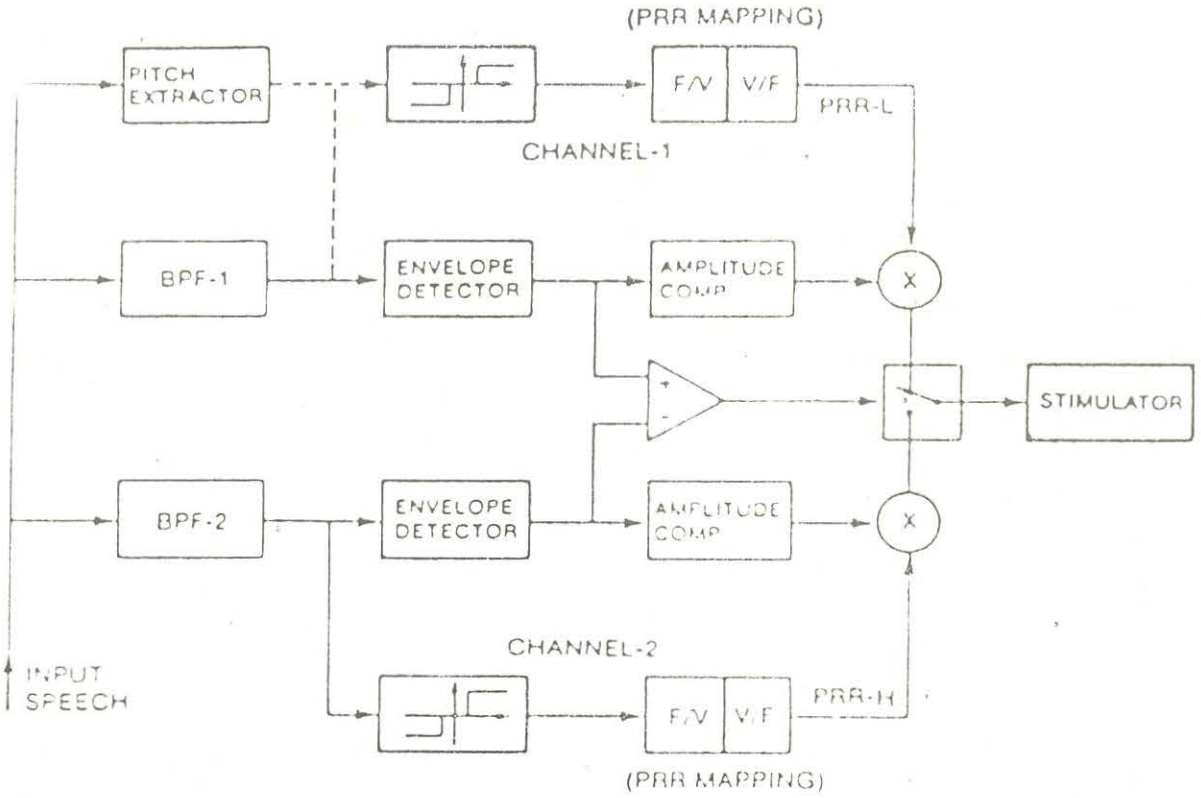


Fig 3.2. A block diagram of speech processing scheme for cochlear prosthesis by Pandey et al (1987). F/V: frequency to voltage converter, V/F: Voltage to frequency converter, COMP: compressor. Channel-1: low frequency periodicity, Channel-2 high frequency bursts. Source: Pandey (1987), Fig. 5.1.

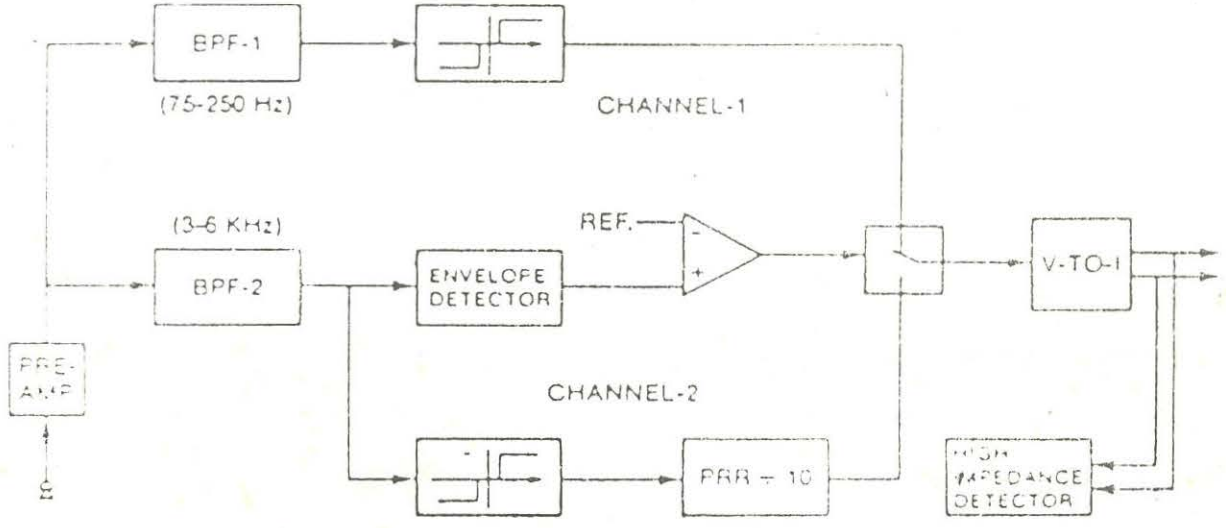


Fig. 3.3 A speech processor based on a simplified scheme. Source: Pandey (1987), Fig. 5.2.

CHAPTER 4

IMPLEMENTATION OF A SINGLE CHANNEL SPEECH PROCESSING SCHEME

4.1 INTRODUCTION

A scheme for single channel cochlear prosthesis was discussed towards the end of Chapter 3. A scheme more suitable for Indian languages is proposed. The scheme and its implementation details are discussed in this chapter.

4.2 THE SCHEME

In Indian languages, a clear distinction of manner of articulation, voicing, and aspiration is important. Four homophonous stops are possible (e.g. /p, p^h, b, b^h/) depending on the presence of voicing and aspiration. These cues could be provided by modifying the scheme that provides information about high frequency fricatives.

The block digram of the scheme is shown in Fig. 4.1. The acoustic signal is processed by two channel. The available frequency range is split into two bands. The channel-1 presents low frequency periodicity information in the lower band and the channel-2 presents the frication information in the upper band.

The filter BPF-1 separates the low frequency information

from acoustic input signal. The pitch extractor computes the pitch and also provides voiced/unvoiced signal a for particular segment. The expected pitch range is mapped on available low frequency band. If a segment is voiced, a periodic waveform with mapped pitch is obtained. The amplitude information is obtained from envelope detector and it is compressed to fit the dynamic range.

During unvoiced sections (unvoiced fricatives, bursts in stops, and aspiration) the channel-2 is activated. The input speech signal is high-pass filtered (typically, cutoff at 1 kHz) by BPF-2. Random noise pulses can be obtained to provide some spectral information about original signal. This pulse is mapped onto the available upper frequency band. The amplitude envelope of high-pass filtered signal is superimposed on the output noise after dynamic range compression.

Thus the output is periodic during voice segments and is aperiodic during the unvoiced segments of the speech. The output can be a square wave or a pulse sequence according to the modality used. The scheme provides the cues about voicing, aspiration, manner of articulation and some spectral information about unvoiced fricatives along with low frequency periodicity information.

4.3 OFF-LINE IMPLEMENTATION

A program PC20 was developed on IBM PC using Turbo-C (Borland, 1988 a, 1988 b) to implement the speech processing scheme. The listing is available in Sapre (1992). The speech data sampled, at 10kHz, are read from specified binary file (an economic way to store integers as two bytes). The parameters for various processing blocks are read from the parameter file or from keyboard, at the beginning. All important parameters can be chosen to get desired specification. The processed data are stored in the output file. The pitch and amplitude envelope information can also be stored in different files.

The data are processed frame-by-frame with a frame length of 100 samples (10 ms). Thus 100 samples are read, output is computed and corresponding 100 output samples are stored in the output file. The blocks of the scheme are implemented as sub-routines. Data buffers of one frame length are used to store the data at intermediate levels and for passing them to sub-sequent blocks. Only the values that are needed for next frame are stored along with the data buffers. Due to the frame-by-frame processing the memory size required is small and it is independent of length of the data file. The processing time is less as disk I/O and computation are done in parallel. The blocks of the scheme are explained below.

(a) Pitch Extractor: The subroutines for pitch extraction

using modified autocorrelation method and dynamic threshold method (two variants) are available. These methods are discussed in detail in section 4.3.

(b) Amplitude Envelope: The amplitude envelope is found by averaging the absolute values of samples in the desired number of latest frames. The sums of absolute values of required number of latest frames are preserved and are added to that of current frame. The average is obtained by dividing the sum by total number of samples. The envelope values are found twice in a frame, i.e., every 5ms.

(c) Filters: IIR second order filters are used. The filter coefficients for various blocks are computed, at the beginning of the program, from the specifications. Butterworth lowpass and highpass filters can be designed by a separate subroutine.

(d) Two-Level Threshold Detector: Random pulses corresponding to frication noise are obtained by the two-level threshold detector. The positive and negative threshold are specified as fraction of the average value for that frame. A pulse is issued when the input waveform crosses the thresholds from the x-axis side. This can be expressed mathematically.

Let θ_1 and θ_2 be the two levels. x(n) and y(n) are input and output respectively.spectively.

$$x_1(n) = \text{sgn}(x(n) - \theta_1)$$

$$x_2(n) = \text{sgn}(\theta_2 - x(n))$$

$$y(n) = u(x_1(n) - x_1(n-1)) + u(x_2(n) - x_2(n-1))$$

where,

$$\text{sgn}(x) = 1 \qquad x > 0$$

$$= 0 \qquad x = 0$$

$$= -1 \qquad x < 1$$

$$u(x) = 1 \qquad x > 0$$

$$= 0 \qquad x \leq 0$$

(e) Random Pulse Mapping: The random pulses from threshold detector are confined to a particular range. A particular lower range is available for presentation The pulse rate is lowered by a suitable divide-by-N counter. A method for precise mapping has been developed to match the available output range. The details of the method are given in Section 4.5.

(f) Waveform Generator-1: This block uses pitch and amplitude envelop values to generate a periodic waveform for channel-1. Two subroutines have been written. The first one generates square wave output and the other

one generates biphasic pulse waveform with negative leading pulse. The square wave output is suitable for residual hearing or for tactile aids whereas biphasic pulse waveform is useful for cochlear prosthesis. The pulse width is adjustable and set to 6 samples.

(g) Waveform Generator-2: The amplitude information is modulated over the mapped pulse sequence to get the aperiodic output for channel-2. Biphasic pulses or square wave (which can be optionally filtered) are available as in the case of channel-1. The square wave is generated by changing the polarity at every pulse at the input. The magnitude is updated every 50 ms. i.e., twice per frame.

The amplitude compression and pitch mapping were implemented. The waveforms at various stages in the two channels are drawn in Fig.4.2 and Fig.4.3 respectively.

The implementation is completely modular and a desired speech processing scheme could be implemented by suitable reconfiguration of the blocks.

4.3 PITCH ESTIMATION

The pitch estimator is an important block in the processing scheme. A large number of method for pitch estimation are reported in the literature. Only few of them are suitable for real-

time application. The modified autocorrelation method (Rabiner & Schafer, 1979) is a computationally efficient method. A method more suitable for implementation on DSP microprocessor was proposed by Sebastian (1991). This was called "dynamic threshold method".

4.4.1 Modified Autocorrelation Method

The speech input is lowpass filtered to get a smooth waveform. The signal is divided into three segments (typically of 10 ms duration). The absolute maximum is found for the first and last segment. If the minimum between the two is less than a specified threshold, the segment is declared unvoiced. Else the entire section is center clipped and infinite peak limited. The clipping level (C_L) is a fixed percentage (60-80%) of the minimum between two absolute maxima. For input x[n], the output takes values the +1, 0, or -1 for $x[n] > C_L$, $-C_L \leq x[n] \leq C_L$, and $x[i] < C_L$ respectively. Then short time autocorrelation is computed for a given range and the peak value and the location are found out. The peak value (r_i) is compared to a fixed threshold of energy (r_o). The section is declared voiced if r_i is above the threshold and the location corresponds to pitch period.

4.4.2 Dynamic Threshold Method

The speech signal is lowpass filtered to retain first few

harmonics. Periodic pulses can be obtained from this smoothened waveform by a comparator with dynamically varying in threshold (Fig.4.4). The threshold is updated to input value if the value is greater than the threshold and a pulse is given at the output. The threshold is kept constant for a blanking period which is a fraction (τ) of recent pitch estimate. After this period the threshold is decayed to fraction β of previous value at every sampling instant. Thus pulses are given during rising edges of stronger section of the smoothed speech wave. The periodicity is retained but only one bit per sample is needed as opposed to two bits (to represent -1, 0,+1) for modified autocorrelation method. Autocorrelation is performed on 300 samples to compute pitch period. The multiplication is replaced by bit-and operation. The interval is declared voiced if the ratio of peak value (r_i) to energy (r_o) is above a set threshold. The typical ranges for the parameters are:

Lowpass filter: Butterworth second order filter

f_c = 900 Hz.

$0.6 < \tau < 0.75$

$0.80 < \beta < 0.90$

$r_i / r_o = 0.25$.

On DSP boards the 300 samples are stored in different registers as bit sequences. It is difficult to shift the

registers by one frame length to create vacancy for the next frame. So a simplification was adopted. Three buffers of 100 samples each are used. The new frame is overwritten on the oldest one. Thus the buffers are filled circularly. This way the periodicity of waveform may get affected especially at larger pitch periods. The above three methods were tested and the details are given in Section 5.2.

4.5 FRICATION INFORMATION

The unvoiced sections of speech are characterized by noise. The sections are to be represented by random waveform confined to a specified band, with certain characteristic similar to the input noise waveform. The scheme for presenting the information is discussed in this Section.

4.5.1 Extraction of Frication Pulses

One way to obtain some information about unvoiced sounds is to use zero crossing detection. This zero crossing rate for unvoiced sounds is quite high, i.e., in the range of 3-8 kHz. Center clipping and infinite peak limiting can be used to reduce the rate. In the present scheme a high-pass filter of 1 kHz cut-off frequency is used to separate the frication part. A zero crossing detector is used to get random pulses. A pulse to pulse mapping scheme is used to get a pulse sequence confined to

desired range.

4.5.2 The Mapping Scheme

The mapping scheme is flow charted in Fig.4.5 The waveforms are plotted in Fig. 4.6. Two counters 'zcount' and 'opcount' are used for tracking pulse sequences at the input and output respectively. At each interval, if a pulse is present at the input the zcount is reset to 1 and average zero crossing time interval (azct) is computed by recursive relation

$$azct = (a_0 * azct + zcount)/(a_0 + 1)$$

else the zcount is incremented.

The counter 'opcount' is a down counter. When it reaches zero output pulse is given. The opcount is loaded with a new value that is computed as mapped value of

$$(b0 * azct + zcount)/c0.$$

The parameters a0, b0, and c0 can be determined for a particular mapping case. In the cases studied the typical values were:

$$a0 = 0.5; \quad b0 = 1.0; \quad 2.0 \leq c0 \leq 2.15.$$

The details of testing the method are given in Section 5.3.

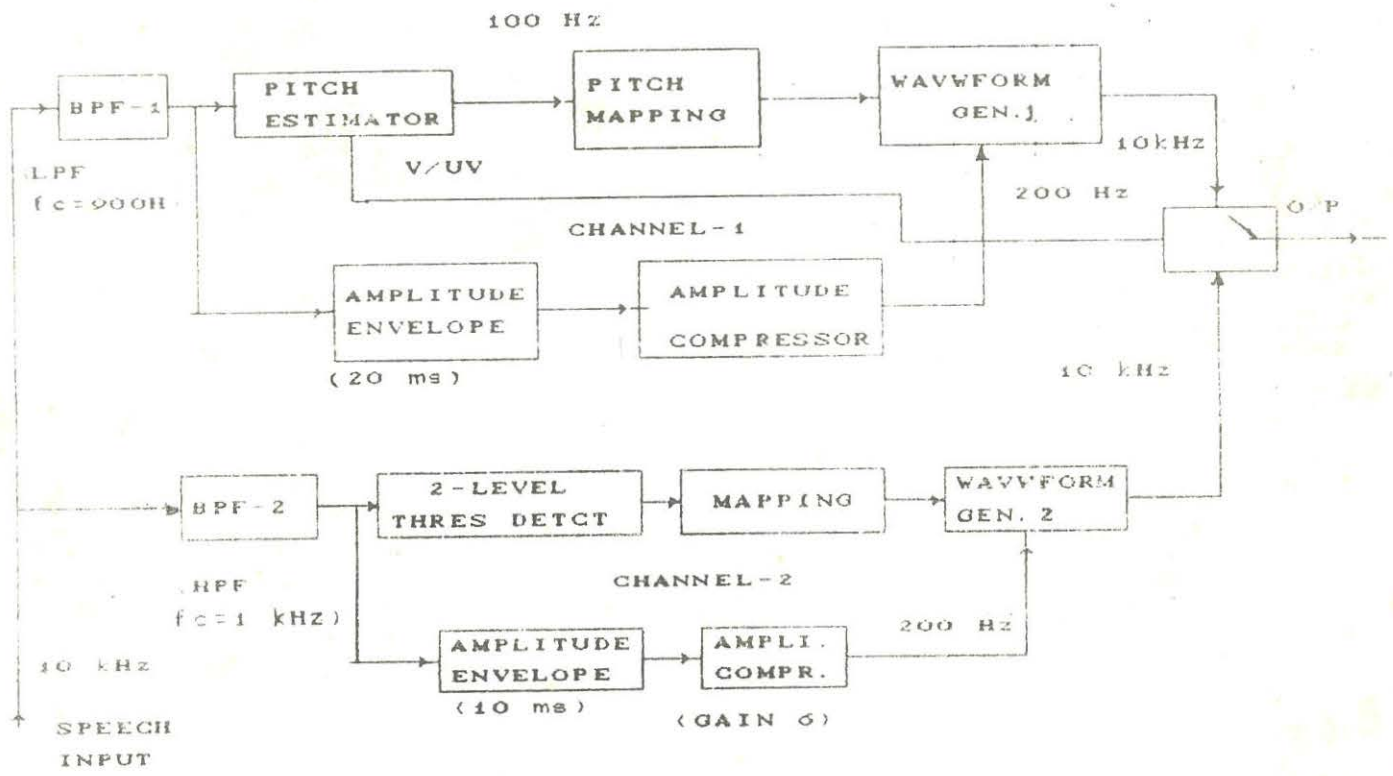
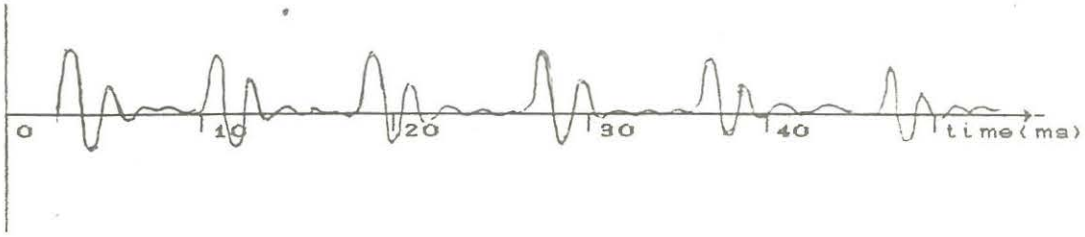
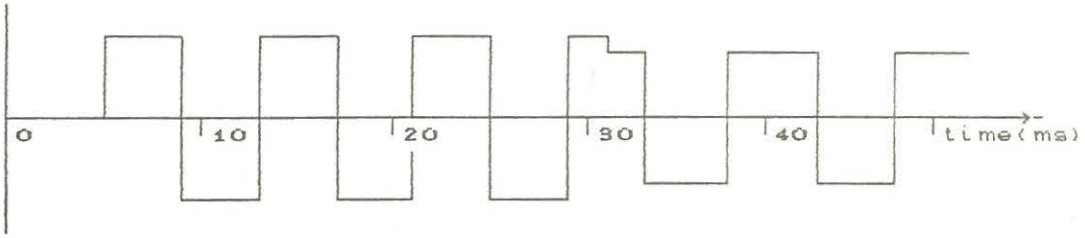


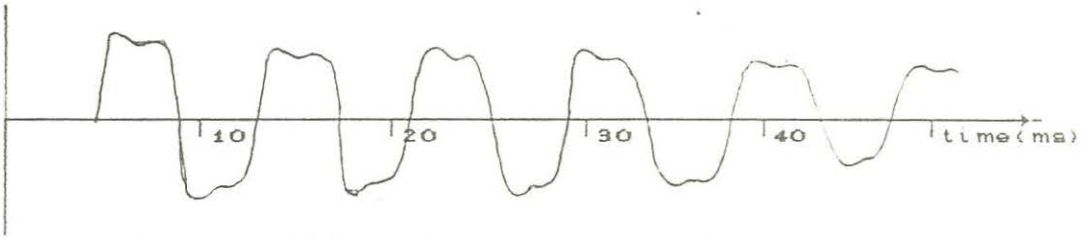
Fig.4.1. A Block diagram of the single-channel speech processing scheme. AMPLI. COMPR: amplitude compressor, BPF: bandpass filter, HPF: highpass filter, LPF: lowpass filter, V/UV: voiced /unvoiced signal, CHANNEL-1: Low fequency prosodic information, CHANNEL-2: frication information. 2-LEVEL THRES DETCT: two level threshold detector.



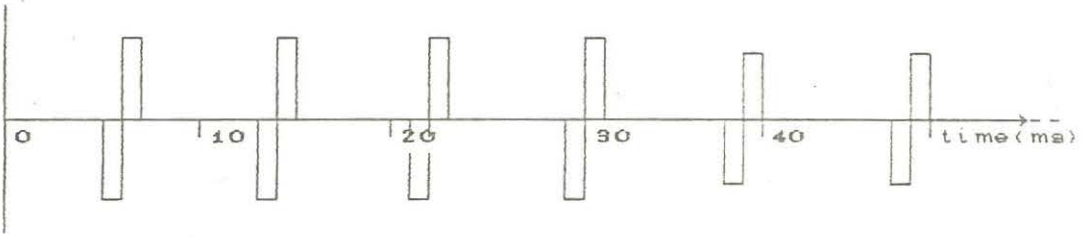
(a) Lovpass Filtered speech vveform



(b) Squarevave output.

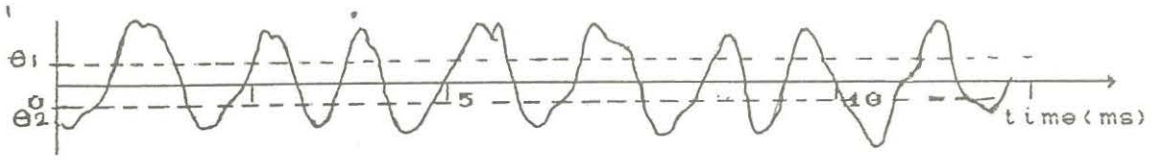


(c) Lovpass Filtered squarevave output.



(d) Biphasic pulse output or cochlear prosthesis.

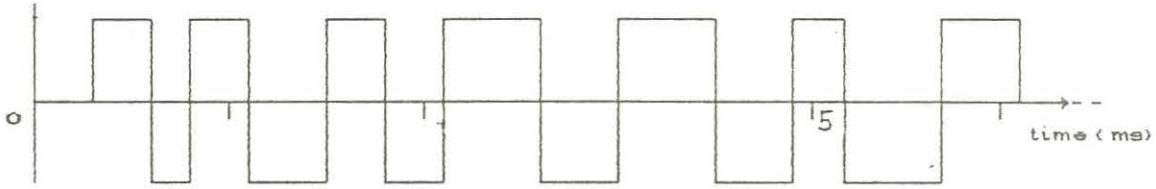
Fig. 4.2: Waveforms for channel-1.



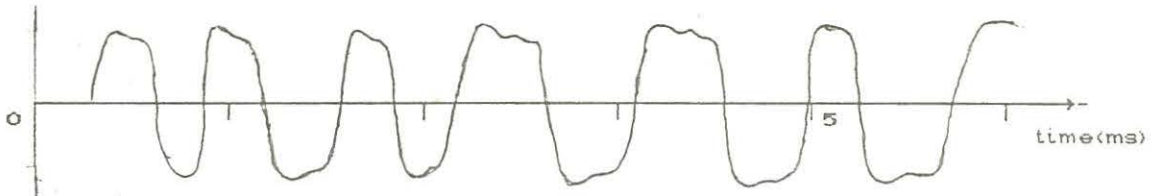
(a) Highpass filtered speech waveform.

(b) Pulses from two-level threshold detector

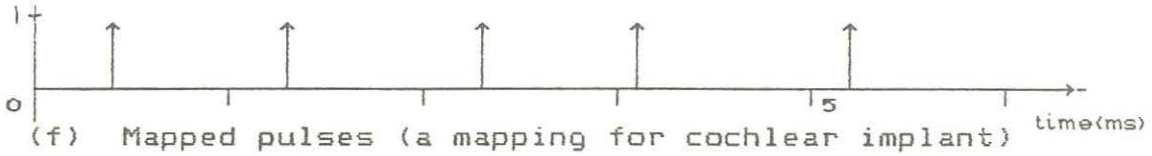
(c) Mapped pulses (a mapping for residual hearing)



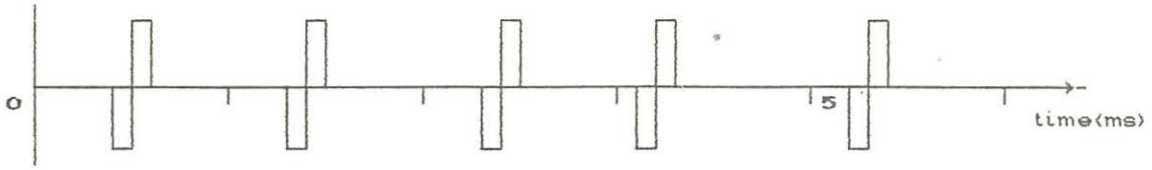
(d) Squarewave output.



(e) Lowpass filtered squarewave output.



(f) Mapped pulses (a mapping for cochlear implant)



(g) Biphasic Pulse output.

Fig. 4.3 Waveforms for channel-2.

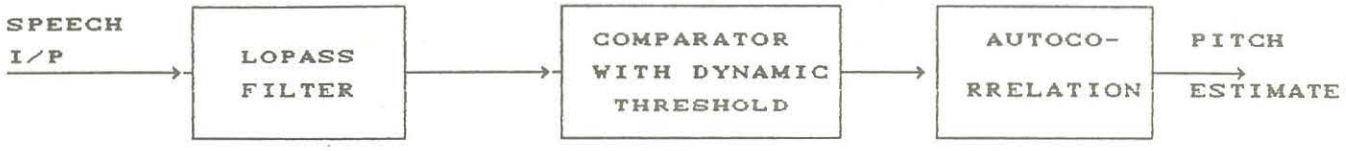
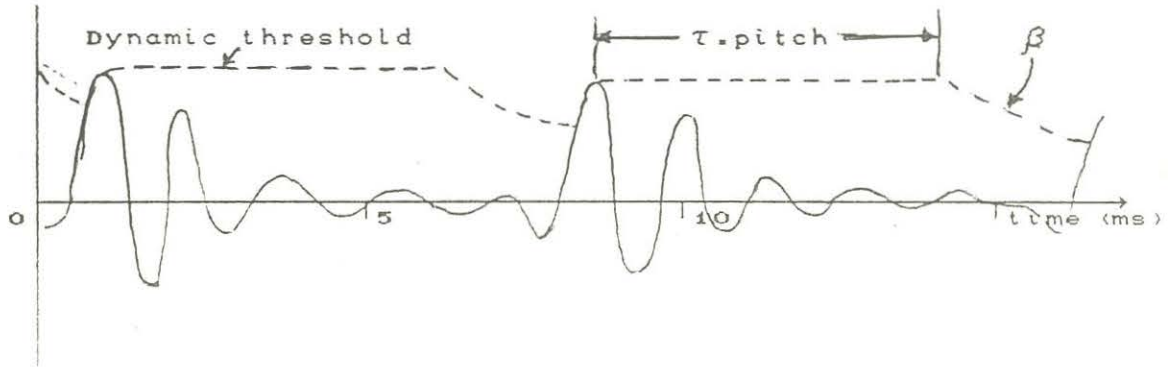
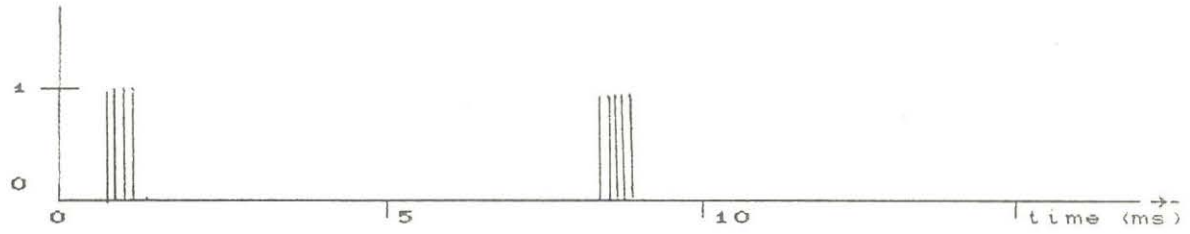


Fig 4.4(a) Dynamic threshold method for pitch estimation.



(i) Lowpass filtered speech signal



(ii) Periodic pulses

(b) Waveform details.

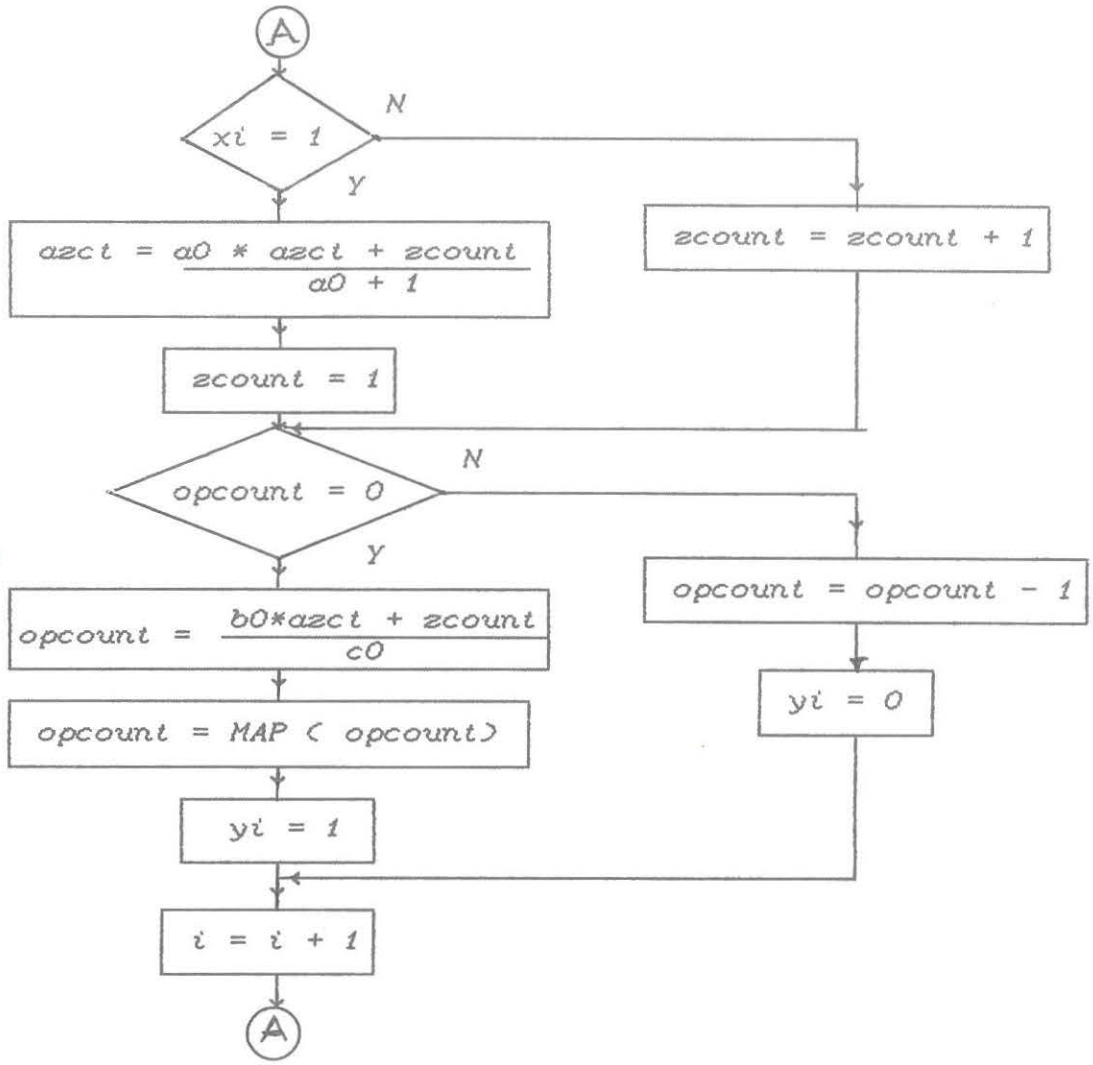


Fig. 4.5 The random pulse mapping method.
{xi}: input pulse sequence, {yi}: output pulse sequence,
zcount: counter at input, opcount: output counter
azct: average zero-crossing time interval,
MAP (.): mapping function
a0, b0, and c0: parameters of the schme.

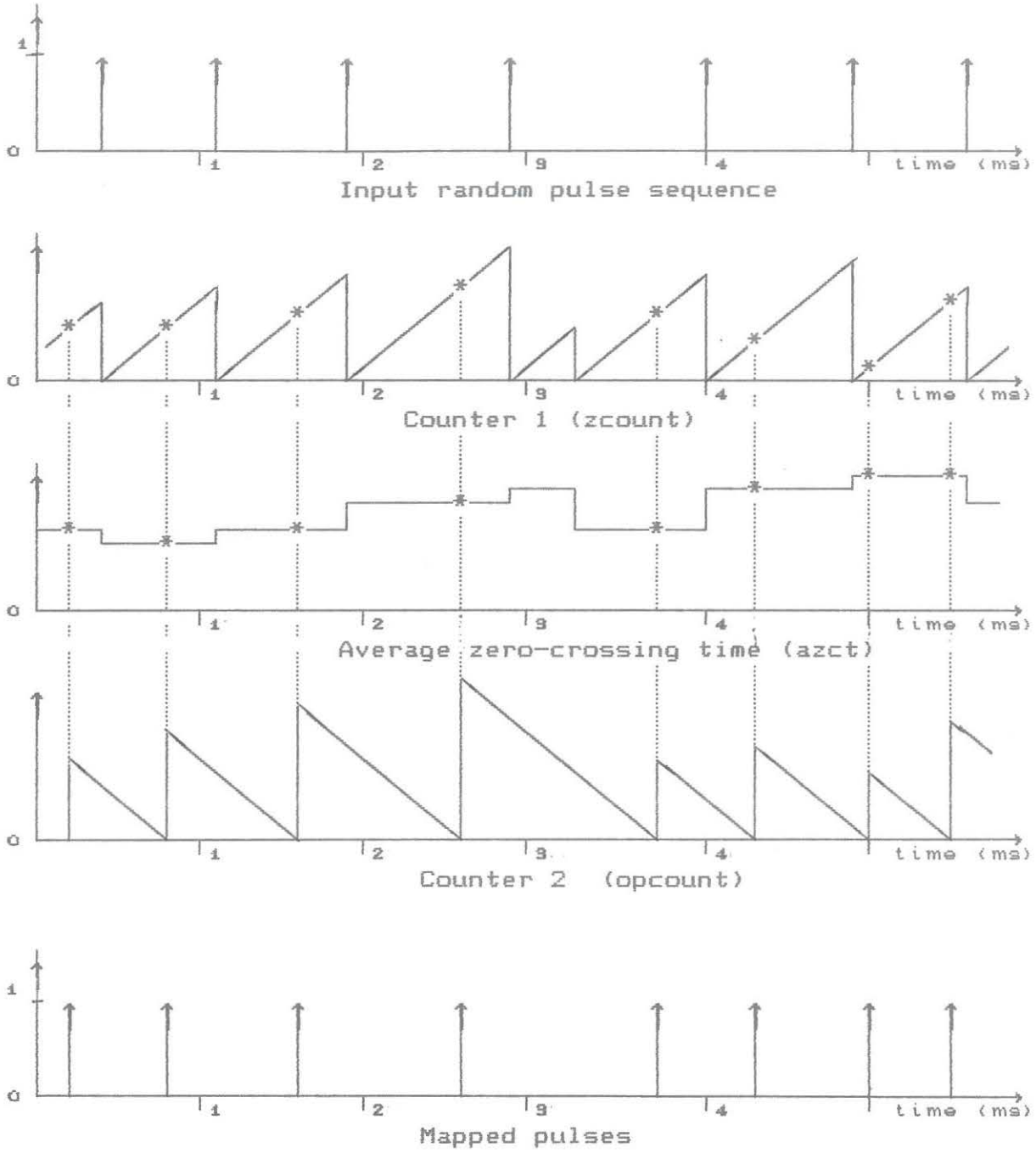


Fig.4.6. Waveforms for the random pulse mapping method.

CHAPTER 5

TESTING OF THE SCHEME

5.1 INTRODUCTION

The implementation of the speech processing scheme along with the pitch estimation methods and noise mapping were discussed in Chapter 4. These methods were thoroughly tested by writing individual test routines or programs. The scheme was tested as a whole from an engineering point of view. The testing part is covered in this chapter.

5.2 TESTING THE PITCH ESTIMATORS

Three methods for pitch estimation were considered: Modified autocorrelation method, dynamic threshold, and dynamic threshold with circular filling. Synthesized vowels /a i u/ for adult male,adult female, and child formants and pitch ranges were used. The cascade pole-zero synthesizer (Kulkarni, 1992) was used. The pitch and amplitude variation is as shown in Fig.5.1. The pitch is constant at value 'pmin' for the first and last 100 ms durations. It doubles within 150 ms, remains steady for next 100 ms and linearly decreases to pmin in 150 ms duration. For adult male, adult female and child the minimum pitch is set to 80 Hz, 150 Hz, and 200 Hz respectively. The formant values

given by Peterson & Barney (1952) were used for synthesis. These vowels form the vowel triangle on the F1-F2 plane which encompasses the formant ranges.

For dynamic threshold methods following parameters were used:

Lowpass filter: Butterworth second order; fc = 900 Hz.

τ = blanking coefficient = 0.75

β = decay ratio = 0.83

autocorrelation threshold = 0.25

For modified autocorrelation method:

clipping level = 0.65

autocorrelation threshold = 0.27

were used.

The results indicate that the dynamic threshold method is as accurate as modified autocorrelation method. But dynamic thresholding with circular filling method may cause small errors. Also for male voice the method fails for pitch periods greater than 10 ms. This is expected because the periodicity is not be retained for periods greater than 1 frame length. For pitch above 100 Hz and for female and child voice, the method was accurate.

5.3 TESTING THE MAPPING SCHEME

The mapping scheme discussed in Section 4.4 was tested for some particular mapping examples. A test signal is obtained by passing random noise through a bandpass filter with suitable center frequency and bandwidth. For zero-crossing rate of 3-7 kHz, a test signal consisting of four segments of 150 ms each with center frequencies at 1.5 kHz, 2.5 kHz, 3.5 kHz, and 4.0 kHz and bandwidths of 10% of center frequencies was generated. A single parallel branch of speech synthesizer was excited with frication source.

A program PC30 was written for testing the scheme. The listing is available in Sapre(1992). It generates an output signal file and a zero-crossings file that contains number of zero-crossings per frame in input and output. The number of zero-crossings is averaged for 150 ms duration or 15 frames. The output signal file is analyzed using LP spectrum analysis program (Rabiner & Schafer, 1979; Kulkarni, 1992). The average center frequency of output is computed. For getting a desired mapping, the mapping function can be adjusted by experimentation to obtain the required average zero-crossing rate or center frequencies, for particular segments.

5.4 MAPPING EXAMPLES

The following mappings were tested.

(1) Input zero-crossing rate of 3-7 kHz is mapped onto

300 to 600 pulses per second (suitable for cochlear implant). A linear mapping, was used. The input zero-crossing rate is scaled down by factor 4 (N = 4)

Zero-crossing rate; 30-70 crossing/100 samples

after dividing by 4: 7.5 to 17.5.

Input zero-crossing time range = (100/17.5 to 100/7.5)

= (5.71 to 13.3)

Output zero crossing time range = $\left[\frac{10000}{600} \text{ to } \frac{10000}{300} \right]$

= (16.67 to 33.33)

A linear mapping is used

$$Y = \frac{[\ 33.33 - 16.67\ ]}{[\ 13.33 - 7.50\ ]} (X - 13.33) + 16.67$$

(5.71)

$$Y = 2.19\ X + 4.2$$

This mapping was tried for synthesized noise data file. The zero crossing rate of output signal is high at the lower side of the band (1.5 kHz center frequency). So a new mapping with 250-600 zero-crossings as output range was tried. This gives the following mapping.

N = 4; Y = 3.06 X - 0.84.

The statistical analysis of zero-crossings for these mappings is shown in Table 5.1.

The new mapping (250-600 pulses/sec) is found more suitable. The mean of o/p zero-crossings per frame in the first segment is 3.3 which is better than the corresponding rate of 3.7 in the first mapping.

(2) A zero-crossing rate of 30-70 pulses/100 samples to be mapped on 400-900 Hz frequency range or 800-1800 pulses per second.

Divide-by-2 counter (N = 2) is used.

The three output ranges were tried:

(a) 800-1800 pulses/sec or Y = 1.82X + 0.348

(b) 700-1800 pulses/sec. Y = 2.37 X - 1.50

(c) 700-2000 pulses/sec. Y = 2.44 X - 1.96

The average zero-crossing rates and standard deviations (SD) and the average values of center frequencies are shown in Table 5.2. The mapping (b) is found most appropriate. The center frequencies are in the range 437 Hz to 871 Hz. Thus fit properly in the assumed range of 400-900 Hz.

(3) Zero-crossing rate of 30-70 pulses/100 samples. mapped on 600-1600 pulses/sec. Divide-by-3 counter was used to cut down the rate. Two mappings were tried:

(a) 600-1600 pulses/sec or Y = 1.83x - 1.50.

(b) 500-1600 pulses/sec or Y = 2.4x - 4.00.

The mean and SD of zero-crossings/frame and center frequencies are compared in Table 5.3. The mapping (b) fits more closely in the required range.

5.5 TESTING THE COMPLETE PROGRAM

The program PC20 has been developed step-by-step. The design is completely modular and the scheme was tested at several levels. The individual subroutines were tested separately and then incorporated in the main program. The interfacing between the subroutines was checked while adding new blocks. The parameters of various blocks were decided finally.

The subroutines for filter, envelope detection were tested separately using test inputs like sine waves, square waves, impulse by checking the outputs. The waveform generation routines were tested with special precaution. There is a processing discontinuity at the frame boundaries. The waveform generation routines were thoroughly checked and modified to ensure that the output waveform is continuous at the frame boundaries.

At the next level, the interfacing among the routines was verified with the help of debugging facility of Turbo-C package.

Finally the values of various parameters of the blocks were decided. This includes the number of samples for averaging for the envelope detectors in the two channels. Digitized speech

files were used for testing. These parameters were adjusted to get smooth envelopes for the signals in the two channels. The relative gain of channel-2 was decided after trying several values and listening to the outputs. The speech file and processed file were plotted together to view the overall operation. Small sections of the output were plotted to verify that the output is undistorted especially at the frame boundaries.

Two sets of 12 consonants in vowel consonant vowel sequence spoken by a male and a female speaker were processed using the off-line implementation. The autocorrelation threshold was found to get proper voiced/unvoiced detection for all the speech segments in the sets. The values of parameters of all the blocks are listed in Table 6.1.

The average time taken for processing a data file of 7000 samples for different PC configurations is as follows.
PC without math coprocessor and floppy drives: 5 min 20 sec
PC with math coprocessor and floppy drives: 1 min 38 sec
PC XT with math coprocessor and hard disk drive: 1 min 5 sec.
Thus for processing a 1 sec segment the machines should take 7 min 30 sec, 2 min 20 sec, and 1 min 30 sec respectively.

Table 5.1. Mapping 3-7 kHz ZCR to 300-600 pulses/sec mean and SD of zero-crossings/100 samples.

			Output Zero-crossing Ranges			
	INPUT		250-600		300-600	
			N = 4		N = 4	
fc			Y=2.19X+4.2		Y=3.06X-0.84	
kHz	mean	SD	mean	SD	mean	SD
1.5	31.1	1.1	3.3	0.5	3.7	0.6
2.5	46.5	2.7	4.5	0.5	4.6	0.6
3.5	64.5	5.5	5.5	0.6	5.7	0.6
4.0	73.4	2.4	6.0	0.5	6.1	0.5

Table 5.2. Mapping 3 kHz to 7 kHz ZCR to 800-1800 Hz range
Mean & SD of zero-crossings per 100 samples.

			O/P Zero-crossing Ranges								
			800-1800 Hz			700-1900 Hz			700-2000 Hz		
INPUT			N = 2 Y=1.82X+0.348			N = 2 Y=2.37X-1.5			N = 2 Y=2.44X-1.96		
fc	mean	SD	mean	SD	fc Hz	mean	SD	fc Hz	mean	SD	fc Hz
1.5	30.7	1.2	9.3	0.5	456	8.7	0.8	437	8.9	0.7	427
2.5	46.1	2.6	12.3	0.7	625	12.5	0.9	637	12.6	1.1	645
3.5	64.1	5.5	15.4	0.9	764	15.8	1.4	826	16.1	1.2	837
4.0	72.9	2.3	17.0	0.8	840	17.6	0.7	871	18.1	0.7	910

Table 5.3. Mapping 3-7 kHz ZCR to 600-1600 Hz range.
Mean & SD of zero-crossings per 100 samples.

			O/P Zero-crossing Ranges					
	INPUT		600-1600 Hz			500-1600 Hz		
			N=3 Y = 1.83X-1.5			N=3 Y = 2.4X-4.0		
fc kHz	mean	SD	mean	SD	fc Hz	mean	SD	fc Hz
1.5	30.7	1.1	7.8	0.7	383	6.7	0.6	320
2.5	46.2	3.0	11.0	1.0	562	10.5	1.2	520
3.5	64.1	5.6	14.3	1.0	716	14.4	1.3	730
4.0	72.8	2.4	15.5	0.6	791	15.5	0.8	784

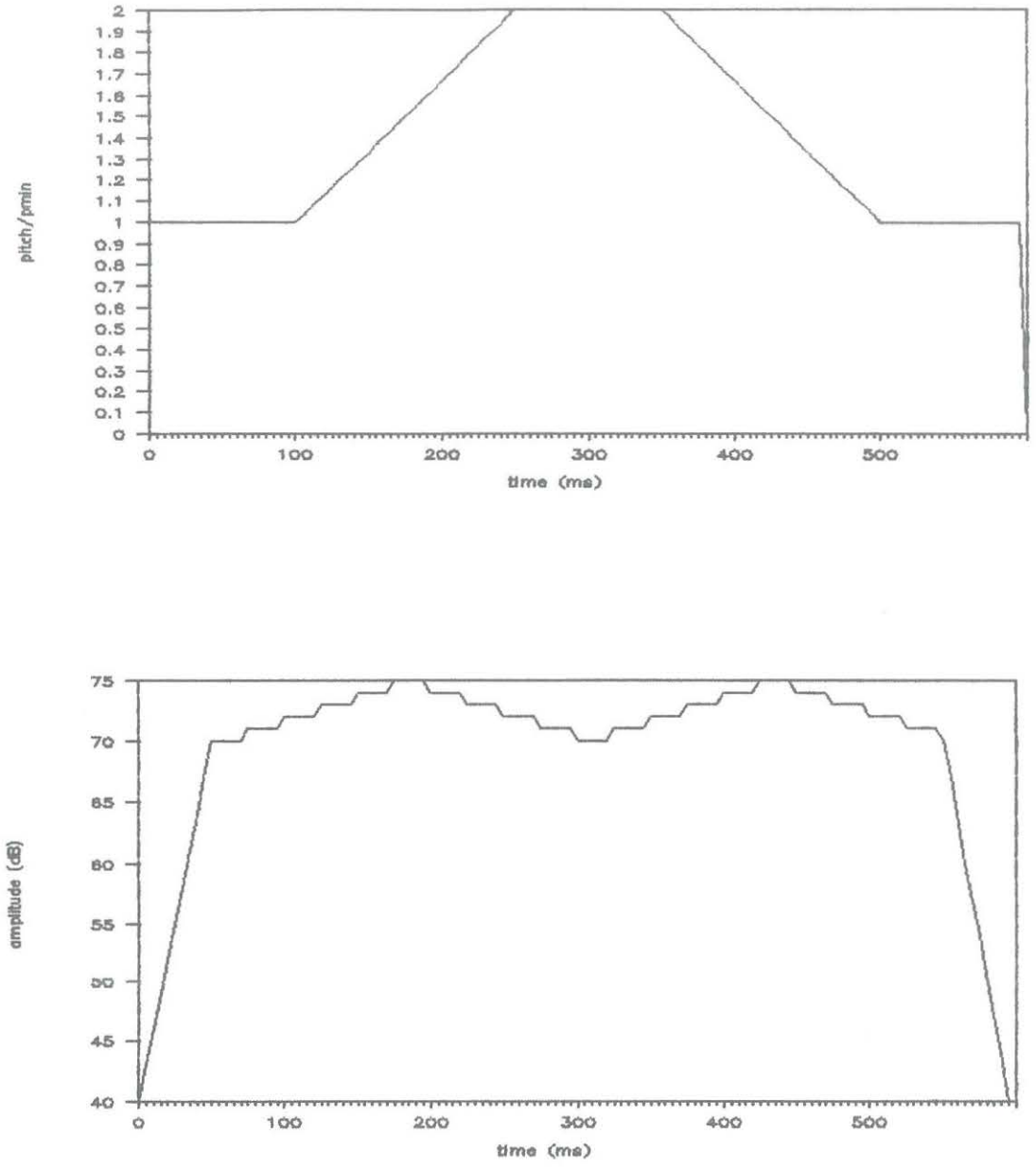


Fig. 5.1 Plot of pitch & amplitude parameter tracks for the synthesized vowels.

CHAPTER 6

LISTENING TEST RESULTS

6.1 INTRODUCTION

A sensory aid is evaluated by performing a series of tests of increasing complexity. One of the basic tests used is the "consonant discrimination test". The subjects under test, are made familiar with processed speech segments consisting of a set of consonants spoken in vowel-consonant-vowel (VCV) context. Then the sounds are presented in a random order. The response for the stimuli are noted, to form a confusion matrix. The confusion matrix has rows corresponding to stimuli and columns corresponding to responses. The diagonal entries show the number of correct responses for stimuli. An entry in the matrix shown the perceptual similarity between the pair of phonemes corresponding to row and column. A statistical analysis is necessary for drawing further conclusions.

In this project work such detailed testing was not carried out due to the lack of time. The scheme was evaluated qualitatively rather than quantitatively, during informal listening tests, by the author. The speech processing details, testing method, and results are presented in this chapter.

6.2 PREPARATION OF DATA SET

Two sets of 12 English consonants /b p m f v d t n z s g k/ in VCV context which /a/ as vowel by one male and one female Canadian speaker were available in digitized form. These sets were used by Pandey et al, (1987) for evaluating their scheme.

The speech was processed assuming a residual hearing of upto 900 Hz. The band 100-350 Hz is used for LF periodicity and band 400-900 Hz is used to mapped the zero-crossing rate of speech signal filtered through 1 kHz highpass filter. The zero-crossing rate was found to lie in 3-7 kHz range for both sets. A mapping function tested earlier (Section 5.4) was used. The parameters of various blocks are listed in Table 6.1. The natural speech and processed output for consonants /g k/ by the female speaker and consonants /z s/ by male the speaker are plotted as Fig 6.1.

6.3 INFORMAL LISTENING TESTS

The processed sounds were scaled to equalize their RMS values. This was done to ensure that overall signal level should not serve as a cue for discrimination.

The experimental set-up consisted of a PC based data acquisition system PCL-208 (Dynalog, 1990), a lowpass filter, an audio amplifier, a headphone, and audio cassette player. The set-up is shown in Fig 6.2. The details of hardware and software are given in Appendix.

The processed sounds were listened to by the author monaurally using cushioned headphone. A program MDA4 was written for presenting any sound out of a given set. The program reads specified data files in the PC RAM. Desired sound can be played by pressing the corresponding key. The key assignment is displayed on the monitor. A segment is played repeatedly with specified inter-presentation gap until any key is pressed. The discriminating factors between all possible pairs were noted down after listening to the pairs in succession to form a "qualitative confusion matrix". The matrices for the two sets are given as Table 6.2 and 6.3.

6.4 CONCLUSIONS

From the informal tests following conclusions are drawn.

1. In all the processed segments the frication information is felt distinct from the periodicity information. The pitch and amplitude variation for voiced sections can be perceived properly.
2. The voiced and unvoiced phoneme pairs can be distinguished clearly. Thus pairs /f v; k g/ etc. can be distinguished on the basis of voicing cue for both the speakers.
3. Stops are felt as complete as complete closure or amplitude drop, followed by a burst, for both the

speakers. For stop /b/ the burst is weak and is not detected but complete closure is felt.

4. In these sets all unvoiced stops are aspirated (/p k t/) The aspiration is perceived as low frequency high amplitude noise. The pairs /b p/, /d t/, /g k/ are thus distinguished by voicing and aspiration cues.

5. The stop pairs /t k/ or /d g/ are perceptually similar (for both speakers). The spectral information about stops is lost after processing. But this information is mostly available from lipreading.

6. The unvoiced fricatives /s,f/ have distinct spectra in digitized male speech. The fricative /s/ is heard as high frequency noise and stands distinct from the rest of the set. The male /f/ is heard as low frequency noise. In digitized female voice, /s/ loses this cue because the dominant band lies above 5 kHz. So /s/ was not heard distinctly in female voice.

Thus the features like manner of articulation, voicing, aspiration, and some spectral information about unvoiced fricatives can be presented. These cues should supplement the information about place, obtained visibly through lipreading.

Table 6.1 Parameters used for processing the data sets.

Processing Blocks:	Parameter values
BPF-1:	Butterworth second order LPF fc = 900 Hz
Amplitude Envelope-1:	200 samples for averaging (20 ms)
Pitch Estimator:	Modified autocorrelation,
	fractonal clipping level 0.65
	autocorrelation threshold: 0.27
Amplitude Envelope-2:	100 samples for averaging (10 ms)
BPF-2:	Butterworth second order HPF fc = 1.0 kHz
Two level threshold detector:	Clipping level = 0.
Pulse mapping:	N = 2; Y = 2.37 X - 1.5;
	a0 = 0.5, b0 = 1.0, c0 = 2.1
Gain of channel-2:	6
Waveform-1:	Square wave LPF fc = 500 Hz
Waveform-2:	Square wave LPF fc = 1 kHz.

Table 6.2. Qualitative confusion matrix for the VCV syllables by the male speaker. Dia gonal cells describe the features of corresponding items. Other cells show discriminating features of coloumn-items w.r.t. those of row-items.

	b	p	m	v	f	d	t	n	z	s	g	k
b	slow decreas-ing; voicing	burst & low freq. uency noise in p	m voiced through out with no gap	low almost constant duration	unvoiced noise duration in /f/	burst present in /d/	silence; burst; noise	constant large voic ing ampli-tude	constant low voic-ing no silence	high freq-noise	strong burst	strong burst; low frequency noise
p	no burst; voicing again.	silence; burst; noise: aspiration	m shows constant voicing	v has voicing duration &no aspir.	total frication no silence	no silence in /d/ also voicing present	/p/ has more silence duration	-,,-	/z/ voiced no noise burst	-,,-	strong burst; voicing follows	strong burst in /k/
m			voicing almost same as /a/ only	small voicing in /v/	noise in /f/	-,,-	burst; followed by noise in /t/	No difference	pitch low in /z/ /z/ longer duration	-,,-	strong burst /g/	burst & following noise in /k/
v			pitch variation felt	low and constant level voicing	/f/ has strong LF noise	burst in /d/	burst; noise interval	/n/ has stronger voicing level	lower pitch in /z/ more duration	-,,-	-,,-	-,,-
f					strong lower freq uency noise	burst in /d/ short duration silence	silence burst noise	/n/ voiced /f/ noisy	slightly different amplitude variation	-,,-	-,,-	-,,-
d						voicing decreases slowly; weak burst; followed by voicing	/t/ has longer silence aspiration	strong voicing in /n/	/z/ voiced without burst	-,,-	HF burst compred to that of /d/	LF noise follows burst
t							long silence; burst; low freq	-,,-	constant voicing in /z/	-,,-	burst only voicing immediate ly follows	strong burst in /t/ else similar
n							noise correspo nding to aspiration	constant voicing pitch lowers	z has less voicing	-,,-	burst in /g/	burst, then noise in /k/
z						-			low level constant voicing	-,,-	burst in /g/	-,,-
s										high frequency high ampli tude long	burst in /g/ short duration	burst then low freq noise
g										duration segment distinct from rest	voicing burst voicing again	silence before burst noise after it
k										of the set		silence, burst then low freq. noise

Table 6.3. Qualitative confusion matrix for the VCV syllables by the female speaker. Diagonal cells describe the features of corresponding items. Other cells show discriminating features of coloumn-items w.r.t. those of row-items.

	b	p	m	v	f	d	t	n	z	s	g	k
b	short silence; weak burst	long silence, aspiration noise	m voiced through out	small frication duration no burst	low freq noise duration in /f/	sharper somwhat higher freq burst	sharper burst then LF noise	n is total voiced -	short weak frication no burst	long duration high freq noise	g has sharper burst voicing follows	strong burst then noise
p		silence, weak burst followed by low	m shows constant voicing level	v has voicing duration &no aspir.	total frication no silence	no aspiration stronger burst	sharp burst in /t/ less noise duration	silence, weak burst aspiration noise	/z/ has low freq frication No silence	-,,-	voicing follows burst in /g/	srong burst in /k/
m		frequency noise corresponding to	voicing level = that of/a/	small voicing in /v/	unvoiced noisy /f/	-,,-	/t/has no voicing /m/voiced through	/m/ & /n/ felt identical	weak frication in z, less & voicing	-,,-	burst is distinguishing factor	-,,-
v		aspiration		decrseased voicing; weak frication	/f/ long duration noise - frication	burst in /d/. /v/ no burst	burst then noise in /t/	/n/ has high level of voicing	/z/ shows longer frication	/s/ has long noise duration	-,,-	burst & LF noise in /k/
f					medium-low frequency, constant amplitude noise	burst in /d/ and no continuous noise	burst then noise in /t/	total voicing in /n/	/z/ has low level of frication	same freq. but pattern differs	-,,-	-,,-
d						small silence; burst; voicing	LF noise discriminates /t/ from /d/	burst absent in /n/	no burst in /z/	burst absent. /s/ longer duration	/g/ has a broader burst	noise follo burst in /k/
t							sharp burst & then LF noise	-,,-	no burst weak frication in /z/	burst absent. /s/	noise absent after burst in g	somwhat lower freq. burst in /k/
n								high & constant voicing amplitude	frication & low voicing level in z	frication no voicing in /s/	low voicing; burst in /g/	burst; aspiration in /k/
z									frication low voic- level in /z/	longer frication duration in /s/	burst is distinguishes /g/	burst & aspiration noise in /k/
s										strong medium frequency frication	burst in /g/ small duration	burst, aspiration noise in /k/
g											decreasing voicing; burst; voicing	silence; aspiration noise in /k/
k												silence; burst; low freq. noise

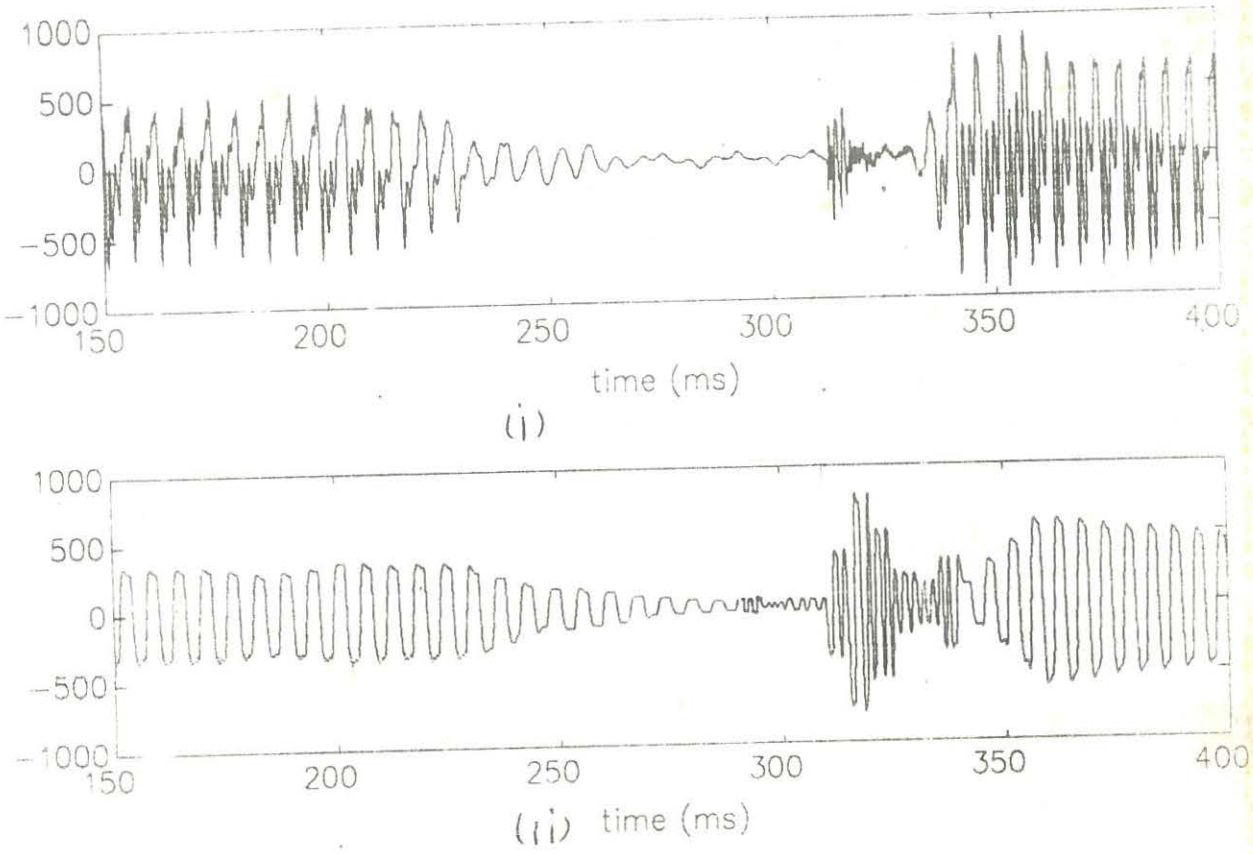


Fig 6.1(a) (i). Speech segment /aga/ by a female speaker.

(ii). Processed output for residual hearing upto 900 Hz.

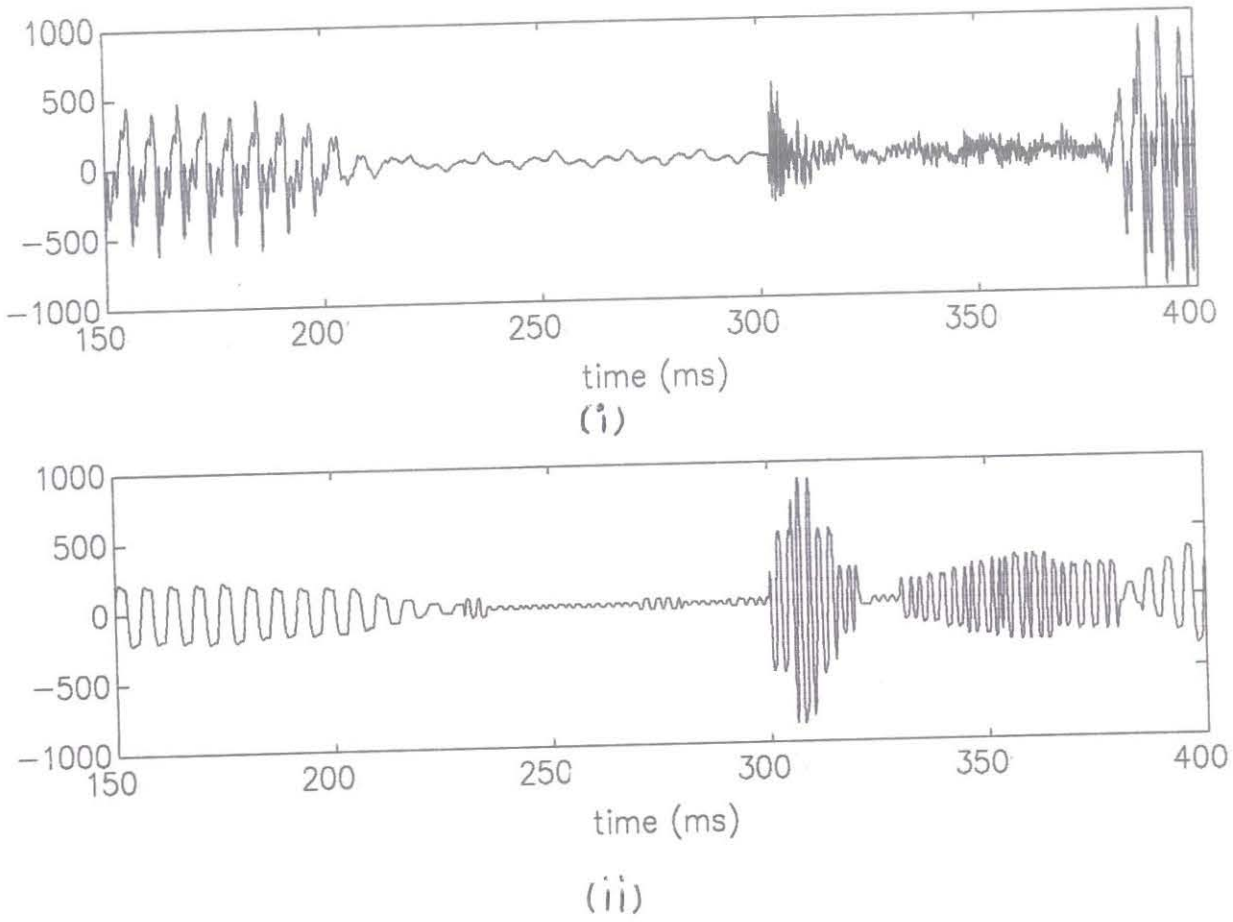


Fig 6.1(b) (i). Speech segment /aka/ by a female speaker.
(ii). Processed output for residual hearing upto 900 Hz

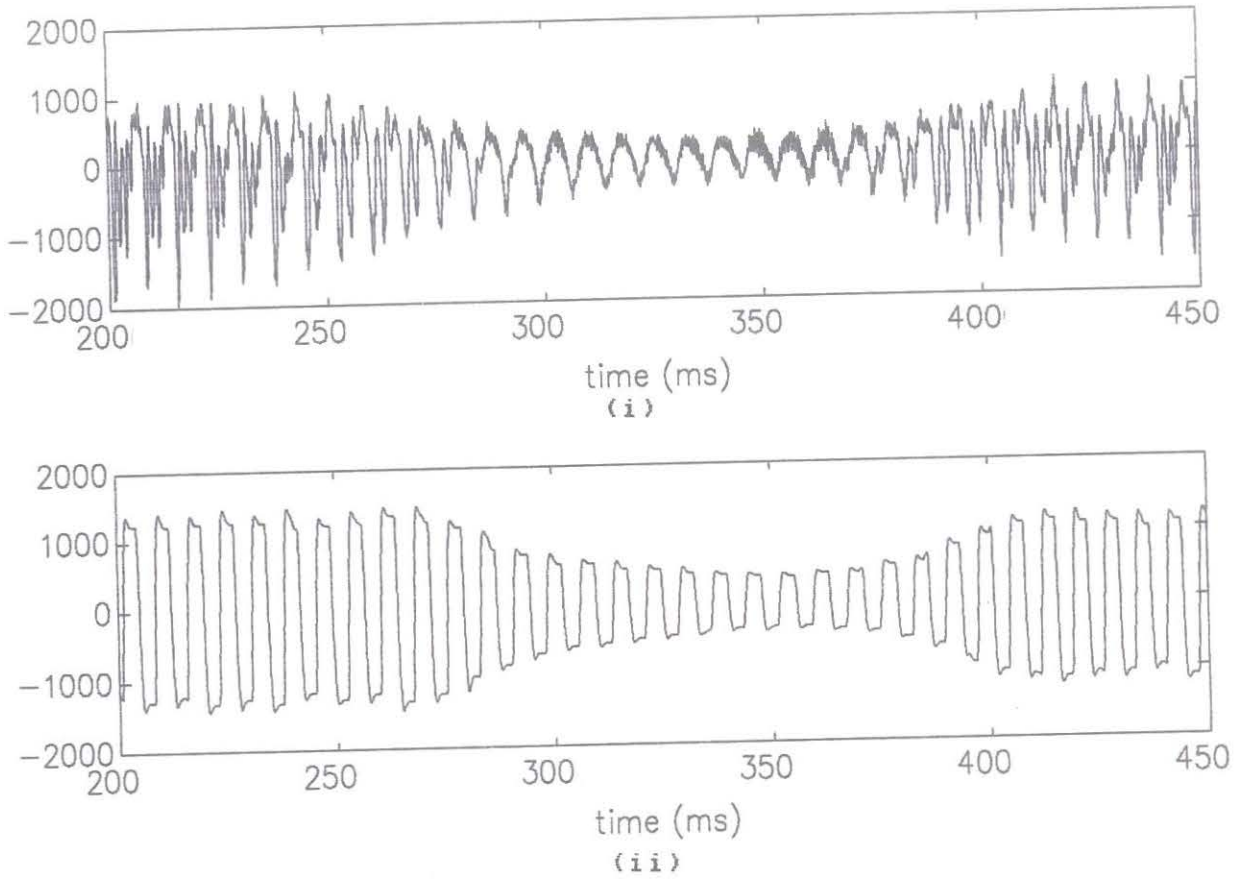


Fig 6.1(C) (i). Speech segment /aza/ by a male speaker.
(ii). Processed output for residual hearing upto 900 Hz.

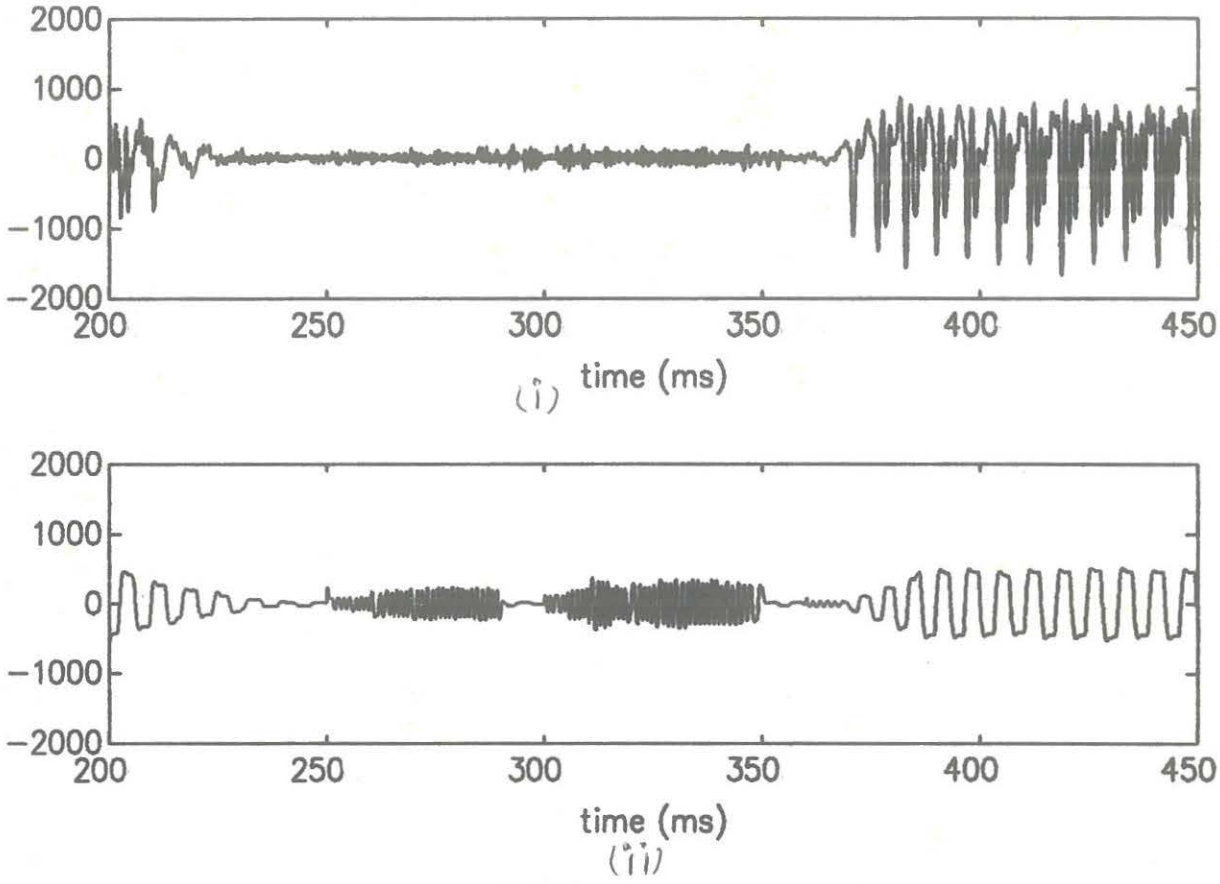


Fig 6.1(d) (i). Speech segment /asa/ by a male speaker.
(ii). Processed output for residual hearing upto 900 Hz.

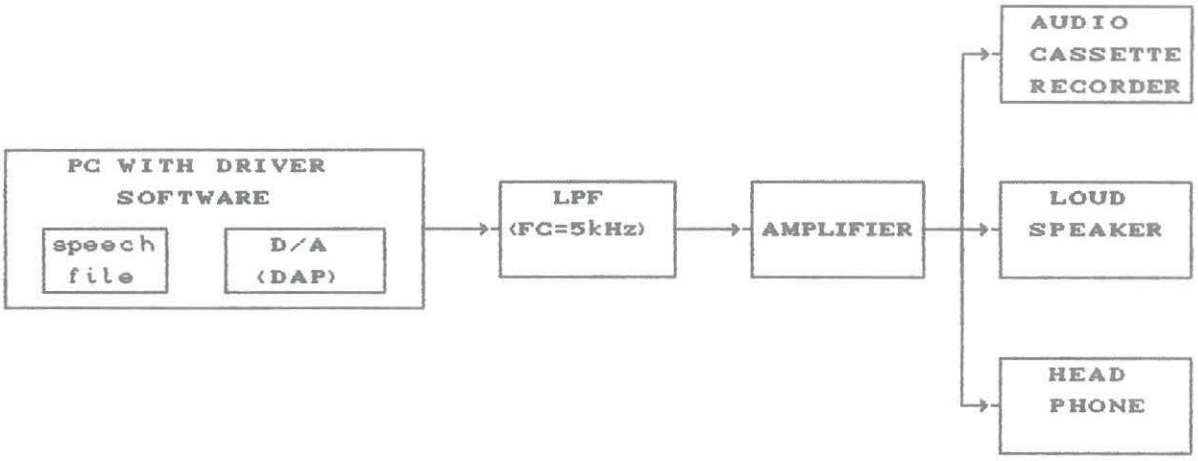


Fig. 6.2 The experimental set-up.

CHAPTER 7

SUMMARY & SUGGESTIONS

7.1 WORK DONE

A scheme that takes into account the phonetic features of Indian languages has been designed during this project work and the scheme is a modification of a scheme for single channel cochlear prosthesis reported earlier in the literature. The scheme was implemented for off-line processing. A simpler pitch estimation method reported earlier was tested for male, female, and child formants and pitch ranges. A random pulse sequence to random pulse sequence mapping is essential for presenting the frication information. A method for this purpose was designed and the typical values of parameters involved were found out by experimentation. Random noise in a given band is mapped on other specified band by using this method. A program for comparing the set of digitized sounds was developed.

Two sets of 12 vowel-consonant-vowel syllables one from a male and another from a female speaker were processed by the speech processing scheme. The sounds were listened to in informal listening tests. The discriminating features between all possible combinations of syllables were noted to analyze the scheme qualitatively. The scheme could be useful for single channel prosthesis.

7.2 SUGGESTIONS FOR FURTHER WORK

In this scheme, voiced/unvoiced detection is done by pitch extactor. The viced/unvoiced detection could be made more relible by using signal magnitude or zero-crossing rate as additional criteria.

The consonant set from Indian languages such as Hindi or Marathi should be selected for testing the scheme. The scheme should be evaluated for normal hearing subjects. This should be followed by tests with hearing impaired subjects. For these tests appropriate amplitude compression should be done. Online version of the scheme should be developed on a DSP microprocessor board to carry out tests with lipreading and speech tracking test.

REFERENCES

Biswas S (1990). Effect of Sinusoidal Magnetic Field on K562 and Hybrid Cell Lines, M.Tech. thesis, School of Biomedical Engg. I I T Bombay.

Blamey PJ (1985). Psychophysical and speech studies with an electrotactile speech processor. In Clark et al (1987) pp 87-89.

Borland (1988a). Turbo C 2.0 Reference Guide, CA: Borland Int.

Borland (1988a). Turbo C 2.0 User's Guide, CA: Borland Int.

Breeuwer M & Plomp R (1984). Speechreading supplemented with frequency selective sound pressure information. J.Acoust. Soc. Am., vol 76(3), pp 686-691.

Breeuwer M & Plomp R (1985). Speechreading supplemented with formant frequency information from voiced speech. J. Acoust. Soc. Am., vol 77(1), pp 314-317.

Clark GM & Busby PA (1987). Papers presented in International Cochlear Implant Symposium and Workshop, Melbourne, 1984 Supplement 128, vol. 96(1) part-2, Anals of Otol. Rhinol. Laryngol.

Cowan RCS, Alcantra JI, Balmey PJ & Clark GM (1988). Preliminary evaluation of a multi channel electrotactile speech processor, J. Acoust. Soc. Am., vol 83(6), pp 2328-2338.

Denes PB (1963). On the statistics of spoken English. J. Acoust.

DMS (1990). PCL-208 Data Acquisition Card User's Manual. Bombay: Dynalog Micro-systems.

Ebrahimi D (1987). Preliminary Research on Peripheral Vision Lipreading Aid. M.S. thesis, Electrical Engineering Department, University of Toronto.

Flanagan JL (1972). Speech Analysis, Synthesis, and Perception New York: Springer-Verlag.

Gelfand SA (1981). Hearing, An introduction Psychological and Physiological Acoustics. New York: Marcel Dekker.

Goldstein Jr.Mh & Procter A (1985). Tactile aids for profoundly deaf children. J. Acoust. Soc. Am., vol 77(1), pp 257-265.

Grant KW, Adrel LH, Kuhl DW, & Sparks DW. (1985). Contribution of fundamental frequency amplitude envelope and voicing duration cues to speechreading in normal hearing subjects. J. Acoust. Soc. Am., vol 77, pp 671-677.

Grant KW (1985). Encoding voice pitch for profoundly hearing impaired listeners. J. Acoust. Soc. Am., vol 78(S1), pp S-42.

Gray RF, Ed. (1985). Cochlear Implants. San Diego, CA: College-Hill Press.

Guttman N & Nelson JR (1968). An instrument that creates some artificial speech spectra for the severely hard of hearing. Amer. Annals. Deaf, vol 113, pp 295-302, Reprinted in Levitt et al (1984 a), pp 211-212.

Ifkube T (1988). A speech coding method for an auditory

prosthesis by extracochlear stimulation. J. Acoust. Soc. Am., vol 84, pp S-41.

Johansson B. (1966). The use of the transposer for the management of the deaf child. Int. Audiol., vol 5, pp 362-372, Reprinted in Levitt et al (1984 a), pp 195-204.

Kiedel WD & Finkenzeller P Eds.(1984). Artificial stimulation theories. Advances in audiology, vol 2, papers presented, Symp. Artificial Auditory Stimulation, Erlanger, FRG, autumn 1982, New York: Karger.

Kirman JH (1974). Tactile communication of speech: A review and analysis. Psychol Bulletin, vol 80, pp 54-74, Reprinted in Levitt et al (1984 a), pp 297-317.

Kulkarni DM (1992). Developement of A Cascade Pole-Zero Speech Synthesizer. M. Tech. thesis, Electical Engg. Dept., I I T Bombay.

Levitt H (1988). Signal processing for sensory aids. J. Acoust. Soc. Am., vol 84(S1), pp S-39.

Levitt H, Pickett JM & Houde RA, Eds. (1980 a) Sensory Aids for the Hearing Impaired. New York: IEEE Press

Levitt H, Pickett JM, & Houde RA (1980 b). Auditory and speech impairments. In Levitt et al (1980 a), pp 3-25.

Lippman RP (1980). Perception of frequently lowered consonants. J. Acoust. Soc. Am., vol 67, pp S-78.

Moore BCJ (1985). Speech coding for cochlear implants in Gray, (1985), pp 163-179.

O'Shaughnessy D (1987). Speech Communication, Human and Machine Massachusetts: Addison Wesley.

Pandey PC (1987). Speech Processing for Cochlear Prostheses Ph.D. thesis, Electrical Engineering Department, University of Toranto.

Pandey PC, Kunov H, & Abel SM (1987). A speech processor providig fricative and low frequency information for single channel cochlear prosthesis. Proc. ICASSP. pp 1422-1425.

Pickles JO (1982). An Introduction to the Physiology of Hearing. London: Academic.

Plant G & Dillon H (1985). Single transducer vibrotactile aid to lipreading and speech production. In Clark et al (1987), pp 89.

Posen MP, Reed CM, Brinda LP, & Durlan NI (1984). Improved frequency lowering techniques. J. Acoust. Soc. Am., vol 75 (s1), pp S-32.

Procter A & Goldstein Jr MH (1988). Development of speech perception and speech production in children who used vibrotactile aids. J. Acoust. Soc. Am., vol 84, pp S-46.

Rabiner LR & Schafer RW (1978). Digital Processing of speech signals. Englewood Cliffs, NJ: Princtice-Hall.

Risberg A (1969). A critical review of work on speech analysing hearing aids. IEEE Trans. Audio Electroacoust, vol 17 pp 240-247, Dec. 1969. Reprinted in Levitt et al (1984 a), pp 213-220.

Sapre RM (1992). Listing of Programs developed during M Tech Project, I I T Bombay.

Sebastian G (1991). A Speech Training Aid for the Deaf. B. Tech. project report, Electrical Engineering Department, IIT Bombay.

Sherrick CE (1984). Tactile aids for the deaf. J. Acoust. Soc. Am., vol 75(5), pp 1325-1340.

Wiessenberger JM & Miller JD (1987). The role of tactile aids in providing information about acoustic stimuli. J. Acoust. Soc. Am., vol 82(3), pp 906-915.

APPENDIX

SIGNAL HANDLING

A.1 INTRODUCTION

A computer aided signal handling system is needed for the present project work. The A/D converter can be used to acquire natural speech segments in digitized form. The segments can be processed by the speech processing program. The listening tests can be conducted by using the A/D conversion. Programs can be developed to handle the test signals, and to control the experiment.

The experimental set-up and the signal handling software are described below.

A.2 HARDWARE SET-UP

Hardware setup used for digitizing natural speech and for listening to the synthesized speech. This include data acquisition card PCL208, a lowpass filter, and an audio amplifier.

PCL208 is a data acquisition card for IBM PC/XT/AT or compatible. This card following features.

(1) Switch selectable 16 single ended or 8 differential analog input channels.

(2) 12 bit successive approximation converter is used to convert analog inputs. The maximum A/D sampling rate is 60 kHz in DMA mode.

(3) Switch selectable, analog input ranges Bipolar:

+/-0.5V, +/-1V, +/-2.5V, +/-5V, +/-10V. Unipolar: +1V, +2V, +5V, +10V

(4) Provides three A/D trigger modes: software trigger, programmable pacer trigger, and external pulse trigger.

(5) A/D converted data can be transferred by program control, interrupt handler routine or DMA transfer.

(6) An INTEL 8254 programmable Timer/Counter provide pacer (trigger pulses) at the rate of 2.5 MHz to 0.00023 Hz to A/D. The time base is switch selectable 10MHz or 1MHz.

(7) Two 12 bit multiplying D/A output channels. Output range of 0-5 V can be created using on-board -5 V reference. External AC or DC reference can also be used to generate other D/A output ranges.

(8) TTL/DTL compatible 16 digital input and 16 digital output channels.

In the present work only one signal channel was needed. A/D channel 0 was used in -2.5 to +2.5 V range and D/A channel 0 was used in +5 V range.

For A/D conversion input signal should be band limited to 5 kHz. Also, a filter is necessary to get a smooth waveform from staircase waveform obtained at the output of D/A converter. A seventh order elliptic filter was used for this purpose. It has a cutoff frequency of 4.6 kHz. It has a

passband attenuation of 0.3 dB and stop band attenuation of 40 dB. The design and hardware details of this filter are given in Sebastian (1990).

The speaker or headphone was driven by an audio amplifier. This amplifier provides two single ended outputs or 1 differential output. The details of this amplifier circuit can be found in Biswas (1991).

A.3 SOFTWARE SET-UP

The software for data acquisition can be written in a high level language with assembly language routines for controlling the time critical I/O tasks. A program MDA was written for listening to multiple sound segments stored in different data files. The program was written in Turbo-C with assembly routine linked for D/A operation.

A signal file should contain the information about the segments to be listened in the following format. The number of files is the first entry. This is followed by the name of the data file and a name (of 1-2 characters) identifying the corresponding sound in each of the following lines. The sampling frequency (upto 50 kHz), the range of data to be presented, inter-presentation gap, and signal file name are specified interactively. The data can be optionally scaled to fit in -2048 to 2047 range. The data files are read into the PC RAM at the beginning.

A key is assigned to each sound and the assignments between keys and names are displayed on the screen. User can select required sound item by pressing the corresponding key. The sound is presented repeatedly until any key is pressed. Two or more sounds can be listened to in succession by rapidly and sequentially pressing the corresponding keys twice.

A program SCLINT (scale integer) was written to scale the data files, i.e., to get required overall signal level. The program can be used to find out the signal range and its RMS value. The samples can be scaled to fit in the specified range or to give specified RMS value. Scaling by absolute factor is also possible. The scaled values are limited to -2048 to 2047 range. A set of data files can thus be matched to get almost equal average intensity. This is essential for listening tests as average intensity may serve as a cue for discrimination. The data files used by various programs are in an economic binary format. A program T2B was written for text-to-binary or binary-to-text format conversion. A program PLOT written by Kulkarni (1992) was used to plot two data files simultaneously on the screen for comparison.