A SPEECH TRAINING AID FOR THE DEAF

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Master of Technology

by

PRASHANT SHASHIKANT GAVANKAR

(93307020)

Guides

Guide: Dr. P.C. Pandey

Co-guide: Prof. S. D. Agashe

Department of Electrical Engineering,

Indian Institute of Technology, Bombay.

January 1995.

INDIAN INSTITUTE OF TECHNOLOGY, BOMBAY

DISSERTATION APPROVAL SHEET

Dissertation entitled **"A Speech training aid for the deaf"** by **Prashant S. Gavankar** is approved for the award of the degree of Master of Technology, in Electrical Engineering.

Guide	:	P.C. Pandey	(Dr. P.C. Pandey)
Co-Guide	:	S. D Agashe	(Prof. S. D. Agashe)
Internal Examiner	:	V.M. Gadre	(Dr. V.M. Gadre)
External Examiner	:	R.S. Sardesai	(Shri. R.S. Sardesai)
Chairman	:	Tarun Kant	(Prof. Tarun Kant)
Date	:	3 Feb. 1995	

CONTENTS

Abstract

Prashant S. Gavankar, "Speech training aid for the deaf," M. Tech. dissertation, Dept. of Elec. Engg., Indian Institute of Technology, Bombay, Jan 1995.

One of the ways of training the profoundly deaf people to acquire proper articulatory features of speech, is by providing a visual or tactile feedback of vocal tract shape, pitch, and energy level. Such a training aid should be capable of estimating and displaying these parameters in real time on the monitor. The aid should also provide the facility of slow motion review.

The hardware setup of the project consists of analog signal conditioning unit, DSP board PCL DSP25, and PC. The speech signal is input through a microphone, passed through preamplifier and 7th order active elliptic low pass filter with a cutoff frequency 4.6kHz. The signal is then fed to the ADC of the DSP board. The signal conditioning unit also provides the facility for four different outputs for microphone, tape recorder, test signal, and speaker.

In this project the vocal tract shape is estimated considering it as a cascade of cylindrical tubes. The Le-Roux Gueguen algorithm is used to estimate the reflection coefficients from the autocorrelation coefficients. The area function is calculated from the reflection coefficients on the PC. The pitch and the energy is estimated from the autocorrelation coefficients on the PC. The center clipping is done to remove the extraneous peaks due to the formant structure of the vocal tract. To increase the accuracy of estimated parameters the block overlap technique is implemented. The size of each frame is 256 samples. The algorithm is then tested for various speech files at different intensity levels.

The system can also display the target vocal tract shape. In addition to the real time mode the system offers the facility for slow motion review of the analysis results.

ACKNOWLEDGEMENTS

I sincerely thank my guides Dr. P.C. Pandey and Prof. S.D. Agashe for constant encouragement and valuable guidance throughout the project. I also thank them for their support extended to me when it was most needed.

I would also like to thank Mr. A.D. Apte of Standards Lab. and my colleagues, who helped me out throughout this project work.

I.I.T. Bombay.
January 1995.

Prashant S Gavankar
(93307020)

LIST OF FIGURES.

LIST OF TABLES.

CHAPTER 1

INTRODUCTION

1.1. OVERVIEW OF THE PROBLEM

The prelingually deaf people have considerable difficulty in uttering an intelligible speech because of the lack of auditory feed back. There is no physiological disorder in their speech production mechanism. A deaf person may be trained to produce normal speech sounds if a feedback is provided by alternative non-auditory means.

Natural methods used to understand and acquire speech by the deaf are lipreading, tactile speech reading, and finger spelling. A deaf person can also be taught to produce correct speech gestures though tactile sensing of teacher's face, neck, and breath stream (Nickerson 1980). Each speech phoneme has its own characteristic vocal tract shape, pitch, intensity level, and voicing pattern. The pitch and intensity variation convey the suprasegmental information. Thus the vocal tract shape, pitch, and intensity are the important features that convey the information present in the speech sound produced (Padro 1982). Hence by means of providing the feedback in terms of these parameters the deaf person can be trained to acquire proper features of speech.

A speech training aid extracts these parameters and display them on a monitor screen, offers a great flexibility in terms of presentation, storage, and quick review of the data (Levitt et al 1980). In such a training aid, the deaf person undergoing training speaks into a microphone and is able to see the vocal tract shape, pitch, and intensity variations for the speech sound uttered in real time on the screen in the graphical form. The speech training aid should be able to extract and display the parameters on the screen with the minimum processing delay in order to establish the correspondence between the sound

produced and the parameters displayed (Pandey 1987).

1.2. SPEECH TRAINING AIDS

Spectrographic display of speech signals have also been experimented with as visual speech training aid(Bolt 1969). Information available in the spectrographic displays include the formant frequencies and their transitions and the fundamental frequency of voicing. Though the formant frequencies are related to the vocal tract shape the information presented is really hard to interpret for achieving the correct articulation.

A system developed by Bernstein et al used a PC screen for the display and the signal processor TMS32010 for extracting pitch information from the speech signal in real time (Bernstein 1986). The drawback of this training aid is that it does not convey any information about the actual effort required to produce speech of intended quality.

Bristow et al have reported a vocal tract shape display aid using a microprocessor fast enough for real time processing and a domestic television set for the display (Bristow et al). Display provides the plot of vocal tract area verses the liner distance along the vocal tract from glottis to the lips. In speech training design by Padro area function was extracted using linear predictive coding (Padro 1982). A special interpolation algorithm is designed to improve the performance of the aid. In a modified form of this aid a realistic vocal tract shape is displayed on the screen and is manipulated according to the area function values available after processing the speech signal frame by frame. Signal processor micro p7720 from NEC is used to perform speech analysis.

A speech training aid developed by Shigenaga and Kubo also intends to train the deaf in articulation of vowels and some consonants (Shigenaga et al 1986). It shows the place and manner of articulation required to produce the intended vowel by displaying the reference vocal tract shape and that of the

actually uttered sound superimposed on each other. The display has more physical understandability in terms of identifying different articulators such as tongue, lips etc.

1.3. THE EFFORTS AT I.I.T. BOMBAY

In earlier efforts at I.I.T. Bombay, a hardware setup and programs are developed for implementing a speech training aid (Gupte 1990, Taklikar 1991, Gracias 1991, Khambete 1992, Bhagwat 1994). Speech segment is uttered through a microphone, processed on the signal processor, and the extracted parameters are transferred to the PC for display.

The input speech signal is processed on frame by frame basis with a frame length of 30 ms. Acquisition of the data for the next 30 ms is done simultaneously when the previous frame is analyzed. These analysis parameters are to be transferred to the PC for the display while the newly acquired frame is being analyzed. Analyzed parameters are to be processed and displayed on the screen by the PC in the following time frame of 30 ms. Thus for a particular frame the display of the vocal tract shape for it will be displayed after a delay equal to 90 ms. This value is within the acceptable limits of the delay which is not disruptive for the perception of the speech while viewing the speaker's face (Pandey 1987). Speech parameters analyzed by the signal processor are transferred to the PC for the display on the screen. The

The overall real time performance can not be achieved by Gupte (1990) as programs running on PC are in higher level language.

The system by Taklikar (1991) gives a graphical display of the vocal tract from the estimated parameters. She tried to make the vocal tract shape display more realistic.

Gracious (1991) developed a software on TMS32010 that estimates the vocal tract and the pitch of the data file on off-line basis. Due to some hardware fault in the board he could not achieve the real time performance.

Khambete (1992) has implemented the speech training aid in real time display and slow motion review modes. But he could not display pitch as well as vocal tract on the monitor simultaneously.

In the system developed by Bhagwat (1994), graphical display of vocal tract shape, pitch, and energy level is possible on the monitor. The vocal tract image is generated in the external memory of the DSP board and is then transferred by the PC to the video RAM. It can capture and store the vocal tract shape for a number of analysis frames and provides the facility to browse through previously captured frames.

The systems developed till now treat the incoming digitized speech signal on block by block basis. Each block is of the size of 256 samples which is the same as the Hamming window length. The parameters estimated in disjoint blocks should be correlated to one another. Also to improve the accuracy in the estimated parameters, the parameter estimation should be carried out on overlapping block basis. For data acquisition and signal conditioning a general purpose unit should be built. Extensive testing of the algorithm for different speech files at various energy levels has to be carried out.

The block diagram of the hardware setup used is as shown in the Fig 1.1. The speech segment uttered through the microphone is amplified by input preamplifier. The output of the preamplifier is passed through a 7th order active elliptic lowpass filter with a sharp cutoff at 4.6 kHz. The signal is then sampled at 10 kHz by on board ADC of the PCL DSP25 add-on card. The Parameters are estimated on the DSP board and the image is generated on the PC. The image is directly transferred to the video RAM for display.

1.4. PROJECT OBJECTIVES

The aim of the project is to develop a real time speech training aid for the deaf and as mentioned earlier it is in continuation with earlier work done at I.I.T. Bombay.The software

in the system is to be tested extensively for different speech files at various energy levels. The system should also be tested for various natural and synthetic sounds.

The present system estimates the speech parameters without window overlapping. So to improve the accuracy of the estimated parameters the overlapping window technique is to be used.

The vocal tract shape in not reliable in case of utterances where there is complete closure. So the algorithm is to be modified accordingly.

1.5. OUTLINE OF THE REPORT

Chapter 2 "Vocal tract shape estimation", describes the model of speech production. Vocal tract shape estimation algorithms are discussed.

In the third chapter "Pitch and Energy estimation", a description about various pitch estimation algorithms is given. The estimation of energy from the autocorrelation coefficients is also discussed. Various non linear transformations for eliminating the effect of formant structure of the vocal tract for accurate estimation of pitch is also discussed.

"Implementation setup", is the fourth chapter which explains the analog signal handling units built as a part of the project. Unit at the input of the DSP board has preamplifier and active elliptic filter. While the unit at the analog output of the DSP board has elliptic filter and the output board.

Chapter 5 "Software development", presents the software description in the form of modules. The allocation of various tasks between PC and the DSP board is discussed.

In the sixth chapter "Test results", the results for various speech files are listed. The speech files generated at various energy levels at the output of the data acquisition card are processed to estimate the parameters. The results are found to be consistent.

The last chapter gives the summary and suggestions for further work. The chapter summarizes the work done in the three stages of the project and suggests the plans for the further work.

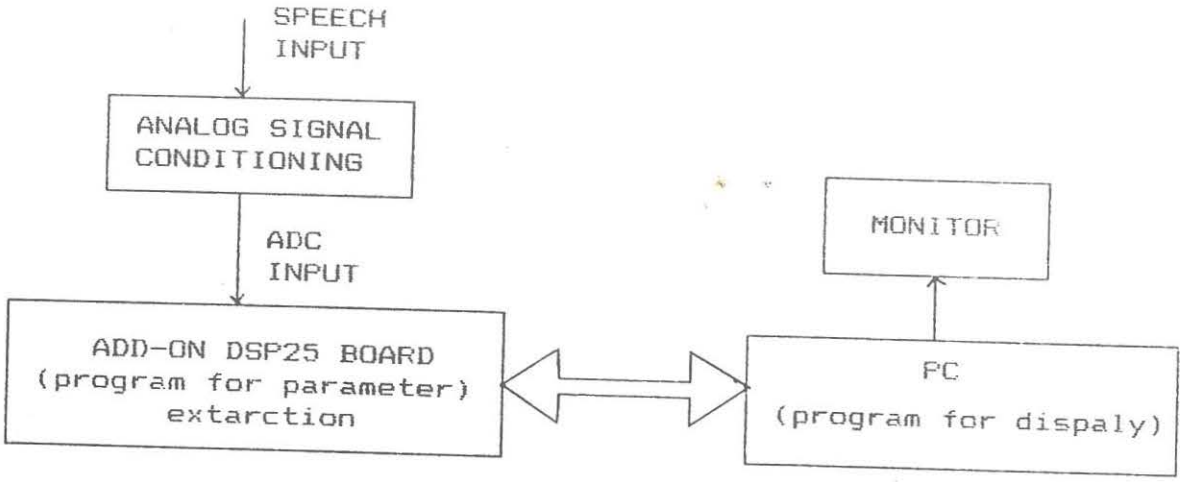


Figure 1.1 Hardware setup of the system.

CHAPTER 2

VOCAL TRACT SHAPE ESTIMATION

2.1. INTRODUCTION

X-ray techniques have been used as direct techniques for the determination of the vocal tract shape. But there are difficulties in the measurement of the lateral dimensions of the vocal tract. The indirect methods for estimating the vocal tract shape are based on the principle of the inverse filter model. This method makes possible the simultaneous extraction of the vocal tract area function. Any such estimation method should have following properties for its real time implementation.

1 Algorithm should be polynomial type.
2 Its implementation should be possible using fixed point arithmetic.
3 It should be accurate.
4 It should be numerically stable even with round off and truncation errors.

2.2. SPEECH PRODUCTION MODEL

A block diagram of the model of the speech production system is as shown in the Fig 2.1. The speech signal is modeled as the output of a cascade of three filters driven by an impulse train for voiced sound or random noise for unvoiced sound (Rabiner 1978).

$$S(z) = V(z)\ G(z)\ U(z)\ R(z) \qquad 2.1$$

where

$S(z)$ = z-transform of the speech signal.
$V(z)$ = vocal tract transfer function.
$G(z)$ = glottis characteristics.

$R(z)$ = radiation characteristics.

$U(z)$ = excitation.

The excitation in this model is either periodic impulse train or random noise for the voiced and unvoiced sounds respectively. Hence $U(z) = 1$.

Lossless tube model:

A widely used model for the speech production is based upon the assumption that the vocal tract can be represented as a concatenation of lossless acoustic tubes of uniform lengths and varying areas of cross-section as shown in the Fig 2.2. The area function of the vocal tract is defined as the variation of its cross-sectional area with respect to its distance from the glottis. If large number of tubes of short lengths is used, it can reasonably be expected that resonant frequencies of the concatenated tubes to be close to those of a tube with continuously varying area function. Since this approximation neglects the loses due to friction, heat conduction, and wall vibration, the bandwidths of the resonances differ from those of detailed model which includes these losses.

Let the vocal tract be divided into an arbitrary number of sections, M each of length l. The solution of the acoustic wave equation results in the waves travelling in the forward and backward directions. The pressure $p_m(x,t)$ and the volume velocity $u_m(x,t)$ at the m^{th} section are given by

$$u_m(x,t) = u_m^+(x,t) - u_m^-(x,t) \qquad 2.2$$

$$p_m(x,t) = \frac{\rho c}{A_m}[u_m^+(x,t) + u_m^-(x,t)] \qquad 2.3$$

where,

$u_m^+(x,t)$ = volume velocity in the forward direction. (Glottis to lips)

$u_m^-(x,t)$ = volume velocity in the backward direction. (Lips to Glottis)

$p_m(x,t)$ = pressure.

ρ = density of air.

c = velocity of sound.

A_m = cross-sectional area of the m^{th} tube.

x = distance from the lips.

The relationship between travelling waves in the adjacent tubes can be obtained by applying the principle that pressure wave and the volume velocity are continuous in both time and space everywhere in the system.

$$u_{m+1}(x_m, t) = u_m(x_m, t) \qquad 2.4$$

$$p_{m+1}(x_m, t) = p_m(x_m, t) \qquad 2.5$$

where x_m is the distance from the lips to the boundary of m^{th} and $(m+1)^{th}$ section. As the tube is assumed to be lossless, $u^+_{m+1}(x_{m+1}, t)$ is same as $u^+_{m+1}(x_m, t)$ delayed by time $\tau = l/c$.

$$u^+_{m+1}(x_m, t) = u^+_{m+1}(x_{m+1}, t - \tau) \qquad 2.6$$

$$u^-_{m+1}(x_m, t) = u^-_{m+1}(x_{m+1}, t + \tau) \qquad 2.7$$

Using equation 2.2 to 2.7 we get,

$$u_m(t) = u_m^+(t) - u_m^-(t) = u^+_{m+1}(t - \tau) - u^-_{m+1}(t + \tau) \qquad 2.8$$

$$p_m(t) = \frac{\rho c}{A_m}[u_m^+(t) + u_m^-(t)] = \frac{\rho c}{A_{m+1}}[u^+_{m+1}(t+\tau)^- + u_{m+1}(t+\tau)] \qquad 2.9$$

By rearranging,

$$u^+_{m+1}(t - \tau) = \frac{1}{1 + r_m}[u_m^+(t) - r_m u_m^-(t)] \qquad 2.10$$

$$u^-_{m+1}(t + \tau) = \frac{1}{1 + r_m}[-r_m u_m^+(t) + u_m^-(t)] \qquad 2.11$$

where,

$$r_m = \frac{A_m - A_{m+1}}{A_m + A_{m+1}} \qquad 2.12$$

If the volume velocity is digitized using the sampling interval $T_s = l/2c = \tau/2$, and writing $u_m^+(n) = u_m^+(t)$ we get,

$$u_{m+1}^+\left(n - \frac{1}{2}\right) = \frac{1}{1 + r_m}\left[u_m^+(t) - r_m\, u_m^-(t)\right] \qquad 2.13$$

$$u_{m+1}^-\left(n + \frac{1}{2}\right) = \frac{1}{1 + r_m}\left[-r_m u_m^+(t) + u_m^-(t)\right] \qquad 2.14$$

Taking z-transform of the above equations,

$$\begin{bmatrix} U_{m+1}^+(z) \\ U_{m+1}^-(z) \end{bmatrix} = \frac{z^{1/2}}{1 + r_m} \begin{bmatrix} 1 & -r_m \\ -r_m z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} U_m^+(z) \\ U_m^-(z) \end{bmatrix} \qquad 2.15$$

At the front end of the vocal tract ie lips the tube opens into an infinite area ie $r_0 = 1$. The entire wave is reflected from the lips. While at the other end of the vocal tract ie at the glottis the vocal tract is terminated into characteristics impedance hence there is no reflection ie $r_0 = 0$.
Hence equation take the form,

$$\begin{bmatrix} U_{m+1}^+(z) \\ U_{m+1}^-(z) \end{bmatrix} = z^{(m+1)/2} K_m \begin{bmatrix} D_m^+(z) \\ D_m^-(z) \end{bmatrix} \left(U_0^+(z) - U_0^-(z)\right) \qquad 2.16$$

$$\begin{bmatrix} D_m^+(z) \\ D_m^-(z) \end{bmatrix} = \begin{bmatrix} 1 & -r_m \\ -r_m z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} D_{m-1}^+(z) \\ D_{m-1}^-(z) \end{bmatrix} \qquad 2.17$$

$$\begin{bmatrix} D_0^+(z) \\ D_0^-(z) \end{bmatrix} = \begin{bmatrix} 1 \\ -z^{-1} \end{bmatrix} \qquad 2.18$$

K_m is a gain factor given by

$$K_m = \prod_{i=1}^{m} \frac{1}{1 + r_i} \qquad 2.19$$

2.3. WAKITA INVERSE FILTERING

Wakita inverse filtering is used for the estimation of the reflection coefficients (Wakita 1973). It is a modification of the Levinson-Durbin algorithm used for LPC analysis. The speech signal $S(n)$ is the output of an all pole filter with impulse train or the random noise as the input excitation. This assumption is valid for all sounds except for a few fricatives and for nasals where velum is lowered and nasal cavity is coupled.

If p is the order of the all-pole filter then its transfer function is given by,

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} = \frac{S(z)}{U(z)} \qquad 2.20$$

$S(z)$ = z-transform of the speech signal.

$U(Z)$ = z-transform of the input.

In time domain,

$$s(n) = - \sum_{k=1}^{p} a_k s(n-k) + G U(n) \qquad 2.21$$

$$\tilde{s}(n) = - \sum_{k=1}^{p} a_k s(n-k) \qquad 2.22$$

the expected value of the squre error with respect to a,

$$\varepsilon = E [s(n) - \tilde{s}(n)] \qquad 2.23$$

The minimum occurs when,

$$\frac{d\varepsilon}{da_k} = 0 \qquad 2.24$$

This results in the following condition,

$$\sum_{j=1}^{p} a_j E[s(n-k) s(n-j)] = - E[s(n) s(n-k)] \qquad 2.25$$

$$E[s(n-k)\ s(n-j)] = R\ (k-j) = R\ (j-k) \qquad 2.26$$

where $R(i)$ is the autocorrelation function.

$$\begin{bmatrix} R(0) & R(1) & R(p-1) \\ R(1) & R(0) & R(p-2) \\ \vdots & \vdots & \vdots \\ R(p-1) & R(p-2) & \cdots\cdots\ R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \\ a_p \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \qquad 2.27$$

i.e $R.a = r$ The matrix R is symmetric and Toeplitz. The solution of these equations can be obtained using Levinson-Durbin algorithm. The recursive algorithm is as follows,

$$a_0(j+1) = a_0(j) = 1 \qquad 2.28$$

$$k_j = \frac{\sum_{i=0}^{j} a_i(j)\ R\ (j+1-i)}{\sum_{i=0}^{j} a_i(j)\ R\ (i)} \qquad 2.29$$

$$a_i(j+1) = a_i(j) + k_j\ a_{j+1-i}(j) \qquad 2.30$$

$$a_{j+1}(j+1) = k_j \qquad 2.31$$

The indices in the parenthesis indicate the iteration number ($0 \le j \le p-1$) and the subscripts are the filter coefficients numbers. Let,

$$A_j(z) = \sum_{i=0}^{j} a_i(j)\ z^{-i} \qquad 2.32$$

$$B_j(z) = -\sum_{i=0}^{j} a_i(j)\ z^{-(j+1-i)} \qquad 2.33$$

From equations 2.32 and 2.33 wet,

$$\begin{bmatrix} A_{j+1}(z) \\ B_{j+1}(z) \end{bmatrix} = \begin{bmatrix} 1 & -k_j \\ -k_j z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} A_j(z) \\ B_j(z) \end{bmatrix} \qquad 2.34$$

$$\begin{bmatrix} A_o^+(z) \\ B_o^-(z) \end{bmatrix} = \begin{bmatrix} 1 \\ -z^{-1} \end{bmatrix} \qquad 2.35$$

it can be shown by induction that $r_{j+1} = k_j$

The intermediate variables in the Levinson-Durbin algorithm correspond to the boundary conditions $r_o = 1$. The cross-sectional area can be calculated as,

$$A_m = \frac{1 + r_m}{1 - r_m} A_{m+1} \quad \text{where } A_{M+1} = 1 \qquad 2.36$$

2.4. LE-ROUX GUEGUEN ALGORITHM

The algorithm discussed above is not suitable for the fixed point arithmetic implementation as the filter coefficients a_i calculated using this algorithm can be any real number. The DSP processor TMS320C25 does not support floating point arithmetic. For fixed point implementation of this algorithm Le Roux-Gueguen algorithm can be used (Roux et al 1977). This algorithm is a modification of the Levinson Durbin algorithm. It calculates the auxiliary variables $e_i(j)$ such that,

$$e_i(j) = \sum_{m=0}^{i} a_m(j) R(i-m) \qquad 2.37$$

The recursive equation is given by,

$$k_j = - \frac{e_{j+1}(j)}{e_0(j)} \qquad 2.38$$

$$e_i(j+1) = e_i(j) + k_j \; e_{j+1-i}(j) \qquad j-p \le i \le p \qquad 2.39$$

j is the iteration number $(0 \le j \le p)$ The auxiliary variables are initialized as $e_i(0) = R(|\, i \,|) - p \le i \le p$

The auxiliary variables are in the range of $[-R(0), R(0)]$. The algorithm can be implemented using fixed point arithmetic. This is a polynomial type algorithm with the number of multiplications on the order of $(NP + P^2)$ where N is the window length.

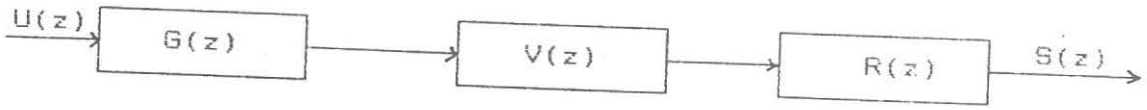


Figure 2.1 Block diagram of speech production system model.

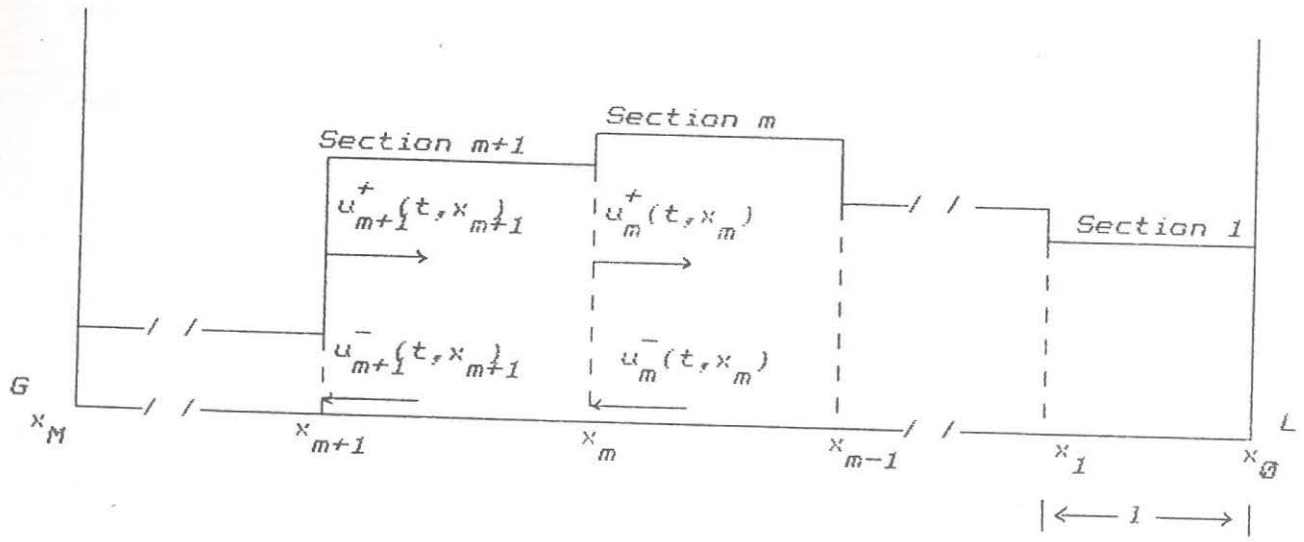


Figure 2.2 Acoustic tube model of the vocal tract.

CHAPTER 3

PITCH ESTIMATION

3.1. INTRODUCTION

An accurate pitch estimation or estimation of fundamental frequency in the speech signal is an important problem in the speech processing. Pitch detectors are used in vocoders, speech identification system, and aids to the handicapped. Because of its importance many solutions have been proposed. All of the proposed systems have their limitations and it can be said that no presently available pitch detection scheme can be expected to give satisfactory results across a wide range of speakers, applications, and operating environments.

There are several reasons which make an accurate and reliable pitch estimation a difficult task (Rabiner 1976) such as,

1. The glottal excitation waveform varies with time and is not a periodic train of impulses.
2. The shape of the glottal waveform is altered significantly by vocal tract filter which introduces the formant structure. The true periodic structure may be masked due to selective amplification of harmonics.
3. The voiced/unvoiced distinction becomes difficult at low levels of voicing.
4. The markers which define the beginning and the end of the pitch period are difficult to identify.
5. Presence of noise can affect the pitch estimation.

3.2. PITCH ESTIMATION METHODS

A number of algorithms are available for estimating the pitch of the speech signal (Gracious 1991, Sapre 1991). These

algorithms can be classified as,

1. Time domain algorithms.
2. Frequency domain algorithms.
3. Hybrid algorithms.

Only time domain algorithms are considered here as the aim of the project is to develop a real time implementation of aid for the deaf. The other two types of algorithms involve the transformation from time to the frequency domain which is computationally very expensive. A true real time performance can not be achieved using these two types of algorithms.

3.2.1. Parallel processing method

This pitch detection scheme has been proposed by Gold and later modified by Gold and Rabiner (Rabiner 1978). It has been used successfully in a wide variety of applications. It is a time domain processing technique.

The pitch signal is processed so as to create a number of impulses which retain the periodicity of the original signal and discard features which are irrelevant to the pitch detection process. The speech waveform is sampled at a rate sufficient to give adequate time resolution. The speech is low pass filtered at 900 Hz to produce a relatively smooth waveform. The peaks and valleys are located in the speech signal and from their location and amplitudes several impulse trains are derived. Each impulse train consists of positive impulses occurring at the location of either the peak or the valley. Impulse train is processed by a time varying nonlinear system. When an impulse of sufficient amplitude is detected at the input, the output is reset to the value of that impulse and then held for a blanking interval, τ (n) during which no pulse can be detected. At the end of the blanking interval the output begins to decay exponentially. When an impulse exceeds the level of the decaying output, the process is repeated.

The rate of decay and the blanking interval depends upon the most recent estimates of pitch period. The pitch period is estimated periodically by measuring the length of the pulse spanning the sampling interval.

This procedure produces very good estimates of the period of the voiced speech. For unvoiced speech distinct lack of consistency is found. When this lack of consistency is detected then the speech is classified as unvoiced.

3.2.2. The short time autocorrelation method

The autocorrelation of a discrete time signal is defined as,

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)\, x(m+k) \qquad 3.1$$

If the signal is periodic with period p samples $\phi(k) = \phi(k+p)$
It is an even function, ie $\phi(k) = \phi(-k)$.
It attains its maximum value at k=Ø, ie $|\phi(k)| \leq \phi(Ø)$ for all k.
The quantity $\phi(Ø)$ is equal to the energy for the deterministic
signals.

The autocorrelation function peaks at the shift corresponding to the time period of the periodic signal as the signal and its shifted version match. With windowed data the peaks of the autocorrelation function are tapering. This helps in avoiding the false detection of secondary peaks at the multiples of the pitch period. The peak of the autocorrelation function are selected and the pitch is estimated (Sondhi 1968).

3.2.3. The short time average magnitude difference method

The computation of autocorrelation function involves considerable arithmetic. A technique that eliminates the need for multiplications is based upon the fact that for a truly periodic

input of period ρ the function

$$d(k) = \sum_{n=-\infty}^{\infty} |x(n) - x(n-k)| \qquad 3.2$$

would be zero for $k=0, \pm \rho, \pm 2\rho \ldots$

Thus for short segments of voiced speech $d(n)$ will be small at the multiples of the period. The AMDF function is implemented with subtraction, addition and absolute value operations, in contrast to addition and multiplication operations for the autocorrelation function (Ross et al 1974). With fixed point arithmetic the AMDF have the advantage. In this case the multiplies often are more time consuming and double precision accumulator is required to hold the sum of the large products. Therefore the AMDF function has been used in many real time speech processing systems.

3.2.4. Linear predictive analysis

The basic discrete time model for the speech production system is as shown in the Fig 3.1. In this case, the composite spectrum effects of radiation, vocal tract, and glottal excitation are represented by time varying digital filter whose steady state response is of the form,

$$H(z) = \frac{S(z)}{U(z)} = \frac{\frac{G}{\rho}}{1 - \sum_{k=1} a_k z^{-k}} \qquad 3.3$$

This system is excited by an impulse train for voiced speech or random noise sequence for unvoiced speech. This simplified all pole model is a natural representation of non-nasal voiced sounds, but for nasals and fricative sounds, the acoustic theory calls for both poles and zeros in the vocal tract transfer function. But if the order P is high enough, the all pole model provides a good representation for almost all sounds of speech. The gain parameter G and the filter coefficients a_k can be

estimated in a very straight forward and computationally efficient manner (Rabiner 1978).

$$s(n) = \sum_{k=1}^{p} a_k \, s(n-k) + G \, u(n) \qquad 3.4$$

A linear predictor with predictor coefficients, a_k is defined as a system with output,

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k \, s(n-k) \qquad 3.5$$

The predictor error, $e(n)$, is defined as

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k \, s(n-k) \qquad 3.6$$

Thus predictor error sequence is the output of a system whose transfer function is

$$A(z) = 1 - \sum_{k=1}^{p} a_k \, z^{-k} \qquad 3.7$$

Thus if the speech signal obeys the model of equation 3.2 exactly, and $a_k = a_k$ then $e(n) = G \, u(n)$. Thus the prediction error filter, $A(z)$ will be an inverse filter for the system, $H(z)$ ie

$$H(z) = \frac{G}{A(z)} \qquad 3.8$$

The basic problem of the linear predictive analysis is to determine the set of predictor coefficients a_k directly from the speech signal in such a manner as to obtain a good estimate of the spectral properties of the speech signal through the use of equation 3.6. Because of the time varying nature of the speech signal the predictor coefficients must be estimated from short segments of the speech signal. The basic approach is to find a set of predictor coefficients that will reduce mean-squared prediction error over a short segment of the speech waveform. The resulting parameters are then assumed to be the parameters of the function $H(z)$ in the model of the speech production.

One of the major limitations of the autocorrelation and AMDF function as that in a sense they retain too much information in the speech signal As a result the autocorrelation function has many peaks and AMDF has many valleys. Most of these peaks and valleys can be attributed to the damped oscillations of the vocal tract response. In the cases when the autocorrelation peaks due to the vocal tract response are bigger than those due to the periodicity of the vocal excitation, the simple procedure of peaking the largest peak in the autocorrelation function will fail. To avoid this problem it is necessary to process the the speech signal so as to make the periodicity more prominent while suppressing other distracting features of the signal. The techniques which perform this type of operations on the speech signal are called "spectrum flatteners" since their objective is to remove the effects of the vocal tract transfer function

In the scheme proposed by Sondhi (Sondhi 1968), the centered clipped speech signal is obtained by a nonlinear transformation,

$$y(n) = C\,[x(n)] \qquad 3.9$$

where C is shown in Fig 3.2 (a). The clipping level C_L is set equal to a fixed percentage of A_{max}, the maximum amplitude. The output of the center clipper is equal to the input minus the clipping level. For the samples below the clipping level the output is zero. For higher clipping levels, fewer peaks will exceed the clipping level and thus fewer pulses will appear at the output, and therefore, fewer extraneous peaks will appear in the autocorrelation function. There is a difficulty with using too high a clipping level. It is possible that amplitude of the speech signal may vary appreciably across the duration of the speech segment so that if the clipping level is set at a higher percentage of the maximum amplitude of the whole segment, there is a possibility that much of the waveform will fall below the

clipping level and be lost. For this reason Sondhi's proposal was to set the clipping level at the 30% of the maximum amplitude.

Another difficulty with the autocorrelation function is the large amount of computation that is required. A simple modification of the center clipping function leads to a great simplification in computation of the autocorrelation function with essentially no degradation in utility for the pitch detection. This modification is shown in the Fig 3.3 (b). The output of the clipper is +1 if $x(n) > C_L$ and -1 if $x(n) < -C_L$. Otherwise the output is zero. This function is called a 3-level center clipper. The computation of the autocorrelation function for a 3-level center clipped signal is simple. The autocorrelation function

$$R_n(k) = \sum_{m=0}^{N-k-1} y(n+m)\, y(n+m+k) \qquad 3.10$$

can have only three different values.

$$\begin{aligned} R_n(k) &= 0 \quad \text{if} \quad y(n+m) = 0 \text{ or } y(n+m+k) = 0 \qquad 3.11 \\ &= +1 \quad \text{if} \quad y(n+m) = y(n+m+k) \\ &= -1 \quad \text{if} \quad y(n+m) = \quad y(n+m+k) \end{aligned}$$

3.3. ENERGY ESTIMATION

The amplitude of the speech signal varies appreciably with time. In particular the amplitude of the unvoiced segments is generally much lower than the amplitude of the voiced segments. The short time energy function is a convenient way of representing the amplitude variations. The short time energy function is defined as,

$$E_n = \sum [x(m)\ w(n-m)]^2 \qquad 3.12$$

where $x(n)$ = speech signal.

$w(n)$ = window function.

If the window function is of smaller duration on the order of the pitch period, energy function will fluctuate rapidly

depending upon the exact variations of the speech signal and will not provide sufficient averaging to produce a smooth energy function. On the other hand if the window function is of longer duration on the order of several pitch periods the short time energy function will change very slowly and will not adequately reflect the changing properties of the speech signal (Rabiner 1978). Thus no single value of N ie window length is entirely satisfactory as pitch period varies from 20 samples for a high pitch female and a child voice up to 250 samples for a low pitch male voice. Therefore a suitable practical choice for N is on the order of 100-200 for 10 kHz sampling rate.

The window function chosen for the short time energy function is Hamming window represented by,

$$h(n) = 0.54 - 0.46 * \cos(2 * \pi * n / (N-1)) \qquad 0 \leq n \leq N-1$$
$$= 0 \qquad \text{otherwise.}$$

3.13

The bandwidth of the Hamming window is twice the bandwidth of the rectangular window. The Hamming window gives much more attenuation outside the pass band than that given by corresponding rectangular window of the same duration. The attenuation of the window is independent of the window duration.

The short time energy function can also be used to locate approximately time at which voiced speech segments become unvoiced and vice versa and for very high quality speech (high signal to noise ratio) the energy function can be used to distinguish speech from silence.

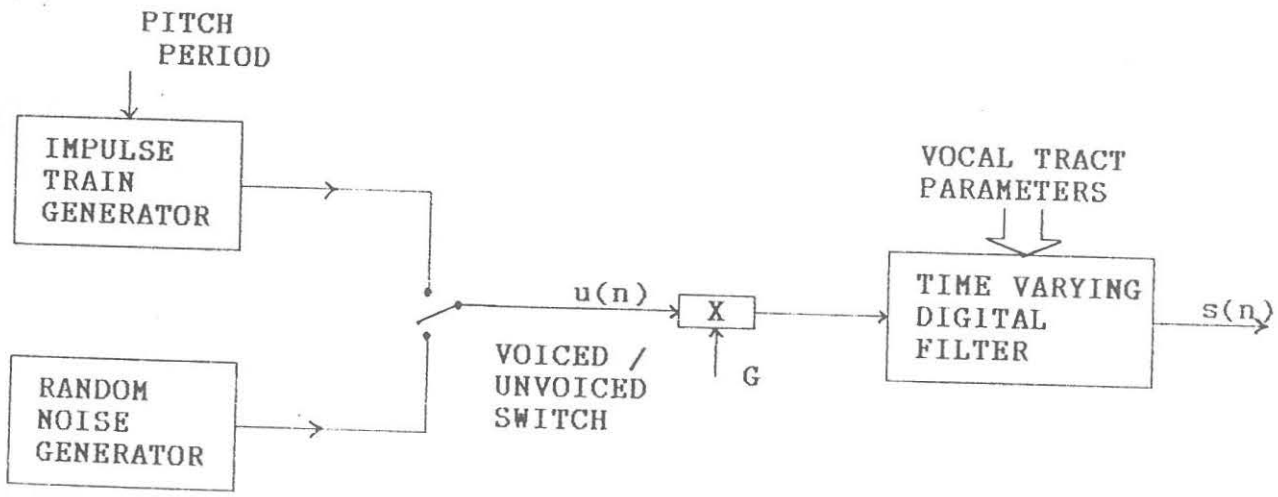
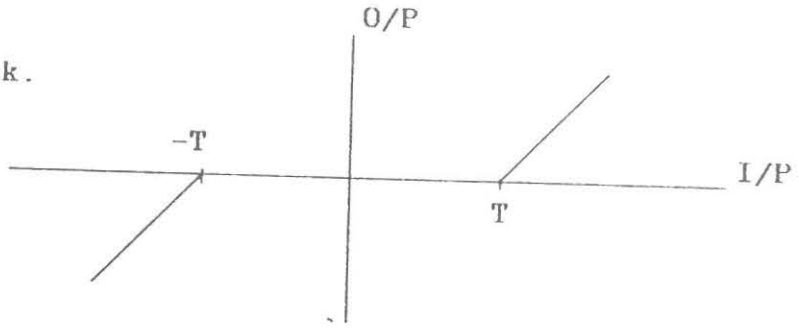
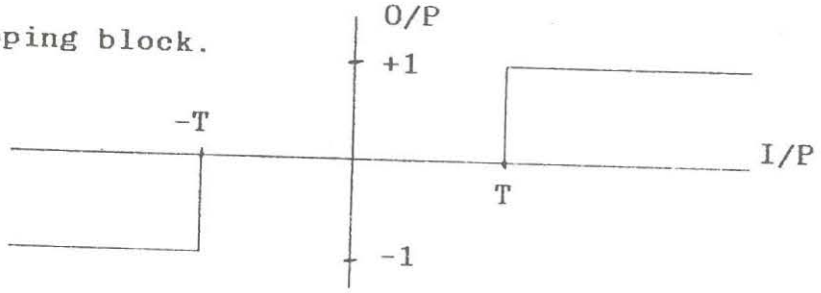


Figure 3.1 Discrete time model for the speech production system.

a) Center clipping block.



b) Infinite center clipping block.



c) Positive center clipping block.

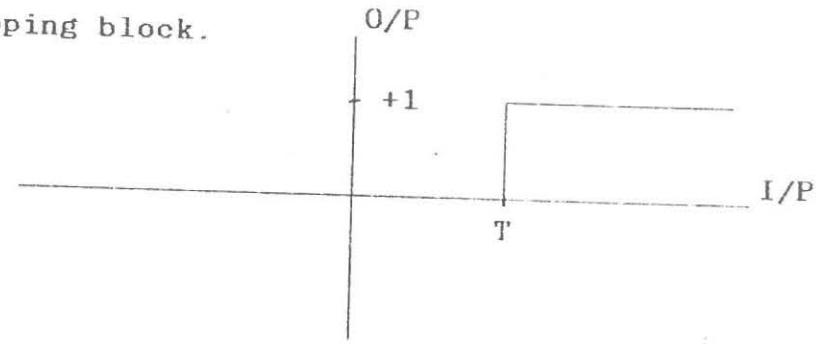


Figure 3.2 Transfer characteristics of various clipping blocks.

CHAPTER 4

ANALOG SIGNAL CONDITIONING.

4.1. INTRODUCTION

As a part of the project a general purpose analog signal conditioning system is built up (Shah 1995). The system consists of two units, one at the input ADC channel of the DSP25 board and the other unit is at the output DAC channel of the DSP board. The overall block diagram of the unit is as shown in the Fig 4.1. The box at the input has a preamplifier and an active elliptic low pass filter. The box at the output has another section of the filter and the output board. The output board has the facility of providing four outputs at different voltage levels for test, tape recorder, microphone, and speaker. The pcb layouts of the preamplifier, filter, and the output board are given in appendix A.

4.2. INPUT PREAMPLIFIER STAGE

The circuit of the input preamplifier stage is as shown in the Fig 4.2. The capacitor coupling at the input of the preamplifier prevents any DC shift that may be present at the output of the microphone. The RC circuit acts as a high pass filter with a cutoff frequency of 16 Hz. Two diodes at the input of the IC protects the IC against excessive voltage at the input. The series resistance limits the current entering the IC input terminals.

The preamplifier circuit provides the facility of two gains for different microphones. This circuit ensures that output of the amplifier is in the range of ±5 volts. The circuit operates on the 12 volts power supply.

4.3. LOW PASS FILTER

All the phonemes and other speech segments are digitized using an ADC which is on the DSP board. These speech segments are sampled at sampling rate of 10 kHz. In order to avoid aliasing, the analog signal must be passed through low pass filter. All the voiced speech segments span the frequency range up to a few hertz while the speech segments in the higher frequency range are unvoiced. Therefore the low pass filter should have a sharp cutoff near 5 kHz. A seventh order active elliptic filter is used for this purpose. The signal frequency components below 4.6 kHz are within 0.3 db of no attenuation at all, while components above 5 k Hz are attenuated by at least 40 db (Pandey 1987). The frequency response of the filter is as shown in the Fig 4.4.

It is a cascade of three biquad sections, each tuned independently. The filter circuit diagram is as shown in the Fig 4.3. For each section the resonant frequency fo, the quality factor q, and the notch frequency fn is as follows,

Section one:

fo = 3.4 kHz q = 1.24 fn = 8.37 kHz

Section two:

fo = 4.35 kHz q = 4.60 fn = 5.57 kHz

Section three:

fo = 4.64 kHz q = 22.8 fn = 5.04 kHz

Each section can be tuned to its fo, q, and fn as follows:

fo = Tune R3 to get a resonance at the bandpass output, node 3 (there should be 180 phase shift between input and output at resonance frequency fo).

q = Tune R1 for unity gain (at node 3) the resonance frequency fo.

fn = Tune R5 to null the node 4 output at the notch frequency fn.

4.4. OUTPUT BOARD

The circuit diagram of the output board is as shown in the Fig 4.5. The output board consists of four modules. It provides four options.

1> ± 5 volts output which can be fedback to the filter input.

2> ± 300 mvolts output option for the recording the signal on the tape recorder.

3> ± 25 mvolts output which can act as an input to the microphone.

4> The output from the power amplifier which can drive the loud speaker.

The output from the DAC is fed to the filter. The filter output is fed to the buffer on the output board. The output of the buffer is fed to all the modules simultaneously. ±5 volts is the direct output of the buffer provided as test signal. Using a suitable gain the inverting amplifier can be used to generate the ± 300 mvolts output for the tape recorder. Similarly the suitable gain of the inverting amplifier can give ± 20 mvolts output for the microphone. A class B push pool amplifier is used to drive the speaker. A complementary symmetry combination of NPN and PNP transistors is used at the output of the opamp. Both the transistors are emitter followers hence have unity gain and provide current boosting. The transistors used are power transistors (HL 100 and HK 100) with 1 ampere rating. Only one transistor conducts at a time. Two emitter resistors eliminate the cross over distortion. Two collector resistors ensure that circuit drives the sufficient current for the speaker. The speaker used is a 0.5 watts 4 ohms speaker.

4.5. PCL DSP-25 BOARD

The block diagram of the DSP-25 board is as shown in the

Fig 4.6 (PCL DSP25). The add-on card DSP board PCL-DSP25 from Dynalog MicroSystems Pvt. Ltd. is suitable for real time implementation of the speech training aid. It is based on TMS320C25 digital signal processor operating at 40 MHz. The DSP card has 64k words of program memory and 64k words of data memory. On board ADC converter has a 35 μs conversion time. The 16 bit programmable timer clocked at 5 MHz is programmed to provide start of conversion pulse to ADC at required sampling interval. The EOC pulse from ADC interrupts the processor for reading the digitized sample. The DSP board fits in the one of the expansion slots of the PC motherboard. The digital signal processor has an instruction cycle of 100 ns for all instructions except for branching and memory exchange instructions. It has 544 words of on chip RAM divided into three blocks. A block of 256 words can be configured either as program memory or data memory. With 32 bit ALU and single machine cycle multiplication it is ideally suited for signal processing applications. The speech signal acquired by the microphone is amplified, filtered, digitized and analyzed by the TMS320C25 to display the vocal tract area function, pitch and energy on the PC screen in real time.

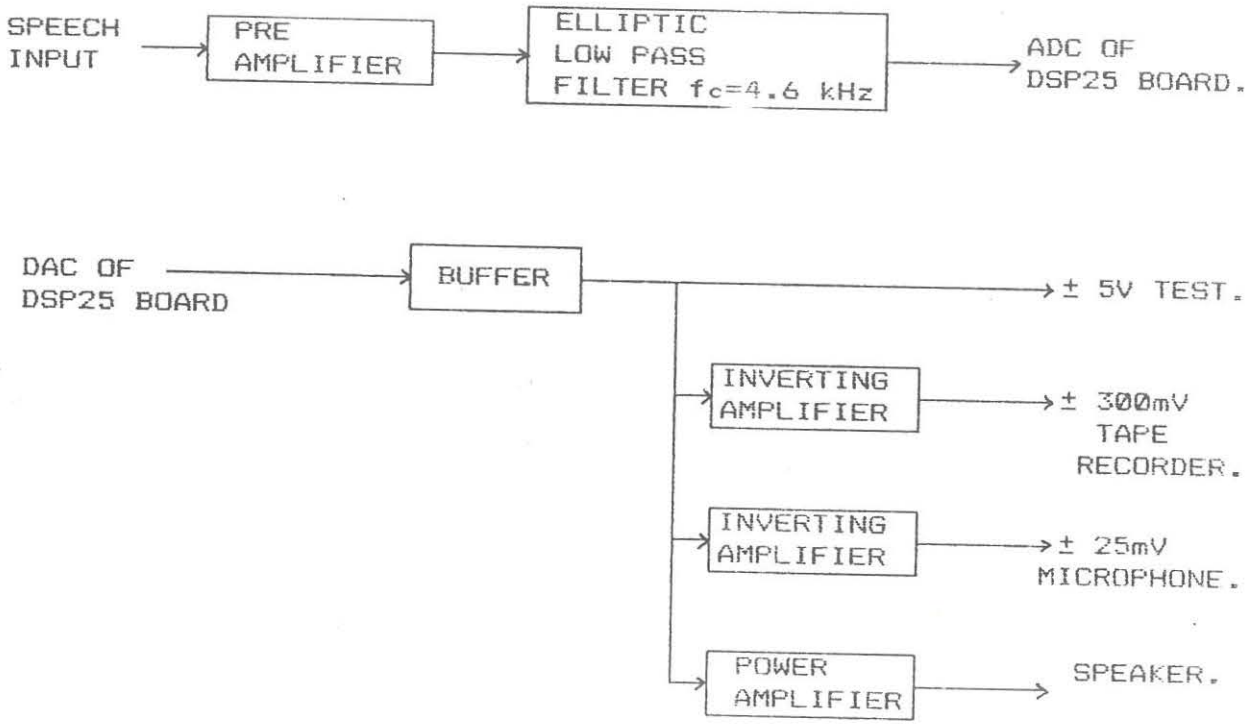


Figure 4.1 Block diagram of the analog signal conditioning unit.

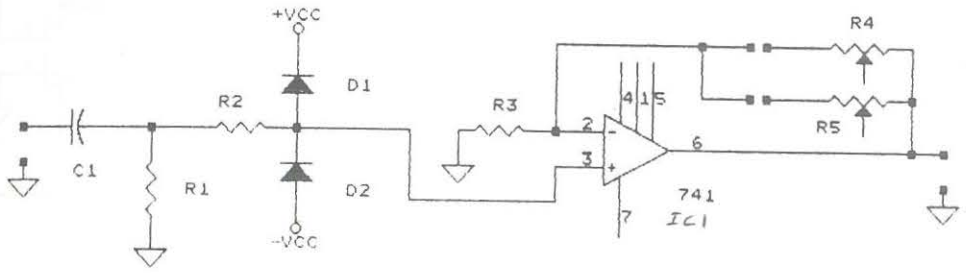


Figure 4.2 The circuit diagram of the input preamplifier

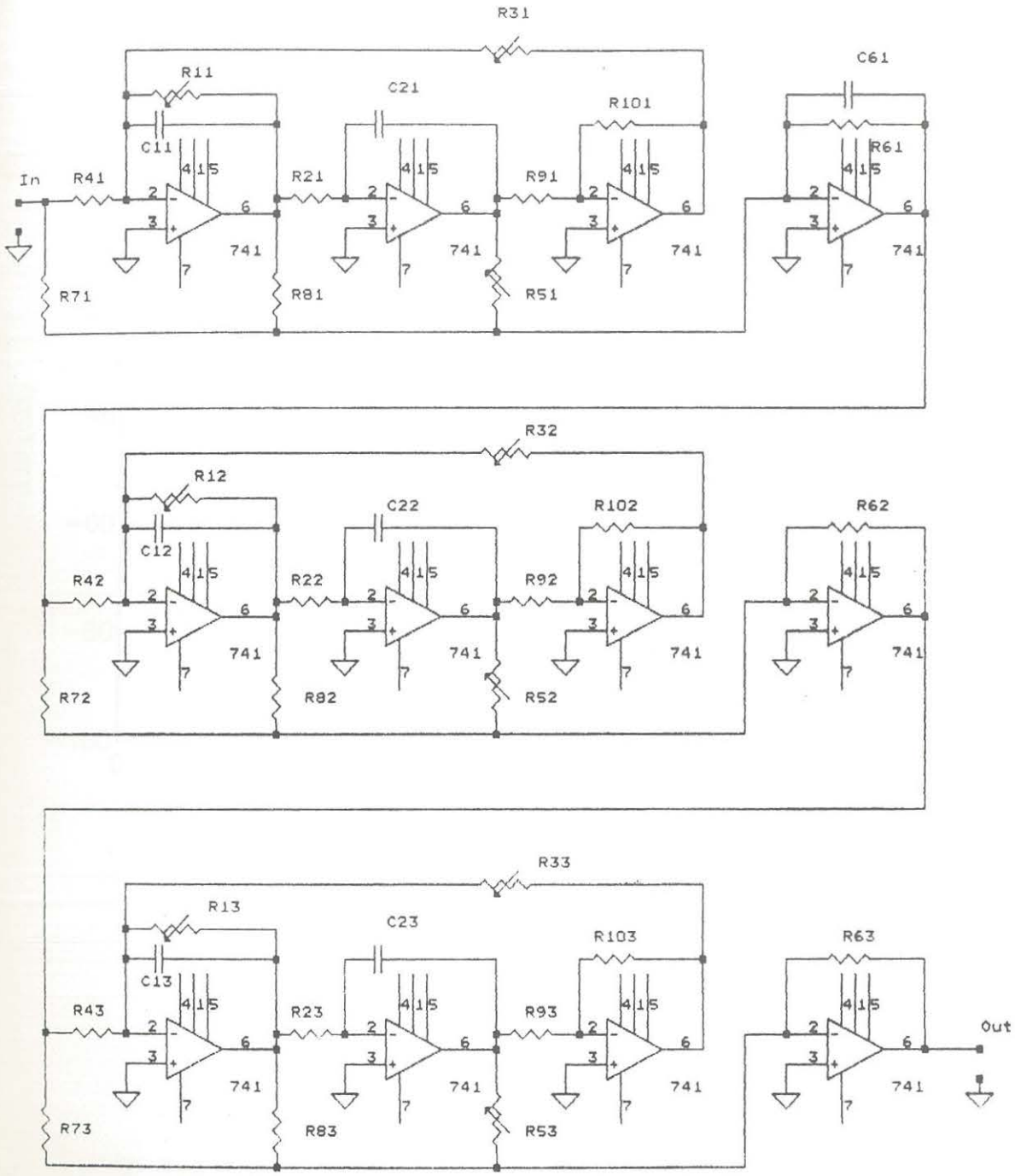


Figure 4.3 Circuit diagram of the 7th order active elliptic lowpass filter

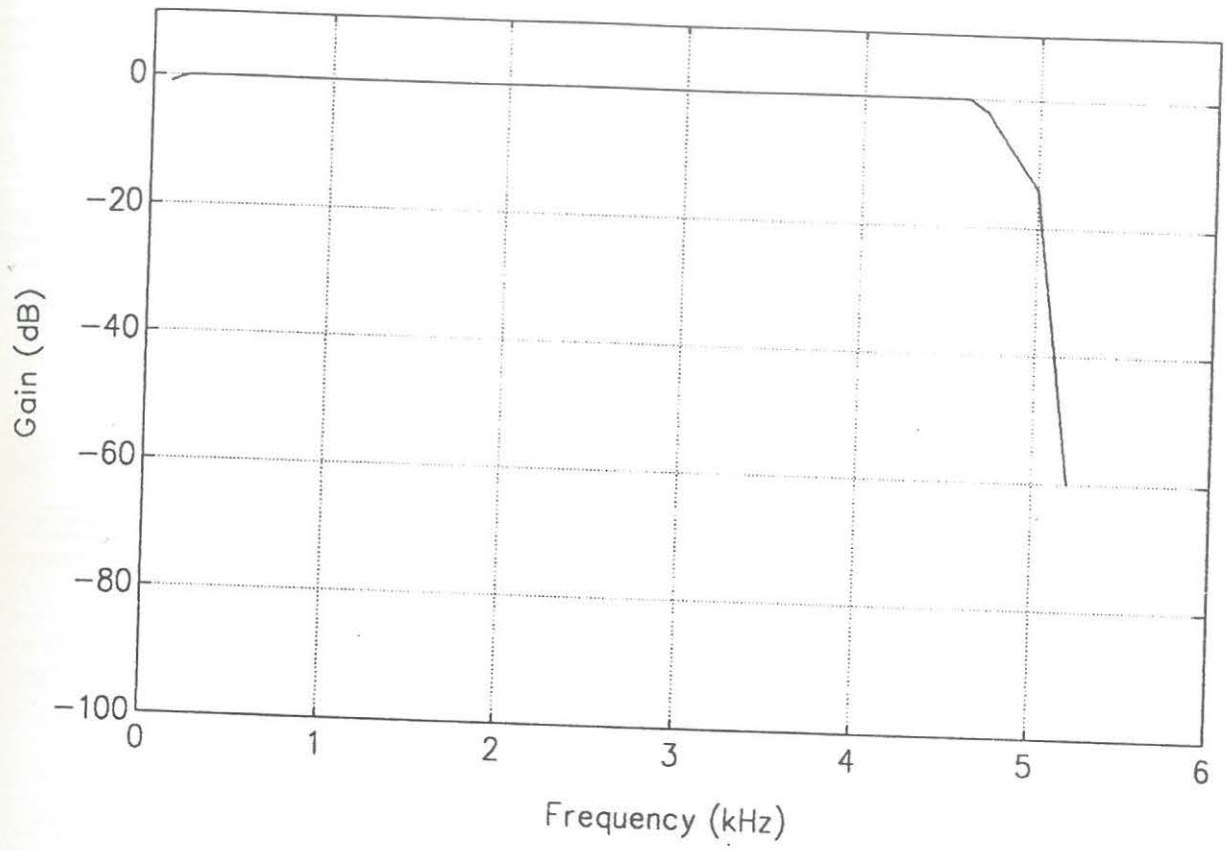


Figure 4.4 The measured magnitude response of the 7th order active elliptic lowpass filter

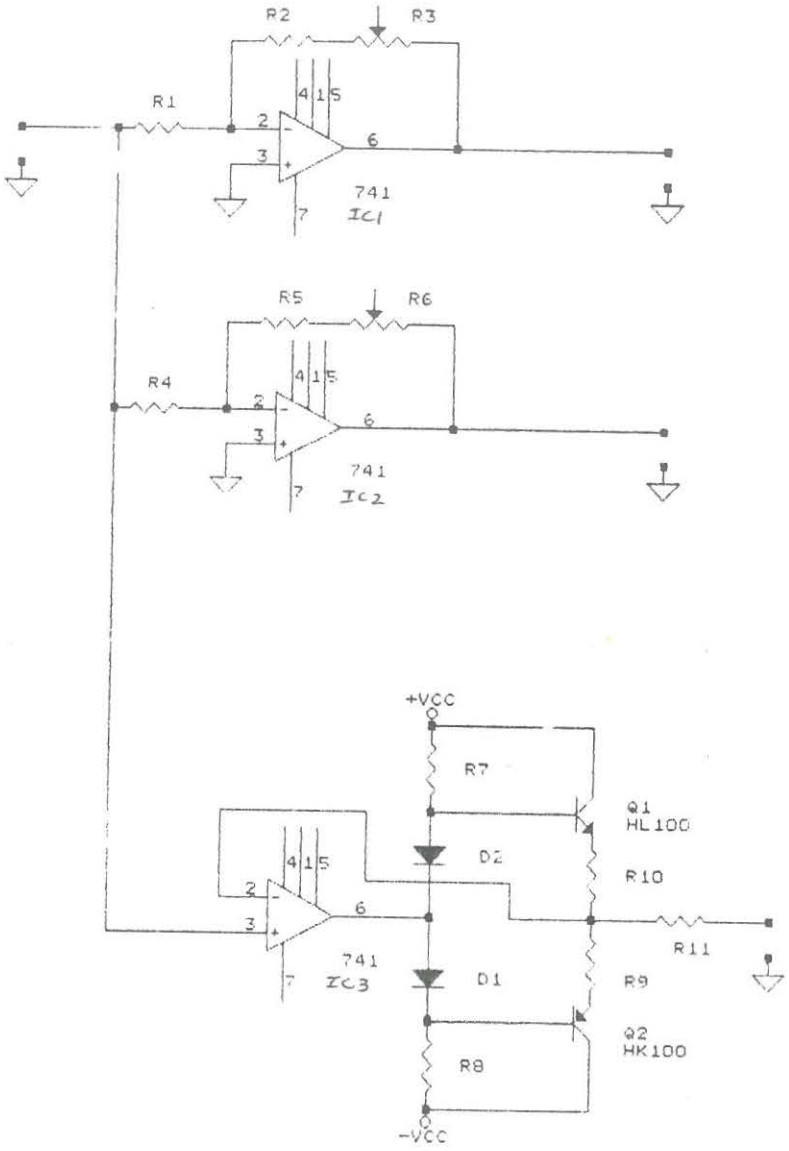


Figure 4.7 The circuit diagram of the output board

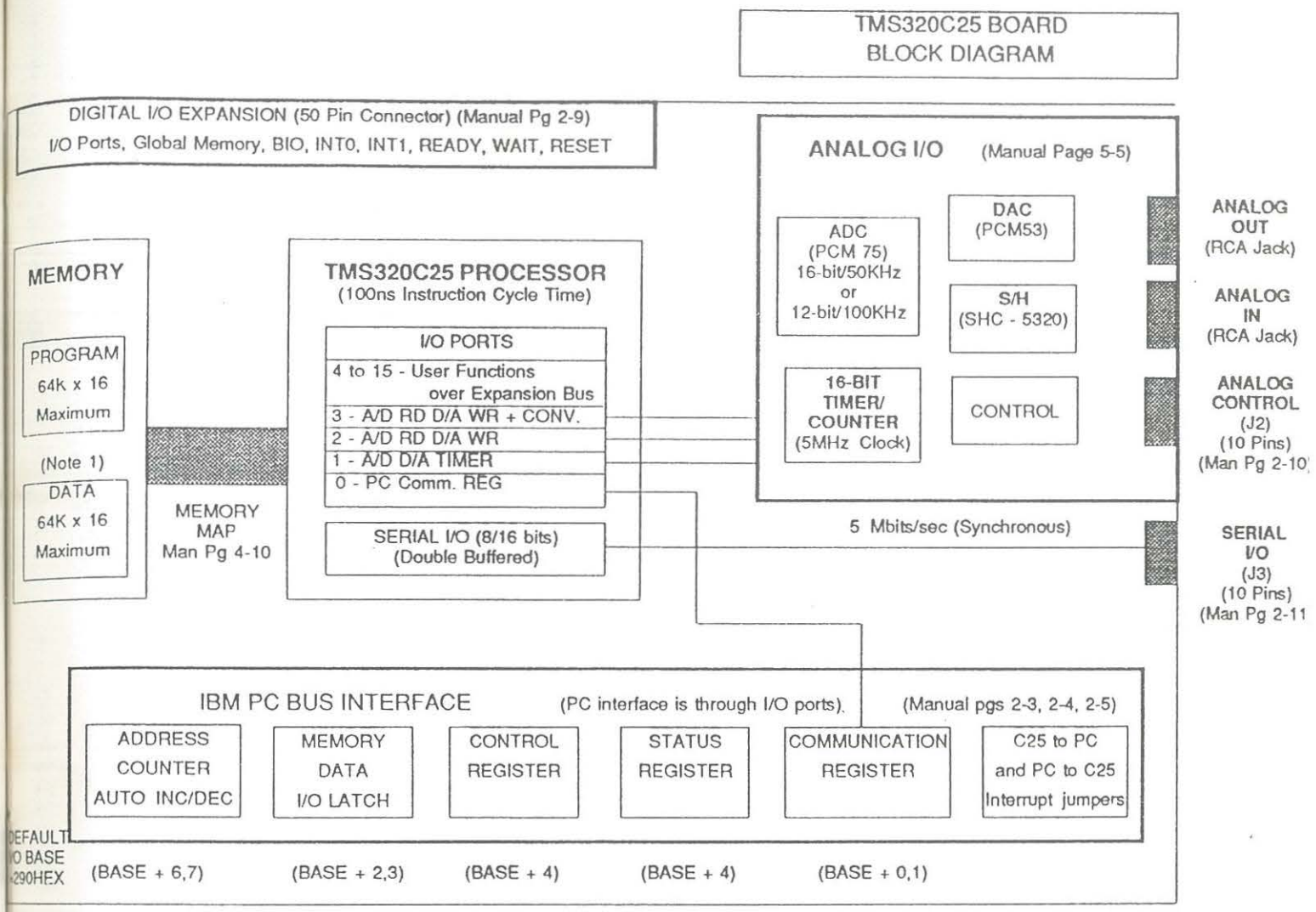


Figure 4.6. Block diagram of PCL DSP25 board

CHAPTER 5

SOFTWARE DEVELOPMENT

5.1. INTRODUCTION

The software developed for the system consists of two modules, one running on the PC and the other on TMS320C25 processor. Both the modules are running simultaneously (PCL DSP25). Speech signal is sampled and digitized by the ADC of the DSP board. The digitized data is processed to obtain the speech parameters namely vocal tract shape, pitch, and energy. The estimated parameters are displayed on the monitor in real time. Facility for storage of analyzed parameters and display them in slow motion review mode also has to be provided.

The analysis of the input speech is carried out by computing autocorrelation coefficients followed by computing the reflection coefficients using Le-Roux Gueguen algorithm, area function is then calculated from these reflection coefficients and is displayed as discussed earlier in the section 2.2.

The pitch is detected from the peaks in the autocorrelation function. The extraneous peaks in the autocorrelation function are removed using center clipping technique. To increase the accuracy in the pitch detection the input speech segment is centered clipped to remove secondary peaks in the autocorrelation function which are the result of the formant frequencies in the vocal tract.

The energy is given by the zeroth autocorrelation coefficient. If the first peak is less than 25% of $R(0)$ or if $R(0)$ is less than certain threshold the frame is declared as unvoiced.

5.2. SOFTWARE ORGANIZATION

The tasks to be carried out are divided in the following steps.

1. Acquisition of the data by sampling the incoming speech signal at a sampling frequency of 10 kHz.
2. Carry out windowing using Hamming window of 256 samples.
3. Calculate the autocorrelation coefficients from the windowed data.
4. Compute the reflection coefficients using Le-Roux Gueguen algorithm.
5. Compute the area function from the reflection coefficients.
6. Estimate the pitch and energy.
7. Display all the three parameters on the PC screen in graphical form.

Out of the tasks specified above the allocation of tasks between PC and the DSP module is very important. As the aim of the project is to display the parameters in real time the time required for executing each task is very important. The TMS320C25 processor works on fixed point arithmetic so the computations involving floating point arithmetic should be preferably assigned to the numeric processor of the PC.

The DSP25 board has on chip ADC. The TMS320C25 processor has in built timer which can be used to sample the ADC at the required frequency. So the task of sampling the input speech segment at 10 kHz frequency is assigned to the processor. In order to avoid the time required for the transfer of data between the processor and the PC, windowing of the data is done on the DSP board. The window coefficients are written to the external data memory of the processor. The autocorrelation coefficients and the reflection coefficients are calculated on the DSP chip.

The Dynalog MicroSystem's PCL DSP-25 board has on board ADC. So sampling of the input speech segment is carried out on the board at a sampling frequency of 10 kHz. The on-board timer can be programmed for different sampling frequencies. The timer is a 16 bit counter, which can be read and loaded as Port-1 of the DSP chip. Writing to Port-1 loads up a register which is reloaded into the counter at the end of each terminal count. Reading from Port-1

reads the current value in the counter. The end of conversion signal from the ADC is used to generate an interrupt to the DSP chip.

The digitized samples are multiplied by Hamming window coefficients for windowing the incoming data. The Window coefficients are written in the external data memory and then transferred to the program memory of the DSP board.

The autocorrelation coefficients, and the reflection coefficients from the autocorrelation coefficients can be calculated using fixed point arithmetic, so this task is assigned to DSP board.

Calculations of area function require floating point arithmetic so this task is assigned to the numeric processor of PC. The reflection coefficients calculated on the DSP board are stored in the shared data memory of the TMS320C25 processor. These values then can be read directly and processed further by PC. This makes the DSP board free for computing pitch and energy.

The pitch estimation algorithm requires computation of autocorrelation coefficients for detecting peaks over the whole window length. Hence this task is assigned to the DSP board.

The energy estimation requires floating point arithmetic. Hence it is calculated on the PC.

Instead of refreshing the whole window the previous image is erased and the new image is put on the screen reducing the time required to display the whole image. All the parameters are written directly to the video RAM to save the time. Using this method the EGA monitor can be used in the high resolution color mode. The image is generated as an array of offsets in the video ram. In the current display 196 bytes are written to the video RAM. When stored image is also displayed on the screen 296 bytes are written to the video RAM.

Thus the allocation of the tasks between the PC and the DSP board is as shown in the Fig 5.1.

5.3. IMPLEMENTATION DETAILS

The software developed for the system has two modules, one running on PC and the other on the DSP board. The DSP chip is a fixed point processor. So the floating point computations can not be assigned to it, instead they are carried out on the PC. The tasks of data acquisition and estimation of parameters is assigned to the DSP chip. The PC module provides the user interface. In order to avoid the communication overhead on the DSP board the image is generated on the PC. Instead of refreshing the whole image the previous image is erased and the new image is put on the screen reducing the time required to display the image.

5.3.1. DSP BOARD MODULE

The program "pra.asm" estimates the reflection coefficients. The input speech segment is sampled at a sampling rate of 10 kHz by on board ADC. 50% overlapping of the window is done to increase the accuracy in estimating the parameters. The strategy used for achieving the block overlap can be well understood from the Fig 5.2.

The sampled data is continuously written to the external data memory of the processor. Incoming data is entered to the locations 4000h and 4100h alteratively. Data memory locations 4000h to 41ffh act as a circular buffer of four sections each of size 128 bytes. Let these sections be called as "A", "B", "C", and "D". Depending upon logic, data from two successive blocks is retrieved to the location 4200h in the data memory for further processing. In the first cycle the data from the adjacent sections "A" and "B", is copied to the location 4200h. Further processing is carried out on the data at these locations. At the end of the computation which takes 10ms another section of the incoming data is available in the block "C". So for the next processing the data from the blocks "B" and "C" is copied to the location 4200h and the process is repeated. Copying of the data to the location 4200h

is carried out using BLKD instruction. This keeps the previous 128 data samples unmodified, that are again used in the next processing cycle. Thus 50% block overlapping of the incoming data is achieved.

The sampled data is windowed by Hamming window coefficients. The size of the window chosen is 256 samples as it is long enough for the autocorrelation function of the windowed data to approximate the signal autocorrelation function. The MAC (multiply and accumulate) instruction of the processor is used to calculate autocorrelation coefficients.

The reflection coefficients, pitch and energy is calculated for every block of 256 samples. The reflection coefficients are written to the shared data memory while pitch and the energy values are outputted through Port-0.

5.3.2. PC MODULE

The PC module "pra.c" provides the user interface and graphical display of the estimated parameters. After estimating the reflection coefficients the DSP chip writes them in the external data memory of the processor. These parameters are them transferred to the PC through "RdBlock" instruction. The area function is estimated on the PC. The zeroth autocorrelation coefficient is transferred to the PC through the Port-0 for each frame. The pitch and the energy is calculated on the PC using the zeroth autocorrelation coefficient. The graphical display of the estimated parameters is achieved by means of writing them as an offset to the video RAM. Various facilities like capturing the images for speech segment, their slow motion display, capturing target waveform and display of the target waveform along with the image generated in real time is also provided.

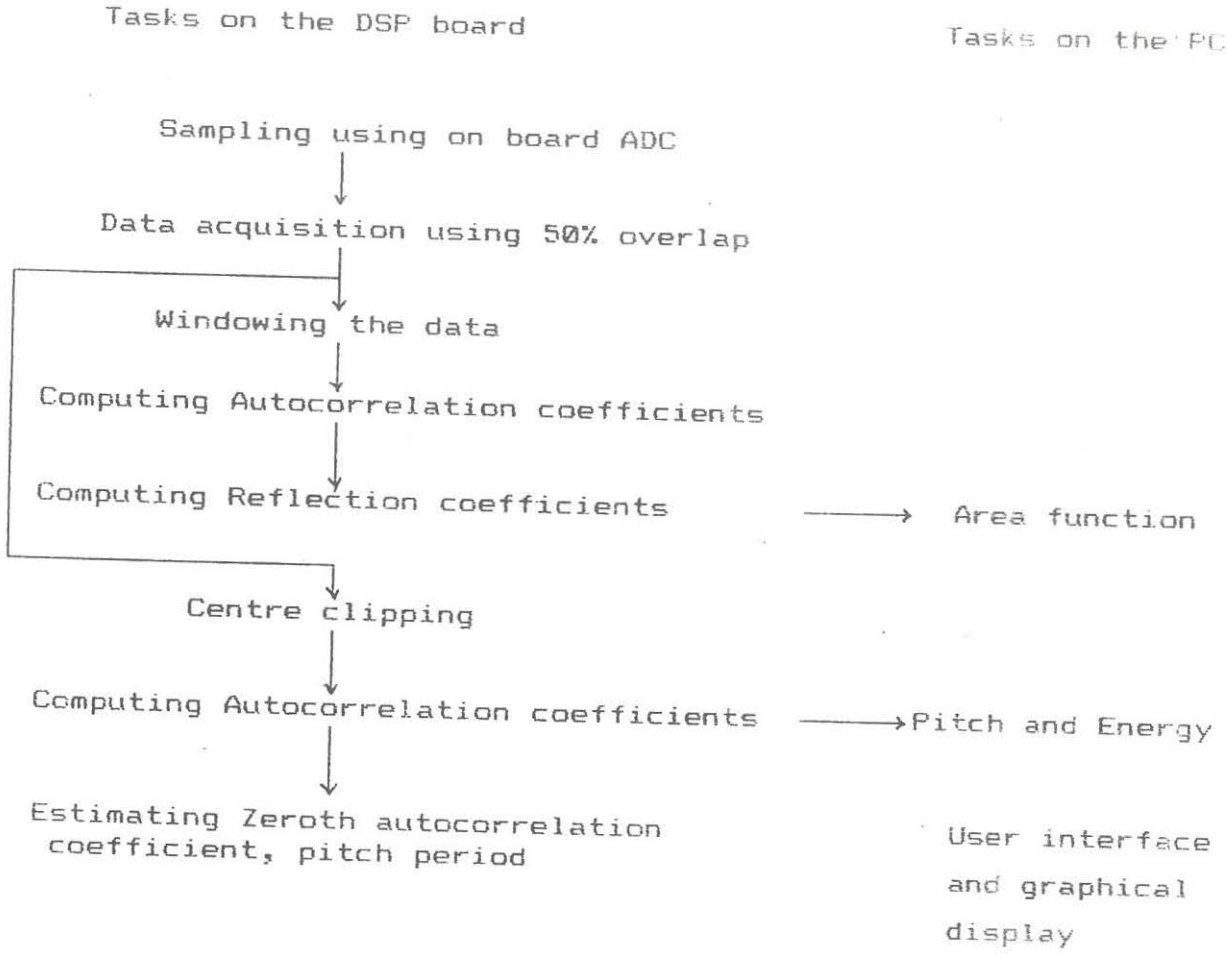


Figure 5.1 The allocation of tasks between PC and DSP board.

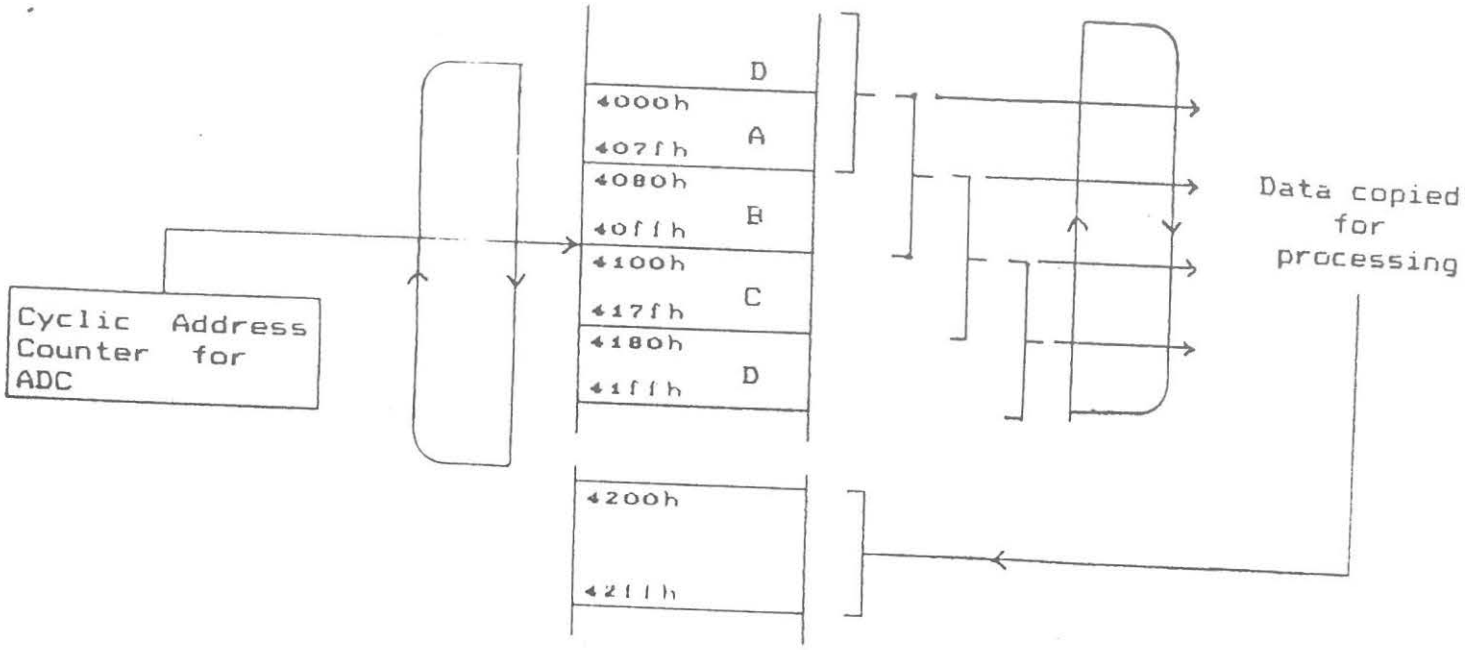


Figure 5.2 Functional diagram of the data acquisition and 50% block overlap

CHAPTER 6

TEST RESULTS

6.1. INTRODUCTION

The software developed for the system is to be tested for its robustness. Testing of the algorithm is carried out by means of generating analog speech files at different energy levels and observing the variations in the vocal tract and pitch. A program called "MAD" developed by Kulkarni is used to generate a speech file at the analog output of the data acquisition card (Kulkarni 1992). The program is modified further to generate the output at five different intensity levels. The attenuation factor offered by different versions of "MAD" program is 1,1/2, 1/4, 1/8, and 1/16 respectively.

The active elliptic filter designed as an analog signal handling unit has the option of three different gain levels. At the input of the filter circuit the inverting amplifier has three different potentiometers. By means of varying the position of the center terminal the gain level can be varied. Thus with the provision of hardware gain and attenuation offered using software it is possible to generate the speech files at various intensity levels.

For testing the algorithm each speech file is fed to one of the versions of the "MAD" program. The gain of the filter is varied from 1 to 4 in steps of 1 unit and the effect of increasing intensity level, on the vocal tract and pitch is carefully observed. The same speech file is then fed to another version of "MAD" program for different scaling factor and the process is repeated. Thus for a single speech file at twenty different intensity levels the vocal tract and the pitch is observed.

The energy level at which the vocal tact shape is not steady is termed as "OK" while the energy level at which the vocal tract shape disappears or a random display is observed in the

window, then it is termed as "BAD". Similarly the pitch window gives steady display of pitch at higher energy levels. As the energy level is reduced the pitch window shows distortion in its continuous display and below a certain range the pitch window is blank. Then it is termed as "0".

6.2. TESTING WITH DIFFERENT SPEECH FILES

The three digitized speech files /a/, /i/, /u/ at four different pitch levels are available. The pitch levels are 100, 125, 200 and 300. At twenty different intensity levels each file is fed and the parameters are observed. The variation in the parameters with respect to the decreasing energy levels is tabulated in the following tables.

A series of tests are carried out on various input data files. The results obtained are as shown in the tables. The energy window is calibrated to measure the voltage at the input of ADC ofthe DSP board.

It can be seen from Tables 6.2 to 6.5 that the pitch level of the input speech file "aa.dat" is increased from 100 to 300. The energy level below which it is not possible to detect the pitch increases from 1.4 to 3.0 vpeak. The energy level below which it is not possible to clearly identify the vocal tract shape increase from 0.6 to 1.4 vpeak..

From the Tables 6.6 to 6.9 it is clear that the pitch level of the input speech file "ee.dat" is increased from 100 to 300. The energy level below which it is not possible to detect the pitch is 2.4 vpeak. The energy level below which it is not possible to clearly identify the vocal tract shape increase from 0.8 to 1.2 vpeak.

The Table 6.10 to 6.13 indicate that the pitch level of the input speech file "oo.dat" is increased from 100 to 300. The energy level below which it is not possible to detect the pitch is 0.6 vpeak. The energy level below which it is not possible to clearly identify the vocal tract shape is 0.6 vpeak.

Table 6.1 The parameter variation with respect to intensity variation for aa100.dat file.
(synthetic vowel with fo = 100 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	2.4	1.2	100 d	good
1	2	5.0	2.5	100	good
1	3	7.0	3.5	100	good
1	4	sat	sat	---	---
2	1	1.2	0.6	0	ok
2	2	2.5	1.25	100 d	good
2	3	3.6	1.8	100	good
2	4	6.0	3.0	100	good
3	1	0.6	0.3	0	bad
3	2	1.2	0.6	0	ok
3	3	2.0	1.0	0	good
3	4	3.0	1.5	100	good
4	1	0.0	0.0	0	bad
4	2	0.6	0.3	0	bad
4	3	1.0	0.5	0	ok
4	4	1.4	0.7	0	good
5	1	0.0	0.0	0	bad
5	2	0.0	0.0	0	bad
5	3	0.4	0.2	0	bad
5	4	0.8	0.4	0	ok

Table 6.2 The parameter variation with respect to intensity variation for aa125.dat file.
(synthetic vowel with fo = 125 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	3.0	1.5	125 d	good
1	2	5.6	2.8	125	good
1	3	8.0	4.0	125	good
1	4	sat	sat	---	---
2	1	1.4	0.7	0	ok
2	2	3.0	1.5	125 d	good
2	3	4.0	2.0	125	good
2	4	6.4	3.2	125	good
3	1	0.8	0.4	0	bad
3	2	1.4	0.7	0	ok
3	3	2.0	1.0	0	ok
3	4	3.2	1.6	125	good
4	1	0.4	0.2	0	bad
4	2	0.6	0.3	0	bad
4	3	1.0	0.5	0	bad
4	4	1.6	0.8	0	ok
5	1	0.0	0.0	0	bad
5	2	0.0	0.0	0	bad
5	3	0.4	0.2	0	bad
5	4	0.8	0.4	0	bad

Table 6.3 The parameter variation with respect to intensity variation for aa200.dat file.
(synthetic vowel with fo = 200 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	4.0	2.0	200	good
1	2	8.4	4.2	200	good
1	3	sat	sat	---	---
1	4	sat	sat	---	---
2	1	2.0	1.0	0	good
2	2	4.0	2.0	200	good
2	3	6.0	3.0	200	good
2	4	1.0	5.0	200	good
3	1	1.0	0.5	0	ok
3	2	2.0	1.0	0	good
3	3	3.0	1.5	200	good
3	4	5.0	2.5	200	good
4	1	0.4	0.2	0	bad
4	2	1.0	0.5	0	bad
4	3	1.4	0.7	0	good
4	4	2.4	1.2	0	good
5	1	0.0	0.0	0	bad
5	2	0.4	0.2	0	bad
5	3	0.8	0.4	0	bad
5	4	1.2	0.6	0	ok

Table 6.4 The parameter variation with respect to intensity variation for aa300.dat file.
(synthetic vowel with fo = 300 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	6.0	3.0	300	good
1	2	12.0	6.0	300	good
1	3	sat	sat	---	---
1	4	sat	sat	---	---
2	1	3.0	1.5	0	good
2	2	6.0	3.0	300	good
2	3	9.0	4.5	300	good
2	4	sat	sat	---	---
3	1	1.4	0.7	0	ok
3	2	3.0	1.5	0	good
3	3	4.8	2.4	300	good
3	4	6.0	3.0	300	good
4	1	0.6	0.3	0	bad
4	2	1.4	0.7	0	bad
4	3	2.4	1.2	0	good
4	4	3.0	1.5	0	good
5	1	0.2	0.1	0	bad
5	2	0.8	0.4	0	bad
5	3	1.2	0.6	0	ok
5	4	2.0	1.0	0	ok

Table 6.5 The parameter variation with respect to intensity variation for ee100.dat file.
(synthetic vowel with fo = 100 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	3.0	1.5	100	good
1	2	6.8	3.4	100	good
1	3	9.0	4.5	100	good
1	4	9.6	4.8	100	good
2	1	1.4	0.7	0	good
2	2	3.4	1.7	100	good
2	3	4.6	2.3	100	good
2	4	6.0	3.0	100	good
3	1	0.8	0.4	0	bad
3	2	1.6	0.8	0	good
3	3	2.4	1.2	0	good
3	4	3.0	1.5	100	good
4	1	0.4	0.2	0	bad
4	2	0.8	0.4	0	bad
4	3	1.2	0.6	0	ok
4	4	1.4	0.7	0	ok
5	1	0.0	0.0	0	bad
5	2	0.4	0.2	0	bad
5	3	0.6	0.3	0	bad
5	4	0.8	0.4	0	bad

Table 6.6 The parameter variation with respect to intensit variation for ee125.dat file.
(synthetic vowel with fo = 125 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	3.4	1.7	125	good
1	2	7.0	3.5	125	good
1	3	9.6	4.8	125	good
1	4	10.0	5.0	125	good
2	1	1.6	0.8	0	good
2	2	3.6	1.8	125 d	good
2	3	5.0	2.5	125	good
2	4	6.8	3.4	125	good
3	1	0.8	0.4	0	bad
3	2	1.8	0.9	0	ok
3	3	2.4	1.2	0	good
3	4	3.4	1.7	125	good
4	1	0.4	0.2	0	bad
4	2	0.8	0.4	0	bad
4	3	1.2	0.6	0	ok
4	4	1.8	0.9	0	good
5	1	0.0	0.0	0	bad
5	2	0.4	0.2	0	bad
5	3	0.6	0.3	0	bad
5	4	1.0	0.5	0	bad

Table 6.7 The parameter variation with respect to intensity variation for ee200.dat file.
(synthetic vowel with fo = 200 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	4.0	2.0	200	good
1	2	9.0	4.5	200	good
1	3	12.8	6.4	200	good
1	4	sat	sat	---	---
2	1	2.0	1.0	0	good
2	2	4.4	2.2	200	good
2	3	6.4	3.2	200	good
2	4	8.4	4.2	200	good
3	1	1.0	0.5	0	ok
3	2	2.4	1.2	0	good
3	3	3.2	1.6	200 d	good
3	4	4.2	2.1	200	good
4	1	0.6	0.3	0	bad
4	2	1.2	0.6	0	bad
4	3	1.6	0.8	0	ok
4	4	2.0	1.0	0	good
5	1	0.0	0.0	0	bad
5	2	0.6	0.3	0	bad
5	3	0.8	0.4	0	bad
5	4	1.0	0.5	0	bad

Table 6.8 The parameter variation with respect to intensity variation for ee300.dat file.
(synthetic vowel with fo = 300 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	4.4	2.2	300	good
1	2	9.6	4.8	300	good
1	3	14.0	7.0	300	good
1	4	sat	sat	---	---
2	1	2.2	1.1	300 d	good
2	2	4.8	2.4	300	good
2	3	7.0	3.5	300	good
2	4	9.0	4.5	300	good
3	1	1.2	0.6	0	ok
3	2	2.4	1.2	300	good
3	3	3.4	1.7	300	good
3	4	4.4	2.2	300	good
4	1	0.6	0.3	0	bad
4	2	1.2	0.6	0	good
4	3	1.6	0.8	0	good
4	4	2.4	1.2	300	good
5	1	0.0	0.0	0	bad
5	2	0.6	0.3	0	bad
5	3	0.8	0.4	0	ok
5	4	1.2	0.6	0	good

Table 6.9 The parameter variation with respect to intensity variation for oo100.dat file.
(synthetic vowel with fo = 100 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	1.2	0.6	100	ok
1	2	2.4	1.2	100	good
1	3	3.6	1.8	100	good
1	4	sat	sat	---	---
2	1	0.6	0.3	0	bad
2	2	1.2	0.6	100	ok
2	3	2.0	1.0	100	good
2	4	2.4	1.2	100	good
3	1	0.0	0.0	0	bad
3	2	0.6	0.3	100 d	bad
3	3	1.0	0.5	100	ok
3	4	1.2	0.6	100	good
4	1	0.0	0.0	0	bad
4	2	0.0	0.0	0	bad
4	3	0.4	0.2	0	bad
4	4	0.6	0.3	0	ok
5	1	0.0	0.0	0	bad
5	2	0.0	0.0	0	bad
5	3	0.0	0.0	0	bad
5	4	0.0	0.0	0	bad

Table 6.10 The parameter variation with respect to intensity variation for oo125.dat file.
(synthetic vowel with fo = 125 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	1.2	0.6	125	ok
1	2	3.0	1.5	125	good
1	3	4.0	2.0	125	good
1	4	sat	sat	---	---
2	1	0.6	0.3	125 d	bad
2	2	1.4	0.7	125	ok
2	3	2.0	1.0	125	good
2	4	2.6	1.3	125	good
3	1	0.0	0.0	0	bad
3	2	0.6	0.3	0	bad
3	3	1.0	0.5	125 d	ok
3	4	1.4	0.7	125	good
4	1	0.0	0.0	0	bad
4	2	0.0	0.0	0	bad
4	3	0.4	0.2	0	bad
4	4	0.8	0.4	125 d	ok
5	1	0.0	0.0	0	bad
5	2	0.0	0.0	0	bad
5	3	0.0	0.0	0	bad
5	4	0.4	0.2	0	bad

Table 6.11 The parameter variation with respect to intensity variation for oo200.dat file.
(synthetic vowel with fo = 200 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	1.6	0.8	200	good
1	2	3.2	1.6	200	good
1	3	4.0	2.0	200	good
1	4	sat	sat	---	---
2	1	0.8	0.4	200 d	ok
2	2	1.6	0.8	200	good
2	3	2.0	1.0	200	good
2	4	2.8	1.4	200	good
3	1	0.4	0.2	0	bad
3	2	0.8	0.4	200 d	ok
3	3	1.0	0.5	200	ok
3	4	1.4	0.7	200	good
4	1	0.0	0.0	0	bad
4	2	0.4	0.2	0	bad
4	3	0.6	0.3	200 d	bad
4	4	0.8	0.4	200 d	ok
5	1	0.0	0.0	0	bad
5	2	0.0	0.0	0	bad
5	3	0.0	0.0	0	bad
5	4	0.0	0.0	0	bad

Table 6.12 The parameter variation with respect to intensity variation for oo300.dat file.
(synthetic vowel with fo = 300 Hz)

DAC scaling Factor	Filter Gain	I/P ADC DSP board	Energy Est	Pitch Est	Vocaltract Shape
1	1	3.0	1.5	300	good
1	2	6.8	3.4	300	good
1	3	9.0	4.5	300	good
1	4	sat	sat	---	---
2	1	1.6	0.8	300 d	good
2	2	3.2	1.6	300	good
2	3	4.4	2.2	300	good
2	4	6.0	3.0	300	good
3	1	0.8	0.4	0	good
3	2	1.6	0.8	300	good
3	3	2.4	1.2	300	good
3	4	3.0	1.5	300	good
4	1	0.4	0.2	0	bad
4	2	0.8	0.4	0	good
4	3	1.2	0.6	0	good
4	4	1.4	0.7	300 d	good
5	1	0.0	0.0	0	bad
5	2	0.4	0.2	0	good
5	3	0.6	0.3	0	ok
5	4	0.6	0.3	0	ok

SUMMARY 7

SUMMARY AND SUGGESTIONS FOR THE FURTHER WORK

7.1. SUMMARY

The aim of the project is to develop a real time speech training aid for the deaf. The system should provide the facility for capturing the images of the speech segment, and slow motion display. This system can be used for training the prelingually deaf person by helping them to learn the movements of the articulators while speaking.

This is in continuation of the on-going efforts at I.I.T. Bombay for developing such a training aid. Software for such a system was already developed but the system is not tested for its robustness.

A general purpose analog signal handling system is to be built fore handling the speech signals at the input and at the output of the DSP board. The software takes the input speech segments on block by block basis each of size 256 samples. To improve the accuracy the block overlap technique is to be implemented. The software is to be tested for different speech files at different intensity levels. The speech parameters are not reliable in case of stop consonants so algorithm is to be modified accordingly. The vocal tract shape can be made more realistic using interpolation technique between discrete area values.

As a first step of the project a general purpose signal conditioning system is designed. Two units, one for handling the signal at the input of the DSP board and the other for handling the signal at the output of the DSP board are designed. The input unit consists of preamplifier and a seventh order active elliptic low pass filter with a cut-off frequency of 4.6kHz. The output unit consists of another seventh order active elliptic lowpass filter and output board which provides the facility for four

different outputs.

The project being the continuation of the on going efforts, the existing system is studied. In the system the analog speech signal is sampled and windowed. The windowed data is used to compute autocorrelation coefficients. The reflection coefficients are then calculated from the autocorrelation coefficients using Le-Roux Gueguen algorithm. The area function, pitch and energy is then calculated using floating point arithmetic. The incoming data is processed using overlap window technique with 50% overlap. The system is tested for various digitized speech files at different energy levels.

7.2. SUGGESTIONS FOR THE FURTHER WORK

The system is tested for various synthesized data files for different energy levels and the results are found to be very consistent. The results may not be accurate in the presence of noise. To make the system more robust the parameter estimation in the presence of noise is to be observed.

It is not possible to correctly estimate the area function in case of the stop consonants due to very low energy level. So the algorithm can be modified to display either previous frame or the next frame.

Presently the system displays the vocal tract shape as sections of cylindrical cascaded tube. The display can be further modified by interpolating between discrete area values.

Once such a system is built and checked for its performance a stand-alone unit can be designed using a more powerful signal processor TMS320C30. The processor should have floating point capability for fast calculation of area function from the reflection coefficients.

REFERENCES

[1] Bernstein L. (1986), James B., "Speech training aids for the profoundly deaf children", *Proc ICASSP*, pp. 633.

[2] Bhagwat Y. (1993), "*Real Time Vocal Tract Shape and Pitch Estimation* ", M.Tech. Dissertation, Dept. Elec. Eng., I.I.T., Bombay.

[3] Bristow G., Brooks S.,"Design and assessment of computer based speech training aids using vocal tract display"(source not available).

[4] Bolt R.(1969). "Speaker Identification by Speech Spectrograms", *Science*, 166, pp.338-343.

[5] Gracias S. (1991). "*A speech training aid for the hearing impaired* ", B.Tech Project Report, Dept. Elec. Eng., I.I.T., Bombay.

[6] Gupte M. (1990). "*A Speech Processor and Display for The Speech Training of The Hearing Impaired* ", M.Tech. Dissertation, Dept. Elec. Eng., I.I.T. Bombay.

[7] Khambete N. (1992)."*A Speech Training Aid for The Deaf*", M.Tech. Dissertation, School of Biomed. Eng., I.I.T., Bombay.

[8] Kulkarni D. (1992). Development of a cascade pole-zero synthesizer. M.Tech. Dissertation, Dept. Elec. Eng., I.I.T., Bombay.

[9] Levitt H., Picket J., and Hounde R. (1980). *Sensory Aids for the Hearing Impaired*, New York: IEEE Press.

[10] Nickerson R. (1980). "Characteristics of the speech of the

deaf person", in *Sensory Aids for The Hearing Impaired*, Levitt and Picket, Eds. New York: IEEE Press.

[11] Padro J. (1982). "Vocal tract analysis for the children", *Proc. ICASSP*, pp 763-766.

[12] Pandey P.(1987), *"Speech Processing for Cochlear Prosthesis"*, Ph. D. Thesis, Dept. Elec. Eng., University of Toronto, Canada.

[13] PCL-DSP25, *User's Manual, Bombay : Dynalog Microsystems.*

[14] Rabiner L., Cheng M., Rosenberg A. and McGonagal C. (1976), "A comparative performance of several pitch detection algorithms", *IEEE Trans. Acoust.,Speech and signal Process.* Vol.22, pp.353-362.

[15] Rabiner L. and Schafer R. (1978), *Digital Processing of Speech Signal*, New Jersey : Prentice Hall.

[16] Ross M., Shaffer H, Cohen A. and Maley H.(1974), "Average magnitude difference function pitch extractor", *IEEE Trans. Acoust., Speech and signal process.* vol.22, pp.353-362.

[17] Roux L.and Gueguen C. (1977), "A fixed point computation of PARCOR coefficients", *IEEE Trans. Acoust., Speech and signal Process.*, vol.27, pp. 257-259.

[18] Sapre R. (1991). *"A Speech Processor for Single Channel Auditory Prosthesis"*, M. Tech Dissertation, Dept. Elec. Eng., I.I.T., Bombay.

[19] Shah. N. (1995). *"A Single Channel Sensory Aid for The Deaf"*, M. Tech Dissertation, Dept. Elec. Eng., I.I.T., Bombay (to be submitted).

[20] Shigenaga M. and Kubo H.(1986). "Speech training systems for

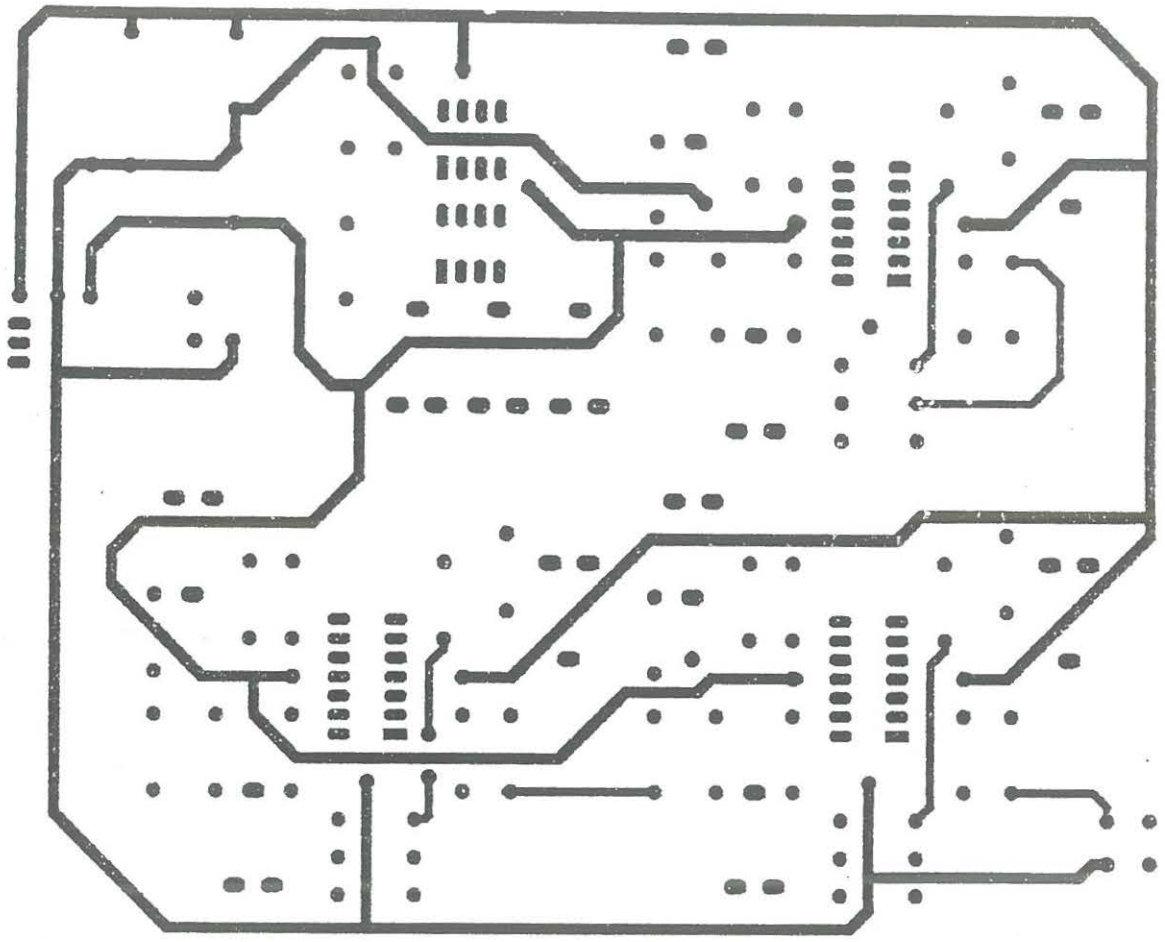
the handicapped children vocal tract lateral shape", *Proc. ICASSP*, pp 637.

[21] Sondhi M.(1968). "New methods of pitch extraction", *IEEE Trans. Audio Electroacoust.*,vol 16, pp 262-266.

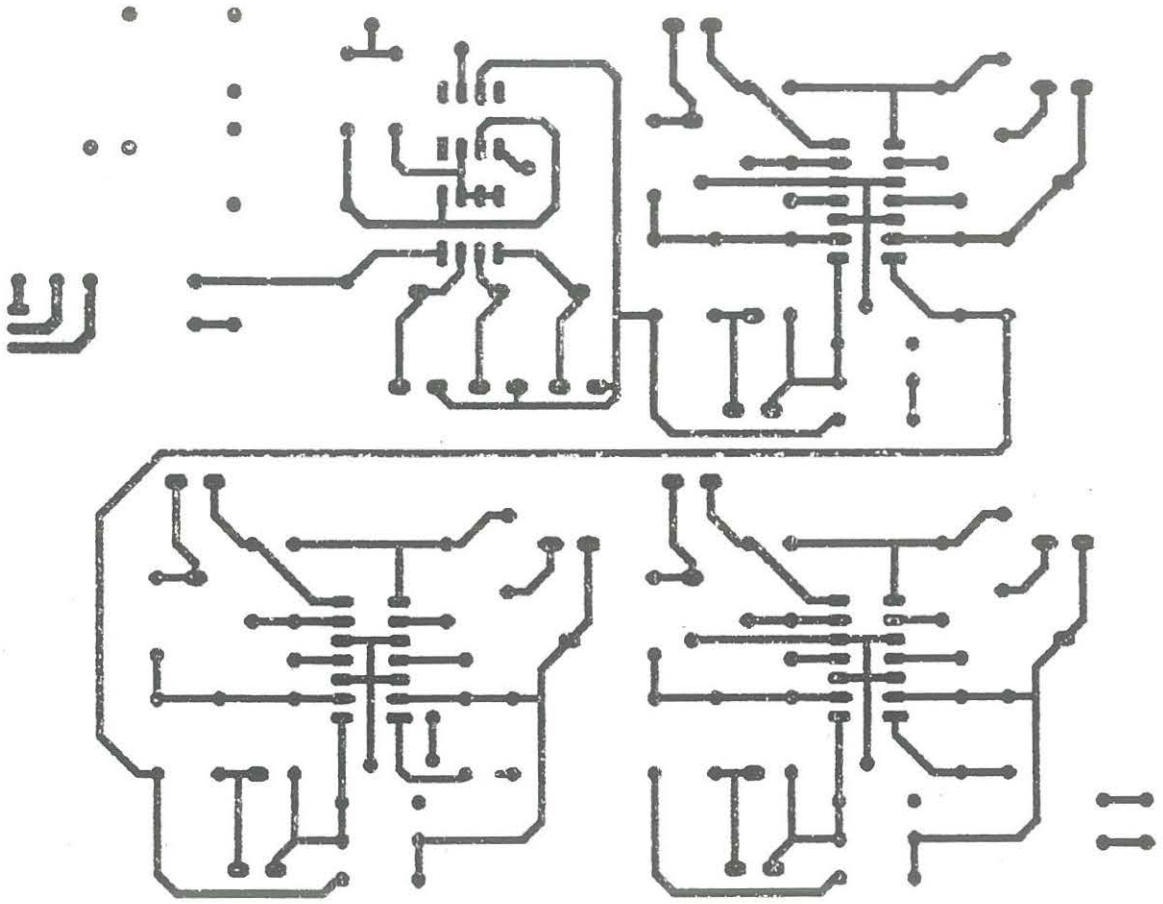
[22] Taklikar K. (1990). "*A Speech Training Aid for The Deaf*", M.Tech. Dissertation, Dept. Elec. Eng., I.I.T. Bombay.

[23] Wakita H. (1973). "Direct estimation of the vocal tract shape by inverse filtering of the acoustic speech waveform", *IEEE Trans. Audio Electroacoust.*,vol 21, pp. 417-426.

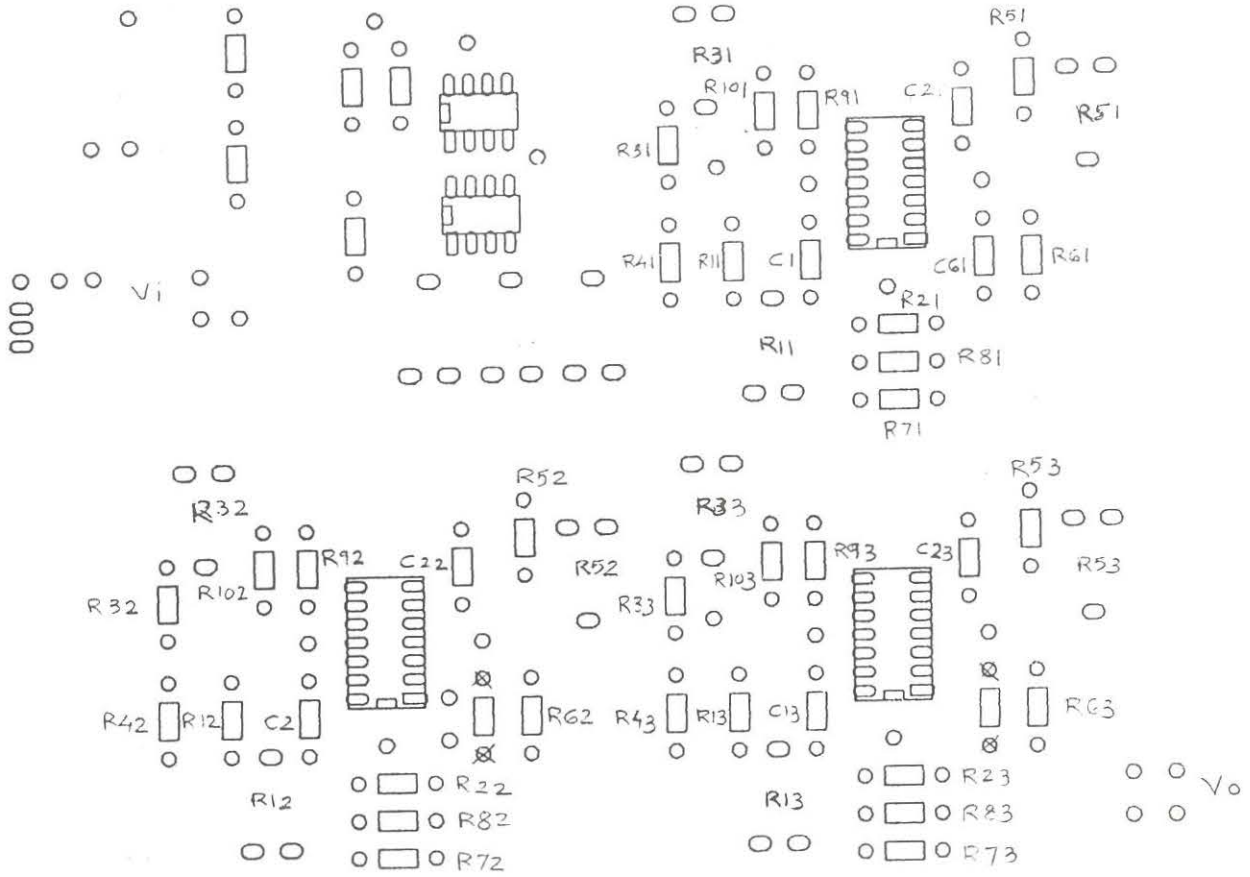
APPENDIX



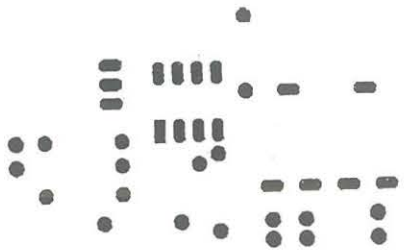
Component side of the 7th order active elliptic lowpass filter



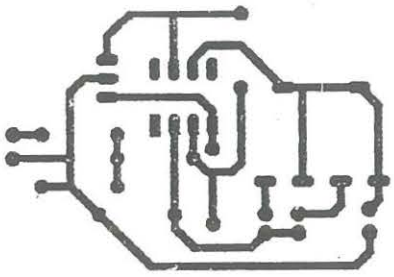
Solder side of the 7th order active elliptic lowpass filter.



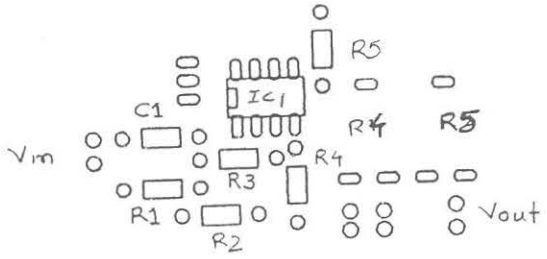
Component placement on the filter board [Figure 4·3]



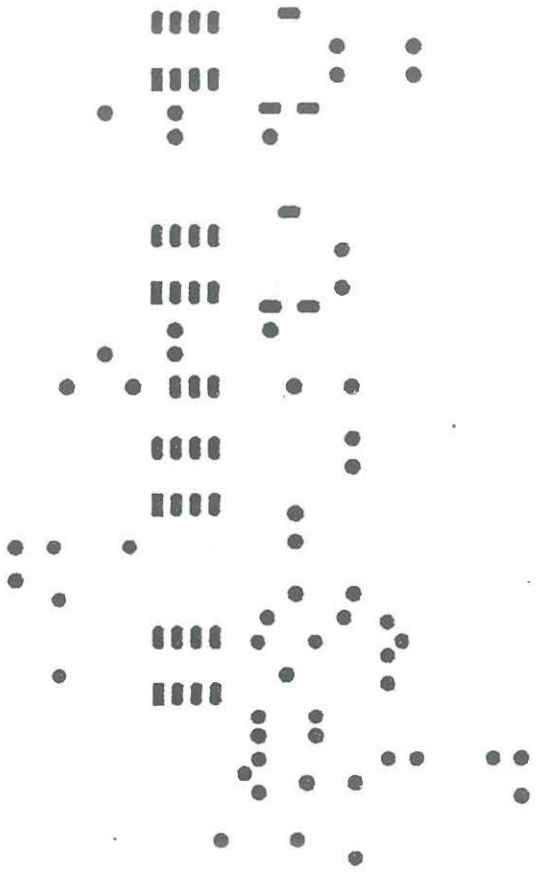
Component side of the input preamplifier.



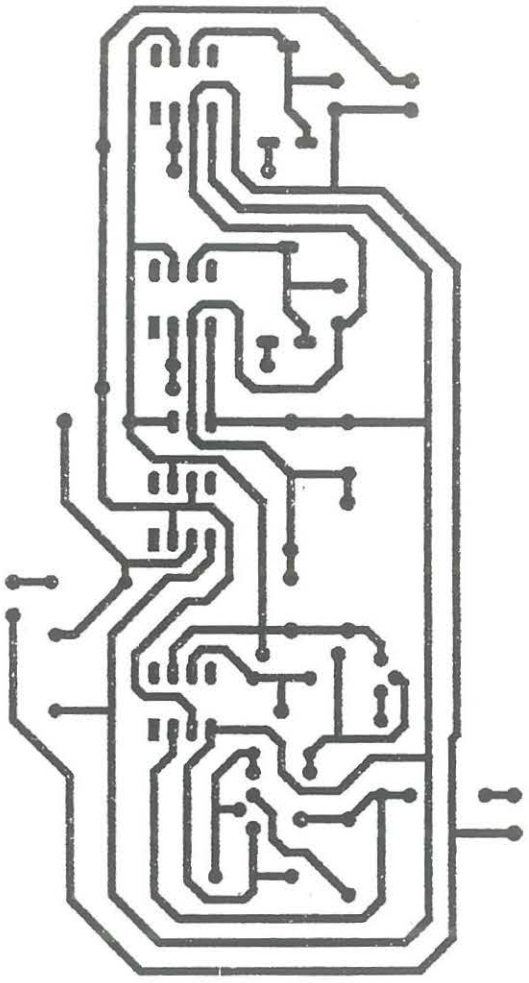
Solder side of the input preamplifier.



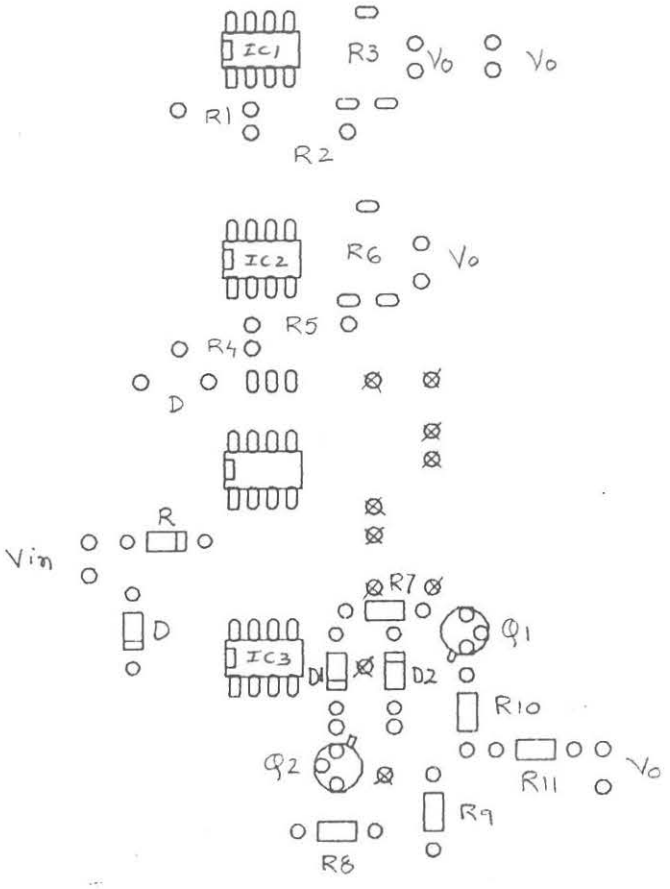
Component placement on the preamplifier board [Figure 4.2]



Component side of the output board.



Solder side of the output board.



Component placement on the output board [Figure 4·5]