DEVELOPMENT OF A CASCADE POLE-ZERO SPEECH SYNTHESIZER

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Master of Technology

by

DARSHANA M. KULKARNI

(90307060)

Guide : Dr. P.C. Pandey

Department of Electrical Engineering

Indian Institute of Technology, Bombay

January 1992

DISSERTATION APPROVAL SHEET

Dissertation entitled "DEVELOPEMENT OF A CASCADE POLE-ZERO SPEECH SYNTHESIZER" is approved for the award of the degree for Master of Technology in Electrical Engineering.

Guide: _____

Internal Examiner: _____

External Examiner: _____

Chairman: _____

21.2.92

Darshana M. Kulkarni: Development of a cascade pole-zero speech synthesizer, M.Tech. dissertation, Department of Electrical Engineering, I.I.T. Bombay, January 1992.

ABSTRACT

A software based synthesizer with flexibility for introducing controlled changes in the characteristics of the speech signal is a useful tool in testing and calibrating various sensory aids for the hearing impaired. Klatt synthesizer is a software synthesizer in which vowels are synthesized by employing a cascade model, and zeros in the spectra of the speech segments with frication are simulated by a bank of formant resonators connected in parallel with amplitude control for individual resonators. This project is aimed at developing a synthesizer as a modification to Klatt synthesizer, which uses a pole-zero cascade model of the vocal tract.

A software based pole-zero cascade synthesizer which uses a cascade model for synthesis of vowels and employs antiresonators to simulate zeros in the spectra of speech segments with frication, along with a program for graphical generation of the parameter tracks, has been developed.

The project also involved the development of software tool for speech analysis and display. Spectral analysis of digitized natural utterances was carried out to obtain parameter tracks for the testing of cascade pole-zero synthesizer for Hindi phonemes, and the approach can be extended to the other Indian languages.

ACKNOWLEDGEMENTS

I take this opportunity to express my gratitude towards my guide Dr. P.C. Pandey for his guidance and constant encouragement throughout this work.

I am thankful to my friends, Mr A.Sharma, Mr A. Kulkarni, Mr C. Bhat, Mr V. Nagarkar, Mr K. Vyas, Mr N. Khambete, Mr S. Chafekar who helped me in several ways during the work on this project.

I wish to express my sincere thanks to Mr A. Vartak and Mr A. Apte from Standards Lab for their co-operation and timely help.

I want to make special mention of my husband Rajesh and my brother Harish for their constant encouragement, right from the beginning of this project.

Last but not the least, I gratefully acknowledge the invaluable help received from Mr R. Sapre through informal discussions and timely suggestions.

Darshana M. Kulkarni

List of abbreviations and symbols

A1	first formant amplitude
A2	second formant amplitude
A3	third formant amplitude
AB	amplitude of bypass path
A/D	analog to digital
ASCII	American National Standard code for Information Interchange
AF	amplitude of frication
AV	amplitude of voicing
B1	bandwidth of first formant
B2	bandwidth of first formant
B3	bandwidth of third formant
BZ1	bandwidth of first zero
BZ2	bandwidth of second zero
BZ3	bandwidth of third zero
D/A	digital to analog
DAP	data acquisition peripheral
DFT	discrete Fourier transform
F0	fundamental frequency of voice pitch
F1	first formant frequency
F2	second formant frequency
F3	third formant frequency
FC	cutoff frequency
FFT	fast Fourier transform
FZ1	first zero frequency
FZ2	second zero frequency
FZ3	third zero frequency
LP	linear prediction
LPF	lowpass filter
UTT	duration of utterance
VCV	vowel-consonant-vowel

CONTENTS

Abstract
Acknowledgements
List of symbols and abbreviations

Chapters

1	Introduction	1
1.1	Overview of the problem	1
1.2	Project objectives	2
1.3	Outline of the dissertation	3
2	Speech synthesis: an overview	5
2.1	Introduction	5
2.2	Classification of speech synthesizers	6
2.3	Cascade/parallel all pole synthesizer	8
2.3.1	Digital resonator	8
2.3.2	Sources of excitation	10
2.3.3	Control of source amplitudes	10
2.3.4	Vocal tract transfer function	11
2.3.5	Radiation characteristics	13
2.4	Cascade pole-zero synthesizer	13
	Tables and figures	15
3	Cascade pole-zero synthesizer	21
3.1	Introduction	21
3.2	Scheme for cascade pole-zero synthesizer	22
3.3	Software implementation	23
3.4	Parameters extraction: zero frequencies computation	26
	Tables and figures	30
4	Speech analysis	42
4.1	Introduction	42

4.2 Software for speech analysis and display 42
4.3 Digitized speech data 44
4.4 Analysis results 45
Tables and figures 48

5 Parameter track generation 51
5.1 Introduction 51
5.2 Program for graphical generation of parameter tracks 51
5.3 Merging parameter files 54
Tables and figures 56

6 Synthesis using cascade pole-zero synthesizer 58
6.1 Introduction 58
6.2 Steps in speech synthesis 58
6.3 Synthesis strategies 59
6.3.1 Synthesis of vowels 59
6.3.2 Synthesis of semivowels 61
6.3.3 Synthesis of fricatives 61
6.3.4 Synthesis of affricates 63
Figures 64

7 Summary and suggestions for further work 73
7.1 Introduction 73
7.2 Work done 73
7.3 Suggestions for further work 74

References 76

Appendices
A Signal handling 79
A.1 Introduction 79
A.2 Hardware set-up 79

A.3 Software tools 81

B Lin Bairstow algorithm for polynomial root solving 84

C Linear predictive analysis of speech 88
C.1 Introduction 88
C.2 Basic principles 88
C.3 Autocorrelation method 90
C.4 Durbin's recursive algorithm 93

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW OF THE PROBLEM

In the areas of psychoacoustic studies and speech sciences, there is often a need for flexible speech generator. A number of speech synthesizers suitable for one or more applications are

(1) Efficient real-time synthesis of speech output from computers or communication equipment.

(2) Synthesis for studying speech production-simulation of vocal tract as a set of interconnected acoustic tube sections and attempt to represent physical variables of the vocal tract.

(3) Formant synthesis for use in study of speech perception.

A flexible synthesizer would be useful in testing and calibration of various sensory aids such as hearing aids, vibrotactile aids, visual aids, and cochlear prosthesis (Levitt, 1980; Pandey, 1987). For example, for testing of pitch estimation schemes, speech segments of known pitch variation are required. Further this performance should be evaluated for male, female, and child voice. This means that, control over formant frequencies and bandwidths is also required. For carrying out psychoacoustic tests using hearing aids, a set of speech segments having particular characteristics is needed, e.g., vowel-consonant-vowel (VCV) sequences with desired duration, constant for all segments,

constant pitch and amplitude for voiced portion, etc. Thus such experiments need a source capable of producing the same sound repeatedly. Also such source should be flexible enough to allow to alter its various parameters as per requirement and one should be able to control its output accurately and precisely. If natural speech segments are used for performance measurement, results will not provide causes for relative success or failure of the device. Also the synthesized signals have advantages over natural signals in the sense that they can be kept very simple, their parameters can be closely controlled, and effect of variation of different parameters can be easily observed. All the above mentioned requirements are satisfied by a software speech synthesizer.

1.2 PROJECT OBJECTIVES

This project is aimed at developing a cascade pole-zero synthesizer for which implementation scheme has been suggested by Chafekar (1990), as a modification to Klatt synthesizer (Klatt, 1980). A software can be written for computation of the control parameters for this synthesizer, i.e., zero frequencies and zero bandwidths from the control parameter data provided by Klatt for cascade/parallel synthesizer.

A program for graphical generation of the parameter tracks, which displays a variable parameter as a function of time and provides facility to edit the track, need to be developed.

For testing the validity of a speech synthesis approach, accurate model parameters for various speech segments are required. Hence to extract parameters from natural speech, an efficient spectral analysis software has to be developed. Linear prediction (LP) analysis is one of the available techniques, which is used for estimation of formant frequencies and bandwidths. With respect to other available techniques such as cepstrum analysis, LP analysis offers advantage of minimum complexity, minimum computation, and maximum accuracy. This spectrum analysis software is further used for extracting parameters from digitized natural speech segments as well as to observe spectrum of synthesized speech.

1.3 OUTLINE OF THE DISSERTATION

Chapter 2 "Speech synthesis: an overview" describes basic models of speech production and different types of speech synthesizers: cascade/parallel speech synthesizer and cascade pole-zero synthesizer.

Chapter 3 "Cascade pole-zero synthesizer" presents scheme for development of cascade pole-zero synthesizer and its software implementation. It also describes the software for computing control parameters, i.e., zero frequencies and zero bandwidths from the parameter data given by Klatt for cascade/parallel synthesizer.

Chapter 4 "Speech analysis" comprises of software for spectral

analysis of speech and graphical display. It also presents the results of the analysis of natural utterances.

Chapter 5 "Parameter track generation" describes the software for graphical generation of the parameter tracks.

Chapter 6 "Synthesis using cascade pole-zero synthesizer" gives general procedure for synthesis. Strategies for synthesis of various classes of speech sounds are discussed. It also gives results of the experiments carried out for synthesis of VCV sequences.

Chapter 7 "Summary and suggestions for further work" comprises a section summarizing work done and a section with suggestions for further improvements.

Appendices provide information about signal handling system: Hardware set-up and software set-up, algorithm for solving for roots of a polynomial, and LP analysis for formants and bandwidth extraction.

Source code listings of the software developed in this project is made available in a separate volume (Kulkarni, 1992). This include programs PZSYNTH (cascade pole-zero synthesizer), PARTRC (graphical generation of parameter tracks), and SPAN (speech analysis and display package).

CHAPTER 2

SPEECH SYNTHESIS: AN OVERVIEW

2.1 INTRODUCTION

Speech synthesis is the process of producing an acoustic signal by controlling the model for speech production with an appropriate set of parameters. One of the first electrical synthesizers which attempted to produce connected speech was the voder, reported by Dudley in 1939 (Flanagan, 1972), following the principle of separation of excitation source and vocal tract. This device used electrical networks which could be selected by finger actuated keys and whose resonances were similar to those of individual speech sounds.

Advent of digital hardware and computers has revolutionized speech synthesis development. A number of special purpose digital signal processing chips, which generate speech by simulating the vocal tract and calculating excitation waveform, have become available. These chips provide high speed speech synthesis and hence they can be used in real-time applications. But, generally they have an in-built set of formant frequencies and hence, lack flexibility required for psychoacoustic and speech perception studies. The advantage of a software implementation over a hardware implementation is that the configuration can be easily changed as new ideas are proposed. Secondly, a software based speech synthesizer, being essentially a programmable synthesizer,

can provide flexibility for controlling the parameters as needed for generation of the test stimuli. Also speed requirement is not very critical as the test stimuli can be synthesized off line.

2.2 CLASSIFICATION OF SPEECH SYNTHESIZERS

Electrical speech synthesizers fall into two broad categories: articulatory and terminal analog. Articulatory synthesizer attempts to duplicate geometry and distributed properties of the tract. It simulates air pressure or flow as a function of time and position in an acoustic tube for which the cross sectional area is a function of position. Thus acoustic tube analog synthesizer attempts to represent physical variables of the vocal tract more directly. Fig. 2.1(a) shows a typical articulatory model, which uses nine co-ordinates to describe area of the vocal tract as a function of the distance from the glottis [Coker, 1976]. Three co-ordinates (K,Y,X) are used to specify the location of a large central portion of the tongue and to regulate jaw movements. Two co-ordinates (W,L) specify closure and rounding of lips, two others (R,B) to regulate raising and curling back of the tongue tip. Another variable (C) serves to control general purpose cross section transformation. A ninth variable (N) represents position of the velum. Model has three variables (G,Q,P_s) to control the manner of excitation of the vocal tract. Several internal variables govern shape of pharyngeal section, area of teeth, etc. Acoustic tube analog of

the vocal tract is shown in Fig. 2.1(b).

Terminal analog synthesizers utilize a system whose transfer function approximates the vocal tract transfer function, but whose implementation bears no direct resemblance to the vocal tract structure. Thus it attempts to duplicate the transmission properties of the vocal tract as viewed from its input and output terminals.

The general structure of a formant synthesizer is illustrated in Fig. 2.2. Synthesizer essentially consists of sources of excitation, model to simulate vocal tract transfer function, and model to simulate radiation load. Sound source may be characterized in the frequency domain by a source spectrum $S(j\Omega)$. Vocal tract is an acoustic cavity which is characterized by a set of resonant frequencies or formants. Since the vocal tract is a linear system, it can be characterized in the frequency domain by a linear transfer function $T(j\Omega)$. Output of the vocal tract model, lip volume velocity $U(j\Omega)$ given by

$$U(j\Omega) = S(j\Omega)T(j\Omega) \qquad (2.1)$$

Finally spectrum of output sound pressure $P(j\Omega)$ is related to the lip volume velocity $U(j\Omega)$ by radiation characteristics $R(j\Omega)$

$$P(j\Omega) = U(j\Omega)R(j\Omega) \qquad (2.2)$$

The vocal tract transfer function can be represented by a linear time varying filter as shown in Fig. 2.3. Following sections describe approach to formant based synthesizers: Cascade/parallel all-pole synthesizer and cascade pole zero synthesizer.

2.3 CASCADE/PARALLEL ALL-POLE SYNTHESIZER

Klatt (1980) developed a cascade/parallel synthesizer, which is a software based synthesizer with its source code available in Fortran. There are 39 control parameters that determine the characteristics of the output. As many as twenty control parameters can be varied as a function of time. Klatt synthesizer has flexibility of utilizing either a parallel or a cascade structure. It can be used in two ways, either in the general cascade/parallel mode, or in a special all-parallel mode as shown in Fig. 2.4. Detailed block diagram of Klatt synthesizer is shown in Fig. 2.5, and the 39 control parameters are listed in Table 2.1. Table gives minimum and maximum values of parameters, their type, and whether variable or constant.

2.3.1 Digital resonator

The basic building block of the synthesizer is a digital resonator as shown in Fig. 2.6(a). Samples of the output of a digital resonator y(nT) are computed from the input sequence, x(nT) by the equation

$$y(nT) = ax(nT) + by((n-1)T) + cy((n-2)T) \qquad (2.3)$$

Where y((n-1)T) and y((n-2)T) are the previous two sample values of output sequence y(nT).

$$a = -\exp(-2\,\Pi\,BW\,T),$$
$$b = 2\exp(-\Pi\,BW\,T)\cos(2\,\Pi\,F\,T) \qquad (2.4)$$
$$c = 1 - a - b$$

where T = 1/sampling rate F = center frequency b = bandwidth

A digital resonator is a second order difference equation. The transfer function of a digital resonator has a sampled frequency response given by

$$T(z) = \frac{a}{1 - bz^{-1} - cz^{-2}} \qquad (2.5)$$

Frequency response of digital resonator is as shown in Fig. 2.6(b). An antiresonator can be realized by slight modifications to above mentioned equations for resonator. The output of an antiformant resonator y(nT) is related to the input x(nT) by the equation

$$y(nT) = a'x(nT) + b'x((n-1)T) + c'x((n-2)T) \qquad (2.6)$$

The constants a', b', c' are defined by the equation

$$a' = 1.0/a, \qquad b' = -b/a, \qquad c' = -c/a \tag{2.7}$$

where a, b, c are obtained by inserting the antiresonance center frequency F and bandwidth BW into Eq. (2.4)

2.3.2 Sources of excitation

For voiced sounds a periodic pulse train lowpass filtered by the resonator RGP produces a typical glottal pulse. For unvoiced sounds, a noise source is used. This is used for frication as well as aspiration. The periodic pulse train is used to modulate the output of the noise source during the production of voiced fricatives and voiced stops.

Sometimes during the production of voiced fricatives a smoothed quasi-sinusoidal voicing is needed. Resonator RGS further filters the glottal pulse to yield such an excitation. The frication source is typically simulated by pseudo-random generator. The noise source is an ideal pressure source and hence its output must be lowpass filtered to give equivalent volume velocity.

2.3.3 Control of source amplitudes

The amplitude of voiced and unvoiced sources are controlled by parameters AV and AF respectively. The amplitude of aspiration is controlled by AH. Voicing amplitudes are adjusted at the onset of each glottal pulse. The noise amplitudes AF and AH are used to

interpolate the intensity of noise sources linearly over 5 ms interval. It is also possible to specify sudden bursts for plosive releases.

2.3.4 Vocal tract transfer function

The acoustic characteristics of the vocal tract are determined by its cross sectional area as a function of distance from the larynx to the lips. This can be represented as a cascade or parallel combination of resonators, each tuned to different formant frequency. Vocal tract transfer function contains only about five complex pole pairs and no zeros as long as articulation is non nasalized and sound source is in the larynx. Hence it can be represented by an all-pole model. Cascade model consists of five formant resonators, and it has a transfer function that can be represented in the frequency domain as a product of transfer functions

$$T_n(z) = \prod_{n=1}^{5} \frac{a_n}{1 - b_n z^{-1} - c_n z^{-2}} \qquad (2.8)$$

where constants a_n, b_n, and c_n are determined by the values of n^{th} formant frequency F_n and n^{th} formant bandwidth BW_n by the relations given earlier in Eq. (2.4). The frequency of the formant peak "n" is determined by the formant frequency control parameter Fn. Formant frequency values are determined by the

detailed shape of the vocal tract. The frequencies of the lowest three formants vary substantially with changes in articulation. The formant bandwidth is a function of energy losses due to heat conduction, viscosity, cavity wall motions, radiation of sound from the lips, and the real part of the glottal source impedance. Nasal murmurs and vowel nasalization are approximated by the insertion of an additional resonator RNP and antiresonator RNS, into cascade vocal tract model. For adult male nasal pole frequency FNP can be set to a fixed value of about 270 Hz for all the time. The nasal zero frequency FNS should also be set to a value of about 270 Hz during non-nasalized sounds, but the frequency of the nasal zero must be increased during the production of nasals and nasalization.

Satisfactory approximation to the vocal tract transfer function for frication excitation can be achieved with parallel set of digital resonators having amplitude controls, and no antiresonators. The presence of transfer function zeros is accounted for by appropriate setting of the formant amplitude controls. There are six formant resonators in the parallel configuration. A sixth formant has been added to the parallel branch specifically for the synthesis of very high frequency of noise in /s z/. A bypass path with amplitude control AB is included because the transfer function for /f v p b/ contains no prominent resonant peaks. Hence the synthesizer should include a means of bypassing all of the resonators to produce a flat

transfer function.

During the production of a voiced fricative, the output of the quasi-sinusoidal voicing source is sent through the cascade vocal tract model, while frication source excites the parallel branch.

2.3.5 Radiation characteristics

Radiation load at the lips acts as a first order high pass filter. It is simulated in the synthesizer by taking the first difference of lip-nose volume velocity

$$p(nT) = u(nT) - u((n-1)T) \qquad (2.9)$$

The radiation characteristics add a gradual rise in the overall spectrum. The radiation characteristics is moved into the source shaping filters. Thus for unvoiced sounds, -6 dB/octave slope in the frequency response of this filter cancels +6 dB/octave radiation effect at the lips, leaving a net flat spectrum, unlike the -6 dB/octave net fall off in voiced sounds.

2.4 CASCADE POLE-ZERO SYNTHESIZER

Another approach to terminal analog synthesizer is to simulate vocal tract transmission in terms of individual poles and zeros, by cascade connection of resonators and antiresonators. A scheme for a speech synthesizer using a

pole-zero model has been suggested by Chafekar (1990). A program PZSYNTH was written to implement this scheme. This scheme along with its software implementation will be described in next chapter.

Table 2.1 Control parameters for cascade/parallel synthesizer.

Source:Klatt(1980), Table 1.

N	V/C	Sym	Name	Min	Max	Typ
1	V	AV	Amplitude of voicing (dB)	0	80	0
2	V	AF	Amplitude of frication (dB)	0	80	0
3	V	AH	Amplitude of aspiration (dB)	0	80	0
4	V	AVS	Amplitude of sinusoidal voicing (dB)	0	80	0
5	V	F0	Fundamental freq. of voicing (Hz)	0	500	0
6	V	F1	First formant frequency (Hz)	150	900	450
7	V	F2	Second formant frequency (Hz)	500	2500	1450
8	V	F3	Third formant frequency (Hz)	1300	3500	2450
9	V	F4	Fourth formant frequency (Hz)	2500	4500	3300
10	V	FNZ	Nasal zero frequency (Hz)	200	700	250
11	C	AN	Nasal formant amplitude (dB)	0	80	0
12	C	A1	First formant amplitude (dB)	0	80	0
13	V	A2	Second formant amplitude (dB)	0	80	0
14	V	A3	Third formant amplitude (dB)	0	80	0
15	V	A4	Fourth formant amplitude (dB)	0	80	0
16	V	A5	Fifth formant amplitude (dB)	0	80	0
17	V	A6	Sixth formant amplitude (dB)	0	80	0
18	V	AB	Bypass path amplitude (dB)	0	80	0
19	V	B1	First formant bandwidth (Hz)	40	500	50
20	V	B2	Second formant bandwidth (Hz)	40	500	70
21	V	B3	Third formant bandwidth (Hz)	40	500	110
22	C	SW	Cascade/parallel switch	0(CASC)	1(PARA)	0
23	C	FGP	Glottal resonator 1 frequency (Hz)	0	600	0
24	C	BGP	Glottal resonator 1 bandwidth (Hz)	100	2000	100
25	C	FGZ	Glottal zero frequency (Hz)	0	5000	1500
26	C	BGZ	Glottal zero bandwidth (Hz)	100	9000	6000
27	C	B4	Fourth formant bandwidth (Hz)	100	500	250
28	V	F5	Fifth formant frequency (Hz)	3500	4900	3750
29	C	B5	Fifth formant bandwidth (Hz)	150	700	200
30	C	F6	Sixth formant frequency (Hz)	4000	4999	4900
31	C	B6	Sixth formant bandwidth (Hz)	200	2000	1000
32	C	FNP	Nasal pole frequency (Hz)	200	500	250
33	C	BNP	Nasal pole bandwidth (Hz)	50	500	100
34	C	BNZ	Nasal zero bandwidth (Hz)	50	500	100
35	C	BGS	Glottal resonator 2 bandwidth	100	1000	200
36	C	SR	Sampling rate	5000	20 000	10 000
37	C	NWS	Number of waveform samples per chunk	1	200	50
38	C	G0	Overall gain control (dB)	0	80	47
39	C	NFC	Number of cascaded formants	4	6	5

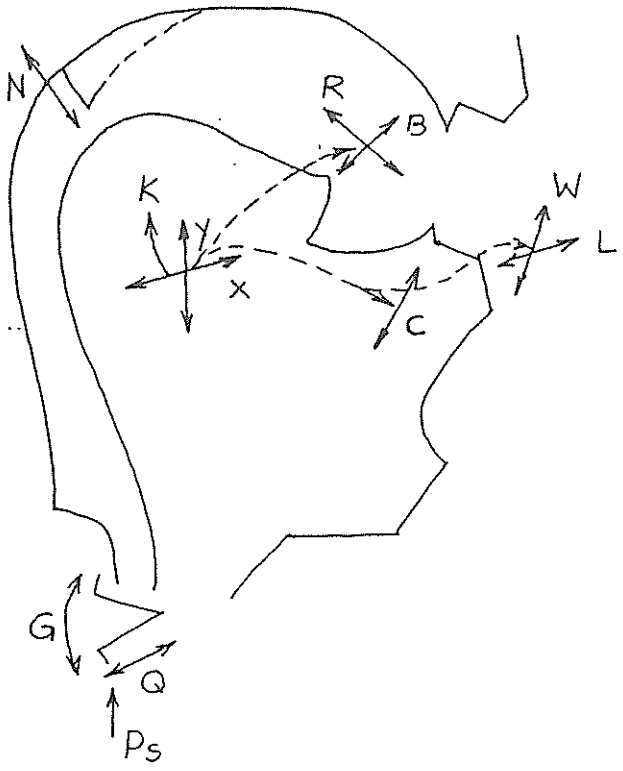


Fig. 2.1(a). Articulatory model of the vocal tract. Source: Coker(1976), Fig. 1

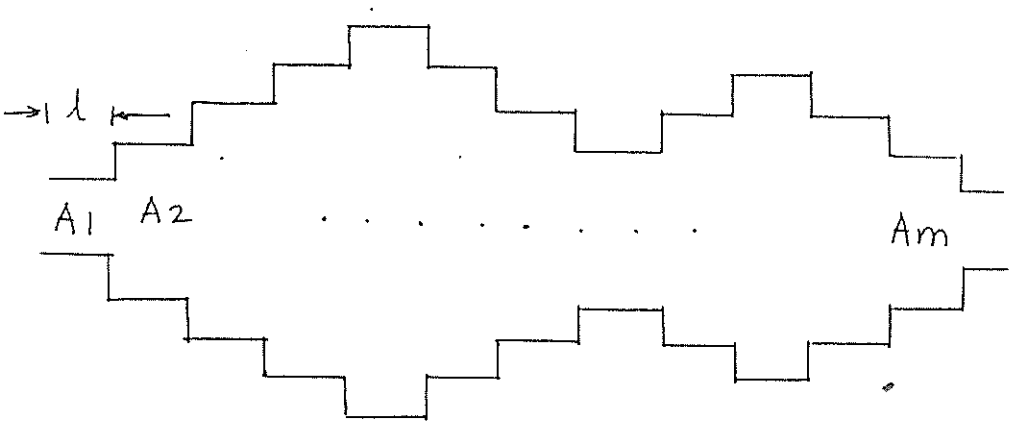


Fig. 2.1(b). Acoustic tube analog of the vocal tract. Source: Rabiner (1978), Fig. 3.7

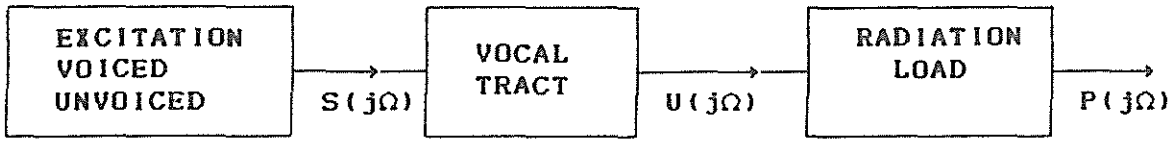


Fig. 2.2 Model of speech production. Source: Rabiner (1978).

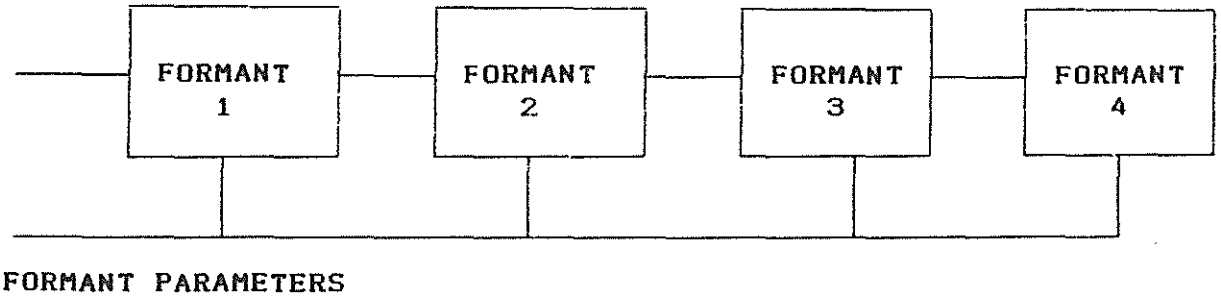
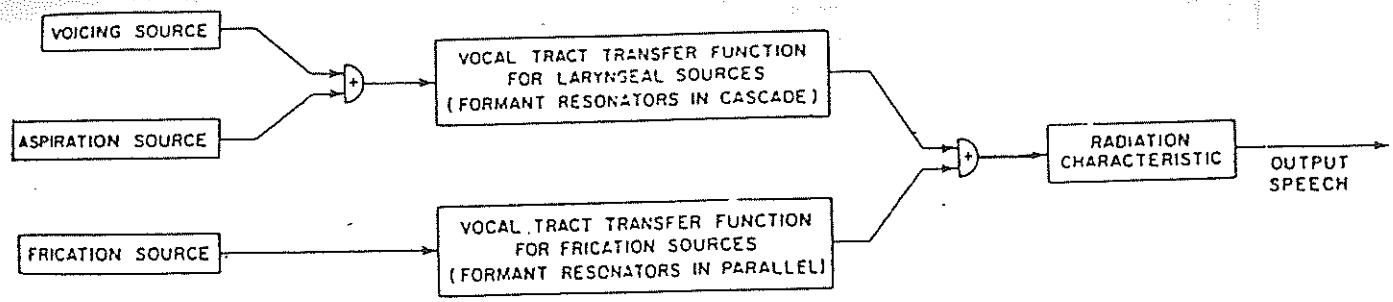
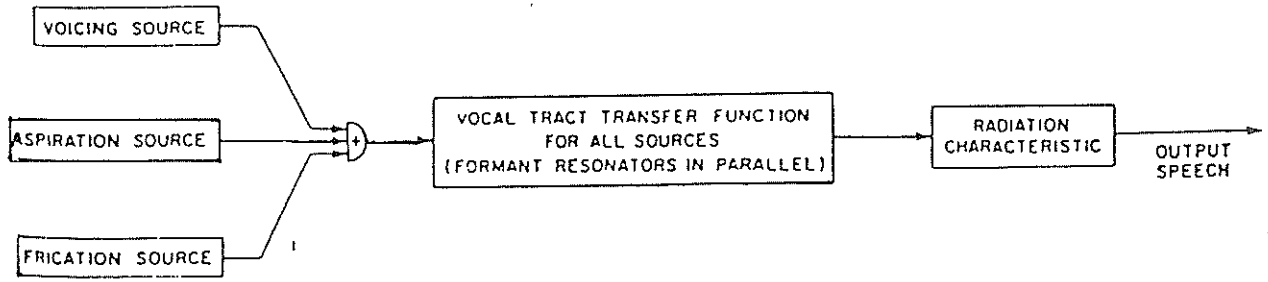


Fig. 2.3 Vocal tract transfer function in a typical formant synthesizer. Source: Rabiner (1978), Fig. 3.5(b)



(A) CASCADE/PARALLEL FORMANT CONFIGURATION



(B) SPECIAL-PURPOSE ALL-PARALLEL FORMANT CONFIGURATION

Fig. 2.4. Configuration of the Klatt synthesizer. Source: Klatt(1980), Fig.4.

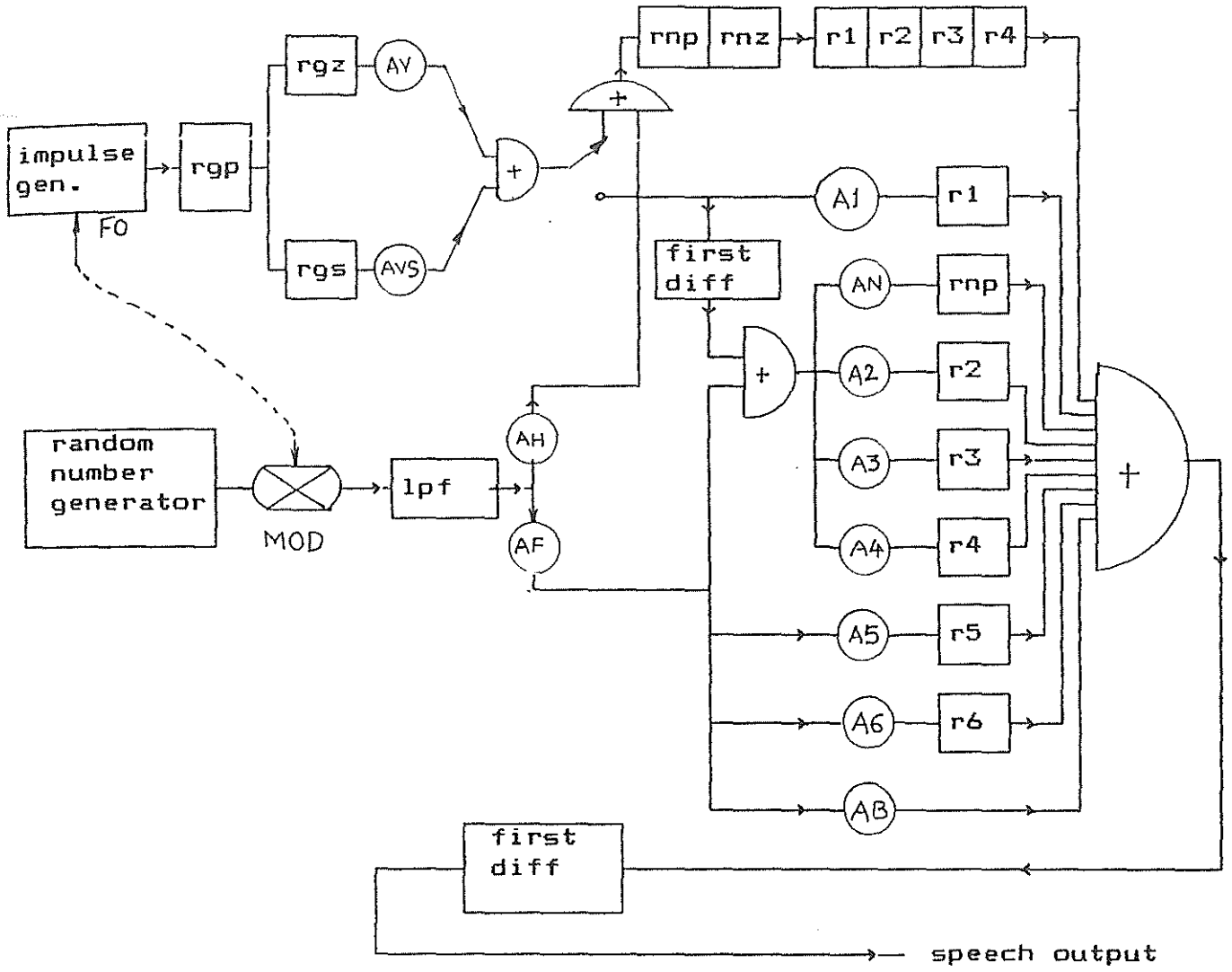


Fig. 2.5. Block digram of Klatt synthesizer. Source: Klatt (1980),fig.6

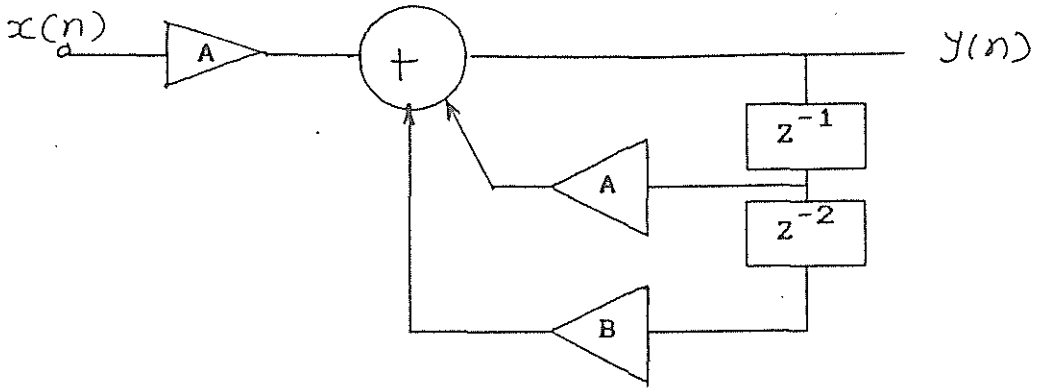


Fig. 2·6(a). A digital resonator. Source: Klatt(1980), Fig. 5

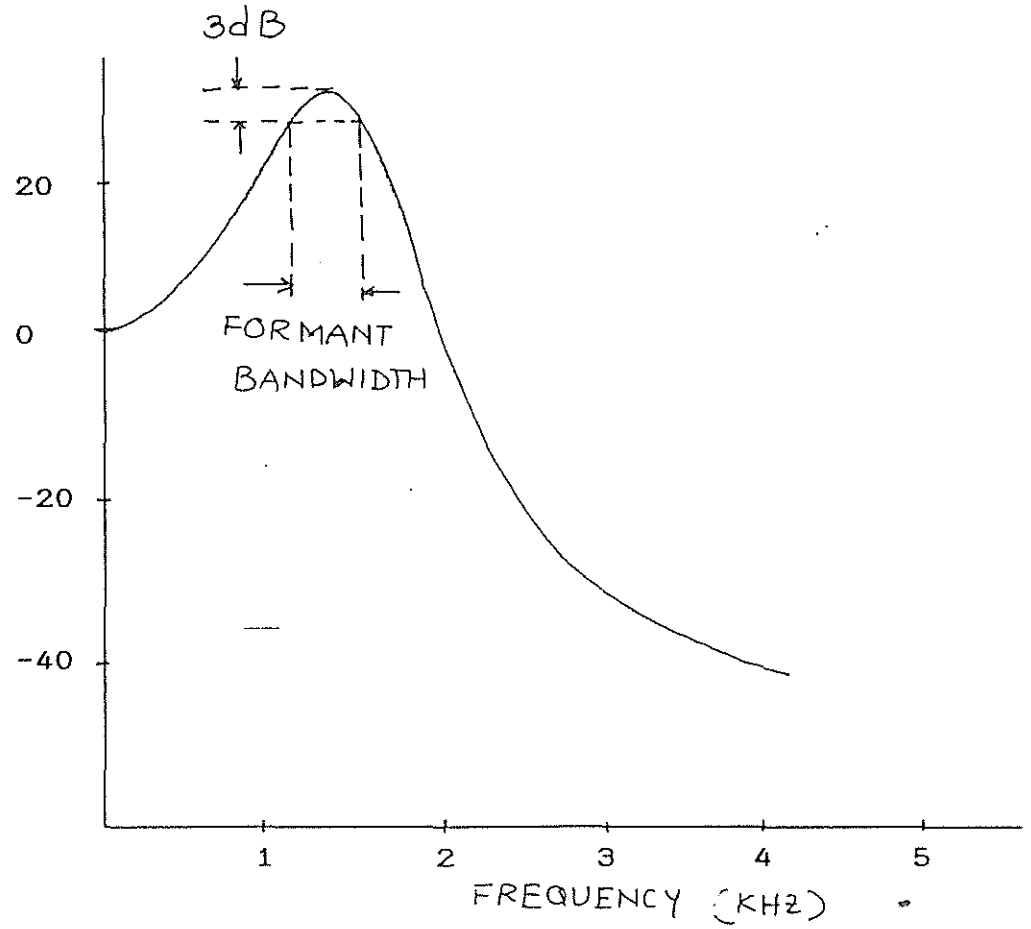


Fig. 2·6(b). Frequency response of resonator. Source: Klatt(1980), Fig. 5

CHAPTER 3
CASCADE POLE-ZERO SYNTHESIZER

3.1 INTRODUCTION

A speech synthesizer, which uses cascade model of vocal tract throughout, and uses antiresonators to simulate zeros in the spectra of sounds with frication excitation, has been developed by Rabiner(1968). But, parameters required for pole-zero model of vocal tract, i.e., pole and zero frequencies and their bandwidths were not published. Klatt (1980) implemented a software based cascade/parallel synthesizer, which uses cascade model for synthesis of vowels, and uses a bank of formant resonators connected in parallel, with amplitude control for individual resonators. Amplitude control data derived from trial and error attempts to match natural frication spectra were also presented. Chafekar (1990) suggested a scheme for cascade pole-zero synthesizer with some modifications to that developed by Rabiner (1968). This model would give more natural representation of speech sounds and would be more efficient in terms of number of blocks to be realized. A program PZSYNTH was written to implement this scheme. Source code listing of this program is available in separate volume (Kulkarni, 1992). Parameters required for synthesis of vowels were obtained by analysis of digitized natural utterances of male and female Hindi speakers, using speech analysis and display software (SPAN), which will be described in Chapter 4. Validity of

pole-zero approach of speech synthesis was tested by extracting parameters for this model, from parameters given by Klatt (1980) for cascade/parallel model. In this chapter, scheme for cascade pole-zero synthesizer, its software implementation, and a method to extract parameters for pole-zero model from parameters for cascade/parallel synthesizer, will be described.

3.2 BLOCK DIAGRAM DESCRIPTION

Fig. 3.1 gives a block diagram of proposed cascade pole-zero synthesizer. Excitation sources are the same as that used in Klatt synthesizer. Voicing is simulated by a pulse generator producing impulse train with period equal to fundamental frequency of voicing. An impulse train is passed through resonator RGP and antiresonator RGZ. Amplitude of voicing can be controlled through parameter AV. Frication source is simulated by a random number generator and a lowpass filter. Noise source is an ideal pressure source, and the volume velocity for frication depends on the impedance seen by this source. Assuming that the volume velocity is proportional to the integral of the source pressure, it is approximated by a first order lowpass digital filter. Amplitude of frication and amplitude of aspiration are controlled through parameters AF and AH respectively. Main difference between cascade pole-zero synthesizer and cascade/parallel synthesizer is in simulation of

vocal tract transfer function. In cascade pole-zero synthesizer six resonators R1 - R6 are connected in cascade. In addition there are five antiresonators RZ1 - RZ5 which simulate zeros in the spectra of speech segments involving frication.

Vowels are simulated by using voiced excitation source and resonators connected in cascade. Unvoiced fricatives and burst portions of unvoiced stops are simulated using frication excitation source. For simulation of voiced fricatives and voiced stops output of the impulse generator modulates ouput of the noise generator producing pitch synchronous excitation.

Nasals and vowel nasalization are approximated by insertion of additional resonator RNP and antiresonator RNZ. For an adult male speaker, nasal pole frequency FNP can be set to a fixed value of about 270 Hz for all the time. The nasal zero frequency FNZ should be set to a value of about 270 Hz during non-nasalized sounds. It is increased during production of nasals and vowel nasalization.

3.3 SOFTWARE IMPLEMENTATION

A program PZSYNTH (pole-zero synthesizer), which runs on IBM PC was developed to implement cascade pole-zero synthesizer. The program reads parameters, their types, default values, synthesis duration, and any time varying values of each of 39 control parameters from the file specified by the user. Generation

of the parameter file is accomplished by a program PARTRC, which displays a parameter track and facilitates its editing. This program will be described in the next chapter. User can also specify output speech file name in which generated speech samples are stored for later use. An option is provided for storing speech samples in either a binary format or as an ASCII format. Program displays the parameters in tabular form as the execution commences. As the computation progresses, corresponding values of parameters are varied on the screen.

Modular programming approach was used. Various procedures used in the program are described below and also outlined in the flowchart given in Fig. 3.2.

(1) TRCREAD: This procedure reads the parameters, their types, default values, synthesis duration, and any time varying values of each 39 parameters from the parameter file specified by the user.

(2) LINAMP: Converts excitation amplitudes in dB to its sample value.

(3) FILTCOEFF: Computes filter coefficients for second order resonators and antiresonators from formant frequencies and bandwidth information read from the parameter file.

(4) CASCADE: Simulates six resonators connected in cascade. Output of the last resonator is given by "velpole".

(5) PARALLEL: Simulates six antiresonators connected in cascade. This branch is activated only when frication excitation is present. Velpole (output of cascaded resonators) is given as input to the first anti-resonator. Output of the final antiresonator is given by "velzero".

(6) EXCITCOEFF: Scales excitation source amplitudes and computes filter coefficients for source shaping filters RGP, RGZ, and RGS. Frequency control of RGP is set to 0 Hz, while bandwidth is varied according to pitch specification. Frequency control of RGS is set to 0 Hz to produce lowpass filter, and bandwidth is set to 200 Hz, which determines the cutoff frequency beyond which the harmonics are strongly attenuated.

(7) EXCITOUT: Generates the output of excitation filters. A pseudo-random number generator is used to generate both burst and aspiration. Noise amplitudes AF and AH are used to interpolate the intensity of the noise source linearly over 5 ms interval. Interpolation permits a more gradual onset for fricatives. A plosive burst involves a more rapid source onset that can be achieved by 5 ms linear interpolation. Therefore if AF is increased by more than 50 dB from its value specified in the previous 5 ms segment, AF is changed instantaneously to its new target

value.

(8) VOCACALOUT: Computes output of the vocal tract filter.

Execution speed on IBM PC without a math co processor is low, i.e., it takes 500 sec for generation of 1 sec speech. But, computational delay is not a serious problem, as for the intended application, stimuli can be generated and stored on disk for later use.

Hardware setup for for listening to synthesized speech consists of a data acquisition card with 12 bit A/D and D/A converter, audio equipment such as amplifier, speaker, tape recorder, and headphone. Steps in speech synthesis are outlined in Fig. 6.1(a) (Chapter 6) and block diagram of hardware setup used for listening to synthesized speech is given in Fig. (6.2)

3.4 PARAMETERS EXTRACTION: COMPUTATION OF ZERO FREQUENCIES

Table 3.1 lists control parameters for cascade pole-zero synthesizer. It gives whether a particular parameter is a variable or constant. It also gives permitted range of values of each parameter. In Klatt synthesizer, presence of any transfer function zero is accounted by appropriate settings of formant amplitude control of resonators in parallel configuration. This amplitude control data was derived from trial and error attempts to match natural frication spectra. These data along with formant frequencies and bandwidths are given in Table 3.2. Parameter data

for cascade pole-zero synthesizer, i.e., zero frequencies and bandwidths, were obtained from amplitude control data given by Klatt. A program ZERDAN was written in Pascal, with following steps in computation of zero frequencies and bandwidths.

(1) Read amplitude control data, formant frequencies and bandwidths given by Klatt, either from keyboard or from a file.

(2) Compute correction factors for parallel resonator amplitudes taking into consideration formant proximities.

(3) Compute filter coefficients of second order filters from corrected amplitudes, formant frequencies and bandwidths

(4) Vocal tract transfer function of the system realized using parallel combination of the filters will be of the form:

$$\sum_{n=1}^{k} \frac{a_n}{1 - b_n z^{-1} - c_n z^{-2}}$$

Expand this equation to obtain numerator polynomial of the form

$$N(z) = a_{2k} z^{-2k} + a_{2k-1} z^{-(2k-1)} + \ldots + a_o \qquad (3.1)$$

$$N(z) = a_{2k}\, z^{-2k} + a_{2k-1}\, z^{-(2k-1)} + \ldots + a_o \qquad (3.1)$$

Now the system transfer function will be of the form

$$\frac{N(z)}{\prod_{n=1}^{k} (1 - b_n z^{-1} - c_n z^{-2})}$$

This transfer function can be realized by cascade connection of resonators and antiresonators.

(5) Roots of the numerator polynomial are computed using Lin Bairstow algorithm of quadratic factors (Hovanessian, 1969). This algorithm is described in the Appendix B.

(6) Zero frequencies and bandwidths are obtained using formula

$$ZF = \frac{1}{2\pi T} \tan^{-1} \frac{\omega}{\sigma} \qquad (3.2)$$

$$BZ = \frac{-1}{2\pi T} \ln [\omega^2 + \sigma^2]$$

where ω and σ are imaginary and real parts of the root. Table 3.2 gives parallel resonators amplitude control data, formants and bandwidths for fricatives ,affricates and stops as given by Klatt for cascade/parallel synthesizer. Table 3.3 lists

bandwidths as computed using program ZEROAN, from data given in Table 3.2. Also details of phonemes in Hindi and English, referred, are tabulated in Table 3.4 and Table 3.5. These details include IPA symbols, features such as manner of articulation, place of articulation, presence or absence of voicing, and keyword for English phonemes.

Table 3.1. Control parameters for pole-zero cascade synthesizer. The list also shows the permitted ranges of values for each parameter and a typical value. V/C indicates whether a parameter is normally variable or constant. Source: Chafekar(1990), Fig. 4.1

N	V/C	Sym	Name	Min	Max	Typ
1	C	NF	Number of formants	4	6	4
2	V	NZ	Number of zeros	0	4	0
3	V	F0	Fundamental freq. of voicing(HZ)	0	500	0
4	V	AV	Ampl. of voicing (dB)	0	80	0
5	V	AF	Ampl. of frication (dB)	0	80	0
6	V	AS	Ampl. of sinusoidal voicing(dB)	0	80	0
7	V	AH	Ampl. of aspiration (dB)	0	80	0
8	V	F1	First formant freq.(Hz)	150	900	450
9	V	F2	Second formant freq. (Hz)	500	2500	1450
10	V	F3	Third formant freq. (Hz)	1300	3500	2400
11	V	BW1	First formant bandwidth (Hz)	40	500	50
12	V	BW2	Second formant bandwidth (Hz)	40	500	70
13	V	BW3	Third formant bandwidth (Hz)	40	500	110
14	V	FZ1	First zero freq. (Hz)	150	5000	-
15	V	FZ2	Second zero freq. (Hz)	150	5000	-
16	V	FZ3	Third zero freq. (Hz)	150	5000	-
17	V	FZ4	Fourth zero freq. (Hz)	150	5000	-
18	V	BZ1	First zero bandwidth (Hz)	-	-	-
19	V	BZ2	Second zero bandwidth (Hz)	-	-	-
20	V	BZ3	Third zero bandwidth (Hz)	-	-	-
21	V	BZ4	Fourth zero bandwidth (Hz)	-	-	-
22	V	FMZ	Nasal zero freq.(Hz)	200	700	250
23	V	FMP	Nasal pole freq.(Hz)	200	500	250
24	C	UPDT	Parameter update int (ms)	2	20	5
25	C	SR	Sampling rate (Hz)	5000	20000	10000
26	C	GO	Overall gain. control(dB)	0	80	0
27	C	F4	Fourth formant freq.(Hz)	2500	4500	3300
28	C	F5	Fifth formant freq.(Hz)	3500	4900	3750
29	C	F6	Sixth formant freq.(Hz)	4000	4999	4900
30	C	BW4	Fourth formant bandwidth(Hz)	100	5000	250
31	C	BW5	Fifth formant bandwidth(Hz)	150	7000	200
32	C	BW6	Sixth formant bandwidth(Hz)	200	2000	1000
33	C	BWNZ	Nasal zero bandwidth(Hz)	50	500	100
34	C	BWNP	Nasal pole bandwidth(Hz)	50	500	100
35	C	FGP	Glottal res. 1 freq.(Hz)	0	600	0
36	C	BWGP	Glottal res. 1 bandwidth (Hz)	100	2000	100
37	C	FGZ	Glottal zero freq. (Hz)	0	5000	1500
38	C	BWGZ	Glottal zero Bandwidth (Hz)	100	9000	6000
39	C	BWGS	Glottal res. 2 bandwidth	100	1000	200

Table 3.2(a). Control parameters for the synthesis of semi-vowels and fricatives before front vowels. Source: Klatt (1980), Table 3. Key words for these phonemes are given in Table 3.5(a).

Semi-vowel	F1	F2	F3	B1	B2	B3
w	290	610	2150	50	80	60
j	260	2070	3020	40	250	500
r	310	1060	1380	70	100	120
l	310	1050	2880	50	100	280

Fric	F1	F2	F3	B1	B2	B3	A2	A3	A4	A5	A6	AB
f	340	1100	2080	200	120	150	0	0	0	0	0	57
v	220	1100	2080	60	90	120	0	0	0	0	0	57
θ	320	1290	2540	200	90	200	0	0	0	0	28	48
ð	270	1290	2540	60	80	170	0	0	0	0	28	48
s	320	1390	2530	200	80	200	0	0	0	0	52	0
z	240	1390	2530	70	60	180	0	0	0	0	52	0
ʃ	300	1840	2750	200	100	300	0	57	48	48	46	0

Table 3.2(b). Parameters values for the synthesis of affricates, nasals and stops before front vowel. Source: Klatt (1980), Table 3. Key words for these phonemes are given in Table 3.5(a).

Affric.	F1	F2	F3	B1	B2	B3	A2	A3	A4	A5	A6	AB
tʃ	350	1800	2820	200	90	300	0	44	60	53	53	0
dʒ	260	1800	2820	60	80	270	0	44	60	53	53	0

stops	F1	F2	F3	B1	B2	B3	A2	A3	A4	A5	A6	AB
p	400	1100	2150	300	150	220	0	0	0	0	0	63
b	200	1100	2150	60	110	130	0	0	0	0	0	63
t	400	1600	2600	300	120	250	0	30	45	57	63	0
d	200	1600	2600	60	100	170	0	47	60	62	60	0
k	300	1990	2850	250	160	330	0	53	43	45	45	0
g	200	1990	2850	60	150	280	0	53	43	45	45	0

Nasals	FNP	FNZ	F1	F2	F3	B1	B2	B3
m	270	450	480	1270	2130	40	200	200
n	270	450	480	1340	2470	40	300	300

Table 3.3(a). Control parameter values for the synthesis of fricatives using pole-zero synthesizer. Zero frequencies and bandwidths are derived from parameter data as given in the table 3.2(a).

Fric	F1	F2	F3	B1	B2	B3	FZ1	FZ2	FZ3	FZ4	FZ5	BZ1	BZ2	BZ3	BZ4	BZ5
f	340	1100	2080	200	120	150	584	1825	3300	3750	4900	184	144	250	200	1000
v	220	1100	2080	60	90	120	524	1822	3300	3750	4900	66	110	250	200	1000
θ	320	1290	2540	200	90	200	472	1520	2640	3300	3750	202	147	257	250	200
ð	270	1290	2540	60	80	170	431	1520	2640	3300	3750	66	135	230	250	200
s	320	1390	2530	200	80	200	320	1390	2536	3300	3750	200	84	203	250	200
z	240	1390	2530	70	60	180	260	1400	2536	3300	3750	70	64	183	250	200
ʃ	300	1840	2750	200	100	300	300	1840	3152	3652	4473	200	100	260	218	927

Table 3.3(b). Control parameters for the synthesis of stops and affricates using pole-zero synthesizer. Zero frequencies and bandwidths are derived from parameter data as given in table 3.2(b).

Stops	F1	F2	F3	B1	B2	B3	FZ1	FZ2	FZ3	FZ4	FZ5	BZ1	BZ2	BZ3	BZ4	BZ5
p	400	1100	2150	300	150	220	622	1870	3300	3700	4900	270	205	250	200	1000
b	200	1100	2150	60	110	130	518	1864	3300	3750	4900	70	122	250	200	1000
t	400	1600	2600	300	120	250	400	1600	2540	2627	3363	300	120	256	266	487
d	200	1600	2600	60	100	170	200	1600	2693	3496	4232	60	100	185	250	754
k	300	1990	2850	250	160	330	300	1990	3179	3671	4482	250	160	270	209	205
g	200	1990	2850	60	150	280	205	1990	3179	3673	4463	66	150	260	214	927
Affricates																
tʃ	350	1800	2820	200	90	300	350	1800	2882	3674	4420	200	90	295	215	895
dʒ	250	1800	2820	60	80	270	260	1800	2882	3672	4420	60	80	268	215	896

Table 3.4(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
æ (bat)	low, front, tense
ʌ (but)	mid, back, lax
ɑ (father)	low, back, lax
ɒ (cot)	low, back, tense
ʊ (foot)	high, back, lax, rounded
u (boot)	high, back, tense, rounded
o (coat)	mid, back, tense, rounded
ɝ (bird)	mid, central, tense
ə (ado)	mid, central, lax
ɔ (all)	mid, back, lax

Table 3.4(b) Classification of Hindi vowels alongwith keywords

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

Table 3.5(a). 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, UN voiced, 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 3.5 (b). Classification of Hindi consonants.

Phoneme Hindi (IPA)	Features manner, voicing, aspiration, place
क (k)	stop, unvoiced, unaspirated, velar
ख (k^h)	stop, unvoiced, aspirated, velar
ग (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 (ñ)	nasal, voiced, unaspirated, alveopalatal
ट (ʈ)	stop, unvoiced, unaspirated, alveolar
ठ (ʈ^h)	stop, unvoiced, aspirated, alveolar
ड (ɖ)	stop, voiced, unaspirated, alveolar
ढ (ɖ^h)	stop, voiced, aspirated, alveolar
ण (ɳ)	nasal, voiced, unaspirated, alveolar
त (t̪)	stop, unvoiced, unaspirated, dental
थ (t̪^{h})	stop, unvoiced, aspirated, dental
द (d̪)	stop, voiced, unaspirated, dental
ध (d̪^{h})	stop, voiced, aspirated, dental
न (n)	nasal, voiced, unaspirated, dental

Table 3.5 (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
थ (θ)	fricative, unvoiced, unaspirated, dental
फ़ (f)	fricative, unvoiced, unaspirated, labiodental
ज़ (z)	fricative, voiced, unaspirated, alveolar
झ़ (ʒ)	fricative, voiced, unaspirated, palatal
ध़ (ð)	fricative, voiced, unaspirated, dental
व़ (v)	fricative, voiced, unaspirated, labiodental

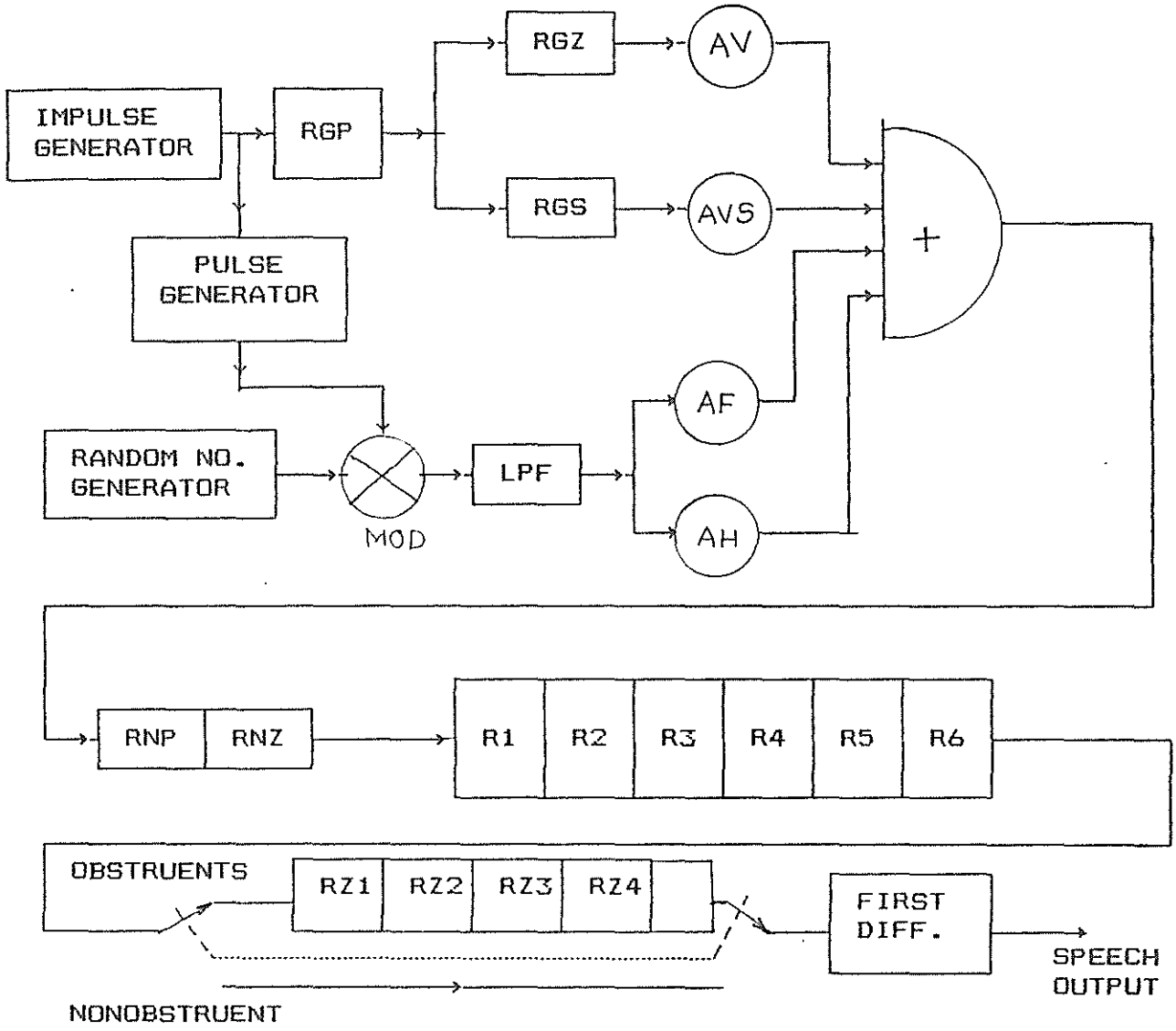


Fig. 3.1. The Cascade Pole Zero Synthesizer. Source:Chafekar (1990), Fig. 4.1.

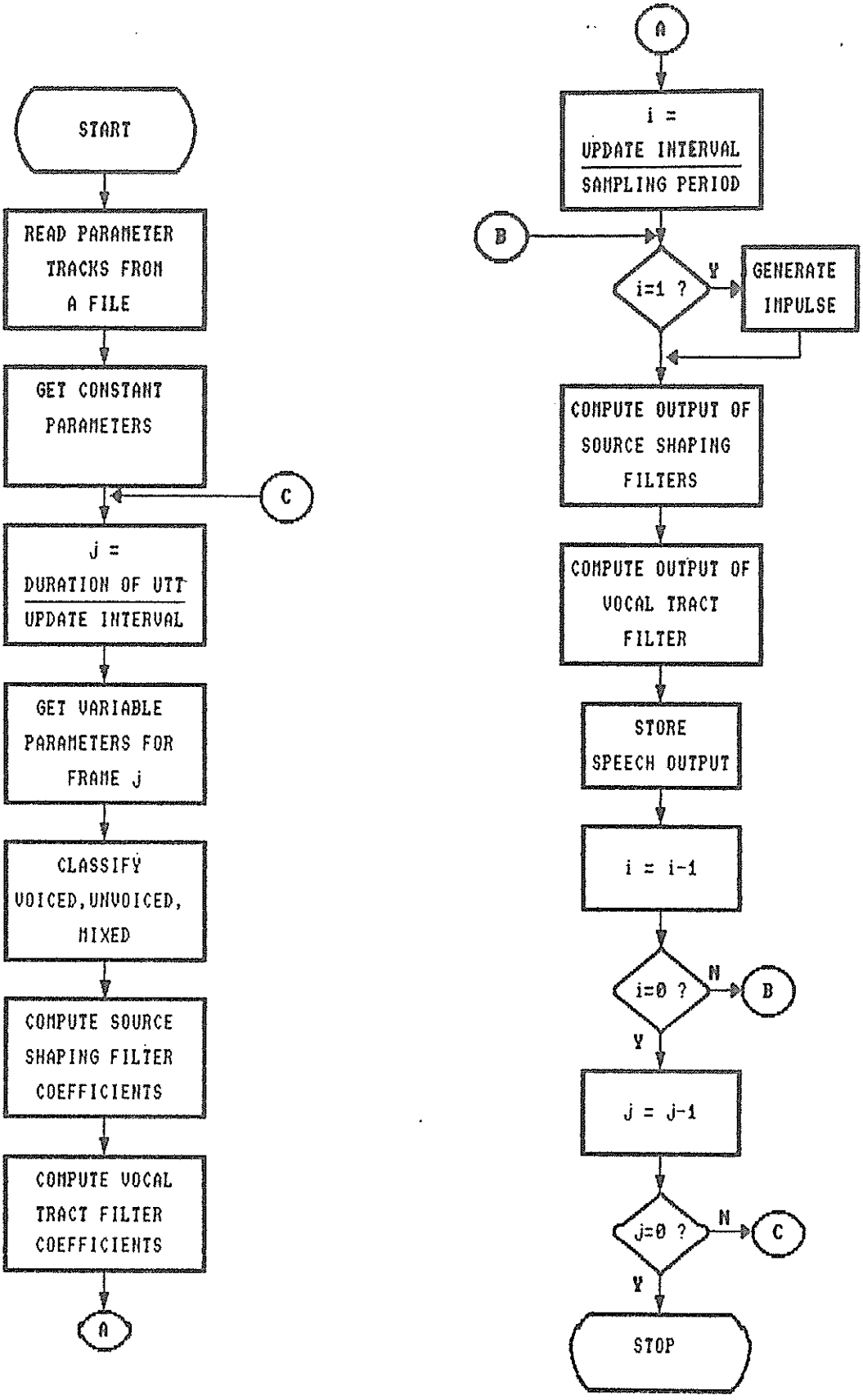


Fig. 3.2. Flow diagram of program PZSYNTH: software implementation of cascade pole-zero synthesizer.

CHAPTER 4
SPEECH ANALYSIS

4.1 INTRODUCTION

Speech analysis capability is needed to extract parameters from the natural speech segments, as well as to verify whether the synthetic waveform has the desired spectral properties. Several speech analysis packages (Klatt, 1980; Pandey, 1987; Sampath, 1990) have been developed. But, these packages run on specific machines with special graphics peripheral, hence are not transportable. Speech analysis and display package, which can run on IBM PC with color graphics adapter, and monochrome or color monitor, was developed. This program (SPAN) will be discussed in this chapter. Source code listing of this program is available in separate volume (Kulkarni, 1992). Results of analysis of natural speech segments carried out for extracting synthesizer parameters are also presented.

4.2 SOFTWARE FOR SPEECH ANALYSIS AND DISPLAY

A program SPAN was written in Pascal for carrying out analysis of digitized speech waveforms analysis of speech waveform. The waveform along with two analysis are displayed on the color graphics monitor.

A segment of user defined length of the waveform is selected at a time and displayed on the upper half of the screen. Analysis is done in the auto mode, i.e., successive segments of specified

window length are selected for analysis. Available analysis facilities are as follows:

(1) First difference of the waveform may be computed to remove any dc component and to give waveform a spectral tilt (emphasize the higher frequency component). This also deemphsizes occasionally present strong fundamental frequency component.

(2) Waveform segment is multiplied by Hamming window.

(3) Log magnitude spectrum: A 512 point FFT of the selected speech segment is computed using FFT procedure available in Turbo Pascal Tool Box. From the real and imaginary parts of the Fourier transform the log magnitude in dB is computed and displayed on the screen below the time waveform.

(4) LP Spectrum : Linear prediction analysis of speech is based on the time varying all-pole linear filter model of speech production. User can specify predictor order. This program uses autocorrelation method of LP analysis, which is described in Appendix C. Durbin's algorithm is used to solve autocorrelation equations. LP spectrum is computed by converting the DFT spectrum of linear prediction coefficients padded with zeros into a log power spectrum and it is overlayed on log magnitude spectrum.

(5) Averaging LP spectrum: For analysis of the speech segments where formants and bandwidths are expected to

remain constant for all frames, e.g., vowels, user can specify this option. In this option point by point averaging of LP spectrum for different frames is performed and final averaged spectrum is displayed. Option of averaging LP spectrum has been incorporated as an additional feature to the earlier developed software.

After the analysis is over, a cursor controlled by arrow keys can be used for locating formant frequencies. As the cursor moves on the screen corresponding frequency and amplitude is displayed on the screen. The co-ordinates of particular location can be stored in an array by pressing 'ENTER' key. Five such peaks can be located. From the known formant frequencies, bandwidths of first three formants are computed from LP spectrum. Two more cursors can be used to mark off consecutive peaks from which pitch for voiced segments can be obtained. Next a 'QUIT' or 'CONTINUE' option is provided. On quitting the results of analysis are displayed. Pitch, formant frequencies, and amplitudes are displayed in tabular as well as graphical form. These parameters can be stored in a file for further processing.

4.3 DIGITIZED SPEECH DATA

For obtaining parameters such as formant frequencies bandwidths, and relative amplitudes, natural speech segments were recorded on a tape recorder. The earphone output of the

recorder was given to a lowpass filter with cutoff frequency of 4.65 kHz in order to avoid aliasing. Lowpass filter is preceded by an amplifier whose gain can be adjusted to 1,2, or 5. Output obtained from a tape recorder is of the order of 200 to 500 mV. This voltage should be amplified to around 2 V so as to utilize full capacity of A/D converter (A/D conversion range was adjusted to -2.5V to +2.5V). Output of the filter was further fed to analog input of data acquisition card PCL208. Speech segments were digitized at a sampling rate of 10 kHz. The digitized data files are then stored on the disk. This card and the driver programs for A/D, D/A conversion are described in Appendix A. Steps in speech analysis are outlined in Fig. 4.1(a) and block diagram of hardware setup is given in Fig. 4.1.(b)

4.4 ANALYSIS RESULTS

Spectrum analysis was carried out for Hindi vowels /अ आ इ ई उ ऊ ए ओ/ (/^ a I i U u e o/) in isolation. Vowels spoken by two male speakers and two female speakers were recorded and digitized for this purpose. Parameters obtained from analysis of vowels for male speakers and female speakers are given in Table 4.1 and Table 4.2 respectively. This data is further used for synthesis of vowels using cascade pole-zero synthesizer.

Table 4.1(a). Control parameters for synthesis of vowels in isolation. These parameters are obtained from the analysis of natural utterances for male speaker 1

Vowel	F1	F2	F3	BW1	BW2	BW3
ʌ	594	1075	2567	118	80	215
ɑ	762	1137	2661	136	100	254
I	302	2223	3037	100	234	567
i	302	2286	3237	60	167	600
U	334	887	2463	80	117	254
u	313	730	2546	45	137	273
e	438	2192	2192	60	156	334
O	459	824	2588	146	128	390

Table 4.1(b). Relative amplitudes in dB of first three formants obtained from the analysis of natural utterances for male speaker 1.

Vowel	A1	A2	A3
ʌ	32.02	22.9	1.37
ɑ	34.46	21.78	1.75
I	20.37	02.75	1.94
i	29.02	02.36	1.7
U	30.05	21.7	1.19
u	33.10	26.45	-0.29
e	22.06	11.07	5.53
O	29.69	27.29	-2.34

Table 4.2(a). Control parameters for synthesis of vowels in isolation. These parameters are obtained from the analysis of natural utterances for male speaker 2.

Vowel	F1	F2	F3	BW1	BW2	BW3
^	594	1252	2432	98	117	273
ɑ	730	1179	2661	78	100	312
I	313	2056	3194	42	234	312
i	313	2192	3267	100	167	312
U	334	908	2166	80	117	254
u	334	845	2546	45	240	360
e	459	1910	2379	60	135	334
O	438	918	2369	200	60	168

Table 4.2(b).Relative amplitudes in dB of first three formants obtained from analysis of natural utterances formale speaker 2

Vowel	A1	A2	A3
^	32 39	20.70	1.22
ɑ	34.48	22.77	0.45
I	22.87	06.78	0.15
i	26.67	07.94	5.58
U	23.31	17.69	0.13
u	32.07	20.18	-0.19
e	23.63	12.75	6.51
O	30.61	29.42	1.13

able 4.3(a). Control parameters for synthesis of vowels in solation. These parameters are obtained from the analysis of atural utterances for female speaker 1.

Vowel	F1	F2	F3	BW1	BW2	BW3
^	657	1273	3152	84	84	500
a	803	1409	3225	312	188	640
I	302	2933	4457	188	220	300
i	280	2912	4436	65	167	600
U	300	1000	3090	50	146	376
u	300	824	3200	85	117	600
e	510	2660	4460	110	470	235
O	490	1000	2235	75	178	300

Table 4.3(b). Relative amplitude in dB of first three formants, obtained from analysis of natural utterances for female speaker 2.

Vowel	A1	A2	A3
^	22.88	10.29	-3.69
a	21.85	19.05	-8.04
I	26.05	6.38	-2.27
i	29.11	8.66	-6.66
U	43.97	22.21	-6.18
u	31.72	29.13	-5.90
e	23.98	6.21	-4.35
O	30.00	17.41	-5.06

Table 4.4(a). Control parameters for synthesis of vowels in isolation. These parameters are obtained from the analysis of natural utterances for female speaker 2.

Vowel	F1	F2	F3	BW1	BW2	BW3
^	657	1450	3256	176	195	330
a	803	1315	3330	230	188	640
I	302	3204	4520	97	126	520
i	240	3110	4592	52	135	220
U	300	1000	3090	50	146	376
u	260	824	3200	85	117	600
e	470	3037	4634	75	386	488
O	490	908	2515	280	60	400

Table 4.4(b). Relative amplitudes in dB of first three formants obtained from analysis of natural utterances for female speaker 2

Vowel	A1	A2	A3
^	24.38	12.97	-4.66
a	26.82	20.49	-8.65
I	28.59	2.47	-5.23
i	36.36	9.24	-4.38
U	35.91	26.83	-9.00
u	31.72	29.13	-5.90
e	30.83	7.29	-6.45
O	30.00	17.41	-5.06

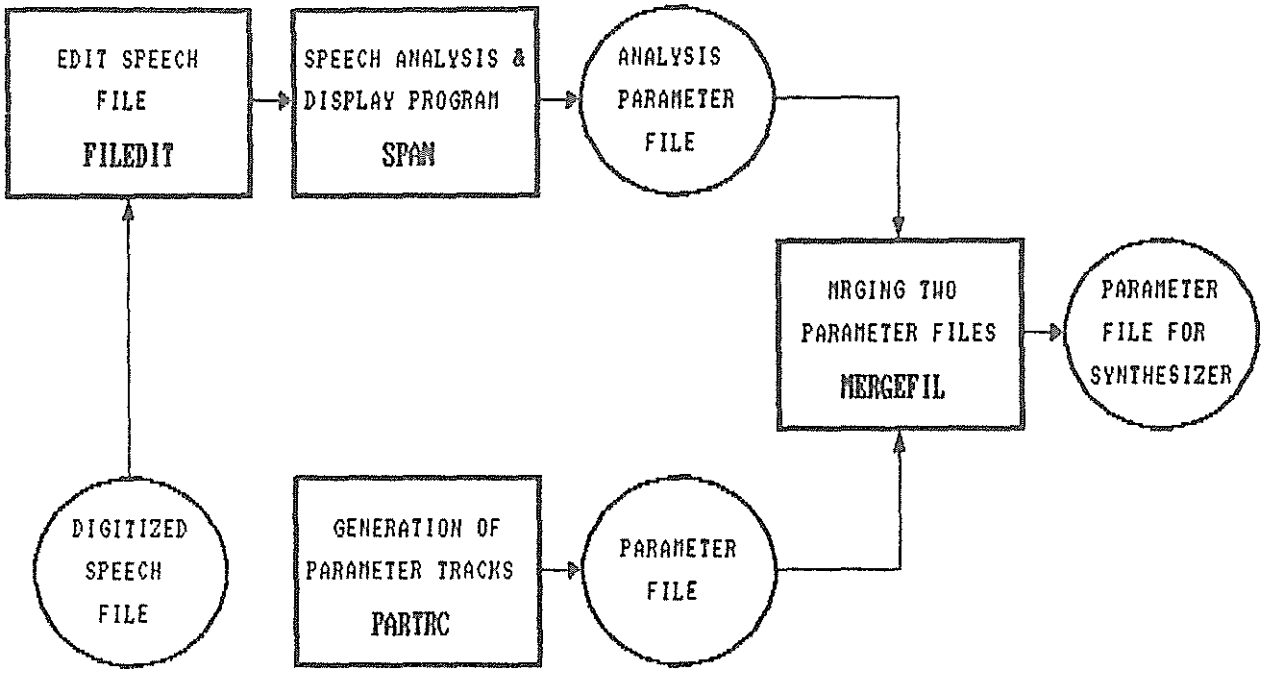


Fig. 4.1(a). Software set-up for analysis of digitized speech.

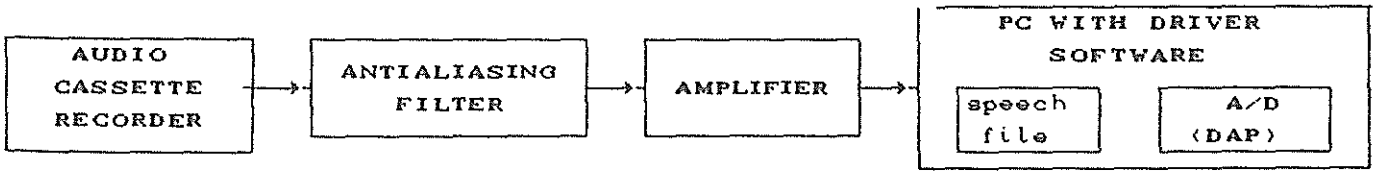


Fig. 4.1(b). Hardware set-up for digitizing natural speech.

CHAPTER 5

PARAMETER TRACK GENERATION

5.1 INTRODUCTION

A user could specify control parameter tracks for a speech utterance to be synthesized, by typing in a sequence of points for each of the variable control parameter, and have the program draw straight lines between them. However this method is time consuming and subject to error if no visual feedback in terms of a time plot of parameter values is provided. A program PARTRC (in Pascal) was developed for graphical editing of parameter tracks, for the speech synthesizer program PZSYNTH described in Chapter 3. Source code listing of this program is available in separate volume (Kulkarni, 1992). This program along with a program for merging two parameter files will be discussed in this chapter.

5.2 PROGRAM FOR GRAPHICAL GENERATION OF THE PARAMETER TRACKS

Program PARTRC displays variable parameters as function of time and its track can be edited. Editing operations are carried out by moving a cursor around the screen. Commands can be given with the help of a set of function keys for storing and erasing points, inserting and deleting line segments, etc. Parameter track is displayed as a set of straight lines through the stored points. When a point is stored or deleted only the parameter track being edited is changed. The position of the cursor, time, and parameter scales, etc are displayed on the screen.

The program is capable of generating parameter tracks for cascade/parallel synthesizer (Klatt synthesizer), as well as cascade pole-zero synthesizer. User has to select required option from the startup menu. Having selected the option from the first menu, it displays next menu providing following options

(1) Use inbuilt configuration

(2) Use phoneme set option

(3) Modify parameter tracks

Once an option from the second menu is selected, 39 control parameters are displayed in a tabular form as shown in Fig. 5.1. This table provides information such as: maximum, minimum, and default value of the parameter, whether a parameter is a variable or a constant, etc. User can specify value of a constant parameter, as well as status (variable or constant) of a parameter. In the first option, user has to specify values of all parameters. In the second option, after the table editing is over, user has to specify number of phonemes, name of phonemes, and their durations. Program uses parameters for various phonems stored as a part of the program.Selection of third option allows to modify the tracks stored in a file. Another menu, which gives following options is displayed on the screen.

(1) Display/modify all parameter tracks

(2) Display/modify single parameter track

(3) Store parameter tracks

Editing previously stored parameter tracks is thus possible. New

parameter tracks are then displayed as straight line interpolation between modified points. By selecting second option, user can specify parameter trajectories for some parameters such as pitch or amplitude of voicing. Such a facility will be useful if slight modifications are required to be done for some parameters, as it displays only specified parameter tracks. At the end of all editing operations, parameters can be saved in a new or an old file. All the editing operations are performed with the help of function keys. Various possible editing operations are as follows:

(1) STP (store point): Stores co-ordinates of a specified point in an array. Parameter track is then displayes as straight line interpolation between this point and its preceding and succeding points.

(2) ERP (erase point): Any point, which is stored using STP command, can be erased using this command. Now the parameter track is a linear interpolation between the points preceding and succeding to the erased point.

(3) CHT (change time width): Allows user to view only a part of the trajectory. New time width is user specified having a value less than the duration of utterance.

(4) INS (insert time segment): Inserts a time segment of 5 ms and effectively increases the duration of utterance. All the tracks will be calculated for newly inserted segment as interpolation between previous points and

inserted segment.

(5) DEL (delete time segment): Pressing the corresponding function key once, delets a time segment of 5 ms, and effectively decreases the duration of utterance.

(6) KON, KOF: Function key help can be turned on or off respectively using these function keys.

Function keys and corresponding editing operations are listed in Table 5.1.

5.3 MERGING PARAMETER FILES

Linear interpolation provides the simplest way to generate a parameter track, using targets for two successive sounds. Linear transitions involving amplitudes (e.g., the intensity of voicing or frication) are frequently adequate. But acoustic theory indicates that formant frequencies must always change slowly, continuously, and slope discontinuities at boundaries should be avoided. Hence in a parameter file obtained from PARTRC, if formant data is replaced by formant tracks obtained from analysis of natural speech data, then synthetic speech will be more natural. A program MERGEFILE was written in PASCAL for this purpose. This program asks for the name of the parameter file generated by program PARTRC and parameter file obtained from analysis of natural speech segments. Analysis program gives formant data for every 25 ms, while update rate is 5 ms. Progrm linearly interpolates the values of the formants using target

frequencies specified at 25 ms. Calculated new formant tracks are written in a new file. Rest of the parameters, which are not specified in parameter file obtained from analysis, are transferred from old file to the new file without any change. This program makes it convinient to utilize the parameters obtained from analysis for resynthesis.

Table 5.1. Function keys for desired operations for graphical generation of the parameter tracks.

Function key	Operation	Description
F1	STP	STORE POINT
F2	CHT	CHANGE TIME SCALE
F3	ERP	ERASE POINT
F4	CHS	CHANGE PARAMETER STEP SIZE
F5	DEL	DELETE TIME SEGMENT
F6	INS	INSERT TIME SEGMENT
F7	KON	FUNCTION KEYS HELP ON
F8	KOF	FUNCTION KEYS HELP OFF
F10	END EDIT	END OF EDITING
Pg up	NEXT TRACK	NEXT TRACK DISPLAYED
Pg Dn	PREVIOUS TRACK	PREVIOUS TRACK DISPLAYED

```
Synthesizer Control Parameters
Format is Parameter, V/C, Min, Max & Default Values.

NF    C  4     6     4      FZ1   C  0     500    0      F4    C  2500  4500  3300
NZ    C  0     5     0      FZ2   C  0     2000   0      F5    C  3500  4900  3750
FO    V  0     500   0      FZ3   C  0     3000   0      F6    C  4000  4999  4900
AV    V  0     80    0      FZ4   C  0     4000   0      BW4   C  100   500   250
AF    C  0     80    0      BZ1   C  0     800    0      BW5   C  150   700   200
AVS   C  0     80    0      BZ2   C  0     800    0      BW6   C  200   2000  1000
AH    C  0     80    0      BZ3   C  0     800    0      BWNZ  C  50    500   100
F1    V  150   900   450    BZ4   C  0     800    0      BWNP  C  50    500   100
F2    V  500   3500  1450   FNZ   C  200   500    250    FZ5   C  1000  4900  4900
F3    V  1300  4500  2400   FNP   C  200   500    250    BZ5   C  50    2000  100
BW1   V  40    500   50     UPDT  C  2     50     5      FGZ   C  0     5000  1500
BW2   V  40    500   70     SR    C  5000  20000  10000  BWGZ  C  100   9000  6000
BW3   V  40    500   110    GO    C  0     80     47     AB    C  0     80    0

    End editing ?   <Y/N> : y

    Duration of utterence in ms < 5..2000 > : 500

  : Curser Control   Back Sp, Del : Delete char   F10 : End edit
```

Fig. 5.1. Display of 39 control parameters, for cascade pole-zero synthesizer, as given by program PZSYNTH.

CHAPTER 6

SYNTHESIS USING CASCADE POLE-ZERO SYNTHESIZER

6.1 INTRODUCTION

Synthesis of various speech segments were carried out using program PZSYNTH, along with speech analysis and display package described in Chapter 4 and software for graphically generating and editing parameter tracks described in Chapter 5. In this chapter general procedure adopted for synthesis is outlined, also strategies for synthesis of various classes of sounds are presented.

6.2 STEPS IN SPEECH SYNTHESIS

General procedure for synthesis using cascade pole-zero synthesizer is as follows:

(1) Parameter tracks were generated using program PARTRC, developed for graphically generating and editing parameter tracks. Target values of parameters listed in Table 3.3 and Table 3.4 were used. Duration of utterance and pitch variation was specified as per requirements. Excitation amplitude controls AV, AF, and AVS were used to adjust overall intensity contour and mixture of periodic voicing to aperiodic noise.

(2) Speech file was created using program PZSYNTH.

(3) Informal listening tests were carried out using hardware setup given in Fig. 6.1(b).

(4) Using spectral analysis and display program (SPAN) it is verified whether synthesized speech has desired spectral characteristics

(5) If quality of the synthesized speech is not as per requirement, necessary changes were made in the parameter specifications and the procedure was repeated.

Several iterations and and adjustments were carried out to obtain perceptual distinctiveness and quality in synthesized samples. Steps in speech synthesis are outlined in Fig. 6.1(a) and hardware setup for carrying out informal listening tests for synthesized speech is given in Fig. 6.1(b).

6.3 SYNTHESIS STRATEGIES AND RESULTS

Following sections discuss strategies for synthesis of various classes of speech sounds. Parameter tracks derived, as a result of several iterations and adjustments carried out to obtain perceptual distinctiveness and quality in synthesized speech, are also presented.

6.3.1 Synthesis of vowels

The production of steady state vowels is described in terms of vocal tract shape, which provides targets when vowels

are articulated in words. The perception of vowels can be usually interpreted in terms of location of first two formants (F1,F2), and a systematic variation of F1 and F2 in synthetic vowel stimuli is sufficient to create distinguishable vowels (o'shaughnessy, 1987).

Control parameters that are varied to generate isolated vowels are: amplitude of voicing (AV), fundamental frequency (F0), lowest three formants (F1,F2, and F3), and bandwidths (BW1,BW2, and BW3). To create natural breathy vowel termination, amplitude of aspiration (AH) and amplitude of quasi-sinusoidal voicing (AVS) can be activated. Table 4.1 and Table 4.2 lists target values for fundamental frequencies and bandwidths for two male speakers. Table 4.3 and Table 4.4 lists target values for formants frequencies and bandwidths for two female speakers. These parameters were obtained from analysis of samples of natural speech for Hindi vowels. Vowels /^ a I i U u e o/ were synthesized for male as well as female speaker. The synthesis duration was 500 ms. Voicing amplitude (AV) was increased from 0 dB to 60 dB in initial 100 ms, retained at 60 dB for next 300 ms, and decreased to 0 dB in last 100 ms, as shown in Fig. 6.2. F0 was kept constant at 140 Hz for male speech, and 200 Hz for female speech.

6.3.2 Synthesis of semivowels

Glides /w j/ were synthesized in VCV syllable with vowel /a/. Synthesis control parameters AV and F0 were similar to those of vowels and formant trajectories were as given in Fig. 6.3. Liquids /r/ and /l/ were synthesized similarly, except that the formant transitions were faster and voicing amplitude was decreased by 10 dB. But perfect discrimination for /r/ and /l/ was not obtained.

6.3.3 Synthesis of fricatives

Distinguishing among the fricatives is mostly based on presence of voicing, amplitude of frication noise, and duration of frication. Formant transitions to and from fricatives provide secondary cues, due to co-articulation of the fricatives with the adjacent phonemes. Target values for variable control parameters for fricatives derived from parameter data given by Klatt for cascade/parallel synthesizer, listed in Table 3.1 were used. For synthesis of unvoiced fricatives value of AV and AF were kept at around 0 dB and 50 dB. Voiced fricatives were synthesized using two sources, periodic glottal pulses and frication noise generated at vocal tract constriction. Hence parameter values of AV and AF were kept at 47 dB and 50 dB respectively. While specifying parameter trajectories for fricatives following points were taken into consideration.

(1) In CV syllables voicing begins during frication, while

voicing is delayed until the offset of frication in voiceless syllables.

(2) Pitch varies with voicing in natural speech. F0 is at its highest value at the onset of vocal cord vibrations, following voiceless consonants, while F0 rises slowly in case of voiced fricatives (Cole & Cooper). Hence for synthesis of /z/ and /ʒ/, F0 was raised from 130 Hz to 160 Hz after onset of vocal cord vibration.

(3) Strong amplitude of frication noise is present for rear places (/s ʃ/), while weak frication signal specifies for front place of articulation (/θ f/). Therefore amplitude of frication was set to 45 dB for synthesis of /s ʃ/ and 35 dB for synthesis of /f θ/.

(4) Shortening the /s/ frication changes its perception to to /z/. By adding frication changes perception of /z/ to /s/ (Domnic & Michail, 1976).

Fricatives /s ʃ z ʒ v θ/ were synthesized in VCV context for vowel /ɑ/. Parameter tracks for some of the fricatives are given in Fig. 6.4. When listening tests were carried out fricatives /s ʃ z ʒ/ were clearly distinguished while there was confusion between /f/ and /v/.

6.3.4 Synthesis of affricates

Affricate parameters given in Table 3.4 refer to fricative portion of affricates. Affricates /dʒ dʒ^{h} tʃ tʃ^{h}/ were synthesized using these parameters. Parameter tracks for /tʃ tʃ^{h}/ are given in Fig. 6.5.

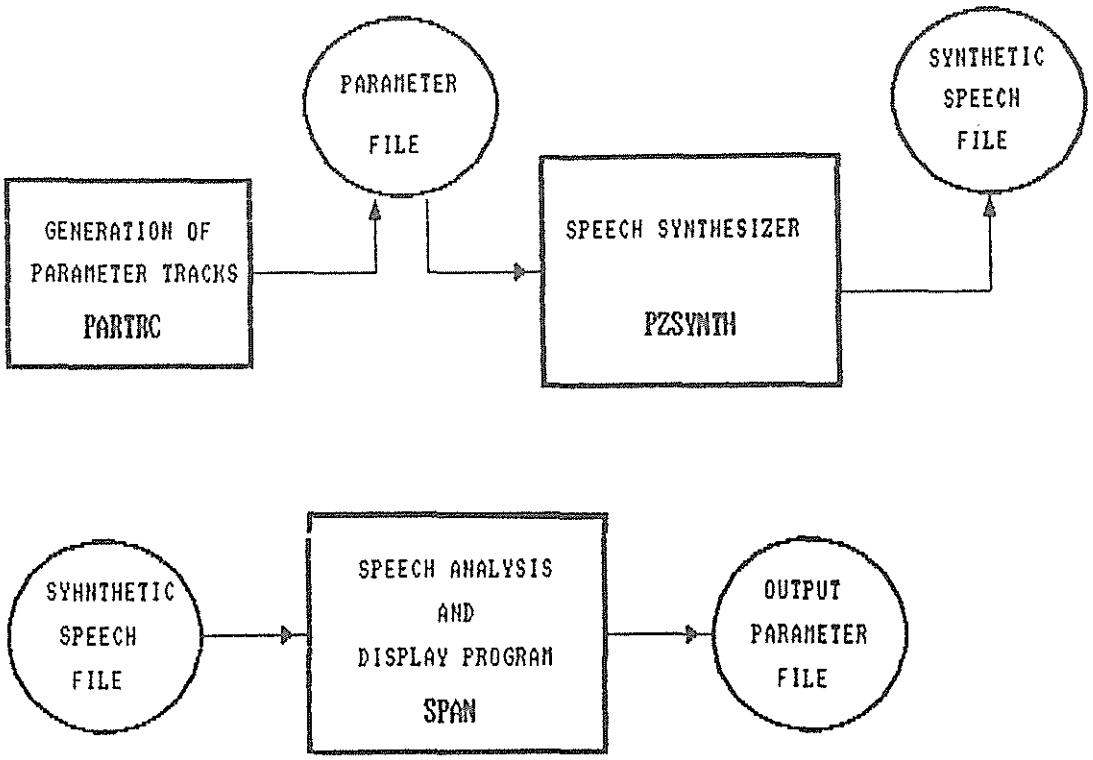


Fig. 6.1(a). Outline of the steps in speech synthesis (software set-up).

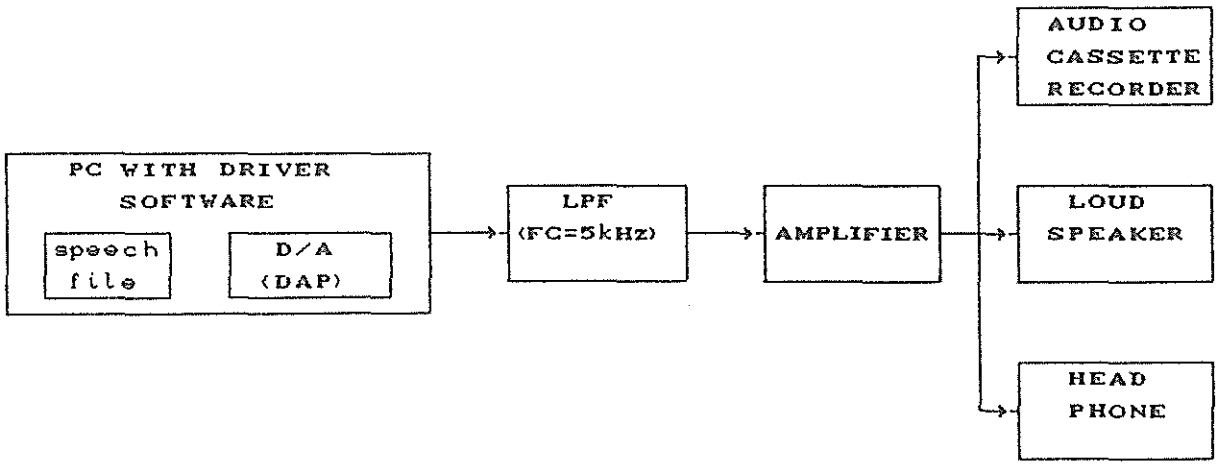


Fig. 6.1(b). Hardware set-up for listening to synthetic speech.

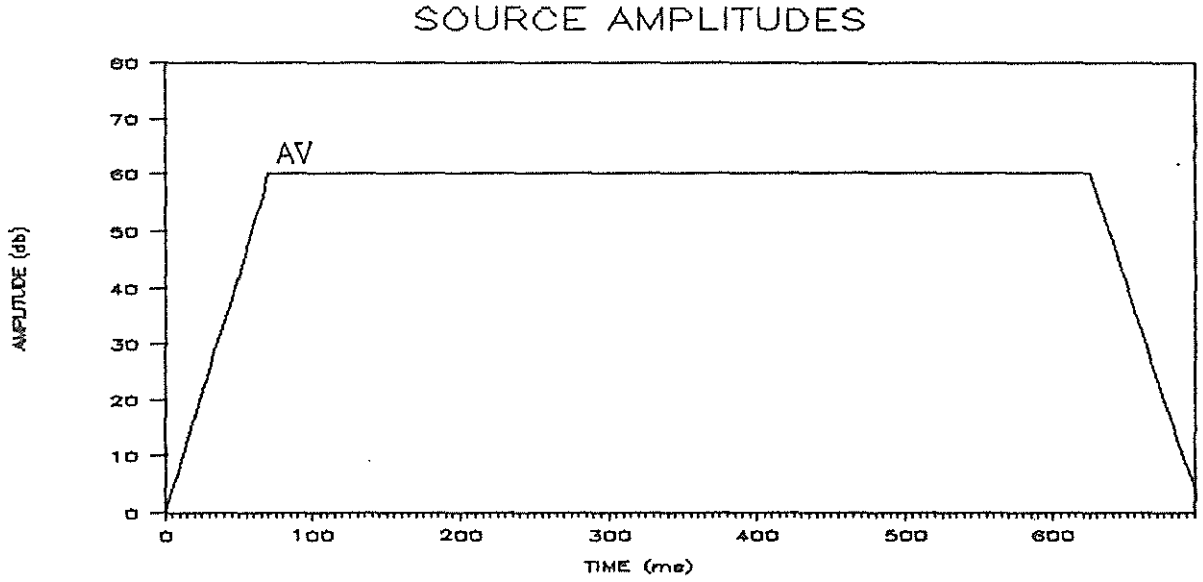


Fig. 6.2. Parameter track for voicing amplitude for vowels and glides in vocalic context.

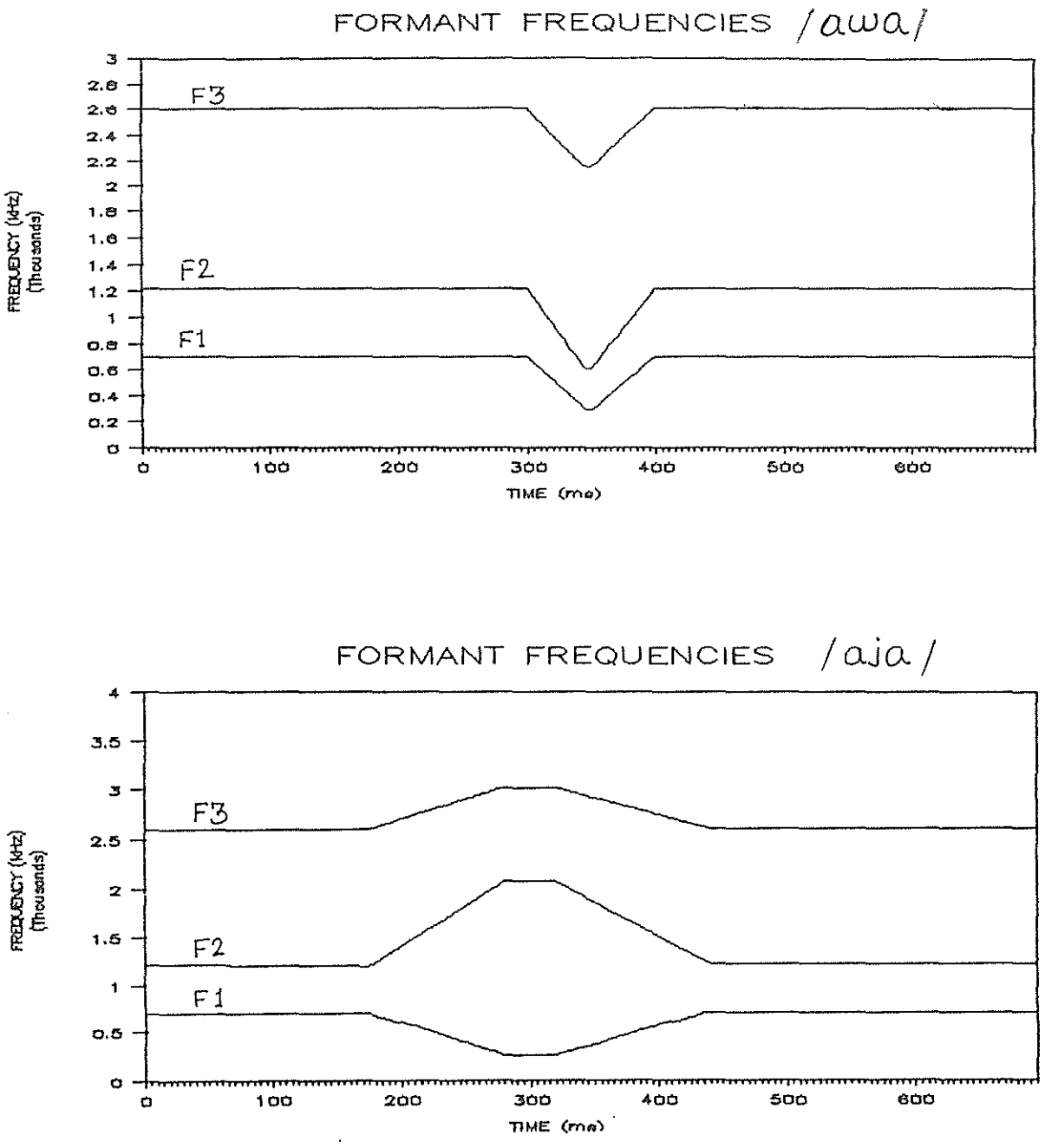
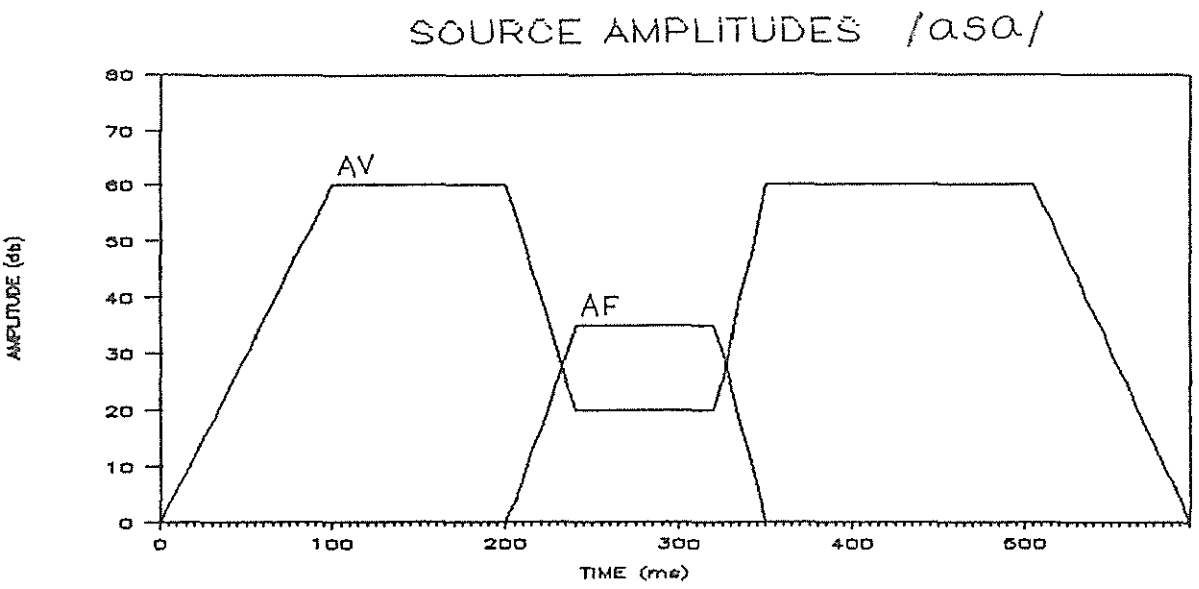


Fig. 6.3. Parameter tracks for formant frequencies for glides in vocalic context: /awa aja/.



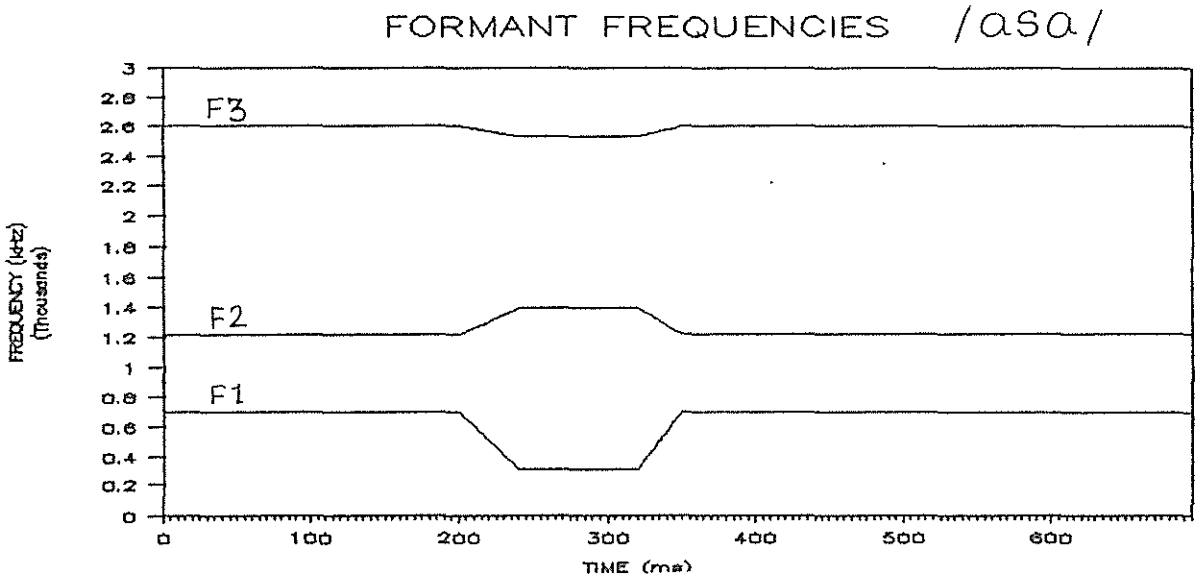
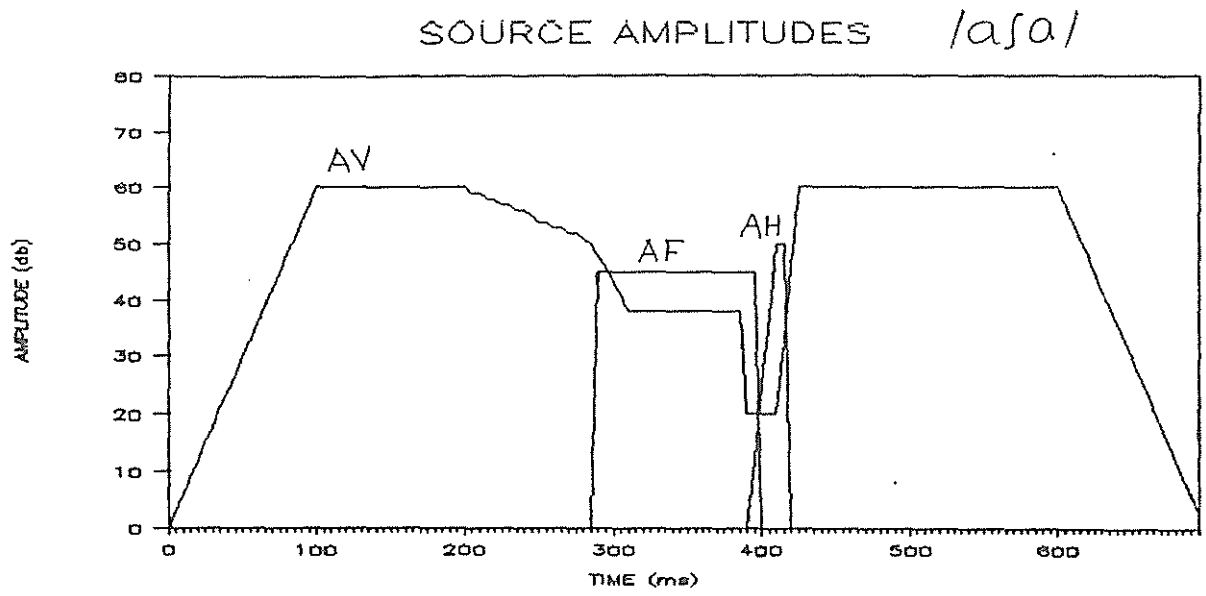


Fig. 6.4. Parameter tracks for excitation source amplitudes and formant frequencies for fricatives in vocalic context: /asa afa aza aza/.



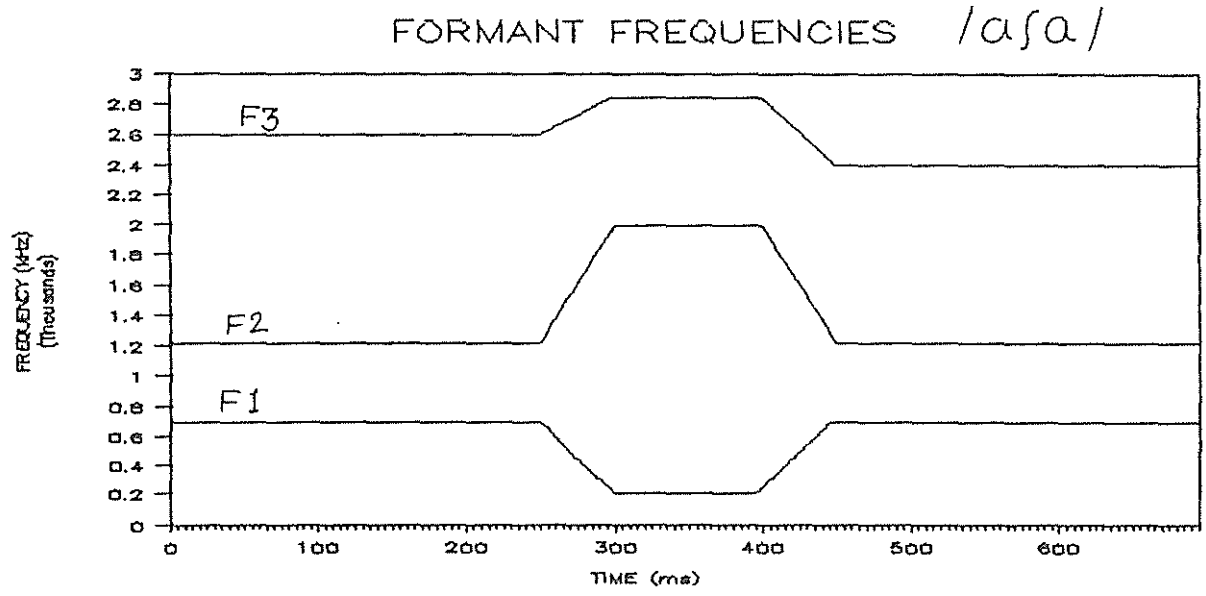
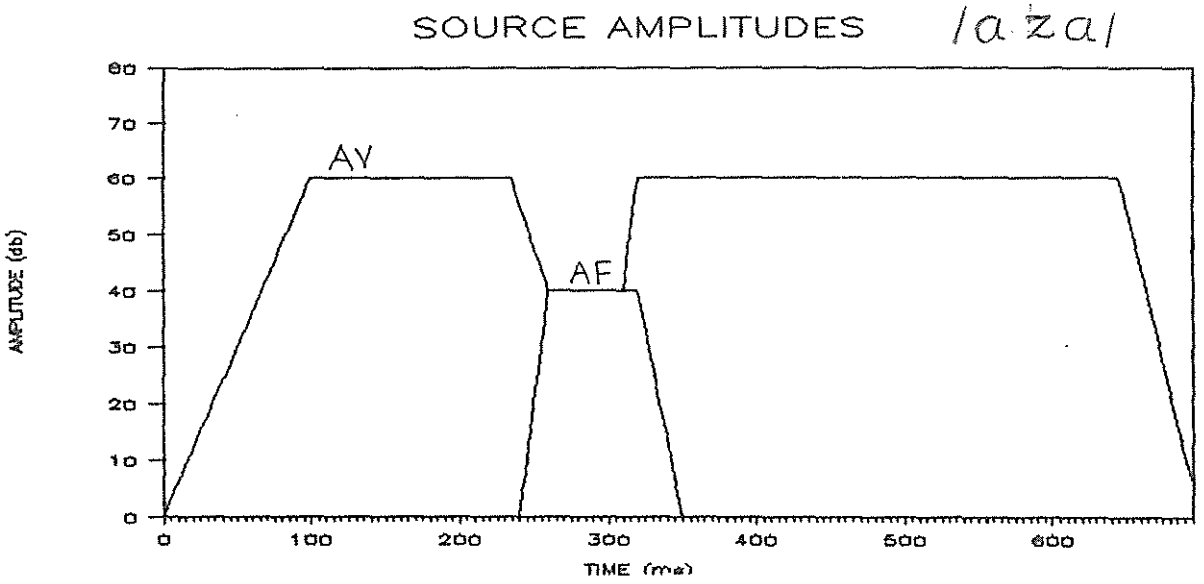


Fig. 6.4. (continued)



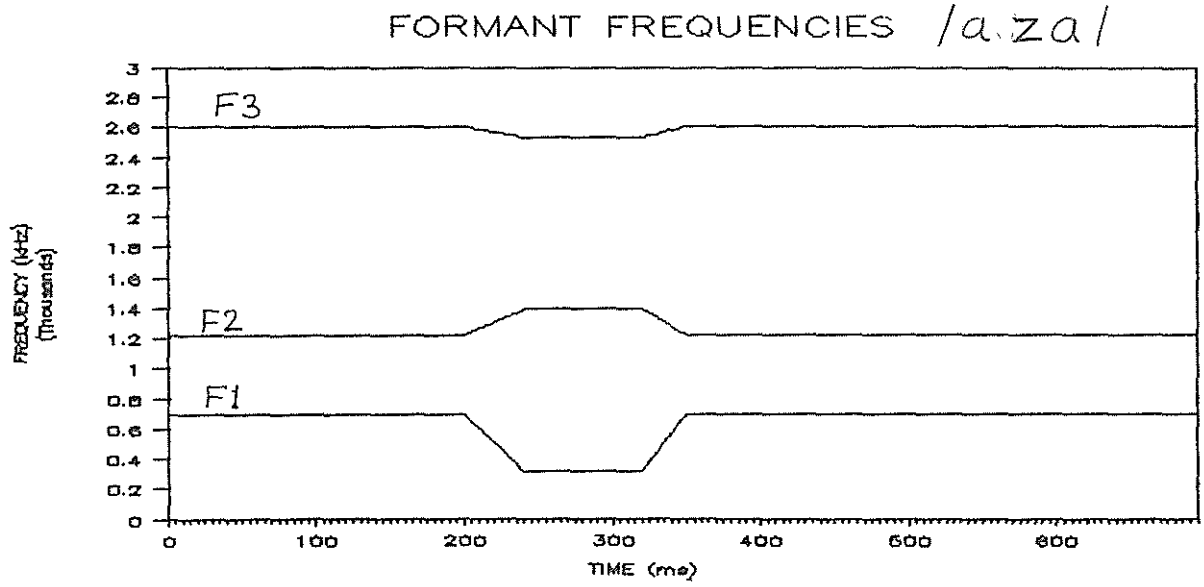


Fig. 6.4. (continued)

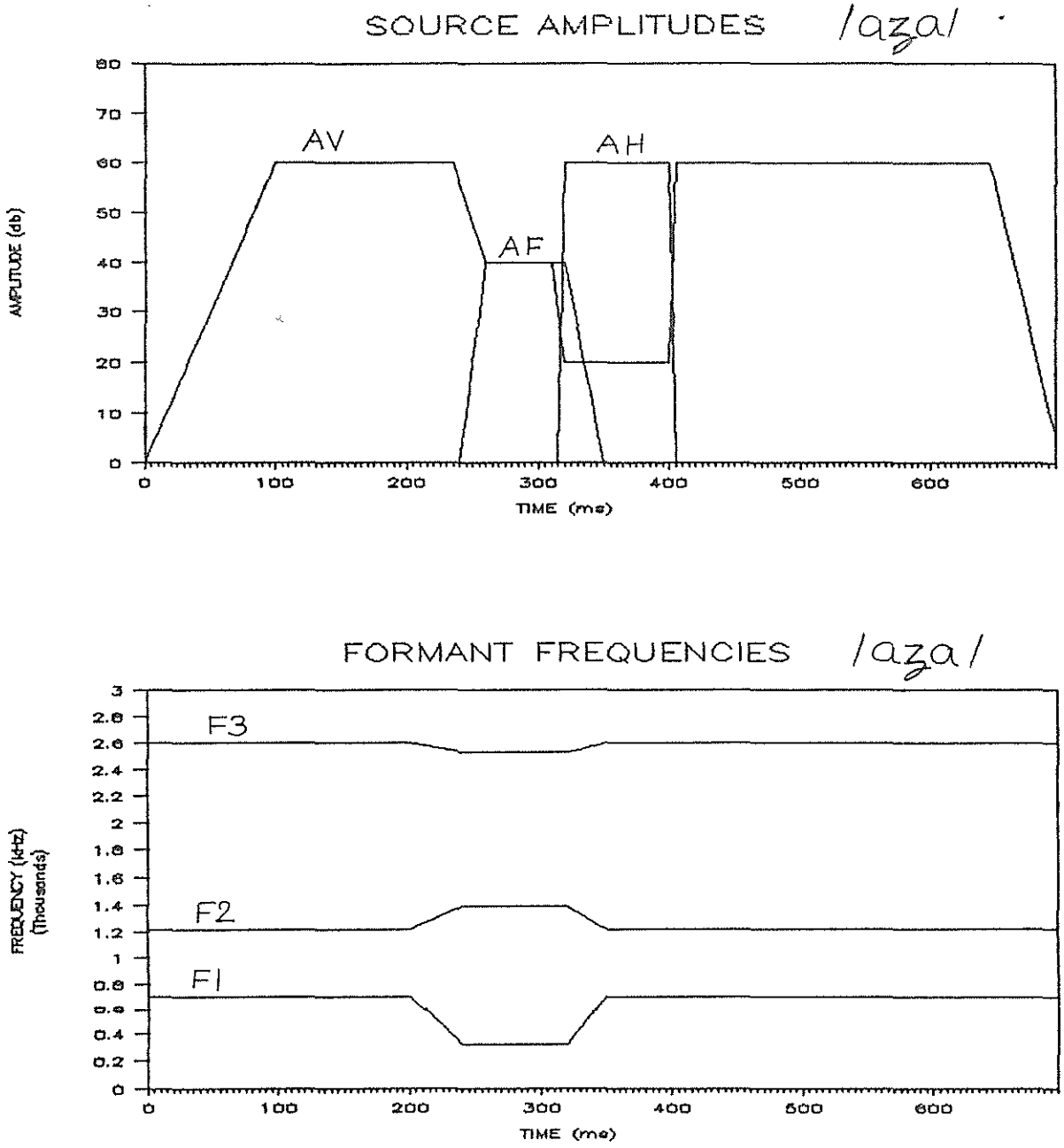


Fig. 6.4. (continued)

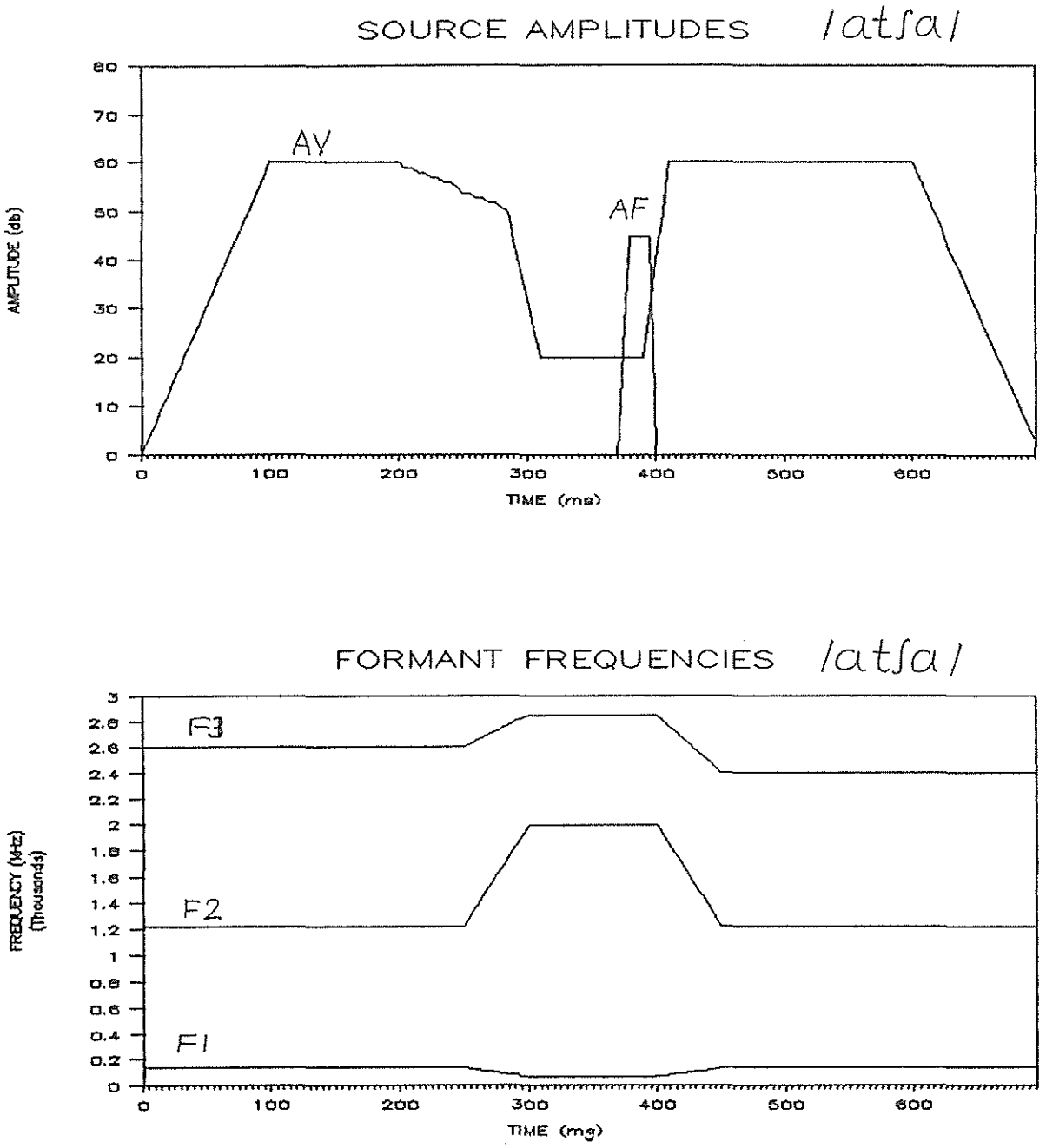


Fig. 6.5. Parameter tracks for excitation source amplitudes and formant frequencies for affricates in vocalic context: /atʃa atʃʰa/.

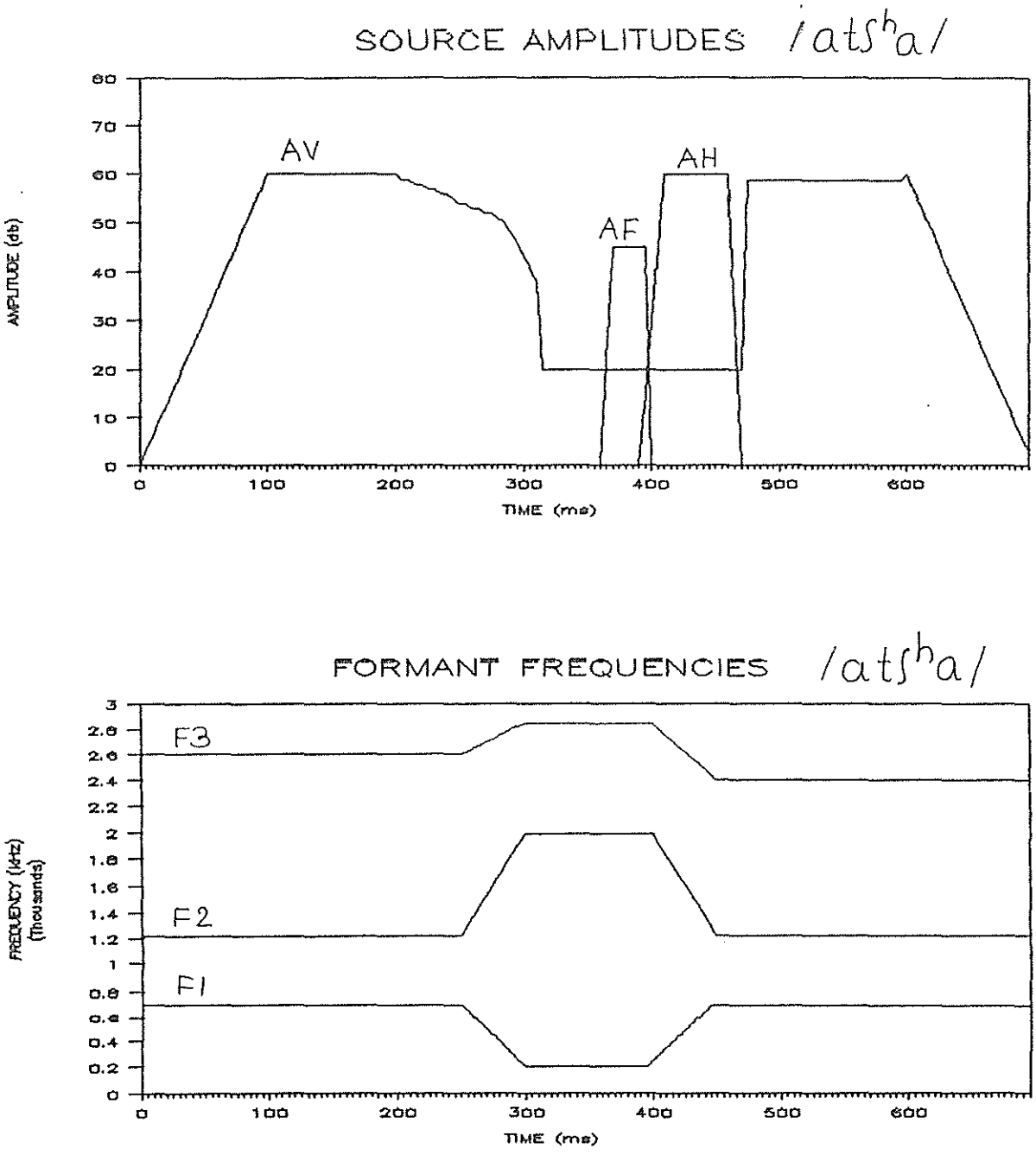


Fig. 6.5. (continued)

CHAPTER 7
SUMMARY AND SUGGESTIONS FOR FUTURE WORK

7.1 INTRODUCTION

A software based speech synthesizer, which uses cascade pole-zero model of the vocal tract, has been developed in this project. In this chapter, work done is summarized and some suggestions for further development are made.

7.2 WORK DONE

A software based pole-zero cascade synthesizer which uses a cascade model for synthesis of vowels and and employs anti-resonators to simulate zeros in the spectra of speech segments with frication was developed. Main application of this synthesizer will be for testing and calibrating various hearing aids and sensory aids for the deaf. This synthesizer gives close control over characteristics of the speech output by providing flexibility to alter its parameters as per requirement.

For testing the validity of the speech synthesis model, availability of accurate model parameters is of utmost importance. This is accomplished by developing a speech analysis software, which plots log magnitude spectrum as well as LP spectrum, hence facilitates the extraction of the parameters such as formant frequencies and bandwidths. For synthesis of

consonants, additional parameters such as zero frequencies and bandwidths are needed. This is achieved by writing a software which computes the required parameters from the data given by Klatt (1980) for cascade/parallel synthesizer. This essentially involves use of numerical techniques for polynomial root solving.

Vowels /^ ɑ I i U u e o/ were synthesized from the parameter data obtained from the analysis of digitized natural utterances for male as well as a female speaker. This analysis was carried out for Hindi phonemes and the approach can be extended to other Indian languages. Consonants /s ʃ z ʒ dʒ $dz_{ʒ}^{h}$ tʃ $tʃ^{h}$ m/, glides /w j/, and whisper /h/ was synthesized in VCV context for vowel /ɑ/. Parameters for these were obtained from the parameters for English phonemes given by Klatt. Informal listening tests were carried out to test discrimination between various consonants.

A program for graphical generation of the parameter tracks (For IBM PC), which displays a variable parameter as a function of time and provides facility to edit the track, was developed.

7.3 SUGGESTIONS FOR FURTHER WORK

Speech synthesis can be viewed as a mechanism for evaluation of the model parameters extracted from the analysis of the speech signal. Pole-zero model has been tested by obtaining parameters from the information given by Klatt for cascade/parallel

synthesizer. Next step should be development of a scheme for extraction of parameters for cascade pole-zero model.

Spectrum matching techniques should be used for pole-zero analysis of natural speech segments. The fitting procedure is initiated by guessing a set of poles and zeros appropriate to calculate real speech spectrum, from the knowledge of their approximate range. Short time spectrum of the digitized speech is computed. Calculated spectrum is then fitted by synthetic spectrum in successive approximation, according to weighted least square error criterion. A weighing function tends to position the poles and zeros so as to fit fluctuations in the speech spectrum. Logarithmic measure of amplitude is chosen to assure equal sensitivities to the movements of poles and zeros. Frequency and damping of individual pole and zero is successively incremented. If good fitting is not obtained even after 10 to 20 iterations, results are examined visually, suitable adjustments in pole-zero pattern are made, and another set of cycle is commenced.

Another improvement that can be made is, incorporation of a variable window length for the analysis of VCV sequences. This is because frame length required for voiced sounds is twice or thrice the pitch period. If same frame length is used for stops, it will cause averaging into the voiced portion following or silence preceding the burst release.

REFERENCES

Biswas S (1991). Effect of Sinusoidal Magnetic Field on K562 and Hybrid Cell Lines, M.Tech. thesis, School of Biomedical Engineering, I.I.T. Bombay.

Borland (1987). Turbo Pascal Owner's Handbook (Ver 4). Scotts Valley: Borland International.

Chafekar SA (1990). Speech Synthesis for the Testing of Sensory Aids for the Hearing Impaired, M.Tech. thesis, School of Biomedical Engineering, I.I.T. Bombay.

Coker C (1976). A model of articulatory dynamics and control. Proc. IEEE, vol 64, pp 452-460.

Cole R & Cooper W (1975). Perception of voicing in English affricates and fricatives. J. Acoust. Soc. Am., vol 58, pp 1280-1287.

DMS (1990). PCL208 Data Aquisition Card User's Manual, Bombay: Dynalog Micro-Systems.

Domnic WM & Michail MC (1976). Contribution of fundamental frequency and voicing onset time to /zi/,/si/ distinction. J. Acoust. Soc. Am., vol 60, pp 704-717.

Dorman MF (1980). Distribution of acoustic cues for stop consonants place of articulation in VCV syllables. J. Acoust. Soc. Am., vol 67, pp 1333-1335.

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

Gold B & Rabiner LR (1968). Analysis of digital and analog formant synthesizers. IEEE Trans. Audio & Electro-acoustics, vol AU-16, pp 81-94.

Hovanessian SA & Pipes LA (1969). Digital Computer Methods in Engineering. New York, McGraw-Hill.

Kenneth NS & Klatt DH (1974). Transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Am., vol 55(3), pp 653-659.

Klatt DH (1980). A software for cascade/parallel formant synthesizer. J. Acoust. Soc. Am., vol 67(3), pp 971-995.

Kulkarni DM (1992). Source code listing of the programs, Developed as a part of M. Tech. project, Department of Electrical Engineering, I.I.T. Bombay.

Ladefoged P (1982). A Course in Phonetics. New York: Harcourt Brace Jovano.

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

Markel JD & Gray AH (1976). Linear Prediction of Speech. New York: Springer-Verlag.

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

Pandey PC (1987). Speech Processing for Cochlear Prosthesis. Ph.D. Thesis, Department of Electrical Engineering, University of Toronto.

Rabiner LR (1968). Digital formant synthesizer for speech synthesis systems. J. Acoust. Soc. Am., vol 43(4), pp 822-828.

Rabiner LR & Shafer RW (1978). Digital Processing of Speech Signals. Englewood Cliffs, New Jersey: Prentice-Hall.

Sampath A (1991). Speech Parameters for Formant Based Synthesis. B.Tech Project Report, Department of Electrical Engineering, I.I.T. Bombay.

Sebastian G (1991). A Speech Training Aid for the Hearing Impaired, B.Tech Project Report, Department of Electrical Engineering, I.I.T. Bombay.

Stevens KN & Blumstein SE (1978). Invariant cues for place of articulation in stop consonants. J. Acoust. Soc. Am., vol 64 pp 1358-1365.

APPENDIX A
SIGNAL HANDLING

A.1 INTRODUCTION

Development of computer aided signal handling system becomes necessary as, A/D conversion is required while digitizing recorded natural speech segments and D/A conversion is required to produce analog speech output from synthesized speech output.

The following sections will describe the hardware, software, and functional operations of this system.

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 TOOLS

The PCL208 has provided software driver routines which can be accessed by BASIC CALL statements. But it was observed that programming using these driver routines give maximum conversion rate of 6 kHz for D/A conversion. For speech output D/A conversion rate required is 10 kHz. Hence assembler subroutines were written for controlling A/D, D/A operations. These routines were linked to program written in PASCAL. Since these routines handle I/O address directly the conversion rate was increased. the maximum conversion rate obtained was 50 kHz which is maximum possible rate for PCL208.

The signal handling task was carried out with the using a program AD_DA.PAS linked to assembly routines AD.ASM for A/D conversion and DA.ASM for D/A conversion. D/A conversion was carried out using the method of D/A conversion on A/D interrupt.

In this program user can select A/D conversion or D/A conversion option. If D/A conversion option is selected, program prompts for the data file containing digitized speech data. Then it displays total number of samples in the data files, scales the samples such that sample values are limited in the range (-2048, +2048), if 'scale' option is selected. The user can specify sampling frequency maximum upto 50 kHz. Data can be edited nondestructively, i.e, user can specify sections of data from files to be played back. This can be done by giving the value of start sample number and the end sample number. The gap between two consecutive presentations can also be specified by the user. The main program then calls assembler routine DA.ASM which controls D/A conversion. After D/A conversion was over program asks whether to repeat presentations, or to edit another section, or to start with a new file. If A/D conversion option is selected from the first menu, program prompts for the sampling frequency and number of samples. Assembly routine AD.ASM is then called to control A/D conversion. After A/D conversion is complete, program asks if digital data is to be tested by A/D conversion. Then data file name can be specified to store digitized speech samples.

This program was used for recording externally generated segments at specified sampling rate and recorded segments were stored on the disk as text files. The digitized natural speech

data or synthesized speech data were played back for listening tests. This analog output can also be displayed on CRO or used for analysis by signal analyzer.

Other software tools used include, software for editing data stored in a file (Program FILEDIT) and software for plotting data stored in two separate files, so as to view two speech segments simultaneously (program PLOT).

APPENDIX B

LIN BAIRSTOW ALGORITHM FOR POLYNOMIAL ROOT SOLVING

This method is used to compute complex roots of the n^{th} order polynomial (Hovanessian, 1969). Polynomial is of the form

$$P_n(x) = A_n x^n + A_{n-1} x^{n-1} + A_{n-2} x^{n-2} + \ldots + A_2 x^2 + A_1 x + A_o \qquad (B.1)$$

It is desired to extract a quadratic factor $(x^2 + r^* x + s^*)$ from this polynomial.

$$P_n(x) = (x^2 + rx + s)(B_n x^{n-2} + B_{n-1} x^{n-3} + B_{n-2} x^{n-4} + \ldots + B_3 x + B_2) + Rx + S \qquad (B.2)$$

Where $(x^2 + rx + s)$ is a trial polynomial and

$$\begin{aligned} A_n &= B_n \\ A_{n-1} &= B_{n-1} \\ A_{n-2} &= B_{n-2} + rB_{n-1} + sB_n \\ &\vdots \\ A_1 &= R + rB_2 + sB_3 \\ A_o &= s + sB_2 \end{aligned} \qquad (B.3)$$

Assume that $x^2 + r^* x + s^*$ is a factor of polynomial $P_n(x)$, then

$$R(r^*, s^*) = 0 \quad \text{and} \quad S(r^*, s^*) = 0$$

Now let

$$r^* = r + \Delta r \quad \text{and} \quad s^* = s + \Delta s \tag{B.4}$$

Then using Taylor series expansion we get

$$\begin{aligned} R(r^* + s^*) &= R(r,s) + \Delta r \frac{\partial R}{\partial r} + \Delta s \frac{\partial R}{\partial s} = 0 \\ S(r^* + s^*) &= S(r,s) + \Delta r \frac{\partial S}{\partial r} + \Delta s \frac{\partial S}{\partial s} = 0 \end{aligned} \tag{B.5}$$

Compute Δr and Δs using Eq. B.3 and B.5. Values of r and s at n^{th} iteration, and roots corresponding to that quadratic factors are given by following equations

$$\begin{aligned} r^n &= r^{n-1} + \Delta r \\ s^n &= s^{n-1} + \Delta s \end{aligned} \tag{B.6}$$

$$\text{root} = \frac{-r \pm (r^2 - 4s)^{1/2}}{2} \tag{B.7}$$

Algorithm is carried out in following steps

(1) Compute initial values of r and s

$$r = \frac{A1}{A2}, \qquad s = \frac{A0}{A2} \tag{B.8}$$

(2) Compute $B_2 \ldots B_n$ from Eq.B.3

(3) Compute $C_2 \ldots C_n$ and $D_2 \ldots D_n$ from following equations.

$$\begin{aligned}
C_n &= 0 \\
C_{n-1} &= -B_n \\
&\vdots \\
C_{n-i} &= -(B_{n-i} + rC_{n-i+1}) - sC_{n-i+2} \\
&\vdots \\
C_2 &= -(B_3 + rC_3) - sC_4
\end{aligned} \tag{B.9}$$

$$\begin{aligned}
D_n &= 0 \\
D_{n-1} &= 0 \\
&\vdots \\
D_{n-i} &= -rD_{n-i+1} - (B_{n-i-2} + sD_{n-i+2}) \\
&\vdots \\
D_2 &= -rD_3 - (B_4 + sD_4)
\end{aligned} \tag{B.10}$$

(3) Compute R,S,Δr,Δs

$$R = A_1 - rB_2 - sB_3 \qquad S = A_0 - sB_2$$

$$\Delta r = \frac{-RW + SU}{TW - UV} \qquad \Delta s = \frac{-TS + VR}{TW - UV} \tag{B.11}$$

where T,U,V,W are given by following equations

$$T = [-(B_2 + rC_2) - sC_3] \qquad V = -sC_2$$

$$U = [-(B_3 + sD_3) - rD_2] \qquad W = -(B_2 + sD_2 \tag{B.12}$$

(4) Correct values of r and s using Eq. B.6. If Δr is sufficiently small goto step (2), otherwise compute roots using Eq. B.7.

(5) Put the polynomial with B coefficients in the form of $P_n(x)$. If new polynomial is of degree 2, then solve the polynomial, otherwise goto step 1 and repeat.

APPENDIX C

LINEAR PREDICTIVE ANALYSIS OF SPEECH

C.1 INTRODUCTION

Linear prediction techniques can be used for accurate formant trajectories and bandwidths estimation. With respect to the other available techniques for the speech analysis such as log magnitude spectrum analysis, cepstrum analysis, linear prediction techniques offer the advantage of minimal complexity, minimal computation time and maximal accuracy for formant estimation. Following sections describe a technique called autocorrelation method of linear predictive analysis.

C.1 BASIC PRINCIPLES OF LINEAR PREDICTIVE ANALYSIS

A speech production model in which composite spectrum effects of radiation, vocal tract and glottal excitation is assumed. Such a model can be represented by a time varying digital filter whose steady state transfer function is of the form

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

For these system the speech samples are related to the excitation

u(n) by difference equation

$$s(n) = \sum_{k=1}^{p} a_k \, s(n-k) + G \, u(n) \qquad \text{(C.2)}$$

A linear predictor with prediction coefficients α_k is defined as a system whose output is

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k \, s(n-k) \qquad \text{(C.3)}$$

The system function for the p^{th} order linear predictor is

$$P(z) = \sum_{k=1}^{p} \alpha_k \, z^{-k} \qquad \text{(C.4)}$$

Prediction error e(n) is defined as

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k \, s(n-k) \qquad \text{(C.5)}$$

Prediction error sequence is the output of the system whose transfer function is

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k \, z^{-k} \qquad \text{(C.6)}$$

Comparing equations (C.2) and (C.3) it can be seen that speech signal obeys the model given by Eq. (C.2) exactly if $\alpha_k = a_k$.

Then $e(n) = G\,u(n)$ and prediction error filter will be an inverse filter for the system, i.e.,

$$H(z) = \frac{G}{A(z)} \qquad (C.7)$$

The basic problem of the linear prediction analysis is to determine a set of predictor coefficients $\{\alpha_k\}$ directly from speech signal in such a manner as to obtain a good estimate of the properties of the speech signal through the use of Eq.(C.7). Because of the time varying nature of the speech signal the predictor coefficients must be estimated from the short segments of the speech signal. The basic approach is to find a set of predictor coefficients that will minimize the mean squared prediction error over a short segment of speech waveform. The resulting parameters are then assumed to be the parameters of the system function H(z).

C.3 AUTOCORRELATION METHOD OF LINEAR PREDICTION ANALYSIS

In this approach it is assumed that the waveform segment is identically zero outside the interval $0 < m < N-1$

$$s_n(m) = s(m+n)\,w(m) \qquad (C.8)$$

where w(m) is a finite window length that is identically zero

outside the interval $0 < m < N-1$.

Short time average prediction error is defined as

$$E(n) = \sum_{m=0}^{N+p-1} e_n^2(m)$$

$$= \sum_{m=0}^{N+p-1} [s_n(m) - \tilde{s}_n(m)]^2$$

$$= \sum_{m=0}^{N+p-1} [\, s_n(m) - \sum_{k=1}^{p} \alpha_k \, s_n(m-k) \,]^2 \qquad \text{(C.9)}$$

We can find the values of α_k that minimizes the E_n in the Eq. (C.9) by setting $\partial E_n / \partial \alpha_i = 0$, $i = 1,2,\ldots,p$ thereby obtaining the equations

$$\sum_{m=0}^{N+p-1} [s_n(m-i) - s_n(m)] = \sum_{k=1}^{p} \hat{\alpha}_k \sum_{k=1}^{N+p-1} s_n(m-i)\, s_n(m-k) \qquad \text{(C.10)}$$

$\hat{\alpha}_k$ are values of α_k to minimize E_n We define

$$\emptyset_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)\, s_n(m-k) \qquad 1 < i < p, \quad 0 < k < p \qquad \text{(C.11)}$$

$s_n(m) = 0$ outside the interval $0 < m < N-1$.

$$\phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m) \cdot s_n(m+i-k) \qquad (C.12)$$

Thus $\phi_n(i-k)$ is short time autocorrelation function evaluated for $(i-k)$.

$$\phi_n(i,k) = R_n(|\, i-k \,|) \qquad (C.13)$$

where

$$R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)\, s_n(m+k) \qquad (C.14)$$

From Eq. (C.10)

$$\sum_{k=1}^{p} \alpha_k R_n|\, i-k \,| = R_n(i) \qquad 1 < i \leq p \qquad (C.15)$$

and minimum mean square error is given by

$$E_n = R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k) \qquad (C.16)$$

In order to effectively implement a linear predictive analysis system it is necessary to solve the linear equations given by Eq. (C.15) in an efficient manner. The most efficient method known for

solving this system of equations is Durbin's recursive algorithm which is described in the next section.

C.4 DURBIN'S RECURSIVE ALGORITHM

Durbin's recursive procedure can be stated as follows

$$E^{(0)} = R(0)$$

$$k_i = \left[R(i) - \sum_{j=1}^{i-1} R(i-j) \right] / E^{(i-1)} \qquad 1 < i < p$$

$$\alpha_i^{(i)} = k_i \qquad \text{(C.17)}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \, \alpha_{i-j}^{(i-1)} \qquad 1 < j < 1$$

$$E^{(i)} = (1 - k_i^2) \, E^{(i-1)}$$

Above equations are solved recursively for i = 1,2....,p and the final solution is given by

$$\alpha_j = \alpha_j^{(p)} \qquad 1 < j < p \qquad \text{C(18)}$$

In process of solving for the predictor coefficients for a predictor order p, the solution for the predictor coefficients of all orders less than p are obtained, i.e., $\alpha_j^{(i)}$ is the j^{th}

predictor of order i.

Finally LP spectrum is computed by converting the DFT spectrum of linear prediction coefficients padded with zeros into a log power spectrum.