Imperial College of Science and Technology

(University of London)

Department of Electrical Engineering

DETERMINATION OF ARTICULATORY PARAMETERS FROM SPEECH WAVEFORMS

by

John Albert Victor Rogers

A Thesis Submitted for the Degree of

Doctor of Philosophy, in the Faculty of Engineering,

University of London.

1974

ABSTRACT

A technique is developed for calculating the vocal tract cross sectional area function from the speech waveform. A digital transversal filter is first defined, having the inverse of the vocal tract transfer function. Next, a method for calculating the coefficients of this inverse filter from a given speech waveform is derived. This is accomplished by minimising the energy of the inverse filter output, during the closed glottis period of phonation. A transmission line analogue of the vocal tract is then developed, consisting of a number of cascaded commensurate sections. An algorithm is derived for calculating the reflection coefficients of this transmission line model from the inverse filter coefficients. The vocal tract area function is then simply calculated from these reflection coefficients.

Speech synthesised by convolution of a known vocal tract transfer function, with a known glottal excitation function is used to test the method. Results show that the method is capable of separating the vocal tract transfer function from the glottal excitation function. Synthetic speech is again used to compare the method with an alternative technique for performing the same analysis, recently reported by Wakita. The comparison shows that the method developed in this thesis is considerably more accurate than the technique reported by Wakita.

Vocal tract area functions are presented which have been calculated from spoken vowel sounds of male and female speakers. In addition to the speech waveform, another parameter required by the method is the time of glottal closure. It is shown how the instant of glottal closure can be determined from the output of the inverse filter, whose coefficients were calculated during the previous analysis interval. Implementation of the algorithm on a small digital computer, shows it is capable of near real time analysis (approximately twenty-five area functions per second can be displayed).

ACKNOWLEDGEMENTS

I am deeply indebted to Professor R.E. Bogner who was my Supervisor during the first 2½ years of this research. He constantly provided both the technical help and encouragement needed, especially when nothing seemed to go right, I am most grateful to him for this. My thanks are also extended to the Staff and Students of the Communications Section. The Staff under the leadership of my Supervisor for the final year, Professor E.C. Cherry have fostered an atmosphere which encourages discussion and exchange of ideas. My fellow students, especially Dr. E.V. Stansfield, Dr. E.J. Hayes, Dr. V. Lawrence, Dr. W. Edmondson and Mr. P. Rashidi, have always been willing to listen to my problems and helped, where possible, to solve them. I am also extremely grateful to Dr. P. Goddard and Dr. M. Apperley for the many occasions when they helped with computer problems.

Many thanks are also due to those who helped in the preparation of this thesis, especially Miss E. Farmer who performed so well, the exacting task of typing this manuscript. The financial support of the Science Research Council and Imperial College is gratefully acknowledged. The Science Research Council supplied both an equipment grant and a maintenance award in the form of a research studentship, and Imperial College employed me as a research assistant during the last eighteen months of my work.

Finally, my deepest gratitude is due to my parents and my wife, who although they often did not understand what I was doing, provided the moral support and encouragement I needed and to whom I dedicate this thesis.

CONTENTS

Page:

GLOSSARY OF MATHEMATICAL SYMBOLS USED IN THIS THESIS

Unless otherwise stated, the symbols used in this thesis have the meanings given below:-

$\underline{a}$	vector of inverse filter coefficients.
a_k	kth. coefficient of inverse filter = kth. coefficient of recursive filter model of vocal tract.
A_k	area of kth. section of vocal tract, (cm^2).
A_L	area of termination at the lips in acoustic tube model, (cm^2).
$A(x)$	vocal tract area function, (cm^2).
b_k	kth. linear prediction coefficient.
B	upper triangular matrix formed by Householder transformation.
c	velocity of sound in the vocal tract, ($cm\ sec^{-1}$).
$\underline{c}$	vector in Householder's method.
c_k	kth. coefficient of Markels "optimum inverse filter".
$C(z)$	inverse of vocal tract transfer function.
C_n	transfer function of n length inverse filter.
d	differential operator.
$\underline{d}$	Householder transform of vector $\underline{c}$.
∂	partial differential operator.
D_o	equilibrium density of air in vocal tract, ($gm\ cm^{-3}$).
e	total error energy (Markels method) or exponent in context.
$E(T/\hat{T})$	energy difference between stochastic and actual speech spectrum.
E_i	ith. sample of error function (Atals method).
F_k	constant in Householder transformation.
F_s	sampling frequency (Hz).
$g(t)$	glottal excitation time function.
g_i	ith. sample of output of the inverse filter.
$G(s)$	glottal excitation spectrum.
$h(t)$	impulse response of vocal tract.
$H(s)$	vocal tract transfer function.

$H_n(s)$	transfer function of n junction vocal tract model.
i	general time index.
I	identity or unit matrix.
j	specific index or $\sqrt{-1}$ in context.
k	general spatial index.
K	general constant.
ℓ_k	length of kth. section of acoustic tube model, (cm).
L	length of vocal tract, (cm).
$L(T/\hat{T})$	logarithmic maximum likelihood function.
ℓn	natural logarithm.
m	number of samples in Householder analysis window.
M	number of coefficients of Markels "optimum inverse filter".
n	number of junctions in vocal tract model [(n+1) sections].
N	number of speech samples used to calculate autocorrelation functions.
p	number of prediction coefficients in Atàls method.
P_G	pressure input from glottis.
P_k^+	forward travelling pressure wave in kth. section of transmission line model.
P_k^-	backward travelling pressure wave in kth. section of transmission line model.
P_k	net pressure in kth. section of transmission line model.
P_L	pressure output at lips.
$P^{(k)}$	elemental Householder matrix.
P(x)	sinusoidal sound pressure.
$p_o^{(m)}$	mth. eigen frequency of Webster horn equation.
Q	Householder transform matrix.
r_k	kth. sample of autocorrelation function.
R_k	kth. reflection coefficient in transmission line model.
R_M	normalised autocorrelation matrix.
R(s)	input reflection function at the glottis.
s	Laplace operator.

s_i	ith. sample of speech waveform.
s_p	pole positions in s plane.
$\underline{s}$	vector of speech samples.
S	matrix of speech samples.
t	time, (sec.).
T	delay, (sec.).
T_k	delay of kth. section of vocal tract model, (sec.).
T_{min}	delay of shortest vocal tract section, (sec.).
T_{tot}	delay of complete vocal tract, (sec.).
$T(\omega)$	speech spectral density.
u_m	constant in Levinsons method.
$\underline{U}$	component vector of Householder transform matrix $P^{(k)}$.
U_G	volume velocity input at the glottis.
U_k^+	forward travelling volume velocity wave in kth. section of transmission line model.
U_k^-	backward travelling volume velocity wave in kth. section of transmission line model.
U_k	net volume velocity in kth. section of transmission line model.
U_L	volume velocity output at lips.
v_m	constant in Levinsons method.
W_k	transmission matrix for kth. section.
x	distance from glottis in vocal tract.
X_{m+1}	constant in Levinsons method.
$Y_L(s)$	admittance of the lip termination, ($gm^{-1} cm^4 sec$).
z	= exp(jω) z-transform operator.
z_p	pole positions in z-plane.
Z	transmission matrix for vocal tract.
Z_k	characteristic impedance of kth. section of transmission line model, ($gm\ cm^{-4} sec^{-1}$).
Z_L	impedance of lip termination, ($gm\ cm^{-4} sec^{-1}$).
Z_G	output impedance of glottal source, ($gm\ cm^{-4} sec^{-1}$).
Z_{IN}	input impedance of vocal tract, ($gm\ cm^{-4} sec^{-1}$).
α	attenuation constant in transmission line model, (nepers cm^{-1}).

α_k	kth. linear regression coefficient.
α_H	attenuation losses due to heat conduction, (nepers cm^{-1})
α_V	attenuation losses due to viscous friction, (nepers cm^{-1})
α_W	attenuation losses due to vocal tract wall vibration (nepers cm^{-1}).
β	phase constant in transmission line model, (radians cm^{-1}).
γ	propagation constant in transmission line model, (cm^{-1}).
δ_{io}	Krönecker delta function.
$\delta(t)$	Dirac delta function.
λ	eigen value of Webster horn equation.
π	= 3.1415927.
ρ_m	reflection coefficient in vocal tract transmission line model (Wakitas method $\rho_m = -R_k$).
ϕ_k	= ϕ_{jk}, kth. sample of autocovariance function.
$\underline{\underline{\phi}}$	covariance matrix.
ω	angular frequency, (rad. sec^{-1})
Σ	summation.
$\prod$	product.
$< >$	average.
$\| \|$	euclidian norm.

PREFACE AND STATEMENT OF ORIGINALITY

In recent years a lot of interest has been shown in the analysis of speech in terms of articulatory parameters. Reasons for this interest have been varied. Some workers hope that articulatory parameters (i.e. the positions of the tongue, lips, jaw and velum used to produce a sound), will prove a useful set of data for automatic speech recognition. Others are mainly interested in learning more about the speech production process, either by analysis in terms of articulator positions, or synthesis from articulatory parameters. Articulatory parameters have two main advantages. Firstly, they contain a low level of redundancy compared to the speech waveform. Secondly, their physical nature placeslimits on their values.

Many different techniques have been employed in attempts to measure articulatory parameters. Some of the first attempts utilised X-rays to take a cineradiograph (film) of the speakers articulators. Although X-ray methods were used to derive most of the present knowledge about articulator positions, they suffer from severe limitations. These limitations include the small exposure to X-rays possible for a single subject, and the poor resolution of the cineradiographs. To avoid the difficulties of using X-rays, attempts have been made to derive articulatory parameters from the speech signal. Analysis of the speech signal for articulatory parameters requires first a model for sound propagation in the vocal tract.

Early attempts considered the propagation of sound to obey the Webster horn equation, and used perturbation techniques to calculate the vocal tract area function. These methods required information not available from the speech waveform to derive unique area functions, and required a lot of computation to calculate a single area function. In MERMELSTEINS method (1967), an iterative process was used to calculate the area function from the driving point impedance at the lips. This impedance was measured by the subject articulating without phonation into an impedance tube.

More recently the vocal tract has been modelled as an acoustic transmission line and the well developed techniques of this field applied to the problem. Initial uses of transmission line theory did not lead to unique area functions, because incorrect terminating conditions were assumed for the lips and the glottis. Later applications chose these terminations correctly, but did not satisfactorily separate the vocal tract transfer

function from the glottal excitation function, before using the former to derive the area function.

In this thesis a technique is developed, which derives unique area functions, based on treatment of the vocal tract as a transmission line. This technique is conveniently considered in two parts. First, the information about the vocal tract transfer function is extracted from the speech waveform by inverse filtering. Secondly, the area function is calculated from the inverse filter coefficients. Separation of the vocal tract transfer function from the glottal excitation function is ensured by analysing only these regions of the speech waveform corresponding to a closed glottis (no input excitation). The inverse filter coefficients are calculated using these short records of speech by Householder transformation. A method is also developed based on commensurate transmission line theory, for calculating the reflection coefficients of the model (and thence the area function) from these inverse filter coefficients.

Synthetic speech is used to test the method and also to compare it with an alternative technique for calculating the area function reported by WAKITA (1972). Results show that my method successfully separates the excitation function from the transfer function, but that Wakita's method only approximately achieves this separation. Furthermore the results prove, my method is only capable of deriving realistic area functions from analysis of closed glottis regions of the speech waveform.

My technique is then applied to the analysis of vowels spoken by four different speakers. The algorithm used for this analysis includes a correction factor for vocal tract losses, which was not used in the synthetic speech work. Area functions derived are realistic, and similar area functions are obtained by analysis of the same vowel spoken by different people. However, it is felt that more results are needed before any sweeping conclusions can be drawn.

The parts of this thesis which, to the best of the authors knowledge, represent original contributions are:-

i) The derivation of the algorithm for calculating the reflection coefficients of a transmission line consisting of cascaded commensurate sections given in section 3.6 (This algorithm is a development of the work of KINARIWALA (1966) and a similar algorithm has been derived for the less general case of equal length sections by ATAL (1971)).

ii) The development of an automatic inverse filtering technique to separate the vocal tract transfer function from the glottal excitation

function, by analysis of only closed glottis periods of the speech waveform (Sections 4.1 → 4.3).

iii) The use of Householder transformation to calculate the inverse filter coefficients (section 4.4). Also to some extent the presentation of Householder transformation given in section 4.4. (This is considerably simplified from the original papers: GOLUB (1965), WILKINSON (1960)).

iv) Inclusion of a correction factor to account for vocal tract losses in the area function mapping algorithm (section 5.2).

v) Comparison of the area functions derived by analysis of the speech waveforms from different speakers, for a given vowel.

CHAPTER ONE

1.1. SPEECH - A UNIQUE HUMAN FACULTY

"The fossil evidence of human evolution implies that a series of changes from the primate vocal tract took place, some of which may have been necessary for the generation of speech". Times Science Report, 11th. June, 1969, page 4.

This statement referred to attempts to synthesise speech using the geometry of the vocal tract of a Rhesus monkey. The work carried out at the Massachusetts Institute of Technology suggested that the Rhesus monkey in particular, and probably other members of the monkey and ape families, are anatomically incapable of imitating the full repertoire of human speech. It seems, therefore, that humans have been endowed with a faculty not found in other animals. Although other animals can transmit emotional signals of pleasure, pain, or danger, only humans can communicate socially by use of the spoken word.

Humans are able to exchange ideas and argue about principles with their neighbours, by use of the spoken word. Writing, developed as a method of recording speech, provides a medium whereby man can transmit his thoughts through time and space. Over the years, communication at this level has forced man to order his thoughts. I would suggest that it was this need for man to order his thoughts that led to the development of the inquisitive scientific mind. Dare I suggest that this development might not have occurred if man had not been given the ability to speak. In that case we would not have the subject of speech production to study, or the ability to study it!

1.2. AIM AND MOTIVATIONS OF THIS RESEARCH

The aim of the research reported in this thesis is to develop a method of determining the vocal tract area function (a measure of the cross sectional areas of the cavities between the larynx and the lips), from the acoustic speech signal detected by a microphone. The outcome of the research at the present time is a technique whereby an approximation to the vocal tract area function can be calculated from samples of the speech waveform. This technique is limited in application to voiced non-nasalised sounds by assumptions made in its derivation.

Motivation behind the work is threefold:-

a) It is believed that if articulatory parameters could be reliably and quickly obtained they would provide useful information for future speech research. At present the main analysis tool available to the speech

research-worker is the spectrogram. While not wishing to detract from the value of the spectrogram, it is felt that an articulogram would be more useful. An articulogram is conceived as a display of the main articulatory parameters. One possibility might be a visual display, similar to that obtained from an X-ray picture of the head, together with some information about the position of the source and, for voiced sounds, the pitch.

The main reason for this belief is that a physically based parameter set, related to the method by which sounds are produced, should be easier to interpret than the somewhat abstract phenomenum of the spectrum. Obviously if an articulograph was to be made available it would not gain popularity overnight. However, its use in conjunction with the spectrograph would prove invaluable to workers in the speech analysis field.

b) It is hoped that articulograms would be useful to deaf people. It is possible that the information contained in an articulogram could provide the feedback necessary to a deaf person learning to talk. Indeed, use of such a facility with the help of a trained speech therapist should help deaf people to understand the required movements of the articulators during speech.

Another possibility is that deaf people might be able to develop their ability to lip read into an ability to read a simple articulogram display. Such a display should not detract from the information about lip shapes presently used, but should make available additional information about other equally relevant articulator positions.

c) It is hypothesised that articulatory analysis might help in solving the problem of automatic speech recognition. This idea is supported by a number of facts about speech. Firstly, classification of speech in terms of articulator positions is the basis for the phonetic description, which is largely talker invariant and contains a low level of redundancy. Thus an articulatory based recogniser might be more useful than one based on, say, acoustic features. Secondly, because of the physical basis for the parameter set, certain constraints about the possible articulator positions and their rate of movement from one position to another can be applied. Finally, articulatory transitions are smoother than formant transitions (FLANAGAN 1972), which makes constraints easier to apply in the articulatory domain than in the frequency domain.

It was with these ideas in mind that the research reported here was

carried out. The next section is a brief review of the subject of this research, namely the speech production process.

1.3. THE SPEECH PRODUCTION PROCESS

"The quality, or 'timbre' of the human voice, I believe, is due in a very minor degree to the vocal chords, and in a much greater degree to the shapes of the passages through which the vibrating column of air is passed". Alexander Graham Bell, 1907.

Although a detailed description of how a speech sound is produced depends to some extent on the particular sound in question, this statement by Bell is still a good generalisation.

A lateral mid-saggital section of the apparatus used in speech production is shown diagrammatically in Figure 1.1. The vocal apparatus consists of three main cavities, namely the pharynx, the mouth and the nasal cavity. These cavities can act as resonators shaping the harmonic content of sounds radiated from the lips and/or the nostrils. The sizes of these cavities are varied by moving the articulators, which primarily consist of the tongue, the lips and the jaw. The various articulator positions give rise to different speech sounds.

Excitation of these cavities may be the result of one or more phenomena. During the production of voiced sounds (e.g. /a/ in father), a quasiperiodic source is formed by forcing air from the lungs through the opening between the vocal folds in the larynx (this opening is referred to as the glottis). Vocal fold oscillation is analogous to that of a relaxation oscillator. The folds are tightened, thus closing the glottis, and pressure builds up in the sub-glottal regions during attempted exhalation. Eventually, this pressure becomes sufficient to force the folds apart, opening the glottis and allowing air to flow from the lungs into the pharynx. This air flow causes a decrease in pressure below the glottis, and a decrease in pressure in the glottal orifice (Bernouilli effect). This decrease in pressure coupled with the tension of the vocal folds causes the folds to snap shut. The process is then repeated, giving rise to a near periodic excitation. VAN DEN BERG et al (1957) studied the physics of this process in detail, and derived equations which accounted for the glottal air flow.

Various computer models simulating the glottal source have been developed. When these are used with realistic values for the glottal dimensions and the subglottal pressure, they give further insight into the workings of the larynx (FLANAGAN et al, 1968 , 1969 ; CRYSTAL et al, 1965).

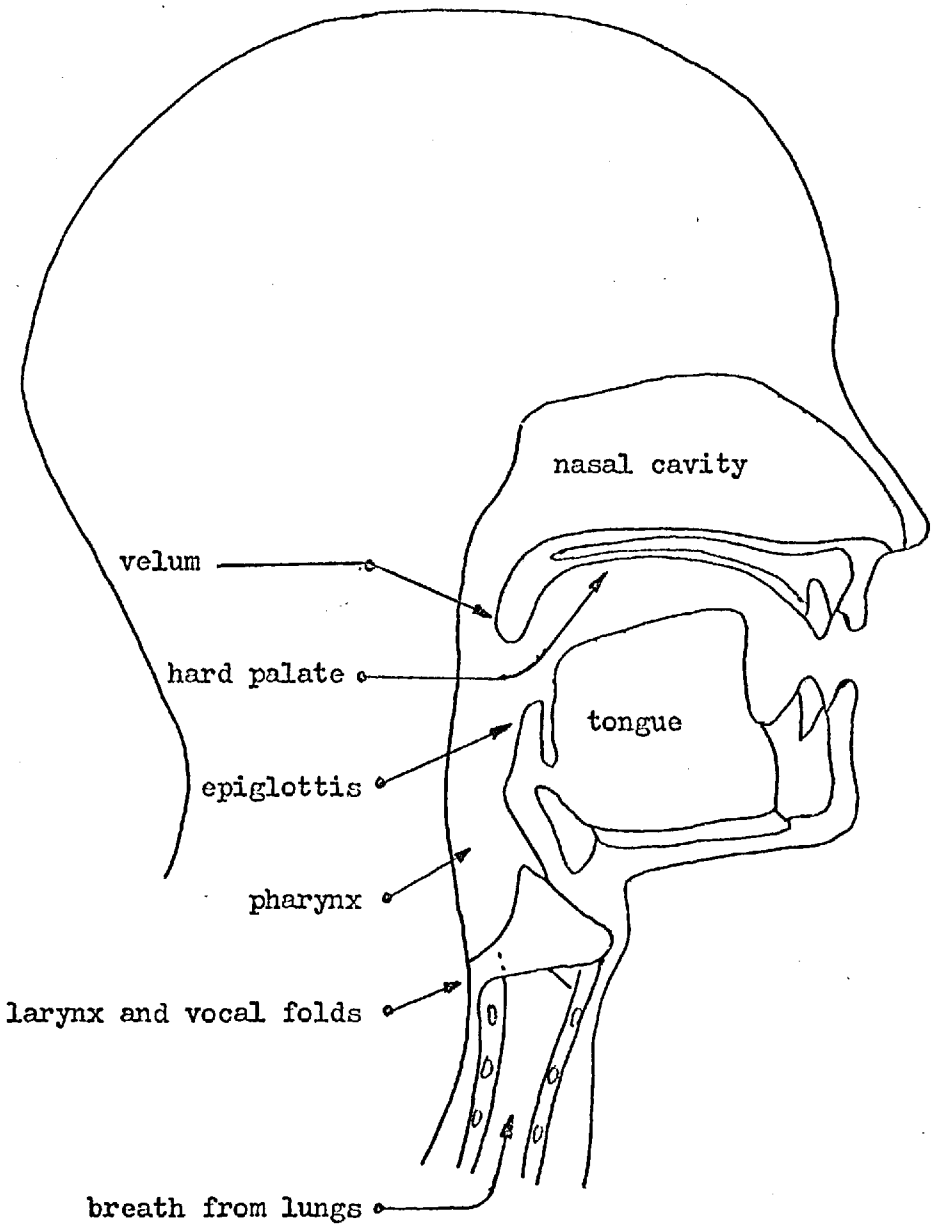


FIGURE 1.1 Lateral cross section through head showing vocal organs

Several techniques have also been developed for measuring the glottal volume velocity source during natural speech. One of the more successful methods has been inverse filtering, where an electrical network is adjusted until its transfer function is the inverse of the vocal tract transfer function. Speech is then passed through this inverse filter, the output being an estimate of the glottal volume velocity waveform (HOLMES 1962, ROTHENBERG 1973).

One important fact about glottal excitation needs to be emphasised. If sufficient vocal effort is used during phonation, a period of glottal closure results in each excitation cycle. During this period there can be no resonant coupling between the vocal tract and the sub-glottal regions. Studies have suggested that during open glottis periods a further resonant cavity, namely the lungs, is added to the speech production apparatus (FANT 1972).

A different form of excitation is used to produce fricated sounds. This takes the form of turbulent air flow at a point of tight constriction in the vocal tract (e.g. /f/ as in for). The turbulence caused by such a constriction consists of rotational flow and eddies, giving rise to a noise-like excitation of the vocal tract. Such sounds are usually phonetically classified in terms of the position of their excitation source.

Stop consonants or plosives are excited by a sudden release of pressure from behind a point of closure in the vocal tract. Such an excitation is analogous to applying a step function to an electrical network. An example of this type of excitation is the lip closure apparent in /p/ as in pay. Many speech sounds have more than one source of excitation. Both voicing and frication are present in /v/ as in vote, and voiced plosive excitation is apparent in /d/ as in day.

Man's repertoire of communicative sounds is made more comprehensive by using different articulatory configurations to shape the source. Consider the example of one class of sounds known collectively as vowels, which are voiced non-nasalised phonemes. The vowels are tabulated in terms of their phonetic description in Figure 1.2, along with the other phonemes of General American English. During the production of most vowel sounds the vocal tract is divided into two resonant cavities, formed by a constriction between the tongue and the roof of the mouth. The sizes of these cavities are determined by the position of the tongue in forming the constriction, and by the positions of the other articulators. The degree of coupling between the cavities is

VOWELS			
Degree of constriction	Tongue hump position		
	Front	Central	Back
High	/i/ eve	/ɝ/ bird	/u/ boot
	/I/ it	/ɚ/ over (unstr)	/U/ foot
Medium	/e/ hate	/ʌ/ up	/o/ obey
	/ɛ/ met	/ə/ ado (unstr)	/ɔ/ all
Low	/æ/ at		/a/ father

FRICATIVE CONSTANTS		
Place of articulation	Voiced	Unvoiced
Labio-Dental	/v/ vote	/f/ for
Dental	/ð/ then	/θ/ thin
Alveolar	/z/ zoo	/s/ see
Palatal	/ʒ/ azure	/ʃ/ she
Glottal		/h/ he

STOP CONSONANTS			NASALS
Place of articulation	Voiced	Unvoiced	Voiced
Labial	/b/ be	/p/ pay	/m/ me
Alveolar	/d/ day	/t/ to	/n/ no
Palatal/velar	/g/ go	/k/ key	/ŋ/ sing

GLIDES		SEMI-VOWELS or LIQUIDS	
Place of articulation	Voiced	Place of articulation	Voiced
Palatal	/j/ you	Palatal	/r/ read
Labial	/w/ we	Alveolar	/l/ let

DIPHTHONGS			
Voiced			
/eI/	say	/aU/	out
/Iu/	new	/aI/	I
/ɔI/	boy	/oU/	go

AFFRICATES	
Voiced	Unvoiced
/dʒ/ jar	/tʃ/ chew

Figure 1.2 Classification of Phonemes of General American English

(After FLANAGAN 1972)

determined by the degree of constriction. The importance of both the degree of constriction and the position of the constriction has led to classification of vowel sounds in terms of the so called vowel quadrilateral (see Figure 1.3).

Nasal phonemes result from lowering of the velum which allows transmission through the nasal tract and radiation from the nostrils. Glides and semivowels to some extent resemble vowels in their method of production. In both cases, voiced excitation is used, the velum is raised and sound is radiated mainly from the lips. The glides (e.g. /j/ in you), are dynamic sounds resulting from movement of the articulators during phonation. Semivowels, on the other hand, are formed by raising the tongue tip thus causing more constriction than is commonly found in vowels (e.g. /l/ in let). Finally, the phonemes of the last two categories are produced by combining two separate speech sounds. Dipthongs result from a combination of two vowels (e.g. /eI/ in say) and affricates result from a combination of stops and fricatives (e.g. /dz/ in jar).

This completes the description of the speech production process. A fuller description may be found in FLANAGAN (1972) and a study in terms of an acoustic theory of speech production is given in FANT (1970).

1.4. DIRECT METHODS OF DETERMINING THE VOCAL TRACT SHAPE

The previous section described the way the articulatory apparatus is used to produce speech. Much of the detailed study of articulator positions during speech has been made with the aid of X-rays. Typically, in these studies a one-dimensional X-ray was taken, with the beam projected perpendicular to the mid-saggital section of Figure 1.1. To obtain the area function from these pictures some method of measuring the lateral dimensions of the vocal tract was needed. Commonly the dimensions of the front buccal or mouth cavity were measured with the aid of palatograms (false palates of the type used by dentists); direct photography was an alternative method. Thin palatograms coated with powder were often used to find the position and extent of tongue contact during production of phonemes in which tongue palate contact occured. The area of the lips could be recorded by cine pictures of the subjects face, and the sound was usually simultaneously recorded for subsequent linguistic or acoustic analysis.

The first X-ray pictures of the vocal tract which included measurements of the lateral dimensions of the pharynx were made by RUSSEL (1929). An

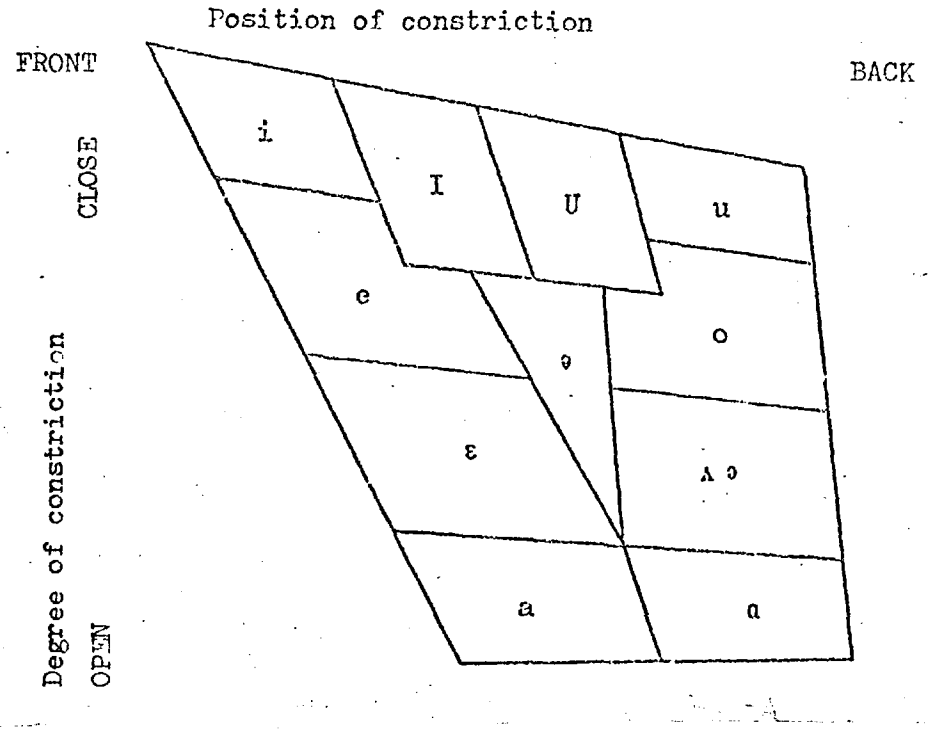


Fig 1.3a Vowel Quadrilateral.

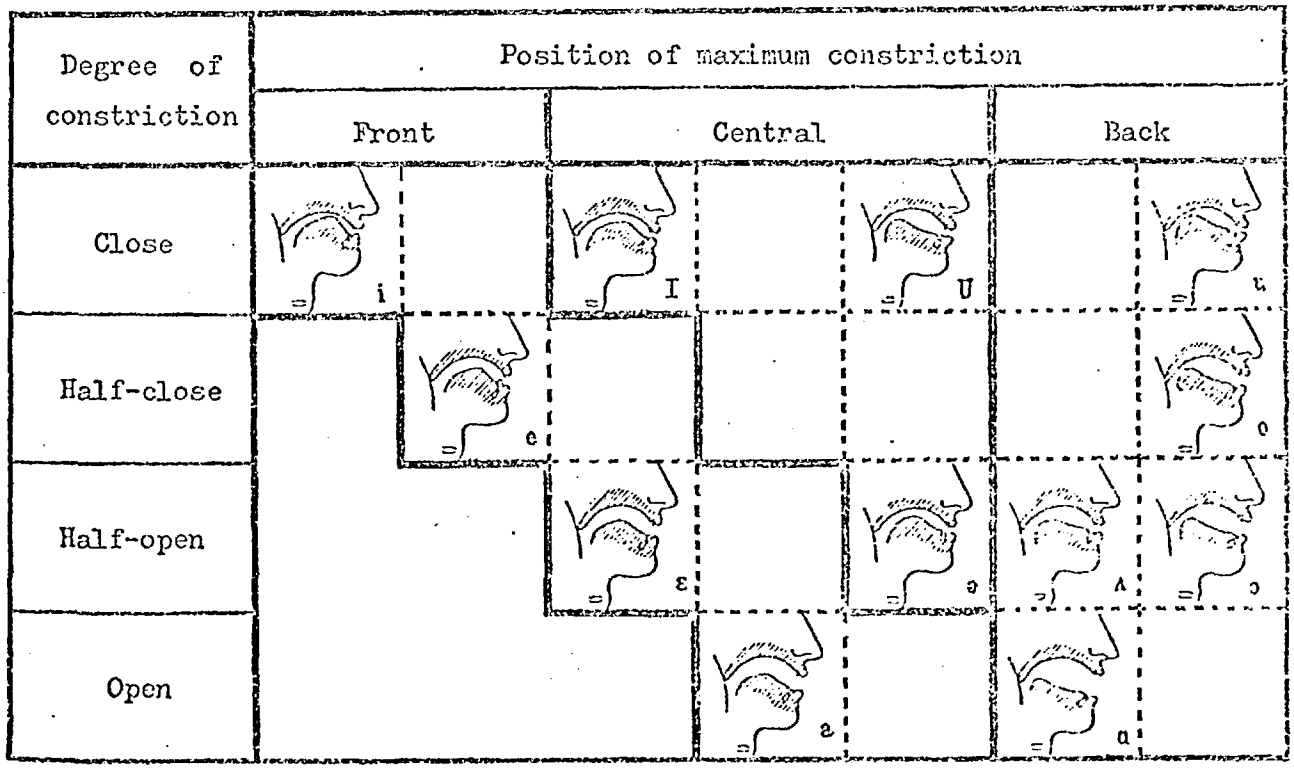


Fig 1.3b Articulator Diagrams.

Figure 1.3 Vowel Quadrilateral and Articulator Diagrams for General American English (Adapted from FLANAGAN 1972)

earlier study made by POLLARD and HALA (1926) did not include these measurements. Russel used a marker to delineate the median line of the tongue, thus making it easier to take the transverse dimensions of the tongue cavity. Lateral dimensions of the front buccal cavities were made with the aid of palatograms calibrated by punching pin holes at the intersections ofalattice, formed by two perpendicular sets of parallel lines 1 cm apart. A laryngo-periskop was used to take the dimensions of the subjects larynx and pharynx regions. In all, some 3000 measurements were made using over 400 different subjects and including measurements of all the phonemes of General American English and some Spanish and German phonemes. However, even with all this data, Russel seemed reluctant to drop the theory apparently prevalent at that time. This theory treated the vocal tract as analogous to a spanish wine decanter. The curved neck of the decanter represented the front mouth cavity, and the belly of the decanter represented the back throat or pharynx cavity.

Chronologically the next improvement in X-ray techniques was made by CHIBA and KAJIYAMA in 1941. They used a fine gold chain attached to the tongue to give the median tongue position on the X-ray. In order to allow the tongue palate constriction to be measured more accurately they also attached a ribbon of tin foil to the palate, or coated the palate with barium sulphate. Clearer pictures were obtained by careful adjustment of the voltages and currents in the X-ray apparatus. Solid palatograms were used, because the planar ones used earlier proved unsatisfactory. Also they made use of a laryngoscope mirror to measure the larynx and pharynx dimensions. Their study led to an articulatory classification of Japanese phonemes. In particular the classification of Japanese vowels by the position and degree of the tongue constriction.

Probably the best known study utilising X-rays was made by FANT and is expounded in his book 'Acoustic Theory of Speech Production' (FANT 1970). It is Fant's data that is commonly used by other workers to verify the applicability of their methods for determining the vocal tract area function. Fants study included six vowels, two liquids, one affricate, eight fricatives and six unvoiced stops, of a single Russian speaker. A mixture of barium sulphate and mucilage of acacia was spread on the subjects palate, the latter additive allowing a denser solution to be made than would be possible if an aqueous solution alone were used; this also gave better adhesion to the palate, and hence clearer outlines of the palate in the X-rays. Plaster of paris was used to take a three dimensional cast of the subjects mouth cavity and a cine

film of the subjects face was made to give the lip openings. Even with all these refinements, Fant commented that some dimensions of the vocal tract were impossible to measure under natural speaking conditions using X-rays.

The next section summarises the difficulties encountered in using X-rays to measure the vocal tract area function during speech.

1.5. DIFFICULTIES IN USING X-RAYS FOR ARTICULATORY MEASUREMENTS

One of the major difficulties in X-ray techniques was the limited dosage which could be used on any one subject. This limited dosage meant that only a few utterances of any given phoneme could be studied, sometimes only one utterance was used; and it also meant that the time resolution of the X-rays was poor. These difficulties were surmounted to some extent by FUJIMURA et al (1968) who used a computer to control the X-ray beam, directing it only at the physiological areas of interest for that phoneme.

Examination of a single frame from an X-ray cine film (cineradiograph) readily reveals the lack of detail available, even if barium sulphate coatings and gold chains have been used to enhance the outline of the articulator positions. This lack of detail, together with the difficulty of transcribing from one dimensional pictures to cross sectional area functions, places a limit on the accuracy of these X-ray techniques. Strictly speaking, three dimensional X-rays are needed to obtain area functions but dosage limitations usually make this impracticable.

The X-ray rooms used had poor acoustic properties which coupled with the noise of necessary proximate apparatus like the X-ray camera resulted in poor sound quality recordings. This poor sound quality made it difficult to carry out meaningful acoustical analyses of the sounds. There is also some evidence to suggest that man makes use of head movements in producing some speech sounds. If this is the case, the necessity for the subject to keep his head still during X-ray studies would mean some of the recorded sounds were unnatural.

Bearing in mind these difficulties, one can only be thankful for the painstaking labours of workers in this field. At the same time however, there is a need for some simpler alternative method of obtaining the area function data. The remainder of this chapter briefly reviews some attempts to obtain articulatory data without resorting to cineradiography.

1.6. ARTICULATORY RELATED SPEECH SYNTHESIS

Research in the area of speech synthesis has undoubtedly made a large contribution to our understanding of the speech production process. Work in this area is not directly relevant to this thesis, therefore it will only be briefly mentioned here.

An early systematic study of speech synthesis from articulatory parameters was made by DUNN (1950). Dunn devised a model consisting of four sections representing the pharynx, the tongue hump, the mouth cavity and the lip constriction. He then used Russell's X-ray data, to derive the parameters of a transmission line analogue consisting of electrical "T" sections. Surprisingly good results were obtained considering the simplicity of the model, and when he later increased the number of sections to thirty, even better results were obtained. FANT (1970) used Dunn's model as a prototype when he developed the LEA electrical transmission line analogue of the vocal tract.

One of the major problems in this type of speech synthesis, is the method of deriving the parameters representing the articulatory configurations. In order to reduce the number of parameters necessary some simple articulatory models for speech synthesis have been proposed. In 1955 STEVENS and HOUSE developed such a model using only three parameters, namely the position of the tongue constriction, the degree of this constriction and the size of the lip opening. COKER (1967) described a model using seven parameters which he used to synthesise speech with the aid of a digital computer. The speech synthesised by COKER is not as good as that currently being synthesised by other methods. One therefore concludes that considerably more effort is needed in this field of speech research.

1.7. PERTURBATION METHODS FOR CALCULATING VOCAL TRACT AREA FUNCTIONS

MERMELSTEIN and SCHROEDER (1965) were the first to report a method for determining the vocal tract area function from measured speech formant frequencies. The inherent ideas of their method were further developed in other papers, namely MERMELSTEIN (1967), and SCHROEDER (1967).

Starting with the Webster Horn equation (Equation 1.1) they considered the problem from a perturbation viewpoint.

$$\frac{d}{dx}\left(A(x)\,\frac{dP(x)}{dx}\right) + \lambda A(x)P(x) = 0 \qquad (1.1)$$

where

x is the distance from the glottis.

$A(x)$ is the vocal tract area function.

$P(x)$ is the sinusoidal sound pressure.

λ is the eigen value of the equation; λ is related to the eigen frequency (i.e. formant) ω via the equation

$$\lambda = \frac{\omega^2}{c^2} \tag{1.2}$$

where c is the velocity of sound in the tract.

Mermelstein and Schroeder used the substitution

$$\frac{1}{A(x)} \frac{d}{dx} A(x) = \frac{d}{dx} (\ell n\ A(x)) \tag{1.3}$$

and rearranged equation 1.1 to give

$$\frac{d^2}{dx^2} P(x) + \frac{d}{dx} (\ell n\ A(x)) \frac{dP(x)}{dx} + \lambda P(x) = 0 \tag{1.4}$$

However, Mermelstein and Schroeder were not interested in solving equation 1.4 for the eigen values, but rather the inverse problem of finding the area function given the eigen frequencies. Assuming the boundary conditions of a closed glottis and zero impedance load at the lips they derived the eigen frequencies of equation 1.4 for a uniform tube as

$$p_o^{(m)}(x) = \frac{\cos(2m - 1)\pi x}{2L} \qquad m = 1,2,3,.... \tag{1.5}$$

and the corresponding eigen values as

$$\lambda_o^{(m)} = \left(\frac{(2m-1)\pi}{2L} \right)^2 \qquad m = 1,2,3,.... \tag{1.6}$$

where L is the total length of the vocal tract.

By considering a first order perturbation to the cross sectional area of the uniform tube, they were able to relate spectral deviations in the eigen frequencies to spatial deviations from the uniform tube. Expansion of the vocal tract area function in terms of a spatial Fourier series allowed quantification of these effects. In his 1967

paper, Mermelstein showed the perturbation of the mth eigen frequency was uniquely related to the (2m-1)th coefficient of the spatial Fourier series

$$\ell n\, A(x) = \ell n\, A_o + \sum_{k=1}^{\infty} A_k \cos\left(\frac{k\pi x}{L}\right) \qquad (1.7)$$

It might appear that knowing the first four formant frequencies for a given sound, one could calculate the deviations of these formants from those of a uniform tract and use Mermelstein's method to calculate the vocal tract area function for that given sound. Unfortunately, two difficulties prevent this from being the case. Firstly, all speech sounds must be considered as being produced by first order perturbations from a uniform vocal tract shape; this is certainly not true for vowels like /u/ in boot. Secondly, the method cannot lead to unique area functions, because only the odd Fourier coefficients of the logarithmic area function can be derived from the formant frequencies. To determine the even Fourier coefficients some further assumptions have to be made.

Another way to look at this uniqueness problem is offered by regarding the Webster Horn equation as a special case of the Sturm--Liouville equation. To solve the Sturm-Liouville equation uniquely, two solutions from two different boundary conditions must be taken and summed (BORG 1946). A suitable second boundary condition might be closed lips and open glottis. Unfortunately, it is rather difficult to measure the eigen frequencies for the closed lips boundary condition, and the problem is likely to elude solution for many years to come.

Realising the limitations of their method, Mermelstein and Schroeder looked for some alternative way of calculating a unique cross sectional area function. They found that two independent sets of eigen frequencies could be calculated from the lip admittance function, which was measured by use of an impedance tube constructed by MERMELSTEIN and WEST (1968). The impedance tube contained a pulse source which excited the vocal tract at the lips, the lip admittance function was calculated from the signals received by two microphones, spatially separated in the impedance tube. To use the tube, subjects were required to articulate, without phonation, with their lips at one end of the tube. In principle, this alternative technique overcomes the second difficulty mentioned earlier. However, the first difficulty mentioned earlier is still not overcome, and the use of an impedance tube to obtain the data is unnatural

and cumbersome.

MERMELSTEIN (1967) partially avoided the difficulty of assuming all vowel sounds to be first order perturbations of the uniform tract. He did this by invoking higher order perturbations and using an iterative algorithm to find the area function. HEINZ (1967) proposed another possible way of avoiding this difficulty by arguing that the spatial Fourier series used by Mermelstein and Schroeder was only one of a number of possible sets of orthogonal functions, anyone of which could equally well be used. In his paper, Heinz was able to show that different sets of orthogonal functions existed for perturbations about each area function; in particular he derived the set applicable to perturbations of the area function for the vowel /a/.

Finally, in this general area of perturbation analysis, another method of solution based on the eigen values of the Webster Horn equation was proposed by GOPINATH and SONDHI (1970). In an extremely mathematical paper they derived a method of calculating the vocal tract area functions from the frequencies and bandwidths of the first four formants, together with the length of the vocal tract. As the bandwidths cannot be reliably estimated from the speech waveform, this method is not thought to be a satisfactory solution to the problem.

We have seen in this section that a number of techniques are known for calculating the vocal tract area function. However, they all require complicated calculations, which are computationally inefficient; also they all require data not easily deducable from the speech waveform. The present author therefore considers none of these methods to be a satisfactory solution of the original problem.

1.8. TRANSMISSION LINE TECHNIQUES FOR CALCULATING THE VOCAL TRACT AREA FUNCTION

The vocal tract can be considered as a cascade of uniform transmission lines, and the well developed methods of transmission theory used to solve the problem. Later in this thesis a technique will be developed based on treating the vocal tract as a transmission line, because of this, the assumptions necessary in modelling the vocal tract as a transmission line are given in Section 3.2.

In 1969 PAIGE and ZUE reported a technique based on RICHARDS algorithm (1948). Richards algorithm is a recursive technique for calculating the characteristic impedances of a number of equal length cascaded transmission lines, from the input impedance function of the

total structure. The characteristic impedance of the first section is calculated from the input impedance function. Knowing this, the input impedance of the sections remaining (after removing the first section), is calculated. This process is then repeated until the characteristic impedances of all the elemental transmission line sections are known. The characteristic impedance of an elemental acoustic transmission line is related to its cross sectional area by

$$Z_k = \frac{D_o c}{A_k} \qquad (1.8)$$

where

D_o is the equilibrium density of air in the tract,
c is the velocity of sound in the tract,
Z_k is the characteristic impedance of the kth. section,
A_k is the area of the kth. section
(STANSFIELD, 1971, page 59).

Hence the vocal tract area function can be calculated from the input impedance function for the vocal tract. PAIGE and ZUE (1969a) proposed precisely this method for finding the vocal tract area function from the driving point impedance at the lips. They tested their method by calculating the poles and zeros of the driving point impedance at the lips based on the data for six Russian vowels (FANT, 1970).

They also proposed a technique for finding the driving point impedance at the lips from the vocal tract transfer function. To do this Paige and Zue considered the transfer impedance matrix for the entire vocal tract, namely,

$$\begin{bmatrix} P_G \\ P_L \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} U_G \\ U_L \end{bmatrix} \qquad (1.9)$$

where

P_G is the pressure at the input of the first section,
P_L is the pressure at the output of the tract (lips),
U_G is the volume velocity in the first section,
U_L is the volume velocity at the lips.

They further assumed the lips to be a short circuit termination ($P_L = 0$) and the glottis to be an open circuit termination. Using these conditions the expression for the transfer function is given by,

$$H(s) = -\frac{U_L}{U_G} = \frac{Z_{21}}{Z_{22}} \qquad (1.10)$$

because P_L is assumed zero. Paige and Zue also stated the general form of the parameters Z_{21} and Z_{22} for a transmission line as

$$Z_{21} = \frac{K}{D(s)} \qquad (1.11)$$

$$Z_{22} = \frac{N(s)}{D(s)} \qquad (1.12)$$

where K is a constant and N(s) and D(s) are all zero functions. Using equations 1.11 and 1.12, equation 1.10 reduces to

$$H(s) = \frac{K}{N(s)} \qquad (1.13)$$

which means that the poles of the transfer function are the zeros of Z_{22} (the driving point impedance at the lips). Therefore the zeros of the driving point impedance at the lips are the poles of the vocal tract transfer function, but the poles of Z_{22} cannot be calculated from the acoustic signal. Paige and Zue suggested that the poles of Z_{22} could be chosen to satisfy the alternation property, but this would not lead to unique area functions, because a large number of possible positions for these poles could equally well be chosen.

It is worth commenting on the choice of boundary conditions used by the various workers. Mermelstein and Schroeder, Heinz, Gopinath and Sondhi, and Paige and Zue all assumed the glottis to be an open circuit termination and the lips to be a short circuit termination. The lips cannot be a short circuit termination in the real vocal tract because no pressure radiation is possible from a short circuit termination. Furthermore it is ironic that the problem of non-uniqueness of the area functions of these methods, has its origin in the choice of these boundary conditions. If, instead of assuming the lips to have a short circuit termination, we consider them to be terminated in a finite impedance, unique area functions can be obtained. We will now show this by adapting an arguement used by STANSFIELD (private communication). For a finite lip impedance the transfer function is given by equation 1.9 as

$$H'(s) = -\frac{U_L}{U_G} = \frac{Z_{21}}{Z_{22}-Z_L} \qquad (1.14)$$

where Z_L is the lip terminating impedance (i.e. $Z_L = P_L/U_L$). Now substituting from equations 1.11 and 1.12 into 1.14 we can see that

$$H'(s) = \frac{K}{N(s) - Z_L D(s)} \qquad (1.15)$$

This means that under the assumption of a finite terminating impedance of the lips both the poles and zeros of the driving point impedance function can be found from the vocal tract transfer function. Stansfield, realised the necessity of a finite resistive lip termination and derived a method using this boundary condition.

In his method STANSFIELD (1971) and STANSFIELD and BOGNER (1973) modelled the vocal tract as a lossless acoustic transmission line consisting of equal length sections, terminated at the lips by a resistance. Starting with the real part of the load impedance at the lips and the squared magnitude of the vocal tract transfer impedance (pressure radiated divided by volume velocity input), they were able to calculate the real part of the input impedance looking in at the glottis, by the equation

$$\mathrm{Re}(Z_{IN}(s)) = \left| \frac{P_L(s)}{U_G(s)} \right|^2 \mathrm{Re}(Y_L(s)) \qquad (1.16)$$

where

$Z_{IN}(s)$ is the input impedance looking in at the glottis
$Y_L(s)$ is the load admittance at the lips

As the vocal tract impulse response is causal, the imaginary part of the input impedance looking in at the glottis follows by Hilbert transformation. A complete description of the input impedance is thus derived from the magnitude of the transfer impedance of the vocal tract, terminated in a finite impedance at the lips. Knowing the vocal tract imput impedance seen at the glottis, and assuming the output impedance Z_G of the glottal source, they were able to calculate the input reflection coefficient as a function of frequency given by

$$R(s) = \frac{Z_{IN}(s) - Z_G}{Z_{IN}(s) + Z_G} \qquad (1.17)$$

where R(s) is the input reflection function at the glottis. Next they calculated the input impulse response by Fourier transformation of R(s). Stansfield then developed an algorithm for finding the reflection

coefficients at each junction in the transmission line, from this input impulse response. Basically, the algorithm was derived by a careful study of the progress of an impulse of pressure travelling from the glottis to the lips. This algorithm developed independently by Stansfield, is also attributable to SHIH (1965). The area function was simply calculated from the reflection coefficients of each junction by using the definition of the reflection coefficient from transmission line theory, i.e.,

$$R_k = \frac{Z_{k+1} - Z_k}{Z_{k+1} + Z_k} \qquad (1.18)$$

(JOHNSON, 1950):
to find the characteristic impedances Z_k, then using equation 1.8 to derive the area function.

Although Stansfield's method leads to unique area functions, the computation of the Hilbert and Fourier transforms is very time consuming. For this reason his method is not considered to be a satisfactory solution to the problem.

In conclusion, none of the techniques described in this chapter are considered to be satisfactory methods of determining the vocal tract area function from the acoustic speech signal. Two other techniques are known, which have been applied to this problem, namely those of ATAL (1971) and WAKITA (1972). These methods will be discussed in the next chapter.

CHAPTER TWO

LINEAR PREDICTION AND RELATED TECHNIQUES

Recently, much interest and research effort in speech analysis has been directed at parametric representations of the speech waveform, because they promise bandwidth savings in digital speech transmission. In this chapter, three of these techniques will be discussed, namely, maximum likelihood analysis (ITAKURA and SAITO 1968), linear prediction (ATAL and HANAUER 1971) and optimum inverse filtering (MARKEL 1972a, WAKITA 1972).

These three methods derive equivalent parameter sets because they all use several periods of the speech waveform to calculate the best all pole approximation to the speech spectrum. The parameters derived contain information about the vocal tract transfer function, the glottal excitation function and the lip radiation impedance. The vocal tract transfer function is known to be an all pole function (FANT 1970), but the glottal excitation has been shown to be an all zero function (MATTHEWS, MILLER and DAVID 1961). Different approximations to the lip radiation effects are possible, depending on the model used for this radiation. However, FLANAGAN (1972) has shown that treating the lip radiation as that of a simple spherical source leads to an all zero approximation. In general, the glottal and lip radiation zeros contribute mainly to the fine detail and general trend of the speech spectrum. However, they can also perturb the pole positions (in the s or z plane) of the vocal tract transfer function (FLANAGAN 1972). Although the vocal tract area function can be calculated from any of these parameter sets, it will have an inherent inaccuracy caused by inclusion of the glottal excitation and the lip radiation effects in the parameter sets. These inaccuracies can be partially avoided if the method derived in chapters three and four of this thesis is used.

2.1. MAXIMUM LIKELIHOOD ANALYSIS OF ITAKURA AND SAITO

ITAKURA and SAITO (1966) first reported their technique in a report written in Japanese. The first English publication of their method was a paper at the Sixth International Congress on Acoustics (ITAKURA and SAITO 1968). In their first paper the mathematical development of their method was very involved and amounted to some forty pages, however in this thesis it is only necessary to outline the technique used.

Basically, Itakura and Saito started by considering the speech samples s_i to be the result of a wide sense stationary stochastic process, governed by Gaussian statistics. They argued that this was a good approximation to the speech waveform even for the deterministic vowel sounds, because of pitch period fluctuations and changing formant patterns.

They considered the set $\underline{S}$ formed by the samples of a short segment of speech given by

$$\underline{S} = (s_1 s_2 s_3 \ldots . s_M) \tag{2.1}$$

These samples were assumed to be the result of a stationary Gaussian process and to have spectral density

$$T(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{\prod_{k=1}^{p} \left| 1 - \frac{z_k}{z^T} \right|^2} = \frac{\sigma^2}{2\pi} \frac{1}{\left| \sum_{k=-p}^{p} \alpha_k z^{-kT} \right|^2} \tag{2.2}$$

where

$z = \exp(j\omega)$,
z_k = kth. pole of speech (i.e. kth. formant),
α_k = kth. linear regression coefficient,
T = sampling period,
ω = angular frequency.

They next defined the autocovariance matrix ($\underline{\underline{\phi}} = [\phi_k]$) of the speech samples ($\underline{S} = [s_i]$), by specifying the elements of this matrix $\underline{\underline{\phi}}$ as

$$\phi_k = \sum_{j=1}^{p} \alpha_j \alpha_{j+|k|} \qquad k = 1,2,3,\ldots,M \tag{2.3}$$

Using this definition they were able to rewrite the equation for the speech spectrum from equation 2.2 as

$$T(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{\sum_{k=-p}^{p} \phi_k \cos k\omega T} \tag{2.4}$$

From this last equation it can be seen that the spectral density $T(\omega)$ of the random gaussian process is specified by the parameters σ and the

ϕ_k's. Itakura and Saito were thus interested in calculating the parameters $[(\sigma, \phi_k)\ k = 1,\ldots M]$, or alternatively the linear regression coefficients (α_k) which gave the best approximation of the stochastic spectrum to the actual speech spectrum. They showed that either of these sets of parameters could be calculated by minimising the likelihood ratio

$$L(T/\hat{T}) = -N \log 2\pi + \frac{N}{4\pi} \int_{-\pi}^{\pi} (\log T(\omega) + \frac{\hat{T}(\omega)}{T(\omega)}) d\omega \qquad (2.5)$$

relating the stochastic spectrum T to actual speech spectrum $\hat{T}$. After much mathematical development they were able to show that minimisation of the previous equation was equivalent to minimising the energy difference between the logarithmic spectra given by

$$E(T/\hat{T}) = \int_{-\pi}^{\pi} 2 \left\{ D(\omega) + \exp(-D(\omega)) - 1 \right\} d\omega \qquad (2.6)$$

where

$$D(\omega) = \log(T(\omega)) - \log(\hat{T}(\omega))$$

They also derived a method for calculating the set of parameters α_k which minimised this energy difference. These parameters α_k are the same as the linear prediction coefficients in Atals work, (see Section 2.2). Hence the vocal tract area function calculated from this parameter set would suffer from the same limitations inherent in Atals method. These limitations are discussed in the next section. Itakura and Saito have not published the application of their method for calculating the vocal tract area function. However, in their method for calculating the linear regression coefficients (α_k), they use as an intermediate parameter set, some partial autocorrelation coefficients (Parcor coefficients). WAKITA (1972) has shown that these Parcor coefficients are equivalent to the reflection coefficients in a transmission line model of the vocal tract. This means the area function could be calculated without the necessity for calculation of the linear regression coefficients. However, the method still suffers from the inherent limitations of Atals method described later, in the next section.

2.2. LINEAR PREDICTION METHOD OF ATAL

In 1968 ATAL and SCHROEDER proposed an alternative technique to that outlined in the previous section, based on linear prediction. This method was developed in another paper ATAL and HANAUER (1971). Atal assumed that speech was linearly predictable, so that the current speech sample could be approximated by a weighted sum of previous speech samples. The weighting factors or linear prediction coefficients, were calculated by minimising the energy difference between the actual speech samples and the predicted speech samples for several pitch periods.

Let us consider the approximations Atal made by considering speech samples to be linearly predictable. The glottal spectrum U_G is known to be an all zero function (MATTHEWS, MILLER and DAVID 1961) so in z transform notation one has

$$U_G(z) = K_1 \prod_{i=1}^{n} (1 + \frac{Z_i}{z^T}) \tag{2.7}$$

where

z = $\exp(j\omega)$
T = sampling interval,
K_1 = a constant dependent on the vocal effort used,
z_i = the zero's of the glottal source.

The radiation impedance at the lips can be approximated by a simple spherical source (FLANAGAN 1972) yielding

$$Z_L(z) = K_2(1 - z^{-T}) \tag{2.8}$$

where K_2 is a constant dependent on the lip area.
This means the combined effects of glottal excitation and lip radiation can be approximated by the all zero function

$$U_G(z)Z_L(z) = K_1K_2(1 - z^{-T}) \prod_{i=1}^{n} (1 + \frac{z_i}{z^T}) \tag{2.9}$$

Atal approximated the all zero function of equation 2.9 by the all pole function

$$U_G'(z)Z_L'(z) = \frac{K_1K_2}{(1 + (1 - z_a)z^{-T})(1 - z_b z^{-T})} \tag{2.10}$$

Further insight into this approximation used by Atal can be obtained by considering the binomial expansion of one of the poles of equation 2.10, i.e.

$$\frac{1}{(1 - z_b z^{-T})} = 1 + \frac{z_b}{z^T} + (\frac{z_b}{z^T})^2 + (\frac{z_b}{z^T})^3 + \ldots\ldots \qquad (2.11)$$

Providing the pole z_b of equation 2.10 lies near the origin, and the pole z_a of equation 2.10 lies near the unit circle, equation 2.10 can be simplified by using the binomial expansion as in equation 2.11 to give

$$U_G'(z)Z_L'(z) = K_1 K_2 (1 + z_b z^{-T})(1 + (z_a - 1)z^{-T}) \qquad (2.12)$$

by neglecting terms of order $(\frac{z_b}{T})^2$. Now if the zeros of equation 2.9 lie near the origin terms of the order $(\frac{z_i}{T})^3$ can be neglected and equation 2.12 is then an approximation to equation 2.9. Hence for Atals approximation (i.e. equation 2.10) to be valid the pole z_a must lie near the unit circle, the pole z_b must lie near the origin and the zeros z_i must lie near the origin, of the z plane. Hence although equation 2.10 may be adequate to account for the perceptual effects of the glottal source and lip radiation in a speech synthesiser (e.g., transmission line synthesiser FANT 1970), I do not believe it is an acceptable approximation for an analysis technique.

However, Atal used equation 2.10 and included conjugate pole pairs to account for the vocal tract transfer function, to approximate the speech spectrum by

$$T(z) = \frac{K_1 K_2}{(1 + (1 - z_a)z^{-T})(1 - z_b z^{-T}) \sum_{k=1}^{n} a_k z^{-kT}} \qquad (2.13)$$

where a_k are the coefficients of a polynomial expansion of the all pole vocal tract transfer function. Now the assumption that the speech spectrum is given by equation 2.13, is equivalent to considering the speech waveform to be the impulse response of a digital recursive filter.

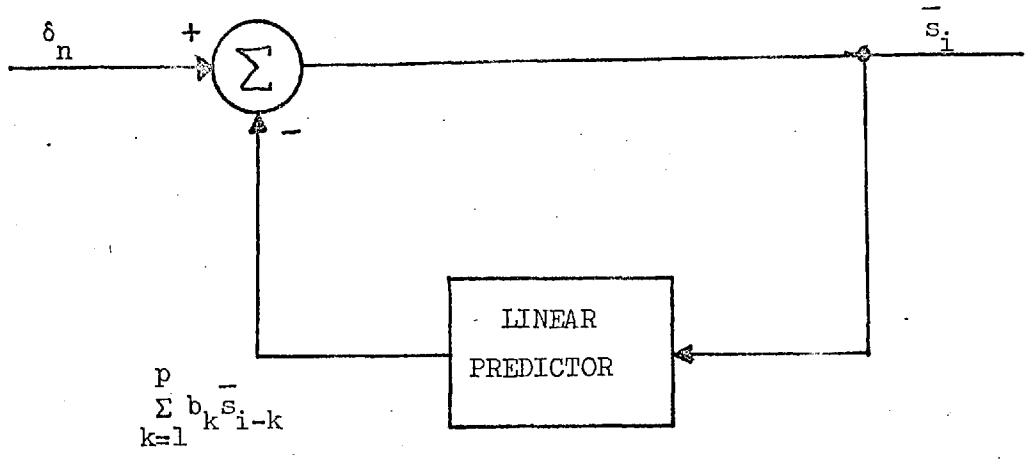


FIGURE 2.1 Model of Speech Production Process Based on Linear Prediction

In Figure 2.1 the speech production process is modelled as the impulse response of a digital recursive filter. The linear predictor, with coefficients b_k, forms a weighted sum of previous speech samples which approximates the current speech sample. The impulse δ_n represents the excitation at the beginning of each pitch period. Hence except for the sample at the beginning of each pitch period the predicted samples $\bar{s}_i$ are given by

$$\bar{s}_i = \sum_{k=1}^{p} b_k s_{i-k} \qquad (2.14)$$

In order to evaluate the predictor coefficients b_k, $k=1,2,3,\ldots.p$ Atal expressed the mean square error between the predicted samples $\bar{s}_i$ and the actual speech samples s_i as

$$\langle E_i^2 \rangle = \langle (s_i - \sum_{k=1}^{p} b_k s_{i-k})^2 \rangle \qquad (2.15)$$

where $\langle\ \rangle$ denotes average.

This expression was minimised to calculate the predictor coefficients, by evaluating the differentials

$$\frac{\partial \langle E_i^2 \rangle}{\partial b_j}$$

for each b_j and then setting these differentials to zero.

Now

$$\frac{\partial < E_i^2 >}{\partial b_j} = -2 < s_i s_{i-j} > + 2 < \sum_{k=1}^{p} b_k s_{i-k} s_{i-j} > \quad (2.16)$$

and setting each of these differentials to zero yields

$$\begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} \cdots\cdots \phi_{1p} \\ \phi_{21} & \phi_{22} \cdots\cdots \phi_{2p} \\ \vdots & \vdots \\ \phi_{p1} & \phi_{p2} \cdots\cdots \phi_{pp} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_p \end{bmatrix} \quad (2.17)$$

where

$$\phi_j = < s_i s_{i-j} >$$

and

$$\phi_{jk} = < s_{i-j} s_{i-k} > = \phi_{kj} \quad (2.18)$$

The matrix ϕ having elements ϕ_{jk} is a symmetric square matrix of order p x p. Furthermore, inspection of its elements shows that it is the autocovariance matrix of the speech samples. Hence the predictor coefficients can be calculated by solving equation 2.17. Atal gives a computationally efficient method for solving equation 2.17 in Appendix C of his paper (ATAL and HANAUER 1971).

Atal used a predictor of order ten (i.e., p = 10) and a speech sampling rate of 10 KHz. He calculated the autocovariance coefficients ϕ_{jk} from equation 2.18, by averaging for one pitch period for voiced speech, and ten milliseconds for unvoiced speech. Based on this technique of linear prediction Atal proposed a speech transmission system. In this system the transmission channel was required to carry the calculated predictor coefficients, the error between the predicted speech sample and the actual speech sample, and the pitch frequency for voiced sounds. At the receiving end of the transmission system the predicted speech samples were calculated and added to the error signal to resynthesise the speech. Atal found that reasonable quality speech was obtained by transmitting 2400 bits/second, when 7200 bits/second were transmitted the

resynthesised signal was of comparable quality to the original speech bandlimited to 5 KHz.

One further point is worth mentioning at this juncture. Atal reported that occasionally the predictor coefficients calculated produced poles in the recursive filter transfer function which were outside the unit circle in the z plane. When this happened an unstable filter resulted. He further proposed a method which shifted any poles which appeared outside the unit circle, back inside the unit circle (Appendix D, ATAL and HANAUER, 1971.). This instability could arise on two accounts.

i) If the model used was not an adequate representation of the speech production process.

or ii) If numerical errors introduced in calculating the predictor coefficients shifted poles which should appear just inside the unit circle outside the unit circle.

I think the former of these two possibilities is more likely, because of the approximation of zeros, by poles, used by Atal (See equations 2.9 and 2.10).

Atal also proposed a technique for calculating the vocal tract area function from the predictor coefficients. In this technique the predictor coefficients (b_k) were assumed identical to the coefficients of a polynomial expansion of the vocal tract transfer function (a_k's in equation 2.13). The vocal tract was then assumed to be an acoustic transmission line consisting of a number of constant cross sectional area sections, each section having equal length. A method similar to that derived in Section 3.6 of this thesis was then used to calculate the reflection coefficients of the transmission line model, from the predictor coefficients. The vocal tract area function was then calculated from these reflection coefficients. In his derivation Atal assumed that the lips were terminated in a unit acoustic resistance. Hence, from the discussion in Section 1.8 it is known that his method will lead to unique area functions.

However, Atal's method is not considered to be a satisfactory technique for calculating the vocal tract area function, because of the following reasons:-

a) The method approximates the whole speech production process by a filter with an all pole transfer function, when it is known that the glottal source for voiced sounds, and the lip radiation are all zero functions.

b) The predictor coefficients calculated contain information about the glottal source, the lip radiation impedance and the vocal tract transfer function. However, when Atal uses them to calculate the area function he implicitly assumes they only contain information about the vocal tract transfer function.

c) Sometimes the calculated predictor coefficients correspond to an unstable filter, which means in these cases they cannot be used to calculate the vocal tract area function.

The next section discusses an alternative technique which is conceptually simpler, but still suffers from some of the limitations mentioned above.

2.3. OPTIMUM INVERSE FILTERING METHOD OF MARKEL

A method which has become known as optimum inverse filtering was first described by MARKEL (1971) and was also published in abbreviated form MARKEL (1972a). Markel developed his method in order to find a better estimate of the speech formant structure than was available directly from the speech spectrum. His method was later adapted by Wakita for estimating the vocal tract area function, (WAKITA (1972), WAKITA (1973)).

Markel defined the optimum inverse filter as the all zero filter whose transfer function is given by

$$C(z) = \sum_{k=0}^{M} c_k z^{-kT} \qquad (2.19)$$

where the c_k's are the optimum inverse filter coefficients.
In order to calculate the coefficients of the optimum inverse filter for a given speech segment, Markel defined the criterion that the inverse filter should transform the input speech samples into the best estimate of white noise at the output.

The samples x_i at the output of the inverse filter are formed by convolution of the filter impulse response with the speech samples s_i at its input, i.e.,

$$x_i = \sum_{k=0}^{i} c_k s_{i-k} \qquad c_0 = 1,\ c_i = 0 \quad i > M \qquad (2.20)$$

Markel also assumed the vocal tract to be excited by an impulse δ_{io}, at the pitch frequency. He therefore defined the error criterion as minimising the mean square difference between the output of the inverse

filter and this impulse, i.e.,

$$e = \sum_{i=0}^{N+M} (x_i - \delta_{io})^2 \qquad (2.21)$$

where δ_{io} is the Krönecker delta, $\delta_{io} = 1$, $i = 0$; $\delta_{io} = 0$, $i \neq 0$.

By assuming the speech signal to be causal and time limited to N samples, substitution from equation 2.20 gives

$$e = \sum_{i=0}^{N+M} \left(\sum_{k=0}^{M} c_k s_{i-k} - \delta_{io} \right)^2 \qquad (2.22)$$

In order to find the values of the inverse filter coefficients c_k which minimised the mean square error, Markel evaluated the differentials.

$$\frac{\partial e}{\partial c_j} = 2 \sum_{k=0}^{M} c_k \sum_{i=0}^{N+M} s_{i-k} s_{i-j} - 2 \sum_{i=0}^{N+M} s_{i-j} \delta_{io} \qquad (2.23)$$

$$j = 0, 1, 2, 3, \ldots, M$$

Markel simplified this equation by identifying the short term autocorrelation functions as

$$r_{k-j} = \sum_{i=0}^{N} s_i s_{i+|k-j|} = \sum_{i=0}^{N+M} s_{i-k} s_{i-j} \qquad (2.24)$$

for a timelimited signal $s_i = 0$, $0 > i > N$.

To find the coefficients which minimise the error, equations 2.23 are set equal to zero. Using this condition and substituting from equation 2.24 yields

$$\sum_{k=0}^{M} c_k r_{k-j} = 0 \qquad (2.25)$$

$$j = 1, 2, 3, \ldots M$$

and

$$\sum_{k=0}^{M} c_k r_k = s_0 \qquad \text{for} \qquad j = 0 \qquad (2.26)$$

Normalising the autocorrelation coefficients by dividing each r_k by r_0 allows equations 2.25 and 2.26 to be expressed as the single matrix relation.

$$(1,0,0,\ldots\ldots,0) = (c_0, c_1, c_2,\ldots c_M)[R_M] \qquad (2.27)$$

where $[R_M]$ is the M + 1 x M + 1 matrix of normalised autocorrelation coefficients r_k/r_0.

Markel used Levinsons solution of equation 2.27, which is reported in ROBINSON (1967). Levinson used the symmetry of the autocorrelation matrix R_M to define an efficient recursive process for solving equation 2.27, this process will now be briefly outlined. At step m in the recursive process the normalised autocorrelation matrix R_m is known, given by

$$[R_m] = \begin{bmatrix} r'_0 & r'_1 & r'_2 & \ldots & r'_m \\ r'_1 & r'_0 & r'_1 & \ldots & r'_{m-1} \\ \vdots & & & & \vdots \\ r'_m & \ldots & \ldots & \ldots & r'_0 \end{bmatrix} \qquad (2.28)$$

where $r'_k = r_k/r_0$

In order to go from step m to step m+1 two auxiliary quantities (u_m and v_m) are defined which satisfy the relation

$$(1, c_1^{(m)}, c_2^{(m)},\ldots,c_m^{(m)}, 0)[R_{m+1}] = (u_m, 0, 0,\ldots,0, v_m) \qquad (2.29)$$

The symmetry of the autocorrelation matrix $[R_{m+1}]$ (see equation 2.28) allows definition of another relation

$$(0, c_m^{(m)}, c_{m-1}^{(m)},\ldots,c_2^{(m)}, c_1^{(m)}, 1)[R_{m+1}] = (v_m,0,\ldots,0,u_m) \qquad (2.30)$$

Equations 2.29 and 2.30 are combined by multiplying 2.30 by a constant X_{m+1} and adding it to 2.29 to give

$$(1, c_1^{(m)} + X_{m+1} c_m^{(m)}, c_2^{(m)} + X_{m+1} c_{m-1}^{(m)}, \ldots, c_m^{(m)} + X_{m+1} c_1^{(m)})[R_{m+1}]$$

$$= (u_m + X_{m+1} v_m, 0, 0, \ldots . 0, v_m + X_{m+1} u_m) \qquad (2.31)$$

Equation 2.31 can be seen to be of equivalent form to equation 2.29 if

$$X_{m+1} = - \frac{v_m}{u_m} \qquad (2.32)$$

and

$$u_{m+1} = u_m + X_{m+1} v_m \qquad (2.33)$$

Equations 2.29 to 2.33 define the complete recursive process for calculating the inverse filter coefficients $c_k = c_k^{(M)}$, from the normalised autocorrelation coefficients $r'_k = r_k / r_0$. A detailed flow diagram of this recursive process is given on page 20 of MARKEL (1972).

Markel used this inverse filtering technique to obtain spectra which contained the gross spectral detail of the formant structure, but not the fine spectral detail of the glottal source. To calculate these "clean" spectra Markel realised the equivalence between equation 2.19 and the definition of the discrete Fourier transform. Using the substitution $z_m = \exp(j2\pi m/N_1)$ for $m=0,1,\ldots,N_1-1$, equation 2.19 can be rewritten as

$$C_M = C(z_m) = \sum_{k=0}^{N_1-1} c_k \exp(- j2\pi mk/N_1) \qquad (2.34)$$

Now equation 2.34 is precisely the definition of the N_1 point discrete Fourier transform of the inverse filter coefficients c_k. The inverse filter transfer function can therefore be calculated by using equation 2.34, and the "clean" spectrum is given by the reciprocal of this inverse filter transfer function. Markel used the peaks of the clean spectra to estimate the formant trajectories of speech.

The autocorrelation function is, strictly speaking, only defined for infinite limits on the summation. Markel has effectively truncated the series to give the short term autocorrelation function of equation 2.24. Truncation of the series in this way is equivalent to analysing the speech with a rectangular window. In order to avoid the well known difficulties of analysis with a rectangular window (BLACKMAN and TUKEY

1958), Markel applied a Hamming window to the speech samples before analysis. In the methods of ITAKURA (1968) and ATAL (1971) the covariance function is normally defined with finite limits, hence they did not need to apply a Hamming window to the speech samples before analysis.

Markel typically used a Hamming window of length 256 samples (i.e. N = 256) and a speech sampling rate of 10 KHz to calculate the first ten to fourteen inverse filter coefficients (i.e. M = 10 → 14). It is obvious from studying his results which typically show only four formant peaks, that many of the inverse filter coefficients are being used to account for the spectral trend of the source (one coefficient is needed for each pole and a conjugate pole pair appears as a formant peak in the speech spectra). This should be expected because the output of the inverse filter is constrained to be an impulse at the pitch frequency embedded in white noise. Therefore any method which uses Markel's inverse filter coefficients to calculate the vocal tract area function will suffer from the same inaccuracies as Atals method described in the previous section.

Wakita reported a method for calculating the vocal tract area function using Markels optimum inverse filtering technique (WAKITA 1972, WAKITA 1973). Wakita was able to show that the constant X_{m+1} of equations 2.29 → 2.33 was the m+1 th. reflection coefficient of an acoustic tube model of the vocal tract. In his model the first reflection coefficient corresponded to the junction nearest the lips, the lips were assumed terminated in a short circuit and the glottis was assumed to have a characteristic impedance $D_o c/A_{M+1}$. In chapter one it was shown that a sufficient condition for obtaining unique area functions was the choice of a resistive lip termination. It can be shown by a similar arguement that the choice of a resistive glottal termination is an alternative sufficient condition for obtaining unique area functions. Therefore, Wakita's choice of boundary conditions leads to a unique area function.

Wakita was able to show the equivalence between the reflection coefficients (ρ_m) of his transmission line model and the constant X_{m+1} in Levinson's solution of the autocorrelation matrix (equations 2.29 - 2.33), by considering the filtering action of the transmission line model. He used a technique similar to that reported in section 3.5 of this thesis to evaluate the filtering action of the vocal tract transmission line model. The derivation will not be repeated here because it is well explained on pages 38-53 of WAKITA (1972), the result will however

be stated. Wakita proved that the equations relating the inverse filter coefficients $c_o^{(M)}$ of an M section vocal tract to the coefficients $c_o^{(M+1)}$ of an M+1 section tract (formed by adding an extra tube of the glottis end) were

$$\begin{aligned} c_o^{(M+1)} &= c_o^{(M)} = 1 \\ c_{M+1}^{(M+1)} &= \rho_M \\ c_k^{(M+1)} &= c_k^{(M)} + \rho_M c_{M+1-k}^{(M)} \end{aligned} \qquad (2.35)$$

where ρ_M is the Mth. reflection coefficient of the vocal tract transmission line model.

Comparison of equation 2.35 with equations 2.31 and 2.29 shows that ρ_M is equal to X_{M+1}. Wakita was thus able to calculate the vocal tract area function directly from the quantities X_{M+1} in Levinson's method. The area function calculated will have an error caused by inclusion of the glottal source function and the lip radiation impedance in the inverse filter coefficient calculations.

The main points made so far in this chapter will now be summarised:-

i) Wakita and Atal have independently derived two equivalent methods of calculating the vocal tract area function from the speech waveform.

ii) Wakita uses Markels autocorrelation method and hence needs to apply a Hamming window to the speech samples.

iii) Atal uses a covariance formulation and does not need to apply a Hamming window.

iv) Both methods include the effects of glottal source and lip radiation in the parameter set, by choosing criteria which minimise the error energy over several pitch periods.

v) Of the two methods Wakita's is computationally the more efficient and probably conceptually the easier.

vi) Both methods lead to unique area functions.

In the next section some of the other uses for these parametric representations of the speech waveform are presented.

2.4. OTHER APPLICATIONS

So far, we have mentioned the main uses which have been suggested for the representation of speech in terms of either linear regression

coefficients, or predictor coefficients, or optimum inverse filter coefficients. However, numerous other papers have been published on these subjects recently. Amongst this multitude of papers have been suggestions for alternative formulations of the problem, modifications which make it easier to extract certain parameters from the methods, or in the majority of cases a different mathematical treatment which supposedly make the methods more easily understood, (perhaps only to the author of the papers in question).

In his paper MARKEL (1971) was able to show that all methods were equivalent solutions of the same problem, by showing that Itakura's linear regression coefficients, Atals' predictor coefficients and his own optimum inverse filter coefficients were equal, to within irrelevant normalisation constants. Markel was also able to show that all methods were basically developments of the method originally attributable to the mathematician Prony, who in the late eighteenth century developed the idea of representing the natural expansion of gasses in terms of sums of exponentials. Prony's method, which must have been difficult for him to prove because of the long hand calculations involved, was extended to a least squares formulation as early as 1924 by Runge and Koenig. Markel was also able to show that all methods were in fact equivalent to a digital formulation of the Wiener Hopf equation which presents the same analysis requirements. Since then, numerous other origins for the methods have been claimed, including digital filter design, and processing of seismic and radar signals.

An alternative formulation of the problem was given by MARKEL and GRAY (1973) who found that an inner product formulation gave further insight to the previously described analysis method of MARKEL (1971). Also using this formulation they were able to estimate the accuracy which could be expected from a fixed point implementation of the method, and in particular the likelihood of instability resulting from the fixed point implementations.

MAKHOUL (1973) was able to show that the autocorrelation method could be derived from the analysis of stationary short time spectra, and that the covariance method was derivable as the analysis of two dimensional time varying short time spectra. Also in this paper another method was proposed for finding the formant frequencies as an alternative to peak picking the clean spectra proposed by Markel. This method proposed by Makhoul consisted of solving the polynomial of equation 2.19 specified in terms of the inverse filter coefficients, for its zeros from which

the poles of the speech spectrum may be calculated.

Another main area of interest in these methods has been pitch determination. MARKEL (1972b) formulated a simplified inverse filter technique for precisely this purpose, which he appropriately called the SIFT algorithm. In this method the speech was low pass filtered to 800 Hz and then an optimum inverse filter with only five coefficients determined. After calculating these coefficients the speech was passed through the SIFT filter and the autocorrelation of the output of the filter taken to give the pitch period estimate. MAKSYM (1973) on the other hand built a hardware simplified linear predictor whose output contained impulses separated by the pitch period.

Speech synthesis from stored predictor parameters has been another area of interest. MORRISS (1972) proposed an algorithm which permitted real time speech synthesis on a moderate sized digital computer and MERMELSTEIN (1972) proposed a method capable of producing synthetic nasalised speech.

Finally, WEINSTEIN and OPPENHEIM (1972) used linear prediction to reduce the data rate necessary in their homomorphic vocoder. Although the survey in this chapter is not claimed to be exhaustive, it is representative of the sort of applications and methods which have been proposed for this area of speech research, which is often referred to by the umbrella term "linear prediction". Regardless of the origin of the particular method, or the details of the method of solution, there is no doubt that these methods have caused considerable upheaval in the field of speech analysis recently. Indeed, no less than ten papers were presented on topics relating to "linear prediction" at the recent 1972 conference on speech communication and processing. It was also interesting to note that in the questionnaire distributed with registration forms for that conference, that out of the 62 research groups who replied 15 were actively engaged in the linear prediction field.

The various applications of these methods which are receiving so much interest are summarised below:-

a) Formant frequency analysis.

b) Low bit rate speech transmission and storage.

c) Pitch extraction.

d) Speech synthesis.

e) Vocal tract area function estimation.

In concluding this chapter the main points will be summarised. Of all the methods described in the literature surveys of this and the

previous chapter none were thought satisfactory for finding the vocal tract area function from the speech waveform, for the various reasons discussed earlier. If a choice has to be made between the previously described methods then based on computational efficiency, ease of understanding of the derivation, and the ability for it to determine unique realistic area functions from the speech waveform the method of Wakita seems to be the best. For these reasons the method of Wakita was chosen as a comparison for my method which is developed in the next two chapters. My method, although basically similar in concept to the methods described in this chapter, has the advantage of excluding the glottal excitation from the parameter set derived by careful choice of the memory of the filter, and analysis of short periods of the speech waveform taken during the closed glottis period of phonation.

CHAPTER THREE

TRANSMISSION LINE MODEL OF THE VOCAL TRACT

3.1. INTRODUCTION

In the first chapter of this thesis the mechanics of the speech production process were discussed and uses for articulatory parameters were presented. The latter half of the first chapter was devoted to a literature survey of cineradiographic studies of speech articulation and some early attempts to measure the vocal tract area function from the acoustic speech signal. Disadvantages of the cineradiographic articulation studies were presented, together with the inadequacies of the early indirect methods using acoustic signals for their starting point. The second chapter was devoted to parametric representations of the speech waveform which allowed the vocal tract area function to be estimated. In particular it was pointed out that all these methods suffered from inaccuracy in the derived area function, mainly caused by inclusion of the effects of glottal excitation in the parameter sets. Of the methods presented WAKITA's (1972) was preferred, mainly because of its computational efficiency.

This chapter is devoted to the derivation of a technique for calculating the vocal tract area function. The vocal tract is first considered as a transmission line. The transfer function of this transmission line is derived and shown to be all pole. Next a purely recursive filter is defined, which has the same transfer function as the transmission line model. A method is then developed for finding the reflection coefficients of the transmission line from the coefficients of the recursive filter. Finally, it is shown how the vocal tract cross sectional area function may be calculated from these reflection coefficients.

A technique for calculating the recursive filter coefficients from samples of the speech waveform is discussed in the next chapter (chapter 4).

3.2. PROPERTIES WHICH ALLOW THE VOCAL TRACT TO BE CONSIDERED AS A TRANSMISSION LINE

In section 1.3 of this thesis it was stated that production of voiced non-nasalised sounds could be thought of as the excitation of an acoustic resonant filter (the vocal tract), by a volume velocity source at the

glottis. The importance of the sizes of the vocal tract cavities in determining the sound produced was stressed. Now we are interested in calculating the vocal tract area function (which defines the position of maximum constriction and the degree of this constriction) from a given speech sound.

At first sight this may seem to be an insoluble problem, because we are only given the output of the system and not its input. However, although we cannot specify exactly the form of the input given only the speech waveform, we can make use of properties of the excitation, and properties of the vocal tract transfer function, to allow us to do the necessary deconvolution. Once we have calculated the vocal tract transfer function by this deconvolution process, we can then calculate the vocal tract area function. The first step in developing a method for calculating the vocal tract area function will be to model the vocal tract as an acoustic transmission line.

The fundamental analogies and relations which allow the vocal tract to be considered as an acoustic transmission line have been known for some time. A comprehensive presentation of these relations can be found in STANSFIELD (1971). Here we merely state, with some justification, the assumptions made about vocal tract transmission; which allow the vocal tract to be treated as an acoustic transmission line.

a) Transmission in the vocal tract takes the form of longitudinal propagation in the frequency range of interest (up to 5KHz). This assumption is valid as long as the greatest cross dimension is less than half a wavelength and the tube does not flare too rapidly (FLANAGAN 1972). For frequencies up to 5 KHz, the geometry of the vocal tract is such that the cross dimensions are appreciably less than a wavelength; hence the assumption of cophasic wavefronts propagating in a longitudinal mode is justified.

b) As will be explained later, records of approximately three milliseconds of the speech waveform are used to measure the vocal tract transfer function. During this period because of the inertia of the articulators, the vocal tract is essentially stationary. This means that the vocal tract can be represented by a sequence of stationary shapes. Consideration of the vocal tract as stationary during an analysis is essential for the development to be presented.

c) Instead of the smoothly varying cross section of the actual vocal tract, a piecewise constant model is used. This piecewise approximation takes the form of abutting tubes of constant cross sectional area. Discreteness of the model is necessary because the data is obtained in

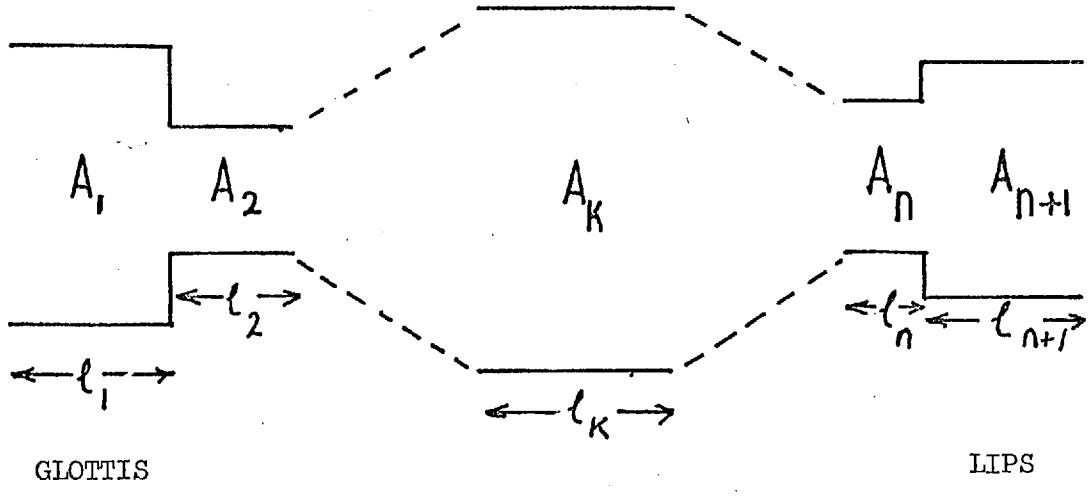
sampled form. As will be shown in section 3.4 the use of sampled data permits definition of only ten equal length tubes for a 17 cm vocal tract and a sampling rate of 10,000 samples per second. Also for a tube of constant cross section the characteristic impedance of the tube is related to its area by equation 1.8 of this thesis.

d) The vocal tract is assumed lossless. This lossless assumption is necessary for the development of the transmission matrix used in this chapter. In the real vocal tract, losses are present. They arise from the passage of air past the vocal tract walls (heat losses), vocal tract wall vibration and radiation from the throat and cheeks. Methods of correcting the area function to account for the losses are discussed in chapter five of this thesis.

e) Individual tube lengths in the model, bear a commensurability relationship with each other, i.e. the length of each tube is an integer multiple of the length of the shortest tube. If the data is available in the form of uniform time samples, it will normally be used to calculate the areas of a number of equal length tubes. However, as will be explained in section 3.4, the sampling rate need only define the minimum length resolution, and tubes with multiples of this minimum length can be included within the framework of the analysis presented here.

3.3. TRANSMISSION LINE MODEL OF THE VOCAL TRACT

The vocal tract is modelled as (n+1) abutting, lossless, constant cross sectional area tubes, as shown below (Figure 3.1)



The kth. tube has cross sectional area A_k and length ℓ_k

FIGURE 3.1. Acoustic Tube Model of the Vocal Tract

It is further assumed that the vocal tract is stationary for the duration of a single analysis. Consider the effect of this structure on a wave of volume velocity injected at the glottis (at time t = 0) as it travels towards the lips. The wavefront will be unaffected until it reaches the first junction, because the component tubes are assumed lossless. At the first junction, some of the wavefront will be reflected and some transmitted, setting up a backward travelling wave in the first section and a forward travelling wave in the second section respectively.

This process is repeated at each junction in the tract model until, after a time,

$$T_{tot} = \sum_{k=1}^{n+1} \frac{\ell_k}{c} \qquad (3.1)$$

(where c is the velocity of sound in the vocal tract)
there will be forward and backward travelling waves in each section of the model.

Looking at the junction between the kth. and (k+1)th. sections as in figure 3.2, we can formulate the effect of a single junction on the waves travelling in the tract model.

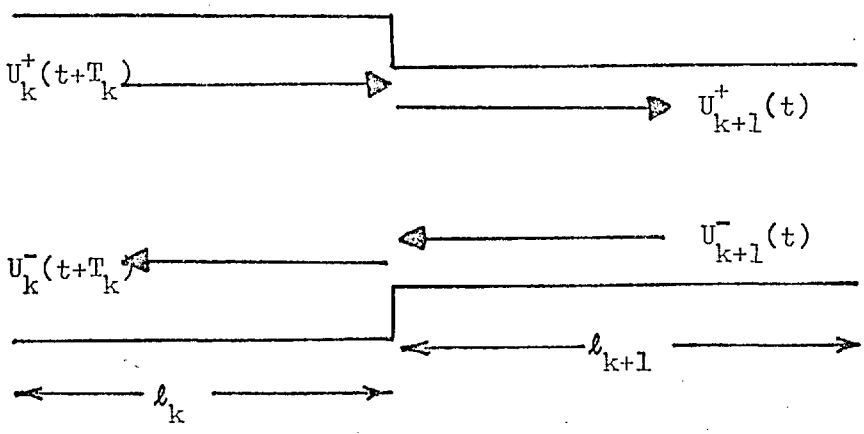


FIGURE 3.2. Volume Velocity Components in the kth and (k+1)th Sections

NOTE ON NOTATION IN THE FIGURE. The notation is such that the subscript refers to the section the wave is travelling in; the superscript

refers to the direction the wave is travelling (+ away from glottis, - towards glottis) and the time in brackets refers to the time that the wavefront was at the beginning (left) of the subscripted section. Also $T_k = \frac{\ell_k}{c}$ is the propagation delay of the kth. section.

Applying the equations of continuity of volume velocity and pressure at the kth. junction shown in figure 3.2, we obtain:

Continuity of volume velocity:-

$$U_k^+(t - T_k) + U_k^-(t + T_k) = U_{k+1}^+(t) + U_{k+1}^-(t) \qquad (3.2)$$

Continuity of pressure:-

$$P_k^+(t - T_k) + P_k^-(t + T_k) = P_{k+1}^+(t) + P_{k+1}^-(t) \qquad (3.3)$$

The vocal tract is excited at the glottis by a volume velocity source. The pressure measured by a microphone is formed by the passage of the output volume velocity ($U_{n+1}^+(t)$), through the radiation load at the lips. In order to define the volume velocity transfer function we need to express equation 3.3 in terms of volume velocity. To do this, we utilise an acoustic formulation of Ohms law where the characteristic impedance is substituted from equation 1.8 and the direction of current or volume velocity is taken into account to give:

$$P_k^+(t - T_k) = U_k^+(t - T_k)\frac{D_o c}{A_k} \; ; \quad P_k^-(t + T_k) = -U_k^-(t + T_k)\frac{D_o c}{A_k}$$

$$P_{k+1}^+(t) = U_{k+1}^+(t)\frac{D_o c}{A_{k+1}} \; ; \quad P_{k+1}^-(t) = -U_{k+1}^-(t)\frac{D_o c}{A_{k+1}} \qquad (3.4)$$

where

D_o = density of air in the vocal tract

Using equations 3.4, the formula of 3.3 can be rewritten as

$$\frac{1}{A_k}(U_k^+(t - T_k) - U_k^-(t + T_k)) = \frac{1}{A_{k+1}}(U_{k+1}^+(t) - U_{k+1}^-(t)) \qquad (3.5)$$

At this point, the equations are simplified by making Kinariwala's transformation $z = e^{j\omega}$. This transformation is used in KINARAWALA's paper (1966) to derive a method of synthesising a transmission line from a required input impedance function. Another algorithm for synthesising a transmission line from its input reflection function was reported by GERSHO and KINARIWALA (1968).

Using Kinariwala's transformation, equations 3.2 and 3.5 become

$$U_k^+(z)\, z^{-T_k} + U_k^-(z)\, z^{+T_k} = U_{k+1}^+(z) + U_{k+1}^-(z) \qquad (3.6)$$

$$U_k^+(z)\, z^{-T_k} - U_k^-(z)\, z^{+T_k} = \frac{A_k}{A_{k+1}} \left(U_{k+1}^+(z) - U_{k+1}^-(z)\right) \qquad (3.7)$$

Returning to figure 3.2, we look at the physical interpretation of equations 3.6 and 3.7 by considering the reflections at the kth. junction. The wavefront U_k^+ is approaching the kth. junction from the left; some of it will be reflected and some transmitted at the junction. Also, the wavefront U_{k+1}^- is approaching the junction from the right, some of which will be reflected and some transmitted. The reflected portion of U_k^+ will be added to the transmitted portion of U_{k+1}^- to give rise to U_k^-; similarly the transmitted portion of U_k^+ will be added to the reflected portion of U_{k+1}^- to give rise to U_{k+1}^+. The transfer matrix for the kth. junction of the model will now be derived in a form which maintains this physical insight. Adding equations 3.6 and 3.7 gives:

$$2U_k^+(z)\, z^{-T_k} = \frac{A_k + A_{k+1}}{A_{k+1}} U_{k+1}^+(z) + \frac{A_{k+1} - A_k}{A_{k+1}} U_{k+1}^-(z) \qquad (3.8)$$

Subtracting equation 3.7 from equation 3.6 gives

$$2U_k^-(z)\, z^{+T_k} = \frac{A_{k+1} - A_k}{A_{k+1}} U_{k+1}^+(z) + \frac{A_{k+1} + A_k}{A_{k+1}} U_{k+1}^-(z) \qquad (3.9)$$

Multiplying equation 3.7 by A_{k+1}/A_k and subtracting the result from equation 3.6 yields

$$\frac{A_k - A_{k+1}}{A_k} U_k^+(z) z^{-T_k} + \frac{A_k + A_{k+1}}{A_k} U_k^-(z)\, z^{+T_k} = 2U_{k+1}^-(z) \qquad (3.9a)$$

Defining the reflection coefficient R_k as

$$R_k = \frac{Z_{k+1} - Z_k}{Z_{k+1} + Z_k} \qquad (3.10)$$

(JOHNSON 1950)

where Z_k is the characteristic impedance of the kth. section. Substituting for the characteristic impedance from equation 1.8 yields

$$R_k = \frac{A_k - A_{k+1}}{A_k + A_{k+1}} \qquad (3.11)$$

Using equation 3.11, equations 3.8 and 3.9a can be simplified and expressed as the single matrix relation

$$\begin{bmatrix} U^+_{k+1}(z) \\ U^-_k(z) \end{bmatrix} = \begin{bmatrix} (1 - R_k)z^{-T_k} & R_k \\ -R_k z^{-2T_k} & (1 + R_k)z^{-T_k} \end{bmatrix} \begin{bmatrix} U^+_k(z) \\ U^-_{k+1}(z) \end{bmatrix} \qquad (3.12)$$

This matrix relation defines a single element in a ladder digital filter model of the vocal tract.

3.4. LADDER DIGITAL FILTER MODEL OF THE VOCAL TRACT

We can see that equation 3.12 specifies the relation between the components of volume velocity in adjacent sections of the transmission line model. By careful study of equation 3.12 we can derive the ladder digital filter model for the vocal tract. We see that for each section two paths are required, one for the forward travelling wave in that section and one for the backward travelling wave in that section. Also, each forward going wave $U^+_{k+1}(z)$ is made up of the summation of the forward going wave $U^+_k(z)$ delayed by T_k and multiplied by $(1 - R_k)$; and the backward going wave $U^-_{k+1}(z)$ multiplied by R_k. Similarly, the backward travelling wave $U^-_k(z)$ is made up of the forward travelling wave $U^+_k(z)$ delayed by $2T_k$, multiplied by $-R_k$; added to the backward travelling wave $U^-_{k+1}(z)$ delayed by T_k and multiplied by $(1 + R_k)$.

Each section of the ladder filter consists of two delays, two adders and four multipliers. A cascade of (n+1) such sections is shown diagrammatically in Figure 3.3.

In order to complete the model, we will consider the terminations at

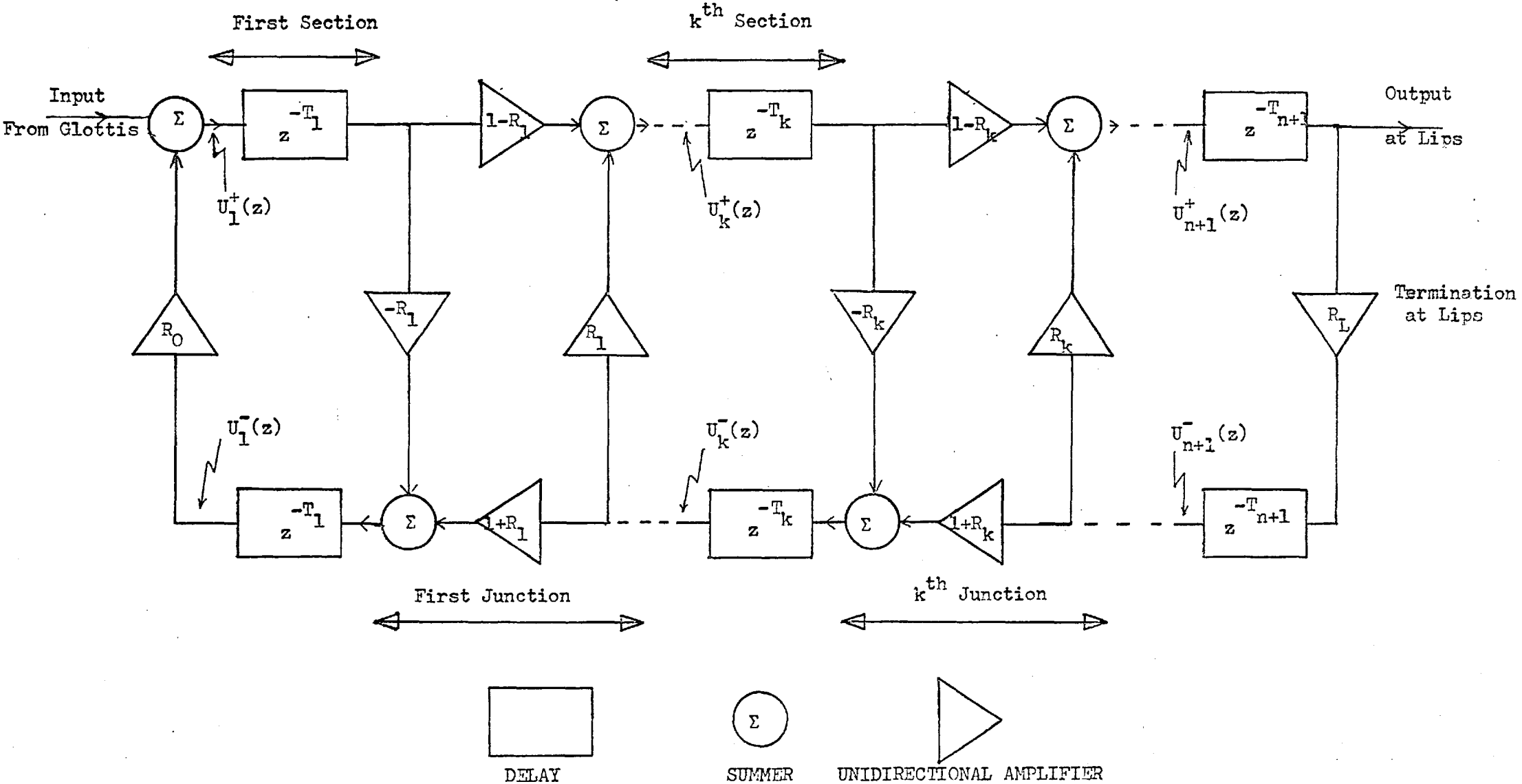


Figure 3.3 Digital Ladder Network Model of the Vocal Tract.

the lips (R_L) and at the glottis (R_o). In chapter one it was shown that termination of the model at the lips with a finite resistance was a sufficient condition for obtaining a unique area function. An alternative sufficient condition was stated in chapter two, namely termination at the glottis with a finite resistance. WAKITA (1972) used the latter of these two boundary conditions, in deriving his method for finding the vocal tract area function, from the coefficients of MARKEL's (1971) "optimum inverse filter". Wakita also chose a short circuit lip termination. The choice of a short circuit lip termination raises the difficulty of physical interpretation, because no pressure radiation is possible from a short circuit lip termination.

It will be shown in the next chapter that analysis of closed glottis periods of speech, avoids inclusion of the effects of sub glottal coupling. Analysis during the closed glottis period (corresponding to an open circuit glottal termination) necessitates the choice of the former of the two sets of boundary conditions.

Putting the closed glottis condition in equation 3.11 gives

$$R_o = \frac{A_o - A_1}{A_o + A_1} = -1 \text{ because } A_o = 0 \qquad (3.13)$$

We must now choose a finite resistance termination at the lips. To simplify the ensueing derivation we choose a matched resistive termination at the lips (i.e. $Z_L = \frac{D_o c}{A_{n+1}}$ or the area of the infinite length lip tube is equal to the area of the section nearest the lips, in our acoustic tube model). Substitution of this condition in equation 3.11 gives

$$R_L = \frac{A_{n+1} - A_L}{A_{n+1} + A_L} = 0 \text{ because } A_L = A_{n+1} \qquad (3.14)$$

The choice of a matched termination simplifies the definition of the transfer function of the transmission line model. Before considering this transfer function we will look at the impulse response of the ladder digital filter model for the vocal tract. Consideration of the vocal tract as a ladder digital filter was first used by KELLY and LOCHBAUM (1962) in their phoneme driven speech synthesiser.

Let us now consider the passage of an impulse of volume velocity injected at the glottis (at time t = 0) as it travels towards the lips.

Before any output can occur the impulse has to travel along the top path of figure 3.3. This means that the first output will occur at time T_{tot} given by

$$T_{tot} = \sum_{k=1}^{n+1} T_k \qquad (3.15)$$

It was stated in section 3.2e that the lengths of component tubes in our model were all integer multiples of the shortest tube length. This means the second output is formed by reflection of the impulse at the junction following the shortest section, passage of the reflected portion of the impulse back through the shortest section, reflection once more at the junction preceeding the shortest section and transmission of the resulting attenuated impulse to the lips. In the more general case, when a number of tubes have the same length as the shortest section, the second output of the filter would be formed by summation of the attenuated components (corresponding to the one considered) for each of the shortest sections.

Studying this behaviour and the time lapse between subsequent outputs in a similar manner, we see that the outputs will be separated in time by $2T_{min}$ (where T_{min} is the propagation delay of the shortest section). If we take samples at the rate F_s per second then we limit T_{min} to the value.

$$T_{min} = \frac{1}{2F_s} = \frac{\ell_{min}}{c} \qquad (3.16)$$

Before we can consider more specifically the magnitudes of the outputs of the ladder filter we must make a further assumption about the ratios of the lengths of the component tubes. We will consider first the simplest case when all the tubes have equal lengths. In this case equation 3.16 reduces to

$$F_s = \frac{c.(n+1)}{2L} \qquad (3.17)$$

where L = total length of the vocal tract = $(n+1)\ell_{min}$.

WAKITA (1972) used the above equation to derive the simple rule of thumb.

number of sections = sampling frequency in KHz (3.18)

for a 17 cm vocal tract, and c = 34000 cm/sec.

3.5. IMPULSE RESPONSE OF THE LADDER DIGITAL FILTER

We will now consider the impulse response of a specific ladder digital filter. From this impulse response we will deduce a method for calculating the reflection coefficients of the vocal tract transmission line model. The case we will consider is a ladder digital filter consisting of three equal length sections, which is shown below (figure 3.4).

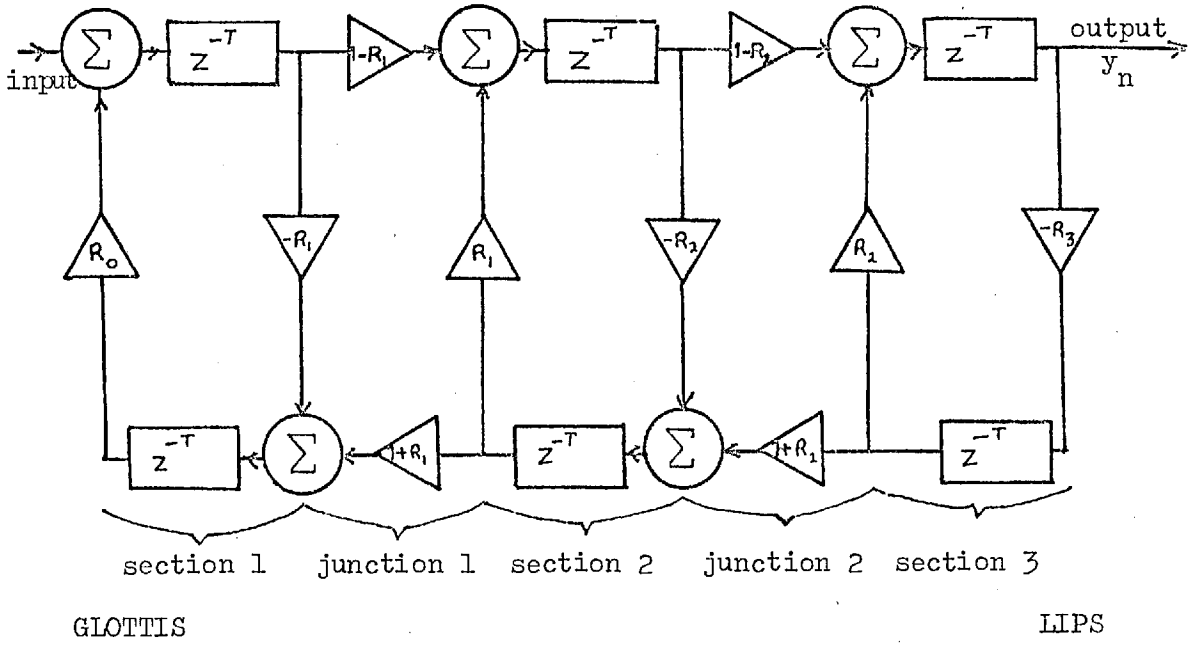


FIGURE 3.4. Three Section Ladder Filter

By studying figure 3.4 (for an impulse input) we can make the following observations:-

a) The first output will occur at time $t = \frac{3T}{2}$

b) All outputs will contain the common factor $(1-R_1)(1-R_2)$.

These two observations make it convenient to consider a time shifted normalised impulse response. The time origin is shifted to the time at which the first output occurs and the magnitude of the first output is assumed to be unity.

The first few terms of the normalised impulse response can be found by summing all the possible paths to the output with the appropriate time delay. These are given by

$$y_0^{(2)} = 1$$

$$y_1^{(2)} = -R_0R_1 - R_1R_2 - R_2R_3 \qquad (3.19)$$

$$y_2^{(2)} = R_0^2R_1^2 + R_1^2R_2^2 + R_2^2R_3^2 + R_0R_1^2R_2 + R_0R_1R_2R_3 + R_1R_2^2R_3$$

$$- R_0R_2(1 - R_1^2) - R_1R_3(1 - R_2^2) \qquad (3.19)$$

where $y_i^{(n)}$ is the output at time iT and the superscript n refers to the number of junctions in the model. Substituting for the boundary conditions at the terminations from equations 3.13 and 3.14 ($R_0 = -1$; $R_3 = 0$) yields

$$y_0^{(2)} = 1$$

$$y_1^{(2)} = R_1 - R_1R_2 \qquad (3.20)$$

$$y_2^{(2)} = R_1^2 + R_1^2R_2^2 - 2R_1^2R_2 + R_2$$

The coefficients of this normalised impulse response will now be equated to the coefficients of a second order purely recursive filter shown in figure 3.5.

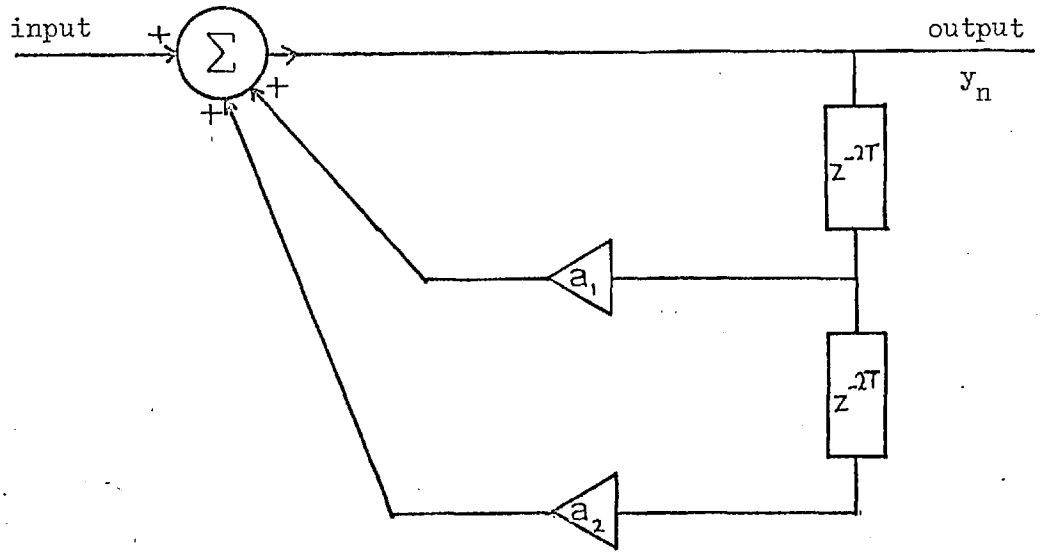


FIGURE 3.5. Recursive Digital Filter

The transfer function of the recursive filter of figure 3.5 is known to be all pole and is given by

$$H_2(z) = \frac{1}{1 - \sum_{k=1}^{2} a_k z^{-k}} \qquad (3.21)$$

The impulse response of the recursive filter of figure 3.5 is given by

$$h^{(2)}(it) = \delta(iT) + \sum_{k=1}^{i} a_k h[(i-k)T] \quad ; \quad a_k = 0 \quad k > 2 \qquad (3.22)$$

Using equation 3.22 we can write down the first three terms of the impulse response of the recursive filter

$$\begin{aligned} h_0^{(2)} &= 1 \\ h_1^{(2)} &= a_1^{(2)} \\ h_2^{(2)} &= (a_1^{(2)})^2 + a_2^{(2)} \end{aligned} \qquad (3.23)$$

Equating formulae 3.20 and 3.23 yields

$$\begin{aligned} a_1^{(2)} &= R_1 - R_1 R_2 \\ a_2^{(2)} &= R_2 \end{aligned} \qquad (3.24)$$

To save having to repeat the calculations for a two section ladder filter model we effectively eliminate the last section by setting $R_2 = 0$ in equations 3.24 to give

$$a_1^{(1)} = R_1 \qquad (3.25)$$

Equations 3.19 were formulated before the terminating conditions were included. This means they also represent the first three terms of the normalised impulse response of a four section ladder filter model. The fourth term of the impulse response is given by

$$\begin{aligned} y_3^{(3)} = &- R_0^3 R_1^3 - R_1^3 R_2^3 - R_2^3 R_3^3 - R_0 R_1^3 R_2^2 - R_0 R_1 R_2^2 R_3^2 - R_1 R_2^3 R_3^2 \\ &- R_0^2 R_1^3 R_2 - R_0^2 R_1^2 R_2 R_3 - R_1^2 R_2^3 R_3 + R_0 R_1^2 R_3 (1 - R_2^2) \\ &+ R_0 R_2^2 R_3 (1 - R_1^2) + 2 R_0 R_1 R_2^2 (1 - R_1^2) + 2 R_0^2 R_1 R_2 (1 - R_1^2) \\ &+ 2 R_1^2 R_2 R_3 (1 - R_2^2) + 2 R_1 R_2 R_3^2 (1 - R_2^2) - R_0 R_3 (1 - R_1^2)(1 - R_2^2) \end{aligned} \qquad (3.26)$$

Equating 3.26 and 3.19 to the impulse response of a third order recursive filter yields

$$\begin{aligned} a_1^{(3)} &= R_1 - R_1R_2 - R_2R_3 \\ a_2^{(3)} &= R_2 - R_3R_1 + R_1R_2R_3 \\ a_3^{(3)} &= R_3 \end{aligned} \qquad (3.27)$$

Equations 3.25, 3.24 and 3.27 relate the coefficients of first, second and third order recursive filters to the reflection coefficients of the vocal tract transmission line model. Comparison of these equations allows us to deduce the following formula relating the coefficients of an (m-1)th order recursive filter to an mth order recursive filter

$$a_i^{(m-1)} = \frac{a_i^{(m)} + R_m a_{m-i}^{(m)}}{1 - R_m^2} \qquad (3.28)$$

We will now prove that this relationship holds for the general case. To do this we will consider the transfer function of the transmission line model of section 3.3.

3.6. DERIVATION OF AN ALGORITHM FOR CALCULATING REFLECTION COEFFICIENTS FOR A GENERAL M SECTION TRANSMISSION LINE MADE UP OF COMMENSURATE SECTIONS

KINARIWALA (1966) developed a method for synthesising a cascaded commensurate transmission line from a required input impedance function. In this section we derive a method for synthesising a cascaded commensurate transmission line from a required volume velocity transfer function. We will use Kinariwala's formulation for the transmission matrix of an elemental transmission line section, which relates the forward and backward travelling volume velocity components on the transmission line. The derivation of the transfer function from this transmission matrix is similar to that given by ATAL (1971, APPENDIX F). Having derived the volume velocity transfer function, we show that it is all pole. Finally, we use induction to derive an algorithm for finding the reflection coefficients of the transmission line model from the all pole transfer function.

Substituting from equation 3.11 into equations 3.8 and 3.9 allows us

to define the following matrix relation

$$\begin{bmatrix} U_k^+(z) \\ U_k^-(z) \end{bmatrix} = \frac{1}{1 - R_k} \begin{bmatrix} z^{+T_k} & -R_k z^{+T_k} \\ -R_k z^{-T_k} & z^{-T_k} \end{bmatrix} \begin{bmatrix} U_{k+1}^+(z) \\ U_{k+1}^-(z) \end{bmatrix} \tag{3.29}$$

Equation 3.29 defines the transmission matrix for the kth section as

$$W_k = \frac{1}{1 - R_k} \begin{bmatrix} z^{+T_k} & -R_k z^{+T_k} \\ -R_k z^{-T_k} & z^{-T_k} \end{bmatrix} \tag{3.30}$$

We can derive the transmission matrix $X^{(m)}$ for a general m junction transmission line (m+1 sections) by using equation 3.29 recursively. In terms of the transmission matrices (W_k) for each elemental section, the transmission matrix for (m+1) cascaded sections is given by

$$X^{(m)} = \prod_{k=1}^{m} W_k = \begin{bmatrix} x_{11}^{(m)}(z) & x_{12}^{(m)}(z) \\ x_{21}^{(m)}(z) & x_{22}^{(m)}(z) \end{bmatrix} \tag{3.31}$$

where

$$\begin{bmatrix} U_1^+(z) \\ U_1^-(z) \end{bmatrix} = \begin{bmatrix} x_{11}^{(m)}(z) & x_{12}^{(m)}(z) \\ x_{21}^{(m)}(z) & x_{22}^{(m)}(z) \end{bmatrix} \begin{bmatrix} U_{m+1}^+(z) \\ U_{m+1}^-(z) \end{bmatrix} \tag{3.32}$$

From equation 3.14 we know that the reflection coefficient corresponding to the lip termination (R_L) is zero. Let us therefore, define in the same manner the termination for our general m section line, (i.e. $R_{m+1} = 0$). This is equivalent to terminating the cascaded transmission line with a matched resistive termination, termination in this manner means $U_{m+1}^-(z) = 0$. Let us further define the termination at the source end of our cascaded transmission line structure according to equation 3.13. With this source termination of the transmission line, inspection of figure 3.3 shows that the input U_G^+ is given by

$$U_G^+(z) = U_1^+(z) - R_o U_1^-(z) = U_1^+(z) + U_1^-(z) \qquad (3.33)$$

Defining the volume velocity transfer function for the m junction transmission line $H_m(z)$, as the ratio of the output volume velocity $U_{m+1}^+(z)$ to the input volume velocity $U_G^+(z)$ gives

$$H_m(z) = \frac{U_{m+1}^+(z)}{U_G^+(z)} = \frac{U_{m+1}^+(z)}{U_1^+(z) + U_1^-(z)} \qquad (3.34)$$

from equation 3.33.

Substituting for $U_1^+(z)$ and $U_1^-(z)$ from equation 3.32 with $U_{m+1}^-(z) = 0$ gives

$$H_m(z) = \frac{U_{m+1}^+(z)}{x_{11}^{(m)}(z)U_{m+1}^+(z) + x_{21}^{(m)}(z)U_{m+1}^+(z)} = \frac{1}{x_{11}^{(m)}(z) + x_{21}^{(m)}(z)} \qquad (3.35)$$

Consider the effect of replacing the matched termination with an extra transmission line section having its own matched termination. From equations 3.32 and 3.29, the input and output volume velocities are now related by

$$\begin{bmatrix} U_1^+(z) \\ U_1^-(z) \end{bmatrix} = \frac{1}{1 - R_{m+1}} \begin{bmatrix} x_{11}^{(m)}(z) & x_{12}^{(m)}(z) \\ x_{21}^{(m)}(z) & x_{22}^{(m)}(z) \end{bmatrix} \begin{bmatrix} z^{+T_{m+1}} & -R_{m+1}z^{+T_{m+1}} \\ -R_{m+1}z^{-T_{m+1}} & z^{-T_{m+1}} \end{bmatrix} \begin{bmatrix} U_{m+2}^+(z) \\ U_{m+2}^-(z) \end{bmatrix} \qquad (3.36)$$

The transmission matrix for the (m+1) junction cascaded transmission line is given by

$$X^{(m+1)}(z) =$$

$$\begin{bmatrix} \dfrac{x_{11}^{(m)}(z)z^{+T_{m+1}} - R_{m+1}x_{12}^{(m)}(z)z^{-T_{m+1}}}{1 - R_{m+1}} & \dfrac{-R_{m+1}x_{11}^{(m)}(z)z^{+T_{m+1}} + x_{12}^{(m)}(z)z^{-T_{m+1}}}{1 - R_{m+1}} \\ \dfrac{x_{21}^{(m)}(z)z^{+T_{m+1}} - R_{m+1}x_{22}^{(m)}(z)z^{-T_{m+1}}}{1 - R_{m+1}} & \dfrac{-R_{m+1}x_{21}^{(m)}(z)z^{+T_{m+1}} + x_{22}^{(m)}(z)z^{-T_{m+1}}}{1 - R_{m+1}} \end{bmatrix} \qquad (3.37)$$

From equation 3.35 the volume velocity transfer function for the (m+1) junction transmission line is given by

$$H_{m+1}(z) = \frac{1}{x_{11}^{(m+1)}(z) + x_{21}^{(m+1)}(z)}$$

which substituting from 3.37 yields

$$H_{m+1}(z) = \frac{1 - R_{m+1}}{(x_{11}^{(m)}(z) + x_{21}^{(m)}(z))z^{+T_{m+1}} - R_{m+1}(x_{22}^{(m)}(z) + x_{12}^{(m)}(z))z^{-T_{m+1}}} \tag{3.38}$$

We will now prove by induction that

$$H_m(1/z) = \frac{1}{x_{22}^{(m)}(z) + x_{12}^{(m)}(z)} \tag{3.39}$$

To prove the above equation we utilise equation 3.37 to evaluate

$$\frac{1}{x_{22}^{(m+1)}(z) + x_{12}^{(m+1)}(z)} = \frac{1 - R_{m+1}}{(x_{22}^{(m)}(z) + x_{12}^{(m)}(z))z^{-T_{m+1}} - R_{m+1}(x_{11}^{(m)}(z) + x_{21}^{(m)}(z))z^{+T_{m+1}}} \tag{3.40}$$

Substituting from equations 3.39 and 3.35; equation 3.40 reduces to

$$\frac{1}{x_{22}^{(m+1)}(z) + x_{12}^{(m+1)}(z)} = \frac{1 - R_{m+1}}{\frac{z^{-T_{m+1}}}{H_m(1/z)} - \frac{R_{m+1}z^{+T_{m+1}}}{H_m(z)}} \tag{3.41}$$

Similarly using equations 3.39 and 3.35, equation 3.38 reduces to

$$H_{m+1}(z) = \frac{1 - R_{m+1}}{\frac{z^{+T_{m+1}}}{H_m(z)} - \frac{R_{m+1}z^{-T_{m+1}}}{H_m(1/z)}} \tag{3.42}$$

Inspection of equations 3.41 and 3.42 shows that if $H_m(1/z)$ is given by

equation 3.39 then $H_{m+1}(1/z)$ is given by

$$H_{m+1}(1/z) = \frac{1}{x_{22}^{(m+1)}(z) + x_{12}^{(m+1)}(z)}$$

which has the same form as equation 3.39. We can therefore see that if equation 3.39 is true for $H_m(1/z)$ it is also true for $H_{m+1}(1/z)$. Finally, inspection of equation 3.39 (with k=1) shows that equation 3.39 holds for the case $H_1(1/z)$. We have therefore shown that equation 3.39 holds for general m and hence equation 3.42 is also true for general m.

We next need to show that the volume velocity transfer function is all pole. Inspection of equation 3.29 (with k=1) allows definition of $H_1(z)$ from equation 3.35 as

$$H_1(z) = \frac{1}{x_{11}^{(1)}(z) + x_{21}^{(1)}(z)} = \frac{1 - R_1}{z^{+T_1} - R_1 z^{-T_1}} \qquad (3.43)$$

We can see from equation 3.43 that $H_1(z)$ is a constant divided by a polynomial in z; hence $H_1(z)$ is all pole. Using induction, from equation 3.42 we can see that the volume velocity transfer function of a general m+1 section transmission line (with an open circuit source termination and a matched load termination) is all pole.

From equation 3.42 we can also deduce that the numerator of the volume velocity transfer function contains the constant factor

$$\prod_{k=1}^{m+1} (1 - R_k)$$

This is equivalent to observation (b) in section 3.5. We also know that the transmission line will have a delay ($z^{-T_{tot}}$) before any output occurs; (from observation(a) of section 3.5).

Using these two facts we deduce the general form for the all pole volume velocity transfer function as

$$H_m(z) = \frac{z^{-\left\{\sum_{k=1}^{m} T_k\right\}} \left(\prod_{k=1}^{m} (1 - R_k)\right)}{1 - \sum_{i=1}^{m} a_i^{(m)} z^{-2v_i}} \qquad (3.44)$$

where v_i is a positive integer multiple of the propagation delay (T_{min}) of the shortest tube in our commensurate model; v_i increases with i; and $v_m = \sum_{k=1}^{m} T_k$.

We will now prove by induction that equation 3.44 is the general form for the volume velocity transfer function of a transmission line consisting of commensurate sections, with an open circuit source termination and a matched load termination.

Substituting from equation 3.44 into equqtion 3.42 gives

$$H_{m+1}(z) = \frac{1 - R_{m+1}}{\left(1 - \sum_{i=1}^{m} a_i^{(m)} z^{-2v_i}\right) z^{+T_{m+1}} - R_{m+1} z^{-T_{m+1}} \left(1 - \sum_{i=1}^{m} a_i^{(m)} z^{+2v_i}\right)}$$

$$z^{-\left\{\sum_{k=1}^{m} T_k\right\}} \left(\prod_{k=1}^{m} (1 - R_k)\right) \qquad z^{+\left\{\sum_{k=1}^{m} T_k\right\}} \left(\prod_{k=1}^{m} (1 - R_k)\right)$$

$$= \frac{z^{-\left\{\sum_{k=1}^{m+1} T_k\right\}} \left(\prod_{k=1}^{m+1} (1 - R_k)\right)}{1 - \sum_{i=1}^{m} a_i^{(m)} z^{-2v_i} - R_{m+1} z^{-\left\{2 \sum_{k=1}^{m+1} T_k\right\}} \left(1 - \sum_{i=1}^{m} a_i^{(m)} z^{+2v_i}\right)} \qquad (3.45)$$

Equation 3.45 is of the same general form as equation 3.44. Also from equation 3.43 we can see that $H_1(z)$ has this same general form. Hence, we have proved equation 3.44 by induction. From equation 3.45 we can also see that $a_{m+1}^{(m+1)}$ as defined by equation 3.44 is given by

$$a_{m+1}^{(m+1)} = R_{m+1} \qquad (3.46)$$

Therefore, given the transfer function of a cascaded transmission line, we can immediately identify the reflection coefficient of the junction nearest the termination, from equation 3.46. We will now derive an algorithm for reducing the transfer function of an (m+1) junction line to the transfer function of an m junction transmission line. The load

termination in both cases being matched to the last tube of the transmission line.

Equation 3.41 is rearranged to give

$$\frac{1 - R_{m+1}}{H_{m+1}(1/z)} = \frac{z^{-T_{m+1}}}{H_m(1/z)} - \frac{R_{m+1} z^{+T_{m+1}}}{H_m(z)} \qquad (3.47)$$

Similarly equation 3.42 is rearranged to give

$$\frac{1 - R_{m+1}}{H_{m+1}(z)} = \frac{z^{+T_{m+1}}}{H_m(z)} - \frac{R_{m+1} z^{-T_{m+1}}}{H_m(1/z)} \qquad (3.48)$$

Multiplying equation 3.47 by R_{m+1} and adding the result to 3.48 yields

$$\frac{R_{m+1}(1 - R_{m+1})}{H_{m+1}(1/z)} + \frac{1 - R_{m+1}}{H_{m+1}(z)} = \frac{z^{+T_{m+1}}}{H_m(z)} - \frac{R_{m+1}^2 z^{+T_{m+1}}}{H_m(z)}$$

which can be rearranged to give

$$H_m(z) = \frac{z^{+T_{m+1}}(1 + R_{m+1})}{\frac{R_{m+1}}{H_{m+1}(1/z)} + \frac{1}{H_{m+1}(z)}} \qquad (3.49)$$

Equations 3.46 and 3.49 enable us to define the algorithm for synthesising a transmission line given a required transfer function. The method of using this algorithm will now be outlined:-

i) Given a required all pole transfer function express it in the form

$$H_n(z) = \frac{K z^{-v_n}}{1 - \sum_{i=1}^{n} a_i^{(n)} z^{-2v_i}} \qquad (3.50)$$

where v_i increases with i; and K is chosen to make the constant term in the denominator unity.

ii) Identify the coefficient of the greatest negative power of z in the denominator as R_n. i.e.,

$$R_n = a_n \qquad (3.51)$$

iii) Derive the transfer function for (n-1) junction transmission line from that of the n junction transmission line, (formed by replacing the last section with a matched termination) as

$$H_{n-1}(z) = \frac{z^{+T_n}(1 + R_n)}{\frac{R_n}{H_n(1/z)} + \frac{1}{H_n(z)}} \qquad (3.52)$$

where z^{+T_n} is chosen such that $H_{n-1}(z)$ reduces to

$$H_{n-1}(z) = \frac{(K/(1 - R_n))z^{-v_{n-1}}}{1 - \sum_{i=1}^{n-1} a_i^{(n-1)} z^{-2v_i}} \qquad (3.53)$$

iv) Repeat steps (ii) and (iii) yielding R_{n-1}, R_{n-2},...,R_1, which are the reflection coefficients for each junction in the transmission line.

This algorithm is difficult to implement efficiently on a digital computer because of the necessity to keep track of the powers of z. If we constrain all sections of the transmission line to be of equal length we can overcome this difficulty.

3.7. ALGORITHM FOR FINDING THE REFLECTION COEFFICIENTS FOR A TRANSMISSION LINE CONSISTING OF N EQUAL LENGTH SECTIONS

In the previous section we derived a method for synthesising a general commensurate transmission line from a required transfer function. This algorithm can be considerably simplified by constraining all the component sections to be of equal length.

Equation 3.44 is simplified in this manner (by substituting $T_k = T$ for all k) to yield the transfer function

$$H_n(z) = \frac{z^{-nT} \prod_{k=1}^{n} (1 - R_k)}{1 - \sum_{k=1}^{n} a_k^{(n)} z^{-2kT}} \qquad (3.54)$$

Using equation 3.54 we can simplify 3.52 as follows

$$H_{n-1}(z) = \frac{z^{+T}(1 + R_n)}{\frac{R_n}{H_n(1/z)} + \frac{1}{H_n(z)}}$$

$$= \frac{z^{+T}(1 + R_n)z^{-nT}\prod_{k=1}^{n}(1 - R_k)}{1 - \sum_{k=1}^{n} a_k^{(n)} z^{-2kT} + R_n z^{-2nT}(1 - \sum_{k=1}^{n} a_k^{(n)} z^{+2kT})}$$

$$= \frac{z^{-(n-1)T}\prod_{k=1}^{n-1}(1 - R_k)}{1 - \frac{\sum_{k=1}^{n-1}(a_k^{(n)} + R_n a_{n-k}^{(n)})z^{-2kT}}{1 - R_n^2}} \qquad (3.55)$$

because $a_n^{(n)} = R_n$

Comparison of equations 3.54 and 3.55 yields the recursive relationship

$$a_i^{(n-1)} = \frac{a_i^{(n)} + R_n a_{n-i}^{(n)}}{1 - R_n^2} \qquad (3.56)$$

This is precisely the relationship we deduced in section 3.5 from consideration of the impulse responses of the first, second and third order ladder networks. The algorithm for calculating the reflection coefficients of a transmission line (consisting of n equal length sections) from its transfer function will now be summarised.

i) Reduce the required transfer function to the form

$$H_n(z) = \frac{Kz^{-nT}}{1 - \sum_{k=1}^{n} a_k^{(n)} z^{-kT}} \qquad (3.57)$$

(This is equivalent to a recursive filter followed by a gain factor K and a delay z^{-nT}).

ii) Identify the recursive filter coefficients $a_k^{(n)}$ and in particular set $R_n = a_n^{(n)}$.

iii) Derive the coefficients $a_i^{(n-1)}$ for an (n-1) section recursive filter

using equation 3.56.

iv) Repeat steps (ii) and (iii) until all the reflection coefficients are known (i.e., $R_1 \rightarrow R_n$).

From the discussion in section 3.4 (in particular equation 3.16) we know that for a sampling rate of 10000 samples per second

$$T_{min} = \frac{1}{20,000} \quad \text{or} \quad \ell_{min} = \frac{c}{20,000} = 1.7 \text{ cm}$$

A typical human vocal tract would be 17 cm long. This means if we choose all tubes to be 1.7 cm long we only have a ten tube model. In general because of the time consuming computation required in keeping track of the powers of z, the method of section 3.6 will prove less efficient when only ten sections are required then the method presented in this section.

Finally in order to derive the vocal tract area function we must calculate the areas of the component tubes from the reflection coefficients of the transmission line model.

This is simply accomplished by rearranging equation 3.11 to give

$$A_k = \frac{1 + R_k}{1 - R_k} A_{k+1} \tag{3.58}$$

Starting with an assumed area for the lips, equation 3.58 can be used recursively to find A_n, A_{n-1},....,A_1, which constitute the vocal tract area function.

In conclusion of this chapter the important results will be repeated. We were able to show that approximating the vocal tract by an acoustic transmission line was equivalent to approximating it by the purely recursive filter of figure 3.5. We also proved that the transmission line approximation had an all pole transfer function. Finally in this section, we presented an algorithm for finding the reflection coefficients (and hence the area function) from the coefficients of a recursive filter having the same transfer function as the vocal tract; (ignoring the gain constant K and delay z^{-nT} in equation 3.57). In the next chapter we will show how the coefficients of this recursive filter can be calculated from samples of the speech waveform.

CHAPTER FOUR

DIGITAL INVERSE FILTER FOR SPEECH

A method of calculating the vocal tract area function for voiced non nasalised speech sounds was derived in chapter three. The starting point for this method was a recursive digital filter having the same transfer function as the vocal tract. It was shown how the reflection coefficients of a transmission line model could be calculated from these recursive filter coefficients. Finally, a method was described for calculating the cross sectional areas of the component tubes in the transmission line model from these reflection coefficients. These cross sectional areas constitute a piecewise approximation to the vocal tract area function.

The next step is to derive a method of calculating the recursive filter coefficients from the speech waveform. When this is done, we will have a method of computing the vocal tract area function from the speech waveform. A digital inverse filter will be defined, having a transfer function equal to the inverse of the recursive filter transfer function. A method will then be derived, for calculating the coefficients of this inverse filter from the speech waveform. Results obtained from synthetic speech are presented in the last two sections of this chapter.

4.1. FORMULATION OF THE DIGITAL INVERSE FILTER

Inverse filtering of speech waveforms is not a new idea. HOLMES (1962) used an inverse filter (where the zeros of the filter were adjusted by hand to cancel the poles of the vocal tract transfer function) to investigate the shape of the glottal pulse. An alternative technique avoiding the necessity of correcting for lip radiation effects, was proposed by ROTHENBERG (1973). In his method the volume velocity at the lips (measured with the aid of a modified pneumotachograph mask) was inverse filtered, instead of the radiated acoustic pressure used by Holmes. MARKEL (1971) calculated his "optimum inverse filter" coefficients by constraining the output of the filter to be the best estimate of white noise, for several pitch periods. He then used the optimum inverse filter coefficients to estimate the vocal tract transfer function (see section 2.3).

In this chapter a technique similar in concept to Markel's is used.

However, we calculate the inverse filter coefficients from samples of the closed glottis region of a single pitch period. It is my belief that calculation of the inverse filter coefficients in this way, avoids the necessity of including the effects of the glottal source in the inverse filter. For this reason the transfer function of the inverse filter is chosen to be the inverse of the vocal tract transfer function.

The transfer function of the vocal tract transmission line was shown to be (equation 3.54)

$$H_n(z) = \frac{z^{-nT} \prod_{k=1}^{n} (1 - R_k)}{1 - \sum_{k=1}^{n} a_k^{(n)} z^{-2kT}} \qquad (4.1)$$

In equation 4.1 the term z^{-nT} in the numerator represents only the delay between the input and the output. The other term in the numerator $(\prod_{k=1}^{n} 1 - R_k)$ is a constant representing the gain (or attenuation) between the input and the output, ignoring the frequency dependance introduced by reflections. We do not require either of these terms to calculate the vocal tract area function. For this reason the inverse filter transfer function is defined as

$$C_n(z) = 1 - \sum_{k=1}^{n} a_k^{(n)} z^{-2kT} \qquad (4.2)$$

A digital transversal filter having this transfer function is shown in figure 4.1.

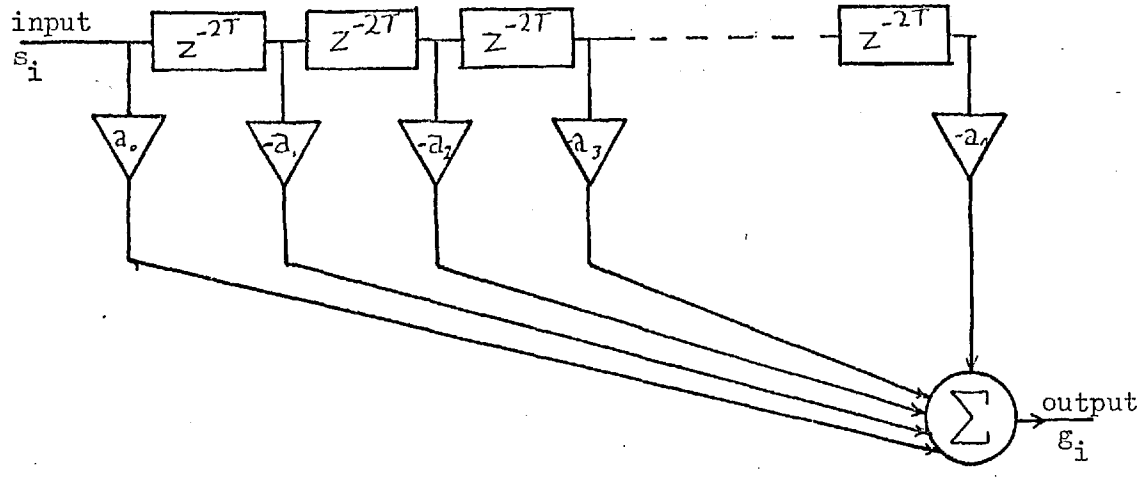


FIGURE 4.1 Digital Inverse Vocal Tract Filter

One fact worth noting at this point is that the filter of figure 4.1 is equivalent to the recursive filter of figure 3.5 with input and output terminals interchanged. This means the concept of the inverse filter is enhanced, because inverse filtering can be thought of as passing the signal generated by the recursive filter backwards through the original filter to obtain the excitation function.

The delay factor T is fixed by the sampling frequency and related to the length of the sections in the transmission line model (Equation 3.17). Finally, it should be noted that calculation of the inverse filter coefficients (a_k) is equivalent to calculation of the coefficients of the recursive filter.

4.2. SOME PROPERTIES OF GLOTTAL EXCITATION

Excitation of the vocal tract by air passing through the vibrating vocal folds was discussed in section 1.3. The properties of this glottal excitation required (in order to apply the inverse filter) will now be summarised.

a) MATTHEWS et al. (1961) were able to show that the spectrum of the glottal excitation function contains only zeros. Their argument was as follows. The Laplace transform of the glottal flow (g(t)) over a single pitch period T is given by

$$G(s) = \int_0^T g(t)e^{-st}dt \qquad (4.3)$$

Now g(t) is an all positive finite function and the limits on the integration are finite. This means the result of the integration must be finite and hence the glottal spectrum must be all zero.

b) During the period of glottal opening these zeros perturb the poles of the vocal tract transfer function. Indeed FLANAGAN (1972), reports that the poles appear to move about in the complex frequency plane. Furthermore, analysis of the vocal tract with a time varying source termination would be difficult. Finally, using an extension of the argument in point (a), we can show that analysis over the closed glottis period introduces no zeros into the speech spectrum.

$$G(s) = \int_0^T g(t)e^{-st}dt = \int_0^{T_c} g(t)e^{-st}dt + \int_{T_c}^T g(t)e^{-st}dt$$

where T_c is the time of glottal closure and $g(t) = 0 (T_c < t < T)$

therefore

$$G(s) = \int_0^{T_c} g(t)e^{-st}dt \qquad (4.4)$$

From equations 4.4 we can see that the zeros of the glottal spectrum are indeed the zeros of the open glottis period, hence the spectrum of the closed glottis period can contain no zeros.

c) If sufficient vocal effort is used a pitch period of speech will always contain a region of zero excitation, caused by closure of the glottis. During weakly voiced speech the vocal folds may not completely close, because the pressure reduction in the glottis which causes the folds to shut is not great enough. Also, during the production of whispered vowel like sounds the folds do not close at all.

d) Rapid closure of the vocal folds usually gives rise to the greatest excitation of the vocal tract, because these instants usually correspond to the sharpest discontinuity of volume velocity.

These four points suggest that a more realistic estimate of the vocal tract transfer function can be made by only analysing closed glottis regions of the pitch period. Work by HOLMES and THORNBER (1973) on estimation of formant frequencies by waveform matching during closed glottis periods also supports this hypothesis. We will therefore calculate the inverse filter coefficients only from samples of the closed glottis period.

4.3. DETERMINATION OF THE INVERSE FILTER COEFFICIENTS

If the vocal tract could be perfectly approximated by an acoustic transmission line consisting of n sections, then the following equations could be solved to determine the inverse filter coefficients

$$\begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_n \end{bmatrix} = \begin{bmatrix} s_0 & s_{-1} & \cdots & s_{-n} \\ s_1 & s_0 & & \vdots \\ \vdots & & \ddots & \\ s_n & s_{n-1} & \cdots & s_0 \end{bmatrix} \begin{bmatrix} +a_0 \\ -a_1 \\ -a_2 \\ \vdots \\ -a_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (4.5)$$

where

g_i are outputs of the inverse filter

s_i are samples of the speech waveform taken during the closed

glottis period.

As the transmission line model can only be an approximation to the vocal tract (for n ~ 10) the n inverse filter coefficients can be optimised by solving the following equations, which express the filtering action of the inverse filter.

$$\begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix} = \begin{bmatrix} s_0 & s_{-1} & s_{-2} & \cdots & s_{-n} \\ s_1 & s_0 & s_{-1} & & \vdots \\ s_2 & s_1 & s_0 & \ddots & \vdots \\ \vdots & \vdots & & \ddots & s_0 \\ s_m & s_{m-1} & \cdots & & s_{m-n} \end{bmatrix} \begin{bmatrix} +a_0 \\ -a_1 \\ -a_2 \\ \vdots \\ -a_n \end{bmatrix} \qquad (4.6)$$

We are analysing only closed glottis periods of speech. Ideally therefore, we require all the outputsof the inverse filter (g_i; $i = 0 \rightarrow m$) to be zero. In practice however (because of the assumptions introduced in section 3.2 which only approximately hold true in practice), we must define some other criterion for the outputs of the inverse filter. The criterion chosen here is minimisation of the energy of the inverse filter outputs i.e.

$$\sum_{i=0}^{m} g_i^2 \quad \text{is minimised or} \quad \| \underline{g} \|^2_{min} \qquad (4.7)$$

where $\| \underline{g} \|^2$ is the Euclidian norm of the vector $\underline{g}$.

$$= \underline{g}^T \underline{g}$$

One possible solution of equation 4.6 under the conditions of relation 4.7 is immediately obvious, namely the trivial one

$$a_i = 0; \quad \text{for} \quad i = 0, 1, 2, \ldots, n \qquad (4.8)$$

This solution is not possible because we formulated the inverse filter with $a_0 = 1$ (see equation 4.2). Using this constraint and substituting from equation 4.6, equation 4.7 can be reformulated as

$$\| \underline{g} \|^2_{min} = \left\| \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ \vdots \\ s_m \end{bmatrix} - \begin{bmatrix} s_{-1} & s_{-2} & \cdots & s_{-n} \\ s_0 & s_{-1} & & s_{1-n} \\ \vdots & \ddots & s_0 & \vdots \\ \vdots & & \ddots & s_0 \\ s_{m-1} & s_{m-2} & \cdots & s_{m-n} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \right\|^2_{min} \qquad (4.9)$$

or alternatively

$$\| \underline{g} \|^2_{min} = \| \underline{c} - S\underline{a} \|^2_{min} = (\underline{c} - S\underline{a})^T(\underline{c} - S\underline{a}) \qquad (4.10)$$

where

$\underline{c}$ is the column vector of m+1 consecutive speech samples in equation 4.9

S is the (m+1) x n matrix of speech samples

$\underline{a}$ is the column vector of filter coefficients

and superscript T denotes matrix transposition.

Equation 4.10 can be solved in a computationally efficient manner by Householder transformation.

4.4. HOUSEHOLDER TRANSFORMATION

Use of Householder transformation to minimise the Euclidian norm of relation 4.10 was first reported by GOLUB (1965). The work presented in this section is based on this paper by Golub and also on another by WILKINSON (1960). This section is included in this thesis, because, perhaps only a few people in the speech discipline will be aware of what Householder transformation involves. However, this section is not essential to the understanding of the remainder of the thesis and may be missed by anyone not interested in the details of the computational methods used.

Householder transformation uses one well known premise of matrix theory, namely premultiplying a matrix by another orthogonal matrix does not alter the Euclidian norm of the original matrix. This can easily be demonstrated for our case. Let Q be an orthonormal square matrix of dimensions

(m+1) x (m+1) then $\| Q(\underline{c} - S\underline{a}) \|^2$ is given by

$$(Q(\underline{c} - S\underline{a}))^T Q(\underline{c} - S\underline{a}) = (\underline{c} - S\underline{a})^T Q^T Q(\underline{c} - S\underline{a})$$

$$= (\underline{c} - S\underline{a})^T (\underline{c} - S\underline{a})$$

because Q is orthonormal and hence $Q^T Q = I$, where I is the identity matrix

$$\therefore \quad \| Q(\underline{c} - S\underline{a}) \|^2 = \| \underline{c} - S\underline{a} \|^2$$

In Householder transformation an orthonormal matrix Q is chosen which has the additional property of sweeping the matrix S into upper triangular form as shown below

$$QS = \begin{bmatrix} B \\ 0 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{21} & b_{31} & \cdots & b_{n1} \\ 0 & b_{22} & b_{32} & \cdots & b_{n2} \\ 0 & 0 & b_{33} & \cdots & b_{n3} \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots\ 0 & \cdots & b_{nn} \\ 0 & 0 & \cdots & & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & & 0 \end{bmatrix} \qquad (4.11)$$

where

[B] is an (nxn) upper triangular matrix i.e., all elements below the leading diagonal are zero

and

[O] is an (m+1-n)xn null matrix.

Let us first assume that we have such a matrix Q and show how the problem may be simplified and solved by this method. Further let

$$Q(\underline{c} - S\underline{a}) = \begin{bmatrix} \underline{d} \\ \underline{d}' \end{bmatrix} - \begin{bmatrix} B \\ 0 \end{bmatrix} \underline{a} \qquad (4.12)$$

where $\begin{bmatrix} \underline{d} \\ \underline{d}' \end{bmatrix}$ is the (m+1) x 1 column vector formed by premultiplying $\underline{c}$ by Q ($\underline{d}$ represents the first n elements of this vector and $\underline{d}'$ the remaining (m+1-n) elements). The Euclidian norm is then given by

$$\| (\underline{c} - S\underline{a}) \|^2 = \left[\begin{bmatrix} \underline{d} \\ \underline{d}' \end{bmatrix} - \begin{bmatrix} B \\ 0 \end{bmatrix} \underline{a} \right]^T \left[\begin{bmatrix} \underline{d} \\ \underline{d}' \end{bmatrix} - \begin{bmatrix} B \\ 0 \end{bmatrix} \underline{a} \right]$$

$$= [\, \underline{d} - B\underline{a} \,]^T [\underline{d} - B\underline{a} \,] + \underline{d}'^T \underline{d}'$$

$$= \| \underline{d} - B\underline{a} \|^2 + \| \underline{d}' \|^2 \qquad (4.13)$$

Equation 4.13 is minimised when

$$\underline{d} = B\underline{a} \qquad (4.14)$$

In equation 4.13 the Euclidian norm of $\underline{d}'$ is a measure of the mean square error of the estimation of the inverse filter coefficients (a_k). We know the matrix B is upper triangular, hence we can simply solve equation 4.14 to find the coefficients a_k by a process called back substitution.

Using equation 4.11, equation 4.14 can be rewritten as

$$\begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_n \end{bmatrix} = \begin{bmatrix} b_{11} & b_{21} & b_{31} & \cdots\cdots & b_{n1} \\ 0 & b_{22} & b_{32} & \cdots\cdots & b_{n2} \\ 0 & 0 & b_{33} & \cdots\cdots & b_{n3} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 \cdots\cdots 0 & & b_{nn} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{bmatrix} \qquad (4.15)$$

Solution of equation 4.15 by back substitution is defined as the following recursive process:-

First find a_n by

$$a_n = \frac{d_n}{b_{nn}}$$

then having found a_n find a_{n-1} by

$$a_{n-1} = \frac{d_{n-1} - b_{nn-1}a_n}{b_{n-1\ n-1}}$$

then find a_{n-2} by a similar method etc. The general formula for finding a_{n-j} given the a_i's $(n-j < i < n)$ is

$$a_{n-j} = \frac{d_{n-j} - \sum_{i=0}^{j-1} b_{n-in-j}a_{n-i}}{b_{n-j\ n-j}} \qquad (4.16)$$

The method for calculating the inverse filter coefficients from samples of the speech waveform has been outlined in principle. We must now investigate the specification of the Householder transform matrix Q. We will use the form of the Householder matrix proposed by GOLUB (1965), because this formulation minimises the number of computational steps necessary.

Golub proposed the following form for the matrix Q. Let

$$Q = P^{(n)}P^{(n-1)}P^{(n-2)}\ldots P^{(1)} \qquad (4.17)$$

where the (m+1) x (m+1) matrix $P^{(k)}$ reduces all the elements below the leading diagonal element, in the kth. column of the matrix S to zero, without effecting the first (k-1) columns of S. The transformation is then formed by premultiplying the speech matrix S first by $P^{(1)}$ then by $P^{(2)}$ etc., where the kth. transformation is represented by

$$P^{(k)}S^{(k)} = S^{(k+1)} \qquad (4.18)$$

where $S^{(k)}$ is the matrix resulting from premultiplying S by $P^{(k-1)}P^{(k-2)}\ldots P^{(1)}$.

Golub proposed the following form for the matrix $P^{(k)}$

$$P^{(k)} = I - F_k \underline{U}^{(k)}\underline{U}^{(k)T} \qquad (4.19)$$

where superscript T denotes matrix transpose and F_k is a constant of the form

$$F_k = \mathrm{sgn}(s_{kk}^{(k)})\left[\sqrt{\sum_{i=k}^{m+1}(s_{ik}^{(k)})^2}\ .\ \left(\mathrm{sgn}(s_{kk}^{(k)})\sqrt{\sum_{i=k}^{m+1}(s_{ik}^{(k)})^2} + s_{kk}^{(k)}\right)\right]^{-1} \qquad (4.20)$$

and the vector $\underline{U}^{(k)}$ is made up of the elements of relations 4.21.

$$
\begin{aligned}
u_i^{(k)} &= 0 \quad ; \qquad \text{for } i < k \\
u_k^{(k)} &= \operatorname{sgn}(s_{kk}^{(k)}) \sqrt{\sum_{i=k}^{m+1} (s_{ik}^{(k)})^2} + s_{kk}^{(k)} \qquad (4.21) \\
u_i^{(k)} &= s_{ik}^{(k)} \quad ; \quad \text{for } i > k.
\end{aligned}
$$

where $s_{ik}^{(k)}$ is the kth. element in the ith. row of $S^{(k)}$

We will now prove that the matrix $P^{(k)}$ is orthonormal. First we will show that the matrix $P^{(k)}$ is symmetric.

Using equation 4.19

$$
\begin{aligned}
P^{(k)T} &= (I - F_k \underline{U}^{(k)} \underline{U}^{(k)T})^T \\
&= I^T - F_k (\underline{U}^{(k)} \underline{U}^{(k)T})^T
\end{aligned}
$$

or

$$
P^{(k)T} = I - F_k \underline{U}^{(k)} \underline{U}^{(k)T} = P^{(k)} \qquad (4.22)
$$

which proves the matrix $P^{(k)}$ is symmetric and hence

$$
\begin{aligned}
P^{(k)T} P^{(k)} = (P^{(k)})^2 &= (I - F_k \underline{U}^{(k)} \underline{U}^{(k)T})(I - F_k \underline{U}^{(k)} \underline{U}^{(k)T}) \\
&= I + F_k^2 \underline{U}^{(k)} (\underline{U}^{(k)T} \underline{U}^{(k)}) \underline{U}^{(k)T} - 2F_k \underline{U}^{(k)} \underline{U}^{(k)T} \\
&= I - (2F_k - F_k^2 \underline{U}^{(k)T} \underline{U}^{(k)}) \underline{U}^{(k)} \underline{U}^{(k)T}
\end{aligned}
\qquad (4.23)
$$

because the quantity $\underline{U}^{(k)T} \underline{U}^{(k)}$ is a scalar and can be commuted. For the matrix to be orthonormal we require $P^{(k)} P^{(k)T} = I$ which we can see is satisfied by equation 4.23 if equation 4.24 is true

$$
\underline{U}^{(k)T} \underline{U}^{(k)} = 2(F_k)^{-1} \qquad (4.24)
$$

We will now evaluate the left hand side of 4.24 using the values of the elements given in 4.21

$$\underline{U}^{(k)T}\underline{U}^{(k)} = \sum_{i=0}^{m+1} (u_i^{(k)})^2$$

$$= \sum_{i=k}^{m+1} (u_i^{(k)})^2 \qquad \text{because } u_i = 0 \text{ for } i < k$$

$$= \sum_{i=k+1}^{m+1} (s_{ik}^{(k)})^2 + \left\{ \text{sgn}(s_{kk}^{(k)}) \sqrt{\sum_{i=k}^{m+1} (s_{ik}^{(k)})^2} + s_{kk}^{(k)} \right\}^2$$

$$= \sum_{i=k+1}^{m+1} (s_{ik}^{(k)})^2 + \sum_{i=k}^{m+1} (s_{ik}^{(k)})^2 + (s_{kk}^{(k)})^2 + 2s_{kk}^{(k)} \text{sgn}(s_{kk}^{(k)}) \sqrt{\sum_{i=k}^{m+1} (s_{ik}^{(k)})^2}$$

$$= 2 \sum_{i=k}^{m+1} (s_{ik}^{(k)})^2 + 2\text{sgn}(s_{kk}^{(k)}) \sqrt{\sum_{i=k}^{m+1} (s_{ik}^{(k)})^2}\ s_{kk}^{(k)}$$

$$= 2\ \text{sgn}(s_{kk}^{(k)}) \sqrt{\sum_{i=k}^{m+1} (s_{ik}^{(k)})^2} \left\{ \text{sgn}(s_{kk}^{(k)}) \sqrt{\sum_{i=k}^{m+1} (s_{ik}^{(k)})^2} + s_{kk}^{(k)} \right\}$$

$$= 2(F_k)^{-1} \qquad \text{from equation 4.20}$$

We have therefore proved that the general matrix $P^{(k)}$ is symmetric and orthonormal. At first a lot of computation may seem necessary in implementing Householder transformation. This is not the case because neither the matrix Q nor the matrices $P^{(k)}$ need be calculated explicitly.

Substituting from equation 4.19 into equation 4.18 gives

$$S^{(k+1)} = S^{(k)} - F_k \underline{U}^{(k)} \underline{U}^{(k)T} S^{(k)}$$

$$= S^{(k)} - \underline{U}^{(k)} \underline{Y}^{(k)T} \qquad (4.25)$$

where the elements of the column vector $\underline{Y}^{(k)}$ are expressed explicitly by

$$y_j^{(k)} = 0 \ ; \qquad \text{for } j < k$$

$$y_k^{(k)} = 1$$

$$y_j^{(k)} = F_k \sum_{i=k}^{m} u_i^{(k)} . s_{ij}^{(k)} \qquad \text{for } j > k \qquad (4.26)$$

from equation 4.21.

Using equations 4.26, 4.25 and 4.21 we can express the relationship between the elements $s_{ij}^{(k+1)}$ of the matrix $S^{(k+1)}$ and the elements $s_{ij}^{(k)}$ of the matrix $S^{(k)}$ as

$$s_{kk}^{(k+1)} = -\operatorname{sgn}(s_{kk}^{(k)}) \sum_{i=k}^{m} (s_{ik}^{(k)})^2$$

$$s_{ik}^{(k+1)} = 0 \; ; \quad i = k+1, k+2, \ldots, m \qquad (4.27)$$

$$s_{ij}^{(k+1)} = s_{ij}^{(k)} - u_i^{(k)} y_j^{(k)} \; ;$$

$$j = k+1, \ldots, n; \; i = k, k+1, \ldots, n \text{ for fixed } j.$$

All other elements of the matrix $S^{(k)}$ remain unchanged. Similarly the transformation of the vector $\underline{c}$ can be expressed specifically as

$$\underline{c}^{(1)} = \underline{c}$$

$$\underline{c}^{(k+1)} = P^{(k)} \underline{c}^{(k)} \qquad (4.28)$$

and

$$c_i^{(k+1)} = c_i^{(k)} - u_i^{(k)} F_k u_i^{(k)} c_i^{(k)}; \; i = 1, \ldots, m$$

where $u_i^{(k)}$ is the ith. element of the vector $\underline{U}^{(k)}$.

Finally from equation 4.27 we can see that the transformation performed by $P^{(k)}$ reduces the elements in the kth. row of $S^{(k+1)}$ below the leading diagonal element to zero as required. Hence premultiplying S by Q reduces S to the upper triangular matrix B.

The main points of this section will now be summarised. The elemental Householder transformation matrices $P^{(k)}$ were shown to be symmetric and orthonormal. Premultiplying the matrix of speech samples S by the Householder matrices $P^{(n)}, P^{(n-1)}, \ldots, P^{(1)}$ was shown to reduce S to the upper triangular matrix B. From the previous section we know how to calculate the inverse filter coefficients from the upper triangular matrix B and the transformed vector $\underline{d}$. Therefore, we now have the complete method for calculating the inverse filter coefficients from samples of the speech waveform by Householder transformation. Finally, from equation 4.9 we can see that (m+n+1) samples of the speech waveform are required for

this analysis. The (m+1) samples of the speech waveform which constitute the vector $\underline{c}$ we will call the analysis interval. This analysis interval is the period of the speech waveform when the output of the inverse filter is constrained to have minimum mean square energy. The n samples previous to this inverval are also required because the inverse filter has "memory"; equivalent to the reflections at the junctions in the transmission line model.

A Fortran program for performing Householder transformation, back substitution and area function mapping is given in Appendix A4.1. This program is used in the next section to test my method using synthetic speech.

4.5. APPLICATION OF THE DIGITAL INVERSE FILTER TO SYNTHETIC SPEECH

In the preceeding sections of this chapter it was explained how the coefficients of an n section inverse filter could be derived from (m+n+1) consecutive samples of the speech waveform. It was also proposed, that if these samples were taken during the closed glottis period of phonation the glottal excitation function could be separated from the vocal tract transfer function. Coefficients of the inverse filter calculated by this method are then identical to the coefficients of a recursive filter model of the vocal tract. Chapter three of this thesis was devoted to a method of using these recursive filter coefficients to derive the vocal tract area function. So we now have a method of finding the vocal tract area function from the speech waveform.

In order to test the method in a controlled manner synthetic speech is used, ensuring that the correct number of filter coefficients are chosen and that the closed glottis region is accurately defined. The speech is synthesised in the time domain, by exciting an all pole filter with a known excitation function. The filter is chosen to consist solely of complex conjugate pole pairs and the values for these pole pairs are calculated from the formant frequencies and bandwidths as given by FANT (1970), tabulated here as figure 4.2.

Starting from the formant frequencies and bandwidths the matched z transform is employed, (GOLDEN 1968, HOLMES 1970) to calculate the transfer function in the z domain. Digital convolution with the excitation function is then used to determine the volume velocity waveform at the lips. In order to calculate the speech pressure waveform the lip radiation is taken to be a simple spherical source (FANT 1970). This is implemented according to equation 2.8 and requires only differencing of the volume velocity samples at the lips to produce the radiated pressure, if K_2 is taken to be unity.

VOWEL	FIRST FORMANT		SECOND FORMANT		THIRD FORMANT		FOURTH FORMANT	
	FREQ.	B-WIDTH	FREQ.	B-WIDTH	FREQ.	B-WIDTH	FREQ.	B-WIDTH
/a/	616	57	1072	72	2430	130	3410	175
/e/	432	39	1959	95	2722	170	3500	325
/i/	222	60	2244	75	3140	240	3700	230
/o/	510	54	900	65	2400	100	3220	135
/u/	231	69	615	50	2375	110	3320	115

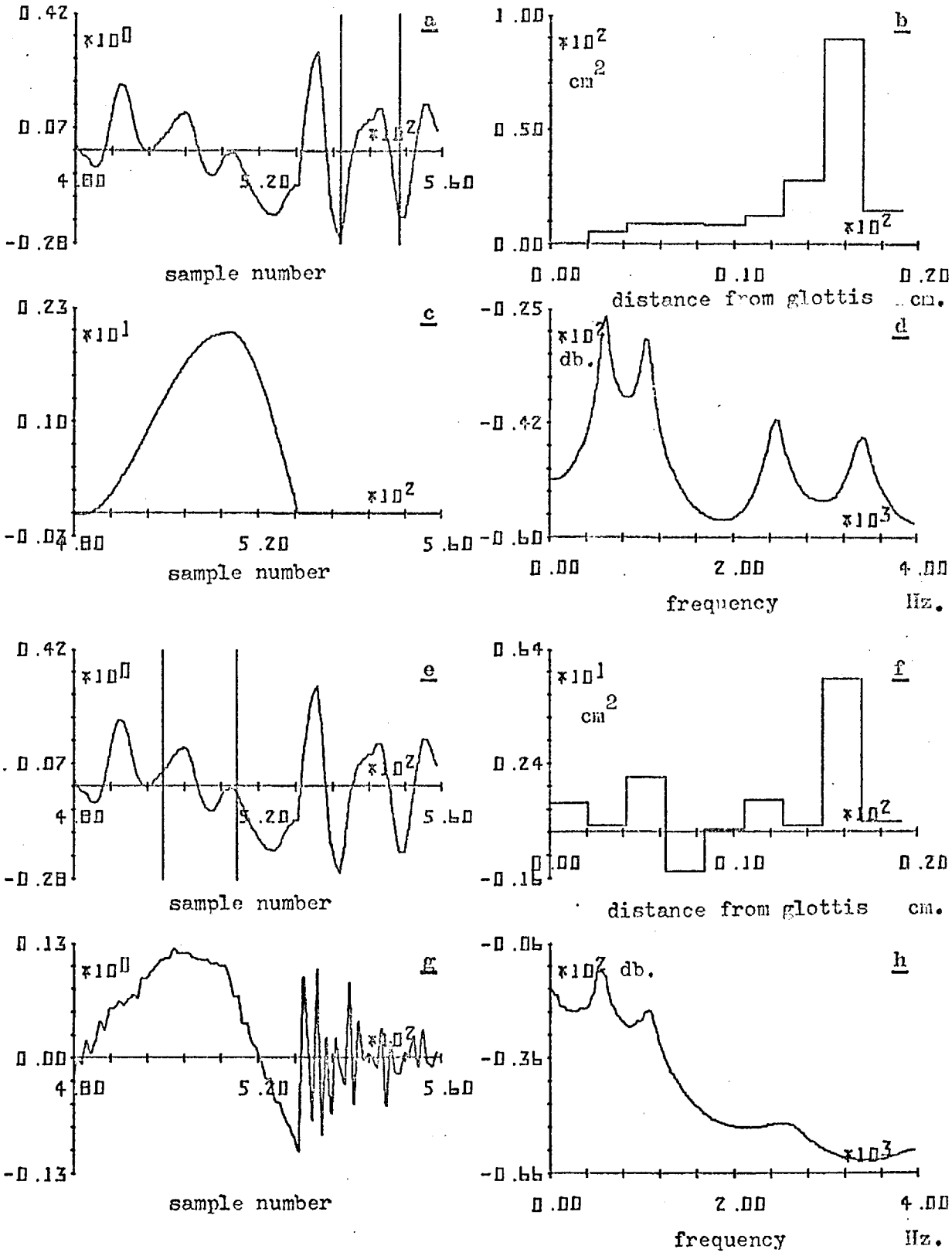
FIGURE 4.2 Formant frequencies and bandwidths used for five vowels (After FANT (1970) based on BESK calculations)(Pages 109 and 126)

Some results from testing the algorithm are given in Figures 4.3 through 4.5. In these cases the excitation function was taken to be a half period cosine in the glottis opening period, a quarter period cosine in the glottis closing period and zero in the closed glottis period; (see Figure 1, APPENDIX A 4.2). The details of the method used to obtain the results of Figures 4.3 to 4.5 will now be explained. (As the graphs 4.3 through 4.5 represent the same analyses for different vowels the term graph a will be used to refer collectively to 4.3a, 4.4a and 4.5a. Similarly graph b etc.).

In each case graphs a - d correspond to analyses carried out using a closed glottis analysis region and graphs e - h correspond to analyses carried out using an open glottis analysis region. Eight periods, each consisting of eighty samples of speech band limited to four kilohertz, were calculated using the digital convolution method previously described. The last period being plotted as graph a and graph e. The two vertical lines on graph a represent the closed glottis analysis region used; (the analysis region is defined as the region over which the filter coefficients are chosen to give minimum energy output of the filter). Similarly on graph e the two vertical lines represent the open glottis analysis region. In all cases the analysis regions were chosen to be greater than eight samples so that $m > n$. The methods of Section 4.4 employing Householder transformation and back substitution were used to determine the inverse filter coefficients a_k. Application of the algorithm summarised in section 3.7 allowed determination of the area function from the inverse filter coefficients. This area function is plotted as graph b (for a closed glottis analysis

FIGURE 4.3 ANALYSIS FOR SYNTHETIC VOWEL /a/

(Showing results of analysis for open and closed glottis regions)

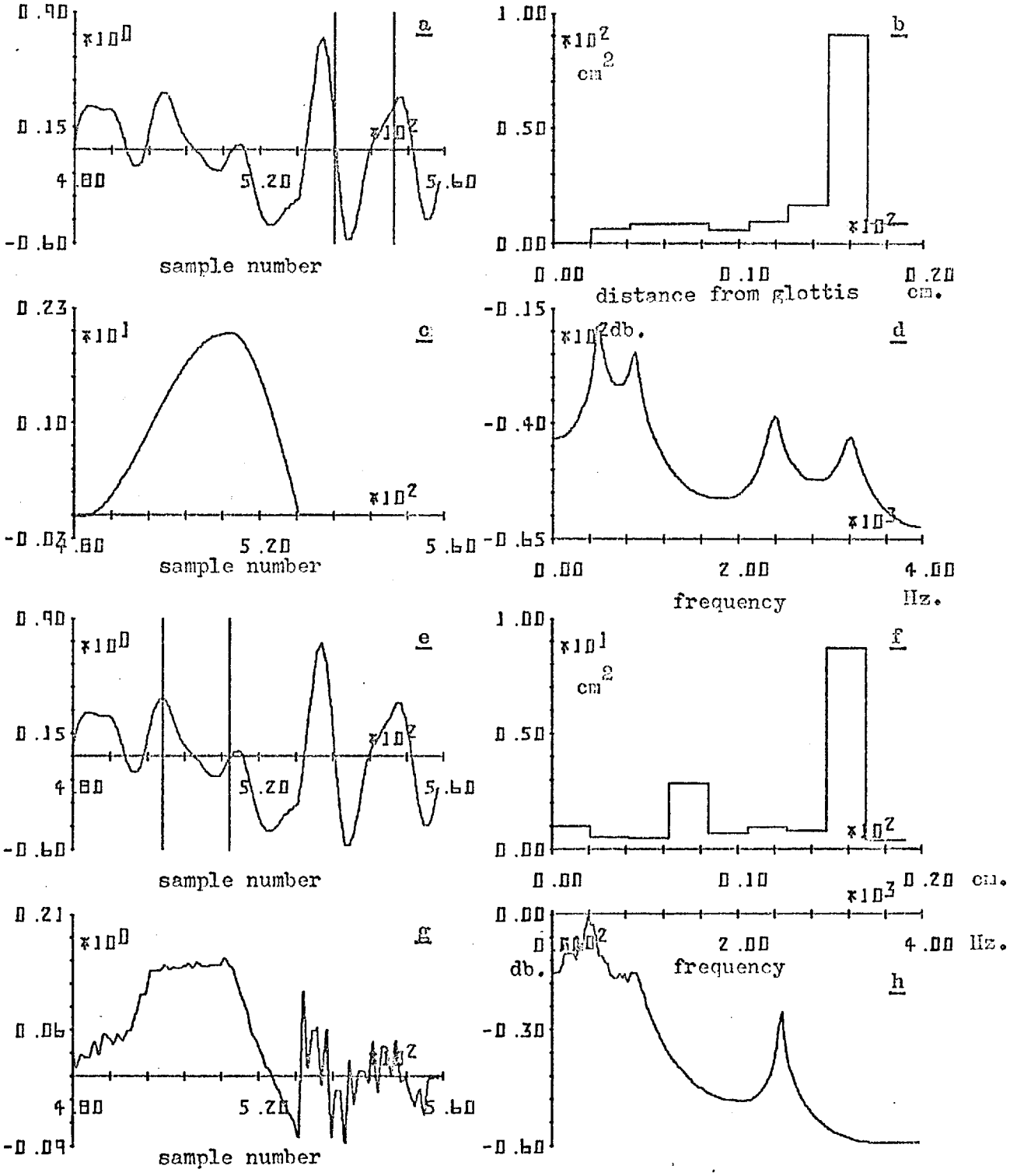


Closed Glottis. Open Glottis.

a	Synthetic Speech Pressure Waveform (Vertical lines denote analysis regions)	e
b	Normalised Area Function (Lips at right)	f
c	Error Signal (Output of inverse filter)	g
d	Reciprocal of Inverse Filter Spectrum	h

FIGURE 4.4 ANALYSIS FOR SYNTHETIC VOWEL /o/

(Showing results of analysis for open and closed glottis regions)

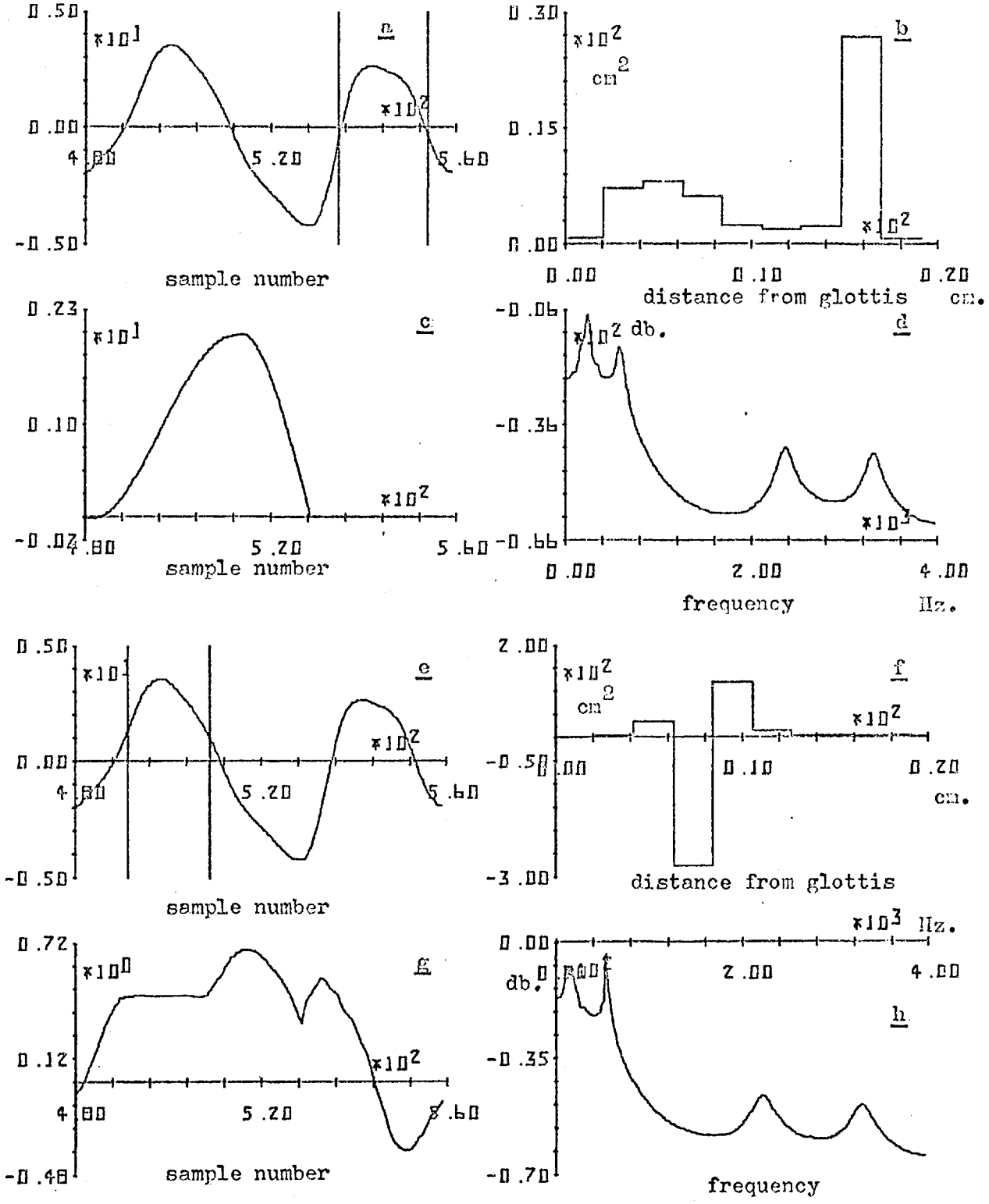


Closed Glottis — Open Glottis

a	Synthetic Speech Pressure Waveform (Vertical lines denote analysis regions)	e
b	Normalised Area Function (Lips at right)	f
c	Error Signal (Output of inverse filter)	g
d	Reciprocal of Inverse Filter Spectrum	h

FIGURE 4.5 ANALYSIS FOR SYNTHETIC VOWEL /u/

(Showing results of analysis for open and closed glottis regions)



Closed Glottis — Open Glottis

a	Synthetic Speech Pressure Waveform (Vertical lines denote analysis regions)	e
b	Normalised Area Function (Lips at right)	f
c	Error Signal (Output of Inverse Filter)	g
d	Reciprocal of Inverse Filter Spectrum	h

region) normalised such that the area of the first tube is unity. To enable evaluation of how well the algorithm separated the glottal excitation function used, the volume velocity at the lips was passed through the inverse filter to give the output of the inverse filter, which we call the error signal. If the inverse filter coefficients have been correctly chosen, the error signal should be the same as the glottal excitation signal, to within a gain constant. To complete the picture, the reciprocal of the inverse filter spectrum was calculated by MARKELS method (1972) and plotted as graph d. Again if the glottal excitation and the vocal tract transfer function have been properly separated, this spectrum should be the vocal tract transfer function used to synthesise the speech.

Choosing the analysis region to be during the open glottis period results in the area function (f), the error signal (g) and the reciprocal of the inverse filter spectrum (h).

From Figures 4.3 through 4.5 the following observations can be made:-

1) In each case, for closed glottis analysis, the output of the inverse filter (graph c) can be seen to be the same as the excitation function used, to within a gain constant (Figure 1, APPENDIX A 4.2).

2) The reciprocal of the inverse filter spectrum (graph d) always has four formants, closer inspection shows that these correspond to the formants used to specify the synthesis filter.

3) The normalised area function, plotted as graph b for the closed glottis analysis region has all its cross sectional areas positive and the position and degree of constriction correspond to those tabulated in the phonemic description of Figure 1.2 (or to those of the vowel quadrilateral of Figure 1.3). As the data used for the formants was taken from calculations based on X-ray data one might expect the area function to be the same as those derived from the X-rays by Fant. This is not the case because the losses of the vocal tract were ignored in the calculation of the area functions. In the next chapter it will be argued that ignoring losses results in larger cross sectional areas at the lip end of the tract. Application of a weighting (inversely proportional

to the distance from the glottis), to the graphs of the area function (b) gives a closer correspondance to the area functions measured by FANT (1970). A first order approximation to the losses will be discussed in the next chapter.

4) The error signal (g) obtained using an open glottis analysis period never corresponds to the glottal excitation function.

5) In figure 4.3f and 4.5f negative cross sectional areas result from using open glottis analysis regions. Negative areas correspond to instability of the recursive filter model of the vocal tract of section 3.6.. (ATAL (1971) has shown that an unstable filter results in negative area functions). I would propose that the filter is unstable because the effects of the glottal excitation have been included in the calculation of the inverse filter coefficients.

6) The first three formants of figure 4.3h approximately correspond to the first three formants of figure 4.3d. It is likely that as the fourth formant is not clearly visible in figure 4.3h that some of the coefficients of the inverse filter have been used to compensate for the general spectrum trend of - 12 db per octave caused by the glottal zeros. In chapter 2 (equation 2.16) it was shown how the zeros of the inverse filter have to be modified to account for the zeros of the excitation function. Similar comments are true for graphs 4.4h and 4.5h.

7) Inspection of the error signal of graphs c and g, shows that if these error signals were differentiated the maximum discontinuity would always occur at glottal closure. As differentiation of graph g corresponds to passing the speech pressure waveform through the inverse filter (instead of the waveform representing the volume velocity at the lips as used to give graphs c and g), this suggests a method for finding the time of glottal closure from the speech waveform. This will be further discussed in the next chapter.

In addition to the results presented here several hundred other analyses have been carried out. For these analyses many values for formant frequencies and bandwidths have been used, together with various excitation waveforms. The excitation waveforms were chosen as triangular waveforms followed by a zero period, short pulses and various ratios of open to closed glottis periods were used for the excitation of figure 1, APPENDIX A 4.2. Varying numbers of formants from one to four were also used.

Providing the closed glottis period was greater than 2n samples the following comments were found to be true:-

a) Of the previously tabulated observations 1, 2, 4 and 7 were always found to be true.

b) The inherent ideas of the observations 3, 5 and 6 were also upheld in general.

c) Use of any analysis portion greater than or equal to n samples taken entirely in the closed glottis region always resulted in the same results being obtained for the same vowel.

Also a number of analyses were carried out directly on the synthetic volume velocity at the lips; these are reported in the paper of APPENDIX A 4.2. The same results were found for any closed glottis analysis region, when either the speech pressure waveform, or the volume velocity at the lips was analysed. This suggests that my method is not affected by the modelling of the lip radiation as that of a spherical source.

I think it is reasonable to conclude that analysis carried out by the method of this chapter is capable of separating the glottal excitation function from the transfer function of the vocal tract, under the following constraints:-

i) The analysis region must be chosen to be greater than or equal to n samples of the closed glottis region which must consist of more than 2n samples, (n is taken as 8 in this section). The reason for the need for 2n samples of the closed glottis region is that the algorithm requires n other samples as well as the analysis region to compute the inverse filter coefficients.

ii) The correct number of filter coefficients must be chosen. In the case of graphs 4.3 through 4.5 the number of filter coefficients was correctly chosen as twice the number of pole pairs of the synthesis filter.

iii) The vocal tract transfer function used in the synthesis must be an all pole function.

The next section describes a comparsion of my method with the method of WAKITA described in section 2.3 of this thesis.

4.6. COMPARISON WITH WAKITAS METHOD

In his paper WAKITA (1972) gives a fortran program for implementing his method, it was this program that was used to allow comparison of the

two methods. Synthetic speech generated in the manner described in the previous section was used, with the formant frequencies and bandwidths chosen as those of figure 4.2. Results obtained are given in figures 4.6 through 4.10. For the analysis method described in this chapter (Rogers method) the analysis region was chosen according to criterion (i) of the previous section. In the method of WAKITA the autocorrelation coefficients were calculated for the whole period of the speech waveform shown in graphs a, because this was found to give the best results.

The normalised area functions were calculated in an equivalent manner (graphs e and f) and the error signals (graphs g and h) were obtained by passing the volume velocity at the lips through the inverse filter. Filter coefficients for the original vocal tract transfer function (graph b) were calculated by polynomial expansion of the denominator of the vocal tract transfer function of equation 4.1 (i.e. they are the a_k's of this equation). These original filter coefficients were calculated to allow comparison with the inverse filter coefficients derived for both methods (graphs c and d).

From the results of graphs 4.6 through 4.10 the following observations can be made:-

1) The inverse filter coefficients obtained by my method (graph c) are always the same (to within the limits of numerical accuracy) as those for the original filter (graph b).

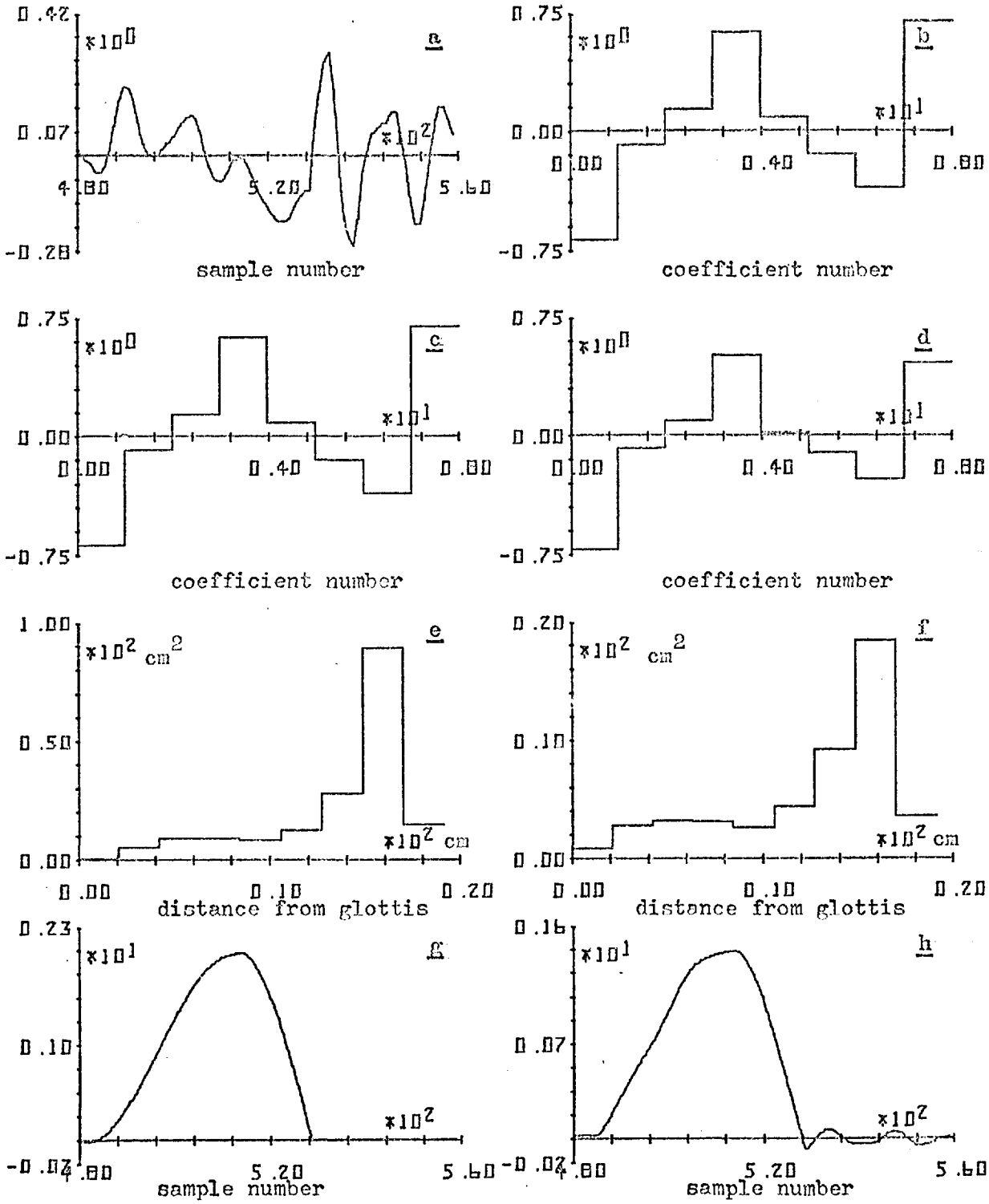
2) Wakitas inverse filter coefficients (graph d) are only a good approximation to the coefficients of the original recursive vocal tract filter (graph b).

3) Comparison of the area functions graphs e and f, especially in figure 4.10 suggests that the area function is less sensitive to errors in the calculation than the inverse filter coefficients are (graphs c and d). (This observation was also made in the two papers given as APPENDICES A4.2 and A4.3).

4) Different values on the ordinate axes ofthe area function of (graphs e and f) result because I consider the area to be normalised such that the area of the section nearest the glottis is unity whereas Wakita normalised the area function such that the area of the glottis is unity).

Figures 4.6 to 4.10 clearly demonstrate that my method produces better results when applied to the closed glottis region of synthetic speech than Wakitas method does when applied to a pitch period of speech.

FIGURE 4.6 COMPARISON WITH WAKITA'S METHOD FOR SYNTHETIC VOWEL /a/

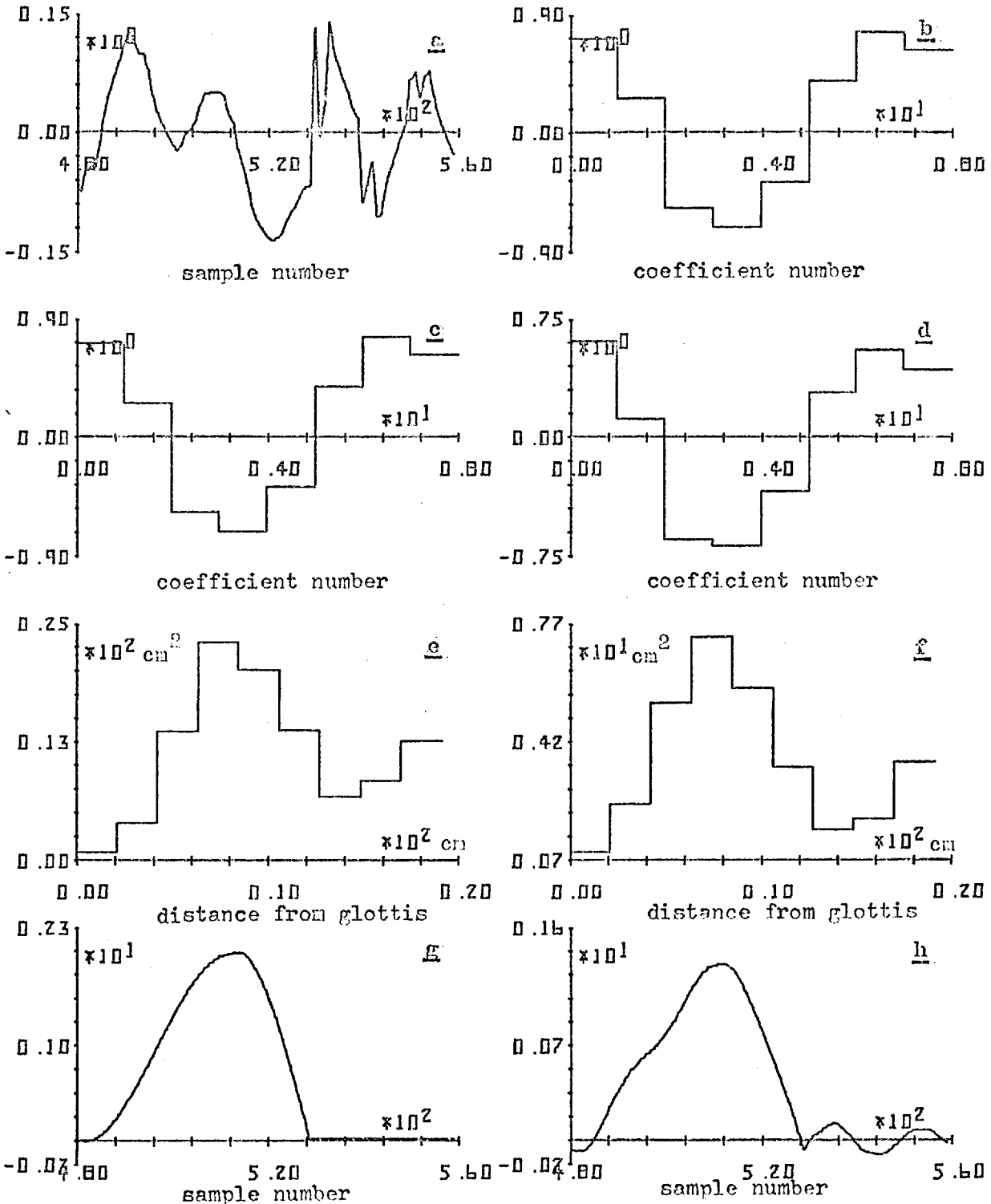


a Synthetic Speech Pressure Waveform

b Filter Coefficients for Original Vocal Tract Transfer Function

Rogers' method		Wakita's method
c	Inverse Filter Coefficients	d
e	Normalised Area Function	f
g	Error Signal (Output of inverse filter)	h

FIGURE 4.7 COMPARISON WITH WAKITA'S METHOD FOR SYNTHETIC VOWEL /e/

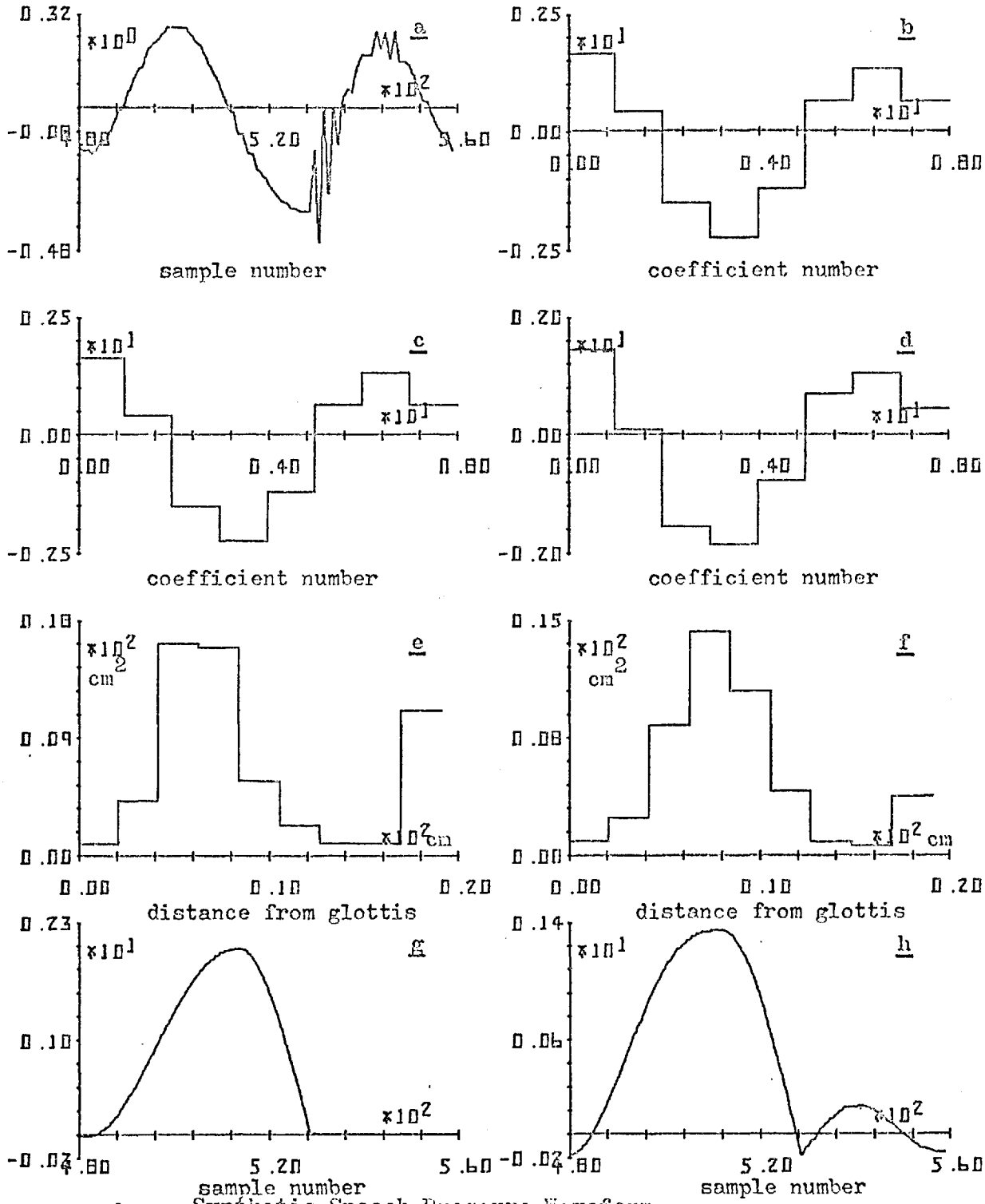


a Synthetic Speech Pressure Waveform

b Filter Coefficients for Original Vocal Tract Transfer Function

Rogers' method		Wakita's method
c	Inverse Filter Coefficients	d
e	Normalised Area Function	f
g	Error Signal (Output of inverse filter)	h

FIGURE 4.8 COMPARISON WITH WAKITA'S METHOD FOR SYNTHETIC VOWEL /i/

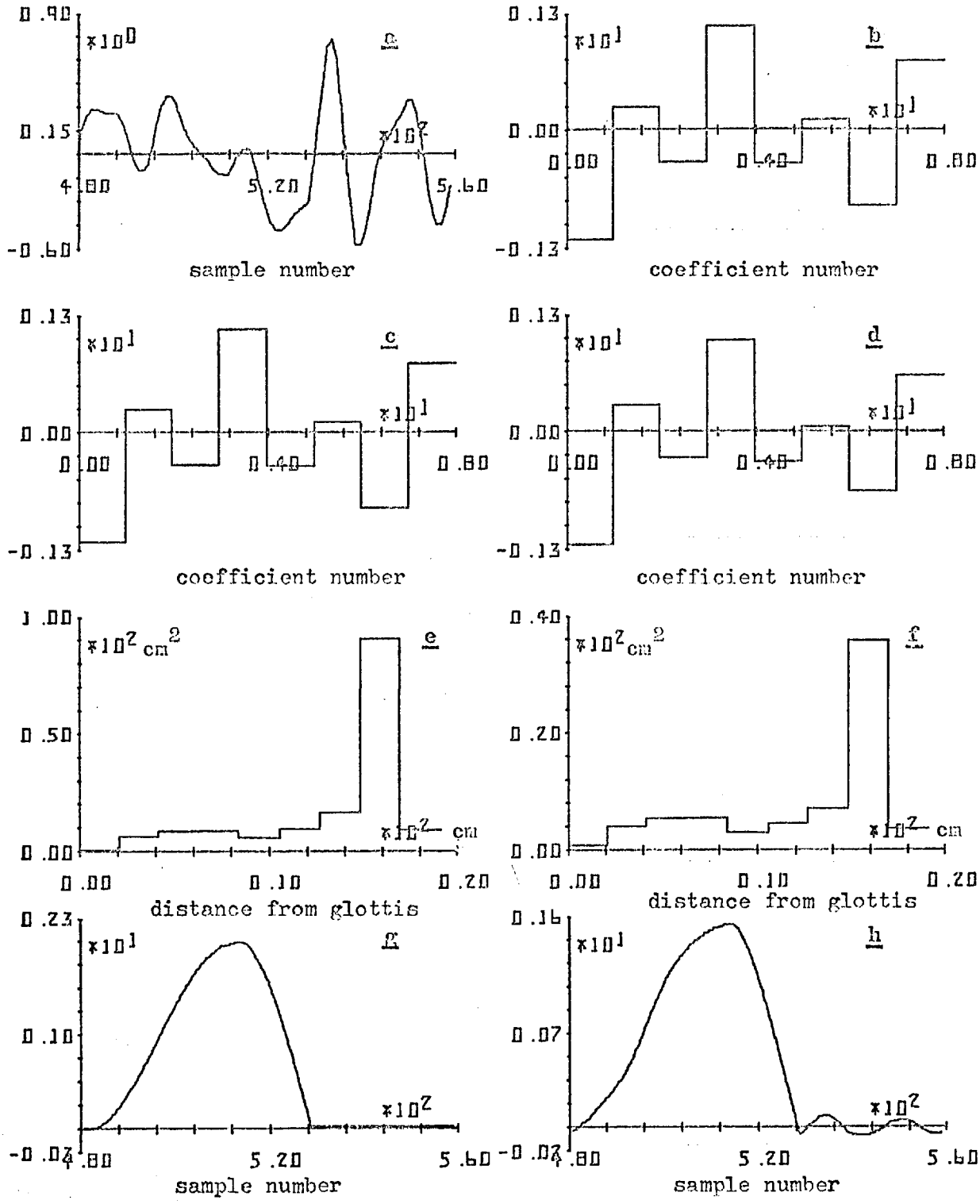


a Synthetic Speech Pressure Waveform

b Filter Coefficients for Original Vocal Tract Transfer Function

Rogers' method		Wakita's method
c	Inverse Filter Coefficients	d
e	Normalised Area Function	f
g	Error Signal (Output of inverse filter)	h

FIGURE 4.9 COMPARISON WITH WAKITA'S METHOD FOR SYNTHETIC VOWEL /o/

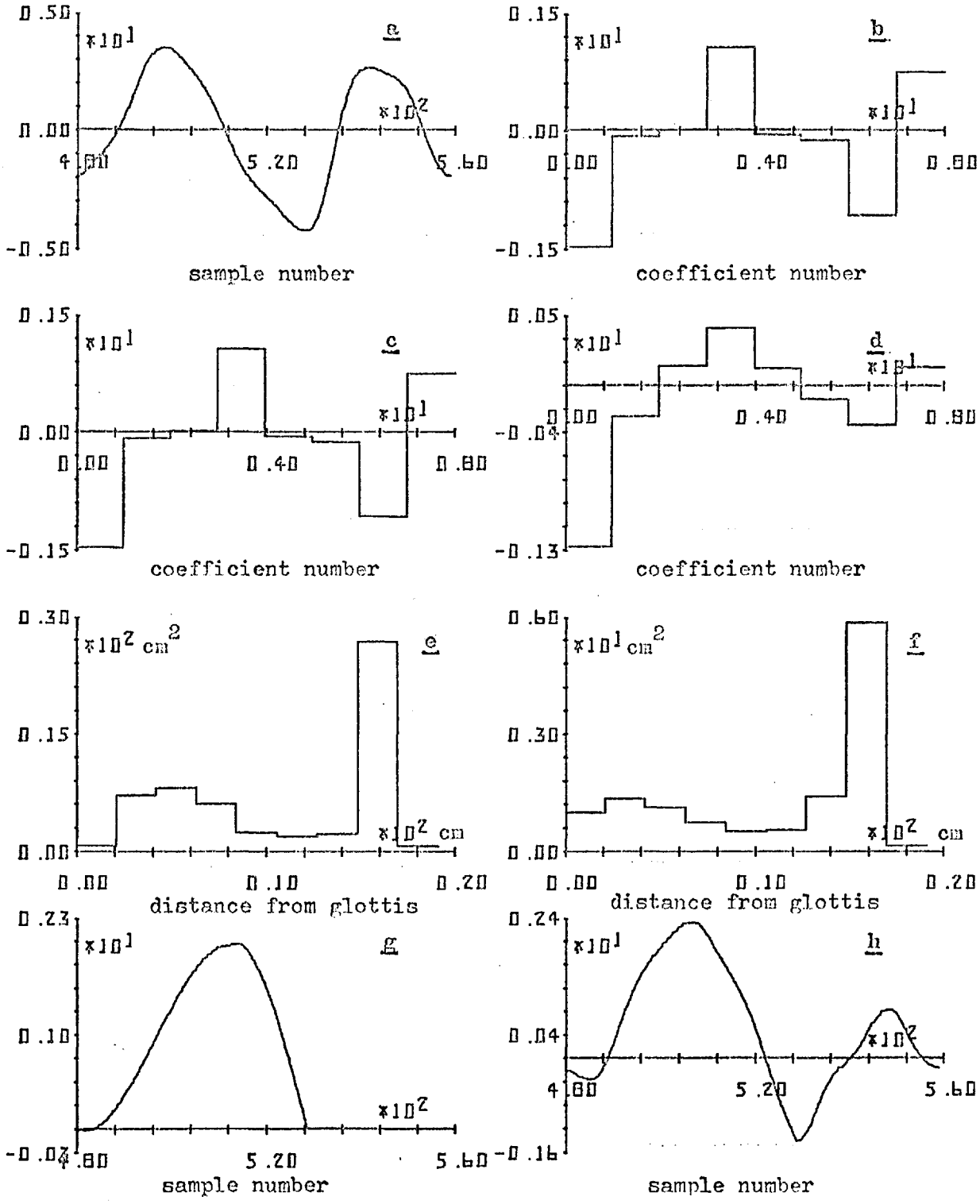


a Synthetic Speech Pressure Waveform

b Filter Coefficients for Original Vocal Tract Transfer Function

Rogers' method		Wakita's method
c	Inverse Filter Coefficients	d
e	Normalised Area Function	f
g	Error Signal (Output of inverse filter)	h

FIGURE 4.10 COMPARISON WITH WAKITA'S METHOD FOR SYNTHETIC VOWEL /u/



a Synthetic Speech Pressure Waveform

b Filter Coefficients for Original Vocal Tract Transfer Function

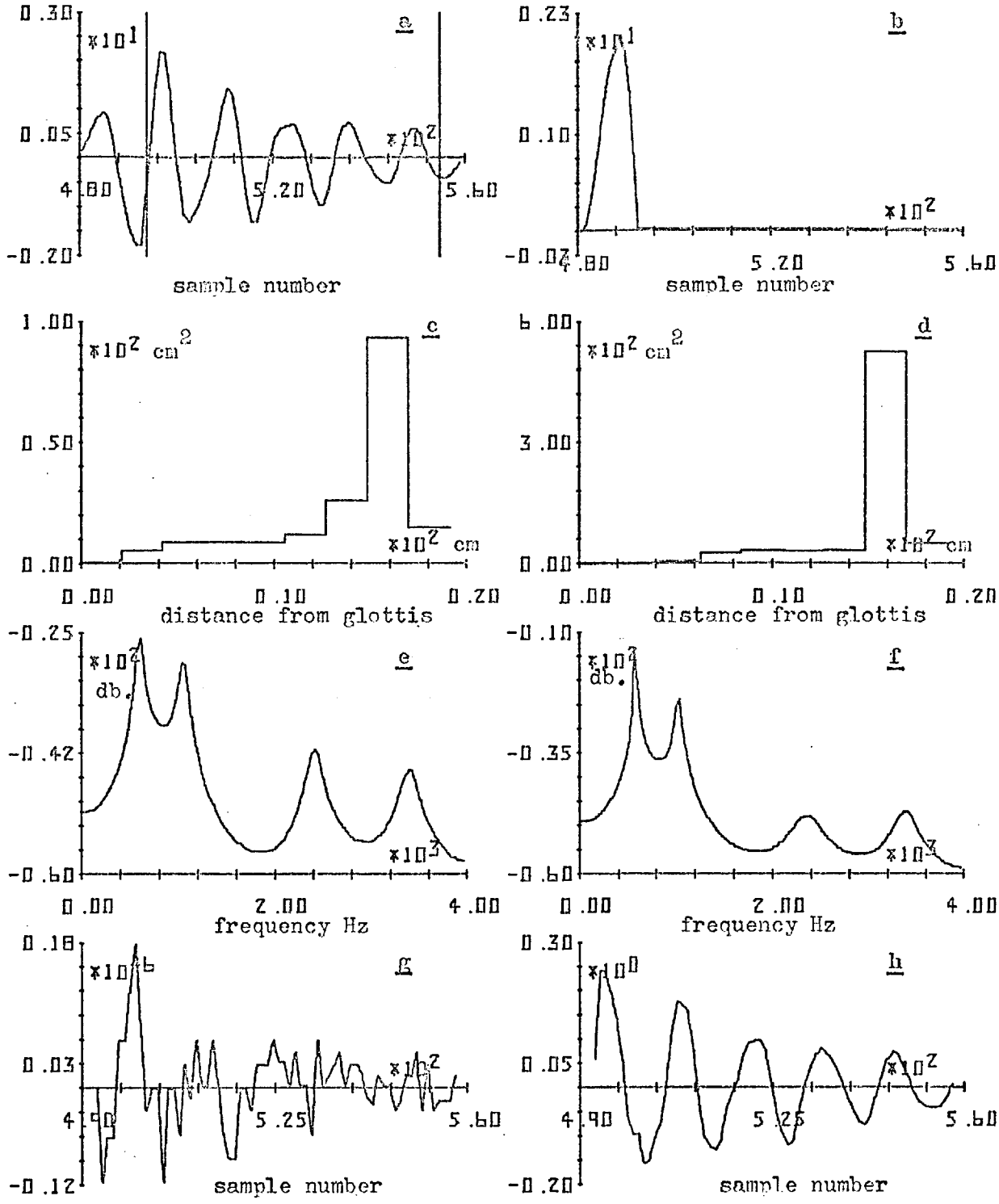
Rogers' method		Wakita's method
c	Inverse Filter Coefficients	d
e	Normalised Area Function	f
g	Error Signal (output of inverse filter)	h

It seems reasonable to ask whether Wakitas method gives better results if it is applied only to closed glottis regions of the speech waveform. Application of Wakitas method to only closed glottis regions was tried and it was found that worse results were obtained, than for application of his method to one pitch period. It was thought that the worse results might be caused by the small number of samples present in the closed glottis region from which one has to calculate the autocorrelation coefficients. In order to test if the small number of coefficients caused the trouble, speech was synthesised for an abnormally high ratio of closed to open glottis periods, the results are shown in Figure 4.11. Synthesis was carried out as before and the analysis region was chosen to be between the two vertical lines for both algorithms. This result suggests that even for such artificially generated synthetic speech Wakitas method is not capable of producing the accuracy of my method. One final point worth mentioning is that the error signals of figure 4.11g and 4.11h have substantially different ordinate axes scales. Inspection of the error signal during the closed glottis region, for analysis during the closed glottis region (figure 4.11g) shows that the error signal calculated by my method is approximately 146 db down on the original speech whereas the error signal for Wakitas method obtained under the same analysis conditions is only 20 db down. This difference can be taken as giving some quantitative feel for the relative accuracies of the two methods.

In concluding this chapter we will summarise the more important results derived. An analysis method has been developed which allows separation of the vocal tract transfer function from the glottal excitation function and this method has been tested under the controlled environment provided by synthetic speech. The analysis method is based on inverse filtering and the filter parameters are chosen automatically by minimising the output of the inverse filter during the closed glottis regions. The method has been found to be considerably more accurate than Wakita's method for separating the vocal tract transfer function from the glottal excitation function.

It should be emphasised that although the method works well for synthetic speech, when applying the method to real speech a number of other factors (like correcting for the lip loading effects, careful choice of the length of the inverse filter and of the analysis region), need to be considered. These factors will be dealt with in the next chapter.

FIGURE 4.11 COMPARISON WITH WAKITA'S METHOD FOR SYNTHETIC VOWEL /e/



N.B. BOTH ANALYSES CARRIED OUT IN CLOSED GLOTTIS REGION

a Synthetic Speech Pressure Waveform (Vertical lines denote closed glottis region)

b Short Open Period Glottal Pulse Used in Synthesis.

Rogers' method		Wakitas method
c	Normalised Area Function	d
e	Reciprocal of Inverse Filter Spectrum	f
g	Error Signal Closed Glottis Region Only (Output of Inverse Filter).	h

APPENDIX A 4.1

FORTRAN PROGRAM FOR IMPLEMENTING HOUSEHOLDER TRANSFORMATION BACK SUBSTITUTION AND AREA FUNCTION MAPPING USED FOR ARTICULATORY ANALYSIS

```
      SUBROUTINE HOUSE(S,C,M,N,B,AREA,REFL)
C     DESCRIPTION OF PARAMETERS
C     S  M BY N COEFFICIENT MATRIX  (DESTROYED)
C     C  M LENGTH RIGHT HAND SIDE VECTOR
C     M  ROW NUMBER OF MATRICES S & C
C     N  COLUMN NUMBER OF MATRIX S, ROW NUMBER OF MATRIX C
C     B  N LENGTH SOLUTION VECTOR OF INVERSE FILTER COEFFICIENTS
C   USES HOUSEHOLDER TRANSFORMATION SEE G.GOLUB METHODS FOR SOLVING
C    LINEAR LEAST SQUARES PROBLEMS, NUMERISCHE MATHEMATIQUE VOL7
C   ISS.  9 (1965)PP206-216
      DIMENSION S(1),C(1),B(1),REFL(1),AREA(1),B1(15),B2(15)
C   DECOMPOSITION LOOP OF HOUSEHOLDER TRANSFORM
      IST=-M
      DO 11 K=1,N
      IST=IST+M+1
      IEND=IST+M-K
C   COMPUTATION OF PARAMETER SIG
      SIG=0.
      DO 1  I=IST,IEND
1     SIG=SIG+S(I)*S(I)
      SIG=SQRT(SIG)
C   GENERATE CORRECT SIGN OF PARAMETER SIG
      H=S(IST)
      IF(H)2,3,3
2     SIG=-SIG
C   GENERATION OF VECTOR UK IN K'TH COLUMN OF MATRIX S AND OF  BETA
3     BETA=H+SIG
      S(IST)=BETA
      BETA=SIG*BETA
      IF(K-N)4,8,8
C   TRANSFORMATION OF MATRIX S
4     ID=0
      JST=K+1
      DO 7 J=JST,N
      ID=ID+M
      H=0.
      DO 5 I=IST,IEND
      II=I+ID
5     H=H+S(I)*S(II)
      H=H/BETA
      DO 6 I=IST,IEND
      II=I+ID
6     S(II)=S(II)-S(I)*H
7     CONTINUE
C   TRANSFORMATION OF RIGHT SIDE VECTOR OF C
8     H=0.
      II=IST
      DO 9 I=K,M
      H=H+S(II)*C(I)
9     II=II+1
      H=H/BETA
      II=IST
      DO 10 I=K,M
```

```
          C(I)=C(I)-S(II)*H
10        II=II+1
11        S(IST)=-SIG
C   BACK SUBSTITUTION
          IEND=(N-1)*N+N
          B(N)=C(N)/S(IEND)
          NMONE=N-1
          DO 13 I=1,NMONE
          NMI=N-I
          H=C(NMI)
          IEND=IEND-1
          IST=IEND
          DO 12 J=1,I
          NMJ=N-J+1
          H=H-S(IST)*B(NMJ)
12        IST=IST-N
13        B(NMI)=H/S(IST)
C   HAVE NOW CALCULATED INVERSE FILTER COEFFICIENTS
C   CALCULATE REFLECTION COEFFICIENTS
          NP1=N+1
          DO 14 I=1,N
14        B1(I)=B(NP1-I)
          DO 17 J=1,N
          REFL(J)=B1(J)
          JP1=J+1
          DENOM=1.-REFL(J)*REFL(J)
          NPJP1=N+JP1
          DO 15 I=JP1,N
15        B2(I)=(B1(I)-REFL(J)*B1(NPJP1-I))/DENOM
          DO 16 I=JP1,N
16        B1(I)=B2(I)
17        CONTINUE
          AREA(1)=1.
          DO 18 I=1,N
18        AREA(I+1)=AREA(I)*(1.+REFL(I))/(1.-REFL(I))
          RETURN
          END
```

BRITISH ACOUSTICAL SOCIETY APPENDIX A 4.2

"SPRING MEETING" at Chelsea College, London, S.W.3 on
Wednesday 25th April / Friday 27th April. 1973.

SPEECH AND HEARING: Session 'B': Speech Analysis and Transmission.

Paper No:
73SHB6

Digital Inverse Filtering of the Speech Waveform.

John Rogers.
Department of Electrical Engineering,
Imperial College, LONDON SW7 2BT.

The method described uses a digital inverse filter to estimate the vocal tract area function and the glottal excitation function for voiced speech. Householder transformation is used to find the inverse filter coefficients which give minimum least squared output during the period of glottal closure. These coefficients uniquely define the area function of the vocal tract model and inverse filtering the speech waveform gives an estimate of the glottal excitation function.

Results are presented for real and synthetic speech, the analysis being carried out in an interactive graphics enviroment on a PDP 15 computer. The results suggest that the area function obtained is insensitive to errors in estimation of the closed glottis period but the deconvolved glottal pulse is very sensitive to errors in this estimation.

INVERSE VOCAL TRACT FILTER.

The vocal tract transfer function for voiced speech is known to be an all pole function, hence the vocal tract can be modelled by a recursive filter. The inverse filter will be a transversal filter whose transfer function $C_n(z)$ is given by

$$C_n(z) = \sum_{i=0}^{n} a_i^{(n)} . z^{-i} \quad , \quad a_o = 1$$

where n is the number of filter coefficients.

For voiced speech the glottal excitation function is known to have a closed period, i.e. period of zero input volume velocity. Using this knowledge the coefficients can be found by minimising the energy output of the filter, for speech input, during the closed glottal period. The vector of outputs $\underline{g}$ is given by

$\underline{g} = S\underline{a}$ where S is a matrix of speech samples used and $\underline{a}$ is the required coefficient vector. To avoid the trivial solution $a_i = 0$ for all i, it is necessary to separate out the first column of S which can be done as a_o is assumed to be unity to give

$$\begin{bmatrix} g_o \\ g_1 \\ \cdot \\ \cdot \\ \cdot \\ g_m \end{bmatrix} = a_o \begin{bmatrix} s_o \\ s_1 \\ \cdot \\ \cdot \\ \cdot \\ s_m \end{bmatrix} + \begin{bmatrix} s_{-1} & s_{-2} & \cdots & s_{-n} \\ s_0 & s_{-1} & \cdots & s_{-n+1} \\ s_1 & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ s_{m-1} & s_{m-2} & \cdots & s_{m-n} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_n \end{bmatrix} \qquad 1$$

for m n.

Equation 1 is most conveniently solved for $\underline{a}$ by minimising the least squared output using Householder transformation (1).

Alternative methods of finding the inverse filter or predictor coefficients proposed by Markel (2) and Atal (3) require the calculation of an autocorrelation or an autocovariance matrix which makes these methods computationally less efficient than using Householder transformation.

Once these coefficients have been found the vocal tract area function can be found using Wakita's algorithm (7) which is repeated here as:-

Given the coefficients of an n length filter normalised such that $a_o^{(n)} = 1$ identify $a_n^{(n)} = R_n$ 2

where R_n is the n^{th} junction reflection coefficient. This is related to the area function by

$$R_k = \frac{A_k - A_{k+1}}{A_k + A_{k+1}} \qquad 3$$

where A_k is the area of the k^{th} section of the vocal tract model. The n^{th} section is then removed and replaced by a termination matched to the $(n-1)^{th}$ section. The coefficients of the (n-1) section model are then found from those of the n section model by using

$$a_i^{(n-i)} = \frac{(a_i^{(n)} - R_n\, a_{n-i}^{(n)})}{(1 - R_n^2)} \qquad 4$$

Equations 2 3 and 4 define a recursive algorithm which allows calculation of the vocal tract area function. This algorithm is a direct development of the work of Kinarivala (10).

The glottal excitation function can be estimated by inverse filtering the speech waveform and the pitch can be found by Markel's method (4).

RESULTS FROM SYNTHETIC SPEECH.

Synthetic speech was formed by convolving the vocal tract transfer function, defined by its formant frequencies and bandwidths, with a synthetic glottal pulse consisting of a half period cosine in the opening period, a quarter period cosine in the closing period and zero in the closed period, (see fig. 1) which is shown to be well matched spectrally to real glottal pulses by Stansfield (5).

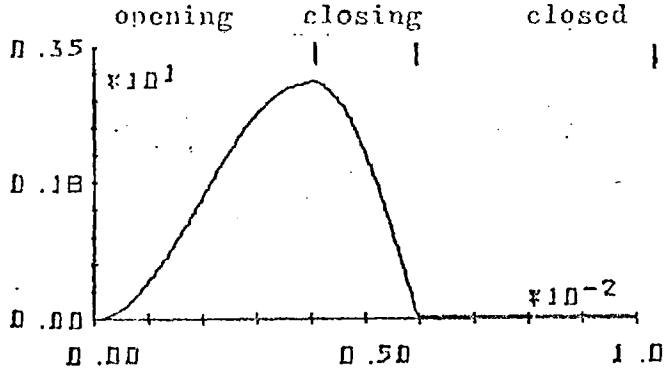
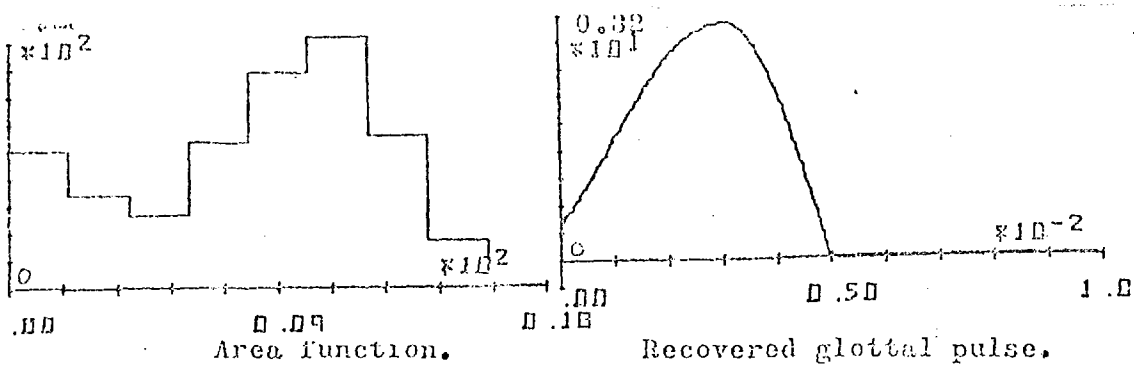


Fig. 1 Synthetic glottal pulse.

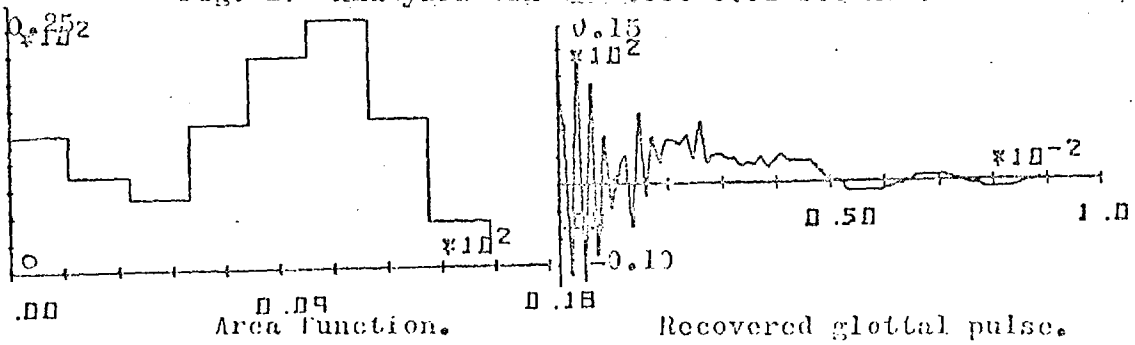
For this type of synthetic speech the glottal pulse is recovered perfectly when the order of the inverse filter is correctly chosen, i.e. equal to twice the number of poles used in the vocal tract, and the analysis is performed over any number of samples entirely in the closed period equal to or greater than the order of the filter. Typical results are shown in fig. 2.

The effect of moving the analysis window to include part of the open period was investigated. It was found that the glottal pulse could no longer be perfectly recovered, but that providing the analysis was carried out entirely during the closed or closing period a very good estimate of the area function is still obtained, see fig. 3. If however the analysis is carried out entirely during the opening period the area function found is vastly different from that obtained by analysis in the closed period. and in some cases the analysis fails, see fig. 4.



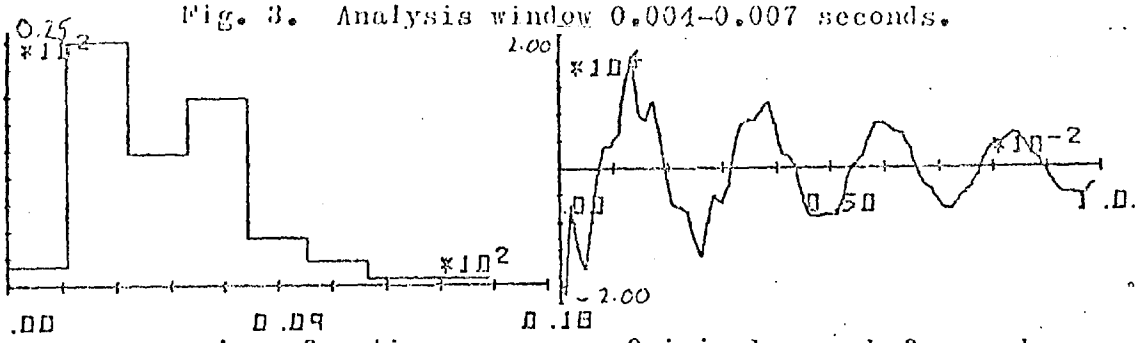
Area function. Recovered glottal pulse.

Fig. 2. Analysis window 0.000-0.01 seconds.



Area function. Recovered glottal pulse.

Fig. 3. Analysis window 0.004-0.007 seconds.



Area function. Original speech for each case.

Fig. 4. Analysis window 0.000-0.003 seconds.

RESULTS FROM REAL SPEECH.

The input for the real speech work was obtained from a capaciter microphone which measures the pressure at a distance from the lips. It has been shown Flanagan (6) that this pressure is approximately given by differentiating the volume velocity at the lips. This means that the speech waveform should be integrated before applying analysis to obtain realistic glottal waveforms. As the contributions due to excitation are specifically excluded from the filter it is not necessary to apply a spectral weighting to account for them as in some methods (7). The results from real speech support those from synthetic speech in the following ways;

1). Realistic area functions can be obtained which are consistent providing the analysis is carried out on the correct portion of the speech waveform.

2). The deconvolved glottal pulse is far more sensitive to changes in the analysis interval than is the area function which is remarkably stable to such changes.

At present the author determines the period of glottal closure experimentally in an interactive graphics enviroment. However, experience suggests that the spikes which are obtained by inverse filtering the speech waveform directly, i.e. without differentiation, give a good estimate of the open glottis period and could provide the basis of an automatic inverse filter program. It is also important to remember that as was pointed out by Holmes (8) reliance can only be placed on the accuracy of the deconvolved glottal pulses if the input speech is obtained from an anechoic chamber and faithfully reproduced, i.e. without phase distortion. No such data is at present available to the author.

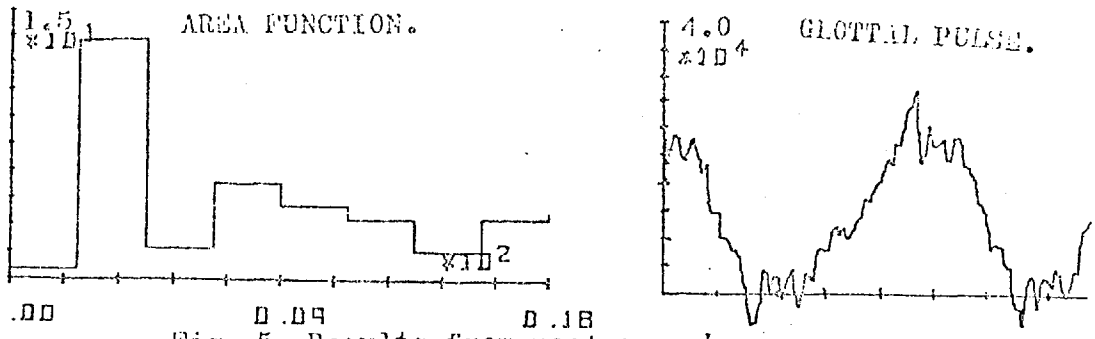


Fig. 5. Results from real speech.

CONCLUSIONS.

A method has been developed which uses Householder transformation and an algorithm consisting of two recursions to give the vocal tract area function. It is the author's belief that more reliable area functions would be obtained if more realistic lip conditions were imposed and losses were included in the analysis. The lip loading can be better approximated if the area of the lips is known. To this end, and to avoid the area normilsation at present necessary, an electronic lip reader based on a television camera has been developed at Imperial College. The losses in the vocal tract can be approximated rather badly at best and one method of including these losses which uses the area of the vocal tract calculated at each stage to modify the transfer function of the remaining sections to avoid accumulative errors is at present under investigation.

The main experimental observations reported here are the facts that the area function is insensitive to errors of the types likely to be encountered in articulatory analysis; this supports earlier work of the author (9), and that the deconvolved glottal pulse is very sensitive to these types of errors. It is hoped now to investigate the usefulness of this method as a feedback loop in teaching deaf people to talk.

REFERENCES.

1). G.H.Golub.1965. Numerische Mathematik 7 (pp206-216)
Numerical methods for solving linear least squares problems.
2). J.D.Markel. 1971. S.C.R.L. monograph No. 7 (pp18-21)
Formant trajectory estimation from a linear least squares inverse filter formulation.
3). B.S.Atal. 1970. J.A.S.A. Vol. 47. No. 1. (pp65?-653)
Speech analysis and synthesis by linear prediction of the speech wave.
4). J.D.Markel. 1972. IEEE. AU20 No. 5. (pp. 367-373).
The SIFT algorithm for fundamental frequency estimation.
5). E.V.Stansfield. 1971. Ph.D. Thesis, University of London.
An articulatory model for speech recognition.
6). J.L.Flanagan. 1972. Springer Verlag.
Speech analysis synthesis and perception.
7). H.Wakita. 1972. S.C.R.L. monograph No. 9
Estimation of the vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion methods.
8). J.N.Holmes. 1962. Paper G13. @ 4th International acoustics congress. An investigation of the volume velocity at the larynx during speech by means of an inverse filter.
9). R.E.Bogner and J.Rogers. 1972.Paper J5 IEEE conference on speech. Determination of the vocal tract area function from a pole description of the vocal tract.
10).B.K.Kinariwala. 1966. B.S.T.J. Vol. 45(pp 633).
Theory of cascaded structures:Lossless transmission lines.

APPENDIX A 4.3

J5

DETERMINATION OF VOCAL TRACT AREA FUNCTIONS FROM A POLE DESCRIPTION OF SPEECH SPECTRA

R.E. Bogner & J. Rogers,

Electrical Engineering Department, Imperial College, London, England.

Summary

We show how unique and consistent area functions for non-nasalized vowels can be synthesized from a knowledge of the magnitude of the transfer impedance of the vocal tract, i.e. (pressure at lips)/(volume velocity at glottis), together with some assumptions about the loading conditions at the lips.

This information is used to determine the input resistance at the glottis as a function of frequency. The input reactance follows by Hilbert transformation. Next the input reflection coefficient in the frequency domain is calculated from the input impedance at the glottis and inverse Fourier transformed to yield the reflection impulse response. This response gives information about the multiple reflections in the vocal tract which is considered as a piecewise constant transmission line. The areas are then found by a recursive processing of the reflection response.

In this paper the transfer impedance is specified by its poles and we examine the sensitivity of the calculated area function to some of the errors liable to be encountered in pole fitting procedures. The effects of different spectral weightings, omission of poles and shifts of pole positions are investigated.

The resulting area functions are encouragingly stable.

Introduction

Our motivations are:

(a) The remapping of acoustic data obtained from real speech into articulatory data could prove to be a profitable step in automatic speech recognition, and

(b) The availability of articulograms (i.e. graphs of area functions displayed during speech) may be helpful in teaching the deaf to speak as this appears to be a very natural way of coding the feedback information.

The signal processing, which is described in detail elsewhere[1], is all being carried out by a small computer (PDP-15). The procedures may be summarised as follows:

1) We use as a starting point the squared magnitude S(f) of the transfer impedance,

$$S(f) = |P_L(f)/U_S(f)|^2 \qquad (1)$$

where $P_L(f)$ is the pressure at the load (lips) and $U_S(f)$ is the volume velocity at the source (glottis). In this paper we study how well the poles fitted to a "measured" speech spectrum serve to describe S(f). The earlier work[1] used a description of S(f) by its values measured at the pitch harmonics.

2) The vocal tract is assumed to be effectively lossless except for the radiation impedance at the lips. It follows that

$$R_{in}(f)|U_S(f)|^2 = G_L(f)|P_L(f)|^2 \qquad (2)$$

where

$R_{in}(f)$ is the real part of the input impedance,

$$Z_{in}(f) = R_{in}(f) + jX_{in}(f)$$

and $G_L(f)$ is the real part of the load admittance,

$$Y_L(f) = G_L(f) + jB_L(f).$$

So far we have been using a constant real value for $Y_L(f)$ to facilitate comparison with experimental results we obtained by speaking into a tube with no reflections. The conductance of the lip loading is in any case substantially independent of frequency (Fig.1) and we are proceeding with a study to include realistic susceptances.

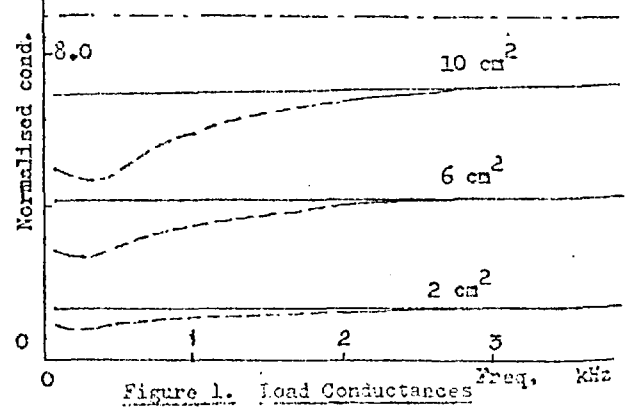


Figure 1. Load Conductances

——— Circular piston in plane baffle
– – – – Circular piston in sphere of rad.9cm.
·—·—· Matched tube of area 9.625cm²

Combining (1) and (2) yields:

$$R_{in}(f) = G_L(f)\, S(f) \qquad (3)$$

and since the network has to be realisable (causal)

$$Z_{in}(f) = R_{in}(f) + j\hat{R}_{in}(f) \qquad (4)$$

where $\hat{R}_{in}(f)$ is the Hilbert transform of $R_{in}(f)$.

3) From $Z_{in}(f)$ we can find the input reflection coefficient $\rho_{in}(f)$ referred to a source resistance of arbitrary, convenient value. This resistance has, for convenience, been taken to be equal to the characteristic impedance Z_0 of the input section of 'the vocal tract', and corresponds to a cross-sectional area of 3cm². The corresponding input impulse reflection response $r_{in}(t)$ follows

by inverse Fourier transformation.

4) The vocal tract is approximated by a cascade of forty equal-length sections of uniform transmission line. The input impulse reflection response, $r_{in}(t)$ is the result of the reflections within this structure. A recursive procedure is used to evaluate, section by section from the input end, the corresponding reflection coefficients and hence the characteristic impedances and areas.

Pole Description of S(f)

The present contribution examines the suitability of a pole description of the transfer impedance spectrum, because there are several advantages available through this description:

1) The pole description is a very compact form of description of the spectrum, requiring relatively few values. It omits the irrelevant pitch frequency information. Also, it is possible to omit from the description details of high order poles without seriously changing the speech quality.
2) There are well-explored techniques for fitting vocal-tract poles to measured speech data[2,3,4].
3) The vocal tract transfer function is all-pole; the excitation all-zero, facilitating separation of the two.

Notice that the real parts of the pole positions or formant bandwidths of the transfer impedance contribute information which is not readily available when area functions are synthesized from lip impedances[5,6]. Inclusion of this information avoids the problem of ambiguities in the synthesised vocal tract shape as it permits the input impulse response to be determined completely.

The present experimental results show the effects of a variety of changes in the transfer impedance spectrum, S(f) described by poles. The changes correspond to the sort of perturbations which could be produced in pole-fitting procedures. They are based on Fant's pole data[7] for /a/ as in father. The poles, before modification are at

Frequency Hz:	616	1072	2470	3410	3820
Bandwidth (twice pole real part) Hz:	57	72	130	175	200

In each case, the lip loading $Y_L(f)$ used in the area synthesis has been real, corresponding to a tube of infinite length and of cross-sectional area 9.625cm^2, which is approximated in our experimental apparatus.

Effects of Spectral Weighting

Discrepancies between a pole-described spectrum and the true spectrum may occur due to the omission of high-order poles, or the mistaking of the number of high order poles. So far we have been considering transfer impedances which are of the form

$$P_L(s)/U_L(s) = \frac{K}{(s-p_1)(s-p_1^*)(s-p_2)(s-p_2^*)\ldots(s-p_N^*)} \qquad (5)$$

for finite N. Fant has examined the spectral weighting introduced by this description in comparison with the spectrum due to a distributed network, and suggests that the differences can be described by suitable weighting functions. The relation is of the form

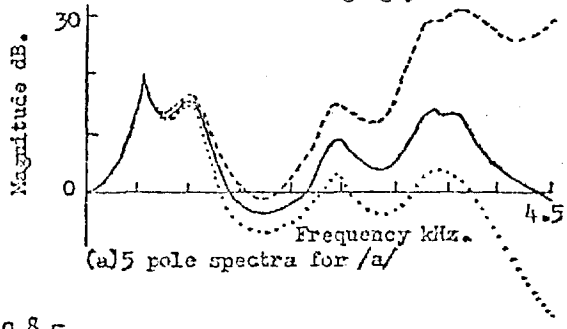
$$20\log_{10}\left|\frac{\text{(Transfer function for distributed system)}}{\text{(Transfer function for four pole system)}}\right| \doteq a_1x^2 + a_2x^4 \qquad (6)$$

where x is an appropriate normalized frequency, and a_1 and a_2 are constants.

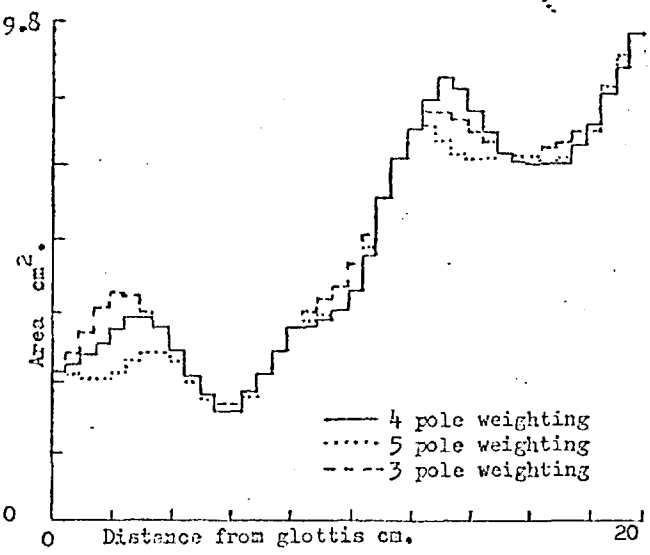
We examined the effects of weightings of this form corresponding approximately to the cases for 3, 4 and 5 poles (Fig.2a). The values used for a_1, a_2 are:

	a_1	a_2
3 poles	0.72	0.0033
4 poles (Fant)	0.54	0.00143
5 poles	0.36	0.0007

These weightings were all applied to the original spectrum to see whether an error in assumption about spectral weighting would cause difficulty. Corresponding area functions are shown in Fig.2b, and are seen to be encouragingly consistent.



(a) 5 pole spectra for /a/



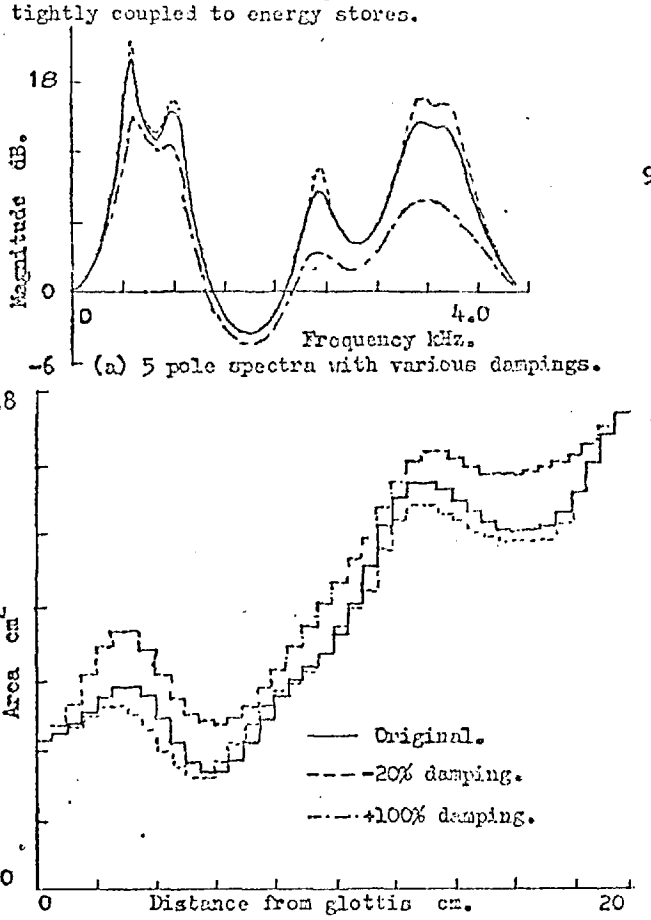
(b) Synthesised area function.

Effects of Changes in Pole Damping

Fig.3a shows the same original 5-pole spectrum incorporating Fant's weighting for 4 poles, together with versions obtained by modifying only

the real part of every pole position by (a) a reduction of 20% and (b) an increase of 100%. We notice that with such broadened peaks, some pole fitting procedures could underestimate the number of poles.

The resultant calculated area functions (Fig.3b) show fairly stable main features. The constrictions in the vocal tract tend to become tighter as damping decreases; this is consistent with the idea that the conductive load at the lips should then be less tightly coupled to energy stores.



(a) 5 pole spectra with various dampings.

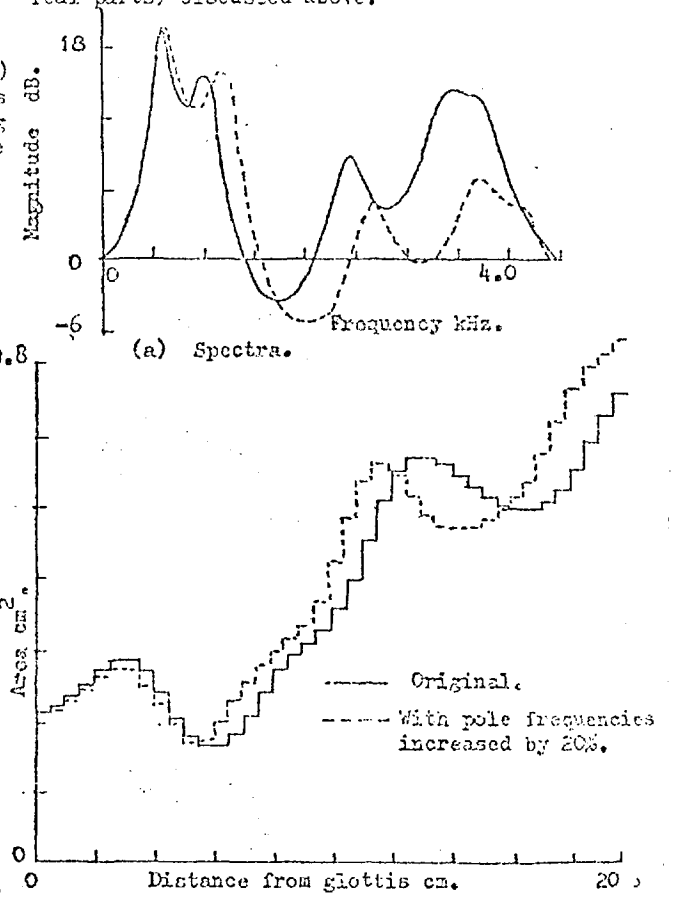
(b) Area functions.

Figure 3. Effects of Pole Damping

Perturbation of Pole Frequencies

As is well known, formant frequencies are not simply relatable to vocal tract cavities and constrictions. Also, as shifts in formant frequency should correspond to changes of vocal tract shape, it is probably not valuable to explore a variety of conditions.

However one important case exists - that corresponding effects on the area function would be an inverse scaling of the distance. Fig.4a and b support this idea. An important implication is evident - automatic adjustment of length of the vocal tract is implicit in the present method of synthesis. Minor modifications of the shape could occur at the same time, to accommodate the slight changes in damping ratio, as we did not simultaneously change the pole dampings in the same proportion. Thus there is superposed an effect similar to the 20% decrease in formant widths (pole real parts) discussed above.



(a) Spectra.

(b) Area functions.

Figure 4. Frequency Scaling Of Pole Positions

Omission of a Pole

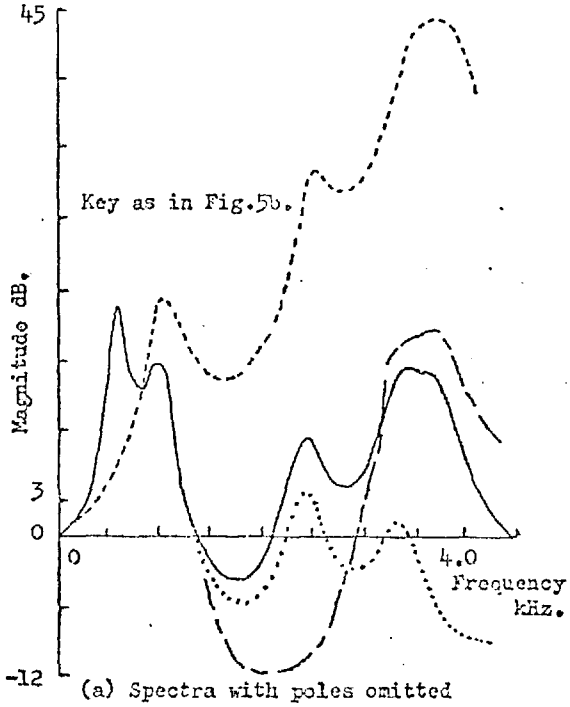
Particularly in spectra where two poles are close, causing merging of formants, some pole-fitting procedures can estimate the wrong numbers of poles. Therefore we examined the effect of removing one of the poles at a time from the spectral description, without other changes (e.g. spectral weighting) - Fig.5a, b. As might be expected, deletion of the higher formants, least significant for intelligibility, results in area functions which are still recognisable. Deletion of the first formant changes the shape radically.

The more broad-minded reader might be able to see the periodic nature of the perturbation of the area function corresponding to the deletion of the formant at 2.47 kHz. The period is about 6cm, corresponding to a half-wavelength at this frequency. This is in agreement with the correlation between energy density and sensitivity to boundary perturbations.

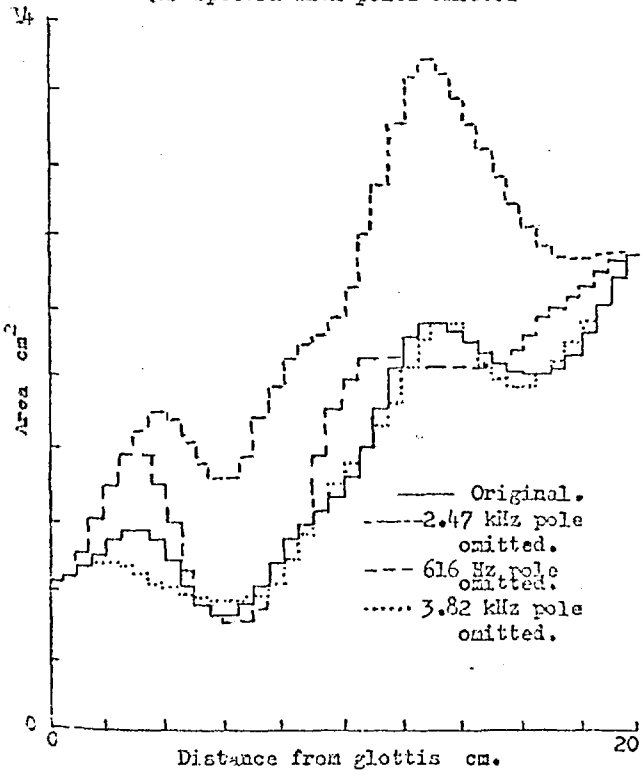
Modification of Termination

One of the assumptions in the method of area

function synthesis is that the real part of the load admittance is known. What if an error is made in this? Fig. 6 shows the effect of a change in the loading condition - a change of the area of the load tube from 9.625 cm^2 to 4 cm^2. The spectrum specified is that of Fig. 2, with 4-pole weighting.



(a) Spectra with poles omitted



(b) Area function.

Figure 5. Effect Of Omitting A Pole

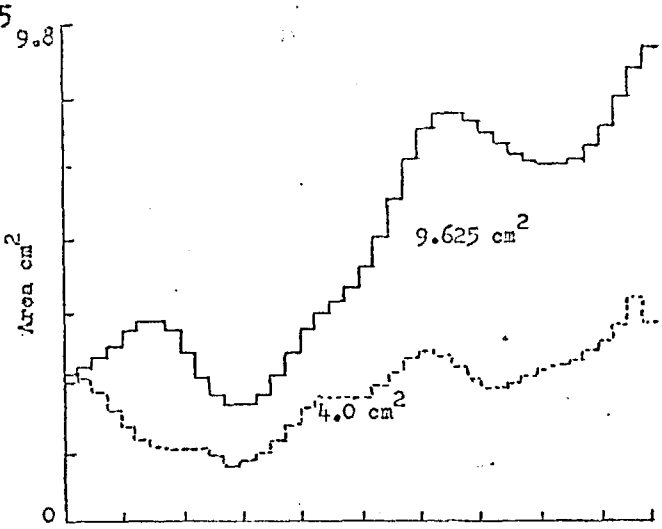


Figure 6. Effect of Changing Load Conductance

Conclusions

These preliminary results, which have been supported by similar ones for other vowels suggest that reliable area functions may be calculated from a pole description of the speech spectrum. Perturbations of the pole specification in ways likely to be encountered are accommodated by the procedure, and in particular, the length of the vocal tract does not have to be specified exactly. The area functions produced are recognisable by eye and speech sounds synthesised from them are also.

The studies are now being extended to include the effects of realistically reactive lip loading, and relaxation of the assumptions about the area of the glottis.

Acknowledgements

We are pleased to acknowledge that Edward Stansfield contributed most of the basic synthesis procedure and programs, and John Holmes emphasised the desirability of the pole description. Financial support came from the Science Research Council and the Joint Speech Research Unit.

References

1. STANSFIELD, E.V. & BOGNER, R.E.:'Determination of Vocal Tract Area Function From Transfer Impedance', To be published.
2. LIEDTKE, C.E.:'Pole-Zero Determination of the Vocal-Tract Transfer Function', IEEE Trans. on Audio & Electroacoustics, V AU-18N4,p.394,1970.
3. PAUL, A.P., HOUSE, A.S. & STEVENS, K.N.:'Automatic Reduction of Vowel Spectra : An Analysis-by-Synthesis Method & Its Evaluation', J. Ac. Soc. A., V.36N2, Feb.1964, p.303.
4. HOLMES, J.N.:Private communications concerning inverse filtering.
5. SCHROEDER, M.R.:'Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements' J. Ac. Soc. Am, V41, Pt.2, P.1002, 1967.
6. PAIGE, A. & ZUE, V.W.:'Computation of Vocal Tract Area Functions', IEEE Trans. on Audio & Electroacoustics, V.AU-18, No.1, p.7, 1970.
7. FANT, G.M.:'Acoustic Theory of Speech Production', Mouton & Co., The Hague, 1960.

CHAPTER FIVE

ANALYSIS OF REAL SPEECH

In chapter four, the method developed in this thesis for calculating the vocal tract area function from samples of the speech waveform was tested using synthetic speech. The results from these tests suggested that separation of the vocal tract transfer function from the glottal excitation function was possible; providing the correct number of inverse filter coefficients were calculated from samples of the closed glottis period of the speech waveform. The area functions calculated from these inverse filter coefficients were similar to those measured by FANT (1970), using X-ray techniques. However, it was stated that some method of accounting for vocal tract losses was necessary before better estimates of the vocal tract area function could be made. It was also found that modelling the lip radiation as a simple spherical source in the synthesiser, produced the same results as ignoring the lip radiation effects in the synthesiser.

In this chapter, methods for finding the closed glottis period will be described and justified. An approximate method for correcting the area function to account for losses in the vocal tract will then be derived. Next, the apparatus used to make the recordings of speech will be described and the interactive computer facilities used for analysis will be outlined. Finally, the results of analyses of five vowels, each spoken by four different subjects will be presented.

5.1. CHOICE OF CLOSED GLOTTIS ANALYSIS REGION

The importance of analysis using only samples of the closed glottis period was argued in section 4.2. In order to verify these arguments, analyses of synthetic speech using both open and closed glottis analysis regions were carried out, the results from these analyses were presented as figures 4.3 → 4.5. From these results it was concluded that my method was only capable of calculating realistic area functions if a closed glottis analysis region was used. In these synthetic speech analyses the time of glottal closure was known, because the speech was synthesised in the time domain by convolution of a known glottal pulse with an all pole approximation of the vocal tract transfer function.

However, before analysis of real speech can be attempted some method of estimating the time and duration of glottal closure is needed. Three

methods of estimating this timing were tried. In the first method the small discontinuity of the speech waveform apparent at glottal closure was found by visual inspection. Although this method worked well in practice sometimes the wrong region was chosen and anyway the method is not suitable for automatic implementation which is the ultimate need.

Secondly, a laryngograph was employed to measure the time of glottal closure. The laryngograph is a machine developed at University College, London, (FOURCIN and ABBERTON 1971) which measures the changing electrical impedance across the larynx region of the neck. Two electrodes are used, one held on either side of the adams apple, an R.F. signal is applied to one electrode and the other acts as a receiver for this signal. The changing electrical impedance path caused by the vibrating vocal folds modulates the received signal. Demodulation of the received signal gives an indication of vocal fold vibration and in particular shows sharp discontinuities at glottal opening and closure. Typical laryngograph outputs are shown as figures 5.1a and 5.2a. On these graphs the region at the bottom of the waveform corresponds to an open glottis, the discontinuity and subsequent sharp rise correspond to glottal closure. The slower falling edge corresponds to the vocal folds parting and the sharp discontinuity at the bottom of this edge corresponds to glottal opening. The laryngograph waveform was recorded simultaneously with the speech waveform on a two channel tape recorder, so it could be used to find the closed glottis region.

Although the laryngograph gives an accurate estimate of the time of glottal closure, this information can only be used if the distance from the subjects larynx to the microphone is known. This distance is needed to calculate the time of arrival at the microphone, of the first wavefront resulting from glottal closure. In the recordings used, to calculate the results presented here, the larynx microphone distance was 50 cm. Assuming the velocity of sound in air to be 35,000 cm/sec and realising that a 10 KHz sampling rate was used, it is easy to show that this delay corresponds to 15 samples (1.5 m Sec).

The third method which was most commonly used, utilises the error signal calculated by inverse filtering the speech waveform. This method consists of first calculating some inverse filter coefficients using an arbitrary analysis region, then visually inspecting the error signal to find the closed glottis region. To confirm the validity of this method the error signal is plotted together with the laryngograph waveform as figures 5.1b and 5.2b. In these cases the error signal plotted has been

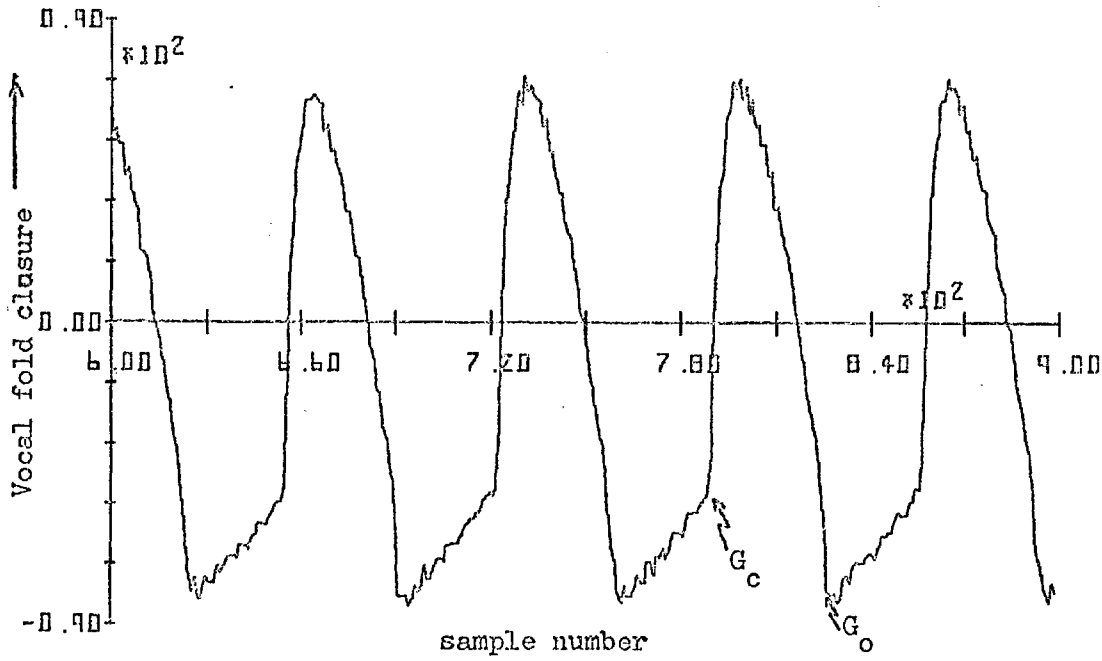


Figure 5 .1a Laryngograph waveform, subject M 1 , vowel /e/

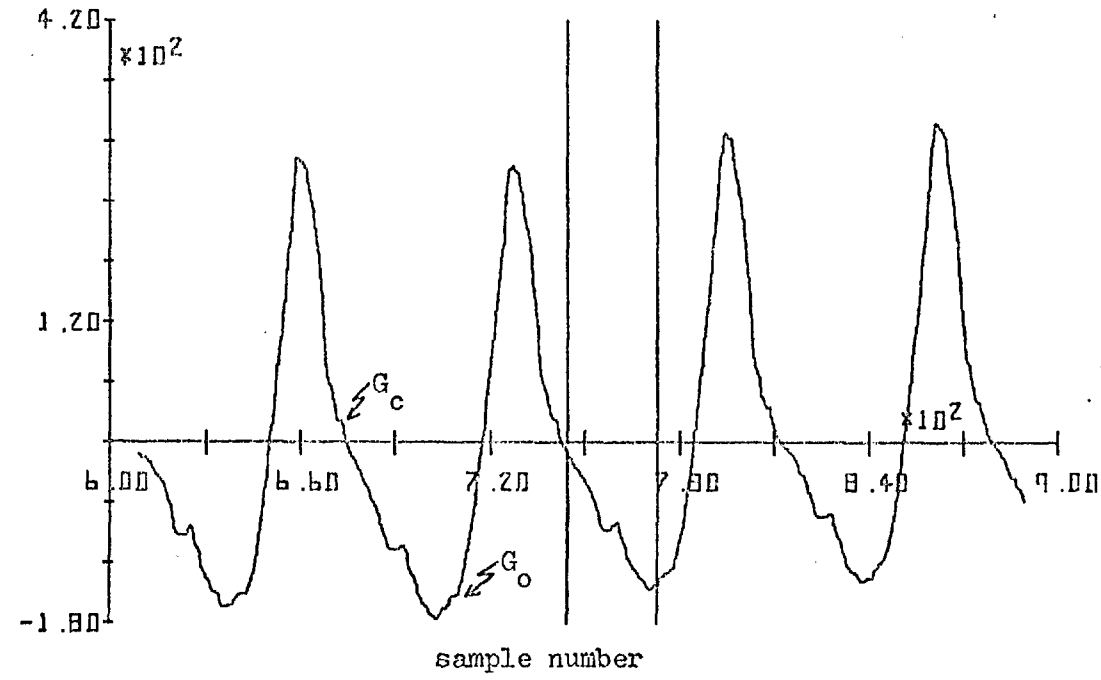


Figure 5.1b Error signal, subject M 1 , vowel /e/

The two vertical lines denote the analysis region used for calculation of the inverse filter coefficients.

G_c = glottal closure

G_o = glottal opening

FIGURE 5.1 COMPARISON OF LARYNGOGRAPH WAVEFORM AND ERROR SIGNAL

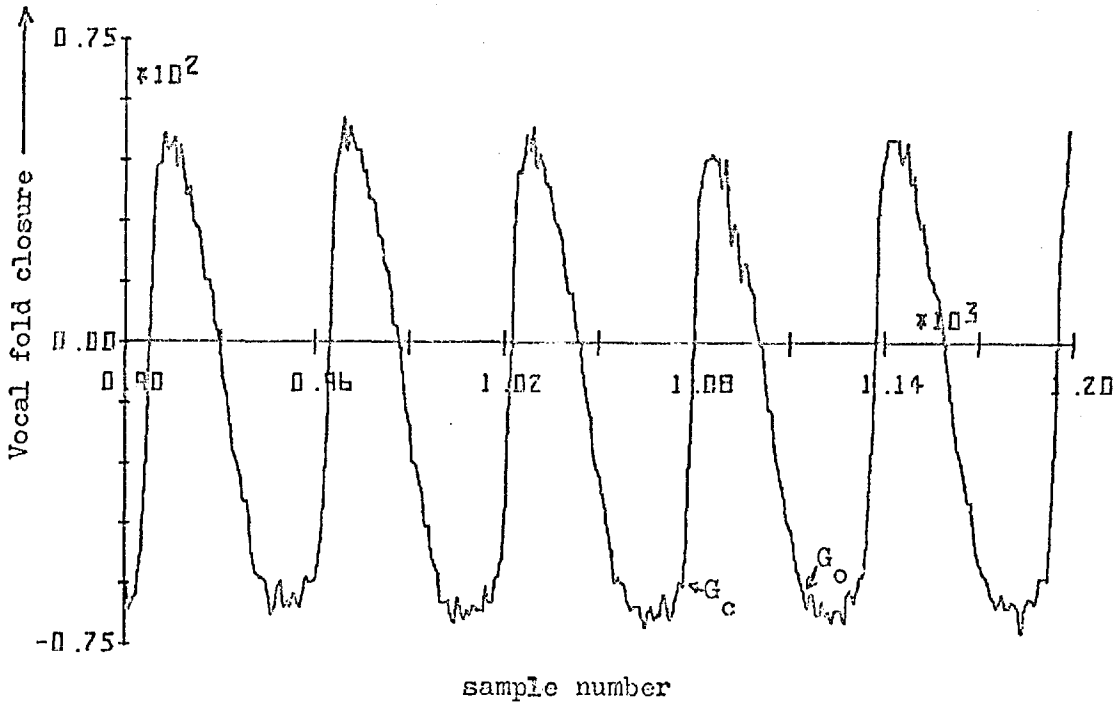


Figure 5.2a Laryngograph waveform subject F 2, vowel /e/

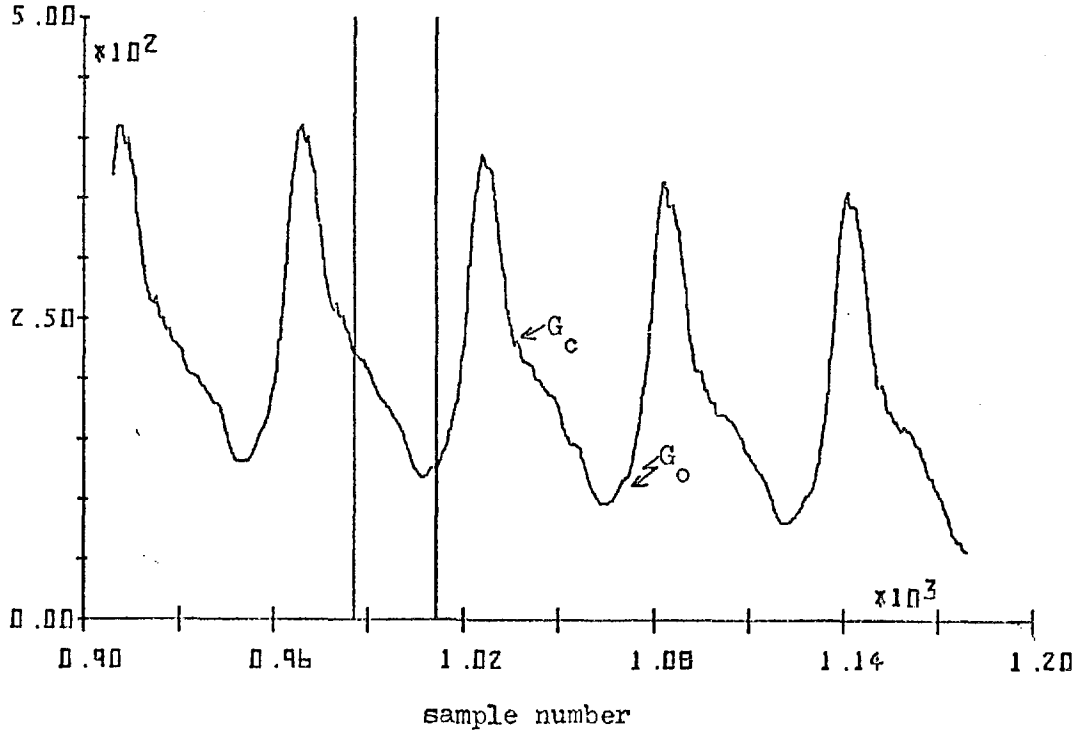


Figure 5.2b Error signal, subject F 2 , vowel /e/.

G_c = glottal closure

G_o = glottal opening

The vertical lines denote the analysis region used to calculate the inverse filter coefficients used.

FIGURE 5.2 COMPARISON OF LARYNGOGRAPH WAVEFORM AND ERROR SIGNAL.

integrated over eight samples to approximate for the lip radiation effects. This integration which smooths the error signal thus making it resemble more closely an actual glottal pulse, means that glottal closure is apparent as the discontinuity on the falling edge of the error signal. Inspection of figures 5.1 and 5.2 shows that this discontinuity occurs approximately fifteen samples after the glottal closure apparent on the laryngograph waveform.

This method using the error signal is preferred because the microphone larynx distance is not needed. In fact no information except the speech waveform is necessary for this method. If automatic analyses of continuous voiced speech were required, the inverse filter coefficients calculated for the previous analysis interval could be used to inverse filter the current region to be analysed, to find the closed glottis period.

5.2. APPROXIMATION OF VOCAL TRACT LOSSES

In chapter three the vocal tract was assumed lossless to allow derivation of the algorithm for calculating the vocal tract area function. A modified method is presented here which corrects the area function derived to approximate for losses in the vocal tract. It must be emphasised that this new method does not synthesise a lossy transmission line, but merely modifies the area function derived for the lossless case to approximate for losses in the vocal tract.

The transmission line model developed in chapter three consisted of a number of abutting constant cross sectional area tubes. In the lossless case the passage of a wavefront through one of these tubes resulted in only a delay, the resonant nature of the vocal tract being accounted for by the reflections at the junctions between the tubes. When the lossy case is considered the attenuation resulting from losses as the wavefront travels through a section must be accounted for. In transmission line theory the losses are incorporated by changing the propagation constant γ. In the lossless case:-

$$z^{T_k} = \exp(j\omega T_k) = \exp(\gamma)\ell_k \qquad (5.1)$$

where γ is the propagation constant per unit length = $j\frac{\omega}{c}$

However, in the lossy case the propagation constant γ is complex, instead of purely imaginary viz

$$\gamma = \alpha + j\beta \tag{5.2}$$

where α = attenuation constant per unit length

β = phase constant per unit length = $\frac{\omega}{c}$

For small losses the phase constant (β) is approximately the same as the phase constant for the lossless case. Inclusion of losses therefore requires the transmission matrix of equation 3.30 to be modified to include the attenuation constant. This modification is equivalent to the change at variable

$$z^{-T_k} \Longrightarrow e^{-\alpha \ell_k} z^{-T_k}$$

and the resulting transmission matrix is given by

$$W_k' = \frac{1}{1 - R_k} \begin{bmatrix} e^{\alpha \ell_k} z^{T_k} & -R_k e^{\alpha \ell_k} z^{T_k} \\ -R_k e^{-\alpha \ell_k} z^{-T_k} & e^{-\alpha \ell_k} z^{-T_k} \end{bmatrix} \tag{5.3}$$

The change of variable merely replaces the delay representing a section in the transmission line model with a delay and an attenuation. Replacing the delay, with a delay and an attenuation factor in equation 3.55 yields

$$H_{n-1}(z) = \frac{e^{\alpha \ell} z^{T}(1 + R_n)}{\frac{R_n}{H_n(1/z)} + \frac{1}{H_n(z)}}$$

$$= \frac{z^{-(n-1)T} \prod_{k=1}^{n-1} (1 - R_k)}{1 - \sum_{k=1}^{n-1} (a_k^{(n)} + R_n a_{n-k}^{(n)}) z^{-2kT}} \Bigg/ e^{\alpha \ell}(1 - R_n^2) \tag{5.4}$$

From equation 5.4 we can see that the modified reflection coefficient mapping algorithm is

$$a_i^{(n-1)} = \frac{a_i^{(n)} + R_n a_{n-i}^{(n)}}{e^{\alpha \ell}(1 - R_n^2)} \qquad (5.5)$$

(c.f. equation 3.56)

Inclusion of losses has the effect of decreasing the magnitude of the reflection coefficients more and more as successive coefficients are calculated using the recursive process of section 3.7. As synthesis is carried out from the lip end of the tract model, this has the effect of applying a weighting inversely proportional to the distance from the glottis to the area function. It was stated in section 4.6 that precisely this type of weighting was required to make the area functions calculated for synthetic speech more realistic.

Now we need to derive the attenuation constant (α) for the case of a lossy acoustic tube. The passage of air past the walls in the vocal tract results in losses caused by; viscous friction between the air and the walls, heat conduction from the air to the walls and vibration of the vocal tract walls. A comprehensive presentation of the losses in a cylindrical transmission line section is given in FLANAGAN (1972, sections 3.2 and 3.73). In these sections Flanagan evaluates the analogous electrical elements for a uniform cylindrical lossy acoustic tube. Assuming that these losses are small they may be expressed as the per unit length attenuations:-

a) Caused by viscous friction:

$$\alpha_V = \frac{S}{2Ac}\sqrt{\frac{\omega\mu}{2\rho}} \qquad (5.6)$$

(FLANAGAN, 1972, Equations 3.8 and 3.33)

b) Caused by heat conduction:

$$\alpha_H = \frac{S(\eta - 1)}{Ac}\sqrt{\frac{\lambda\omega}{2C_p\rho}} \qquad (5.7)$$

(FLANAGAN, 1972, Equations 3.8 and 3.33)

c) Caused by cavity wall vibration:

$$\alpha_W = \frac{r_s S \rho c}{2(r_s^2 + x_s^2)A} \qquad (5.8)$$

(FLANAGAN, 1972, Equations 3.80 and 3.81)

where

S = circumference of the cylindrical acoustic tube

A = cross sectional area of the acoustic tube

c = velocity of sound in the vocal tract (3.5×10^4 cm sec^{-1}, moist air 37^oC)

ω = angular frequency (rad sec^{-1})

μ = coefficient of viscosity (1.14×10^{-3} gm cm^{-3}, moist air 37^oC)

η = adiabatic constant, ratio of specific heats (1.4)

λ = coefficient of heat conduction (5.5×10^{-5} cal cm^{-1}, sec^{-1} $^oC^{-1}$)

C_p = specific heat of air at constant pressure (0.24 cal gm^{-1} ^oC)

r_s = real part of mechanical impedance of skin (6500)

x_s = imaginary part of mechanical impedance of the skin ($0.4\ \omega$)

Substituting these values into equations 5.6 → 5.8 for a cylindrical cross section tube and a frequency of 1 KHz yields

$$\alpha = \alpha_V + \alpha_H + \alpha_W \qquad \frac{1.21 \times 10^{-2}}{\sqrt{A}} \qquad (5.9)$$

It should be noted that no attempt is made to include the frequency dependence of the losses, or to account for the change in phase constant (β) caused by the losses.

The method of correcting the area function to account for the attenuation losses of equation 5.9 will be summarised. First the inverse filter coefficients (a_k) are calculated by Householder transformation, in exactly the same way as for the lossless case. Next, the algorithm of section 3.7 is used to find the reflection coefficient and thence the cross sectional area of the section nearest the lips (as for the lossless case). The losses associated with a tube having this area are then calculated using equation 5.9. Having calculated these losses, the inverse filter coefficients of the transmission line formed by removing this lossy section are calculated using equation 5.5. This process is then repeated until the area function is known. In essence the algorithm including the correction for losses, is given by section 3.7 with equation 3.56 replaced by equation 5.5 where equation 5.9 is used to evaluate the attenuation constant α. This modified algorithm is used to calculate all the area functions presented in this chapter.

5.3. APPARATUS USED FOR ANALYSIS

The apparatus used for real speech analysis can be conveniently divided into two categories, the equipment used to record the speech and the computer facilities used to perform the analyses.

Recordings were made of the speech of three male and three female subjects. Results are presented in the next section for two of the male and two of the female subjects. The subjects were asked to articulate individual phonemes of approximately a seconds duration, several practice runs being made before recording.

A Brüel and Kjaer F.M. tape recorder (model 7001) was used to make the recordings at a recording speed of 15"/sec.. Used at this speed the recorder incorporates a low pass filter in the recording path, (with cut-off frequency 5 KHz, passband ripple < 0.2 db, and attenuation outside the passband of 15 db/octave). On playback another low pass filter is incorporated in the recorder, having a passband ripple < 0.1 db and an attenuation outside the passband of 50 db/octave. The speech waveform and laryngograph output were recorded simultaneously, one on each channel of the tape recorder.

The speech was detected with a half inch capacitor microphone (Brüel and Kjaer type 4134) incorporating a cathode follower preamplifier (type 2615) and was further amplified by a Brüel and Kjaer microphone amplifier (type 2603). The microphone was placed 12" from the subject's mouth and used with its protection grid at a grazing incidence of 90°. Used in this way the microphone has a flat frequency response from D.C. to 20 KHz. Recordings were made in the anechoic chamber belonging to the Department of Phonetics at University College London.

Analysis was carried out on a PDP 15 computer incorporating a VT 15 vector drawn, continuously refreshed display. An interactive graphics program was used for analysis, control being provided by use of a light pen. The tape recordings were digitised using the computer's analogue to digital converter sampling the speech and laryngograph waveforms simultaneously. A playback speed of 1.5"/sec was used with a sampling rate of 1 KHz. This effectively gives a sampling rate of 10 KHz on real time speech because the F.M. tape recorder is capable of frequency transposition. The speech waveform samples (10 bits) and the laryngograph samples (8 bits) were packed into one word and stored on the computer's disk.

Any portion of approximately six seconds of speech, stored on disk in this manner, could be recalled for analysis by typing in the sample numbers

of the first and last samples required. Further selection of the portion of the speech waveform to be analysed was possible by pointing the light pen at the graph of the speech waveform. Having selected the portion to be analysed a number of analysis parameters could be changed and a number of graphical outputs were possible. The analysis region (i.e. the region over which the inverse filter coefficients are calculated to minimise the inverse filter output), could be chosen by pointing the light pen at the limits of this region on the speech waveform graph or by typing the sample numbers of the required upper and lower limits. The length of the inverse filter (i.e. number of inverse filter coefficients to be calculated) could be varied using either the light pen or the teletype.

Possible graphical outputs were:-

a) Speech waveform (from storage on disk)
b) Speech spectrum (by Fourier transformation of the Hamming windowed speech waveform selected for analysis)
c) Laryngograph waveform (corresponding to current portion of the speech waveform).
d) Inverse filter coefficients (calculated by Householder transformation)
e) Reflection coefficients of acoustic transmission line (calculated by the algorithm described in section 5.2)
f) Vocal tract area function (calculated as described in section 5.2)
g) Error signal (result of inverse filtering the speech waveform)
h) Reciprocal of the inverse filter spectrum (calculated from the inverse filter coefficients by Markels method described in section 2.3).

In addition to the interactive facilities a hard copy facility was available. To use this facility a push button on the display was pressed and the picture currently displayed on the screen was plotted on an analogue plotter using the digital to analogue converter.

5.4. RESULTS FROM REAL SPEECH

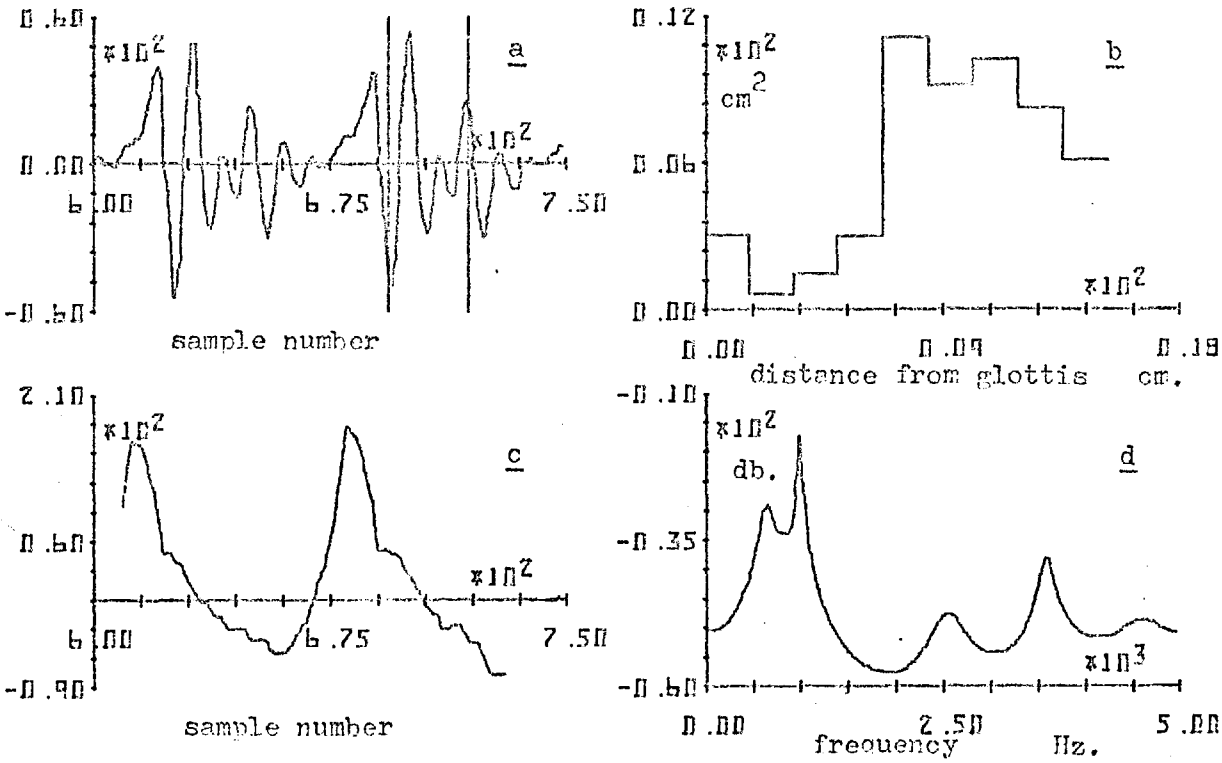
In the preceeding sections of this chapter, the method of recording the speech waveform, the technique used to find the closed glottis region and the computer facilities used to analyse samples of the closed glottis region of the speech waveform were described. Here, the results of analyses

of the five vowels /a/, /e/, /i/, /o/ and /u/ are presented as graphs 5.3 → 5.22. For each vowel, four different sets of results are presented, one set for each of four speakers. All the speakers use English for their natural language, speakers M1 and M2 are male, and speakers F1 and F2 are female.

Four graphs are given for each analysis. In each case graph a represents the segment of the speech waveform chosen for analysis. This segment was chosen during a steady portion of the speech waveform, and was typical of the speech waveform for the particular utterance. The two vertical lines on graph a represent the closed glottis analysis region chosen (i.e. the period of the speech waveform when the energy of the inverse filter output was minimised for the purpose of calculation the inverse filter coefficients).

The speech waveform was pre-emphasised by differencing successive samples before analysis. This differencing applies a spectral weighting of approximately + 6 db/octave and helps to counteract the low energy content of the higher frequency parts of the speech waveform. The area function calculated for each analysis is given as graph b and is normalised such that the area of the lips is 6 cm^2. For the four vowels /a/, /e/, /i/ and /o/ eight inverse filter coefficients were calculated and for the fifth vowel /u/ nine inverse filter coefficients were calculated. These numbers of inverse filter coefficients were chosen because it was found that on average they gave the most realistic area functions. However, it is known that for a 10 KHz sampling rate each vocal tract section will be 1.75 cm long (see equation 3.16). Hence for 8 inverse filter coefficients (9 sections) the vocal tract calculated will be 15.75 cm long and for 9 inverse filter coefficients (10 sections) the vocal tract will be 17.5 cm long. An extra section is included for /u/ to account for the increase in vocal tract length caused by the lip protrusion used in production of that vowel.

The error signal (graph c) was calculated by passing the speech waveform (graph a) through the inverse filter and summing the previous eight samples of the output to give the current sample of the error signal. This "integration" of the error signal was used to counteract the effect of the lip radiation, modelled as a simple spherical source (equation 2.8). A better estimate of the glottal excitation waveform should be given by using a more realistic model of the lip radiation impedance. Unfortunately these more realistic models require a knowledge of the lip area which was not measured in these studies.



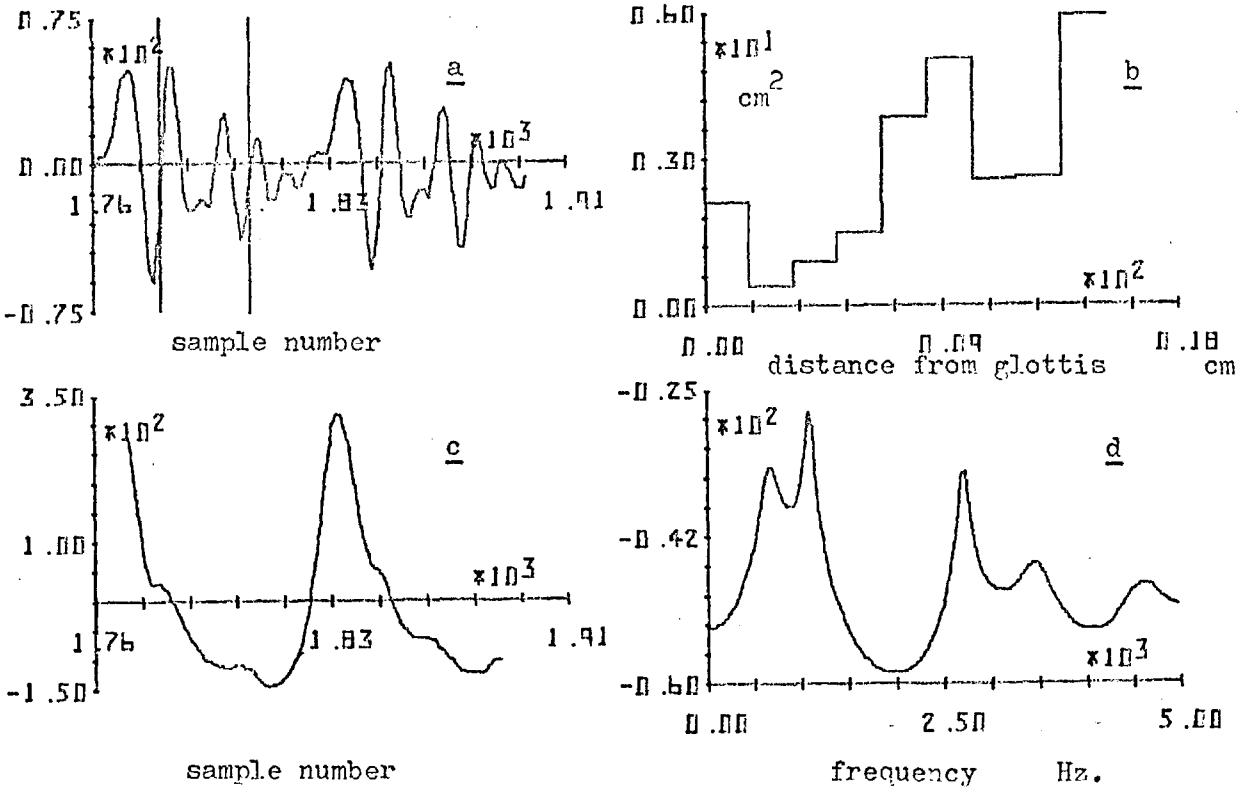
a Speech Pressure Waveform

b Normalised Area Function

c Error Signal

c Reciprocal of Inverse Filter Spectrum

FIGURE 5.3 ANALYSIS FOR VOWEL /a/ SPEAKER M1.



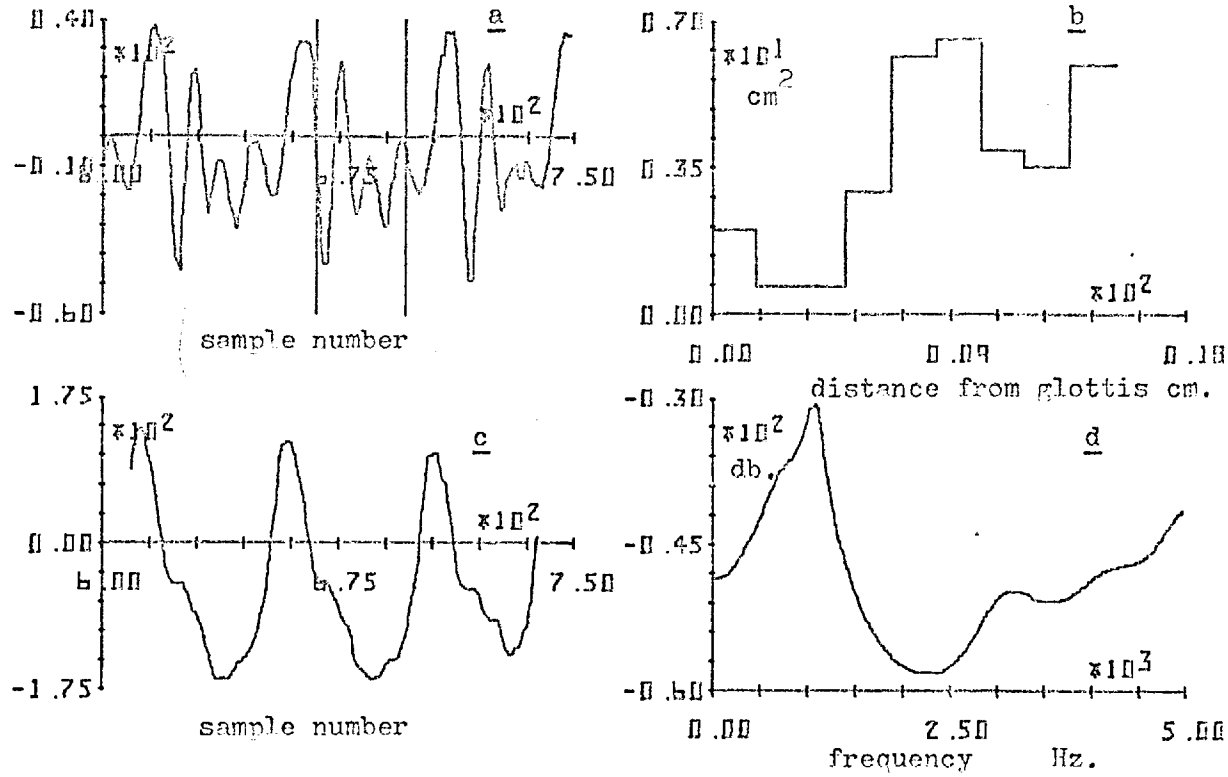
a Speech Pressure Waveform

b Normalised Area Function

c Error Signal

d Reciprocal of Inverse Filter Spectrum

FIGURE 5.4 ANALYSIS FOR VOWEL /a/ SPEAKER M2



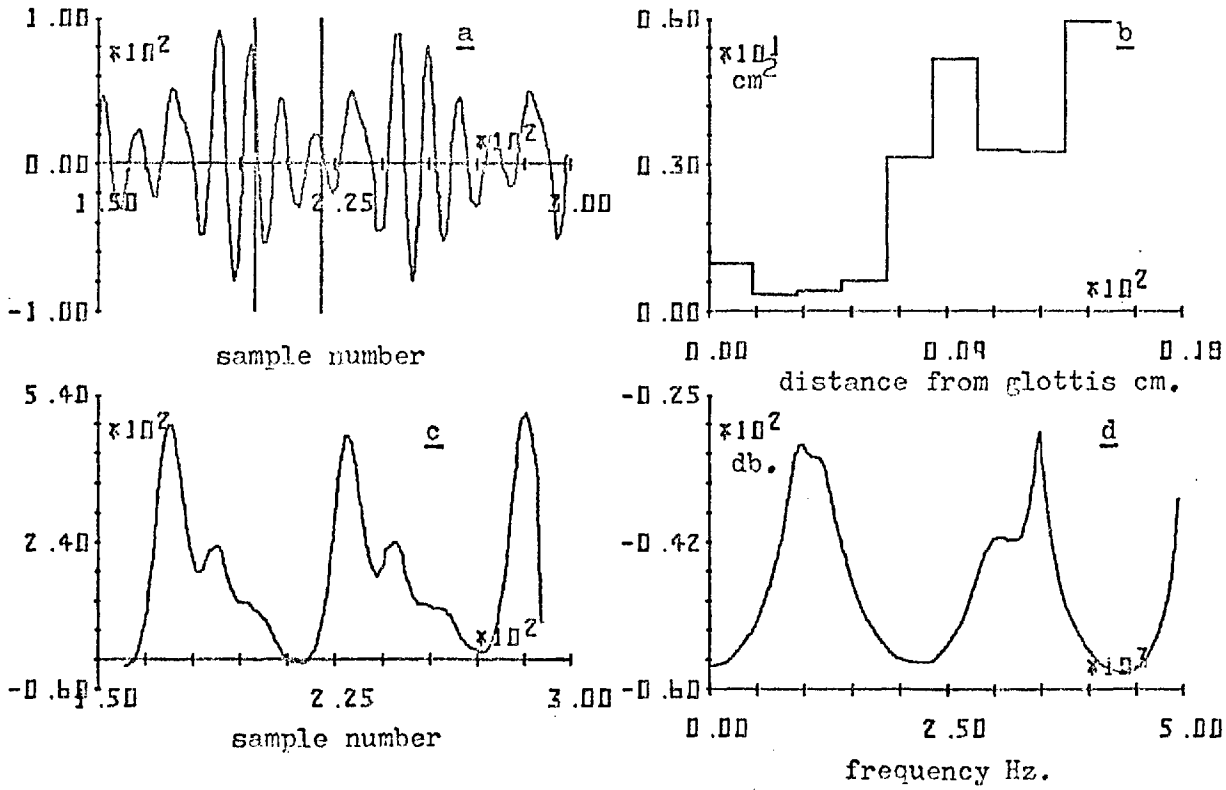
a Speech Pressure Waveform

b Normalised Area Function

c Error Signal

d Reciprocal of Inverse Filter Spectrum

FIGURE 5.5 ANALYSIS FOR VOWEL /a/ SPEAKER F1



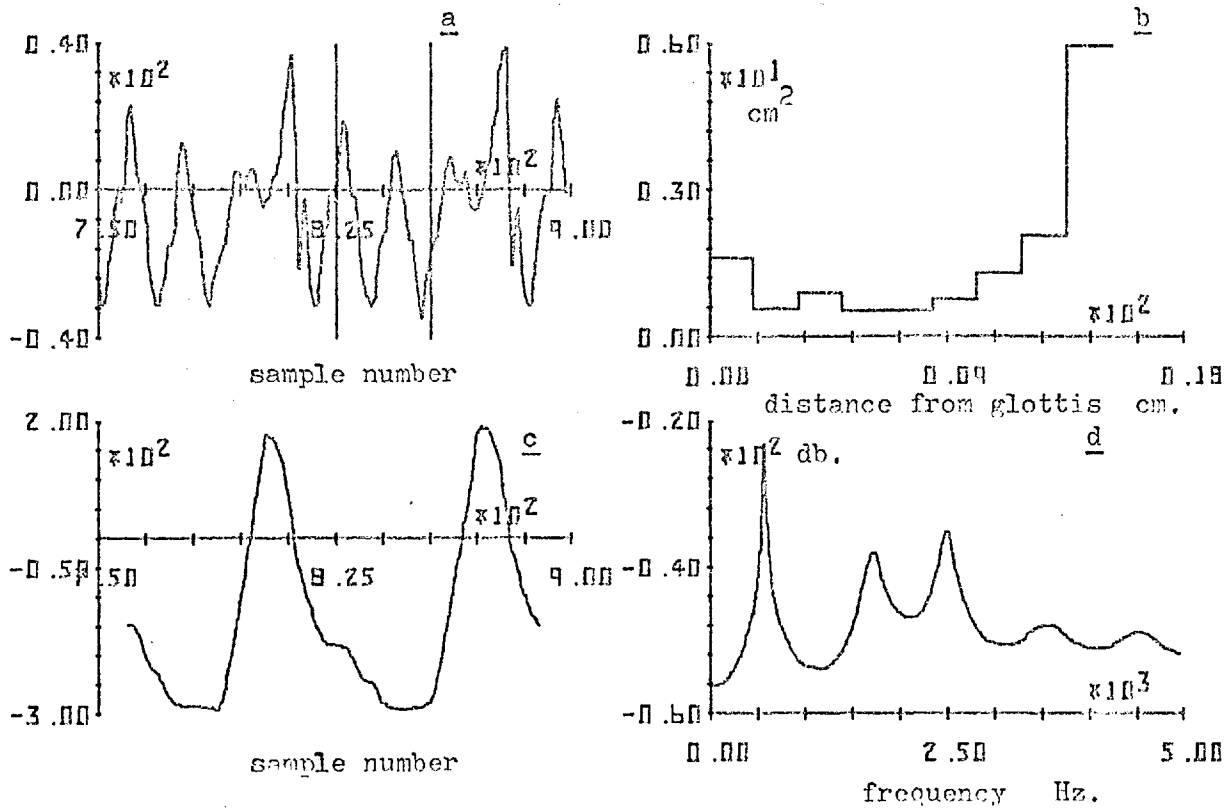
a Speech Pressure Waveform

b Normalised Area Function

c Error Signal

d Reciprocal of Inverse Filter Spectrum

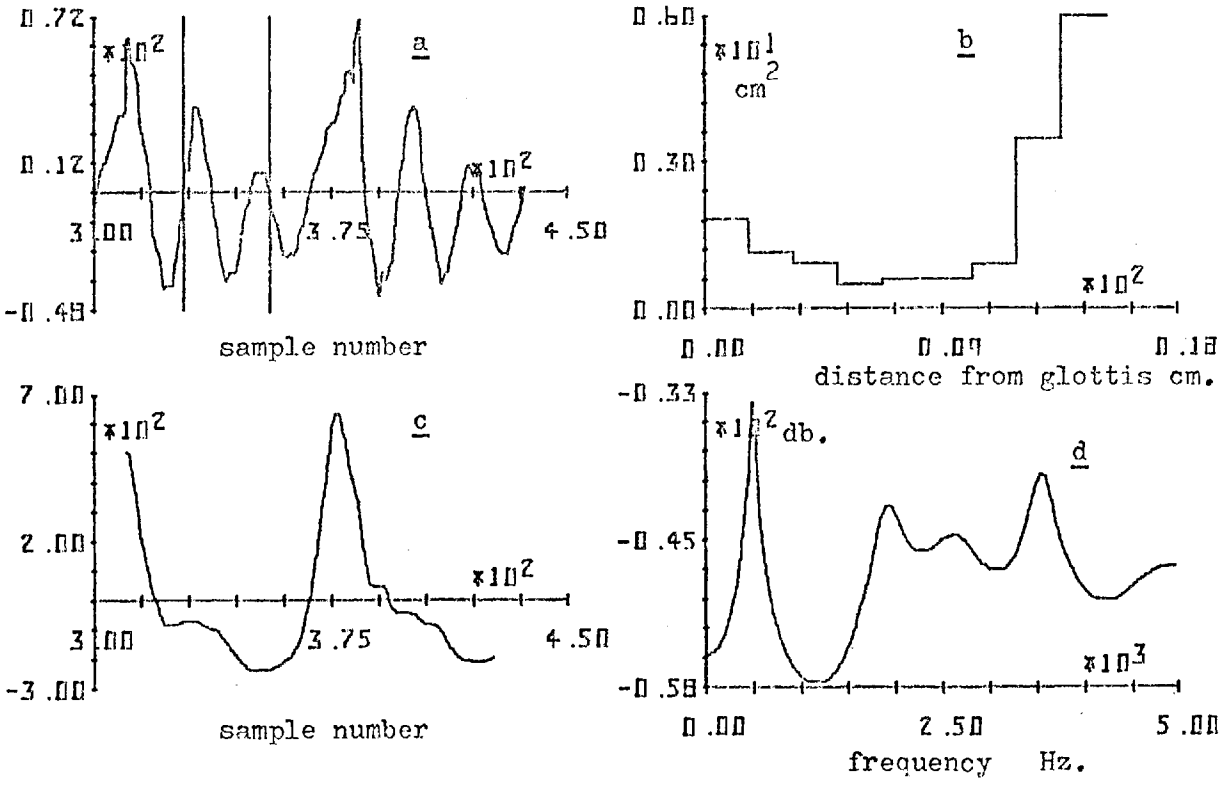
FIGURE 5.6 ANALYSIS FOR VOWEL /a/ SPEAKER F2



a Speech pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

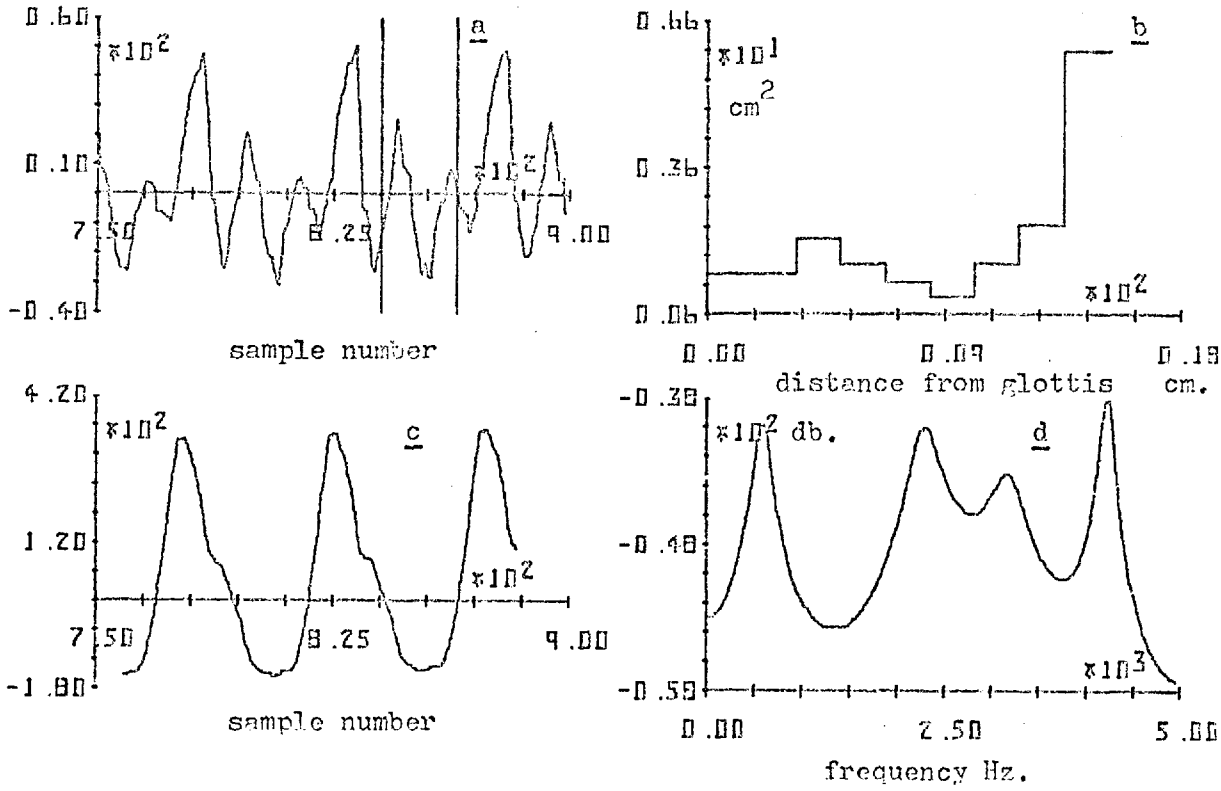
FIGURE 5.7 ANALYSIS FOR VOWEL /e/ SPEAKER M 1



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

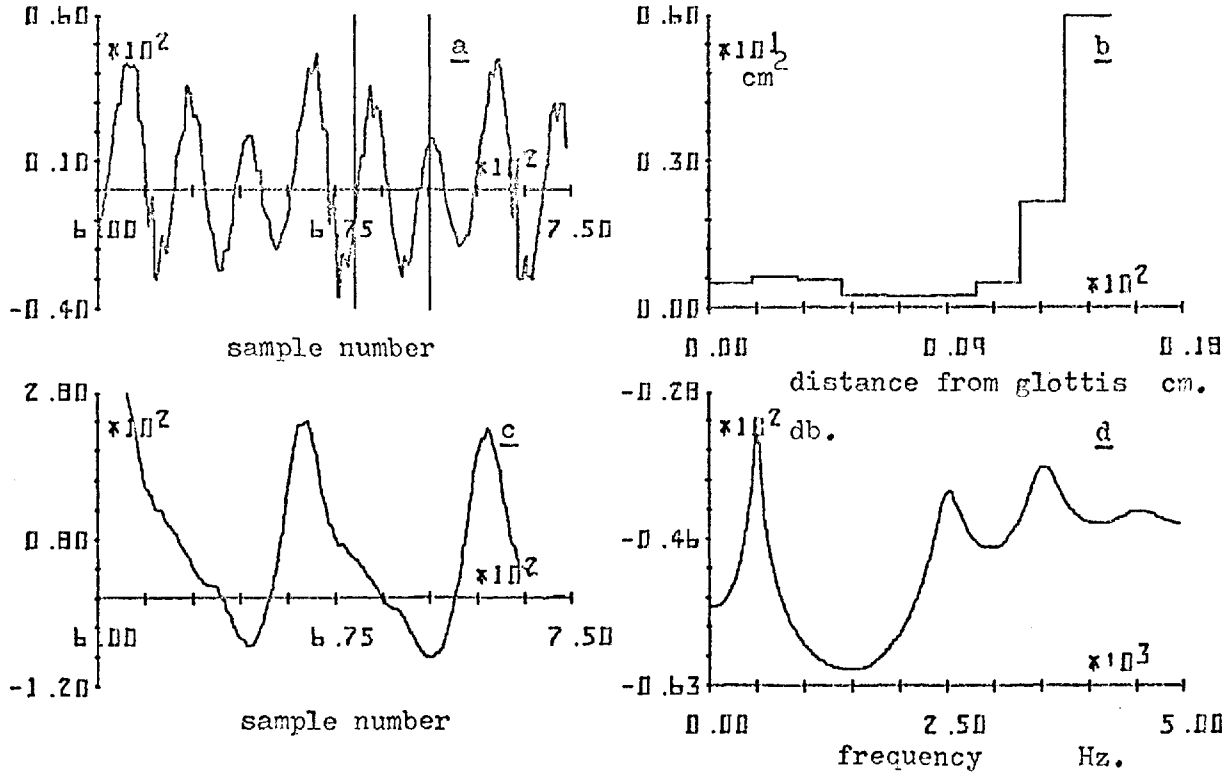
FIGURE 5.8 ANALYSIS FOR VOWEL /e/ SPEAKER M 2



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

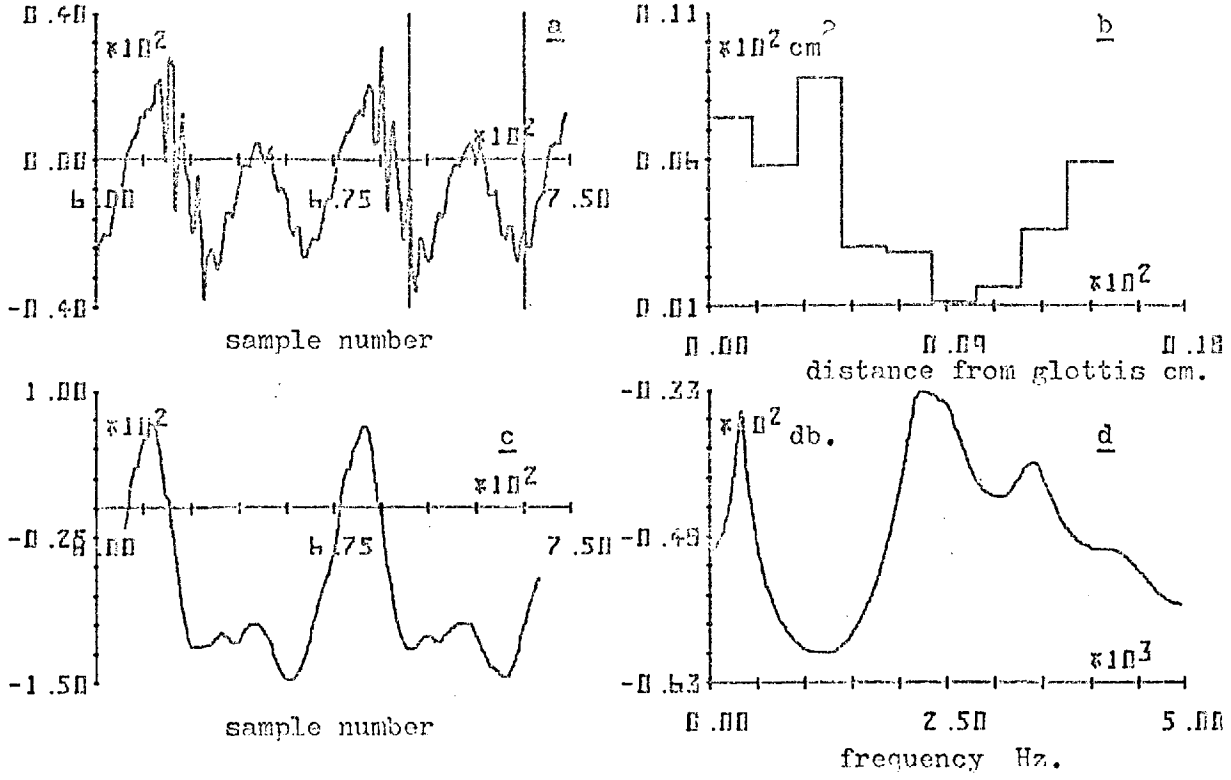
FIGURE 5.9 ANALYSIS FOR VOWEL /e/ SPEAKER F 1



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

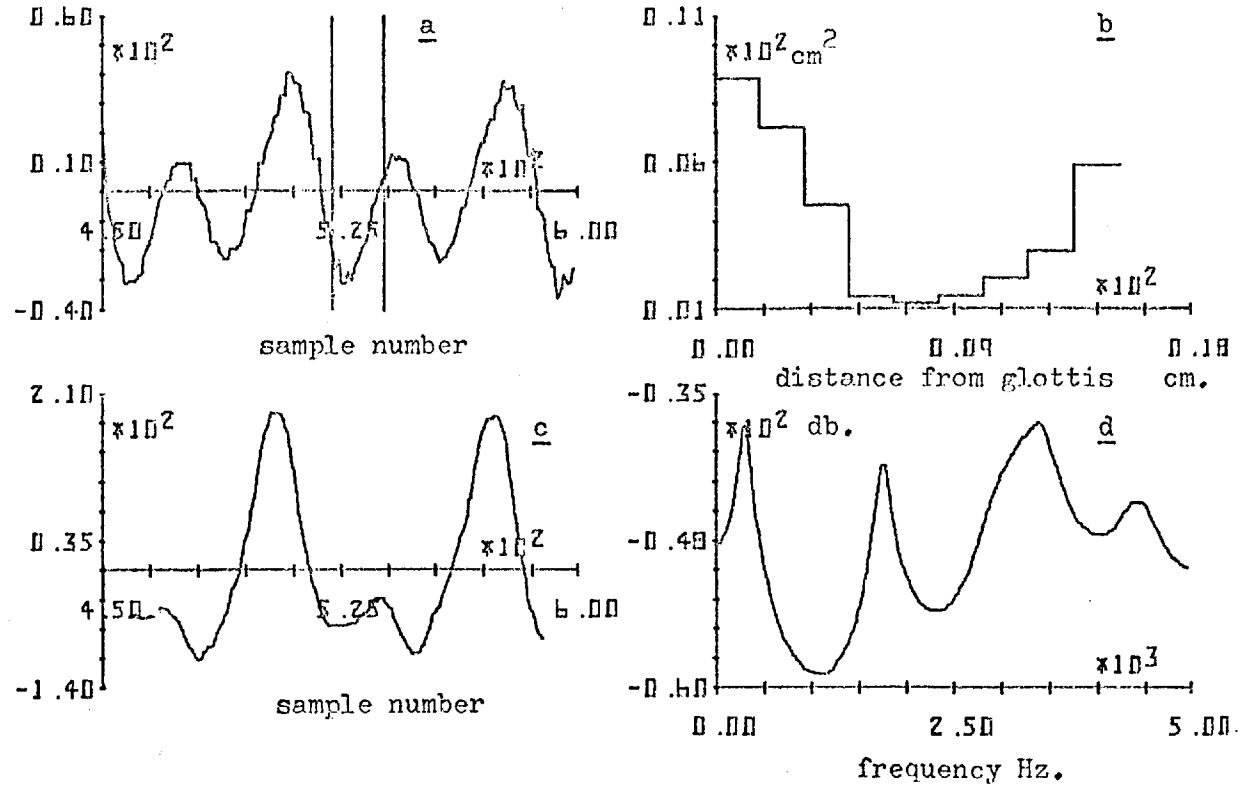
FIGURE 5.10 ANALYSIS FOR VOWEL /e/ SPEAKER F 2



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

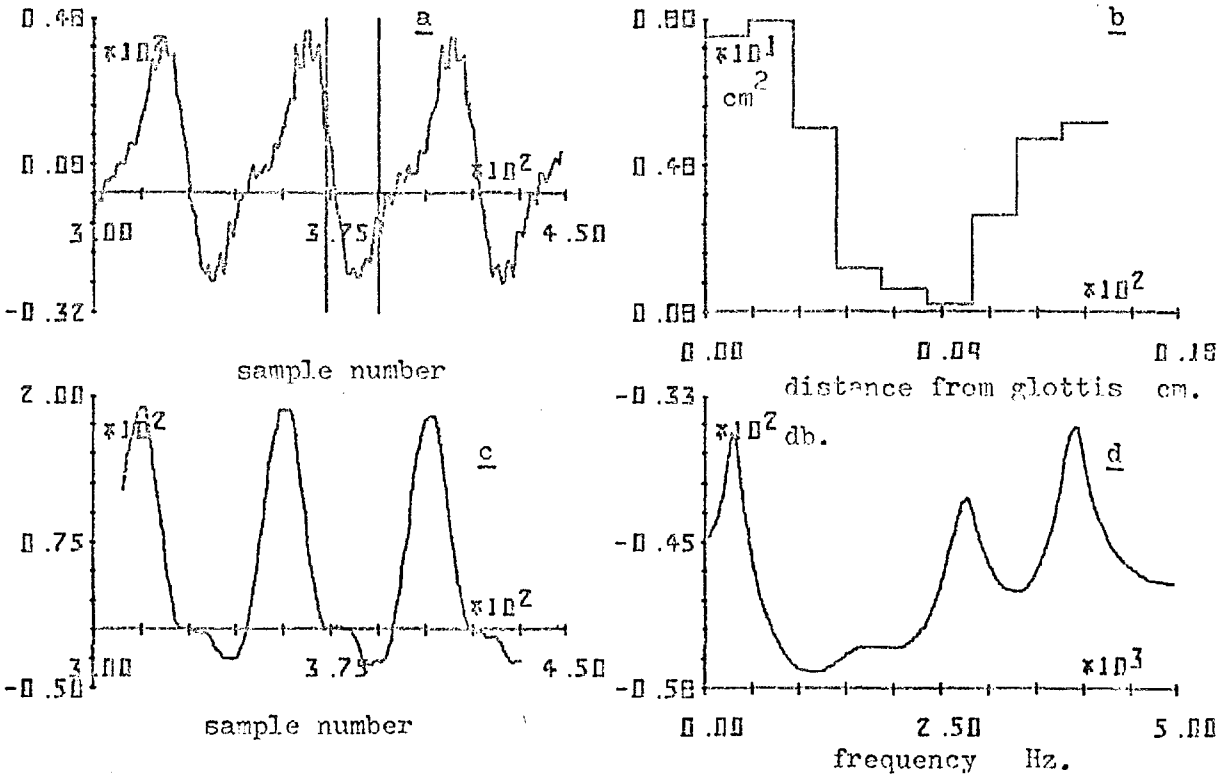
FIGURE 5.11 ANALYSIS FOR VOWEL /i/ SPEAKER M 1



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

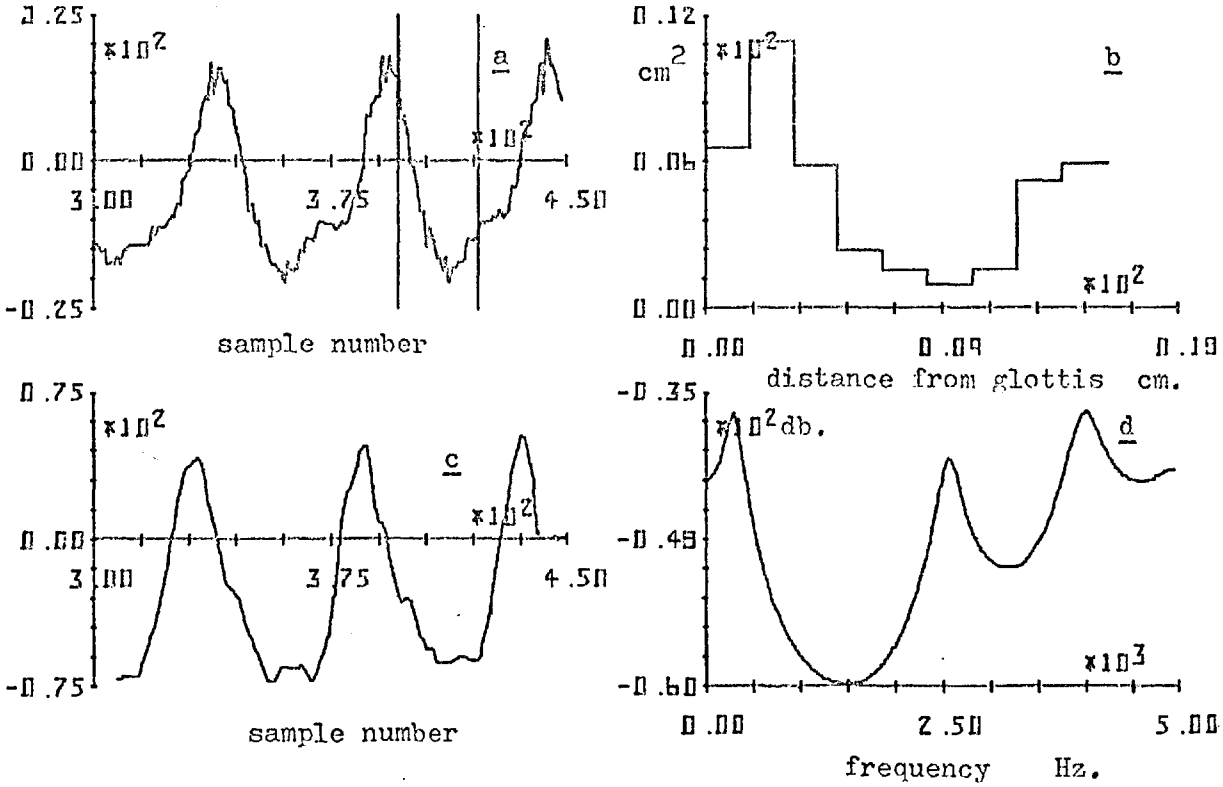
FIGURE 5.12 ANALYSIS FOR VOWEL /i/ SPEAKER M 2



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

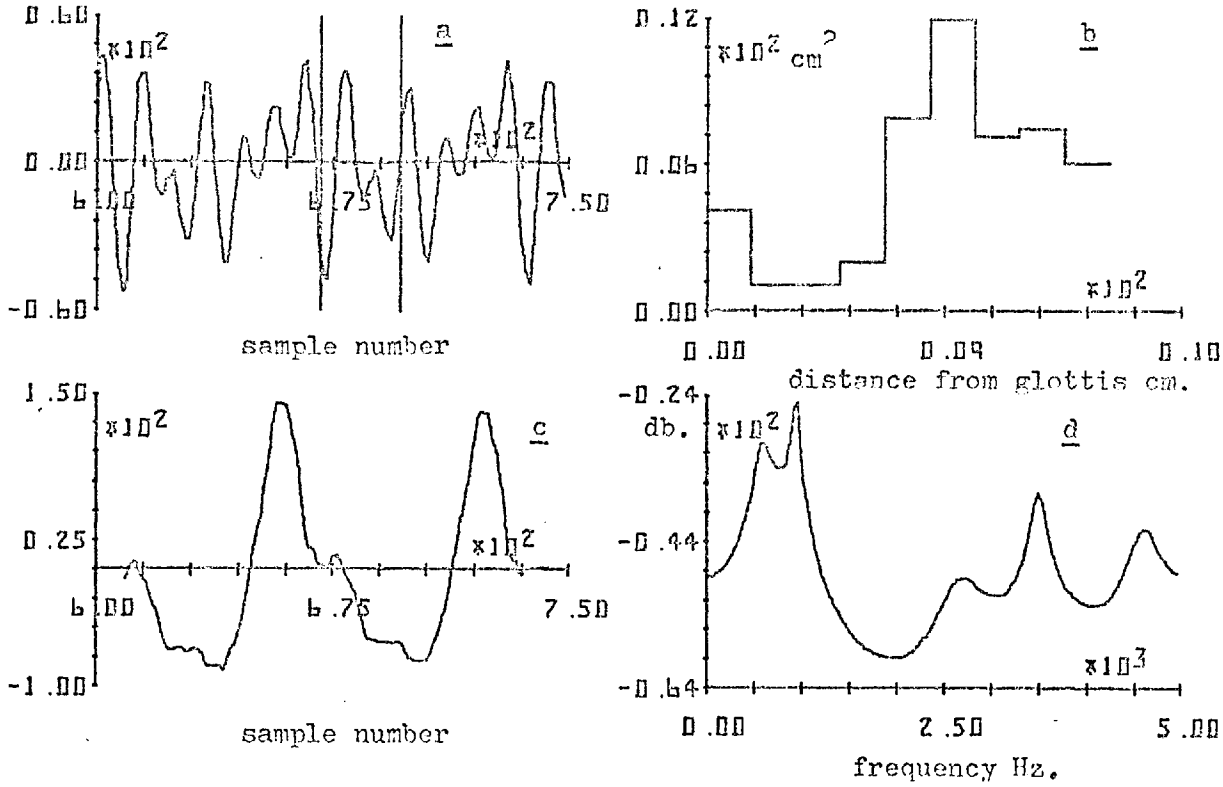
FIGURE 5.13 ANALYSIS FOR VOWEL /i/ SPEAKER F 1



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

FIGURE 5.14 ANALYSIS FOR VOWEL /i/ SPEAKER F 2



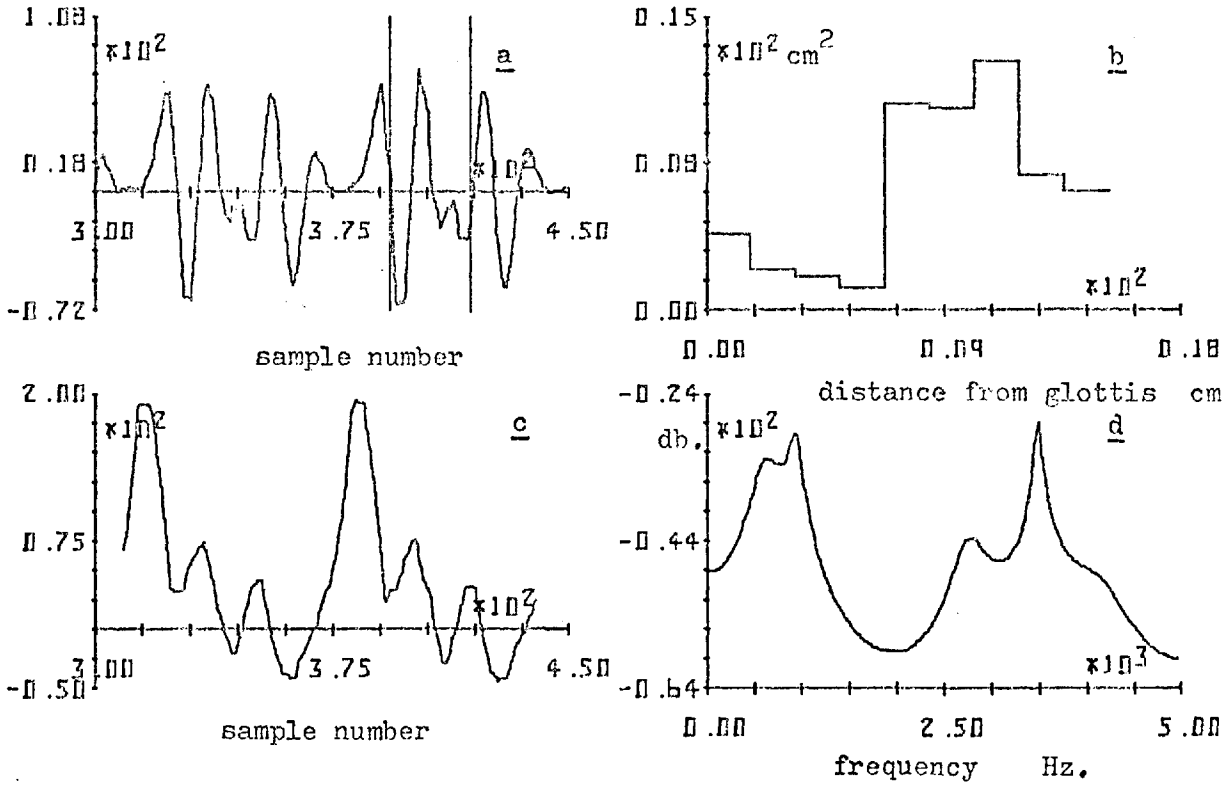
a Speech Pressure Waveform

b Normalised Area Function

c Error Signal

d Reciprocal of Inverse Filter Spectrum

FIGURE 5.15 ANALYSIS FOR VOWEL /O/ SPEAKER M 1



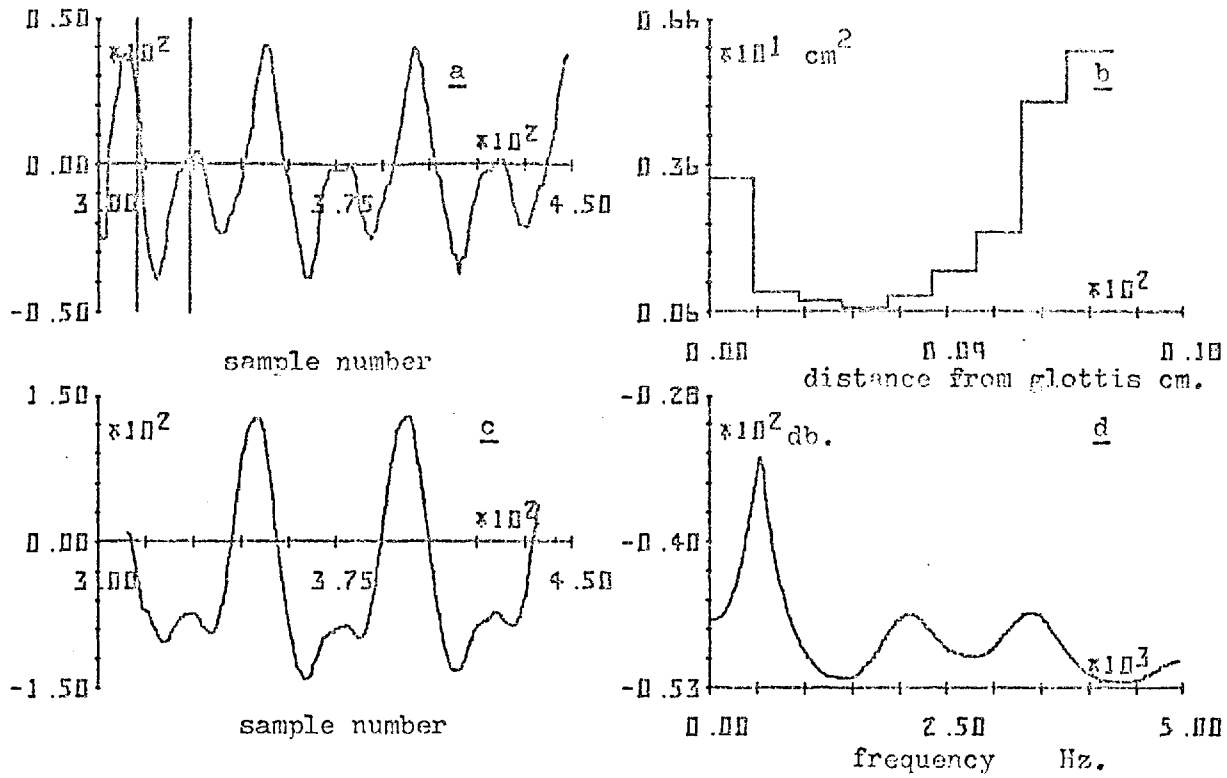
a Speech Pressure Waveform

b Normalised Area Function

c Error Signal

d Reciprocal of Inverse Filter Spectrum

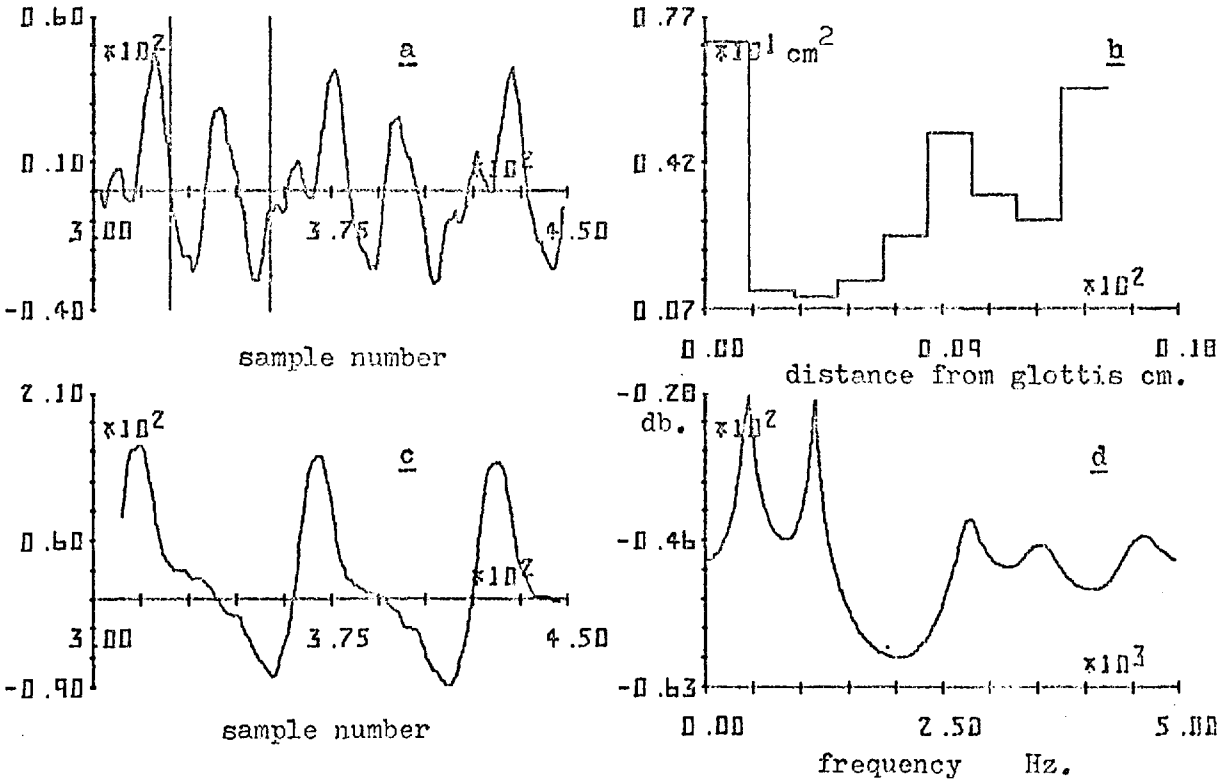
FIGURE 5.16 ANALYSIS FOR VOWEL /O/ SPEAKER M 2



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

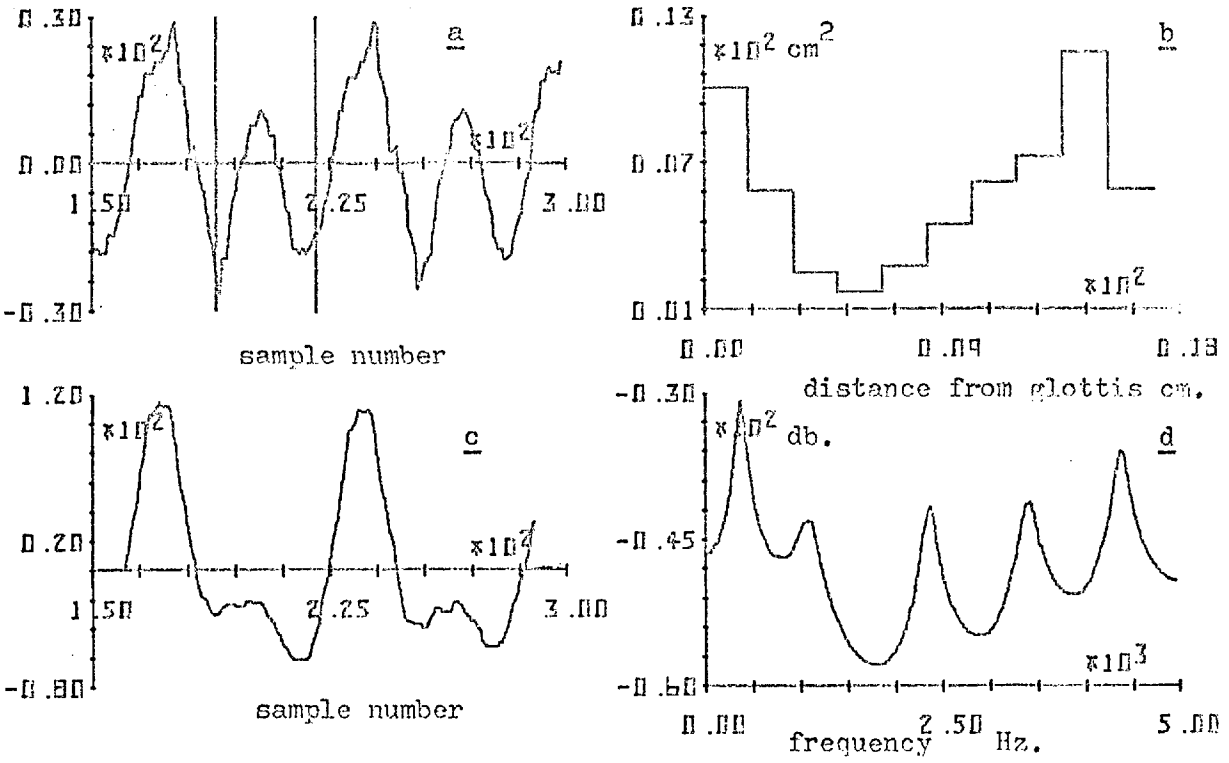
FIGURE 5.17 ANALYSIS FOR VOWEL /O/ SPEAKER F 1



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

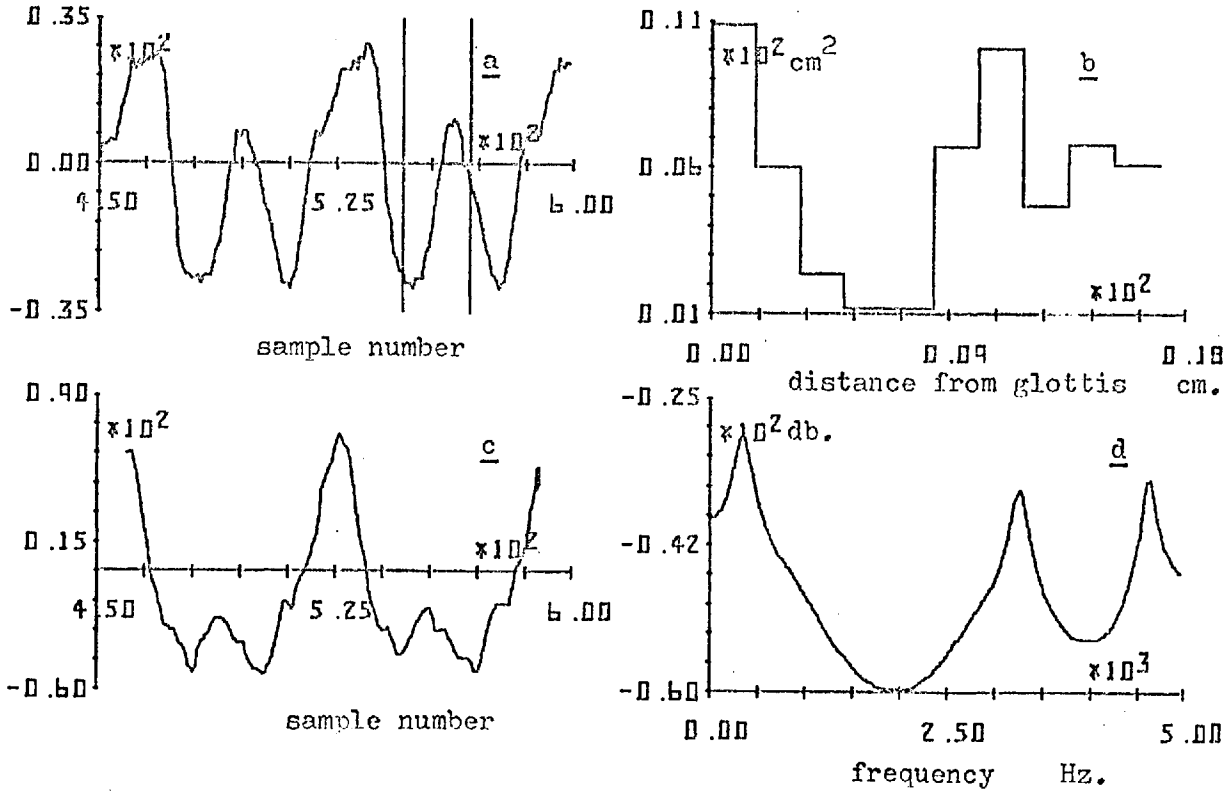
FIGURE 5.18 ANALYSIS FOR VOWEL /O/ SPEAKER F 2



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

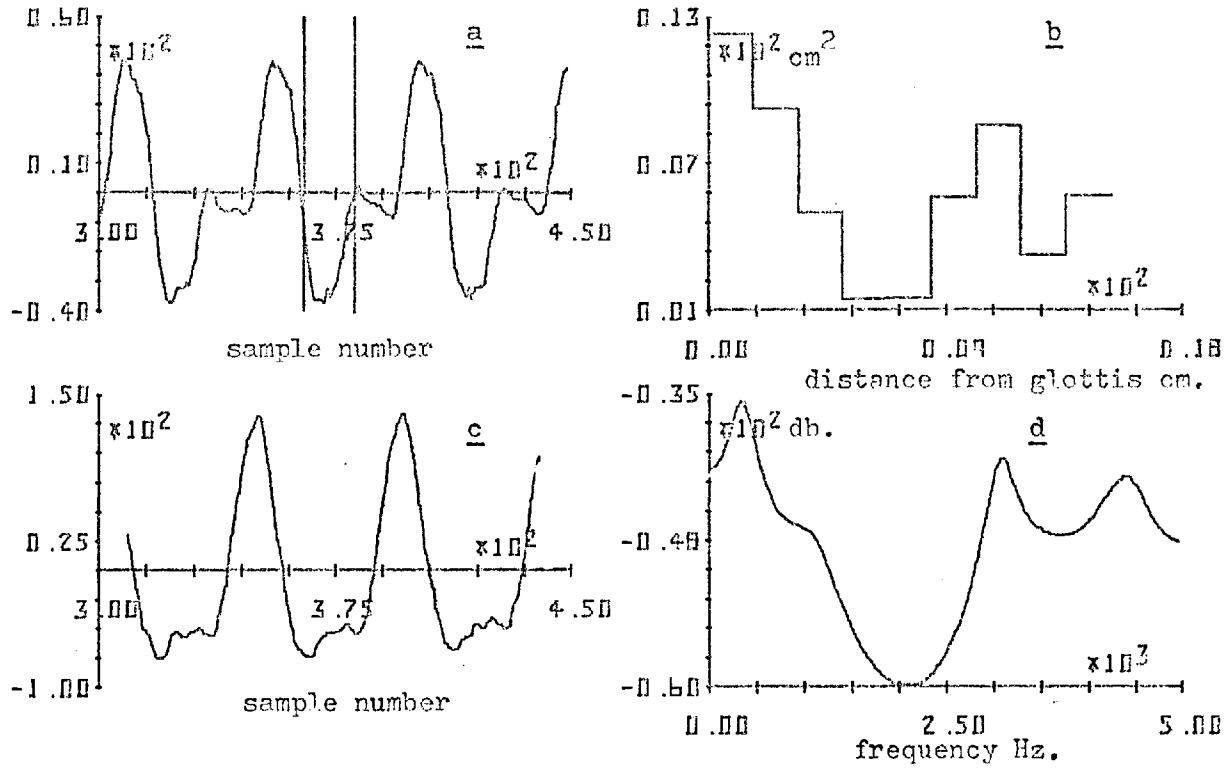
FIGURE 5.19 ANALYSIS FOR VOWEL /u/ SPEAKER M 1



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

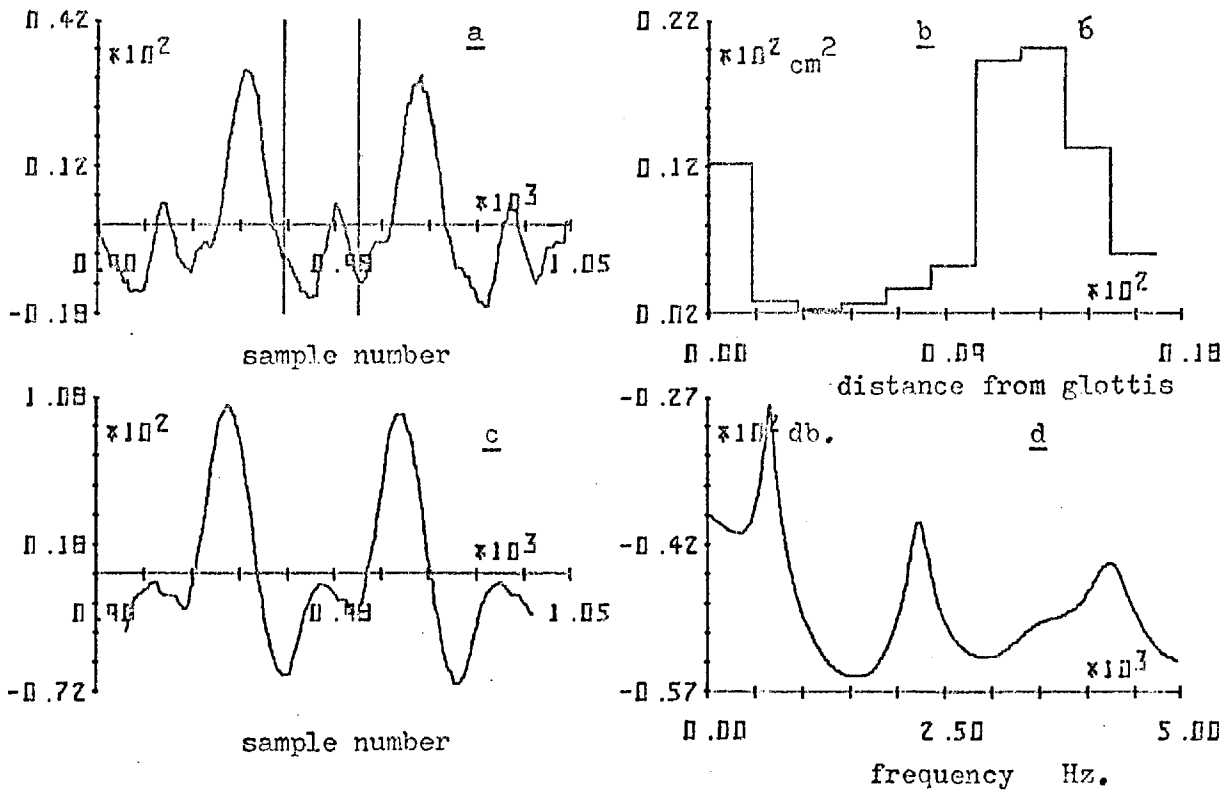
FIGURE 5.20 ANALYSIS FOR VOWEL /u/ SPEAKER M 2



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter Spectrum

FIGURE 5.21 ANALYSIS FOR VOWEL /u/ SPEAKER F 1



a Speech Pressure Waveform b Normalised Area Function

c Error Signal d Reciprocal of Inverse Filter spectrum

FIGURE 5.22 ANALYSIS FOR VOWEL /u/ SPEAKER F 2

The "clean speech spectrum" (i.e. the spectrum with the source contributions removed was calculated and is given as graph d). This "clean spectrum" (reciprocal of inverse filter spectrum) was calculated using the analysis interval shown on graph a, but ten inverse filter coefficients were used in each case. Householder transformation was again used to calculate the inverse filter coefficients and Markels method (section 2.3) was used to calculate the "clean spectrum" from these coefficients. Ten coefficients were chosen because often five formants are present in the region 0 - 5 KHz of the spectrum.

Only broad conclusions will be drawn about the results for two reasons. Firstly, they represent only a small sample and hence should not be used to draw sweeping conclusions. Secondly, it is felt that additional refinements to the analysis technique are needed, (e.g., some method of accurately estimating the lip radiation impedance) before more analyses are made. However, some conclusions are possible from these results:-

i) The area functions derived are realistic and approximate those measured during X-ray studies (e.g. FANT 1970).

ii) In all cases (except figure 5.17b) the area functions calculated for a given vowel, spoken by different speakers are similar. From figure 5.17d we can see that only three formants were found in the clean spectra, even though ten inverse filter coefficients were used. Also comparison of figures 5.17a and 5.24a shows that two distinctly different types of speech waveform were present during this particular utterance. In the analysis presented as figure 5.24 twelve inverse filter coefficients were calculated. The area function derived in this case corresponds more closely to the area functions derived for the same vowel for the other speakers. (figures 5.15b, 5.16b and 5.18b).

iii) The area function calculated for the same phoneme but different speakers, are more similar than the spectra calculated for the same phoneme and different speakers.

iv) The error signals calculated show the general character that would be expected of the glottal excitation function.

v) In each case the error signal could be used to determine the closed glottis region of the speech waveform.

An alternative method of comparing the articulatory data is given as figure 5.23. In this figure the position of maximum constriction has been plotted against the area of this constriction for each utterance. (These values were taken from figures 5.3 → 5.22). One graph is plotted

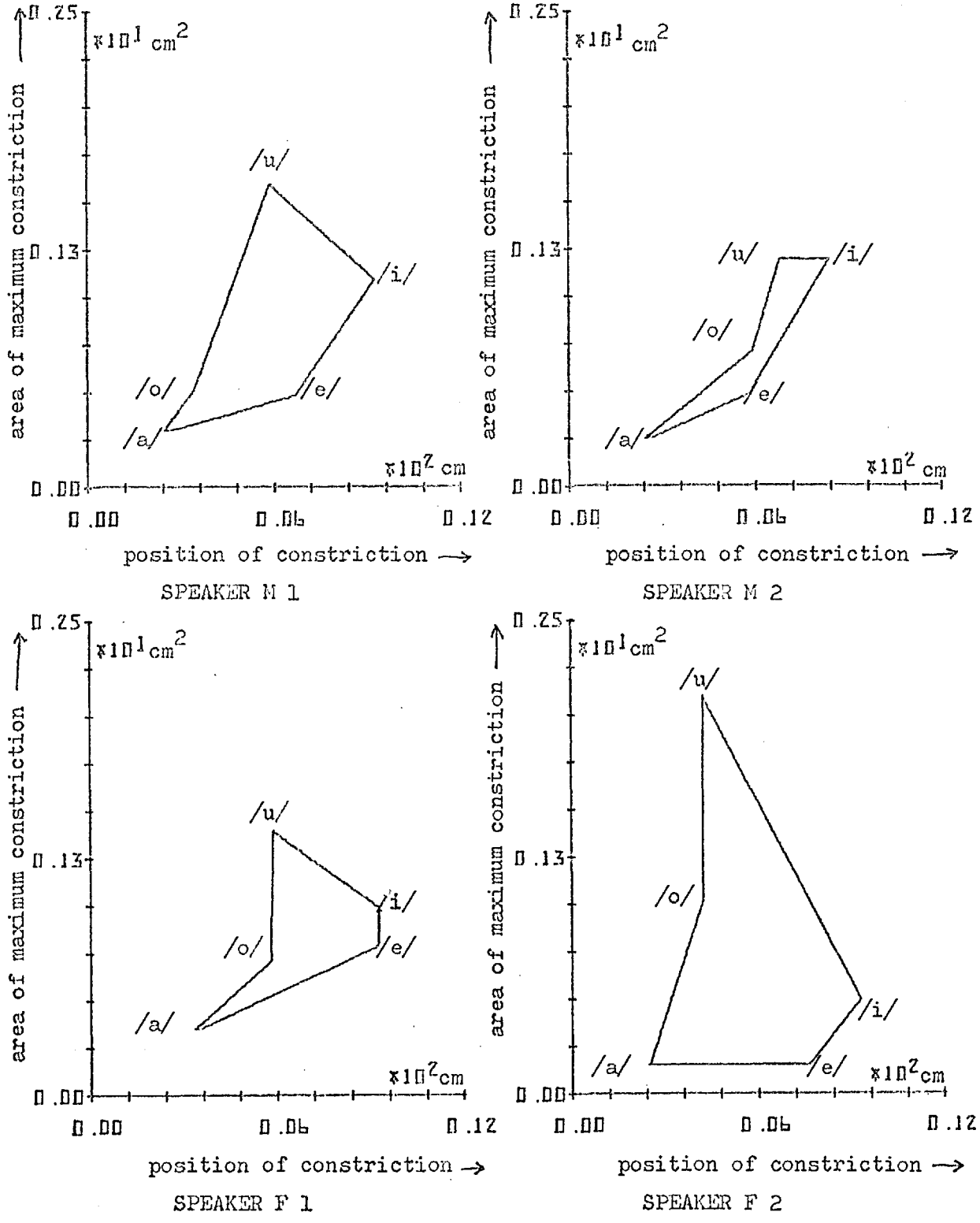
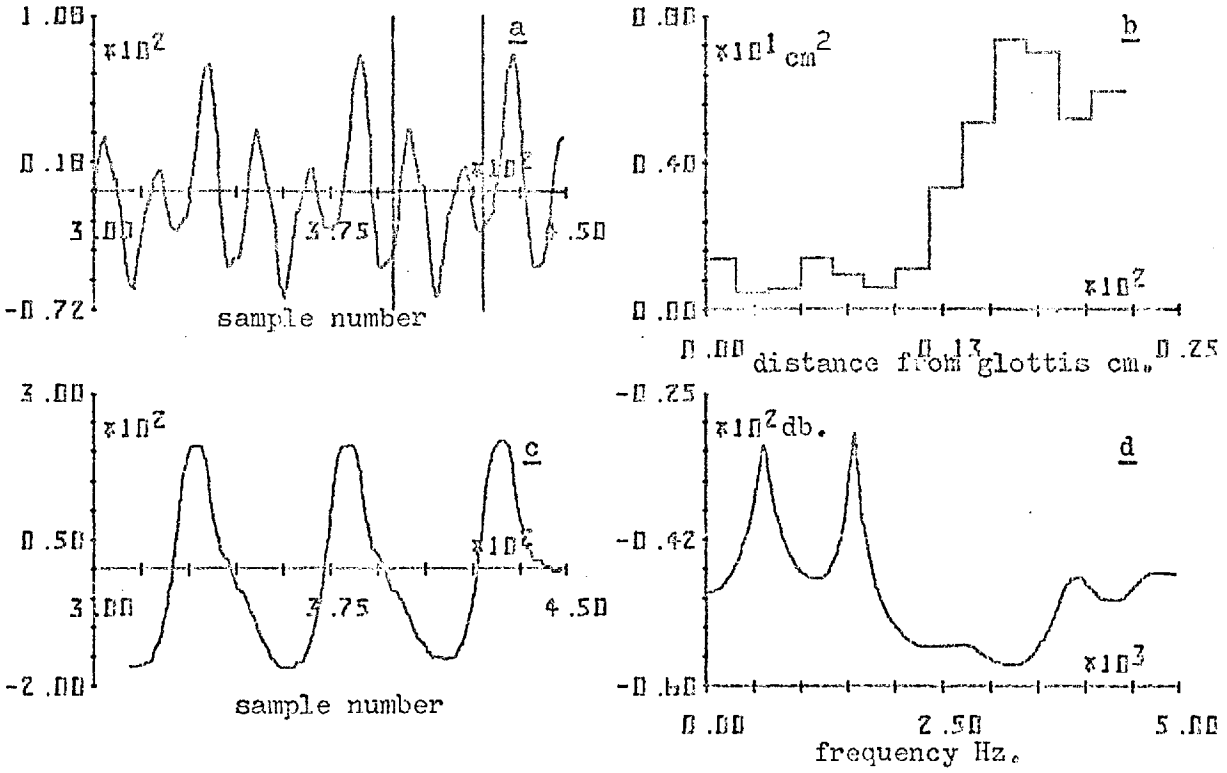


FIGURE 5.23 COMPARISON OF 'VOWEL QUADRILATERALS' CALCULATED FROM AREA FUNCTIONS FOR THE FOUR SPEAKERS.



a Speech Pressure Waveform

b Normalised Area Function

c Error Signal

d Reciprocal of Inverse Filter Spectrum.

FIGURE 5.24. ANALYSIS FOR VOWEL /o/ SPEAKER F 1 (12 inverse filter coefficients).

for each speaker and the five points corresponding to the five phonemes have been joined by straight lines. The resulting shape is analogous to the vowel quadrilateral commonly plotted. However in the vowel quadrilateral, regions are usually given for each phoneme instead of a single point, and the degree of constriction is usually plotted instead of the area of the constriction. The similarity of figures 5.23a, b, c and d suggest that a phoneme recogniser based on articulatory parameters is feasible; how successful such a scheme would be can only be predicted with the aid of more results.

The results in this section do show that realistic area functions can be determined by my method. However, the accuracy of the calculated area functions can only be determined if the area function of the vocal tract is known. Probably, the only independent method of determining the area function would be to use the X-ray techniques described in sections 1.4 and 1.5. A criterion does exist for determining the accuracy of the inverse filter coefficients, namely the similarity between the error signal and the known glottal pulse characteristics. Unfortunately in this study, the lip radiation impedance could not be approximated well enough to allow a meaningful comparison, because the lip area was not known. The necessity to know the lip area can be avoided if the volume velocity waveform at the lips is used, instead of the speech pressure waveform, to calculate the inverse filter coefficients. A technique for measuring the volume velocity at the lips has been reported recently, by ROTHENBERG (1973).

The conclusions which can be drawn from the results presented in this section are:-

a) The method produces realistic area functions and warrants further investigation.

b) Some experiments aimed at measuring the accuracy of the derived area functions should be carried out.

5.5. DEVELOPMENT OF A FAST ALGORITHM

An algorithm capable of calculating twenty five area functions per second has been developed and will be described briefly in this section. The algorithm was written in machine code and implemented on a PDP 15 computer. Unfortunately no results are available from this algorithm at the time of writing, because of computer faults. (For the past three months, an interrupt on the computer has destroyed the contents of the main machine register, the accumulator).

A laryngograph was employed to determine the closed glottis period of the speech waveform. The laryngograph signal was passed through an auxilliary piece of hardware for this purpose. This hardware first differentiated the laryngograph signal and then used a level detector to find the instant of glottal closure. A monostable was then used to introduce the delay equivalent to the passage of sound from the larynx to the microphone. The output of this monostable was used with a signal from the computer (saying it was ready to receive data) to gate the sampling clock into the computer. Use of this facility ensured that the computer only received data corresponding to samples of the closed glottis period of the speech waveform. The computer cancelled the ready signal when it had sufficient samples for analysis.

The speech waveform was sampled to 10 bit accuracy, then stored in the array and vector needed for Householder transformation. Householder transformation was carried out using 18 bit scaled fixed point arithmetic, but 36 bit products were stored when the next operation was a division or a square root. The inverse filter coefficients calculated by back substitution were stored in exponent and mantissa form because the area function mapping employed floating point arithmetic. The resulting area function was converted to integer form and used to plot a cross section through the head (see figure 5.25).

The results, from the limited usage of the facilities possible before computer failure, were promising and the picture rate 25 frame/second was fast enough , if not too fast.

During the development of the program, area functions were calculated simultaneously using a fortran program and the machine code program described here. Visual comparison of the results showed no discrepancy between the fortran and machine code programs after several hundred analyses.

It is therefore concluded that the analysis method, derived in this thesis, is capable of near real time working on a moderate sized digital computer. (Real time analysis requires 1 analysis/pitch period, approximately 100 analyses/second for male speech). It is hoped that this facility might be useful to deaf people, because it would allow comparison of their own articulatory configurations with the desired configurations. Extensive testing of the method is needed before any conclusions about its usefulness can be drawn.

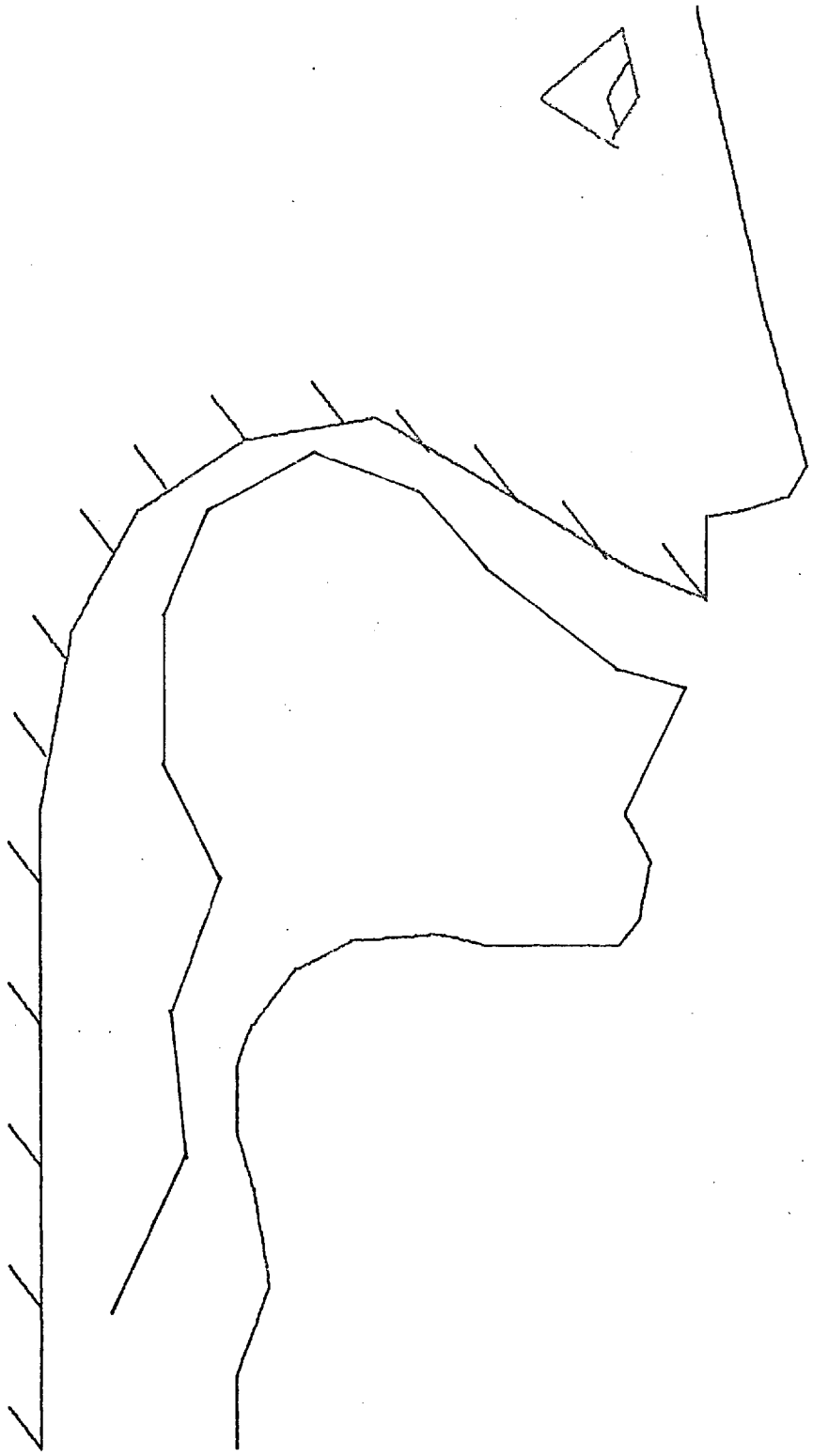


FIGURE 5.25 AN EXAMPLE OF THE DISPLAY RESULTING FROM THE FAST ALGORITHM, VOWEL /i/ SPEAKER M 1.

CHAPTER SIX

CONCLUSIONS AND SUGGESTIONS FOR FURTHER WORK

The aim of the research reported in this thesis is to develop a method for calculating the vocal tract area function from the speech waveform. A technique has been developed for calculating a piecewise approximation to the vocal tract area function, from samples of the speech waveform of voiced non-nasalised sounds.

This technique, which is derived in chapters three and four of this thesis, first uses an inverse filter to calculate the vocal tract transfer function from the speech waveform. The coefficients of this inverse filter are calculated by minimising the energy of the inverse filter output over the region of the speech waveform corresponding to a closed glottis. Householder transformation is employed to calculate these inverse filter coefficients, which are also the coefficients of a polynomial expansion in the z domain, of an all pole transfer function. For voiced non-nasalised speech sounds, the vocal tract transfer function is all pole and providing the correct number of inverse filter coefficients are calculated they constitute a good approximation to the vocal tract transfer function in the z-plane. In section 4.5 synthetic speech has been used to test the inverse filtering process. It has been found that the technique successfully separates the vocal tract transfer function from the glottal excitation function, providing a closed glottis analysis region is used to calculate the correct number of filter coefficients.

The vocal tract area function is then calculated from these inverse filter coefficients. This is achieved by modelling the vocal tract as an acoustic transmission line, and calculating the reflection coefficients of the model by a process which effectively successively removes sections of the model, replacing them each time with a matched termination. The normalised area function is simply recursively calculated from these reflection coefficients, having assumed a value for the lip area. Area functions calculated by this technique from synthetic speech are presented in section 4.5. These area functions are judged to be realistic approximations to the results expected from X-ray studies.

In chapter four it has also been proposed that the error signal resulting from use of an arbitrary analysis interval, can be used to determine the instant of glottal closure. Furthermore, a correction factor has been derived in chapter five, which approximates for the effect of losses, when calculating the area function. Utilising these ideas, the method has been applied to the analysis of five vowels, each

vowel being spoken by four people. The area functions derived for the same vowel spoken by different people are similar, and are reasonable approximations to the results expected from X-ray studies.

Results show that reasonable approximations to the vocal tract area function can be calculated from samples of the speech waveform, using the method derived in this thesis. The accuracy of the method cannot be evaluated unless an independent method is used to measure the area function. It is therefore suggested that the following work could be carried out to estimate the accuracy of the method:-

a) A technique should be developed for measuring the lip area during speech. If the lip area is known a correction for the lip radiation effects can be included in the analysis. Also the lip area would provide the correct normalisation factor for the area functions.

b) Alternatively, the volume velocity measured at the lips could be used as data instead of the speech pressure waveform currently used. (A method for measuring this volume velocity has been reported by ROTHENBERG (1973)). This would avoid the necessity to correct for radiation effects.

When the lip radiation effects have been corrected for, the error signal (which should then approximate the glottal excitation function) could be used to determine the accuracy of the inverse filtering process. To determine the accuracy of the area functions calculated, X-ray techniques could be used. The area functions measured by X-rays could then be compared with those calculated from the speech waveform.

Once the accuracy of the method has been verified a number of applications exist for it:-

i) It could be used to provide a visual display of articulator positions during speech, which might prove useful in training deaf people to talk.

ii) The articulatory information available, which would include information about vocal fold activity as well as the area function, should prove useful for speech analysts. It is possible that the error signal might have medical applications in diagnosing laryngeal disorders.

iii) The data which could be accrued from use of the method, could be used to evaluate the possibility of using an articulatory description as the first step in a speech recognition strategy.

iv) Use of the algorithm on continuous speech might enable estimation of target positions of non-voiced phonemes.

Also research on removal of some of the limitations of the method would

be very useful. An investigation aimed at including the nasal path in the articulatory mapping would immensely increase the potential of the method. Further work is also needed on the approximations of losses present in the real vocal tract, and on determination of the effect of the radiation load at the lips. If it is possible, the method should be modified to enable analysis of sounds when the source is forward in the tract, (i.e. fricatives and stops) instead of at one end of the tract as is the case for vowels.

REFERENCES

ATAL, B.S. and SCHROEDER, M.R. 1968
Predictive Coding of Speech Signals, paper C-5-4,
Proc. 6th. International Congress on Acoustics, Tokyo.

ATAL, B.S. and HANAUER, S.L. 1971
Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, pages 637-655,
J.A.S.A., Vol. 50, No. 2, Pt. 2.

BELL, A.G. 1907
The Mechanism of Speech,
Volta Bureau, Washington D.C.

BLACKMAN, R.B. and TUKEY, J.W. 1958
The Measurement of Power Spectra,
Dover Publications, New York.

BORG, G. 1946
pages 1 - 96,
Acta Math., No. 78.

CHIBA, T. and KAJIYAMA, M. 1941
The Vowel, its Nature and Structure,
Tokyo-Kaiseikan Publishing Co. Ltd., Tokyo.

COKER, C.H. 1967
Synthesis by Rule from Articulatory Parameters,
Proc. Conf. on Speech Communication and Processing,
Joint I.E.E.E. A.F.C.R.L. Conference, Cambridge, Massachusetts.

CRYSTAL, et al. 1965
A Model of Larynx Activity During Phonation, pages 212-219
M.I.T., Q.P.R., No. 78.

DUNN, H.K. 1950
The Calculation of Vowel Resonances and an Electrical Vocal Tract,
pages 740 - 753,
J.A.S.A., Vol. 22.

FANT, G. 1970
Acoustic Theory of Speech Production,
Mouton and Co., The Hague (First Edition, 1960).

FANT, G., ISHIZAKA, K., LINDQVIST, J. and SUNDBERG, J. 1972
Subglottal Formants, pages 1 - 12,
S.T.L. Q.P.R., KTH Stockholm.

FLANAGAN, J.L. and LANDGRAF, L. 1968
Self Oscillating Source for Vocal Tract Synthesisers, pages 57-64,
I.E.E.E., AU-16, No.1.

FLANAGAN, J.L. 1969
Use of an Interactive Laboratory Computer to Study an Acoustic Oscillator Model of the Vocal Cords, pages 2 - 6,
I.E.E.E., AU-17.

FLANAGAN, J.L. 1972
Speech Analysis Synthesis and Perception,
Springer Verlag, Berlin (First Edition, 1964).

FOURCIN, A.J. and ABBERTON, E. 1971
First Applications of a New Laryngograph, pages 172-182,
Medical and Biological Illustration, Vol. 21, No. 3.

FUJIMURA, O., ISHIDA, H. and KIRITANI, S. 1968
Computer Controlled Dynamic Cineradiography, pages 6-10,
Annual Bulletin (Research Inst. of Logopedics and Phoniatrics),
No. 2, University of Tokyo.

GERSHO, A. and KINARIWALA, B.K. 1968
A Synthesis Algorithm for Cascaded Transmission Lines, pages 889-897,
Sixth Allerton Conference on Circuits and Systems Theory, Illinois.

GOLDEN, R.M. 1968
Digital Filter Synthesis by Sampled Data Transformation,
pages 321-329,
I.E.E.E., AU-16.

GOLUB, G.H. 1965
Numerical Methods for Solving the Linear Least Squares Problem,
pages 206-216,
Numerische Mathematik, No. 7.

GOPINATH, B. and SONDHI, M.M. 1970
Determination of the Shape of the Human Vocal Tract from Acoustical Measurements, pages 1195-1214,
Bell Systems Technical Journal, No. 49.

HEINZ, J.M. 1967
Perturbation Functions for Determination of Vocal Tract Area Functions from Vocal Tract Eigen Values, page 1,
S.T.L., Q.P.R., No. 2, KTH Stockholm.

HOLMES, J.N. 1962
An Investigation of the Volume Velocity Waveform at the Larynx During Speech by Means of an Inverse Filter, paper G.13,
4th. International Congress on Acoustics, Copenhagen.

HOLMES, J.N. 1970
The Advantages and Disadvantages of the Matched z-Transform for Digital Filtering.
Symposium on Digital Filtering, Imperial College, London.

HOLMES, J.N. and THORNBER, E.M. 1973
Formant Frequency Measurement by Waveform Matching During Closed Glottis Periods, paper 73SHBJ,
British Acoustical Society Spring Meeting, Chelsea College, London.

ITAKURA, F. and SAITO, S. 1966
The Theoretical Consideration of Statistical Optimum Methods For Speech Spectral Density, [In Japanese],
Electrical Communication Laboratories, NTT, Tokyo, Japan,
Report 3107.

ITAKURA, F. and SAITO, S. 1968
Analysis Synthesis Telephony Based on the Maximum Likelihood Method, paper C-55,
Proceedings 6th. International Congress on Acoustics, Liege.

JOHNSON, W.C. 1950
Transmission Lines and Networks,
McGraw-Hill, London.

KELLY, J.L. and LOCHBAUM, C. 1962
Speech Synthesis.
Proceedings Stockholm Speech Communication Seminar,
KTH, Stockholm.

KINARIWALA, B.K. 1966
Theory of Cascaded Structures : Lossless Transmission Lines, pages 631-649,
Bell Systems Technical Journal.

LINDBLOM, B. and SUNDBERG, B. 1971
Acoustic Consequences of Lip, Tongue, Jaw and Larynx Movement,
Papers from the Institute of Linguistics, University of Stockholm.

MAKHOUL, J. 1973
Spectral Analysis of Speech by Linear Prediction, pages 140-148,
I.E.E.E., AU-21, No. 3.

MAKSYM, J.N. 1973
Real Time Pitch Extraction by Adaptive Prediction of the Speech Waveform, pages 149-154.
I.E.E.E., AU-21, No. 3.

MARKEL, J.D. 1971
Formant Trajectory Estimation from a Linear Least Squares Inverse Filter Formulation, Report No. 7,
Speech Communication Research Laboratories,
University of Santa-Barbara, California.

MARKEL, J.D. 1972a
Digital Inverse Filtering : A New Tool for Formant Trajectory Estimation, pages 129-137,
I.E.E.E., AU-20, No.2.

MARKEL, J.D. 1972b
The SIFT Algorithm for Fundamental Frequency Estimation, pages 367-377,
I.E.E.E., AU-20, No. 5.

MARKEL, J.D. and GRAY, A.H. 1973
On Autocorrelation Equations as Applied to Speech Analysis, pages 67-79,
I.E.E.E., AU-21, No. 2.

MATTHEWS, M.V., MILLER, J.E. and DAVID, E.E. 1961
Pitch Synchronous Analysis of Voiced Sounds, pages 179-186,
J.A.S.A., Vol. 33, No. 2.

MERMELSTEIN, P. and SCHROEDER, M.R. 1965
Determination of Smooth Cross-Sectional Area Functions of the Vocal Tract From Formant Frequencies, paper A24,
Proceedings 5th. International Congress on Acoustics, Liege.

MERMELSTEIN, P. 1967
Determination of the Vocal Tract Shape From Measured Formant Frequencies, pages 1283-1294.
J.A.S.A., Vol. 41, No. 5.

MERMELSTEIN, P. and WEST, J.E. 1968
The Impedance Tube Method for Acoustic Measurement of Area Functions of Non Uniform Tubes,
Bell Laboratories Internal Memorandum.

MERMELSTEIN, P. 1972
Speech Synthesis with the Aid of a Recursive Filter Approximating the Transfer Function of the Nasalized Vocal Tract, paper D7
Joint I.E.E.E., A.F.C.R.L. International Conference on Speech Communication and Processing, Newton, Massachusetts.

MORRISS, L.R. and PAILLET, J.R. 1972
Real Time Software Speech Synthesis, paper D9,

Joint I.E.E.E., A.F.C.R.L. International Conference on Speech Communication and Processing. Newton, Massachusetts.

PAIGE, A. and ZUE, V.W. 1969a
Computation of Vocal Tract Area Functions, pages 7-18, I.E.E.E., AU-18, No. 1.

PAIGE, A. and ZUE, V.W. 1969b
Calculation of Vocal Tract Length, pages 268-270, I.E.E.E., AU-18, No. 3.

POLLARD, B. and HÁLA, B. 1926
Les Radiographies de l'articulation des sons Technéques, Prague.

RICHARDS, P.I. 1948
Resistor Transmission Line Circuits, pages 217-220, Proceedings I.R.E., Vol. 36.

ROBINSON, E.A. 1967
Statistical Communication and Detection with Special Reference to Digital Data Processing of Radar and Seismic Signals, Griffin, London.

ROTHENBERG, M. 1973
A New Inverse Filtering Technique for Deriving the Glottal Air Flow Waveform During Voicing, pages 1632-1645, J.A.S.A., Vol. 53, No. 6.

RUSSEL, G.O. 1929
The Mechanism of Speech, pages 83-109, J.A.S.A., Vol. 1.

SCHROEDER, M.R. 1967
Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements, pages 1002-1010, J.A.S.A., Vol. 41, No. 4.

SHIH, S.T. 1965
Synthesis of Optical Filters by Transmission Line Analogues, M.S. Thesis, Massachusetts Institute of Technology.

STANSFIELD, E.V. 1971
An Articulatory Model for Speech Recognition, Ph.D. Thesis, University of London.

STANSFIELD, E.V. and BOGNER, R.E. 1973
Determination of Vocal Tract Area Functions from Transfer Impedance, pages 153-158, Proceedings I.E.E., Vol. 120, No. 2.

STEVENS, K. and HOUSE, A. 1955
Development of a Quantitative Description of Vowel Articulation, pages 484-493,
J.A.S.A., Vol. 27.

VAN DEN BERG, J.W., SANTEMA, J.T. and DOORNENBAL, P. 1957
On the Air Resistance and the Bernouilli Effect of the Human Larynx, pages 626-631,
J.A.S.A., Vol. 29.

WAKITA, H. 1972
Estimation of the Vocal Tract Shape by Optimal Inverse Filtering and Acoustic Articulatory Conversion Methods.
Report No. 9, Speech Communication Research Laboratories, University of Santa-Barbara, California.

WAKITA, H. 1973
Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms, pages 417-427,
I.E.E.E., AU-21, No. 5.

WEINSTEIN, C.J. and OPPENHEIM, A.V. 1971
Predictive Coding in a Homomorphic Vocoder, pages 243-249,
I.E.E.E., AU-19, No. 3.

WILKINSON, J.H. 1960
Householders Method for Solution of the Algebraic Eigen Problem, pages 23-27.
Computer Journal, Vol. 3.

Abbreviations

AFCRL	=	Air Force Cambridge Research Laboratories.
IEE	=	Institute of Electrical Engineers (England).
IEEE	=	Institute of Electrical and Electronic Engineers (America)
IRE	=	Institute of Radio Engineers.
IEEE AU	=	Journal of Audio and Electro-Acoustics Institute of Electrical and Electronic Engineers.
JASA	=	Journal of the Acoustic Society of America.
MIT QPR	=	Massachusetts Institute of Technology, Quarterly Progress Report.
STL QPR	=	Speech Transmission Laboratory Quarterly Progress Report. Royal Institute of Technology, Stockholm.