A SYSTEMATIC ASSESSMENT OF THE ACCURACY OF VOCAL TRACT AREA FUNCTION ESTIMATES MADE FROM THE SPEECH WAVEFORM

Thesis by

Paul H. Milenkovic

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1981

(Submitted June 30, 1980)

To my brother Victor,
who is next in line

ACKNOWLEDGEMENTS

Foremost among the people guiding me in this undertaking has been John Pierce, my advisor at Caltech. Like those chess masters capable of playing simultaneous games, Dr. Pierce has an uncanny talent for suggesting the correct move after obtaining only a glance at the situation. Another person with a major role has been Edward Posner, also of Caltech. Dr. Posner has helped me with some of the mathematical derivations and has had the patience to sit through my seminar presentations of this material.

Bishnu Atal, Peter Denes, and Max Mathews of Bell Telephone Laboratories were instrumental in getting this project off the ground by having me work under them in the summer of 1979. Dr. Atal saved me at least a year's work by getting me to zero in on the important research issues in vocal tract area function estimation.

Special thanks go to Walden Barcus and Mark Dolson, graduate students at Caltech, for providing moral support at crucial moments. Also to be mentioned are Veljko Milenkovic and Harry Johnson of Ford Motor Company Manufacturing Research. Dr. Milenkovic raised important questions on the use of LPC analysis on periodic signals, and it was Mr. Johnson that insisted that I not work for Ford but go to graduate school instead.

ABSTRACT

By performing Linear Predictive Coding (LPC) analysis on the speech waveform, it is possible to determine the cross sectional areas, or area function, of a discrete section acoustic tube model of the vocal tract. It is a matter of controversy, however, as to whether the areas of the acoustic tube model accurately estimate the areas of the actual vocal tract. There are several sources of error which cause the estimated areas to differ from the true areas. A procedure for estimating the spectrum of the vocal tract response in terms of LPC derived formant frequencies and bandwidths is discussed; the areas of the acoustic tube model can be calculated from these frequency and bandwidth values. The accuracy with which formant frequencies and bandwidths can be estimated is evaluated by experiments where the frequency and bandwidth of a one resonator vocal tract model are estimated. The accuracy of the complete procedure for estimating the area function from speech is evaluated by experiments where the area function is estimated from synthetic speech sounds. These speech sounds are synthesized from known vocal tract shapes against which the estimated area function can be compared.

Table of Contents

I. INTRODUCTION

Calculating the Vocal Tract Area Function From Speech Sounds

The shape of the vocal tract has an important influence on the perceptual qualities of speech sounds. The vocal tract consists of the cavity that connects the vocal cords at the glottis to the lips. This cavity includes what is commonly regarded as parts of the windpipe, the throat, and the mouth. A side branch connects the vocal tract to the nose. Except for nasalized portions of speech, this side branch is closed and can be disregarded. During voiced speech, air forced past the glottis by lung pressure causes the vocal cords to vibrate. This vibration is carried acoustically to the lips, where sound radiates to the surrounding air. The resonances of the vocal tract together with the transfer function of the radiation response shape the spectrum of the vocal cord vibration to give the spectrum of the speech waveform. The changes in the positions of the tongue, lips, and jaw that occur during articulation of speech produce changes in the shape of the vocal tract. Changing the shape of the vocal tract produces changes in the resonance frequencies and dampings of the vocal tract, which in turn produce changes in the envelope of the speech spectrum.

Given the shape of the vocal tract, it is possible to make acoustical

calculations to obtain the vocal tract transfer function [Refs. 1,2,3,4]. Combining this transfer function with assumptions about the spectrum of the excitation and the transfer function of the radiation characteristic relating sound vibration at the lips to sound in the far field, one can synthesize speech sounds. The question arises whether one can reverse this process - estimate the speech spectrum from the speech wave, remove the influences of radiation response and excitation spectrum, and calculate the shape of the vocal tract. The shape would be expressed as the vocal tract area function, the cross sectional area as a function of position along the length of the vocal tract. The process of estimating the area function from speech can be regarded as "inverting the vocal tract." So the question remains, can one "invert the vocal tract."

There are several motivations for wanting to calculate the vocal tract area function from speech. First of all, being able to do it would be useful in studying speech. By calculating the area function for successions of speech sounds, one would be able to observe first hand the dynamics of the articulation of these speech sounds. One would avoid the medical hazards of techniques such as x-ray photography [Ref. 3] or the use of cumbersome procedures such as requiring subjects to phonate with their lips placed against impedance tubes [Refs. 5,6]. Secondly, the area function may be an efficient code for speech recognition, speech synthesis, and speech compression applications. Because the area function is directly connected to the production of speech sounds, it can characterize speech sounds. The area function also has direct physical meaning in terms of the structure of the actual speech

production mechanism. Physiological constraints limit the rates of change and degrees of freedom of the position of structures such as the tongue and the lips. The area function could potentially be described by few parameters which only need to be updated at a slow rate as a speech sound is articulated.

One of the most promising approaches to estimating the area function from speech is the LPC (Linear Predictive Coding) derived acoustic tube model [Refs. 4,7,8,9,10]. Prior to performing LPC analysis, the speech signal is low pass filtered and sampled. LPC analysis operates on the resulting time series to estimate the speech spectrum as the magnitude squared of the frequency response of a discrete time filter [Refs. 4,7,9,11,12,13]. If proper corrections have been made for the radiation characteristic and the excitation spectrum, the coefficients of this discrete time filter, the LPC coefficients, can specify the cross sectional areas of an acoustic tube model of the vocal tract [Appendix A]. The cross sectional areas of the model are a parameter set that is quite robust in both speech compression [Ref. 14] and speech synthesis [Ref. 15] applications. It is a matter of controversy, however, as to whether the acoustic tube areas accurately estimate the actual vocal tract areas [Refs. 16,17]. For nasalized sounds and for consonant sounds involving fricative excitation in the middle of the vocal tract, the acoustic tube model no longer approximates the acoustic structure of speech production; one would naturally expect the area function estimates that result to be physically meaningless. But for vowel sounds, where the acoustic tube model is physically meaningful, there remain sources of error which can cause the area function estimate to

differ significantly from the actual vocal tract area function.

Sources of Error in Area Function Estimation

There are two major sources of error in estimating the vocal tract area function as the cross sectional areas of the LPC derived acoustic tube model. The first source of error is the difficulty of obtaining LPC coefficients that accurately estimate the spectrum of the vocal tract. By doing LPC analysis and applying corrections, one wants to obtain a set of coefficients that describe a discrete time filter where the magnitude squared of its frequency response accurately represents the magnitude squared of the frequency response, or spectrum, of the vocal tract. The second source of error is in the acoustic tube model which is obtained from the LPC coefficients. Many simplifying assumptions are made in formulating the acoustic tube in such a way as to permit calculating its cross sectional areas directly from the LPC coefficients. These assumptions make the acoustics of the vocal tract model markedly different from the acoustics of the actual vocal tract.

The error in estimating the vocal tract spectrum breaks down into three major components. The first component is the problem of the undersampling of the vocal tract spectrum [Appendix C]. During voiced speech, the glottal excitation consists of a nearly periodic series of pulses. The periodicity of the voicing source gives speech a line spectrum - the vocal tract spectrum is then known only at multiples of the pitch frequency of the excitation. The higher the fundamental frequency, or

pitch, of a speech sound, the more widely spaced are the spectral lines. The second component is the problem of correcting the speech spectrum for the effects of the radiation characteristic and the glottal spectrum to obtain the vocal tract spectrum. The radiation characteristic can be reliably determined by acoustical calculations [Refs. 1,3,4], but considerable variation can occur in the shape (hence also the spectral envelope) of the glottal pulses [Ref. 18]. The third component is the problem of lack of information about the high frequencies of the vocal tract spectrum. The speech spectrum contains reliable information about the vocal tract spectrum only out to around 3.5 to 4 KHz [Refs. 17,19]. At higher frequencies the area function ceases to be sufficient to specify vocal tract acoustics as wave propagation in the vocal tract is no longer a one dimensional problem along the length axis of the vocal tract - the assumption of plane wave propagation breaks down.

Applying simple solutions to these three problem areas can result in errors in estimating the vocal tract spectrum. LPC models the speech spectrum as the spectrum of the discrete time filter of the impulse response. The spectrum, the magnitude squared of frequency response, of this filter can be evaluated at any frequency from the LPC coefficients, giving a smooth, continuous representation of the speech spectrum envelope [Ref. 7]. The fact that LPC analysis was performed on a periodic signal, however, introduces significant systematic error [Appendix C]. As for correcting the speech spectrum to obtain the vocal tract spectrum, the glottal spectrum can be regarded as asymptotically attenuating the speech spectrum 12 dB/octave and the radiation

characteristic as boosting the speech spectrum 6 dB/octave. Prior to performing LPC analysis, the speech signal can be preemphasized by differencing neighboring samples of the sampled speech signal; the resulting 6 dB/octave boost corrects for the net 6 dB/octave rolloff of glottal spectrum and radiation characteristic [Ref. 10]. The trouble is that the asymptotic spectral envelope of the glottal spectrum can differ considerably from a 12 dB/octave attenuation - it could be as much as 18 [Ref. 18]. On the matter of the lack of high frequency information, one can low pass filter the speech signal at 4 KHz and sample it at 8 KHz prior to performing LPC analysis. The acoustic tube model is comprised of discrete sections, the length of each section being a little over 2 cm for this sampling rate. In addition to having the area function specified at this course interval along the length of the tube, the length of the acoustic tube model comes only in multiples of the section length. The spectrum of the acoustic tube model, however, will not accurately represent the vocal tract spectrum if the length of the acoustic tube model differs significantly from the length of the vocal tract [Appendix B].

The second major error source concerns the simplifying assumptions made in the acoustic tube model of the vocal tract. There are two formulations of the acoustic tube model that permit direct calculation of the area function from the LPC coefficients [Appendix A]. Both formulations have hard, lossless walls. The vocal tract, however, has yielding walls, and there are several loss mechanisms associated with the walls. It is possible to apply a correction to the vocal tract spectrum estimate to convert it to that spectrum that would have resulted

if the walls were hard and lossless [Ref. 20]. This correction is based on the assumption that the wall loss is scaled by vocal tract area so that attenuation and dispersion of acoustic waves do not vary with vocal tract area. The difficulty is that the amount of loss cannot be measured from the vocal tract spectrum. Where the amount of loss one should correct for is not reliably known, one may just as well not correct for wall losses at all.

The Wakita formulation of the acoustic tube has the mouth, or lip, end of the tube terminated in a short. All loss is lumped into the source resistance of the volume velocity source exciting the tube at the glottal end. The Atal-Hanauer formulation has the lips end terminated in a a resistive load, where all the losses are lumped. The tube is excited with a perfect volume velocity source, having an infinite acoustic resistance, at the glottal end of the tube. Neither formulation realistically represents the vocal tract, having loss at the glottis which some represent as being resistive [Ref. 1] and having a radiation load that has inductive and frequency dependent resistive components. To complicate matters, the area function obtained from the Wakita formulation is that of the Atal-Hanauer formulation where one has taken the reciprocals of all the area values and turned the tube around. The two formulations give the same area function for antisymmetric vocal tract shapes; they give radically different answers for symmetric shapes [Ref. 16].

Thesis Overview

In light of all of the error sources in area function estimation, there remain those who claim that the LPC method can still give acceptable area function estimates under certain conditions [Refs. 8,10,21]. It is the purpose of this study to examine these claims. The effects of the error sources mentioned are examined individually and in combination. First of all, one can determine whether or not reliable area function estimates can be obtained from speech sounds. Secondly, if the area function estimates are not reliable, one can see what causes them to fail so that remedies can be suggested.

Section II of this report contains a review of the major work that has been done within the past twenty years on the problem of estimating the area function from acoustic information. Particular emphasis is given to the LPC derived acoustic tube estimates made from speech sounds, but alternate approaches are also reviewed. These alternate approaches can give insight into the nature of problems in calculating areas from speech. Section III of this report reviews the initial experimental work done during this project. These experiments expose reasons for doing things certain ways that are not fully explained or perhaps fully understood by authors of existing papers on this subject. The outcomes of these experiments were a major influence on the design of the area function estimation procedure that was used to give the results presented in this report.

Section IV describes a set of experiments performed to evaluate the accuracy of LPC analysis in estimating the frequencies and bandwidths of speech formants. The frequencies and dampings of the poles of the

vocal tract transfer function are represented in the frequencies and bandwidths of the resonant peaks, or formants, of the speech spectrum. Accuracy in estimating formant frequencies and bandwidths is the key to determining the vocal tract spectrum from the speech spectrum.

Section V presents the results of a set of experiments to estimate the area function from synthesized speech sounds. Speech sounds are synthesized from known vocal tract areas. Different error sources can be selectively incorporated into the sound synthesis. The effects of these error sources can be evaluated by comparing the estimated area function with the known vocal tract shape. Though no results for estimating area functions from real speech are presented in this report, results on the synthetic speech are presented for the combination of all the known error sources, indicating the kinds of errors one can expect.

The appendices contain detailed discussions of mathematical and physical issues involved in the area function estimation problem. Appendix A provides a review of the two acoustic tube formulations which permit one to compute the area function directly from the LPC coefficients. Appendix B explains the need for matching the length of the acoustic tube model to the length of the vocal tract if the acoustic tube model spectrum is to accurately estimate the vocal tract spectrum. The choice of sampling rate used in the LPC analysis of the speech signal is involved because the sampling rate determines the length of sections in the LPC derived acoustic tube model. Appendix C provides a discussion of the systematic error which occurs in applying LPC analysis to

periodic waveforms such as voiced speech. Possible avenues of correcting this error are presented. Appendix D describes a model of the glottal excitation of the vocal tract that was used to obtain the results on formant frequency and bandwidth estimation that are presented in Section IV. Appendix E contains a discussion of a correction for some of the error which results from the simplifying assumptions made in the acoustic tube model.

II. HISTORICAL DEVELOPMENT OF AREA FUNCTION ESTIMATION

Vocal Tract Area Function from Acoustic Data

The earliest literature on the subject of estimating the vocal tract area function addresses the problem of calculating the vocal tract areas from acoustic measurements made on the vocal tract [Refs. 5,6,19,22,23]. What is measured is the impedance of the vocal tract seen at the lips looking down into the vocal tract. Impedance can be measured by having subjects phonate specified sounds and by placing a measuring device against their lips. One can either measure impedance as a function of frequency by applying a sweep tone to the lips, or one can measure the impedance in terms of the impulse response that results from exciting the vocal tract by a spark impulse source applied to the lips.

The first developments on the subject include the work of Schroeder [5] and Mermelstein [22]. Smooth approximations to the area function are expressed as the first few terms of the Fourier series expansion of the logarithm of the area function. If the length of the vocal tract is known and the area function is expressed in terms of the first N even and odd Fourier coefficients, the vocal tract shape is constrained to 2N degrees of freedom. The absolute scaling of the area function is left unspecified - absolute scaling relates to adding a constant term to the log area function. The Fourier coefficients of the area func-

tion estimate are determined by an iterative procedure that matches the first N resonant frequencies (poles) and the first N antiresonant frequencies (zeroes) of the lip impedance calculated from the estimate to corresponding values measured from the vocal tract. Specifying these 2N quantities was found to uniquely constrain the 2N degrees of freedom of the area function estimate in all experiments run, but the uniqueness of the area function estimate was not established theoretically.

The zero frequencies of the vocal tract impedance correspond to the resonant frequencies of the vocal tract with the lips wide open, and the pole frequencies of the impedance correspond to the resonant frequencies of the vocal tract with the lips closed [Ref. 24]. The formant frequencies observed from the speech spectrum can be used to specify zero frequencies of vocal tract impedance. Vocal tract pole frequencies do not have any readily identifiable component in the speech spectrum; they have to be determined by acoustical measurement.

Additional work on area function estimation from acoustic measurements has been done by Paige and Zue [Ref. 23] and by Gopinath and Sondhi [Ref. 19]. Paige and Zue developed a procedure for estimating the vocal tract area by a discrete section acoustic tube model. The cross sectional areas of the model are calculated directly from the measured frequencies of impedance singularities - poles and zeroes. Smooth estimates of vocal tract area are obtained by specifying more tube sections than there are measured singularity values and by supplying the additional singularity values as those of a uniform acoustic tube. The frequencies of the additional singularities are determined from the

vocal tract length, which has to be independently estimated. Gopinath and Sondhi developed a procedure for calculating continuous, smooth estimates of vocal tract area. Vocal tract length and the low order impedance singularities are supplied. The high order impedance singularities out to infinite frequency are assumed to be those of a uniform acoustic tube. Gopinath and Sondhi extended their method to work directly from impulse response measured at the lips [Ref. 5]. Vocal tract length no longer needs to be independently estimated as it is contained in the impulse response measurement.

LPC Derived Area Function Estimates

The development of the technique of estimating the vocal tract area function directly from LPC coefficients opened the way for estimating area functions from speech. Atal and Hanauer [Ref. 7] describe an acoustic tube model that has hard, lossless walls, a resistive termination at the lips, and an ideal volume velocity source exciting the tube at the glottis. A means for computing the cross sectional areas of this tube directly from the LPC coefficients is presented. The LPC coefficients are computed by analysis of the speech waveform. The LPC coefficients are presented as a parameter set that can describe the speech signal, and the acoustic tube model is presented as a physical realization of the linear filter specified by the LPC coefficients. The question of whether the acoustic tube areas accurately estimate the actual vocal tract areas is not addressed. The claim is made that the acoustic tube may sound like the vocal tract but no claim is made that

it looks like the vocal tract.

Additional work on the LPC acoustic tube model was presented by Wakita [Ref. 8]. Here, an alternate model is developed, one having the lips terminated in a short circuit and having all loss lumped into the source resistance of the volume velocity source that excites the vocal tract. This model could be thought to more realistically represent the vocal tract inasmuch as the impedance of the lip termination becomes very small for low frequencies and loss at the glottal termination is resistive to a first order approximation. In addition, preemphasizing the speech signal to apply a 6 dB/octave boost prior to performing LPC analysis is suggested as a means of correcting for the effects of the radiation characteristic and the asymptotic glottal spectrum on the speech spectrum. The preemphasis and the new acoustic tube model are applied to a limited number of examples of actual speech, and area function estimates for vowel sounds are shown that look quite reasonable. They look reasonable inasmuch as they look like vocal tract shapes for vowels and one has nothing else to compare them against.

Formant Based Extension of the LPC Method

Correcting the speech spectrum for the effects of the radiation characteristic and the glottal spectrum with a fixed +6 dB/octave preemphasis is a rather crude form of correction. Suppose one used LPC analysis with enough coefficients to very accurately estimate the speech spectrum with the spectrum of the discrete time filter specified by the LPC

coefficients. As the LPC filter is all pole, the denominator of its transfer function can be expressed as a polynomial, the complex roots of which are the poles of the filter. Some of these poles will be complex conjugate pole pairs which correspond to vocal tract formants: the frequency coordinate of a pole pair corresponds to formant frequency and the damping coordinate of a pole pair corresponds to half the 3 dB bandwidth of the formant [Refs. 7,9]. By factoring the LPC polynomial, one could retain the roots corresponding to vocal tract formants and discard the rest, effectively removing the elements of the speech spectrum not belonging to the vocal tract spectrum.

The use of LPC derived formant frequencies and bandwidths to characterize the vocal tract spectrum for purposes of calculating the area function is described by Wakita [Refs. 10,25]. The complementary application of using LPC derived formant frequencies and bandwidths to remove the vocal tract characteristic from the speech waveform so that one may observe the glottal waveform is described by Wong and others [Ref. 26]. One should keep in mind that this technique involves polynomial root solving, which provides both frequency and bandwidth. Simpler procedures for obtaining formant frequencies by way of LPC involve picking the peaks of the LPC spectrum estimate [Refs. 27,28,29]; these procedures do not give formant bandwidths.

From the polynomial roots corresponding to the first four formants, one can recompute the LPC coefficients and then calculate the cross sectional areas of an eight section acoustic tube model. The length of each section, and hence the length of the eight section tube, is deter-

mined by the sampling rate inherent in the formulation of the discrete time LPC filter [Appendix B]. The accuracy of the area function estimate is degraded if the length of the acoustic tube model is not equal to the length of the vocal tract [Ref. 25]. The transformation from LPC polynomial roots to frequencies and bandwidth is dependent on the sampling rate [Appendix B]. In converting back from formant frequencies and bandwidths to polynomial roots, and hence to LPC coefficients, it is possible to substitute a new value for the sampling rate which will make the length of the acoustic tube model equal that of the vocal tract. In the case where the length of the vocal tract is not known, one can try different sampling rates and select that sampling rate that gives the smoothest area function estimate. Finding the smoothest area function is a means of estimating the vocal tract length [Ref. 25].

An additional refinement proposed for obtaining the vocal tract spectrum from the speech signal is closed glottis interval analysis. The opening and closing of the glottis during voiced speech can be observed in the LPC prediction residual of the speech signal [Refs. 30,31]. During the portion of the excitation waveform cycle that the glottis is closed, the speech waveform is the result of free decaying vocal tract oscillations, modified by the radiation characteristic with initial conditions determined by the past history of excitation. It is thought that by doing covariance method LPC over short analysis intervals within the time of glottal closure, one can avoid the effects of glottal spectrum on the LPC coefficients [Refs. 10,26]. The radiation characteristic is easily corrected for because it is close to being a fixed +6 dB/octave boost [Refs. 1,3].

The most recent method proposed by Wakita for estimating the area function [Ref. 10] involves performing closed glottis interval LPC and then determining formant frequencies and bandwidths. One needs to go to the trouble of solving for the roots of the LPC polynomial to obtain formant frequencies and bandwidths in order to scale the length of the acoustic tube model. Solving for roots, however, makes the difficulty of identifying the closed glottis interval for selecting the analysis interval redundant, as vocal tract roots can also be obtained when the LPC analysis interval contains entire pitch periods. The claim is made that the estimated formant frequencies and bandwidths are the values for the vocal tract with the glottis closed. Then the proposal is advanced that because LPC analysis does not estimate bandwidths very accurately, the bandwidths should be discarded, and predetermined bandwidths for similar speech sounds should be substituted. It is very interesting that this reason is advanced for substituting bandwidths because even if the LPC bandwidths were accurate, they would represent conditions of no glottal loss. The acoustic tube model has all loss lumped at the glottis, suggesting that the use of closed glottis bandwidths is wrong to begin with. Bandwidth substitution is not a very practical procedure because one needs to know a priori what vocal tract shape one is calculating to know what bandwidth values to substitute.

Adaptive Preemphasis Based Extension of the LPC Method

A different approach to the matter of matching the length of the

acoustic tube model to the vocal tract length and the matter of correcting for the influence of radiation characteristic and glottal spectrum on the speech spectrum is proposed by Nakajima and others [Ref. 21]. The LPC sampling rate in increased to 17.5 KHz, where the section length of the acoustic tube model is down to 1 cm. The number of acoustic tube sections is increased to allow for the maximum vocal tract length. In estimating the area function of vocal tract shapes of less than the maximum length, one expects under the Wakita formulation of the acoustic tube that the first M sections starting at the lip end would estimate the area function and the remaining section would be of uniform area, an area giving the glottal end of the tube a characteristic impedance matched to the source resistance of the glottal termination. The estimate of vocal tract length is M cm where M is the number of nonuniform acoustic tube areas. A one centimetre increment is fine enough to reduce the errors resulting from a mismatch between acoustic tube and vocal tract length.

The matter of correcting the speech spectrum is handled with an adaptive preemphasis filter. An implicit assumption is made that the vocal tract shape is sufficiently smooth that its acoustics take on characteristics of a uniform tube. A uniform tube has a spectrum that is asymptotically flat because its resonances are evenly spaced in frequency. It is assumed that the vocal tract is smooth enough that its spectrum is asymptotically flat, even though its resonances depart from uniform spacing. The speech signal is preprocessed by passing it through an adaptive filter that is constrained to flatten the speech spectrum asymptotically without removing any of the resonant peaks.

This adaptive filter serves to undo the effects on the speech spectrum of the radiation characteristic and the glottal spectrum, both effects having considerable spectral slope.

By sampling speech at 17,500 Hz, one is taking a speech bandwidth of 8750 Hz, beyond the frequency where area function is a sufficient description of vocal tract acoustics. Though the higher order formant frequencies and bandwidths from speech may not be the correct ones with which to specify the acoustic tube because of the three dimensional nature of vocal tract acoustics at the higher frequencies, they may be as good as any other values that one could replace them with. The effectiveness of the adaptive filter is another matter. The LPC coefficients form an estimate of the spectrum of the speech signal after it is run through the adaptive filter. The excess length of the acoustic tube model corresponds to excess LPC coefficients beyond those required to specify the vocal tract spectrum. These excess coefficients will pick up any artifact in the speech spectrum that the adaptive filter failed to remove. The presence of the excess length of the acoustic tube model together with artifacts in the corrected speech spectrum will cause large errors in the area function estimate [Ref. 31]. One would suspect that Nakajima's method for estimating area function, though demonstrated to be effective in his paper, may be difficult to get to work properly, especially if one is having any trouble with the adaptive filter.

Multiparameter Approaches

The LPC based methods of estimating area function make use of an acoustic tube model that has two possible formulations, neither of which matches the actual vocal tract acoustics very accurately. This problem can be avoided by going to a multiparameter approach. The area function of an acoustic tube model can be specified by a set of parameters. The spectrum of the acoustic tube model can then be calculated where the tube incorporates more realistic loss and boundary conditions. The parameters of the tube are varied by an iterative procedure which maximizes the match between the acoustic tube spectrum and the vocal tract spectrum estimated from the speech waveform. The problem of properly identifying the vocal tract spectrum from speech remains as in any other method to estimate the area function from speech. The multiparameter approach has the added problem of solving a multiparameter optimization problem. If one designates the parameter set as the areas of acoustic tube model sections, one has a rather large number of parameters to work with. One could reduce the number of parameters by going to an articulatory model. This however is not very useful because one of the purposes of estimating area functions is to obtain data from which one can develop articulatory models.

The method developed by Strube [Ref. 16] is based on an acoustic tube model similar to the one used in the LPC method for estimating area function. The acoustic tube model has hard lossless walls, and the wall losses are lumped into a resistance at the glottal termination. The lip termination, however, contains the correct radiation load. The inverse filter to this acoustic tube model is calculated. If the

acoustic tube model matches the vocal tract, putting the speech signal through this inverse filter should recover the glottal waveform modified by the radiation characteristic. The parameters of the acoustic tube are adjusted to give minimum energy in the inverse filtered speech signal during the closed glottis interval, a time when the glottal waveform should be zero.

Atal and others have developed a computer sorting technique for estimating the area function [Ref. 2]. Formant frequencies, bandwidths, and amplitudes are calculated for a vast number of different vocal tract shapes. The acoustical calculations take into account wall losses, a proper radiation load, and a resistive glottal termination. The collection of acoustic data is sorted to identify collections of vocal tract shapes which correspond to a set of formant frequencies. The philosophy taken is that formant frequency is all that one can properly estimate about the vocal tract spectrum from the speech spectrum given the many error sources involved. Formant frequencies are not changed very much either by errors in spectrum estimation or by variance in the loss and boundary conditions of the acoustic tube model from the actual vocal tract. Formant bandwidths are subject to errors from both sources [Refs. 1,33]. For a given set of formant frequencies, with formant bandwidths left unspecified, one has a multiplicity of vocal tract shapes to select from. Further refinement of this technique to develop a practical means of estimating area function from speech may impose a constraint of smooth variation of the vocal tract shape in time during the articulation of a speech sound in order to resolve this ambiguity.

The vocal tract shape estimation technique described by Ladefoged and others [Refs. 34,35] is based on estimating the vocal tract shape from a set of formant frequencies. X-ray studies were performed to determine an articulatory model with as many degrees of freedom as one has formant frequencies. Formant frequencies of speech sounds are correlated with parameter values of the articulatory model for the different x-ray pictures of the vocal tract. The shape of the vocal tract can be obtained from formant frequencies by a statistical regression technique. This method gives the shape of the vocal tract as seen in an x-ray picture - the area function has to be inferred from this representation of shape. Heavy reliance on an articulatory model does not make this method very useful.

III. INITIAL APPROACHES TAKEN

Supplying Missing High Frequency Information

The speech waveform lacks information about the area function at high frequencies. Estimating the area function from the vocal tract spectrum estimate obtained from speech is based on acoustic models where sound propagates as plane waves down the length of the vocal tract; this plane wave assumption breaks down at frequencies beyond 4 KHz [Ref. 17]. One can utilize only the low frequency portion of speech by low pass filtering the speech signal to 4 KHz, sampling at 8 KHz, and estimating the vocal tract spectrum by means of LPC analysis. The resulting acoustic tube model, however, then consists of discrete sections that are over 2 cm long. The area function estimate is only specified intervals of over 2 cm, and the length of the acoustic tube model comes only in integer multiples of this interval.

An approach to be considered is one of extending the estimated vocal tract spectrum by supplying missing information about the high frequency portion [Ref. 32]. The vocal tract spectrum out to 4 KHz can be estimated from the speech signal. From 4 KHz out to 17,500 KHz, the spectrum is specified as having a constant value. The rational for using a constant is that by specifying a quantity for which information is lacking, one wants to add as little structure, or information, to the spectrum as possible.

The results of an experiment testing this idea are shown in Figure 3.1. The vocal tract spectrum was calculated from a known vocal tract shape. The autocorrelation of vocal tract impulse response was computed by taking the inverse Fourier transform of the spectrum; the vocal tract spectrum was then estimated by computing LPC coefficients from the autocorrelation function. The estimate of the vocal tract shape is based on the Wakita formulation of the LPC derived acoustic tube. Actual area function and actual spectrum are shown in solid lines; estimated area function and LPC based spectrum estimate [Refs. 11,12] are shown in dashed lines. The plot for area function estimate was obtained by drawing straight lines between the midpoints of acoustic tube segments. The mouth, or lip, end of the area function plots is on the left. The spectrum is computed from the known shape using the same acoustic tube loss and boundary conditions used to estimate the area function.

One can see from the displays in Figure 3.1 that the high frequency portion of the vocal tract spectrum has been set to a constant value. Because the spectrum is specified out to 17.5 KHz, the area function estimate is specified every half centimetre. The number of acoustic tube sections has been overspecified beyond the length of the area function being estimated. In the case where the constant value has been specified too high, the LPC spectrum estimate does not match the actual spectrum. The resulting area function estimate contains a great deal of jaggedness. In the case where the constant value has been set to a value that allows for an accurate match by the LPC spectrum, the resulting area function estimate fits the actual area function quite

accurately, is smooth, and accurately estimates vocal tract length as the number of acoustic tube sections having detail in them.

The significant result seen here is that the accuracy of the area function estimate is related to the accuracy of the vocal tract spectrum estimate. It goes without saying that if the estimated spectrum does not match the actual spectrum, the estimated area function will not match the actual area function. This result, however, indicates that if missing information has been incorrectly specified on the target spectrum, the estimated spectrum will be unable to match the target spectrum, causing the area function estimate to be in error. What is especially significant is that under these conditions, the estimated spectrum does not accurately match the target spectrum at low frequencies as well as at high frequencies.

The primary difficulty run into with this approach is the problem of correcting for the effects of glottal spectrum and radiation characteristic on the speech spectrum. To allow for the variation in vocal tract lengths that occur between talkers and between different sounds from the same talker [Refs. 36,37], the number of sections in the acoustic tube model has to be overspecified. These additional sections correspond to additional LPC coefficients, additional poles of the LPC filter model. Any artifacts left by imperfectly correcting the speech spectrum for the effects of glottal spectrum and radiation characteristic will be incorporated into these poles in the LPC filter model representing the vocal tract. As a consequence, significant errors will be made in the area function estimate [Ref. 32].

Factoring the LPC Polynomial to Correct for Source Spectrum

The source spectrum, the spectrum of the glottal signal operated on by the radiation characteristic, is a major source of error in area function estimation. The radiation characteristic may be determined fairly reliably by acoustical calculations, but considerable variation may occur in the glottal spectrum. One means of removing the effects of the source spectrum from the speech signal is to either differentiate the speech signal, or difference the sampled speech signal [Ref. 8]. This boosts the speech signal a fixed 6 dB/octave; the correction is based on the assumption that the radiation characteristic boosts the vocal tract spectrum 6 dB/octave and the glottal spectrum attenuates it 12 dB/octave. The glottal spectrum can depart from a 12 dB/octave asymptotic slope [Ref. 18]. A fixed correction can introduce large errors in the estimate of vocal tract spectrum, and hence the area function estimate.

Figure 3.2 shows the result of an experiment to observe the effects of errors in correcting for source spectrum. Area function and vocal tract spectrum estimates are compared against known values. The mouth, or lip, end of the area function plots is on the left. The plot for area function estimate is obtained by connecting the midpoints of acoustic tube segments with straight lines. The known vocal tract spectrum is computed from a known shape using acoustic tube loss and boundary conditions matched to the estimation procedure. Vocal tract spectrum is specified over a range of a bit over 4 KHz; an acoustic

tube model with 2 cm long sections is used. A 6 dB/octave attenuation is imposed on the target vocal tract spectrum to represent the type of error that could occur if one uses a fixed source spectrum correction.

The first case shown in Figure 3.2 is based on an acoustic tube model that has enough sections to match its length to that of the target vocal tract shape. The LPC spectrum estimate fails to match the target spectrum, and the area function estimate fails completely. In the second case, the number of LPC coefficients has been overspecified. The polynomial in the denominator of the transfer function of the LPC filter was factored into roots. The roots corresponding to narrow band complex conjugate pole pairs were retained; roots corresponding to wide band pole pairs or real poles were discarded. The spectrum and area function estimates shown are based on the LPC coefficients corrected by the root removal. The resulting area function estimate is quite good. The spectrum estimate does not contain the 6dB/octave rolloff, but the formant frequencies and bandwidths of the vocal tract spectrum are represented quite accurately. One has in effect corrected for source spectrum effects by identifying the vocal tract spectrum in terms of formant frequencies and bandwidths estimated by the narrow band LPC pole pairs.

The type of failure that occurred in the area function estimate has been previously reported [Ref. 9 p.81]. The technique of identifying the vocal tract spectrum by the LPC pole pairs that correspond to vocal tract formants has been previously presented [Refs. 7,10,25], but the robustness of this technique for avoiding errors associated with source

spectrum has never been emphasized. The effects of source spectrum may cause errors in the area function estimate, but catastrophic failures as shown in Figure 3.2 can be avoided.

Role of Vocal Tract Length in Area Function Estimates

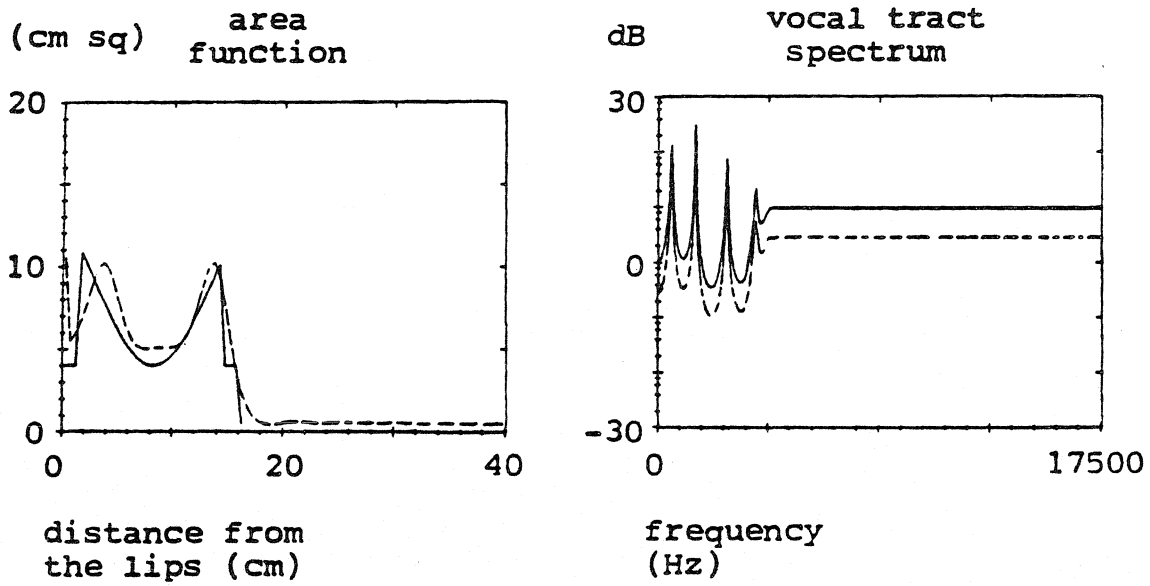
The robustness of the formant based area function analysis method in dealing with the source spectrum problem is clearly desirable. Working with the first four formant frequencies and bandwidths, however, one has eight coefficients to specify an eight coefficient tube. The length of this tube is fixed so one is back to the problem of a mismatch between acoustic tube length and vocal tract length.

The effects of vocal tract - acoustic tube model length mismatch are shown in Figure 3.3. Again, one starts with a known vocal tract shape target and calculates the vocal tract spectrum target using acoustic tube model loss and boundary conditions. Known values are compared against estimated values. In the first case, the vocal tract length matches the acoustic tube model length. The area function and spectrum estimates are quite accurate. In the second case, the vocal tract length is too short. The formant frequencies and bandwidths of the estimate are correct, but the formant amplitudes are wrong. The estimated area function has been made jagged.

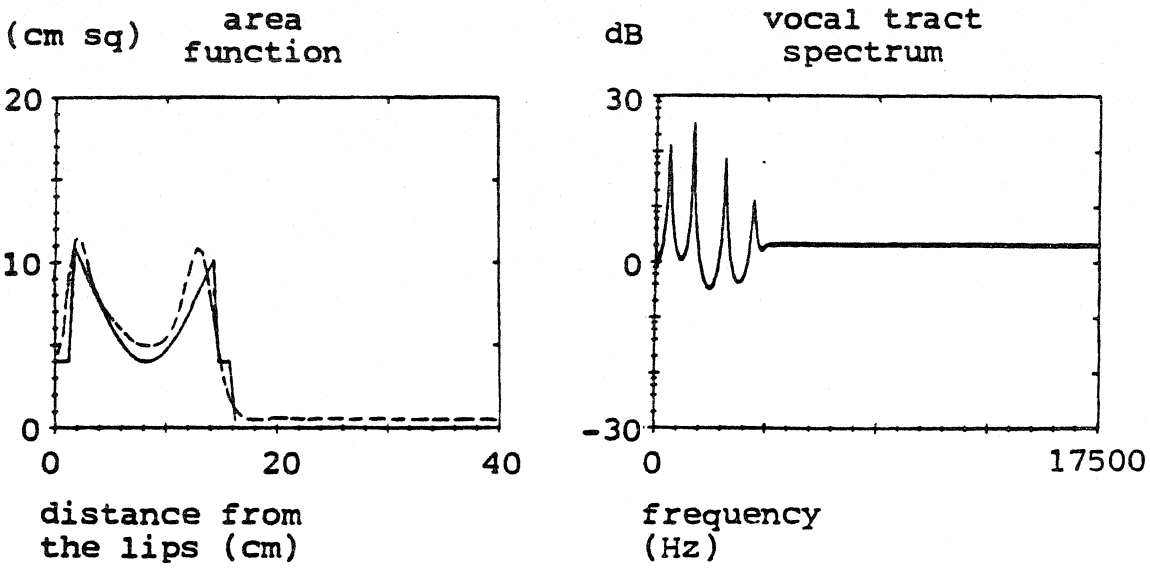
The key to getting formant amplitudes right [Appendix B] as well as getting accurate area function estimates is to match vocal tract length and acoustic tube length. Aside from the Procrustean approach of

varying vocal tract length to fit the acoustic tube, some means has to be used to vary the length of the acoustic tube. Because one is already working with formant frequencies and bandwidths, it is a simple matter to introduce a change in LPC filter sampling rate in converting back from frequencies and bandwidths to LPC poles [Appendix B]. A change in LPC filter sampling rate corresponds to a scaling of the length of acoustic tube sections. One would be tempted to determine the correct scaling by finding the sampling rate that gives a good match between vocal tract spectrum and the LPC spectrum. This, however, is circular reasoning as one uses the LPC estimate to determine what the vocal tract spectrum is to begin with. One can take advantage of the jaggedness of the incorrect area function estimate. The vocal tract length can be estimated by finding that acoustic tube that gives the smoothest area function estimate [Ref. 25].

Constant set too high:



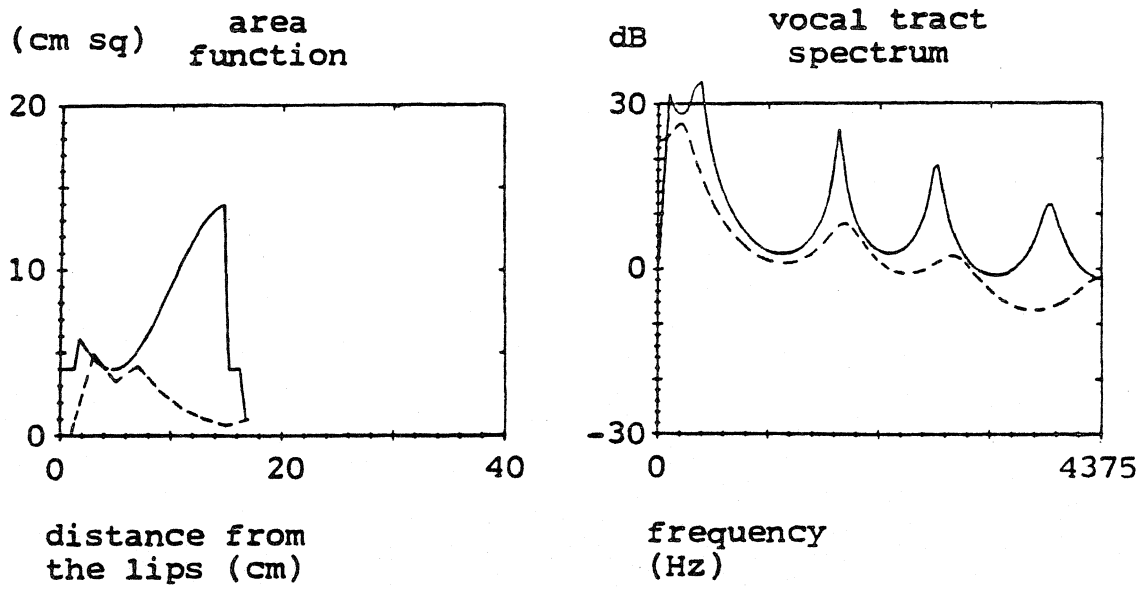
Constant set at the correct value:



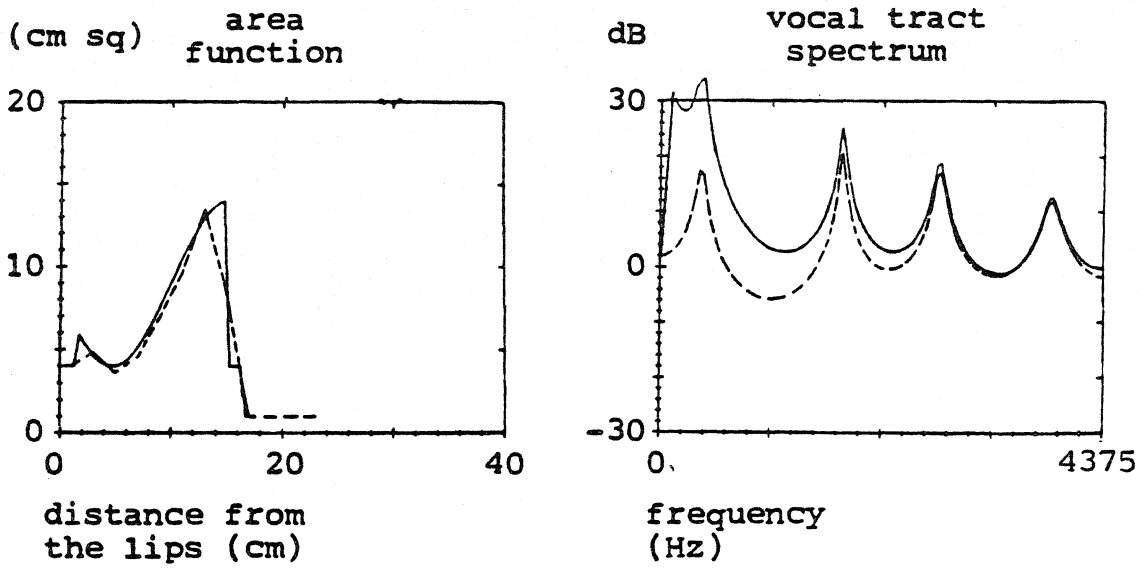
Estimating vocal tract area function by specifying the high frequency spectrum of the vocal tract as a constant.

Figure 3.1

Too few LPC coefficients used:



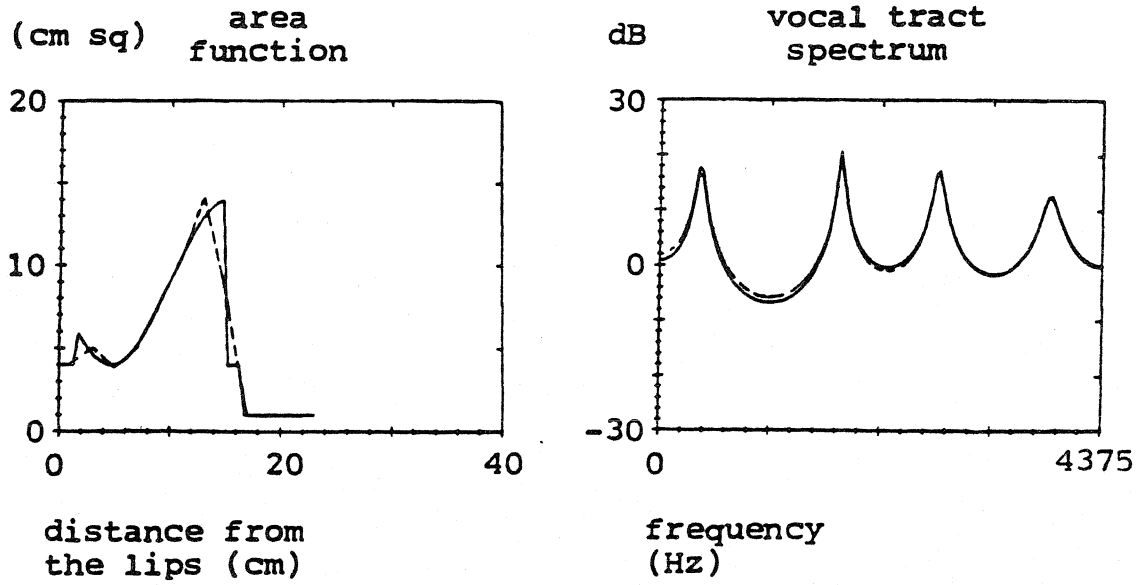
Additional LPC coefficients specified - vocal tract spectrum corrected by pole removal:



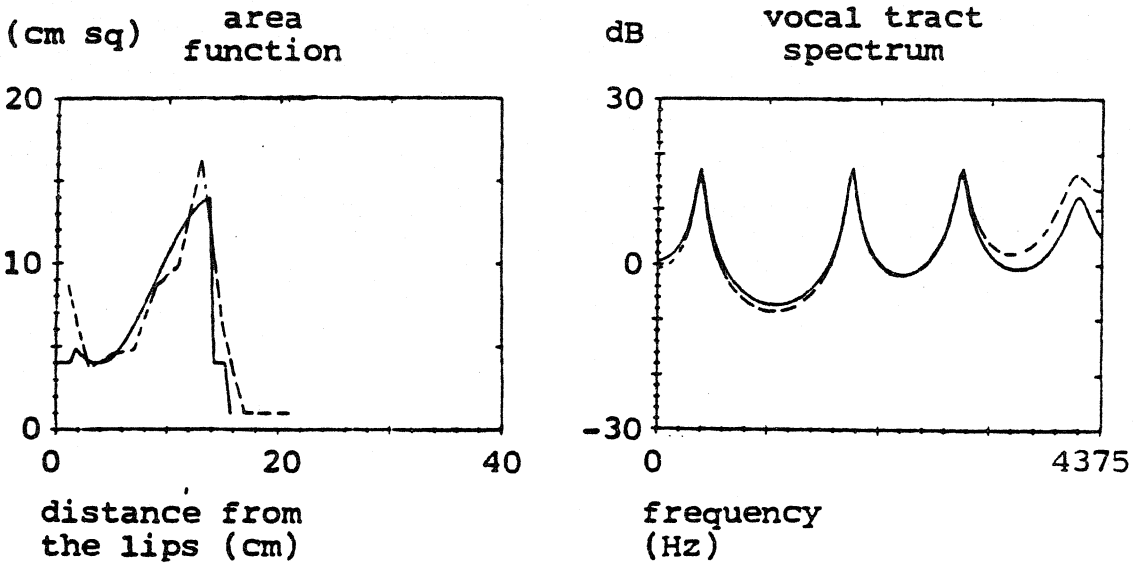
Estimating vocal tract area function and spectrum under conditions of a 6 dB/octave error in the vocal tract spectrum.

Figure 3.2

Vocal tract length matched to the acoustic tube model:



Vocal tract length too short:



Effect of vocal tract length on the accuracy of estimating vocal tract area function and spectrum.

Figure 3.3

IV. EVALUATING THE ACCURACY OF LPC FORMANT FREQUENCY AND BANDWIDTH ESTIMATES

Need for Estimating Both Formant Frequency and Formant Bandwidth

The experiments discussed in Section III demonstrated the value of basing the estimation of area function estimates on formant frequencies and bandwidths. The LPC derived formant frequencies and bandwidths are a robust means of extracting the vocal tract spectrum from the speech waveform by means of LPC analysis. The resulting area function estimates will be resistant to errors introduced by the glottal spectrum and the radiation characteristic. Estimates of both formant frequency and bandwidth need to be known because formant frequencies alone are not sufficient information to uniquely specify the cross sectional areas of an acoustic tube vocal tract model [Ref. 2].

The reliability of area function estimates is then limited by the reliability with which formant frequencies and bandwidths can be estimated. This provides the motivation for taking a detailed look into the accuracy of using the narrow band poles of the LPC filter to estimate the frequencies and bandwidths of vocal tract formants.

Problems in Frequency and Bandwidth Estimation

There are three main problem areas that influence the accuracy of LPC derived formant frequency and bandwidth estimates. The first problem is the effect of voicing periodicity - the periodic nature of the glottal waveform exciting the vocal tract. During voiced speech, the type of speech that the LPC derived area function estimates are good for, the vocal tract is excited by a series of nearly evenly spaced glottal pulses. This makes the resulting speech waveform nearly periodic, and applying LPC analysis to a periodic waveform results in significant errors in frequency and bandwidth estimates [Refs. 9 pp.188-189,33]. The second problem is the influence of the glottal pulse shape. The glottal pulse shape influences the envelope of the glottal spectrum. The glottal spectrum departs considerably from a constant dB/octave attenuation slope, especially for low frequencies. The locations of LPC poles estimating formants may by shifted by the structure of the envelope of the glottal spectrum. The third problem is the time variation of vocal tract damping during the progress of the glottal pulse. As the opening of the glottal constriction varies with vibration of the vocal cords, the contribution of glottal loss to vocal tract damping varies [Ref. 38]. It is necessary to understand what influence this has on estimated formant bandwidth values.

Closed Glottis Interval Analysis

The closed glottis interval analysis technique has been proposed as a means for avoiding errors introduced by glottal pulse shape and the variation in glottal damping from affecting formant frequency and bandwidth estimates [Refs. 10,26]. Covariance method LPC [Refs. 7,9,13] is applied to an analysis interval wholly contained within that portion of the glottal waveform that the glottis is closed. During this time, the speech waveform, corrected for the effects of the radiation spectrum, is purely determined by free decaying oscillations of the vocal tract. Furthermore, the acoustic of the vocal tract during this interval are based on a constant, zero, glottal damping. The LPC analysis should directly characterize the vocal tract under these conditions. In light of the claims made for this technique, it bears closer scrutiny.

One Resonance Vocal Tract Model

The problem areas in formant frequency and bandwidth estimation were investigated by use of a one resonance vocal tract model. This way the study of this problem is kept simple; one has only one resonance to identify in the analysis. This resonance can be excited either by periodic impulses or by periodic glottal pulse shapes. The quantities that can be controlled are resonance frequency and bandwidth under

closed glottis conditions, the pitch (fundamental frequency) of the excitation, the analysis interval, glottal pulse duty cycle, and the amount of added damping that occurs during glottal opening. The number of LPC coefficients used in the analysis can be specified. The LPC analysis is conducted by a stablized covariance method that is constrained to give a stable LPC filter model [Ref. 39]. A stable LPC filter insures that all pole pairs have positive values of damping. The LPC polynomial is factored into quadratic terms using Bairstow's method [Ref. 40]. These quadratics are factored into pole pairs from which frequency and bandwidth values can be computed and displayed. The resonance is characterized by the pole pair with the smallest bandwidth value; the other poles correspond to artifacts introduced by the excitation waveform.

The one resonance vocal tract model is implemented with a two section acoustic tube model. The lip end is terminated in a short and the glottal end is terminated in the linear time varying excitation source described in Appendix D. The section length of the acoustic tube is set to 1.75 cm, a length corresponding to a sampling rate of 10 KHz. Varying the ratio of areas of the two sections varies the resonant frequency of the acoustic tube over a range of 0 to 5 KHz. Damping is introduced into the sections to give a non zero closed glottis bandwidth for the model. The acoustic tube sections impose both a delay and an attentuation on forward and backward going wave components passing through them [Ref. 9 p.88]. The delay and attentuation are invariant with frequency, meaning that the acoustic tube model remains nondispersive. The output of this model is obtained by running a

dynamical simulation of the passage and reflection of forward and backward travelling wave components thru the model. The advantage of the acoustic tube model is that it gives a stable, discrete time formulation of a resonator that can be easily tuned over the range of possible frequencies and bandwidths.

In the case of impulse excitation, LPC analysis is performed directly on the output of the resonator. In the case of glottal pulse excitation, the output is boosted +12 dB/octave prior to doing LPC analysis, +6 dB corresponding to vocal tract radiation characteristic and +6 dB corresponding to preemphasis. Correcting for the 12 dB/octave rolloff of the glottal spectrum reduces the dynamic range of the spectrum of the signal being analysed. This reduction in dynamic range improves the numerical accuracy of LPC [Refs. 9,13,41]. Even though preemphasis was originally advanced as a means of correcting for source spectrum, it is still useful in obtaining improved numerical properties in LPC analysis.

Figure 4.1 demonstrates the use of LPC on the output of the one resonance vocal tract model to recover the formant oscillation from the output signal. The pole pair corresponding to the formant has been removed from the LPC polynomial, and analysed signal has been passed thru the LPC inverse filter, removing all other components except for the formant oscillation. This formant oscillation has been plotted alongside the glottal pulses exciting the resonance. The tick marks represent the bounds of the LPC analysis interval. The second and third tick marks indicate the bounds of the error minimizing interval;

the first and third indicate the extent over which signal samples enter into the calculation of the correlations. The glottal pulse shape used is one suggested by Fant [Ref. 38]. The use of such a shape, having a smooth opening and hard glottal closure, is justified by glottal shapes that are observed in impedance tube observations [Refs. 18,42], inverse filter observations [Refs. 43-47], and perceptual judgements on the effects of glottal pulse shape on speech synthesis [Refs. 48,49].

The first plot in Figure 4.1 shows results for a frequency 500 Hz, bandwidth 100 Hz resonance excited by a glottal pulse where there is no glottal damping. The pitch rate is 100 Hz. The plot beside it shows the effect of introducing a large amount of glottal damping. One should note how the formant oscillation is quenched when the glottis opens. The remaining plots in the figure show the same amount of glottal damping where the closed glottis damping is reduced to 50 Hz and then to 10 Hz. The amount of glottal damping is quite heavy - a glottal coupling coefficient of .75. The glottal coupling coefficient gives the amount of wave reflection occurring at the glottal interface during maximum glottal opening area. A coefficient of 0 corresponds to perfect reflection - no glottal loss. A coefficient of 1 corresponds to perfect transmission - a matched glottal termination. A coefficient of .75 can result from lung pressure of 8 cm H_2O, air density of $1.15(10^{-2})$ gm/cm^3, sound velocity of 35,000 cm/sec, maximum glottal opening area of 20 mm^2, and a pharynx cavity area of 5 cm^2 [Appendix D Eqn D.12].

Results of inverse filtering to recover the glottal waveform from the

the output of the one resonance vocal tract model are shown in Figure 4.2. All roots except for the roots corresponding to the complex conjugate resonance pole pair are removed from the LPC polynomial, and the vocal tract model output is run thru the LPC inverse filter. Integrating the signal twice corrects for the radiation characteristic and preemphasis boost, recovering an estimate of the glottal waveform. The first plot shows the effect on recovered glottal waveform where heavy glottal damping occurs - a glottal coupling coefficient of .75 is used. A pronounced slope has been placed on the closed glottis portion of the waveform. This plot was obtained using an LPC analysis interval containing an entire pitch period. Closed glottis interval analysis does not give any improvement - in fact contamination of the waveform by formant oscillation is increased. The plot below shows the effects of no glottal damping. The closed glottis portion of the recovered waveform is now flat. The slope placed on estimated glottal wave under conditions of heavy glottal damping may be an artifact of the approximations made in formulating the glottal model [Appendix D]. A glottal damping value of .75 is quite large in light of linearizations made in the model. The slope in the closed glottis interval portion of the waveform, however, occurs in Joan E. Miller's inverse filtering data from actual speech [Ref. 50], indicating that the slope may have its origin in the time varying nature of the glottal damping.

It should be noticed that in the closed glottis interval analysis case, the analysis interval includes the moment of glottal closure. Placing the analysis frame entirely within the closed glottis interval results in numerical instability of the LPC analysis. The origin of this

instability comes from the fact that the linear predictor (LPC inverse filter) gives zero output, or prediction residual, if the analysis frame is totally within the closed glottis interval. A zero prediction residual makes calculating LPC coefficients from correlations of the data ill conditioned, imposing a 0/0 uncertainty [Ref. 51]. Having to include the moment of glottal closure in the analysis interval, however, defeats the aim of the closed glottis interval technique of analysing free decaying vocal tract oscillations.

Experimental Results

The one resonance vocal tract model is now applied in a series of experiments. Atal's experiment on the effects of voicing periodicity [Ref. 9 pp.188-189,33] was redone; results are shown in Figure 4.3. Estimated formant frequency and estimated formant bandwidth are plotted against formant frequency. The actual formant bandwidth was fixed at 100 Hz and the pitch frequency F_0 was fixed at 200 Hz. The one resonance vocal tract model was excited by a train of impulses. A refinement on Atal's experiment was to compare results of the first impulse with the fifth impulse. The first impulse starts with the vocal tract model at rest before exciting it. The fifth impulse excites the vocal tract in a condition where it has approached the steady state response to periodic excitation. LPC analysis of the first impulse gives very accurate frequency and bandwidth estimates, even though the analysis interval is only one pitch period long. The accuracy is insensitive to the number of LPC coefficients used. The case of fifth impulse shows

systematic error in frequency and very large systematic error in formant bandwidth estimation, even though the same pitch period long analysis interval was used. The error is worse as the number of LPC coefficents is increased. One interpretation to these results is that the pitch period long analysis interval applied to the first impulse contains the pure impulse response of the vocal tract model. Pitch period long frames of succeeding pitch periods contain a systematic series of overlaps of the tails of impulse response from previous pitch period contaminating the impulse response in the pitch period being analysed. The frequency domain interpretation is contained in Appendix C.

Figure 4.4 shows the results of correcting for the effects of voicing periodicity. The vocal tract model is excited with periodic impulses; pitch frequencies of 100, 200, and 400 Hz are considered. The line spectrum of the resonator output is computed from a pitch period long interval under steady state periodic conditions. This line spectrum is interpolated by treating the reciprocal values of the line spectrum as samples of a bandlimited function [Appendix C]. Interpolating the line spectrum to specify the spectrum at a denser set of points is equivalent to the effect of having a lower pitch frequency, hence lower voicing periodicity error. The number of line spectrum values specified is expanded to eight for every one. The interpolated power spectrum is inverse Fourier transformed to give autocorrelation coefficients from which one computes LPC coefficients. The plots indicate that this technique is effective in correcting for voicing periodicity errors for pitches up to 400 Hz. The case of 400 Hz pitch runs into

trouble, however, where the frequency of the resonance approaches and goes below 400 Hz, which is consistent with expectations.

Unfortunately, this method of correcting voicing periodicity errors has not been successfully applied to the case where a glottal pulse train is exciting the resonator. The combination of the glottal spectrum and the 12 dB/octave boost influences the output of the resonator in a way to make the assumptions that went into the method invalid [Appendix C].

Figure 4.5 shows results for applying closed glottis interval LPC to the case where the vocal tract resonator model is excited with periodic glottal pulses of 33 per cent duty cycle and 200 Hz pitch frequency. There is no glottal coupling occurring here. It is evident that the closed glottis interval technique is not immune to pitch periodicity errors. Both 2 and 4 coefficient LPC give comparable results.

Figure 4.6 shows formant estimation in the frequency range of the first formant; Figure 4.7 shows formant estimation in the frequency range of the second formant. A 33 per cent duty cycle glottal pulse train with no glottal damping excites a resonance of 100 Hz bandwidth and variable frequency. The pitch frequency F_0 is fixed at 200 Hz. The LPC analysis is done with 4 coefficents; LPC done with a pitch period long analysis frame is compared against closed glottis interval LPC. The closed glottis method gives no improvement in estimation accuracy over analysing over the whole pitch period. Both methods suffer from the effects of voicing periodicity. The voicing periodicity errors in the whole pitch period case, however, vary more consistently with variation in formant frequency than the errors for

closed glottis interval analysis. Increased error in frequency and bandwidth estimates occur for formant frequencies approaching the pitch frequency. This error is attributable to the use of glottal pulses instead of impulses. Note that both the whole pitch period and closed glottis interval methods suffer from this error to the same degree.

Figure 4.8 shows the effect of changing duty cycle from 33 to 50 per cent. The added frequency and bandwidth error that occurs at low frequencies changes with duty cycle, confirming that this error results from the influence of glottal pulse shape.

Figure 4.9 shows results where time varying glottal damping is introduced. The pitch frequency 100 Hz and the glottal duty cycle is 33 per cent. Whole pitch period analysis is used. Values of glottal coupling up to .5, a moderately high value, were tried. The results are extremely surprising; time varying glottal damping actually improved both frequency and bandwidth estimates. Furthermore, the estimated damping values converged to the closed glottis value. The contribution of glottal damping to the resonance bandwidth seen by LPC is practically nil, even with glottal coupling wide open. The answer is found in the plots of Figure 4.1. Glottal closure excites the resonance. By the time the glottis opens, the resonance oscillation has largely decayed. LPC analysis is keying into the closed glottis interval because that is where most of the energy of the oscillation is found. The opening of the glottis serves to quench the oscillations, resetting the resonance to the quiescent condition by the time the next glottal closure occurs. This resetting action avoids the errors of voicing

periodicity.

Figure 4.10 gives worst case results for glottal damping. The pitch frequency is 200 Hz, the duty cycle is 50 per cent, and glottal coupling has been increased to .75, a very large value. Whole pitch period and closed glottis analysis are compared. Whole pitch period gives decreased frequency estimates and greatly increased bandwidth estimates. Closed glottis interval analysis does not improve matters - frequency estimates are erratic and bandwidth estimates are too low by the same amount as whole pitch period estimates are high.

In Figure 4.11, the contribution of glottal damping with a glottal coupling coefficient of .75 to estimated bandwidth for various different closed glottis bandwidths is examined. The glottal duty cycle is 50 per cent and pitch frequency is 100 Hz. The contribution to bandwidth with closed glottis bandwidth at 100 Hz is practically nil. At 50 Hz the contribution is an additional 10 Hz, and for a closed glottis bandwidth of 10 Hz, the bandwidth added by glottal damping is 30 Hz. For the lower values of glottal bandwidth, the formant oscillation has not decayed as much at the time the glottis opens. Returning to Figure 4.1, one sees that for the 10 Hz closed glottis bandwidth case, glottal opening puts a large kink in the envelope of the formant oscillation, a kink that is reflected in an increased bandwidth estimate.

Figure 4.12 addresses the question of what happens if the LPC analysis interval does not contain exactly one pitch period. Using one and a half pitch periods caused considerable degradation of accuracy in formant bandwidth estimates. Increasing the interval past two pitch

periods decreased the influence of the fractional pitch period. In cases where pitch period is not being estimated, one should use an analysis interval using as many pitch periods as possible to reduce the degradation of including fractional pitch periods. The desirability of long analysis intervals is borne out by experience in the use of LPC derived areas [Ref. 52].

Conclusions

Accurate formant frequency and bandwidth estimates are necessary for obtaining accurate area function estimates. The periodicity of the glottal waveform for voiced speech is the major source of error in formant estimation, especially bandwidth values. The problem is worse with increased pitch frequency; a pitch of 200 Hz can give threefold variation in bandwidth estimates.

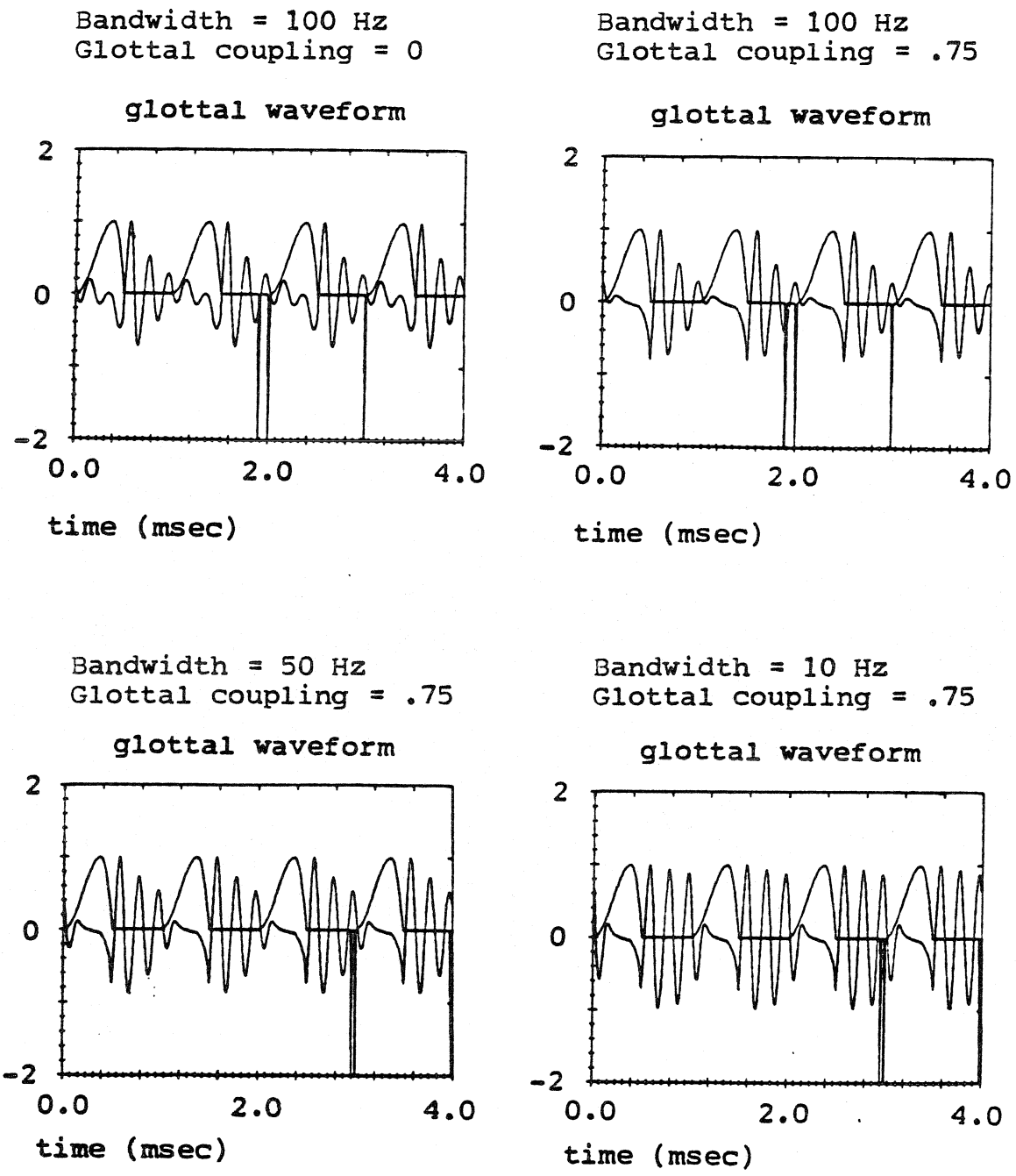
The voicing periodicity errors could be corrected if the vocal tract were excited by periodic impulses. Work is needed to develop a method of correction that works on speech, given that the vocal tract is excited by glottal pulses, not impulses.

The effects of glottal pulse shape are much less of a problem than voicing periodicity. The errors introduced by the glottal pulse shape are confined to the low frequency range of first formant, so the entire vocal tract spectrum is not affected. Methods such as Nakajima's adaptive preemphasis [Ref. 21] correct for variations in the slope of the glottal spectral envelope - they bear looking into if errors introduced

by glottal pulse shape prove to be a problem.

Time varying glottal damping actually reduces the errors of voicing periodicity. Glottal opening serves to reset the vocal tract oscillation to the quiescent condition before the next oscillation is excited by glottal closure. The improvement may not be as good at higher frequencies because glottal damping is reduced by glottal inductance [Ref. 1]. Glottal damping contributes much less to formant bandwidth, or rather LPC estimated formant bandwidth, than previously thought [Ref. 1]. As closed glottis bandwidth is increased, the contribution of glottal damping vanishes because the formant oscillation is practically gone by the time the glottis opens.

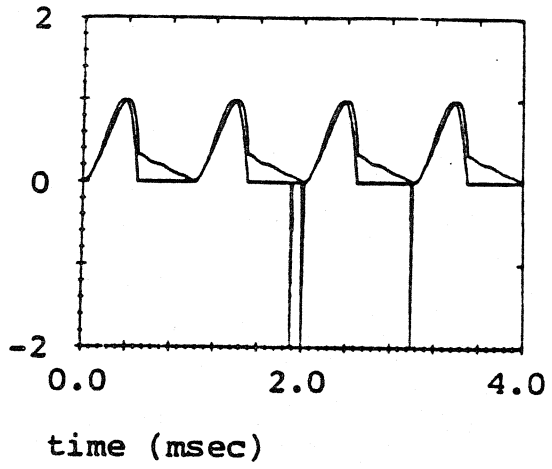
Closed glottis interval analysis is completely useless because there is no improvement in performance over analysing over entire pitch periods. There is no reason then to introduce the complexity of having to locate the times of glottal closure in the speech waveform.



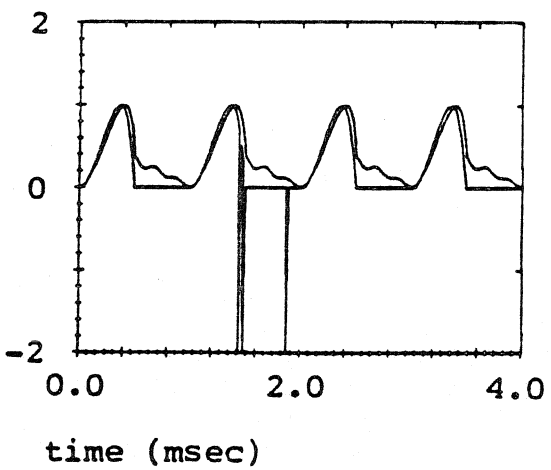
Inverse filtered formant waveforms for different closed glottis bandwidths are compared with glottal excitation waveforms.

Figure 4.1

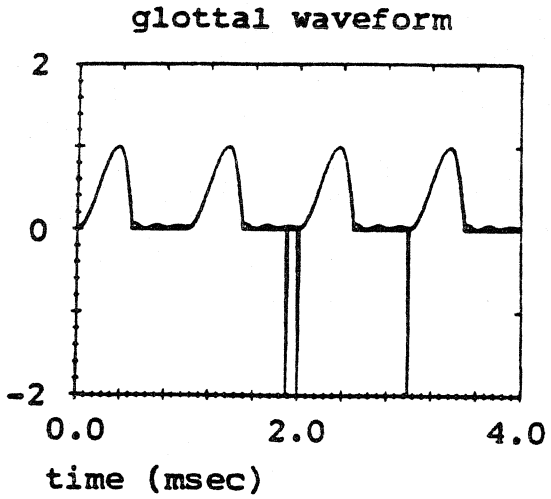
Whole pitch period method.
Glottal coupling = .75



Closed glottis interval method.
Glottal coupling = .75



Whole pitch period method - no glottal damping.



Inverse filtered glottal waveforms compared with actual glottal waveforms.

Figure 4.2

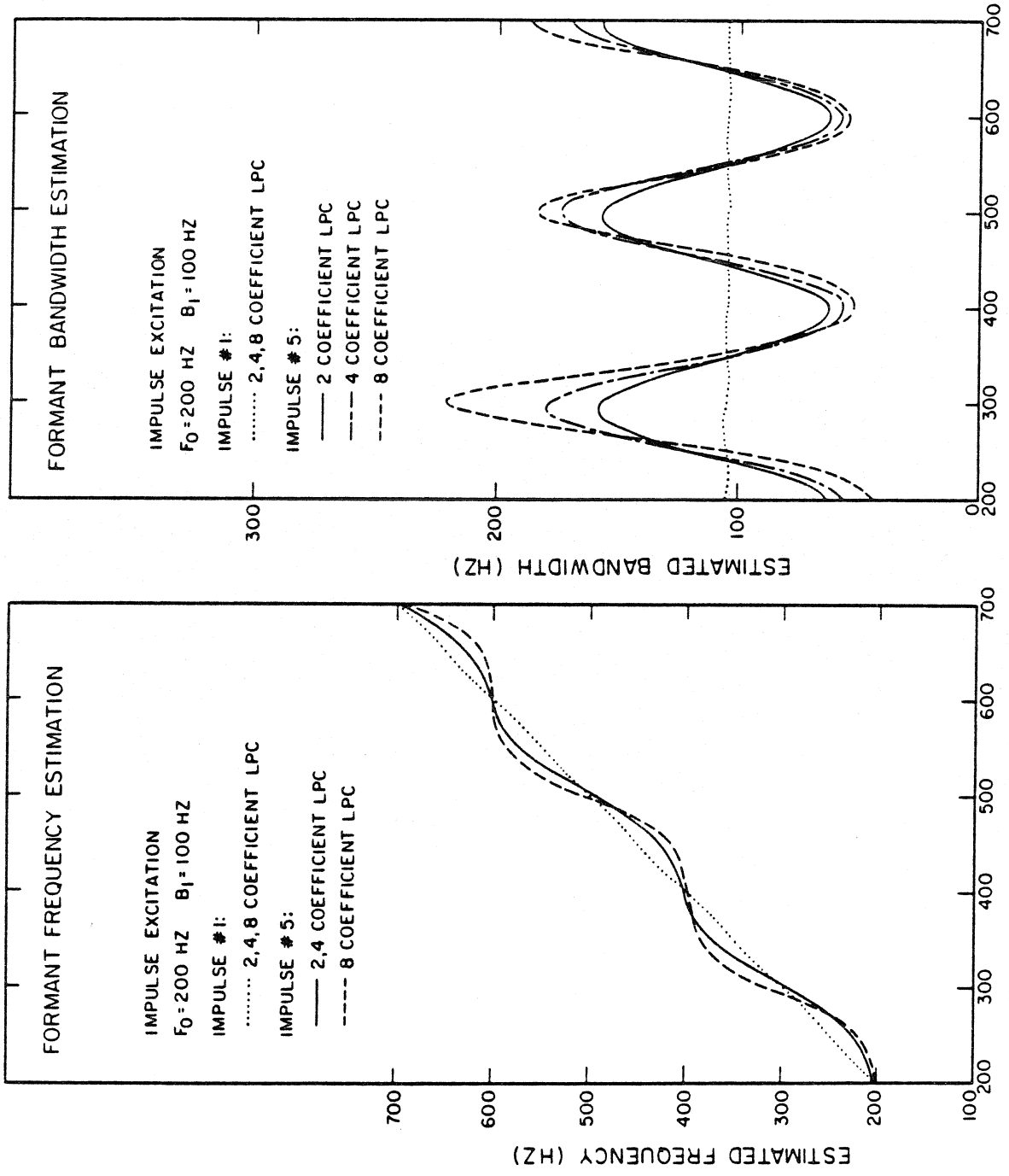


Figure 4.3

Using LPC to estimate the frequency and bandwidth of a resonance excited with a periodic train of impulses.

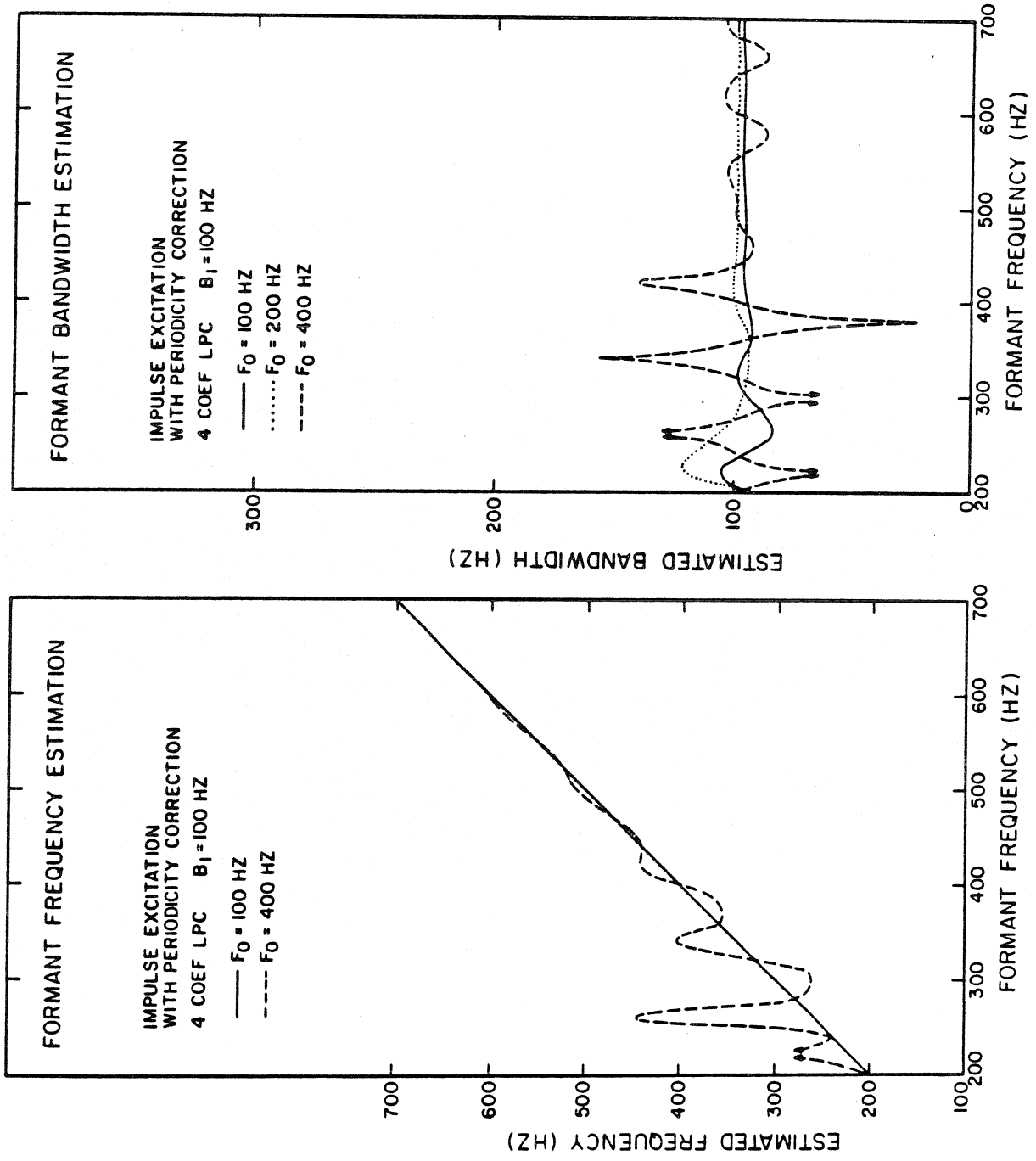


Figure 4.4

Correcting for the errors of voicing periodicity by interpolating the spectrum of the LPC inverse filter.

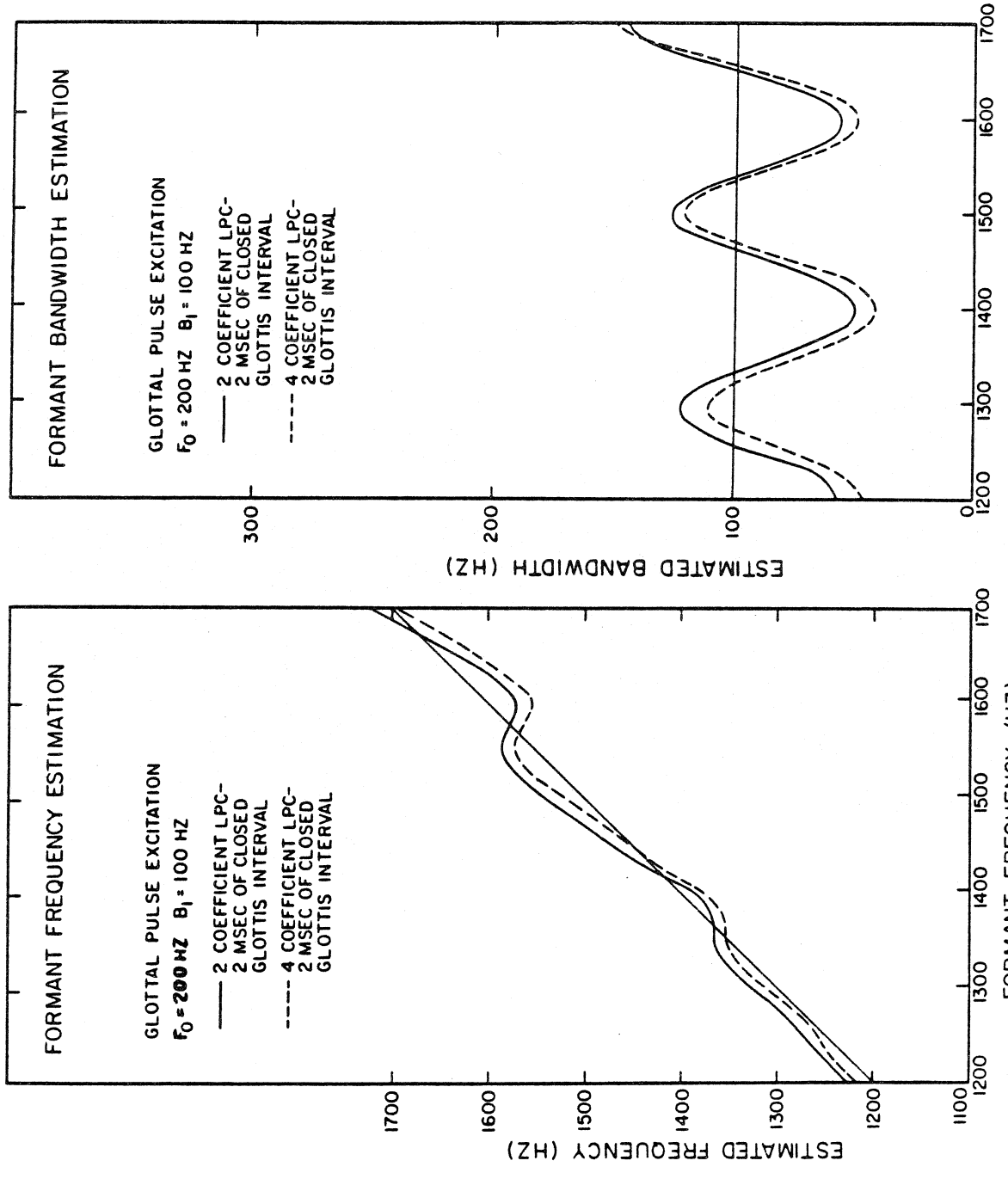


Figure 4.5

Effect of the number of LPC coefficients on closed glottis interval analysis of a single resonance excited by glottal pulses.

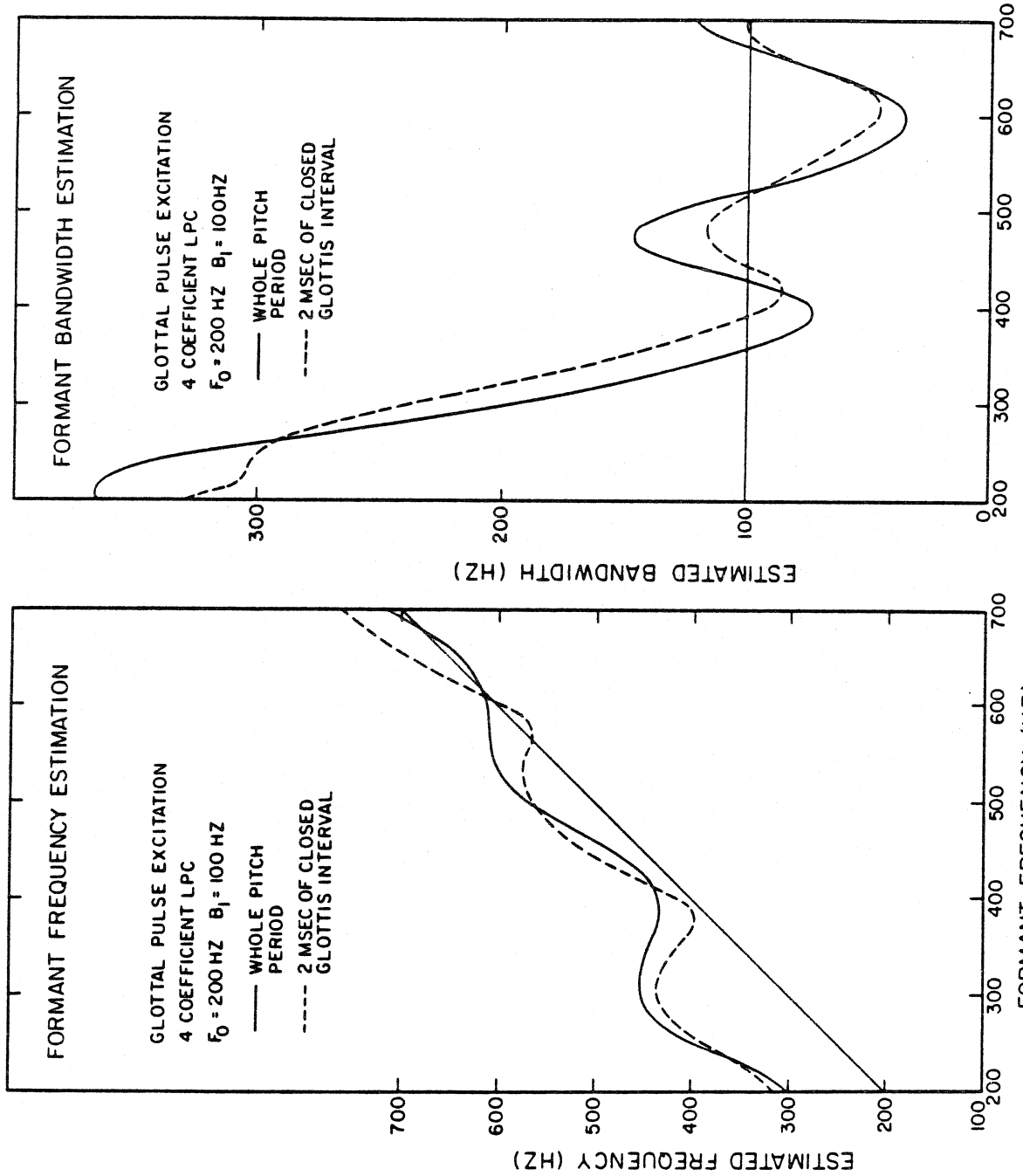


Figure 4.6

Closed glottis interval analysis compared against whole pitch period analysis of a single resonance excited by glottal pulses in the frequency range of first formant.

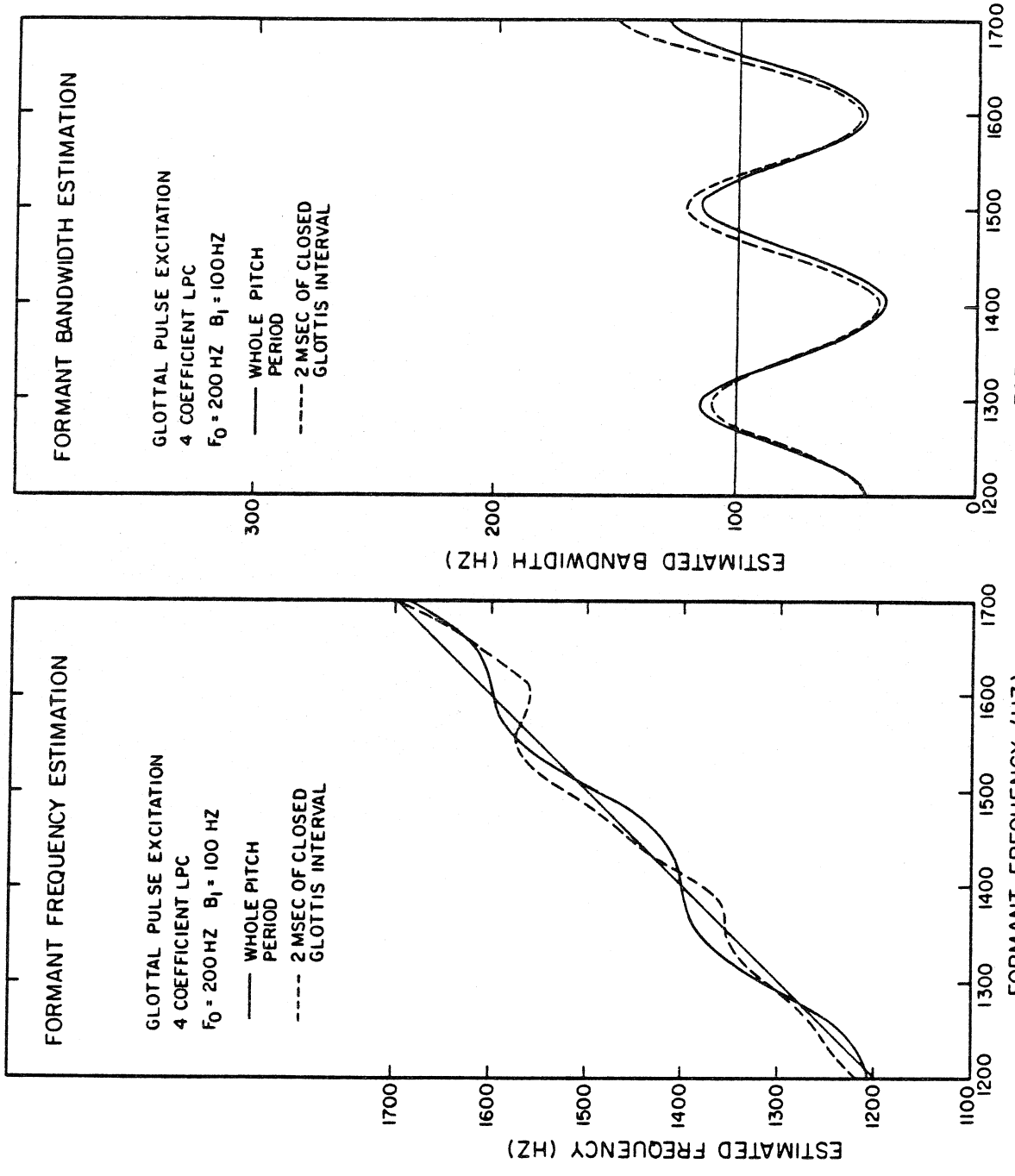


Figure 4.7

Closed glottis interval analysis compared against whole pitch period analysis in the frequency range of the second formant.

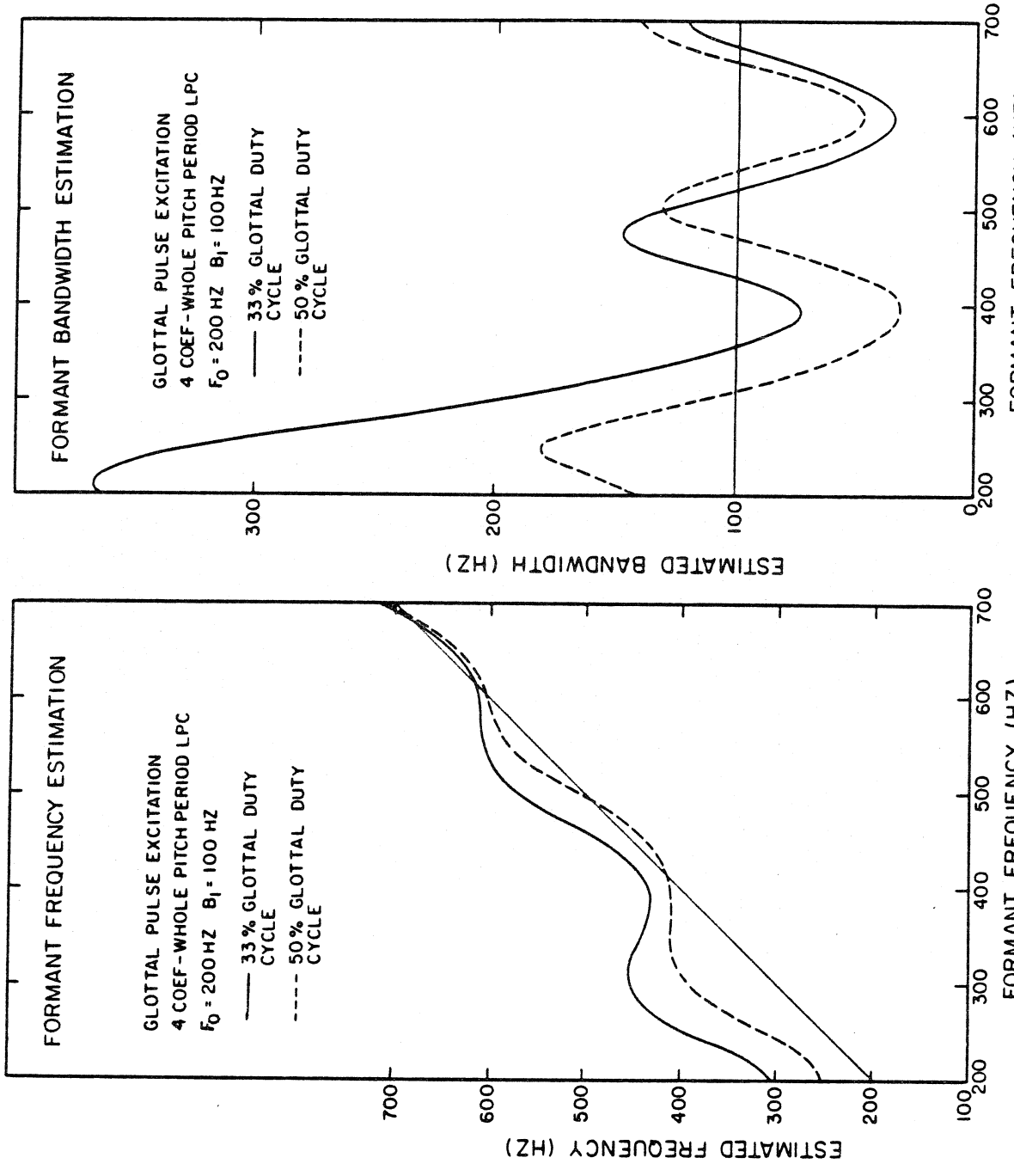


Figure 4.8

Effect of glottal duty cycle on frequency and bandwidth estimation of the first formant.

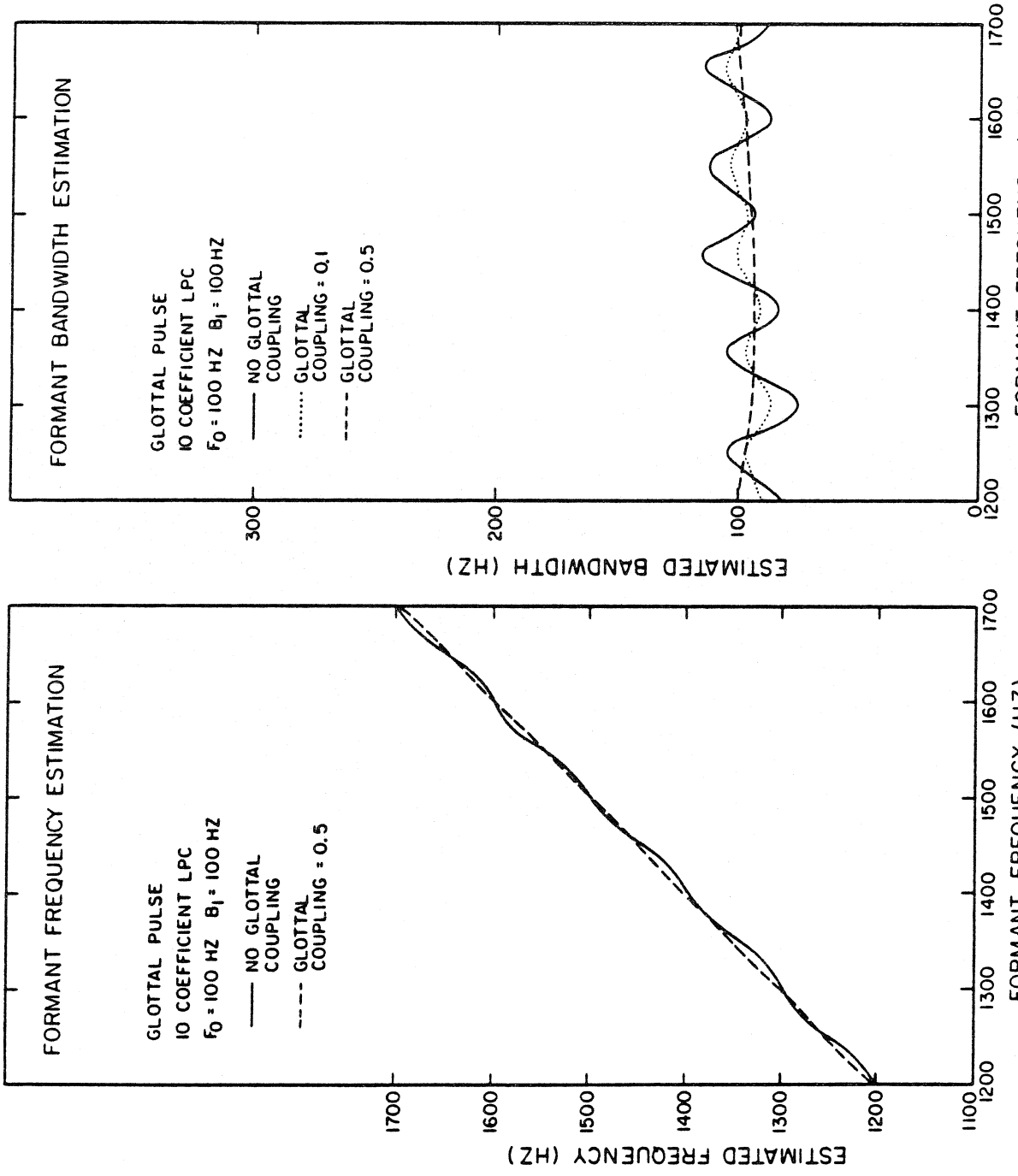


Figure 4.9

Best case results under conditions of time varying glottal damping - glottal duty cycle = 33%, glottal coupling coefficient = .5

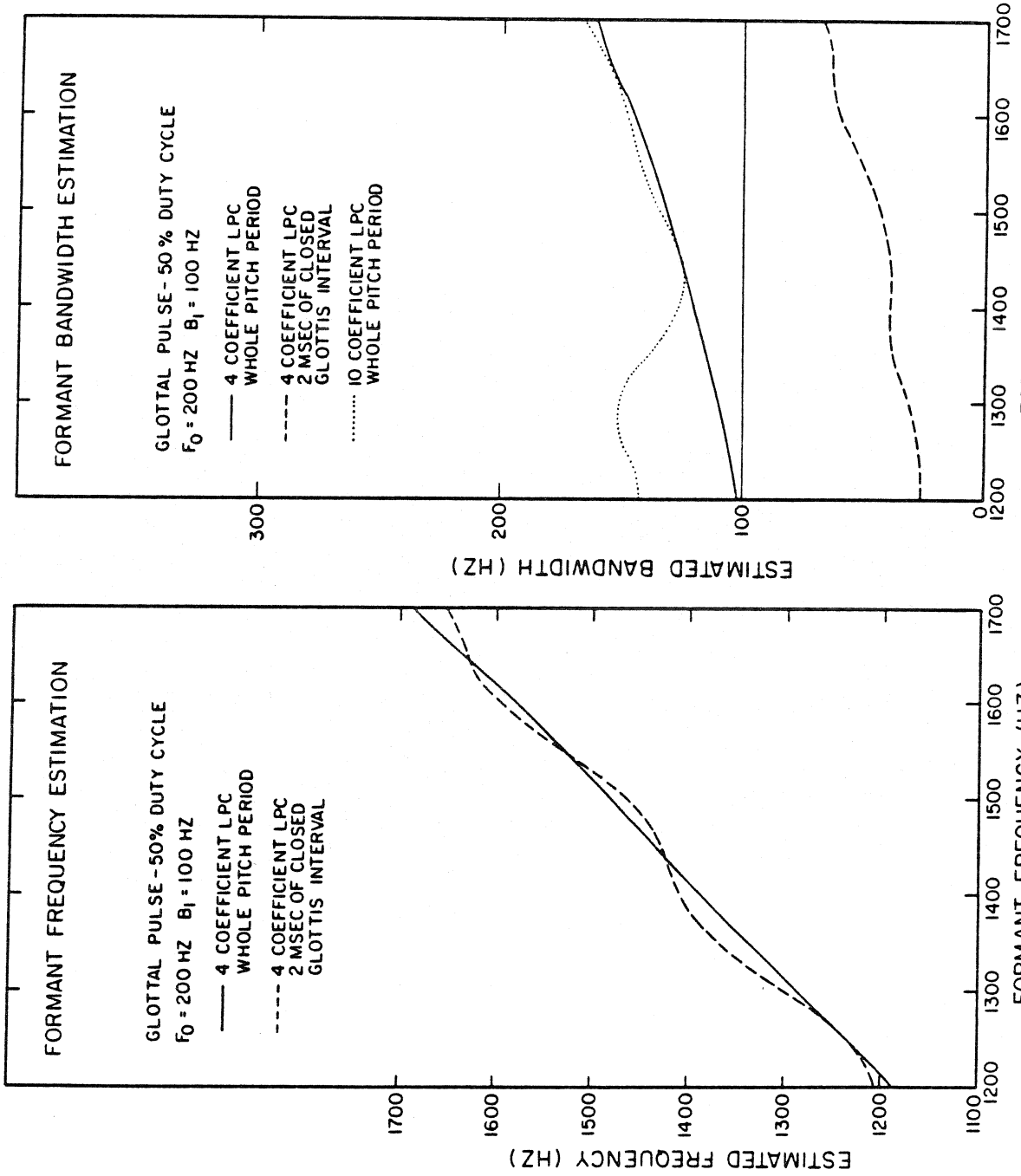
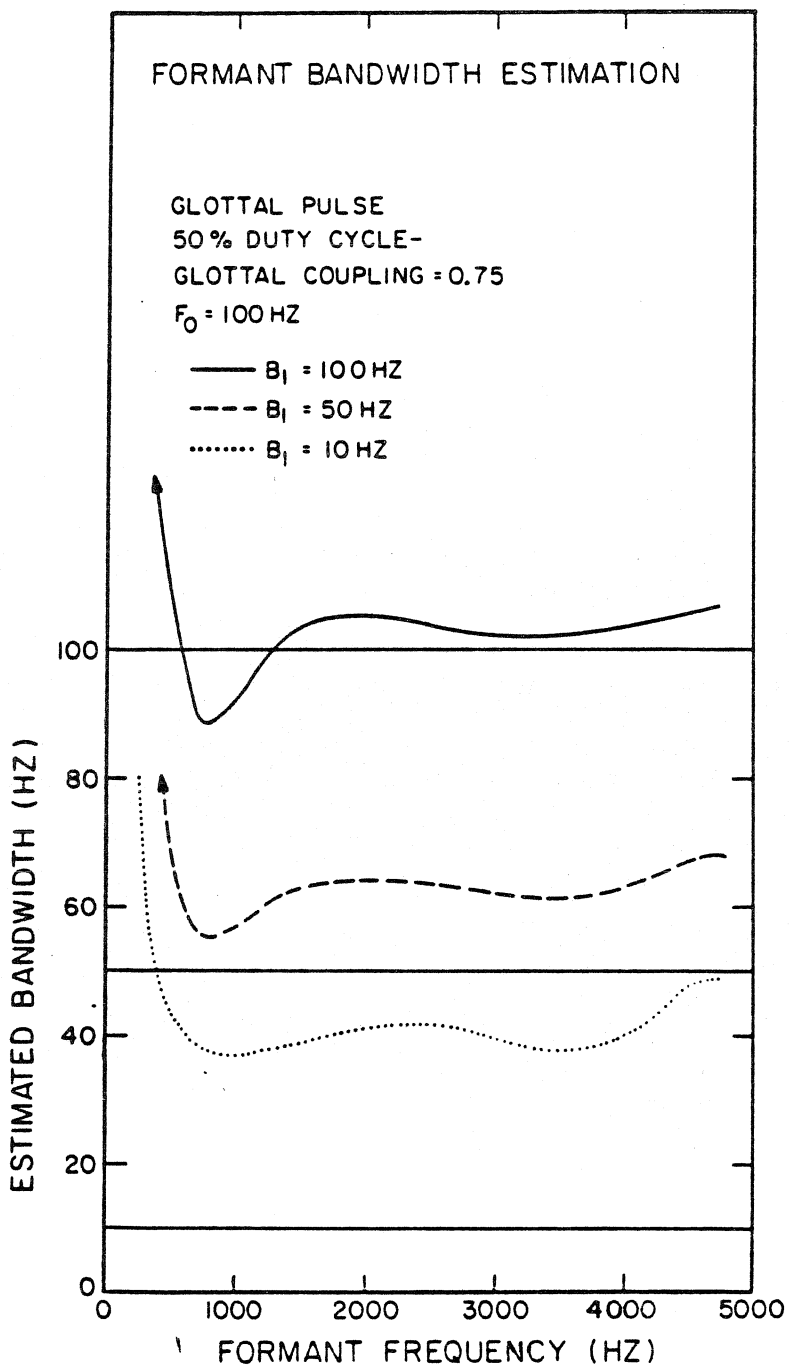


Figure 4.10

Worst case results under conditions of time varying glottal damping - glottal duty cycle = 50%, glottal coupling coefficient = .75



Effect of heavy glottal damping on resonances of varying closed glottis bandwidth.

Figure 4.11

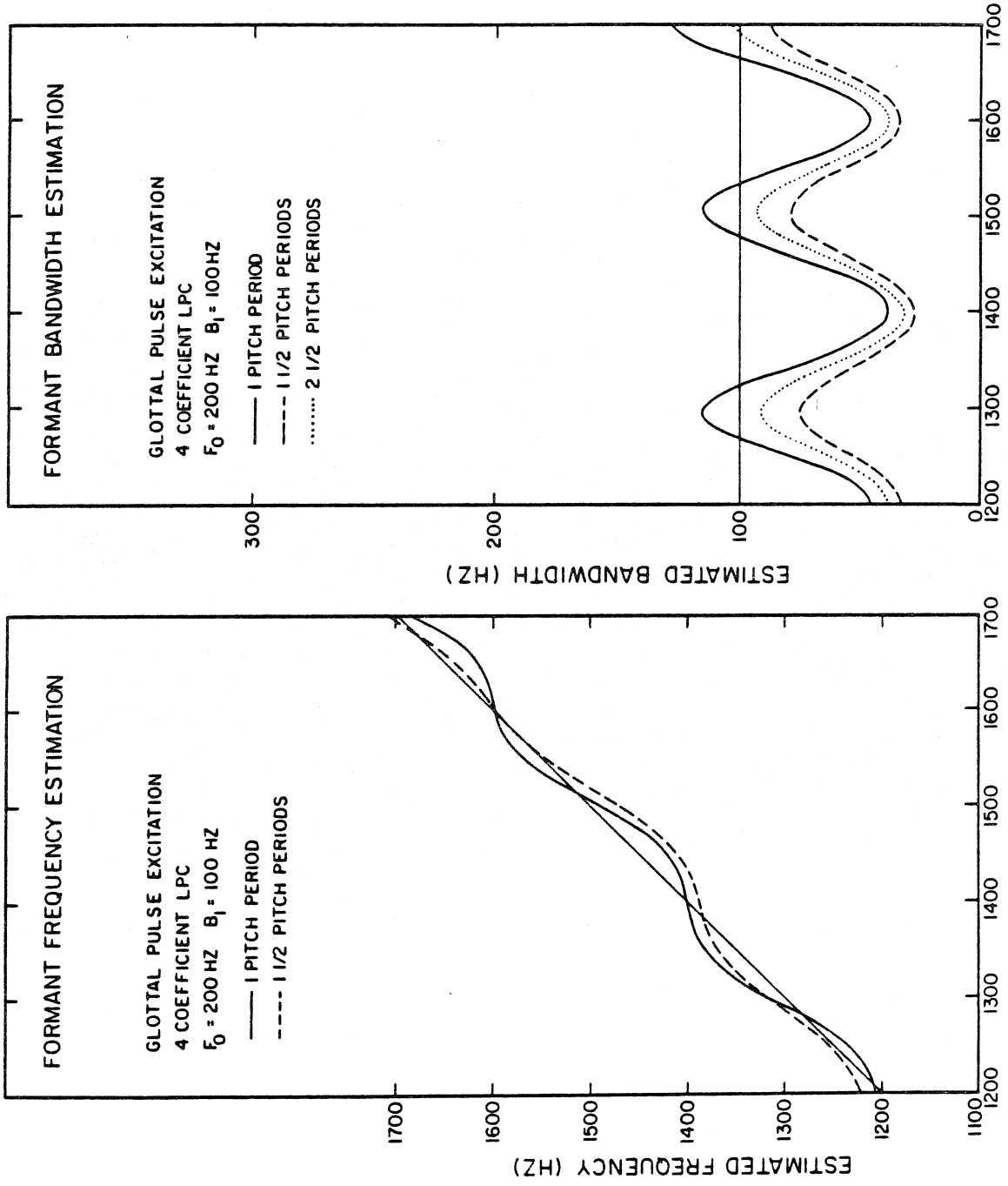


Figure 4.12

Effect of a non-integral number of pitch periods in the LPC analysis interval.

V. ESTIMATING VOCAL TRACT SHAPES FROM SYNTHETIC SPEECH SOUNDS

Purpose of Using Synthetic Speech

The effectiveness of the LPC technique for estimating the vocal tract area function is evaluated by using synthetic speech. One can start with a known vocal tract shape and perform acoustical calculations to determine the spectrum of the vocal tract response. This spectrum can be changed to represent the effects of the glottal spectrum and the radiation characteristic. This gives the spectrum of the synthetic speech sound corresponding to the vocal tract shape. Taking the inverse Fourier transform of this speech spectrum gives the autocorrelation coefficients that one would have obtained if one were analysing a pitch period long interval of voiced speech. LPC coefficients can be calculated from these autocorrelations; the LPC coefficients can then in turn be processed to calculate the area function of an acoustic tube vocal tract model. The spectrum of the synthetic speech signal is specified by discrete sample values that are uniformly spaced in frequency, the frequencies of the sample values being the pitch frequency and its harmonics.

The advantage of synthetic speech is that one can compare estimated area function against a known vocal tract shape. The other advantage is that one can control the number and severity of the different error sources that act on the area function estimate; the error sources can

be studied one at a time. The real test of area function estimation is to apply it to actual speech. The purpose of this study, however, is to understand the effects of the error sources so that when experiments are conducted on actual speech, the results can be properly interpreted. With actual speech, one has nothing to verify the accuracy or reliability of area function estimates except intuition.

Speech Synthesis Model

Speech sounds are synthesized by taking the area function values for vowel sounds given by Fant [Ref. 3] and by calculating the spectrum of the response of an acoustic tube model having section areas specified by the Fant areas. The vocal tract area function is specified every half centimetre so the acoustic tube section length is .5 cm and the length of the acoustic tube in centimetres is the number of area function values specified divided by two. The spectrum of the vocal tract model so specified is calculated by the method described by Atal [Ref. 2]. An equivalent circuit model of the acoustic tube and its terminations is formulated. This formulation allows for the heat conduction and viscosity loss in acoustic wave propagation, the lossy mechanical impedance of the vocal tract walls, the proper radiation load lip termination, and a time invariant resistive glottal termination. For a given frequency, all the acoustic impedances in the circuit formulation are known, so one can solve for the transfer function of the network. Taking the magnitude squared of transfer function give the spectrum. The network has to be solved for each frequency that one needs to

compute the spectrum because the impedances are frequency dependent.

The technique used for synthesizing a speech sound is based on time invariance. The vocal tract shape is time varying, but the shape changes very slowly compared to speech waveform analysis intervals chosen. The glottal termination, however, is time varying synchronously with pitch periods. Ideally one should use a dynamical simulation of the acoustic tube vocal tract model to compute a synthetic speech waveform that reflects the effects of a time varying glottal termination. Flanagan has developed a dynamic model of vocal tract and vocal cords [Ref. 53]. A circuit model formulation of the vocal tract is expressed as a system of differential equations acting on the state variables of the vocal tract - pressures and volume velocities in the acoustic tube sections. The solution to these differential equations is obtained in the time domain by approximating the equations as difference equations. The Flanagan model, however, does not have as accurate a circuit model for all the loss mechanisms as the Atal model does. Incorporating a better circuit model into the Flanagan model is difficult because there are stability considerations to be dealt with in the difference equations used to obtain the model output.

Estimated on a time invariant basis, it is shown in Section IV that the glottal contribution to formant damping is for the most part frequency independent but nonlinear, increasing as closed glottis bandwidth decreases. The Atal model, used in the experiments described in this section, has a resistive termination that has a damping contribution that is both frequency independent and linear. The resistance value

can be chosen to represent the glottal damping contribution for the average value of formant bandwidth occurring.

The wall impedance values used in this study differ significantly from those used in the study done by Atal [Ref. 2]. Flanagan [Ref. 1 p.68] quotes a specific wall resistance of 6500 dyne-sec/cm^3 and a specific wall mass of .4 gm/cm^2. When this amount of wall loss is combined with the glottal loss estimates made by Flanagan, the resulting formant bandwidths are much too large. Atal took the approach of reducing the specific resistance to 600 dyne-sec/cm^3 and increasing the wall mass value to 2 gm/cm^2. The wall loss is reduced while large values of glottal loss are retained. The results of Section IV indicate that even with very heavy glottal damping, the formant bandwidths estimated by LPC may be very little changed from closed glottis values. The approach taken here is to accept Flanagan's wall impedance values and doubt Flanagan's estimates of glottal contribution to formant damping rather than the other way around. The matter of where the losses are located is a very serious matter. The Wakita acoustic tube model has all loss placed at the glottis and lossless walls. Area function estimates taken from synthetic speech where the glottal losses in the speech synthesizer are made large and the wall losses made small will give area function estimation accuracies that are overly optimistic.

The beneficial effect of the quenching of vocal tract oscillations that occurs when the glottis opens is not incorporated in the calculation of the spectrum of the synthetic speech sound. This quenching greatly improves the accuracy of formant bandwidth estimates. It should be

noted, however, that this quenching only occurs for the oscillations of low frequency formants - for high frequency formants the amount of glottal damping that occurs is greatly reduced by the increased impedance of glottal inductance. Even so, the errors observed from voicing periodicity will be overly pessimistic. It is better, however, to err on the side of pessimism and be conservative in the acceptance of area function estimation results than to err on the side of optimism and believe erroneous results to be correct.

Types of Error Sources That Can Be Simulated

A primary emphasis in the synthetic speech study of area function estimation is the selection of error sources that are to be incorporated. The first error source one has control of is the difference between vocal tract and LPC derived acoustic tube loss and boundary conditions. One can specify the loss and boundary conditions of the speech synthesis model to match the acoustic tube model of the analysis to remove this type of error. The vocal tract wall losses and radiation load can then be incorporated into the synthesis model to observe what changes occur in the area function estimate.

Other error sources concern the problem of estimating the vocal tract spectrum from the speech spectrum. Reducing the density of computed spectrum values can introduce the effects of voicing periodicity. The less dense the spectrum, the higher the apparent pitch frequency. The vocal tract spectrum can be modified by the spectral envelope of a

glottal pulse of the type described in Section IV, or one can remove this source of error by estimating area function directly from the vocal tract spectrum. The effect of a lack of high frequency information about area function in the spectrum can be incorporated by computing and analysing only the low frequency portion of the spectrum of the synthetic sound. The effects of the antialiasing filter used prior to sampling the speech signal have not been considered as high quality filters with steep cutoffs are available for use in speech analysis.

Procedure for Estimating the Vocal Tract Area Function

The procedure for estimating area function used is basically an extension of the method developed by Wakita [Ref. 25]. LPC analysis is performed on the speech signal. The LPC polynomial is factored into roots; the roots corresponding to narrow band poles of the LPC filter model are used to estimate the frequencies and bandwidths of vocal tract formants. The sampling rate that corresponds to an acoustic tube model having the required length is used to convert the formant frequencies and bandwidths back into LPC coefficients. The area function of the acoustic tube model is computed from the corrected set of LPC coefficients. This area function provides an estimate of vocal tract area function.

A flow diagram of how one starts with the known area function, synthesizes the speech spectrum, and then estimates the area function is shown in Figure 5.1. The vocal tract spectrum is evaluated at harmon-

ics of the pitch frequency. If a glottal pulse spectrum was applied to the vocal tract spectrum, a 12 dB/octave boost is applied to account for the combination of radiation characteristic and the preemphasis applied to flatten the speech spectrum. Taking this preemphasized speech spectrum, computing autocorrelation coefficients, and then computing LPC coefficients by the autocorrelation method is equivalent to having done covariance method LPC on a pitch period long frame of the speech waveform [Appendix C]. It should be remembered that preemphasis of the speech signal serves to improve the numerical accuracy of LPC analysis.

LPC poles corresponding to formants are identified by displaying the frequencies and dampings corresponding to complex conjugate poles and the dampings of real poles. These values are displayed on a computer terminal, and an interactive program permits one to specify which frequency and bandwidth values are those of the vocal tract formants. An automatic procedure for identifying the vocal tract formants was not developed as one needs to test such a procedure on real speech, not a predetermined set of vocal tract shapes, to verify the procedure as being effective. For the vocal tract shapes tested, it was found that one could get good area function estimates by considering those complex conjugate pole pairs having bandwidth values below 1000 Hz and by selecting those pole pairs having the four lowest frequency values as representing the first four formants. The selected LPC poles are then scaled to account for the sampling rate specified by the length of the acoustic tube model, LPC coefficients are computed from the poles, and then areas are calculated from the LPC coefficients.

Vocal tract length is estimated by employing a binary search procedure over the range of possible length values to find that length that gives the smoothest area function. The search estimates length to an accuracy of over a tenth of a per cent after ten iterations. The Newton-Raphson technique can give much faster convergence, but the type of variation of area function smoothness with length that occurs tends to make it unstable. The smoothest area function is one that minimizes the normalized energy in the second central difference of the area function values:

$$E = \sum_{i=2}^{N-1} \left(\frac{S_{i-1} - 2S_i + S_{i+1}}{S_i} \right)^2$$

The maximum length in the search procedure is limited to that which specifies a sampling rate equal to twice the highest formant frequency out of the formants selected for estimating area function. Suppose one were using four formants to estimate area function. If the acoustic tube length were too long, the fourth formant frequency would be greater than half the sampling frequency; an aliased image of the fourth formant would then be placed between the third and fourth formant.

Estimating the first four formant frequencies and bandwidths specifies the areas of an eight section tube. Suppose one doubled the sampling rate, cut the length of the sections in half, and wanted to specify the areas of a sixteen section tube. The first four formants can come from the speech signal because they are within the frequency range where the

formants in speech contain information about the area function. The next four formants are specified as having the frequencies that the acoustic tube would have if it were a uniform tube. These frequencies are at 9/32, 11/32, 13/32, and 15/32 of the sampling frequency. The bandwidths are specified as the average of the bandwidths of the first four formants. The maximum permissible length of the acoustic tube can be 1/8 more than before without having formant number four placed past formant number five at 9/32 of the sampling frequency.

Experimental Results

What follows is a discussion of experimental results for estimating area function from synthetic speech. Six vowels shapes from Fant [Ref. 3] are used, corresponding to the Russian vowels /A/, /O/, /U/, /I:/, /I/, and /E/. These vowels represent a progression from vowels with the tongue constriction in the back to those having the constriction in the front. Figures 5.2 thru 5.16 deal with the matter of having to estimate area function using only the first three or first four formants. The consequences of the lack of high frequency information in the speech spectrum about the area function are shown. All other error sources - radiation load at the lips, wall losses, high pitch frequency, and glottal spectrum - have been removed. The effects of these other error sources are shown in Figures 5.17 thru 5.25. In these cases area function is estimated with the sixteen coefficient acoustic tube model. Four of the eight formants needed come from the speech spectrum; the other four are specified as the formants of a uniform

tube. The acoustic tube length is set to the target vocal tract length. Figures 5.25 thru 5.28 show the effects of all error sources combined - radiation load, wall loss, voicing periodicity, glottal spectrum, and having to estimate the length of the vocal tract. The Wakita formulation acoustic tube is used throughout these experiments except where otherwise noted. The formant frequencies and bandwidths that were estimated in all of these experiments are summarized in Table 5.1, and the results of estimating vocal tract length are shown in Table 5.2.

Figure 5.2 shows plots of estimated area function and estimated vocal tract spectrum compared against their target values for the vowel /A/. Target values are shown in solid lines, estimated values in dashed lines. The base ten logarithm of area function is shown - the value zero corresponds to one square centimetre area. The estimated areas have been scaled to match the scale of the target area function. The vocal tract loss used in synthesizing the vocal tract spectrum is based on a glottal resistance value equivalent to the characteristic impedance of a lossless tube that is .7 cm^2 in diameter. The sampling rate used in the LPC analysis is matched to the vocal tract length. The top plots in the figure correspond to estimating the area function with an eight section tube; the bottom plots correspond to estimating area function with a sixteen section tube where uniform tube values are specified for the high order formants.

Figure 5.3 is based on the same vowel sound, using the sixteen section acoustic tube. This time a fixed sampling rate of 12 KHz is used in

the LPC analysis, and then the LPC poles corresponding to formants are scaled to match the sampling rate specified by vocal tract length. In the top plot, the correct vocal tract length has been specified. The area function estimate is unaffected by going to a fixed sampling rate and scaling the poles rather than using a variable sampling rate matched to length. The estimated spectrum matches the first four formants of the target spectrum quite well. The bottom plot shows the effect of estimating vocal tract length by finding the smoothest area function estimate. This procedure has introduced error in both the area function and spectrum estimates because vocal tract length is underestimated. The amplitude of fourth formant relative to first formant is about 3 dB too low.

Figure 5.4 shows results for the vowel /O/. In the case where the LPC sampling rate is matched to vocal tract length, the fourth formant of the speech spectrum is past 1/2 the sampling rate. A uniform tube of the same length would have its fourth formant at only 7/16 the sampling rate. In the case of variable rate sampling one can only specify a six section tube as in the top plots, or a twelve section tube as in the bottom plots where the number of formants is doubled by filling in uniform tube formants. With fixed rate sampling, one can specify the sampling rate high enough to locate the first four formants. The procedure of filling in extra formants allows enough leeway that the fourth formant is not placed past where the fifth formant should go. The area function estimate using a sixteen section tube is shown in the top of Figure 5.5. An eight section tube based on four formants cannot be used as an acoustic tube of the correct length would alias an image

of the fourth formant between the third and fourth formant. The bottom plot of Figure 5.5 demonstrates the need for using the first four formants in estimating areas. In this case not enough LPC coefficients were used so the fourth formant was skipped over and the fifth formant was used in the area function estimation. Note the large errors in the estimated area function.

Figure 5.6 shows the effect of having to estimate vocal tract length. In the top plot, the correct length was specified, and in the bottom plot, length was estimated. Note that there is some degradation of the estimate.

Figure 5.7 shows variable sampling rate results for the vowel /U/ , and Figure 5.8 shows fixed sampling rate results. Here also, one can only estimate three formants in the variable sampling rate results. The formant pattern reflects the heavy departure from a uniform tube that this vocal tract shape represents. In estimating vocal tract length, the length was rather heavily underestimated and errors in formant amplitude of up to 6 dB are made in the estimated spectrum as a result. The shape of the vocal tract estimate, however, is still correct though shortened.

Figure 5.9 gives variable sampling rate results for /I:/; Figure 5.10 gives fixed sampling rate results. The results here are quite good; but the very narrow second formant should be noted. This formant will be the source of trouble later on. Figures 5.11 thru 5.13 give results for /O/. In all the cases where uniform tube formants are specified, the lip opening is too small. Figure 5.13 gives fixed sampling results

for using only the first three formants instead of the first four. Where vocal tract length is specified, the results are quite improved. Where length has to be estimated, the area function estimate fails completely. Figures 5.14 thru 5.16 give results for /E/. In Figure 5.15 one sees that in going from fixed sampling rate length specified to fixed sampling rate length estimated, the area function estimate deteriorates. Figure 5.16 compares this last result with estimating the area function on the eight section tube. The vocal tract length is overestimated by the same amount as length is underestimated with the sixteen section tube, but the resulting area function and spectrum estimates are much improved.

The sixteen section acoustic tube does not provide any big gains in accuracy over the eight section tube. The motivation for using the sixteen section tube is that the results are more consistent. For some vowel sounds, the formant spacing departs from the uniform tube spacing so much that specifying the correct length on the eight section tube would give the wrong spectrum, introducing an aliased formant. The sixteen section tube works for the use of four formants in all the cases tested.

Figures 5.17 thru 5.19 address the question of glottal spectrum and voicing periodicity. The top plots show results for a pitch frequency of 120 Hz and impulse excitation of the vocal tract; the bottom plots show results for 120 Hz pitch and glottal pulse excitation combined with radiation and preemphasis spectrum boost. Most of the degradation comes from the voicing periodicity; adding the glottal pulse effect

shows little added degradation. The vowels /U/ and /I:/ show the most degradation. These vowels have the constriction located more in the center; these shapes are most symmetric of the six shapes. Losses affect the formant bandwidths most. Antisymmetric vowels tend to be less sensitive to these bandwidth errors as the Atal-Hanauer and Wakita acoustic tube formulations, having completely different loss distributions, give the same estimated area function. Using impulse excitation and a pitch of 240 Hz begins to affect the more symmetric shapes /A/, /O/, /U/, and /E/ - this is shown in Figure 5.20.

Figures 5.21 thru 5.23 show results of vocal tract loss and boundary conditions. Voicing periodicity, glottal pulse, and length estimation errors are eliminated so as to concentrate on errors originating in the acoustic tube model formulation. The top plots come from having heat conduction, viscosity, and soft wall losses. The glottal loss has been reduced to a very small value - the loss resulting from the characteristic resistive impedance of a lossless acoustic tube of area .05 cm^2. There is some degradation of the area function estimates, notably /I:/ and /E/. The bottom plots show results for adding the vocal tract radiation load to the speech synthesis model. The degradation is much larger, changing mouth opening and shifting the point of constriction. The vowel /I:/ suffers most. Checking the formant estimates in Table 5.1, one sees that both the frequency and bandwidth of the second formant of this vowel are changed a great deal. With the radiation load acting on this vowel, one no longer has the correct formant values to estimate the correct area function.

In Figure 5.24, the Atal-Hanauer acoustic tube was applied to the more nearly antisymmetric vowels where the synthesis model has both wall losses and a radiation load at the lips. The radiation load loss is not frequency independent as the lip loss in the Atal-Hanauer model is, but it could be argued that the Atal-Hanauer model might be appropriate given the degree to which glottal loss has been reduced and the fact that loss is now occurring at the lips. The Atal-Hanauer acoustic tube, however, does not seem to give results that are better than the Wakita model, but the results are not much worse either. In the top of Figure 5.25, wall loss and radiation load were retained, but glottal loss was greatly increased to see if sufficient glottal loss would improve Wakita formulation estimates. This method of cheating can improve the results for /U/, but /I:/ resists this treatment for improving one's data. In the bottom plots, glottal loss was eliminated and the Atal-Hanauer formulation was used in the area function estimate. The resulting closed glottis condition formants gave an improved estimate for /I:/, but /U/ was left in very bad shape.

Figures 5.26 thru 5.28 show results for including everything - 120 Hz pitch, the effects of the glottal pulse, length estimation, wall loss, and vocal tract radiation load. The lip opening came out too small in all cases except for /E/. The resulting area function estimates for the most part still look like the shapes of the vowel sounds they belong to, though the estimates do not track the target shapes very well at all. The estimation procedure fails for the vowel /I:/. This vowel is sensitive to the kinds of changes in formant bandwidth that come from the various error sources.

Conclusions

Can one estimate vocal tract area function accurately from the speech waveform? The answer is, sometimes. First of all, the LPC derived estimates are restricted to non-nasalized, voiced speech sounds. At least being able to analyse these sounds is a start. One cannot determine the area function at a speech stop, but one may be able to infer the articulation of a stop by observing succeeding frames of vowel articulation before and after the stop. Even for vowel sounds, however, the answer is still sometimes, as there are vowel sounds for which the method fails.

For the most part, one can estimate the general shape of a vowel correctly. The area function estimate generally gives a distorted version of the vocal tract shape, yet a shape that one could identify as being associated with the vowel sound in question.

The most serious sources of error are the effects of voicing periodicity and the radiation load at the lips of the vocal tract. The effects of the glottal spectrum envelope and the wall losses are of much less importance. The estimation of vocal tract length can give trouble on occasion - some fine tuning of the criterion for selecting the estimated area function of maximum smoothness may help.

The problem of the lip radiation load not being represented properly in any of the acoustic tube model formulations that exist is difficult to deal with until other formulations are discovered. The problem of the

voicing periodicity errors in formant bandwidth estimation is something that one ought to be able to solve. LPC analysis is really not configured for use on periodic waveforms. Some of the pre-LPC analytical techniques such as Miller's Pitch Synchronous Analysis (PISA) [Refs. 46,47,54] avoid some of the sources of systematic error in LPC. These earlier procedures have been for the most part abandoned because LPC requires much less computational complexity. These procedures should perhaps be reexamined.

Table 5.1

Formant Frequency and Bandwidth Estimates

Legend:

f_s sampling frequency (Hz)

F0 fundamental (pitch) frequency (Hz)
N number of LPC coefficients used
a impulse excitation
b glottal pulse (50% duty cycle) with fixed preemphasis
c viscosity, heat conduction, soft wall loss; reduced glottal loss
d radiation load at the lips

	f_s	F0	N	abcd	F1	B1	F2	B2	F3	B3	F4	B4
A	8000	<30	8	x	670	416	1083	162	2489	83	3694	128
	12000	30	16	x	673	391	1087	152	2486	90	3714	121
	12000	120	16	x	667	380	1085	141	2498	100	3713	101
	12000	120	16	x	651	258	1078	154	2503	95	3711	95
	12000	240	16	x	714	363	1096	251	2451	146	3711	658
	12000	30	26	x x	663	100	1129	60	2501	47	3666	35
	12000	30	16	x xx	615	105	1034	102	2418	121	3500	248
	12000	120	18	xxx	607	66	1017	131	2422	106	3490	242
O	7368	<30	8	x	518	233	876	82	2391	38	----	---
	12000	30	16	x	502	194	871	72	2394	37	3498	624
	12000	30	12	x	548	511	849	145	2405	45	3769	130
	12000	120	16	x	503	207	859	81	2400	20	3512	567
	12000	120	16	x	520	149	853	70	2401	22	3639	612
	12000	240	16	x	512	137	916	132	2400	9	3499	587
	12000	30	16	x x	510	70	889	47	2404	38	3452	61
	12000	30	16	x xx	476	79	818	70	2361	67	3439	91
	12000	120	18	xxx	483	37	852	49	2378	84	3448	106

	f_s	F0	N	abcd	F1	B1	F2	B2	F3	B3	F4	B4
U	7000	<30	8	x	231	64	590	54	2365	21	----	----
	12000	30	16	x	231	59	589	52	2368	19	3782	682
	12000	120	16	x	239	31	598	29	2388	53	3821	581
	12000	120	16	x	247	22	600	20	2388	55	3899	167
	12000	30	16	x x	231	46	595	41	2388	55	3694	51
	12000	30	16	x xx	219	49	589	44	2385	55	3695	59
	12000	120	18	xxx	245	21	600	16	2398	38	3707	66
I:	7180	<30	8	x	283	49	1613	8	2333	123	3383	354
	12000	30	16	x	284	53	1616	14	2345	118	3390	217
	12000	120	16	x	274	79	1600	82	2342	164	3394	220
	12000	120	16	x	278	69	1604	74	2350	166	3401	223
	12000	30	16	x x	285	48	1623	43	2343	41	3379	42
	12000	30	18	x xx	278	49	1413	171	2325	58	3349	91
	12000	120	18	xxx	271	61	1425	189	2320	92	3356	76
I	8236	<30	8	x	225	51	2282	233	3183	123	3643	869
	12000	30	16	x	225	53	2288	211	3178	118	3670	788
	12000	120	16	x	238	30	2290	204	3173	145	3547	795
	12000	120	16	x	246	21	2305	229	3174	140	3602	628
	12000	240	16	x	261	179	2241	234	3151	137	3781	513
	12000	30	16	x x	227	47	2277	47	3179	57	3737	70
	12000	30	16	x xx	223	50	2264	73	2868	502	3654	165
	12000	120	18	xxx	245	18	2775	55	2828	438	3650	182
E	8236	<30	8	x	427	65	2023	186	2928	428	3773	489
	12000	30	16	x	426	61	2015	190	2902	462	3693	549
	12000	120	16	x	432	98	2017	183	2894	463	3685	574
	12000	120	16	x	451	78	2005	191	2867	470	3678	548
	12000	240	16	x	467	68	1988	229	2881	544	3715	605
	12000	30	16	x x	428	50	1996	52	2863	57	3739	52
	12000	30	16	x xx	409	55	1939	131	2746	390	3389	672
	12000	120	18	xxx	421	83	1928	121	2725	328	3364	652

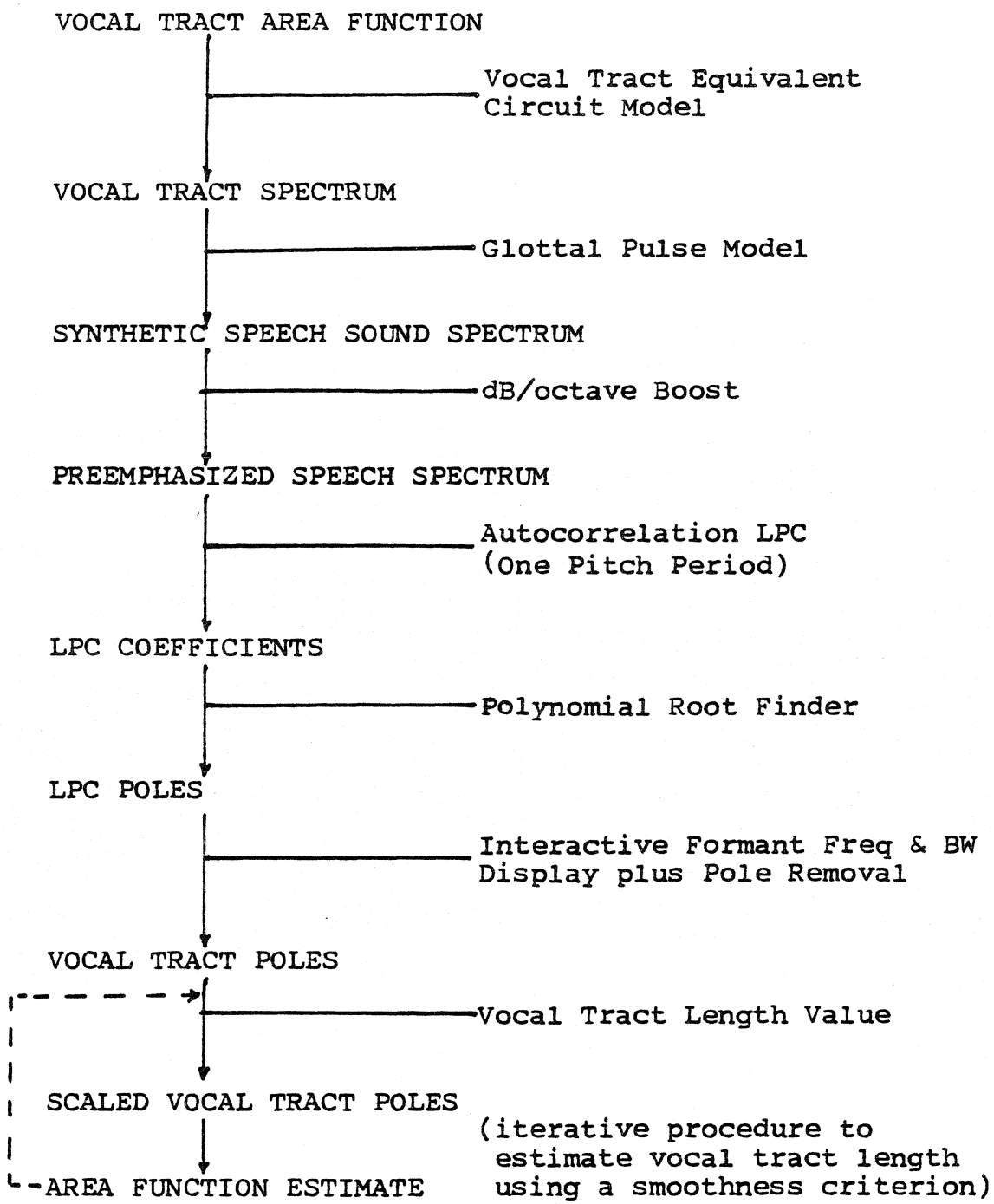
Table 5.2

Vocal Tract Length Estimation

Estimated Length and Per Cent Error in Estimated Length

	Actual length	4 formant		3 formant		4 formant, eight section tube		4 formant, other error sources	
		L	%	L	%	L	%	L	%
A	17.5	16.7	-4.6					18.0	2.9
O	19.0	18.3	-3.7					18.0	-5.3
U	20.0	17.7	-11.5					17.4	-13.0
I:	19.5	18.6	-4.6					18.1	-7.2
I	17.0	16.5	-2.9	12.9	-21.8			16.5	-2.9
E	17.0	16.4	-3.5			17.5	2.9	18.1	6.5

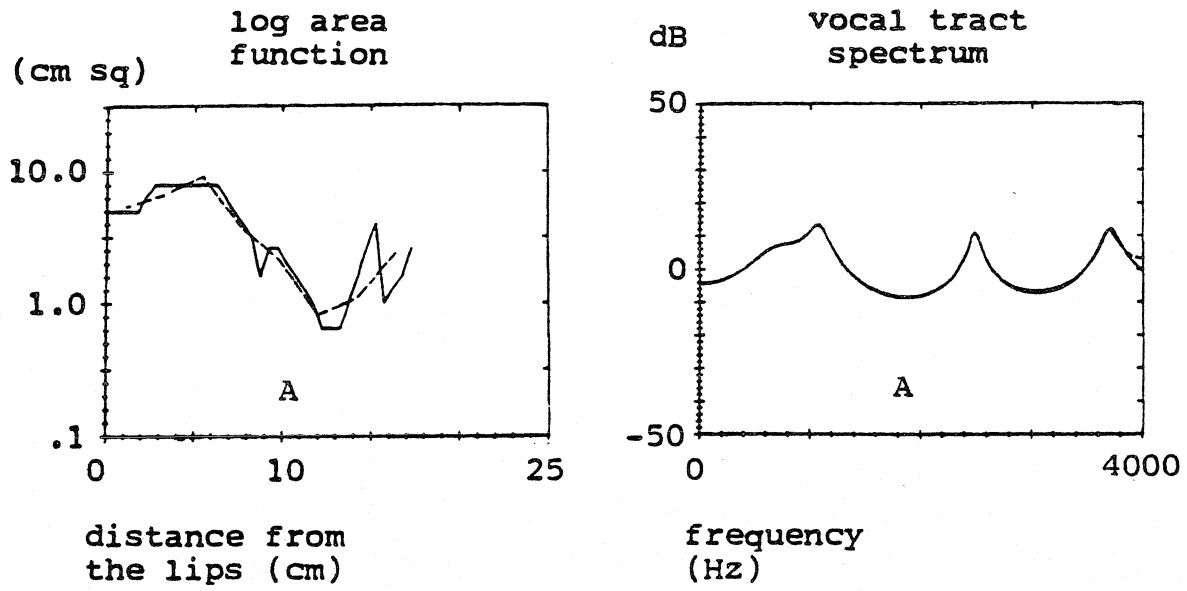
(lengths in cm)



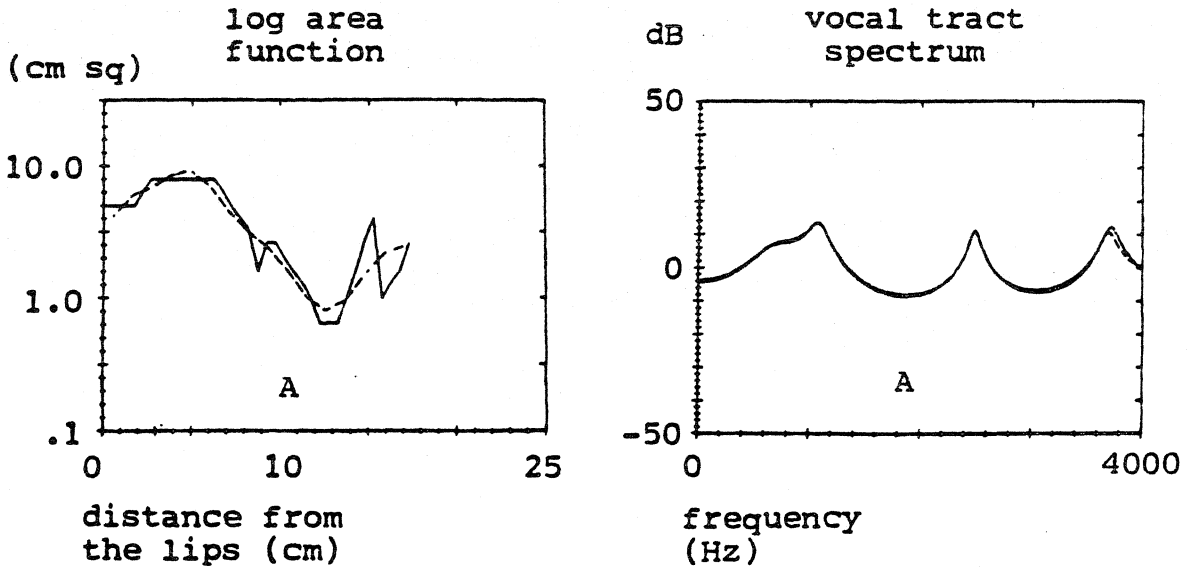
Procedure for estimating the area function from synthetic speech sounds.

Figure 5.1

8 section acoustic tube:

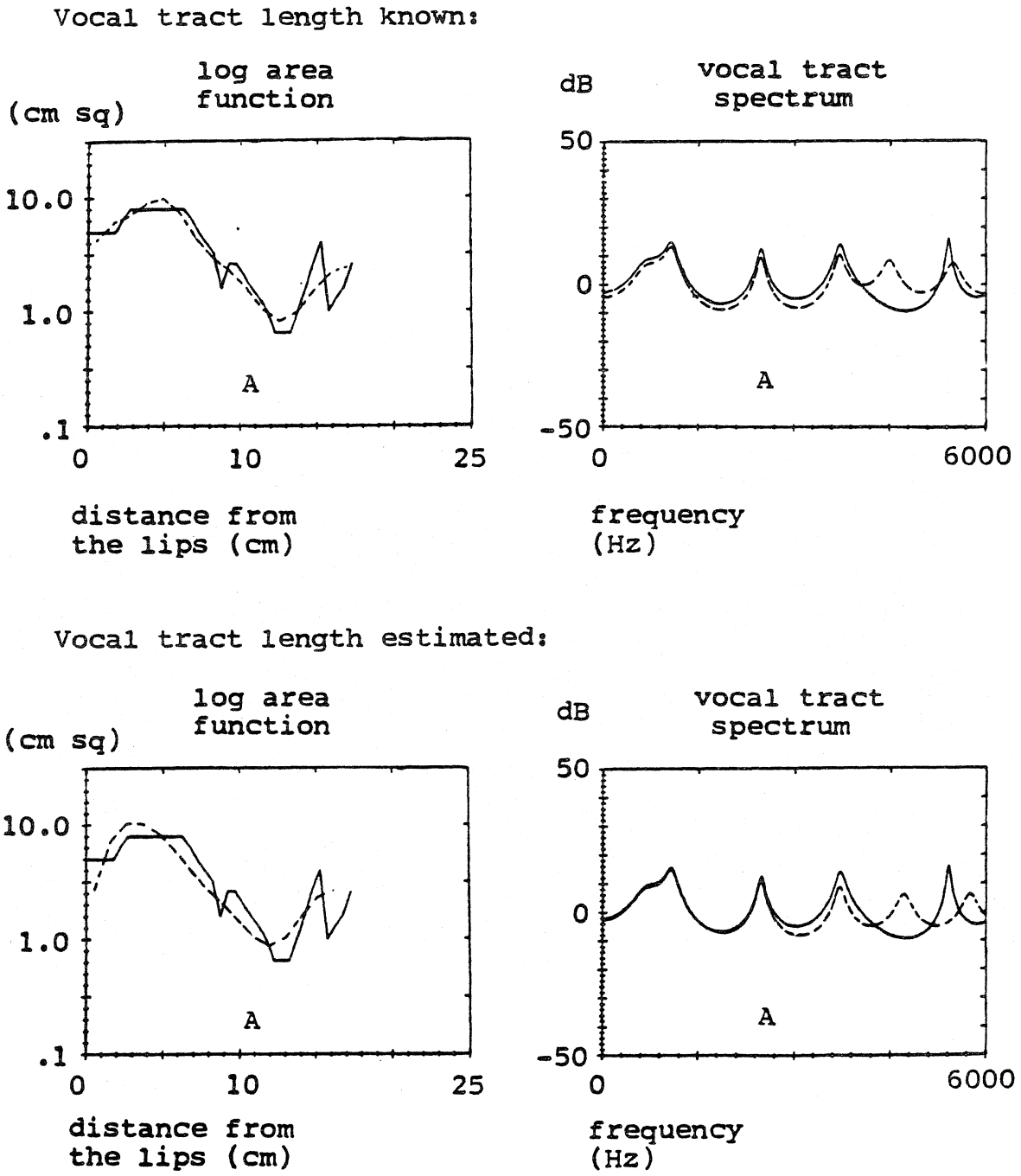


16 section acoustic tube:



Sampling rate matched to vocal tract length.

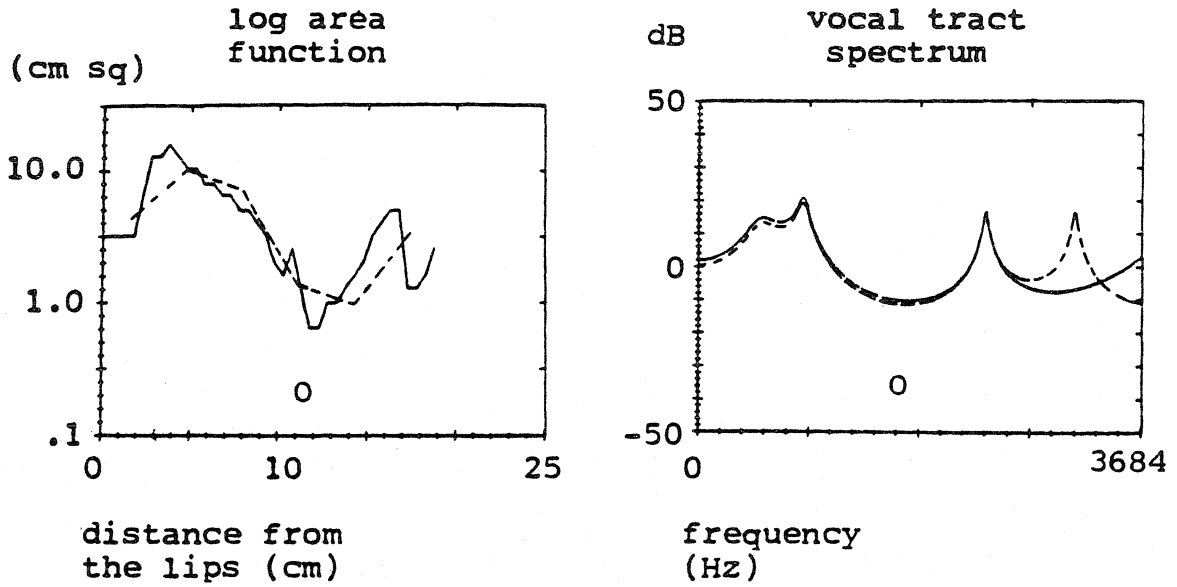
Figure 5.2



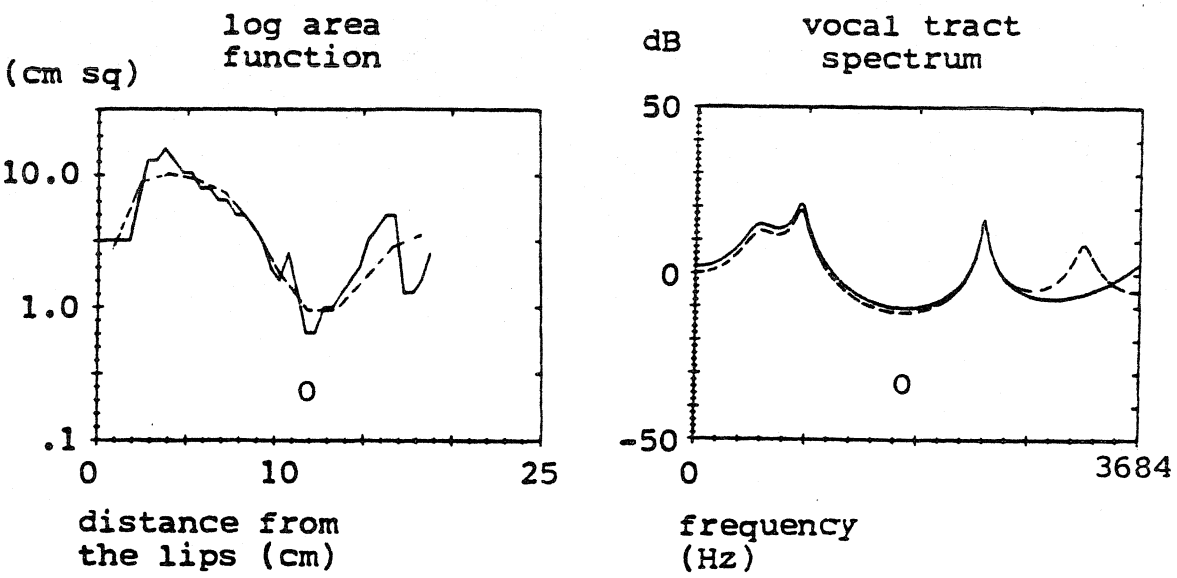
Fixed sampling rate, 16 section acoustic tube.

Figure 5.3

6 section acoustic tube:



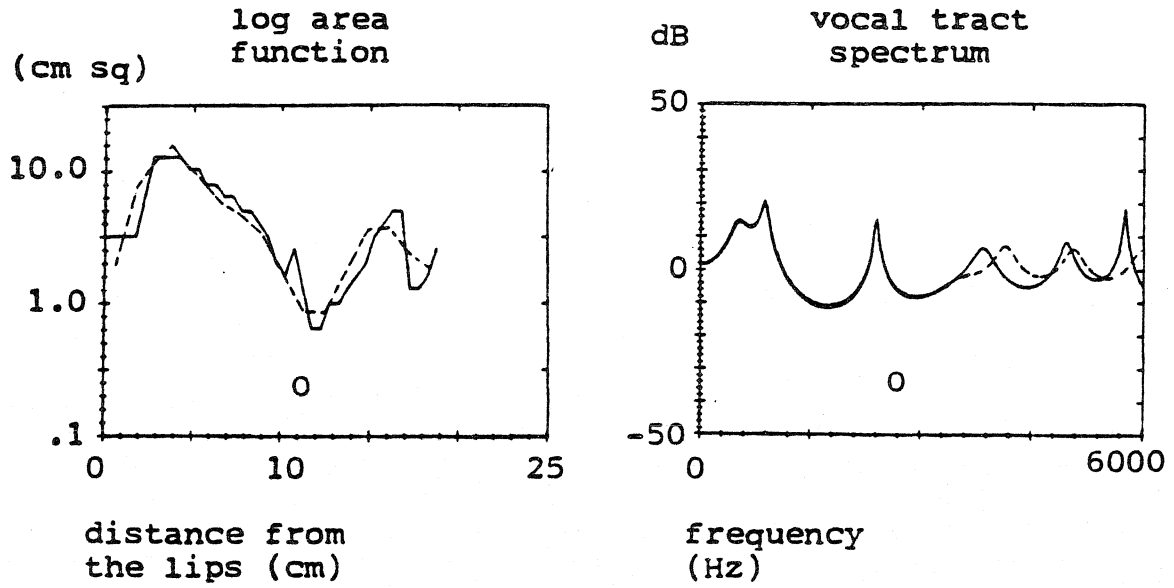
12 section acoustic tube:



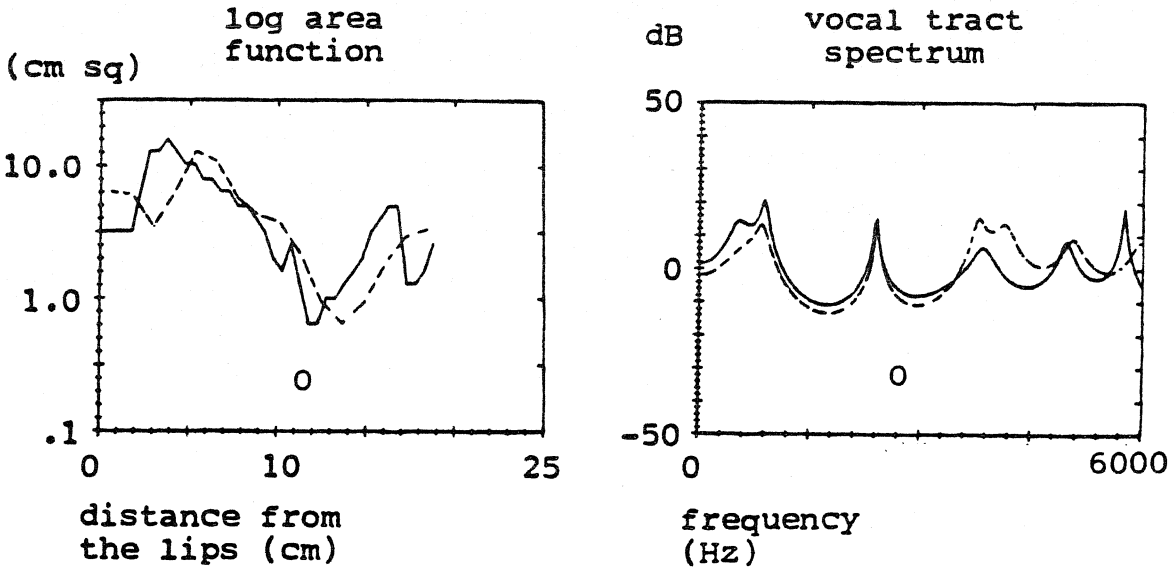
Sampling rate matched to vocal tract length.

Figure 5.4

16 coefficient LPC analysis:



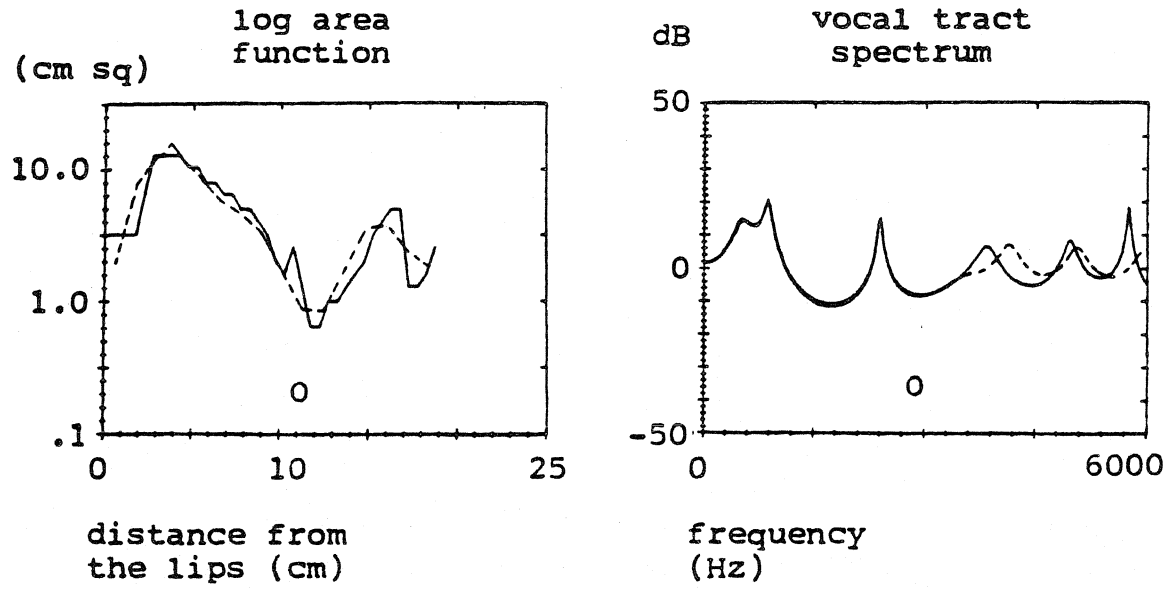
12 coefficient LPC analysis:



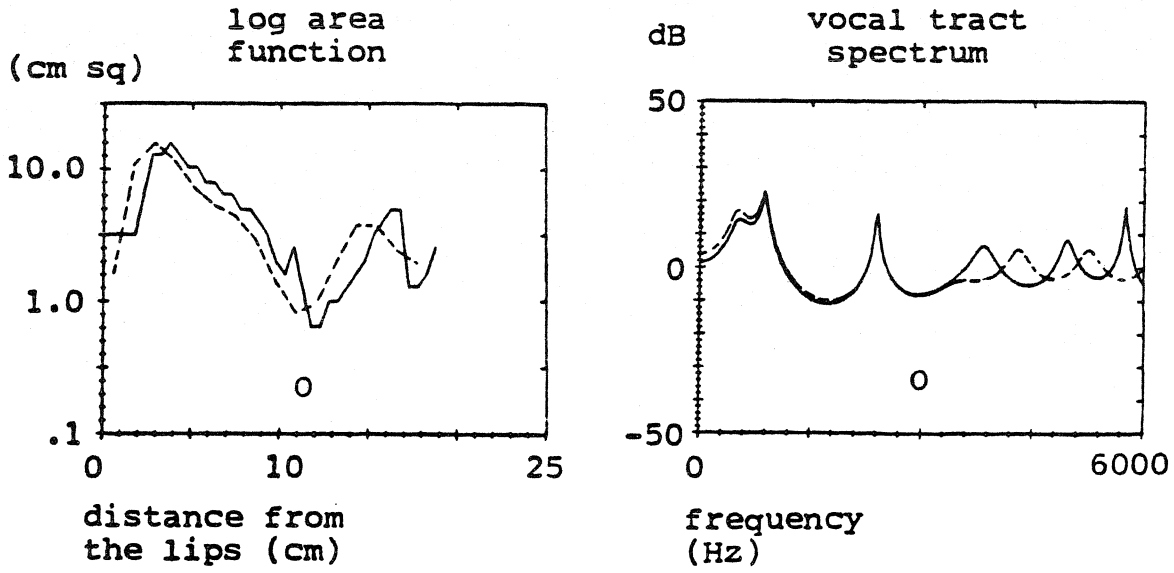
Fixed sampling rate, 16 section acoustic tube, vocal tract length known.

Figure 5.5

Vocal tract length known:



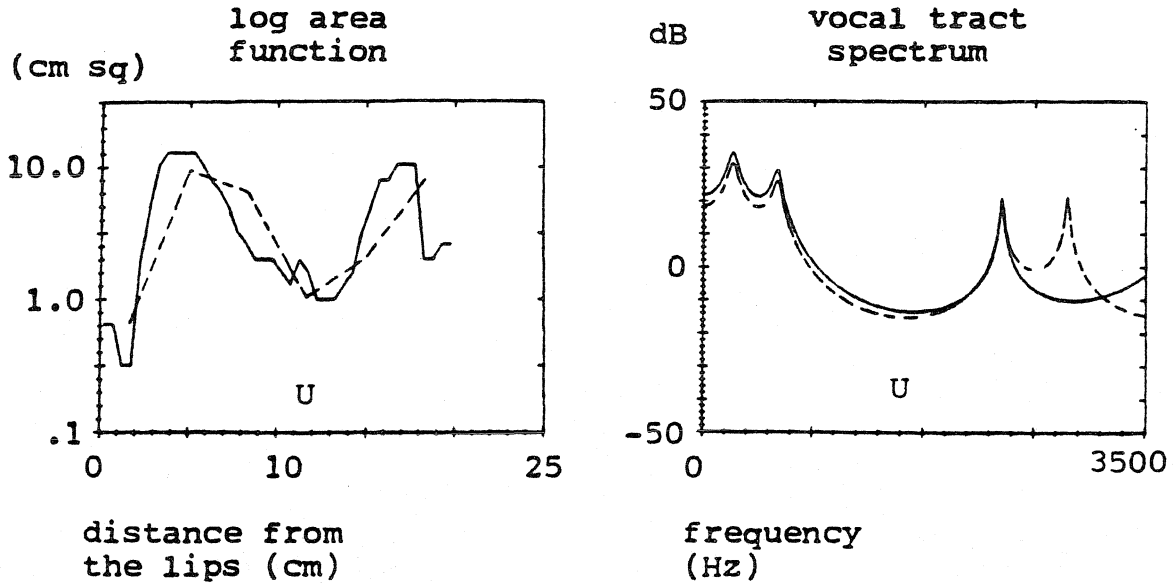
Vocal tract length estimated:



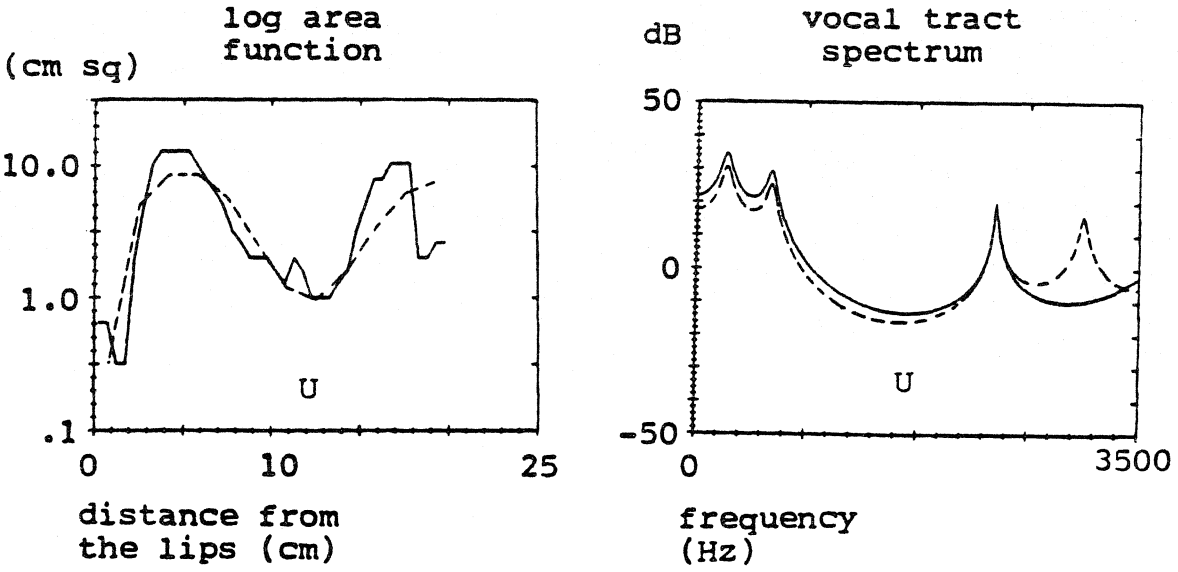
Fixed sampling rate, 16 section acoustic tube.

Figure 5.6

6 section acoustic tube model:



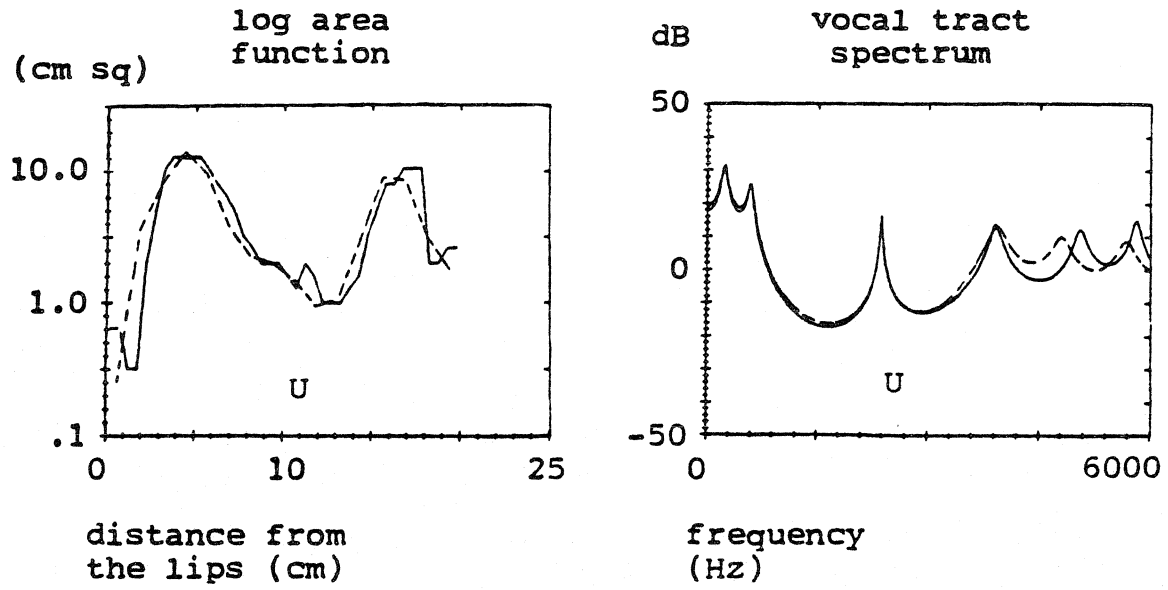
12 section acoustic tube model:



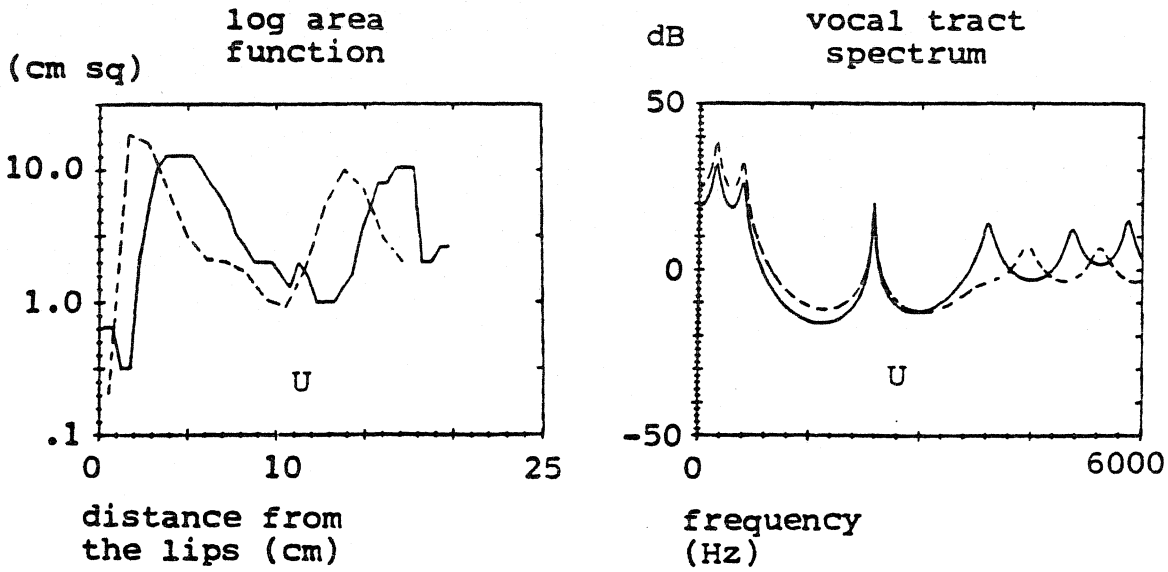
Sampling rate matched to vocal tract length.

Figure 5.7

Vocal tract length known:



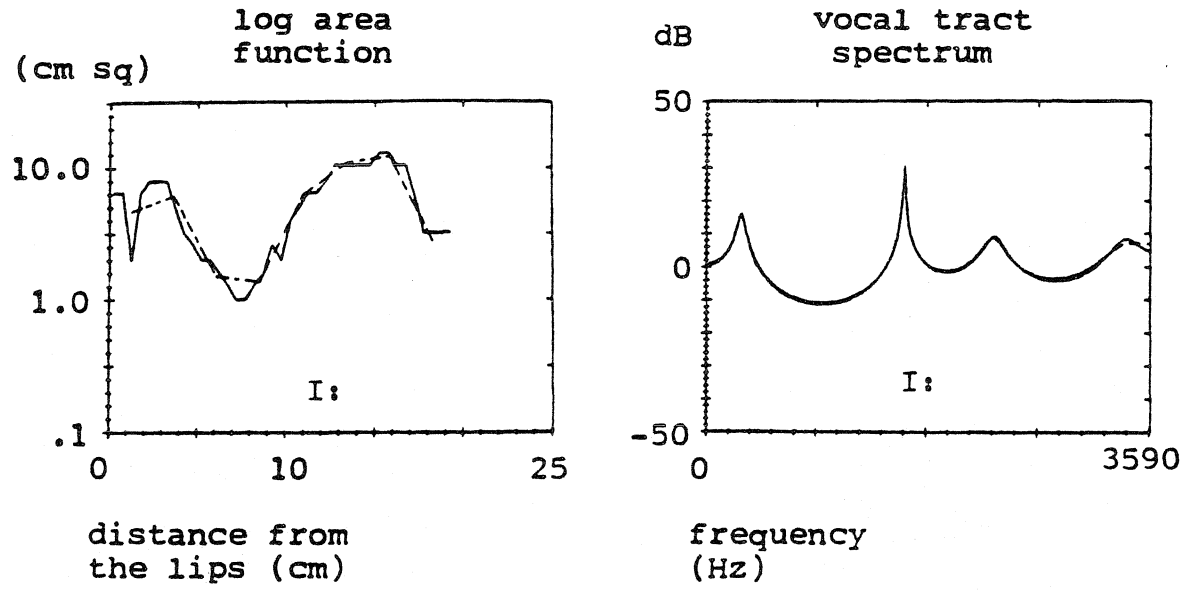
Vocal tract length estimated:



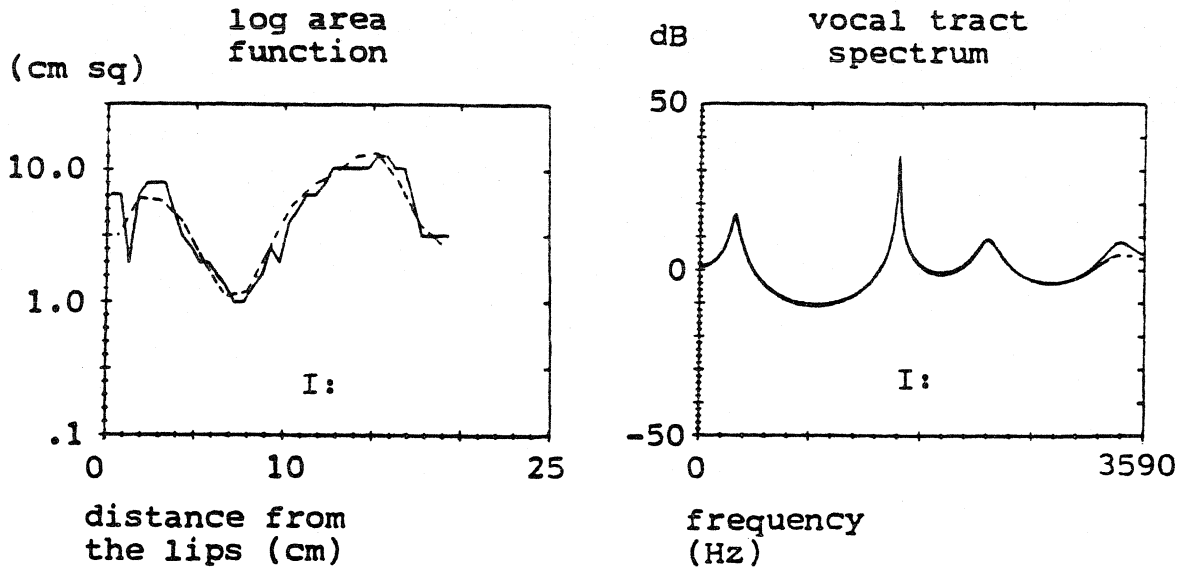
Fixed sampling rate, 16 section acoustic tube.

Figure 5.8

8 section acoustic tube:



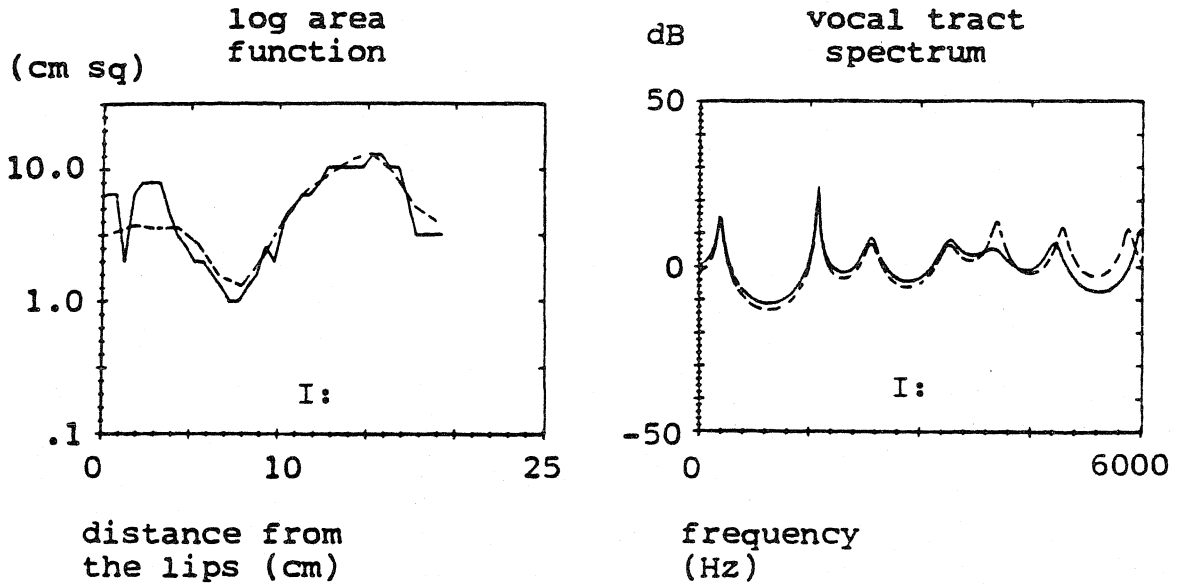
16 section acoustic tube:



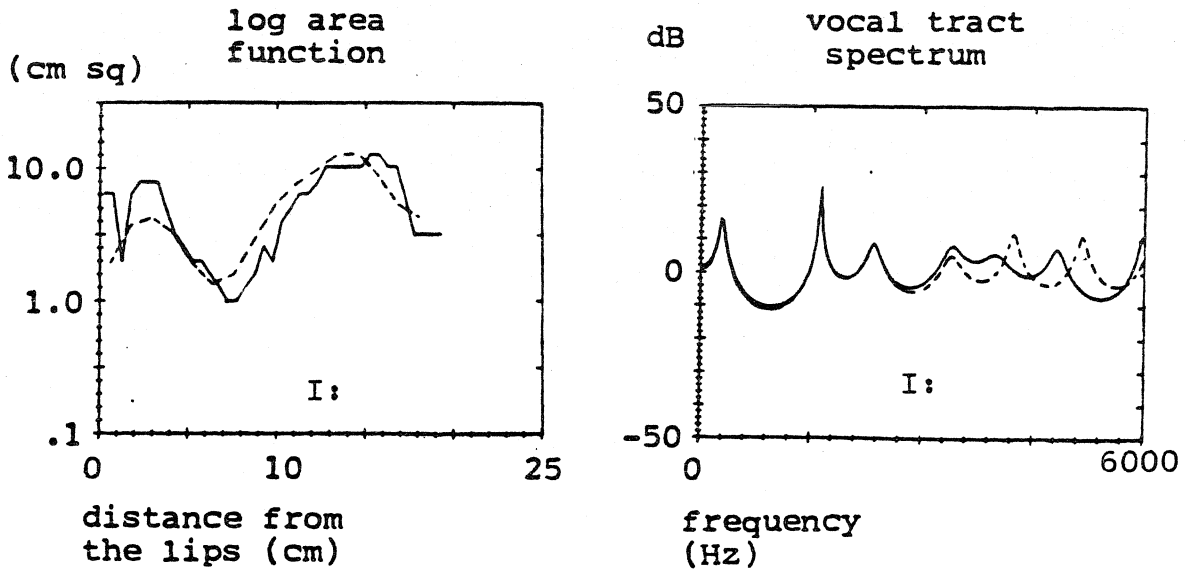
Sampling rate matched to vocal tract length.

Figure 5.9

Vocal tract length known:



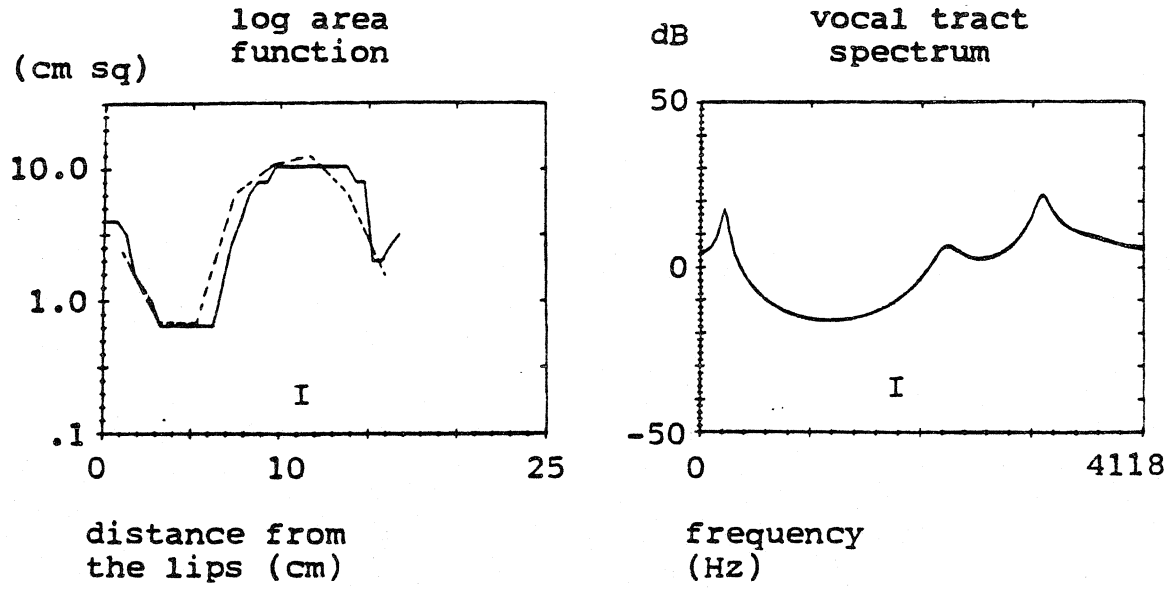
Vocal tract length estimated:



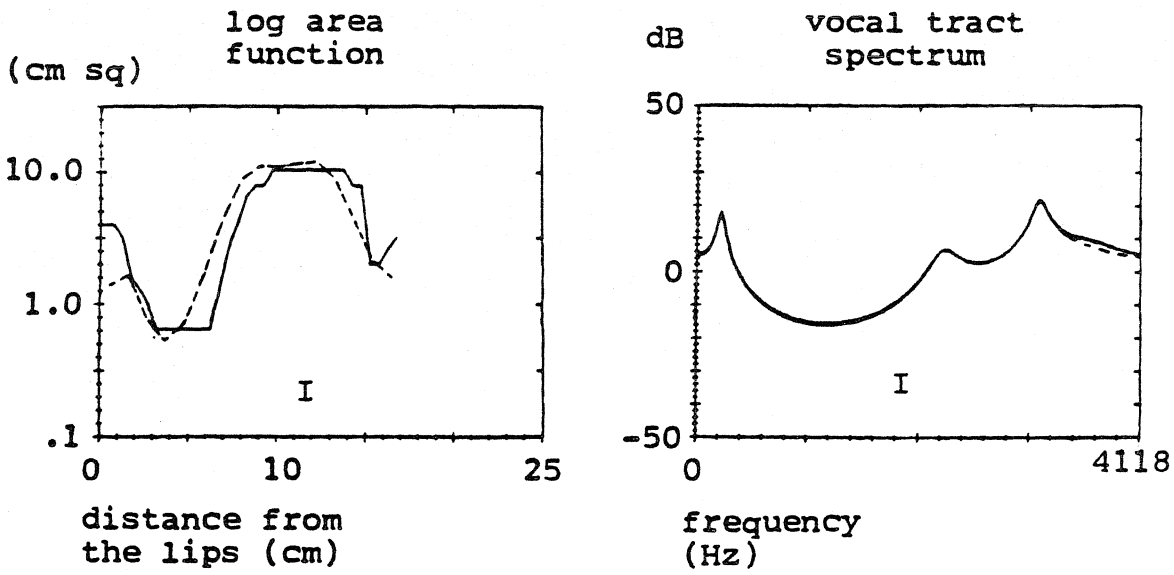
Fixed sampling rate, 16 section acoustic tube.

Figure 5.10

8 section acoustic tube:



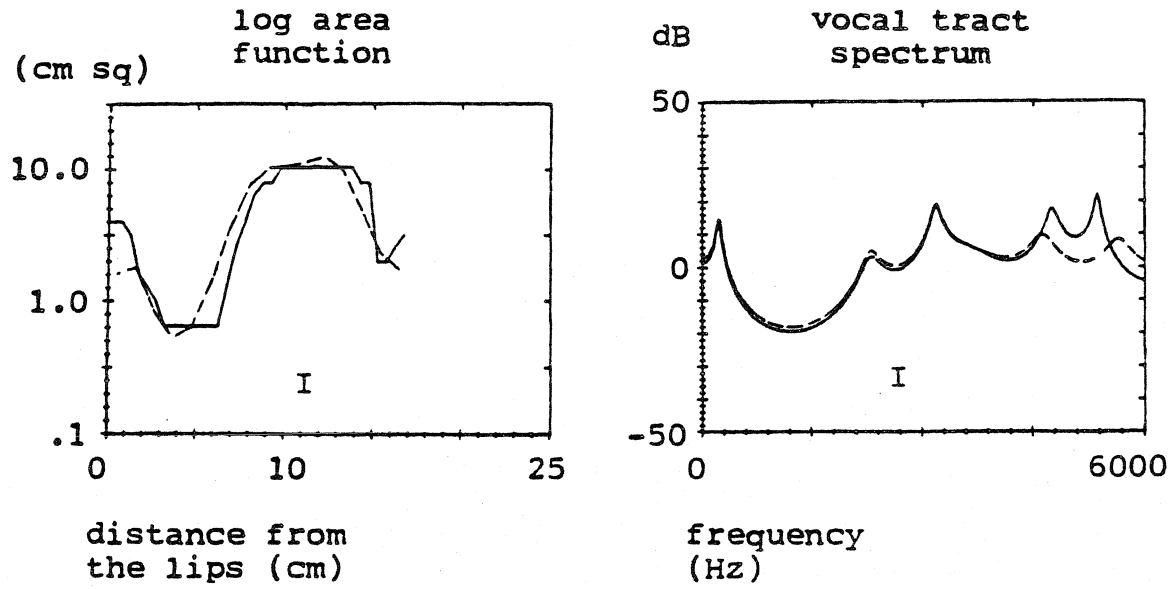
16 section acoustic tube:



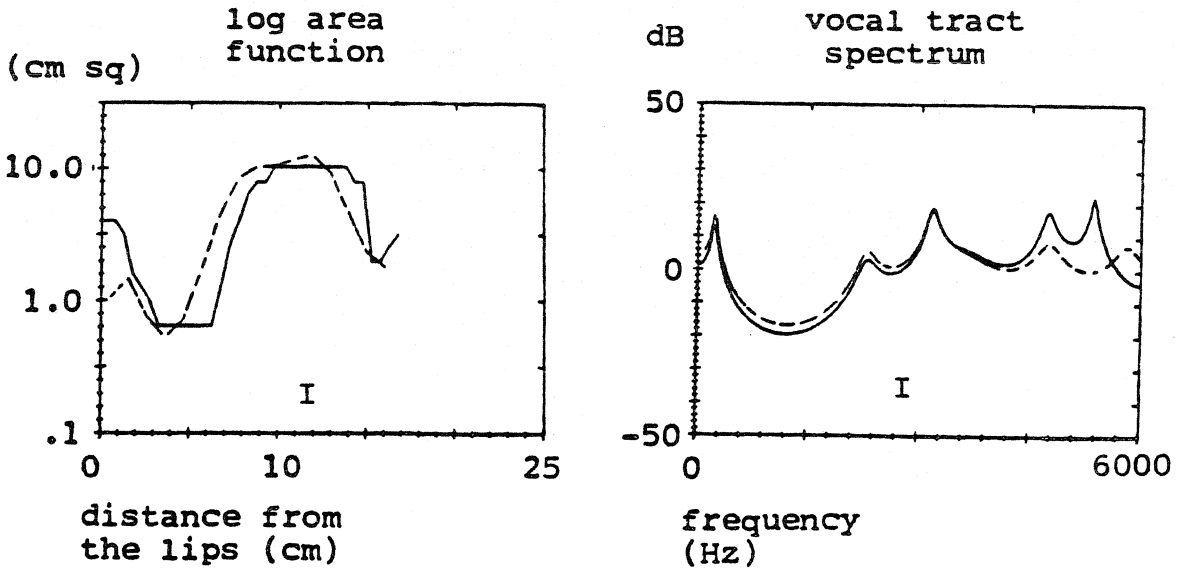
Sampling rate matched to vocal tract length.

Figure 5.11

Vocal tract length known:



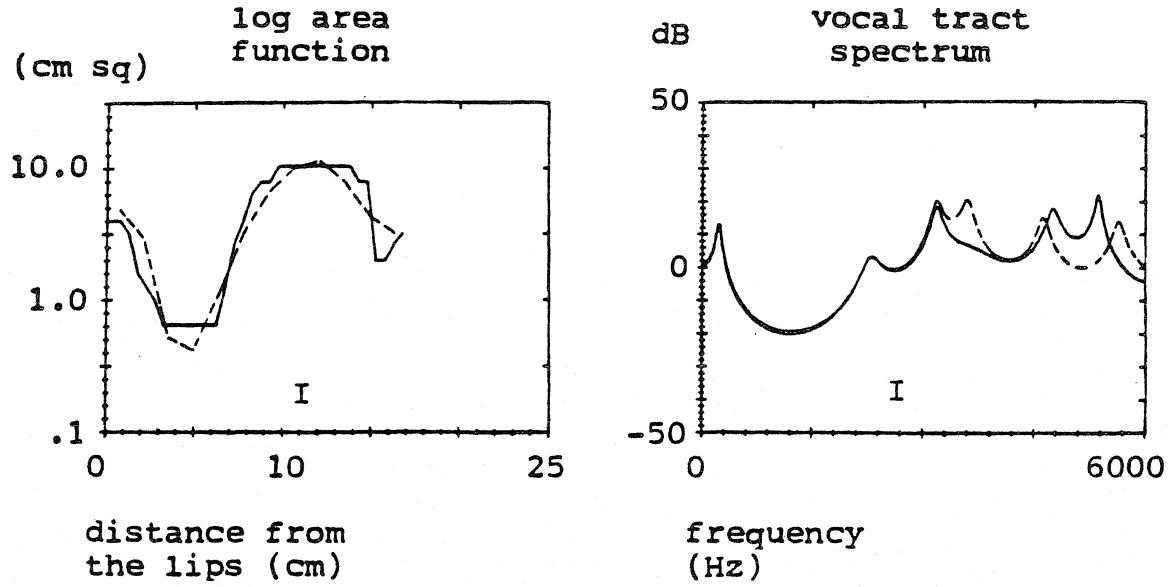
Vocal tract length estimated:



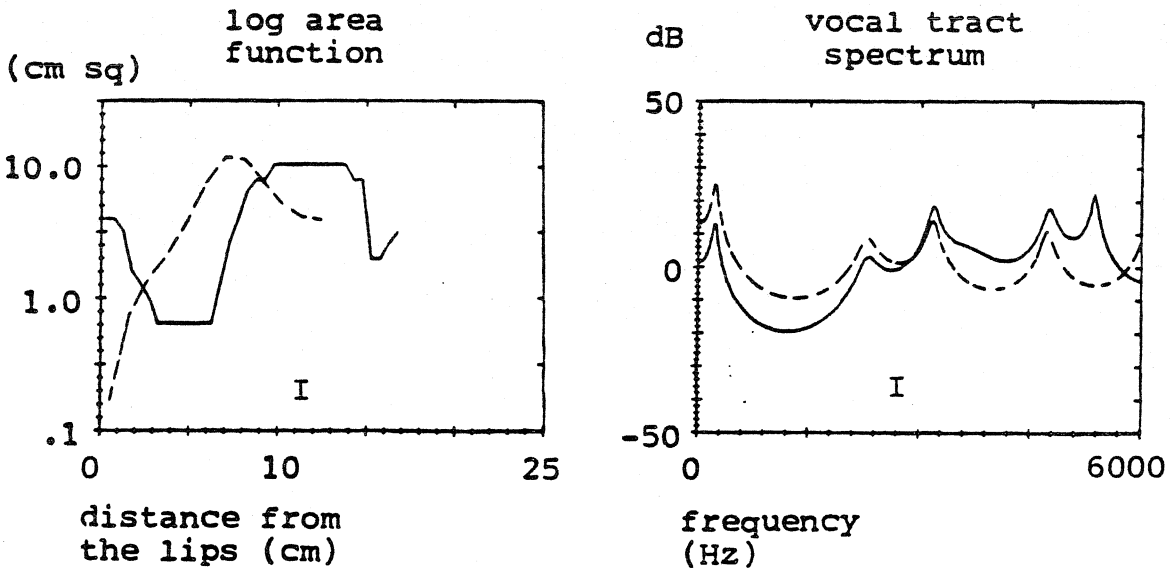
Fixed sampling rate, 16 section acoustic tube.

Figure 5.12

Vocal tract length known:



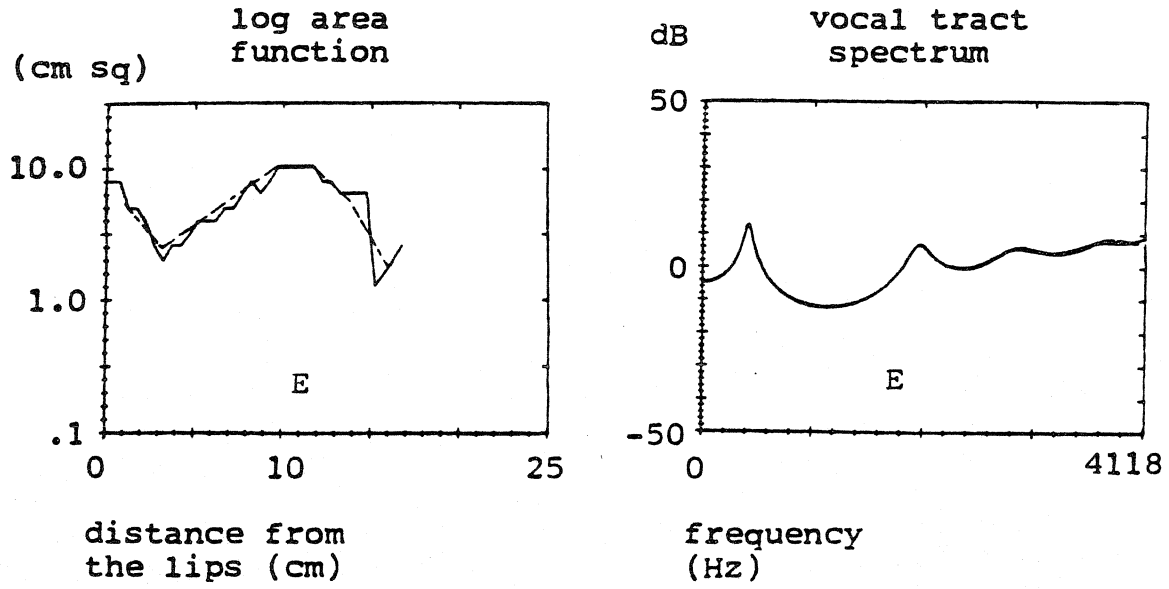
Vocal tract length estimated:



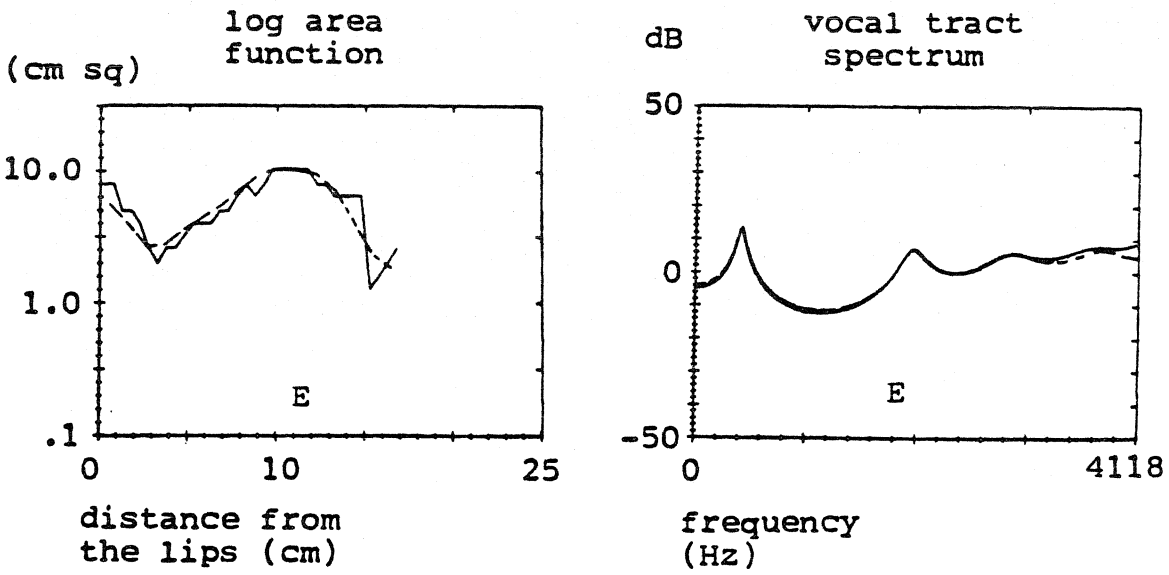
Fixed sampling rate, 12 section acoustic tube - area function estimated from three formants.

Figure 5.13

8 section acoustic tube:

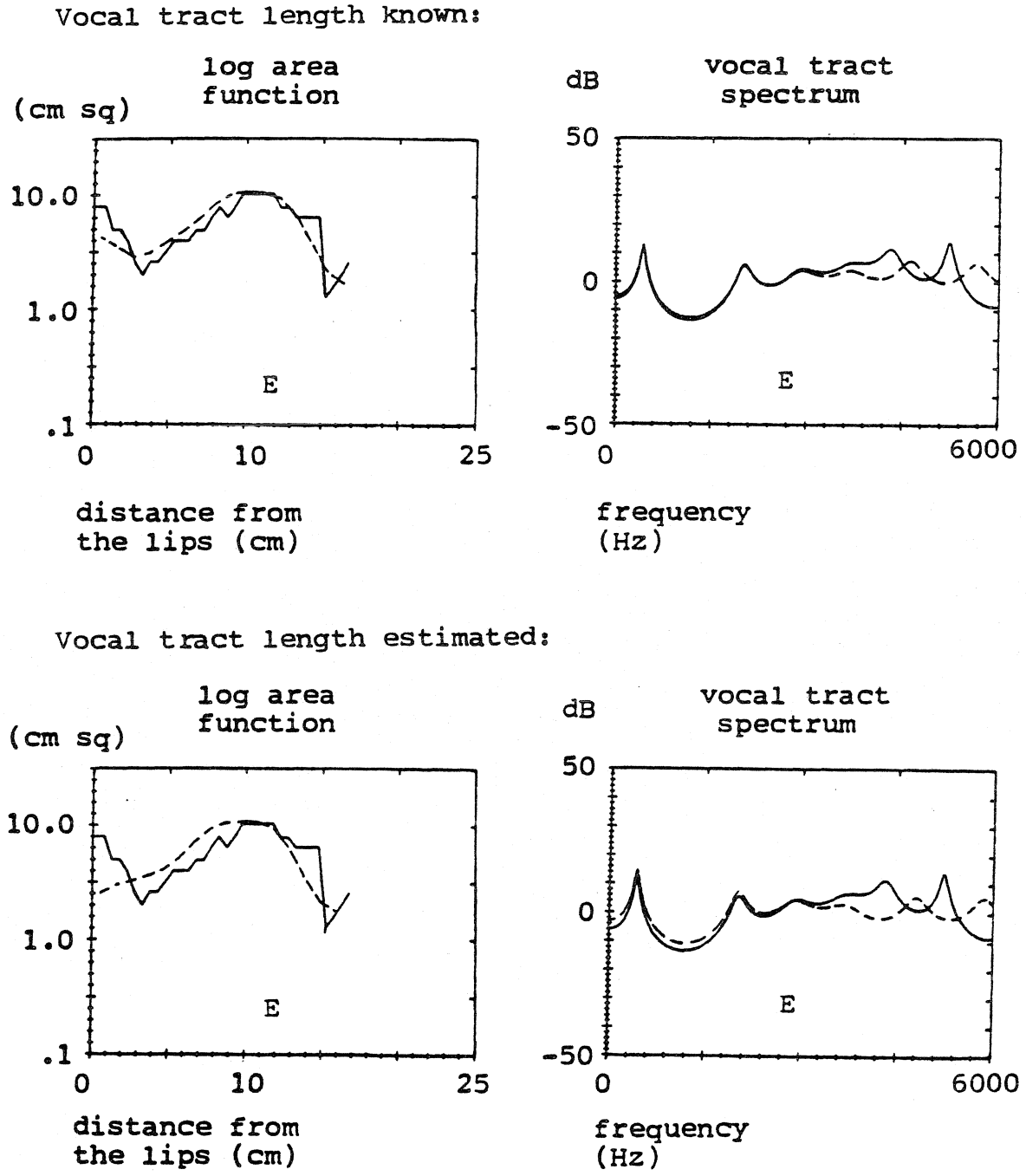


16 section acoustic tube:



Sampling rate matched to vocal tract length.

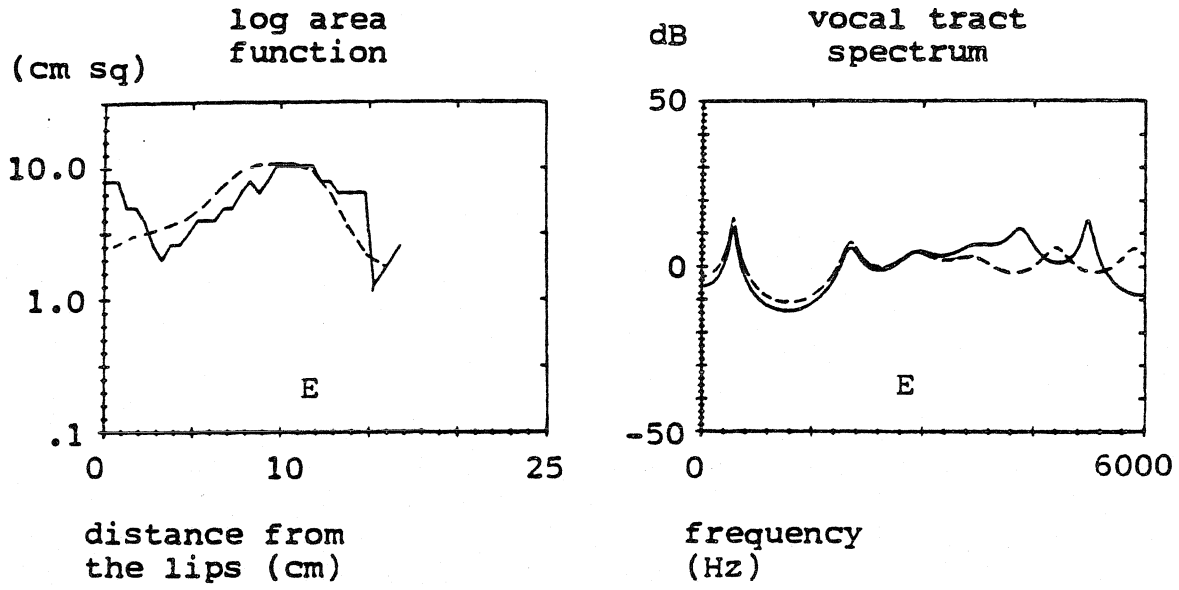
Figure 5.14



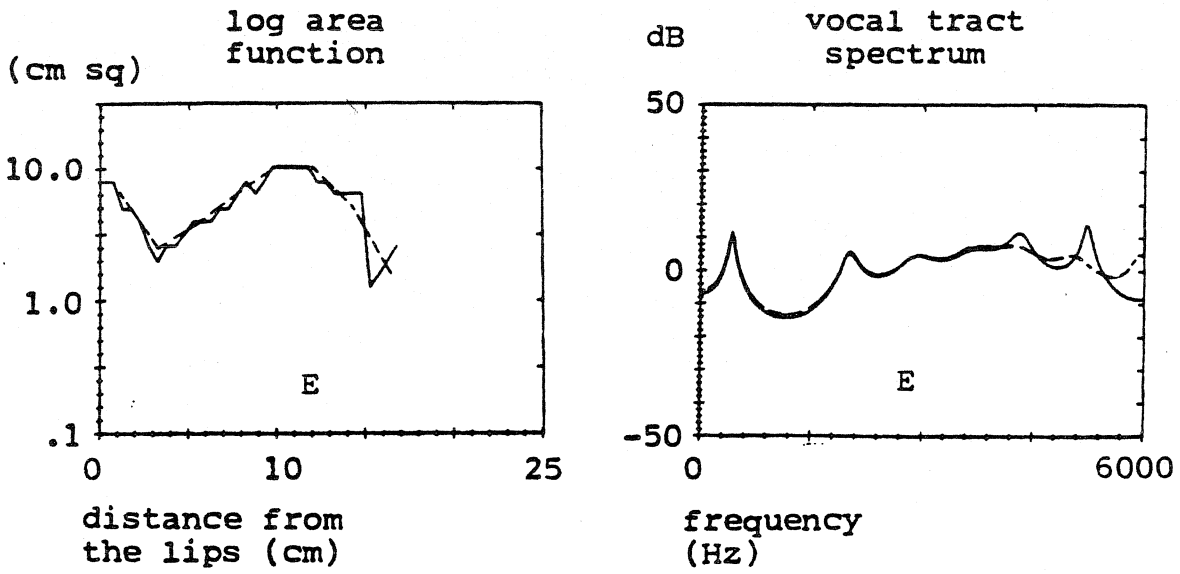
Fixed sampling rate, 16 section acoustic tube.

Figure 5.15

16 section acoustic tube:



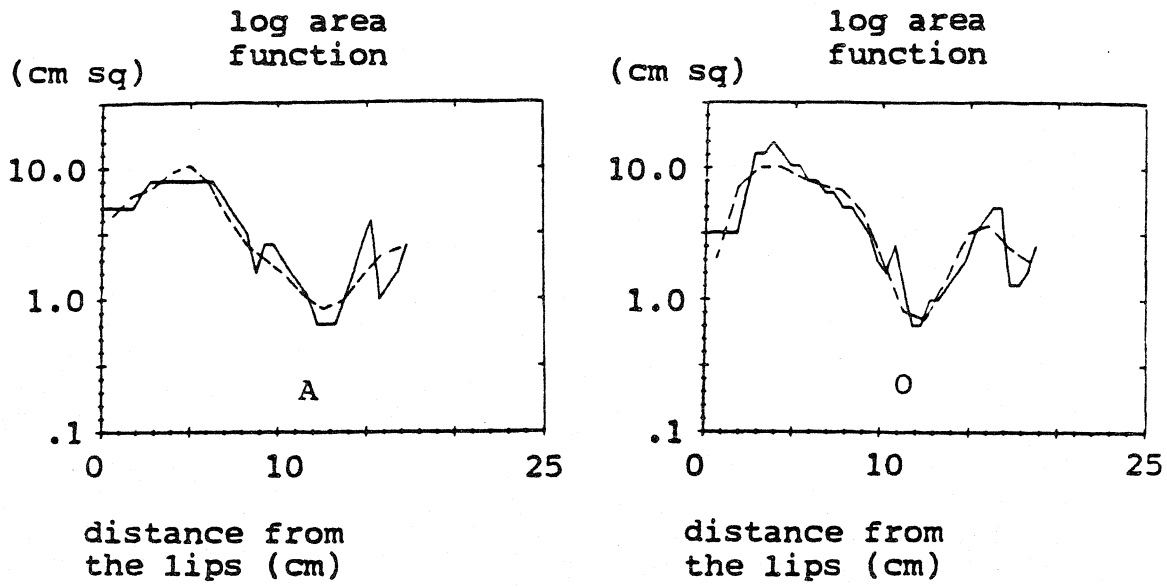
8 section acoustic tube:



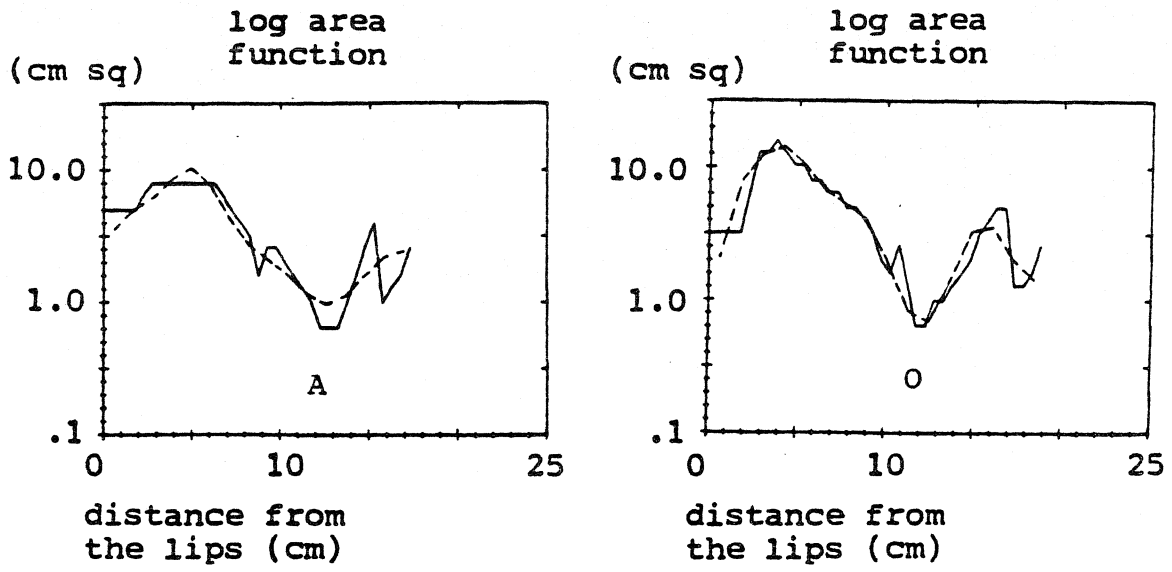
Fixed sampling rate, vocal tract length estimated.

Figure 5.16

Impulse excitation - pitch = 120 Hz:



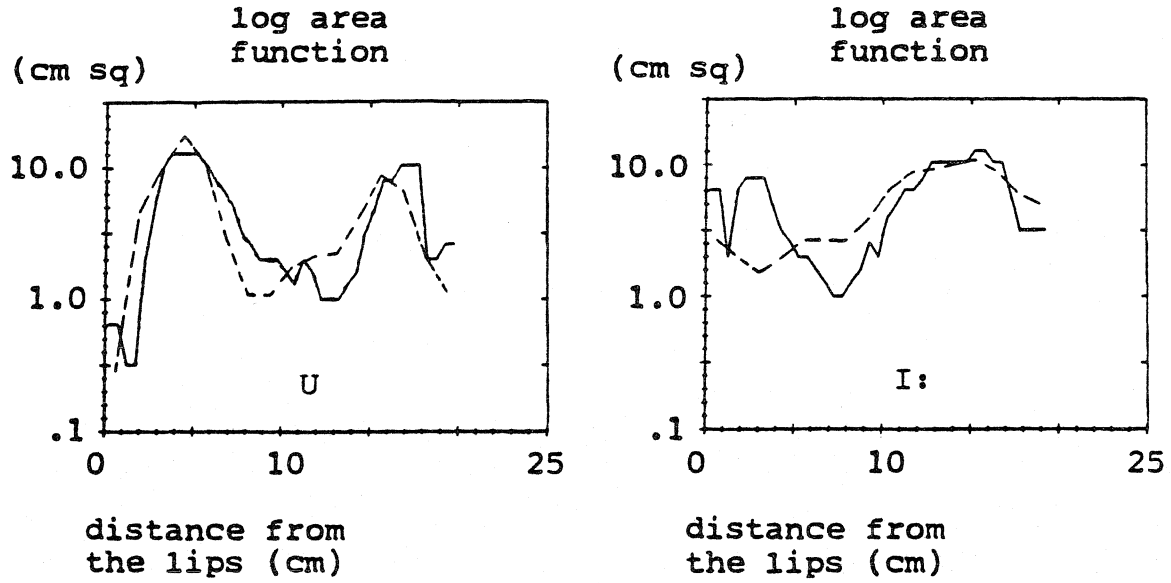
Glottal pulse excitation - pitch = 120 Hz:



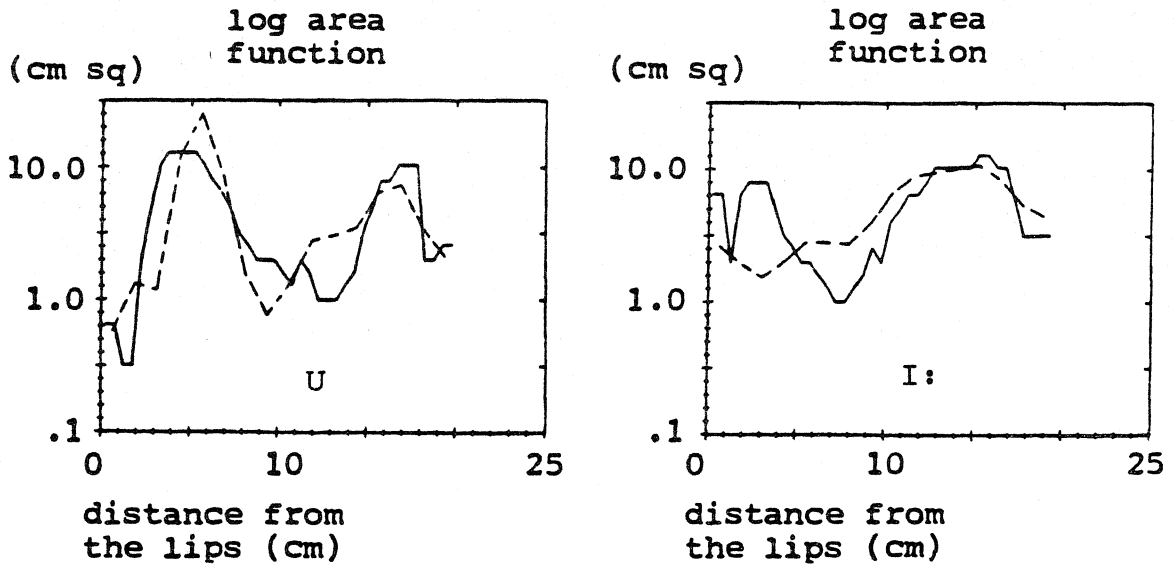
Effects of voicing periodicity and glottal spectrum.

Figure 5.17

Impulse excitation - pitch = 120 Hz:



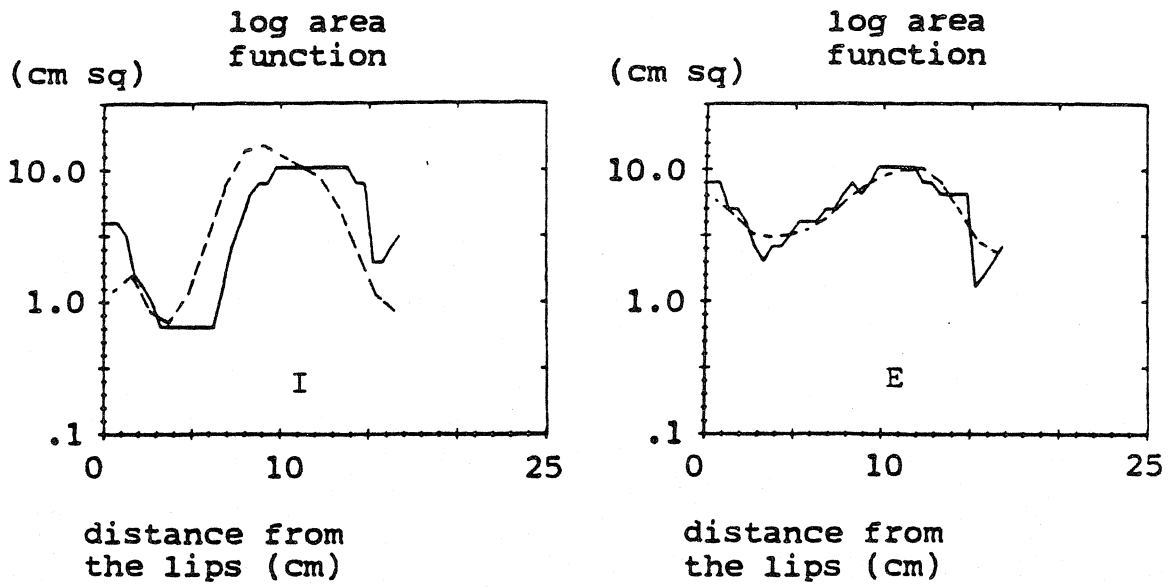
Glottal pulse excitation - pitch = 120 Hz:



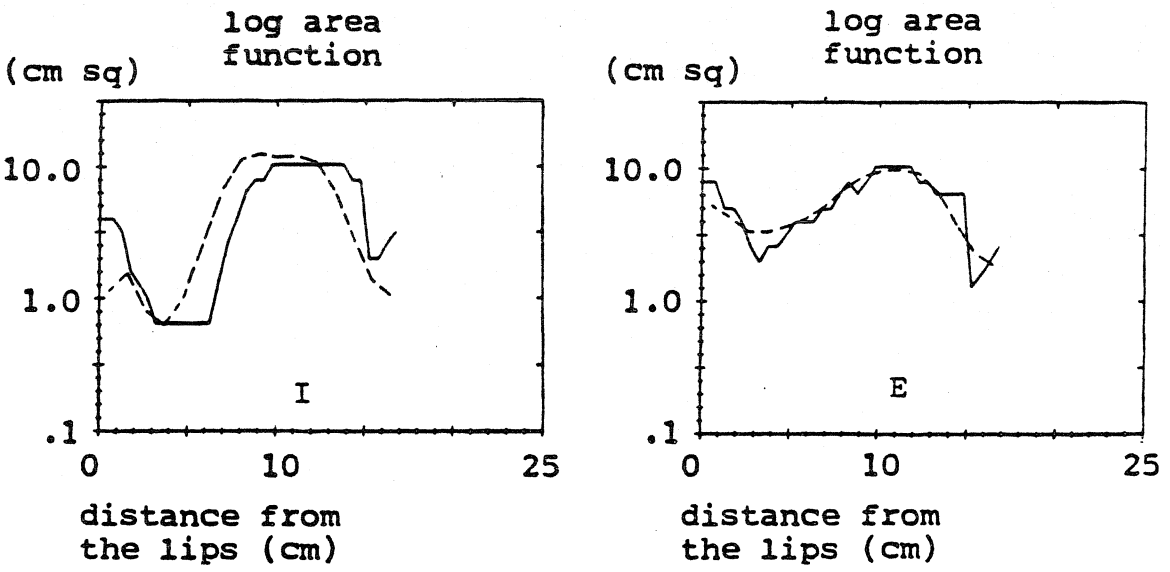
Effects of voicing periodicity and glottal spectrum.

Figure 5.18

Impulse excitation - pitch = 120 Hz:

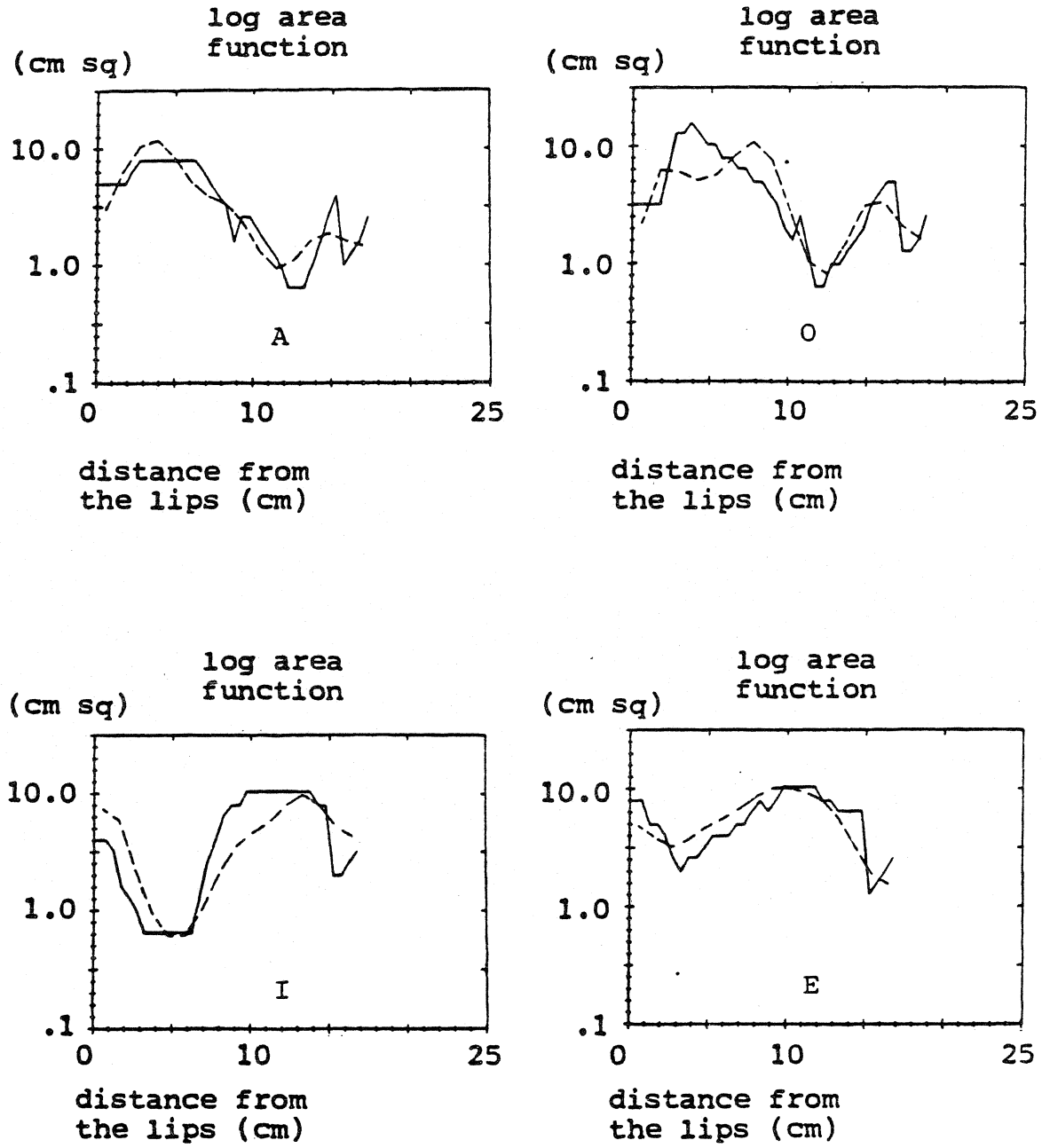


Glottal pulse excitation - pitch = 120 Hz:



Effects of voicing periodicity and glottal spectrum.

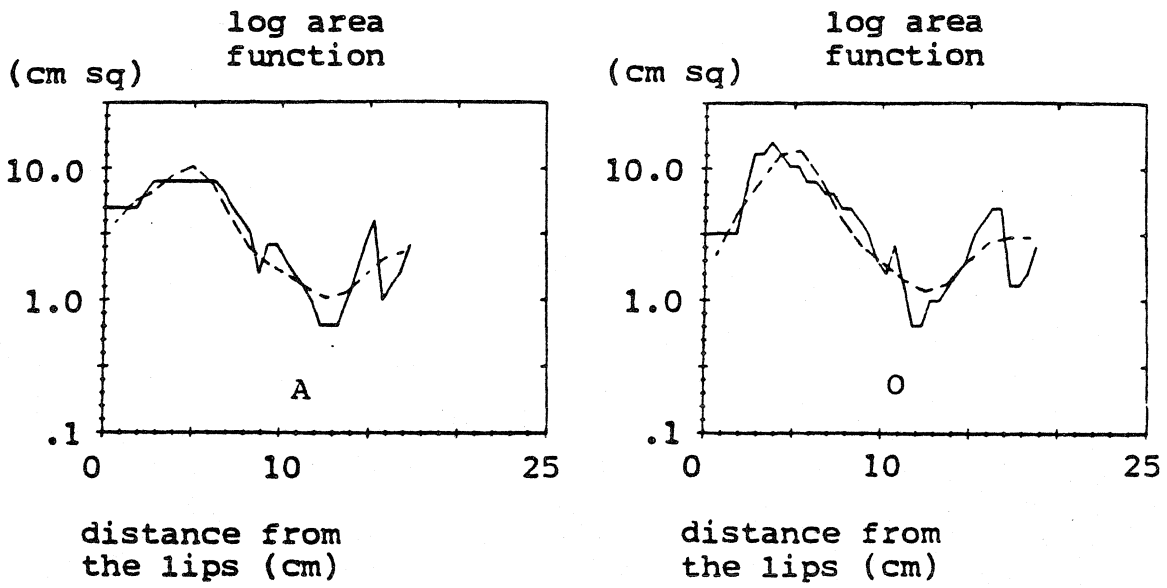
Figure 5.19



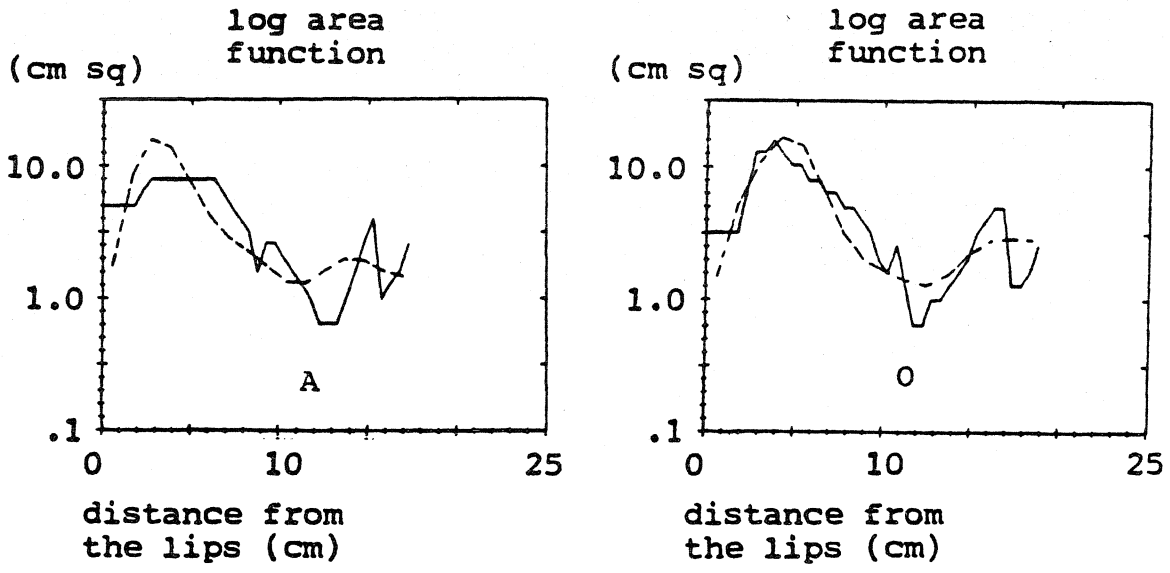
Effects of voicing periodicity - pitch = 240 Hz, impulse excitation.

Figure 5.20

Vocal tract wall losses:



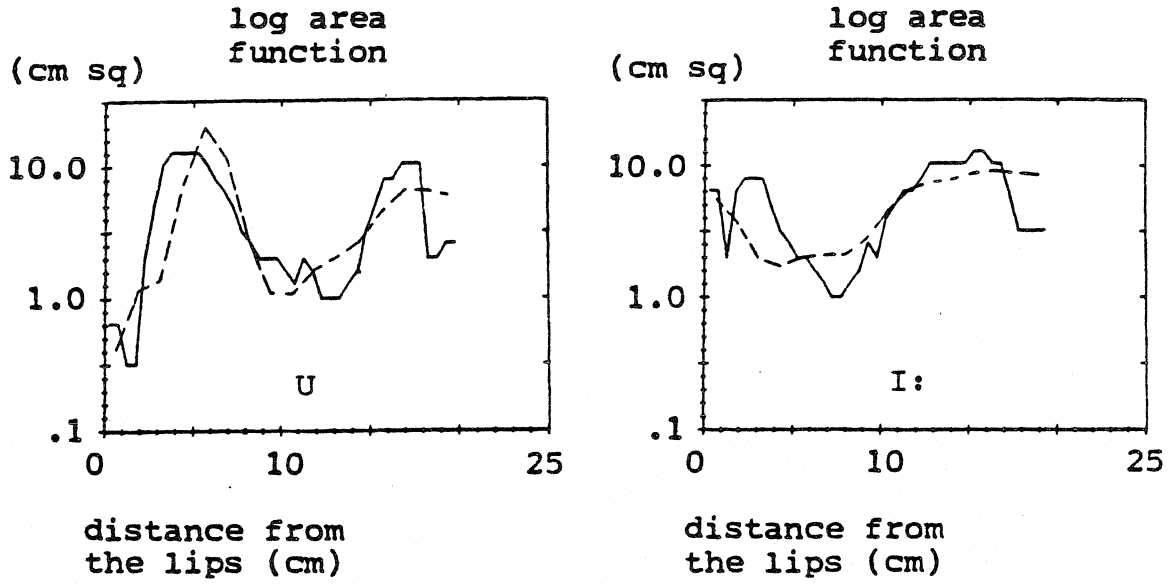
Wall losses plus radiation load:



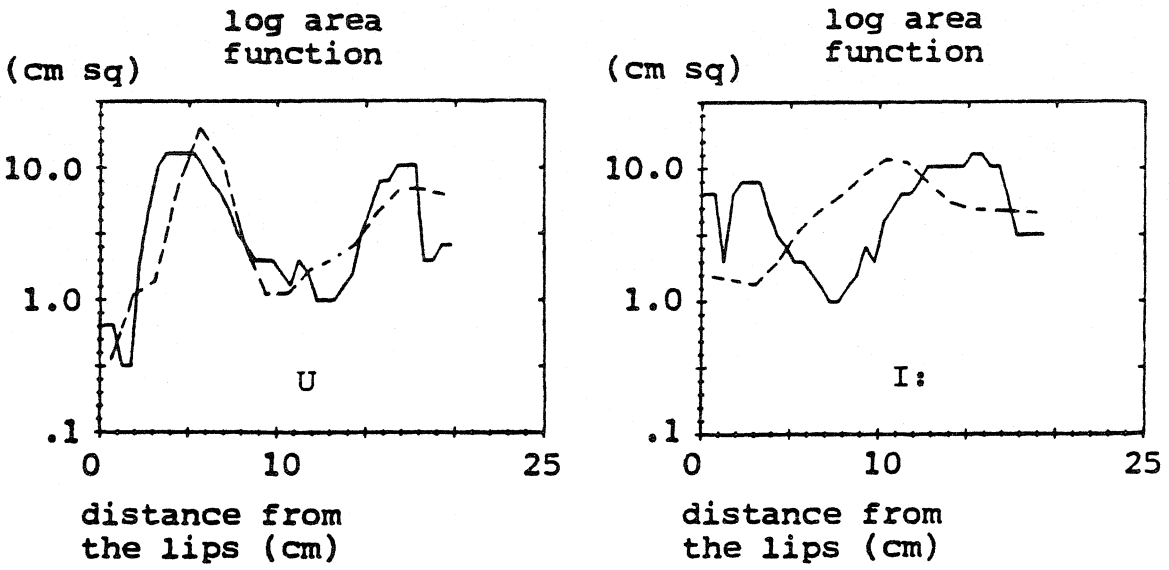
Effects of vocal tract loss and boundary conditions.

Figure 5.21

Vocal tract wall losses:



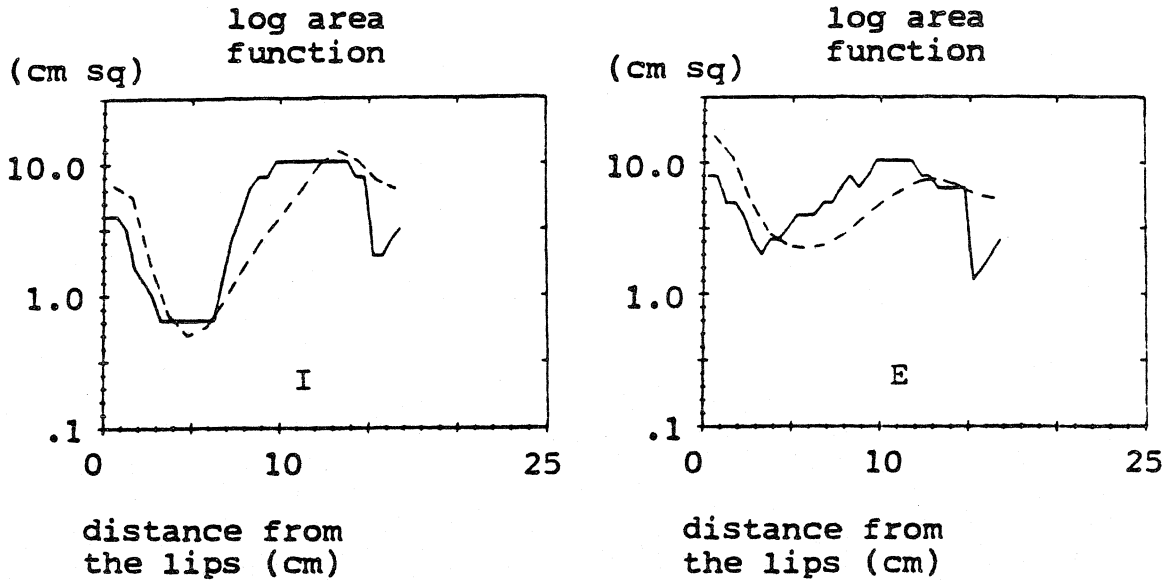
Wall losses plus radiation load:



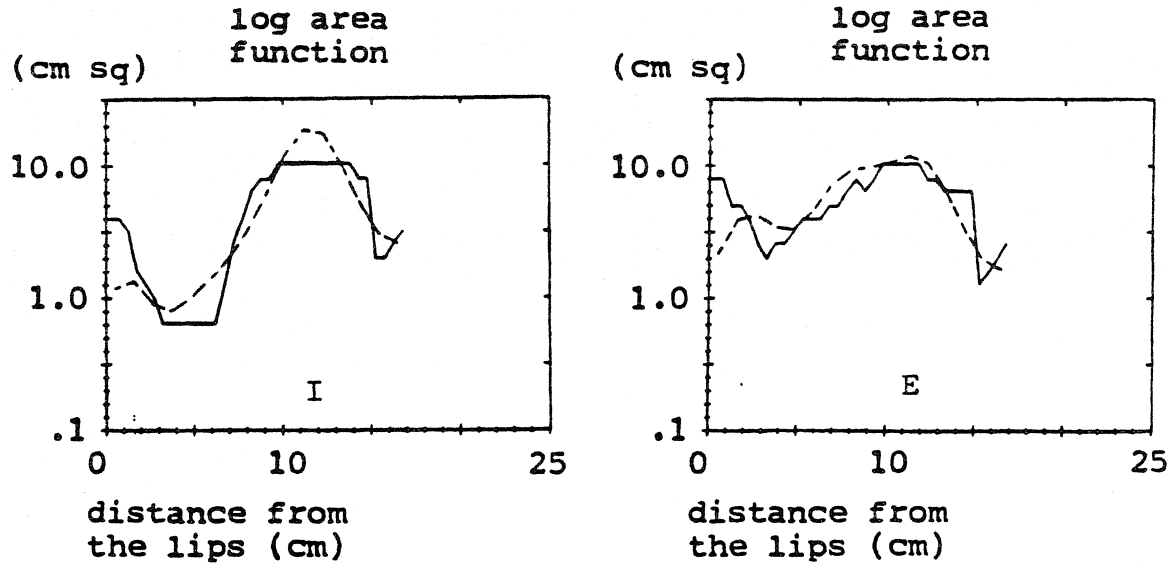
Effects of vocal tract loss and boundary conditions.

Figure 5.22

Vocal tract wall losses:

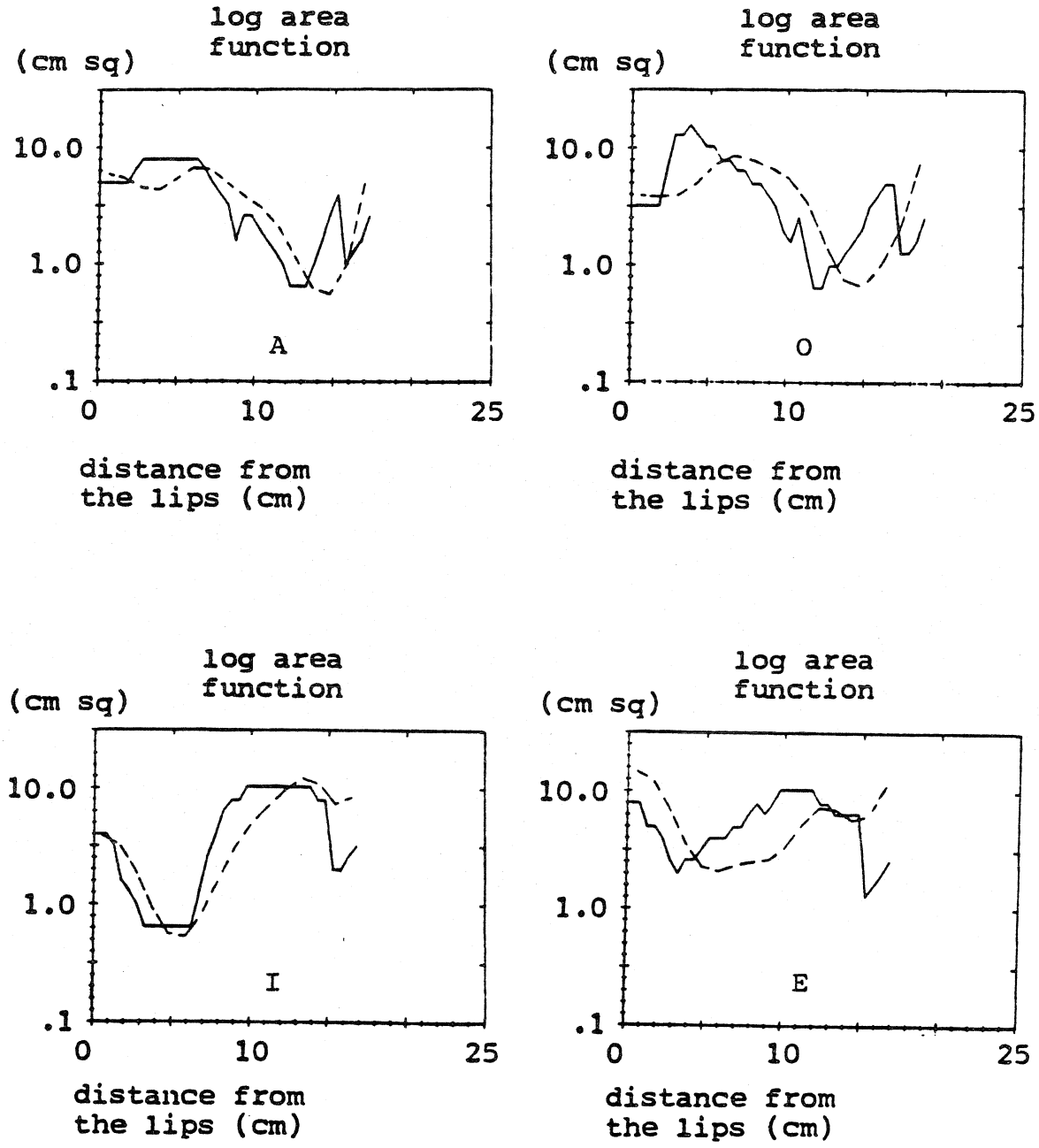


Wall losses plus radiation load:



Effects of vocal tract loss and boundary conditions.

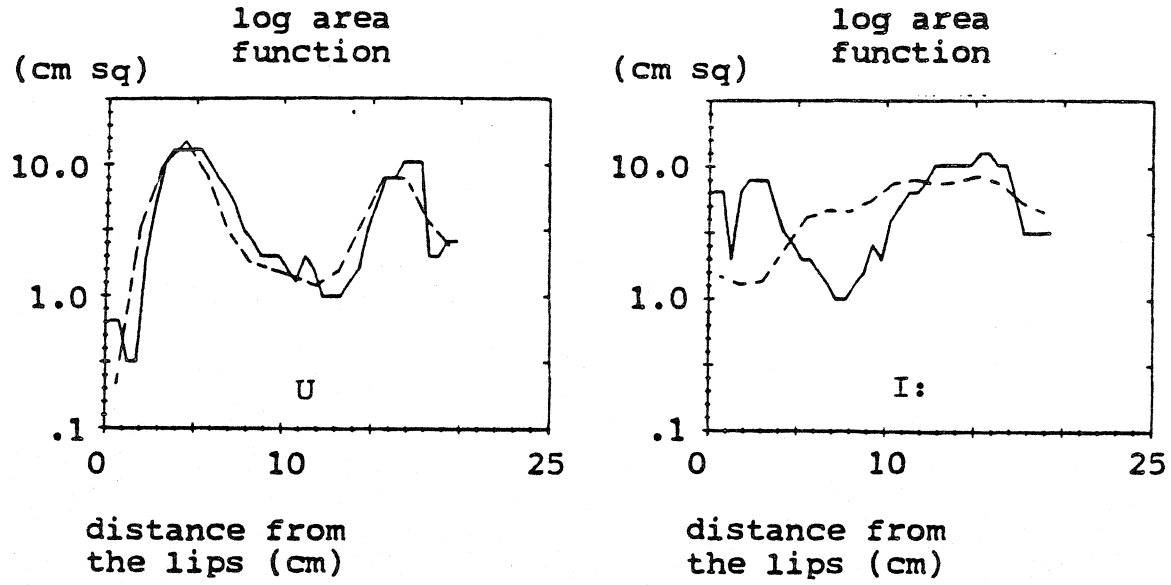
Figure 5.23



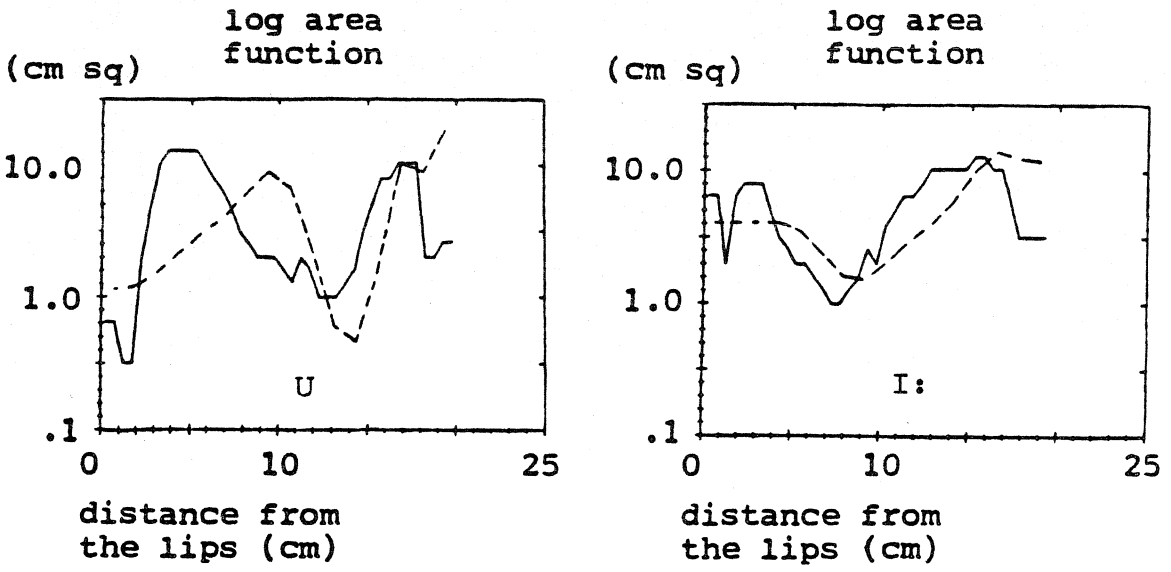
Atal-Hanauer acoustic tube applied to the case of vocal tract wall loss and radiation load.

Figure 5.24

Wakita formulation acoustic tube applied to the case of wall loss, radiation load, and greatly increased glottal loss:

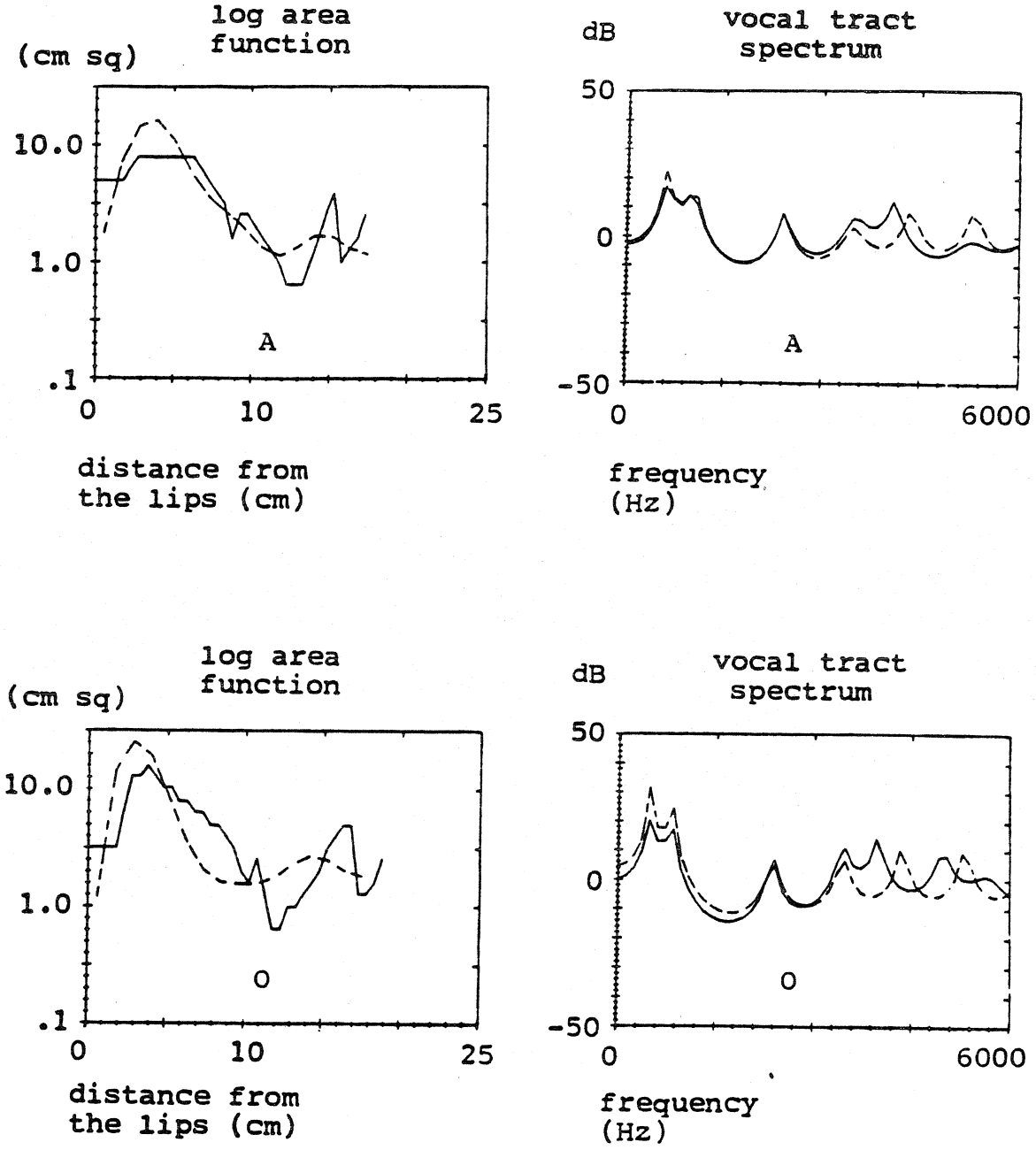


Atal-Hanauer formulation acoustic tube applied to the case of wall loss, radiation load, and no glottal loss:



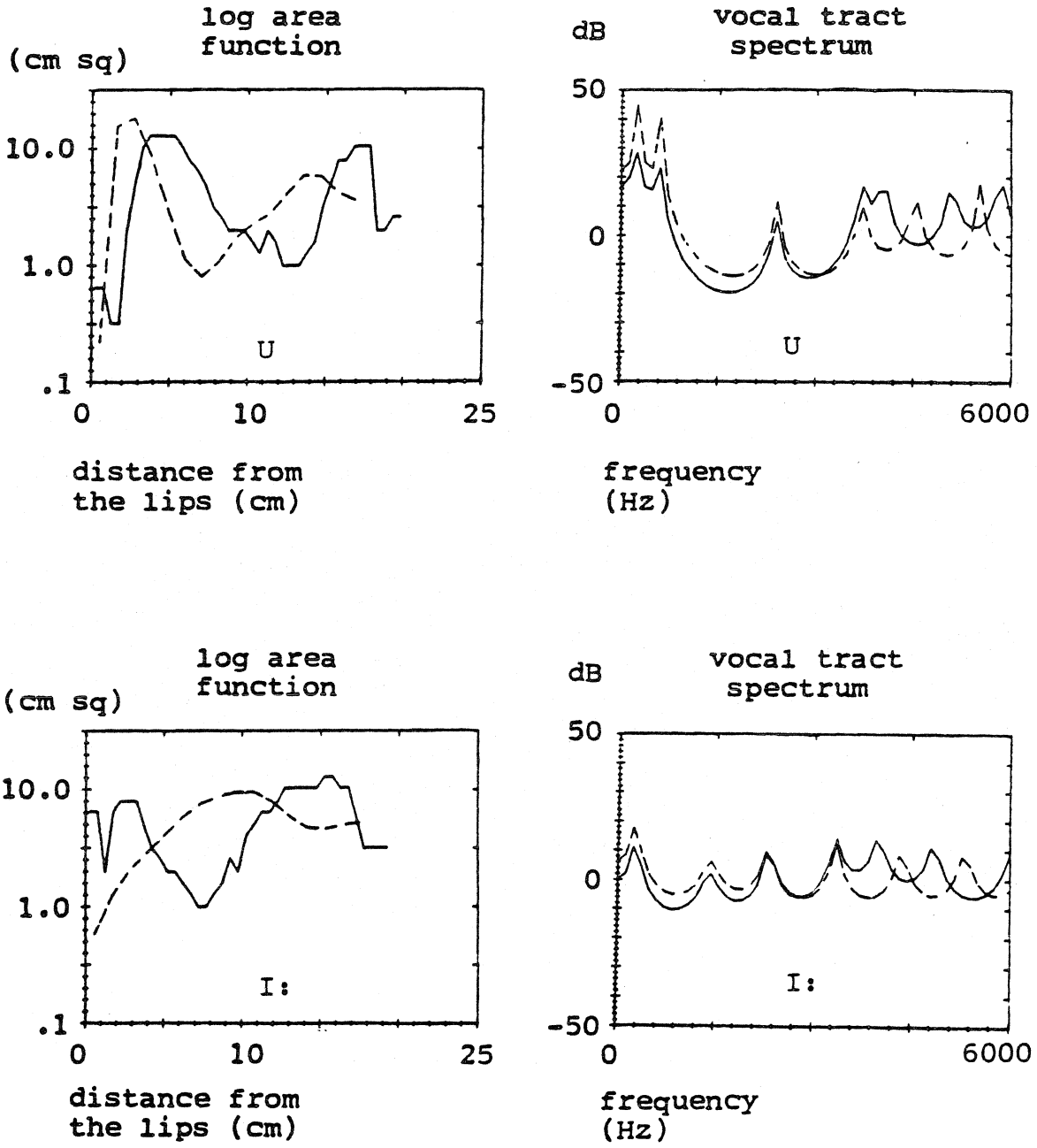
Circumstances under which one acoustic tube formulation may be more valid than the other.

Figure 5.25



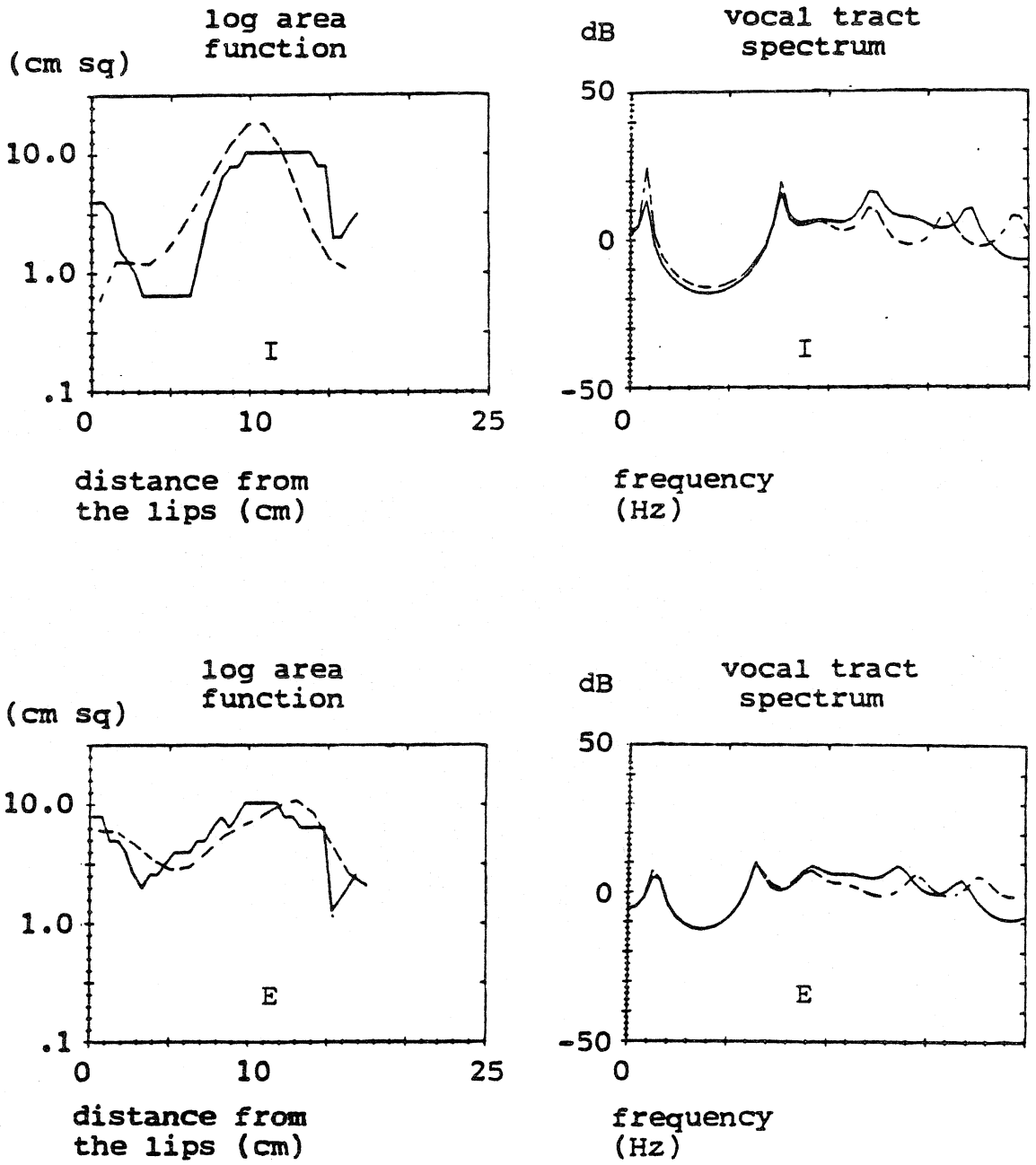
Area function estimation applied in the case where all error sources are in effect.

Figure 5.26



Area function estimation applied in the case where all error sources are in effect.

Figure 5.27



Area function estimation applied in the case where all error sources are in effect.

Figure 5.28

APPENDIX A:

Relationship Between the Acoustic Tube Vocal Tract Model and the All Pole Discrete Time Filter

Linear predictive analysis of a sampled speech signal can be used to characterize the transfer function of the vocal tract by that of an all pole discrete time filter. The coefficients of this filter, the LPC coeffficients, can be used to specify the cross sectional areas of an acoustic tube model of the vocal tract [Refs. A1-A4]. The transformation between LPC coefficients acoustic tube areas is presented for two versions of the acoustic tube model.

The acoustic tube model is a pipe formed by connecting a series of N discrete sections [Ref. A5]. Each section is of uniform cross sectional area. Though the cross sectional area of each section can differ from that of its neighbors, each section has the same length. An appropriate glottal source model is specified at one end of the tube, and an appropriate acoustic load is specified at the other end of this tube. This load specifies vocal tract boundary conditions at the lips. The electrical analog of this acoustic tube model is that of a two port network, one port at the lips and the other port at the glottis. The model depicts the vocal tract acoustics in the absence of nasalization or fricative sounds.

There are two formulations for the acoustic tube terminations that

permit direct transformation back and forth between cross sectional areas and LPC coefficients. The two cases are shown in Figure A1. In interpreting the electrical analog representation of the acoustic loads shown in the figure, one should keep in mind that current is the analog of acoustic volume velocity and voltage is the analog of pressure. The acoustic tube formulation proposed by Atal and Hanauer [Ref. A1] places a resistive load at the lips. The lips can be thought of as connected to a semi- infinite pipe of uniform cross sectional area S_{N+1}. The load will equal the characteristic impedance of the pipe, which is purely resistive if the pipe is lossless with rigid walls. The glottal source is an ideal acoustic volume velocity source, having infinite acoustic impedance. The formulation proposed by Wakita [Ref. A2], however, terminates the mouth end of the tube with an acoustic short. The semi-infinite pipe attached to the mouth is regarded as having infinite cross sectional area, giving it an impedance of zero. The glottal volume velocity source now has a non zero resistive impedance. This impedance again can be regarded as that of a semi-infinite pipe of area S_{N+1}, this time attached to the glottis. These semi-infinite pipes are physical realizations of resistive loads. The actual vocal tract is not connected to semi-infinite pipes. As such, the actual vocal tract terminations are not pure resistances.

Each section of the acoustic tube model is a fixed length of lossless pipe with rigid walls. Wave propagation in each section is lossless and dispersion free. The volume velocity in each section is the difference between the volume velocities of forward and

backward travelling acoustic waves. The length of each tube section is:

$$\Delta x = \frac{1}{2} c T_s \qquad \text{(A.1}$$

where c is the velocity of sound and T_s is the round trip propagation time for sound in each section. Each section imparts a delay of $T_s/2$ on forward and backward travelling waves passing thru the section. The juncture between sections results in an acoustic impedance mismatch that causes reflections. Knowing the delays and the reflections, the signal flow graph for the acoustic tube model can be constructed. Figure A2 shows signal flow graphs for both the Wakita and the Atal-Hanauer boundary conditions. Signal flow from left to right represents the forward going acoustic wave. Flow from right to left represents the backward going wave. Signal flow up and down represents the reflections.

The quantities U, F, and B shown on the graphs represent transforms of total, forward going, and backward going volume velocity components. Because this model will be specified by analysing a sampled speech signal, the z-transform and the s-transform (Laplace transform) are used interchangeably according to the relationship $z=\exp(sT_s)$ where the propagation delays in the sections have the transform value $z^{-1/2}$ because each section delays a signal $T_s/2$ and imposes no attenuation. The reflection coefficients for the

Wakita case are:

$$\mu_i = \frac{S_i - S_{i+1}}{S_i + S_{i+1}} \qquad \text{(A.2}$$

where i = 1,...,N and for the Atal-Hanuaer case they are:

$$\mu_i = \frac{S_{i+1} - S_i}{S_{i+1} + S_i} \qquad \text{(A.3}$$

The S's are the cross sectional area values. These relationships specify the reflection coefficients from the areas but only the ratios between areas of neighboring sections from the reflection coefficients. Knowing the reflection coefficients specifies the shape of the acoustic tube but the absolute scale is unknown. For a given set of reflection coefficients, the acoustic tube shapes under the two formulations are not the same, however, but are related by:

$$S_i = \frac{1}{S_{N+2-i}} \qquad \text{(A.4}$$

Atal-Hanauer Wakita

where i = 1,...,N+1.

First, the relationship between the LPC coefficients and the Wakita formulation of the acoustic tube is shown. Forward and backward going wave components on each side of the i'th section of the tube

are related by:

$$F_i = \frac{z^{1/2}}{(1+\mu_i)} (F_{i-1} - z^{-1}\mu_i B_{i-1}) \qquad \text{(A.5}$$

$$B_i = \frac{z^{1/2}}{(1+\mu_i)} (-\mu_i F_{i-1} + z^{-1} B_{i-1}) \qquad \text{(A.6}$$

These relationships can be read from the signal flow graph. By specifying:

$$A_0(z) = 1 \qquad \text{(A.7}$$

the following holds true:

$$F_0 = A_0(z)\, F_0 \qquad \text{(A.8}$$

$$B_0 = -F_0 = -A_0(z^{-1})\, F_0 \qquad \text{(A.9}$$

by virtue of the boundary conditions. Knowing equations A.5 thru A.9,

one can establish by induction that:

$$F_n = \frac{z^{n/2}}{\prod_{1}^{n}(1+\mu_i)} A_n(z)\, F_0 \qquad \text{(A.10}$$

$$B_n = \frac{-z^{-n/2}}{\prod_{1}^{n}(1+\mu_i)} A_n(z^{-1})\, F_0 \qquad \text{(A.11}$$

for all n = 1,...,N where the polynomial A_n is recursively constructed as:

$$A_n(z) = A_{n-1}(z) + \mu_n z^{-n} A_{n-1}(z^{-1}) \qquad \text{(A.12}$$

for n = 1,...,N. The transfer function relating volume velocity at the lips to volume velocity at the glottis is of the form:

$$T(z) = \frac{U_L}{U_G} = \frac{2F_0}{2F_N} = \frac{z^{-N/2}\prod_{1}^{N}(1+\mu_i)}{A_N(z)} \qquad \text{(A.13}$$

This transfer function is the same as that of an all pole filter modified by a fixed gain and placed in series with a delay. This delay equals the time it takes sound to travel the full length of the acoustic tube.

Next, the same analysis is carried out for the Atal-Hanuaer formulation. From the signal flow graph one obtains:

$$F_{i-1} = \frac{z^{1/2}}{(1+\mu_i)} (F_i - \mu_i B_i) \qquad \text{(A.14}$$

$$B_{i-1} = \frac{z^{1/2}}{(1+\mu_i)} (B_i - \mu_i F_i) \qquad \text{(A.15}$$

Specifying that:

$$A_0(z) = 1 \qquad \text{(A.16}$$

it holds that:

$$F_0 - B_0 = A_0(z)\, F_0 - A_0(z^{-1})\, B_0 \qquad \text{(A.17}$$

By induction it can be established that:

$$F_0 - B_0 = \frac{z^{n/2}}{\prod_1^n (1+\mu_i)} A_n(z)\, F_n - \frac{z^{-n/2}}{\prod_1^n (1+\mu_i)} A_n(z^{-1})\, B_n \qquad \text{(A.18}$$

for all $n = 1,\ldots,N$. The polynomial A_n is specified by equation A.12 as in the previous formulation. Given that:

$$B_n = 0 \qquad \text{(A.19}$$

which simply means that speech sound exits the mouth but is not put back in again, the transfer function of the acoustic tube has the form:

$$T(z) = \frac{F_N}{F_0 - B_0} = \frac{z^{-N/2} \prod_{1}^{N} (1 + \mu_i)}{A_N(z)} \qquad \text{(A.20}$$

This means that by specifying the same reflection coefficients one obtains the same transfer function under the two formulations.

Finally, one can convert between the polynomial A_n and the reflection coefficients. If we have:

$$A_{n-1}(z) = 1 + \sum_{i=1}^{n-1} a_{(n-1)i} z^{-i} \qquad \text{(A.21}$$

then:

$$A_n(z) = 1 + \sum_{i=1}^{n-1} (a_{(n-1)i} + \mu_n a_{(n-1)(n-i)}) z^{-i} + \mu_n z^{-n}$$

$$= 1 + \sum_{i=1}^{n} a_{ni} z^{-i} \qquad \text{(A.22}$$

This means the coefficient:

$$a_{ni} = a_{(n-1)i} + \mu_n a_{(n-1)(n-i)} \qquad \text{(A.23}$$

for i = 1,...,n-1 and:

$$a_{nn} = \mu_n \qquad \text{(A.24}$$

for i = n. The reflection coefficients can be then obtained from A_N by remembering that for each n we can obtain a [Eqn. A.24], and the order of A_n can be reduced by specifying:

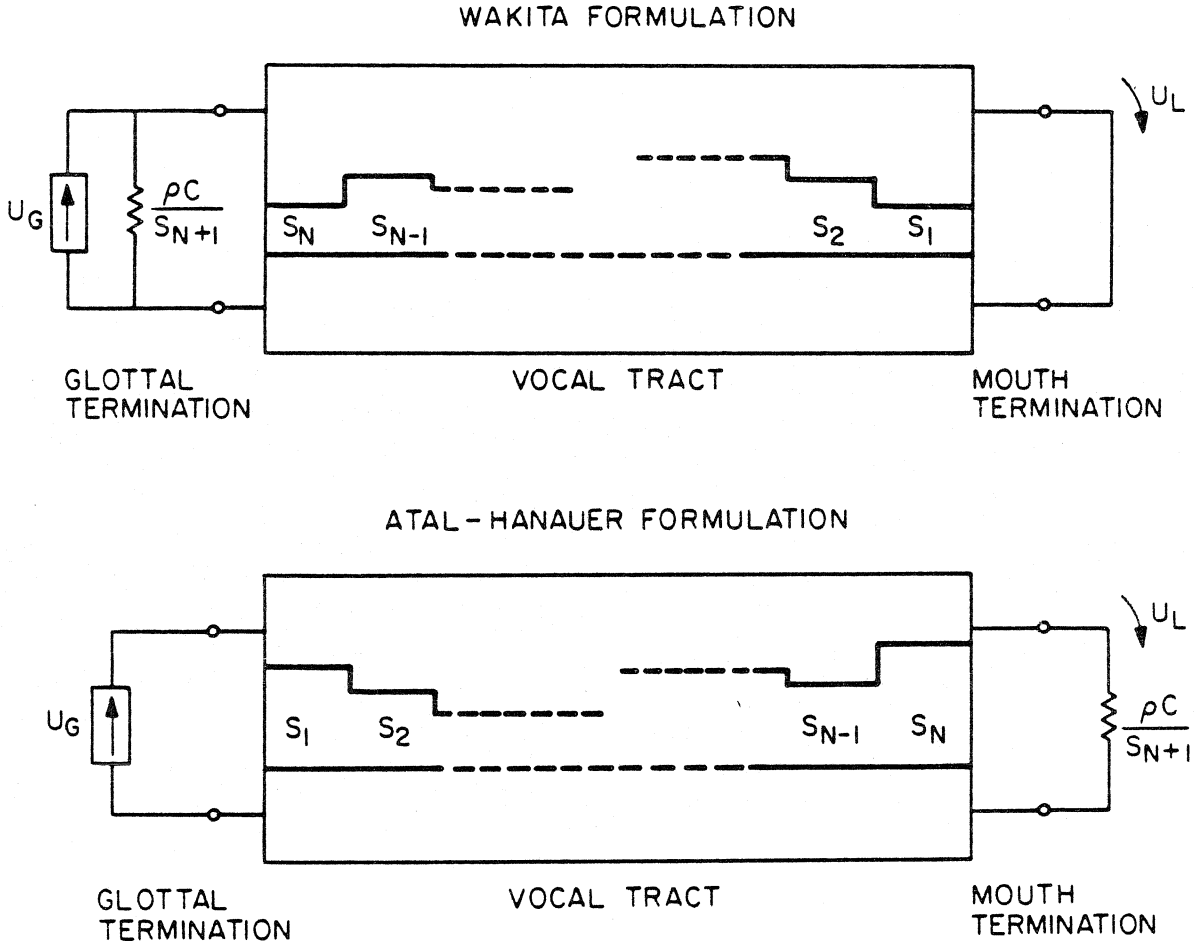
$$a_{(n-1)n} = 0 \qquad \text{(A.25}$$

and:

$$a_{(n-1)i} = \frac{a_{ni} - a_{nn}a_{n(n-i)}}{1 - a_{nn}^2} \qquad \text{(A.26}$$

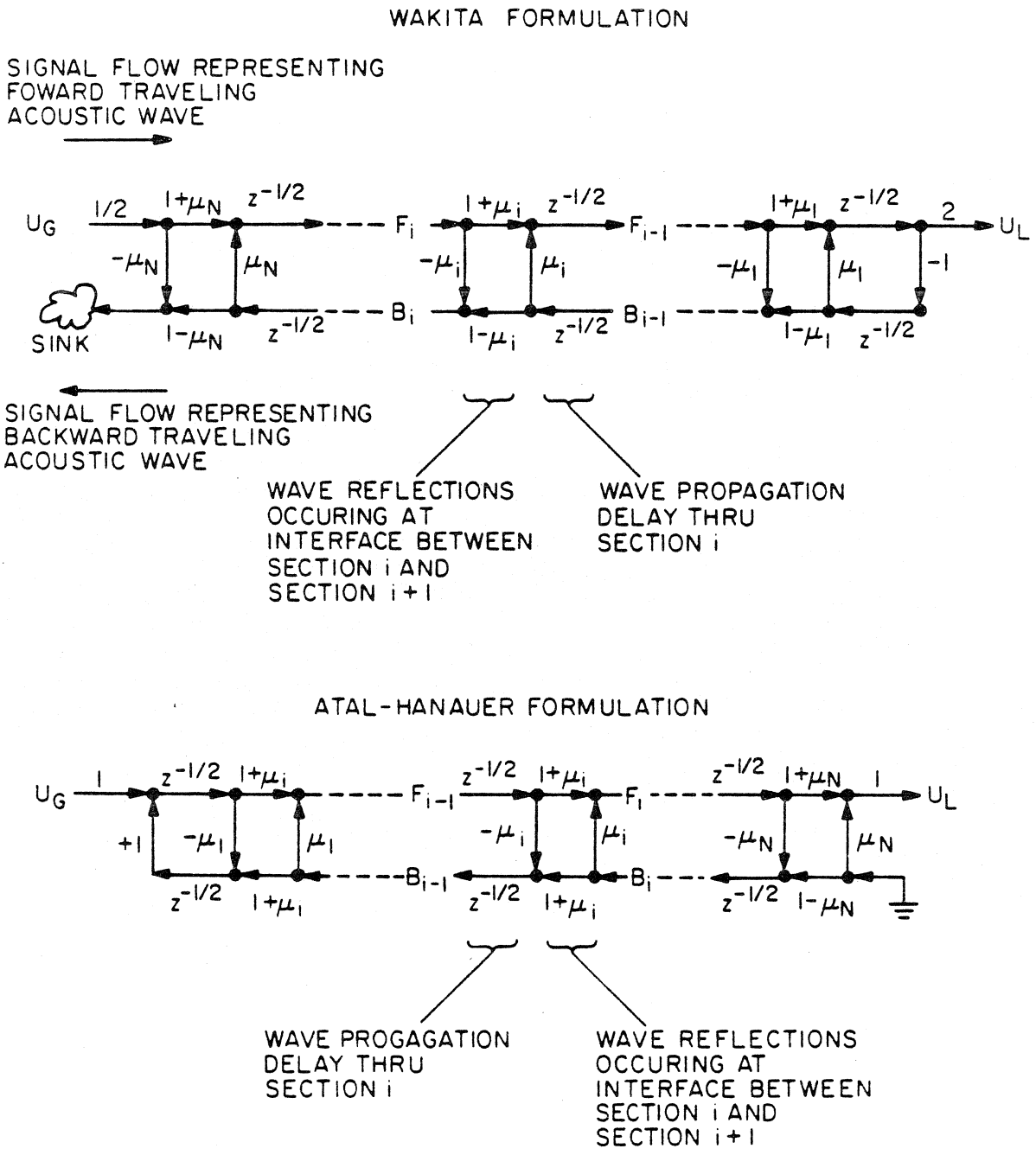
for i = 1,...,n-1.

The transformation between coefficients of the polynomial A_N and the reflection coefficients is bidirectional. Likewise one can transform both ways between a particular acoustic tube model, with its ratios of areas in neighboring sections, and the reflection coefficients. The coefficients of A_N are obtained from LPC analysis as LPC specifies the coefficients of an all pole discrete time filter model. As a result of all of this, LPC analysis can specify acoustic tube cross sectional areas to within an unknown scaling.



Discrete section acoustic tube models of the vocal tract.

Figure A1



Signal flow graphs for discrete time filter realizations of the acoustic tube models.

Figure A2

APPENDIX B:

Relationship Between Sampling Rate and Vocal Tract Length

Estimating the cross sectional areas of the vocal tract from the speech wave is a two part process. In the first part, the transfer function of the vocal tract is modelled as that of an all pole discrete time filter. The coefficients of this filter are obtained by doing LPC analysis on the speech wave. The second part of the process involves calculating the areas of an acoustic tube model from the filter coefficients one has just obtained.

The LPC analysis is performed on sample values of the speech signal. One of the crucial quantities to be specified is the sampling rate. The LPC forward filter is an all pole discrete time model of the vocal tract, which is a continuous time all pole system. The accuracy to which a discrete time model can represent a continuous time system is sensitive to the sampling rate selected.

The power spectrum of the speech signal, or speech spectrum, can be considered to be the product of three parts based upon the linear model of voiced speech production depicted in Figure B1. These three parts are the glottal spectrum, the magnitude squared of the vocal tract frequency response, and the magnitude squared of the frequency

response of the radiation from the vocal tract:

$$S(f) = G(f) \left|V(j2\pi f)\right|^2 \left|R(j2\pi f)\right|^2 \qquad \text{(B.1}$$

The magnitude squared of frequency response is referred to as the spectrum as in vocal tract or radiation spectrum. The vocal tract spectrum is obtained from the transfer function relating lip volume velocity to glottal volume velocity; the radiation spectrum comes from the transfer function relating far field acoustic pressure to the lip volume velocity:

$$V(s) = \frac{U_L(s)}{U_G(s)} \qquad \text{(B.2}$$

$$R(s) = \frac{P(s)}{U_L(s)} \qquad \text{(B.3}$$

Prior to performing the LPC analysis, the speech signal is low pass filtered and sampled at a rate f_s. The effect on the speech spectrum of these operations is shown in Figure B2. The spectrum of the sampled speech signal has the following form:

$$\hat{S}(f) = \sum_{i=-\infty}^{\infty} S(f - if_s)\, LPF(f - if_s) \qquad \text{(B.4}$$

It is desirable for the low pass filter to have very steep skirts so as to avoid aliasing. However, the 6 dB frequency of the filter is

placed right at half the sampling frequency, which introduces some amount of aliasing [Refs. B1,B2 pp.153-154]. Some degree of aliasing must be tolerated to avoid introducing a notch in the spectrum of the sampled speech signal. This notch is undesirable as it would interfere with accurately modelling the speech spectrum.

Linear predictive analysis is performed on the speech signal to specify coefficients of an all pole discrete time filter having the transfer function:

$$\bar{V}(z) = \frac{K}{1+\sum_{i=1}^{N} a_i z^{-i}} = \frac{K}{\prod_{i=1}^{N/2} (1-\frac{z_i}{z})(1-\frac{z_i^*}{z})} \qquad \text{(B.5}$$

This transfer function has been expressed as a z-transform. The spectrum, or magnitude squared of the frequency response, of this filter:

$$\left| \bar{V}(e^{j2\pi fT_s}) \right|^2 = \left| \frac{K}{\prod_{i=1}^{N/2} (1-\frac{z_i}{e^{j2\pi fT_s}})(1-\frac{z_i^*}{e^{j2\pi f\ T_s}})} \right|^2 \qquad \text{(B.6}$$

forms an estimate of the envelope of the sampled speech signal - S(f). This is a consequence of the spectrum matching properties of LPC [Refs. B2 pp.129-163,B3,B4].

The first consequence of the choice of sampling rate is whether the LPC spectrum accurately represents the sampled speech spectrum. Accuracy here is crucial if the LPC spectrum is to accurately

represent S(f), the actual speech spectrum, in the frequency range 0 to $f_s/2$, the lower frequencies of speech being crucial to speech perception. If $f_s/2$ occurs in a trough of S(f), the LPC spectrum will give a good fit. But if $f_s/2$ occurs on the shoulder of a resonance peak of S(f), the sampled speech spectrum will not vary smoothly across the junctures $f_s/2 + nf_s$, $n = -\infty$ to $+\infty$. In this case the LPC spectrum will not give a good fit to the sampled speech spectrum at these junctures.

Since it is the vocal tract response one needs to model, one needs to remove the effects of G(f) and R(s) [Eqns. B.1,B.3] from the LPC transfer function $\overline{V}(z)$. This is done by factoring out poles from $\overline{V}(z)$, leaving only the narrow bandwidth conjugate pole pairs corresponding to the vocal tract response [Ref. B5 p.646]. Once the extraneous poles have been removed from $\overline{V}(z)$, the LPC spectrum [Eqn. B.6] will form an estimate of the vocal tract spectrum.

The vocal tract transfer function has an infinite set of conjugate pole pairs corresponding to an infinite set of resonances. These resonances are referred to as the formants. The vocal tract poles are separated into the first N/2 pole pairs, corresponding to the perceptually important low order formants, and everything else:

$$V(s) = \frac{K_G}{\prod_{i=1}^{\infty}\left(1-\frac{s}{s_i}\right)\left(1-\frac{s}{s_i^*}\right)} = \frac{K_G\, K_{FN}(s)}{\prod_{i=1}^{N/2}\left(1-\frac{s}{s_i}\right)\left(1-\frac{s}{s_i^*}\right)} \qquad (B.7$$

From the transfer function the vocal tract spectrum is computed as:

$$|V(j2\pi f)|^2 = \left| \frac{K_G \, K_{rN}(j2\pi f)}{\prod_{i=1}^{N/2} \left(1 - \frac{j2\pi f}{s_i}\right)\left(1 - \frac{j2\pi f}{s_i^*}\right)} \right|^2 \qquad (B.8$$

The term $K_{rN}(s)$ corresponds to the effect of "everything else," that is all the high order formants [Ref. B6 p.42]. The exact frequencies and bandwidths of the high order poles are of minor interest because the high order formants are of minor perceptual significance in speech. But the gross effect of the high order poles cannot be neglected as their combined effect greatly influences the amplitudes of the low order formants [Ref. B6 p.50].

The form of equation B.6, the LPC spectrum, does not lend itself to direct comparison with equation B.8, the vocal tract spectrum. The form can be changed by use of a theorem in complex variables [Ref. B7] that states that if F(s) has simple zeroes at a_1, a_2, a_3, ..., and the limit of $|a_n|$ as n goes to ∞ is ∞, and F(s) is an entire analytic function then:

$$F(s) = F(0) e^{\frac{F'(0)}{F(0)} s} \prod_{i=1}^{\infty} \left\{ \left(1 - \frac{s}{a_i}\right) e^{s/a_n} \right\} \qquad (B.9$$

Next, we make the z-transform to s-transform transformation:

$$z = e^{sT_s} \qquad (B.10$$

which applied to equation B.5 gives:

$$\bar{V}(z) = \frac{K}{\prod_{i=1}^{N/2} \left(1 - e^{(s_i - s)T_s}\right)\left(1 - e^{(s_i^* - s)T_s}\right)} \qquad (B.11$$

Considering the term containing the i'th z-transform pole pair and applying the theorem we get:

$$(B.12$$

$$F_i(s) = \left(1 - e^{(s_i - s)T_s}\right)\left(1 - e^{(s_i^* - s)T_s}\right)$$

$$= \left(1 - e^{s_i T_s}\right)\left(1 - e^{s_i^* T_s}\right) e^{s\left[\frac{e^{s_i T_s}}{1 - e^{s_i T_s}} + \frac{e^{s_i^* T_s}}{1 - e^{s_i^* T_s}}\right]}$$

$$\prod_{n=-\infty}^{\infty} \left[\left(1 - \frac{s}{s_i + \frac{j2\pi n}{T_s}}\right)\left(1 - \frac{s}{s_i^* - \frac{j2\pi n}{T_s}}\right)\right.$$

$$\left. e^{\left(\frac{s}{s_i + \frac{j2\pi n}{T_s}} + \frac{s}{s_i^* - \frac{j2\pi n}{T_s}}\right)}\right]$$

for i = 1,...,N/2. Taking the magnitude squared of the frequency response of this term gives:

$$(B.13$$

$$|F_i(j2\pi f)|^2 =$$

$$|K_i|^2 \cdot |1|^2 \cdot \left| \prod_{n=-\infty}^{\infty} \left(1 - \frac{s}{s_i + \frac{j2\pi n}{T_s}}\right)\left(1 - \frac{s}{s_i^* - \frac{j2\pi n}{T_s}}\right) \right|^2 \cdot \prod_{n=-\infty}^{\infty} |1|^2$$

where:

$$K_i = \left(1 - 2\,Re\left(e^{s_i T_s}\right) + \left|e^{s_i T_s}\right|^2\right) \qquad (B.14$$

This means that the spectrum of the LPC filter model has the expansion:

$$\left|\bar{V}\left(e^{j2\pi f T_s}\right)\right|^2 = \left|\frac{K}{\prod_{i=1}^{N/2} K_i}\right|^2 \cdot \left|\frac{1}{\prod_{i=1}^{N/2}\left(1-\frac{j2\pi f}{s_i}\right)\left(1-\frac{j2\pi f}{s_i^*}\right)}\right|^2 \qquad (B.15$$

$$\cdot \left|\frac{1}{\prod_{n=1}^{\infty}\left[\prod_{i=1}^{N/2}\left(1-\frac{j2\pi f}{s_i+\frac{j2\pi n}{T_s}}\right)\left(1-\frac{j2\pi f}{s_i-\frac{j2\pi n}{T_s}}\right)\left(1-\frac{j2\pi f}{s_i^*+\frac{j2\pi n}{T_s}}\right)\left(1-\frac{j2\pi f}{s_i^*-\frac{j2\pi n}{T_s}}\right)\right]}\right|^2$$

This equation is of the form of an arbitrary gain factor multiplied by the magnitude squared of the frequency response of a transfer function of an infinite number of pole pairs. Comparing equations B.8 and B.15 one sees that the first N/2 pole pairs of the vocal tract are represented by the first N/2 pole pairs of the LPC model. The high order pole pairs of the vocal tract are represented by the periodic replications of the first N/2 pole pairs of the LPC model into the higher frequencies.

The key feature of using a discrete time digital filter model to represent a continuous time vocal tract is that one obtains the formant amplitude correction of having the high order poles while specify-

ing only a finite number of model coefficients. The effect on formant amplitude varies with the spacing of the high order s-transform poles of the model, the spacing being controlled by the sampling rate. The amount of amplitude correction can be varied by varying the coefficients of the filter model to represent a change in sampling rate [Ref. B8]. The complex s-transform variable s is represented as having frequency and damping coordinates:

$$s = 2\pi(-d + jf) \qquad \text{(B.16}$$

Frequency and damping both have the units of Hertz. Employing equation B.10, we use the z-transform pole z_i of the filter model to specify:

$$d_i = \frac{-1}{2\pi T_s} \log z_i = \frac{-f_s}{2\pi} \log z_i \qquad \text{(B.17}$$

$$f_i = \frac{1}{2\pi T_s} \arg(z_i) = \frac{f_s}{2\pi} \arg(z_i) \qquad \text{(B.18}$$

The quantity d_i is an estimate of one half the 3 dB bandwidth of the i'th vocal tract formant. The relation:

$$z_i = e^{s_i T_s} = e^{s_i / f_s} = e^{\frac{2\pi(-d_i + jf_i)}{f_s}} \qquad \text{(B.19}$$

$$\{ j = \sqrt{-1} \}$$

recovers the value of z_i. By substituting a new value of f_s in

recovering z_i, however, one can scale all the z_i's of the LPC filter model to reflect a change in sampling rate.

The vocal tract length relates to the need for changing the sampling rate. The length of the vocal tract not only varies from talker to talker, but it varies from frame to frame in the course of articulation [Refs. B9,B10]. The length of the LPC acoustic tube model is given by:

$$L = N \frac{c}{2f_s} \qquad \text{(B.20}$$

where c is the speed of sound and N is the number of coefficients. For fixed sampling rate, the length of the acoustic tube model can only change in discrete steps by changing N, not continuously as it should [Ref. B11]. Being able to scale the LPC coefficients to account for a change in sampling rate permits the length of the acoustic tube to vary continuously to match the vocal tract length.

How does vocal tract length then relate to the high order poles? A uniform tube open at one end and closed at the glottal end has resonances at:

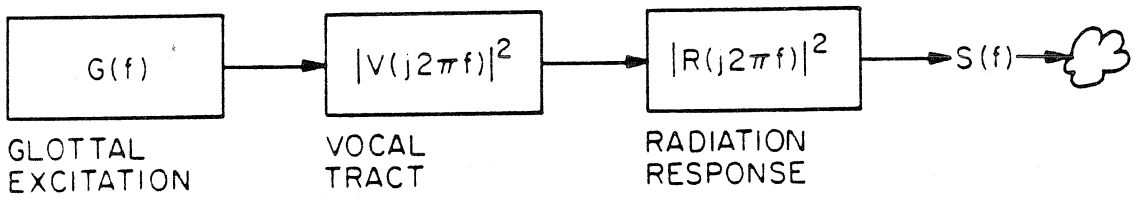
$$f_i = (i + \frac{1}{2}) \frac{c}{2L} \qquad \text{(B.21}$$

for $i = 0,1,2,\ldots$. Articulation causes the vocal tract to be a smooth departure from uniformity - the resonances (formants) are

locally displaced in frequency from the positions indicated by equation B.21 [Ref. B12]. But the average spacing of the resonances remains the same - it depends on length. In fact for the smooth, lossless tube of varying cross section, the resonance frequencies approach equation B.21 asymptotically [Ref. B13]. By selecting a change in sampling frequency such that the length of the acoustic tube model equals the length of the vocal tract, one is scaling the frequencies of the high order s-transform poles of the LPC model to equate pole densities between the LPC model and the vocal tract. The exact locations of the LPC high order poles differ from the vocal tract high order poles. But the density of poles in the LPC model is chosen to be the correct value - this provides the proper high order contribution to amplitudes of low order formants.

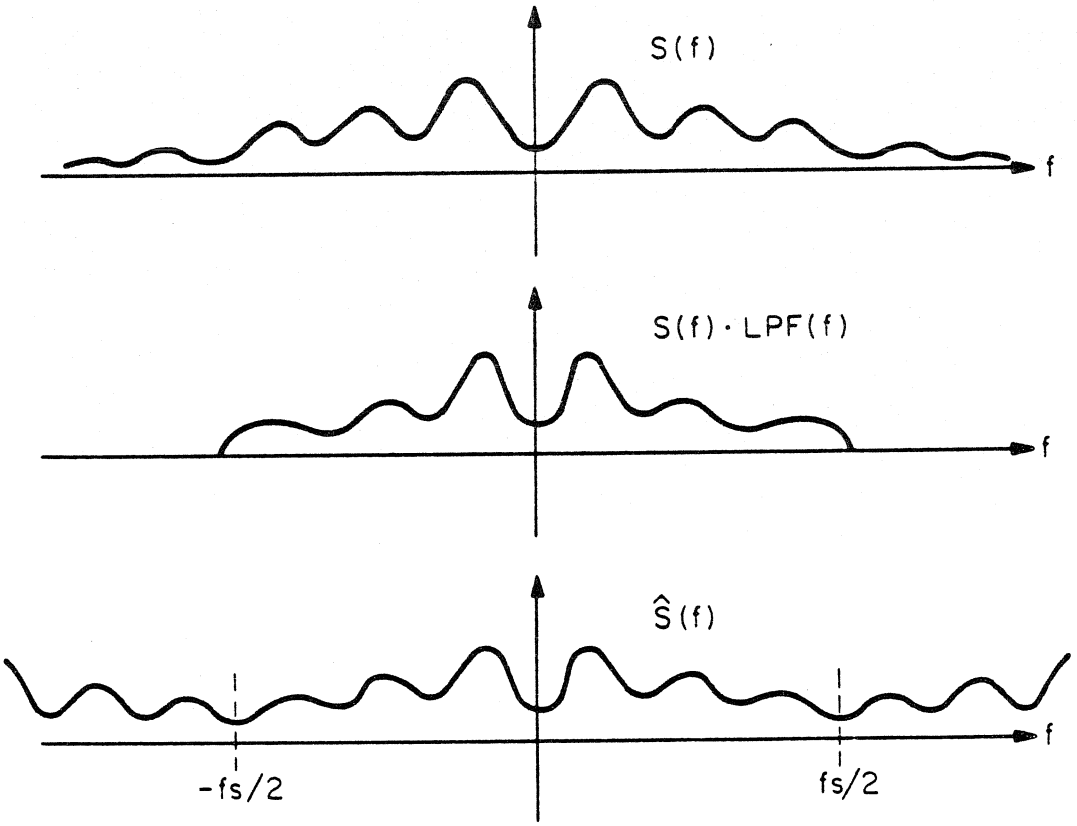
So we are able to scale the coefficients of the LPC filter model to allow for a change in the implicit sampling rate of the discrete time model. Changing the sampling rate results in a proportional change in length of the acoustic tube model obtained from the LPC coefficients. Ideally, the sampling rate is scaled to equate the acoustic tube length with the actual vocal tract length. In the s-plane representation of the LPC filter model frequency response, scaling the sampling rate controls the spacing in frequency of the high order poles. By equating lengths, the high order s-plane poles of both the model and the vocal tract have the same average spacing in frequency. This means that the high order poles of the model give the correct contribution to the amplitude peaks of the low order formants, giving a good match between the model spectrum

and the vocal tract spectrum.



Linear model of voiced speech production.

Figure B1



The speech spectrum after low pass filtering and sampling.

Figure B2

APPENDIX C:

The Use of LPC Analysis on Periodic Waveforms such as Voiced Speech

Linear predictive coding (LPC) analysis is a technique for characterizing the transfer function of a discrete time system [Refs. C1-C3]. One attempts to determine the transfer function by observing the output of the system and by inferring the input to the system. The inference about the input is based upon assumptions about the statistical properties of the input.

Consider the time series e(n) to be the input and the series s(n) to be the output of a discrete time linear system. We assume that the input samples e(n) are uncorrelated from each other. The correlations between samples of s(n) are assumed to result from the linear system. The input e(n) is recovered by uncorrelating s(n):

$$e(n) = s(n) - \hat{s}(n) \qquad \text{(C.1}$$

$$\hat{s}(n) = -\sum_{i=1}^{N} a_i \, s(n-i) \qquad \text{(C.2}$$

The input e(n) is taken to be what remains of s(n), or the residual, after predicting what s(n) should be by taking a linear combination of the previous N samples. The correct values for the a_i's are

those that minimize the squared error of the prediction residual, the energy in e(n). The a_i's are obtained by solving the linear system:

$$\sum_{i=1}^{N} a_i c_{ik} = -c_{0k} \qquad \text{(C.3}$$

for k = 1,...,N. For an analysis frame consisting of samples n_0 thru n_1, inclusive, the c_{ik}'s are calculated as the autocorrelations of the data:

$$c_{ik} = \sum_{n=n_0}^{n_1} s(n-i)\, s(n-k) \qquad \text{(C.4}$$

and the quantity being minimized is:

$$\alpha = \sum_{n=n_0}^{n_1} e(n)^2 \qquad \text{(C.5}$$

This procedure is called the covariance method of LPC analysis. If the sequence s(n) has been multiplied by some window which causes s(n) to be of finite energy, we can let $n_0 \rightarrow -\infty$ and $n_1 \rightarrow +\infty$ in equation C.4, causing the c_{ik}'s to have the property:

$$c_{ik} = c(|i-k|) \qquad \text{(C.6}$$

The procedure of windowing the data and applying an infinite analysis interval to the windowed data is called the auto-

correlation method.

The filter which recovers e(n) from s(n), the inverse filter, has a transfer function:

$$A(z) = 1 + \sum_{1}^{N} a_i z^{-i} = \sum_{0}^{N} a_i z^{-i} \qquad \text{(C.7}$$

where $a_0=1$. As this is a discrete time filter, its transfer function has been expressed as the z-transform of its impulse response. The forward filter, that which obtains s(n) from e(n), has transfer function:

$$F(z) = \frac{1}{\sum_{i=0}^{N} a_i z^{-i}} \qquad \text{(C.8}$$

The function F(z) forms an estimate of the transfer function of the linear sysem being analysed. F(z) is an all pole filter, and A(z) is an all zero filter.

Next, the use of the LPC analysis technique on voiced speech signals is considered. During voiced speech, the vocal tract is excited by a series of glottal pulses. The rate at which the pulses occurs is the pitch frequency and the time between pulses is the pitch period. Consider the hypothetical situation of a voiced sound of constant pitch for all time from $-\infty$ to $+\infty$. In this case the resulting speech signal is a periodic waveform. Assume also that the pitch period is an integral multiple of the interval

between samples of the sampled speech waveform. The sampled speech waveform is a time series with the property:

$$s(n) = s(n + kM) \qquad \text{(C.9}$$

where k is an arbitrary integer and M is the number of samples in a pitch period. The analysis frame is chosen to have M samples - to be one pitch period long. In this case the coefficients c_{ik} satisfy equation C.6, meaning that the covariance method is exactly the same as the autocorrelation method with a pitch period long rectangular window.

An important question to consider is how accurately LPC analysis characterizes a linear system excited periodically as in the case of idealized voiced speech. References C1 (pp.188-189) and C4 describe an experiment where a one resonance vocal tract model is excited with a periodic train of impulses. LPC analysis is performed over a pitch period, and the LPC polynomial A(z) is factored to locate the pole pair of F(z) [Eqn. C.8] which corresponds to the resonance in question. Significant errors occur in estimating the resonance frequency, and large errors occur in estimating the resonance bandwidth. These errors are worse as the pitch frequency of the excitation is increased.

The power spectrum of a periodic time series is a line spectrum cal-

culated as:

$$S(f) = \sum_{i=-\infty}^{\infty} \delta(F - i f_p) \left| \sum_{k=0}^{M-1} s(k) e^{-j 2\pi \frac{ki}{M}} \right|^2 \qquad (C.10$$

where $j = \sqrt{-1}$, and f_p is the pitch frequency. As the pitch frequency is increased, the lines of S(f) are spaced farther apart. It would appear that errors would be made in estimating the frequency and bandwidth of a resonance peak of S(f) because as f_p increases, S(f) is more and more undersampled. But more explanation of the source of these errors is required.

The maximum entropy formulation of LPC analysis [Refs. C5,C6] is useful in giving insight into the source of the voicing periodicity errors and what can be done to correct them. Maximum entropy analysis uses the the magnitude squared of the amplitude response, or spectrum, of the LPC forward filter to estimate the power spectrum of a signal s(t). S(f) is the estimate of the power spectrum of a signal s(t), which is sampled as a time series s(n) for purposes of the LPC analysis:

$$S(f) = \left| g F\left(e^{j 2\pi f T_s}\right) \right|^2 = \left| \frac{g}{\sum_{i=0}^{N} a_0 e^{-j 2\pi f T_s}} \right|^2 \qquad (C.11$$

where T_s is the interval between samples of s(n), and g is a gain factor.

Information theoretic considerations suggest that the "best" spectral estimate S(f) is one that maximizes an "entropy gain" criterion:

$$H = \int_{-f_s/2}^{f_s/2} \text{Log}\left[S(F)\right] \, dF \qquad (C.12$$

The spectral estimate is subject to the constraints:

$$\rho_n = \int_{-f_s/2}^{f_s/2} S(F) e^{-j 2\pi F n T_s} \, dF \qquad (C.13$$

where n = -N to N, and $j = \sqrt{-1}$. The quantities ρ_n are the autocorrelation lags estimated from the data. Equation C.13 essentially asserts that the first N autocorrelation lags of the spectral estimate should equal the first N lags estimated from the data. The N+1 lag and beyond of S(f) are left unconstrained - the determination of the high order lags is made in maximizing H. If the ρ_n's are estimated by windowing and correlating s(n), the solution to the maximization problem is to make S(f) proportional to the spectrum of the LPC forward filter, the coefficients of which are evaluated by the autocorrelation method.

Consider a periodic signal to be the output of a linear system excited by a periodic train of impulses. In this case autocorrelation lags estimated from a pitch period long analysis frame

obey the relationship:

$$\rho_n = \int_{-f_s/2}^{f_s/2} S(f)\left[\sum_{i=0}^{M-1} \delta\left(f - F_p \frac{i}{M}\right) \frac{F_s}{M}\right] e^{-j2\pi f n T_s} \, dF \qquad (C.14$$

where n = -N to N, $j = \sqrt{-1}$, f_s is the pitch frequency, and S(f) is now the spectrum of the linear system. But if the spectrum of the LPC forward filter is to estimate the spectrum of the linear system, the values of the ρ_n's should be those specified in equation C.13. The autocorrelation lags calculated from s(n) only approximate the autocorrelation lags of the spectrum of the linear system. The higher the pitch period, the cruder the approximation.

Can the maximum entropy formulation be rederived by basing the constraints on equation C.14? Substituting the proper formulation of the constraints would insure that the line spectrum of the LPC model would properly estimate the line spectrum of the data. To solve for the spectral estimate S(f), a variational technique is employed. The maximum of H occurs for a value of S(f) where the variation in H is 0:

$$0 = \delta H = H\left(S(F) + \varepsilon(F)\right) - H\left(S(F)\right) \qquad (C.15$$

Employing equation C.12:

$$0 = \delta H = \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} \mathrm{Log}\left(1 + \frac{\mathcal{E}(f)}{S(f)}\right) df \sim \int_{-\frac{f_s}{2}}^{\frac{f_s}{2}} \left[\frac{\mathcal{E}(f)}{S(f)}\right] \quad (C.16$$

For both S(f) and S(f)+ (f) to satisfy equation C.14, the constraints, we have:

$$0 = \int_{-\frac{f_s}{2}}^{f_s/2} \mathcal{E}(f)\left[\sum_{i=0}^{M-1} \delta\left(f - F_p\frac{i}{M}\right)\frac{f_s}{M}\right] e^{-j2\pi f\, nT_s}\, df$$

$$= \int_{-\frac{f_s}{2}}^{f_s/2} \mathcal{E}(f)\left[\sum_{i=0}^{M-1} \delta\left(f - F_p\frac{i}{M}\right)\frac{f_s}{M}\, e^{-j2\pi\frac{ni}{M}}\right] df \quad (C.17$$

where $j = \sqrt{-1}$ and n = -N to N. Equation C.16 has 1/S(f) perpendicular to $\mathcal{E}(f)$ - their inner product measure is zero. Now $\mathcal{E}(f)$ is perpendicular to a set of δ pulse trains [Eqn. C.17]. This means 1/S(f) is colinear with the δ pulse trains:

$$\sum_{i=0}^{M-1} \delta\left(f - F_p\frac{i}{M}\right)\frac{f_s}{M}\, e^{-j2\pi\frac{ni}{M}} \quad (C.18$$

where n = -N to N. So:

$$\frac{1}{S(f)} = \sum_{n=-N}^{N} c_n \left[\sum_{i=0}^{M-1} \delta\left(f - F_p \frac{i}{M}\right) \frac{f_s}{M} e^{-j 2\pi \frac{ni}{M}} \right] \qquad (C.19$$

which means that:

1) The maximum entropy spectrum estimate S(f) is a line spectrum, the lines occurring at multiples of the pitch frequency.

2) 1/S(f) is the Fourier tranform of an autocorrelation function having no more than N lags, the values of the lags being proportional to the $c_{|i-k|}$'s.

That 1/S(f) has a finite number of autocorrelation lags means that 1/S(f) can be expressed as the spectrum of an all zero linear discrete time filter having a finite number of coefficients. Then S(f) can be expressed as an all pole filter having a finite number of coefficients. But solving for these coefficients involves the difficult matter of substituting the spectrum of this all pole filter into equation C.14. All is not lost. We have at least determined the proper form of the filter model which has spectrum S(f). An all pole discrete time filter of a finite number of coefficients can be modelled as an acoustic tube having a finite number of sections [Appendix A]. As the speech spectrum results from exciting a vocal tract of finite length, there is physical justification for estimating the spectrum as that of an acoustic tube of a finite number of sections. The number of sections N, also the number of filter coefficients, should be large enough that the

length of the acoustic model is at least as long as the vocal tract. The maximum entropy argument merely specifies the use of N coefficients in the filter model because N is the number of autocorrelation lags given. Physical arguments uphold the use of the type of filter specified by maximum entropy and go a step beyond, specifying what N should be.

Information as to the number of autocorrelation lags of 1/S(f) can be put to use in solving the voicing periodicity problem. Suppose we calculate the line spectrum of an estimate S(f) of the linear system spectrum. This is done by taking the magnitude squared of the Fourier transform of a pitch period frame of s(n). The voicing periodicity errors would be diminished if additional spectral lines of S(f) could be determined; it would be as if one had a much lower pitch frequency making the lines closer together. It is not clear how to interpolate S(f). But one is able to interpolate 1/S(f) correctly. Just as a sampled time function can be completely recovered if it is bandlimited, 1/S(f) can be completely recovered from its sample values because it is limited in time - the autocorrelation lags are finite in number. The autocorrelation lags can be extended by zeroes, and by taking the Fourier transform, one can obtain 1/S(f) (hence also S(f)) at an arbitrarily dense set of sample values. Taking the inverse transform of the denser line spectrum representation of S(f) gives values for the ℓ_n's from which one can calculate LPC coefficients. Calculating coefficients from these corrected ℓ_n's results in reduced voicing periodicity error.

Atal proposes a different solution to the voicing periodicity error problem [Ref. C4]. It is suggested that one increase the number of LPC coefficients to the point that each spectral line is represented by an LPC pole pair. The LPC spectrum would no longer estimate the envelope of the line spectrum of the data; the LPC spectrum would follow the fine structure, the spectral lines. The pole pairs corresponding to vocal tract resonances would be found by factoring out the spectral line pole pairs. But increasing the number of LPC coefficients to this degree would reduce the energy in the prediction residual to the point where numerical instabilities would occur in the calculations [Ref. C7]. Atal has not shown that this method actually works, suggesting this to be the case.

The concept of interpolating the spectral lines of $1/S(f)$ can be viewed as an application of spectral root homomorphic deconvolution, or SRHD [Ref. C8]. One cannot recover $S(f)$ from its sample values because the impulse responses of the linear system being excited by periodic impulses overlap. But we raise $S(f)$ to the -1 power, $1/S(f)$ can be completely recovered from sample values because the impulse responses of the filter having spectrum $1/S(f)$ do not overlap. Taking a spectral root by evaluating $S(f)^{\gamma}$, $\gamma = -1$ in this instance, deconvolves the system impulse response from the periodic excitation function. A value of γ is used where the linear system having spectrum $S(f)^{\gamma}$ has a finite duration impulse response.

All along it has been assumed that the linear system having the spectrum $S(f)$ is all pole. The LPC analysis models this system as being

all pole. But because a system is modelled as being all pole does not necessarily mean that it is. The vocal tract is all pole, but the speech spectrum is also influenced by the glottal and radiation spectra as well as the vocal tract spectrum. If speech were all pole, the proper value of γ is -1. If S(f) is the spectrum of a system having only zeroes, the value of γ to use is +1. If the system has both poles and zeroes, some value of γ has to be chosen between +1 and -1. At this stage there is only limited empirical evidence as to what value of γ is best for speech [Ref. 8].

It has been discovered that LPC analysis results in systematic errors in modelling a linear system excited to give a periodic output. These errors are of concern because voiced speech is locally periodic. These errors become more severe as the pitch frequency is increased. The source of the error can be regarded as an undersampling of the spectrum S(f) of the linear system. The undersampling comes from the line spectrum imposed by the periodic nature of the waveforms. Having the spectrum sampled at periodic multiples of the pitch frequency as in this instance does not represent a fundamental lack of information however. If the linear system were all pole, the spectral lines of 1/S(f) could be interpolated, enabling S(f) to be more completely specified. Given that the speech spectrum has additional components besides the all pole vocal tract spectrum, one can interpolate $S(f)^{\gamma}$ where $1 > \gamma > -1$. What value of γ is effective is an open research question at this point.

APPENDIX D:

A Linear Time Varying Model of Glottal Excitation of the Vocal Tract

The sound produced in voiced speech sounds originates as the vibration of the glottal folds, the vocal cords. The vocal cords are brought to vibrate by air forced thru them by lung pressure. The vibration of the glottal folds excites acoustic waves in the vocal tract. The vocal tract has a multiplicity of resonant modes which shape the spectrum of the resulting sound.

The time invariant linear model is the most basic representation of the coupling between the glottis and the vocal tract. Vibration of the vocal cords results in a certain wave shape of the acoustic volume velocity at the glottal end of the vocal tract. The vocal tract is regarded as having a linear time invariant transfer function relating volume velocity at the mouth to volume velocity at the lips. This transfer function operates on the glottal wave shape as a linear filter would to give the acoustic wave at the lip end of the vocal tract. Another linear time invariant transfer function, the radiation response, relates lip volume velocity to far field acoustic pressure, the quantity measured by a recording device.

This model ignores two important factors: the effect on the glottal volume velocity wave of acoustic loading of the vocal tract and the variation in vocal tract damping with glottal opening [Ref.

D1]. The first factor refers to the fact that the glottal waveform is influenced by the impedance of the vocal tract as well as lung pressure, vocal cord mass, and vocal cord tension. The linear model is still valid if one takes whatever volume velocity waveform occurs at the glottis to be the excitation waveform, and then allows for the variation of this waveform with changes in vocal tract loading that occur with articulation. The problem, however, of estimating the glottal waveform so that one can determine the vocal tract transfer function from the speech waveform is made more difficult. One cannot now know a priori what the glottal pulse shape is. The second factor is one that may invalidate the linear model entirely. As the vocal cords vibrate, the area of the glottal opening is changing. The damping of vibrations in the vocal tract depends on the degree of opening. The transfer function is not time invariant, but it is varying in synchronism with the vocal cord vibration.

What is being discussed in this section is a linear time varying model of the vocal cord - vocal tract coupling. This model can be incorporated into a dynamical simulation of an acoustic tube model of the vocal tract which can produce synthetic speech sounds. The resulting synthetic speech can be analysed using methods that assume a linear time invariant method of speech production so that the degree of error that results from this assumption can be determined. The linear time varying model is not intended to be a very accurate representation of the actual speech production mechanism. It is intended to produce results similar enough to the actual speech

mechanism that similar errors will result in the analysis methods that are based upon time invariance.

A model for vocal cord - vocal tract coupling is shown in Figure D1. This model and related acoustical calculations shown are based on the analysis done by Flanagan [Ref. D2 pp.43-53]. The lungs are modelled as a constant pressure source. This pressure forces a volume velocity flow thru the glottal constriction, a constriction which varies as the vocal cords vibrate. The glottal volume velocity flow feeds directly into the pharynx cavity at the base of the vocal tract. It is possible to approximate this model of the voicing mechanism with the model shown in Figure D2. The voicing source is regarded as a volume velocity source that produces glottal pulses. This source is in parallel with a time varying glottal resistance.

The pharnyx cavity has associated with it a characteristic waveguide impedance:

$$Z_p = \frac{\rho c}{A_p} \qquad \text{(D.1}$$

where ρ is the density of air, c is the speed of sound, and A_p is the cross sectional area of the pharynx cavity. The impedance value given is all resistive, a fact that is based on assuming the pharynx to be lossless and having hard walls. The glottal constriction has a resistance associated with it that is a function of the volume velocity thru

the glottis and the area of the glottal constriction:

$$R_G(U_G, A_G) = \frac{U_G}{2\ A_G^2} \qquad \text{(D.2}$$

This resistance is aerodynamic in origin; it is based on the pressure required to force an incompressible fluid past a constriction. The pharynx cavity impedance, however, is sonic in origin; it relates the magnitude of pressure variation to volume velocity variation in a travelling sound wave. The glottis also has an inductance associated with it resulting from the effects of the air mass in the glottal constriction. This inductance will be ignored in this analysis, something that can be done for frequencies up to several hundred Hertz.

During phonation, there will be forward and backward going acoustic wave components in the pharynx cavity. The acoustic volume velocity and pressure at the juncture with the glottis can be expressed in terms of the wave components at this point:

$$U_G = U_F - U_B \qquad \text{(D.3}$$

$$P_G = Z_p\ (U_F + U_B) \qquad \text{(D.4}$$

The quantity P_G, total pressure at the juncture between glottis and pharynx, can also be expressed in terms of lung pressure minus the

pressure drop past the glottal constriction:

$$P_G = P_S - U_G R_G = P_S - \frac{\rho}{2A_g^2} U_G^2 \qquad \text{(D.5}$$

Combining equations D.4 and D.5 gives:

$$P_S = \frac{\rho}{2A_G^2} (U_F - U_B)^2 + Z_p (U_F + U_B) \qquad \text{(D.6}$$

Solving this quadratic relation for the forward going volume velocity wave component gives:

$$U_F = U_B - \frac{A_G^2}{\rho} Z_p + A_G \sqrt{\frac{2P_S}{\rho}} \sqrt{1 - \frac{2Z_p U_B}{P_S} + \frac{A_G^2 Z_p^2}{2\rho P_S}} \qquad \text{(D.7}$$

If the pharynx cavity has a sufficiently large cross sectional area, the quantity Z_p, the pharynx cavity characteristic impedance, will be sufficiently small that one can approximate equation D.7 by:

$$U_F = U_B - \frac{A_G^2}{\rho} Z_p + A_G \sqrt{\frac{2P_S}{\rho}} \left(1 - \frac{Z_p U_B}{P_S} + \frac{A_G^2 Z_p^2}{4\rho P_S}\right) \qquad \text{(D.8}$$

Saving the terms up to order A_G^2 one obtains:

$$U_F = (1-A_G\sqrt{\frac{2}{\rho P_S}}\,Z_p)U_B + (1-A_G\sqrt{\frac{1}{2\rho P_S}}\,Z_p)A_G\sqrt{\frac{2P_S}{\rho}} \qquad \text{(D.9}$$

$$U_F = C\,U_B + \frac{1}{2}(1 + C)\,A_G\sqrt{\frac{2P_S}{\rho}} \qquad \text{(D.10}$$

where:

$$C = (1 - \frac{A_G}{A_G^{max}}\,Q) \qquad \text{(D.11}$$

and:

$$Q = A_G^{max}\sqrt{\frac{2}{\rho P_S}}\,Z_P = \frac{A_G^{max}}{A_p}\,c\sqrt{\frac{2\rho}{P_S}} \qquad \text{(D.12}$$

C: reflection coef. c: sound velocity

Q being the glottal coupling coefficient. The quantity A_G^{max} refers to the maximum area of the glottal constriction in the course of a glottal pulse.

Consider now the glottal model depicted in Figure D2. If the volume velocity source has the value:

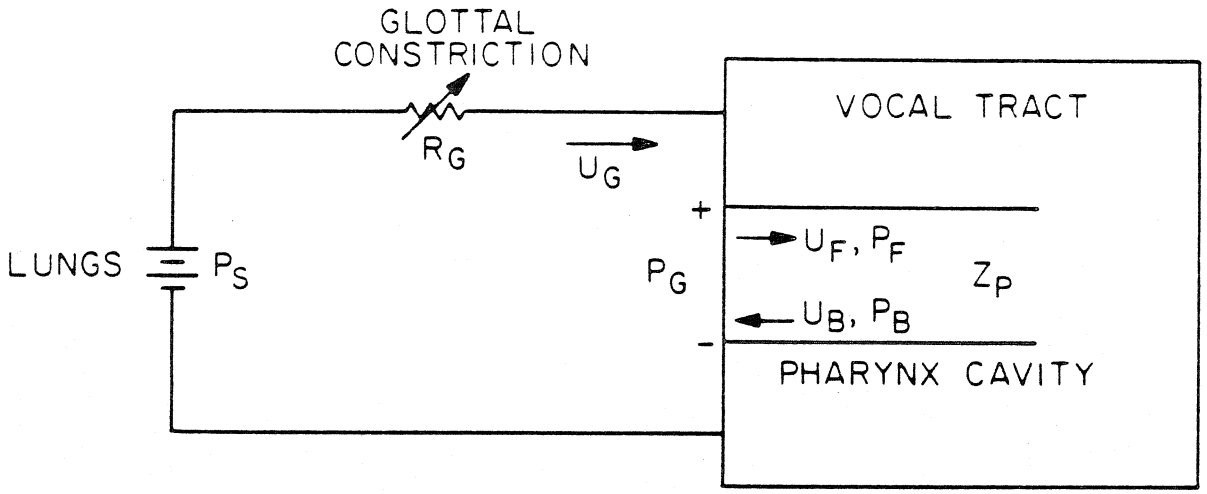
$$U_G + A_G\sqrt{\frac{2P_S}{\rho}} \qquad \text{(D.13}$$

and the source resistance has the form:

$$\hat{R}_G = \left(\frac{1+C}{1-C} \right) Z_p \qquad \text{(D.14}$$

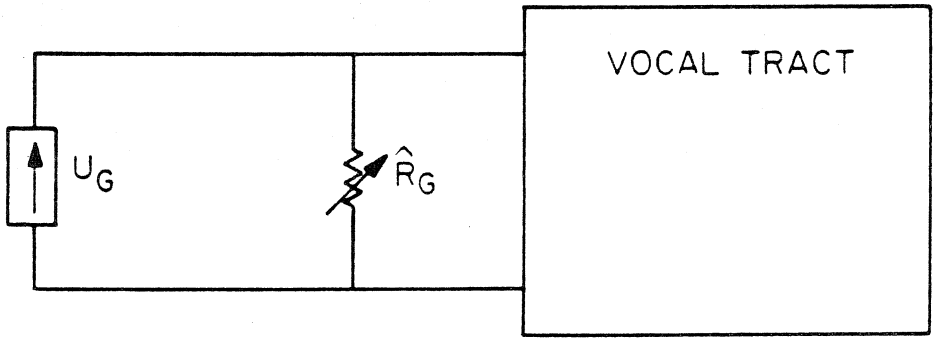
where C and Q are given by equation D.12 and D.13, then the vocal tract boundary conditions will be given by equation D.10 [Appendix A]. This means that the model in Figure D2, the time varying linear model, is a valid approximation to the physical model in Figure D1. The glottal volume velocity source is proportional to glottal area and the glottal source resistance is a rational function of glottal opening area.

What has been established is that by making certain approximations, the time varying linear model of voicing can represent the physical model of voicing production. The approximations are made under the assumption that the pharynx cavity area is large compared to the glottal opening. This is equivalent to assuming that the glottal damping factor Q is small compared to 1. The independent quantity in the model is the area of the glottal constriction. The variation of this area with time to produce glottal pulses is taken as a given. Glottal volume velocity is proportional to glottal area, and glottal source resistance is a function of glottal area. This linear time varying model of the voicing source can be used in synthesizing speech sounds so that speech analysis techniques that assume time invariance can be tested for accuracy.



Equivalent circuit vocal cord model.

Figure D1



Linear time varying glottal termination.

Figure D2

APPENDIX E:

Correcting the Area Function Estimate for the Effects of the Radiation Load at the Lips

There are two formulations of acoustic tube models of the vocal tract [Appendix A] that permit direct calculation of cross sectional areas from speech sounds. Unfortunately both formulations have boundary conditions that differ significantly from actual vocal tract boundary conditions, raising doubts as to whether the cross sectional areas of either acoustic tube model accurately represent the actual vocal tract areas. The Wakita formulation, in particular, has all losses lumped into the source resistance of the voicing source located at the glottal end of the tube; the walls are hard and lossless. The lip, or mouth, end of the tube is terminated in an acoustic short circuit. The actual vocal tract has wall losses, and it has a lesser amount of loss at the glottis. In addition, the lip end of the vocal tract is terminated in an acoustic radiation load that has both inductive and resistive impedance.

Estimating vocal tract areas using the the Wakita acoustic tube model gives values for the area of the lip opening that are too small. It is possible to adjust the area of the acoustic tube model section at the lip end of the tube to allow for the added inductance of the radiation load [Ref. E1]. This correction is only valid for the Wakita formulation of the acoustic tube vocal tract model. The following is a

description of the acoustics and formulas behind the correction.

The radiation load at the lips can be regarded as a series combination of a resistance that goes as the square of frequency and an inductive reactance that is proportional to frequency [Ref. E1 P36]. At zero frequency this load is simply an acoustic short circuit, the same as the termination in the Wakita model. For low frequencies, however, the inductive reactance is very low but the resistance is even lower. For the low frequency range of the speech spectrum the radiation load is best characterized as having an inductance:

$$L_R = \frac{8\rho}{3\pi^{3/2}\sqrt{A_M}} = \frac{k}{A_M} \qquad \text{(E.1}$$

where ρ is the density of air and A_M is the area of the mouth, or lip, opening. Consider now the section of the acoustic tube model in the Wakita formulation that is at the lip boundary. The combination of this segment and the short circuit termination has an inductance that results from the inertia of the air column in this last segment:

$$L_T = \frac{\rho \Delta x}{A_M} \qquad \text{(E.2}$$

The quantity Δx is the length of the acoustic tube section.

In the Wakita model, the inductance looking into the last section toward the lips only includes the effect of the air column inertia in

that section because the lip termination is a short. To be consistant with the actual vocal tract, this lip termination ought to be an inductance, a termination that is asymptotically correct for low frequencies. The inductance looking into the last section would then be the sum of air column plus radiation load inductance. This relationship can be expressed as:

$$L = L_R + L_T \qquad \text{(E.3}$$

where L is air column inductance corresponding to the uncorrected lip opening of the acoustic tube model, L_R is the radiation load inductance corresponding to the true lip opening, and L_T is the air column inductance corresponding to the true lip opening. Substituting values obtained by equations E.1 and E.2 gives:

$$\frac{\rho \Delta x}{\hat{A}_M} = \frac{k}{\sqrt{A_M}} + \frac{\rho \Delta x}{A_M} \qquad \text{(E.4}$$

where $\hat{A}_M$ is the uncorrected lip opening and A_M is the true lip opening. Solving this quadratic relationship for the true lip opening from the uncorrected lip opening area gives:

$$\sqrt{A_M} = \frac{kA_M}{2\rho \Delta x} + \sqrt{\left(\frac{kA_M}{2\rho \Delta x}\right)^2 + \hat{A}_M} \qquad \text{(E.5}$$

where k has been defined in equation E.1.

An estimate of the true lip opening area can be calculated from the uncorrected lip opening area obtained from the Wakita formulation of the acoustic tube model. This correction is based on considering the low frequency asymptotic impedance of the radiation load - that of an inductance. The difficulty with this method is that in calculating the acoustic tube model corresponding to speech sounds it is area ratios, not absolute area values, that are specified [Appendix A]. To effectively apply this correction, the proper scaling of the acoustic tube model areas to relate area ratios to absolute area values is needed. This correction has been demonstrated to be effective in a case where the scaling is known [Ref. E1]. To apply this correction to area function analysis of actual speech, however, some means of determining the absolute scale of the acoustic tube model areas needs to be developed.

REFERENCES:

1. J. L. Flanagan, Speech Analysis Synthesis and Perception, Second Edition, Springer-Verlag, New York 1972.

2. B. S. Atal, J. J. Chang, M. V. Mathews, J. W. Tukey, "Inversion of Articulatory-to-Acoustic Transformation in the Vocal Tract by a Computer Sorting Technique," Journal of the Acoustical Society of America, Vol. 63, No. 5, May 1978, pp. 1535-1555.

3. G. Fant, Acoustic Theory of Speech Production, Mouton, The Hague, 1970.

4. L. R. Rabiner, R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

5. M. R. Schroeder, "Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements," J. Acoust. Soc. Am., Vol. 41, 1967, pp. 1002-1010.

6. M. M. Sondhi, "Determination of Vocal Tract Shape from Impulse Response at the Lips," J. Acoust Soc. Am., Vol. 49, 1971, pp. 1867-1873.

7. B. S. Atal, S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., Vol. 50, No. 2 part 2, August 1971, pp. 637-655.

8. H. Wakita, "Direct Estimation of the Vocal-Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, October 1973, pp. 417-427.

9. J. D. Markel, A. H. Gray, Jr., Linear Prediction of Speech, Springer-Verlag, New York, 1976.

10. H. Wakita, "Estimation of Vocal-Tract Shapes from Acoustical Measurements: The State of the Art," IEEE Transactions on Acoustics Speech and Signal Processing, Vol. ASSP-27, No. 3, June 1979, pp. 281-285.

11. J. Makhoul, "Spectral Linear Prediction: Properties and Applications," IEEE Trans. on ASSP, Vol. ASSP-23, No. 3, June 1975, pp. 283-296.

12. J. Makhoul, "Spectral Analysis of Speech by Linear Prediction," IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No. 3, June 1973, pp. 140-148.

13. J. Makhoul, "Linear Prediction: A Tutorial Review," Proceedings of the IEEE, Vol. 63, April 1975, pp. 561-580.

14. R. Viswanathan, J. Makhoul, "Quantization Properties of Transmission Parameters in Linear Predictive Systems," IEEE Trans. on ASSP, Vol. ASSP-23, No. 3, June 1975, pp. 309-311.

15. J. P. Olive, N. Spickenagel, "Speech Resynthesis from Phoneme Related Parameters," J. Acoust. Soc. Am., Vol. 59, No. 4, April 1976, pp. 993-996.

16. Hans Werner Strube, "Can the Area Function of the Human Vocal Tract be Determined from the Speech Wave," Drittes Physikalishes Institut, Universitat Gottingen, Germany.

17. M. M. Sondhi, "Estimation of Vocal-Tract Areas: The Need for Acoustic Measurements,"IEEE Trans. on ASSP, Vol. ASSP-27, No. 3, June 1979, pp. 268-273.

18. R. B. Monsen, A. M. Engebretson, "Study of Variations in the Male and Female Glottal Wave," J. Acoust. Soc. Am., Vol. 62, No. 4, October 1977, pp. 981-993.

19. B. Gopinath, M. M. Sondhi, "Determination of the Shape of the Human Vocal Tract from Acoustical Measurements," Bell System Technical Journal, Vol. 49, 1970, pp. 1195-1214.

20. M. M. Sondhi, "Model for Wave Propagation in a Lossy Vocal Tract," J. Acoust. Soc. Am., Vol. 55, 1974, pp. 1070-1075.

21. T. Nakajima, H. Omura, K. Tanaka, S. Ishizaki, "Estimation of Vocal Tract Area Functions by Adaptive Filtering Methods and Identification of the Articulatory Model," Proceedings of the Speech Communication Seminar, Stockholm, August 1-3, 1974.

22. P. Mermelstein, "Determination of the Vocal-Tract Shape from Measured Formant Frequencies," J. Acoust. Soc. Am., Vol. 41, 1967, pp. 1283-1294.

23. A. Paige, V. W. Zue, "Computation of Vocal Tract Area Functions," IEEE Trans. Audio Electroacoustics, Vol. AU-18, 1970, pp. 7-18.

24. Hisashi Wakita and Augustine H. Gray, Jr., "Numerical Determination of the Lip Impedance and Vocal Tract Area Functions," IEEE Trans. on ASSP, Vol. ASSP-23, No. 6, December 1975, pp. 574-580.

25. Hisashi Wakita, "Normalization of Vowels by Vocal-Tract Length and its Application to Vowel Identification," IEEE Trans. on ASSP, Vol. ASSP-25, No. 2, April 1977, pp. 183-192.

26. David Y. Wong, John D. Markel, Augustine H. Gray, Jr., "Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform," IEEE Trans. on ASSP, Vol. ASSP-27, No. 4, August 1979, pp. 350-355.

27. S. P. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Trans. on ASSP, Vol. ASSP-22, No. 2, 1974, pp. 135-141.

28. Chong Kwan Un, "A Low-Rate Digital Formant Vocoder," IEEE Transactions on Communications, Vol. COM-26. No. 3, March 1978, pp. 343-354.

29. John D. Markel, "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation," IEEE Trans. Audio Electroacoustics, Vol. AU-20, No. 2, June 1972, pp. 129-137.

30. L. R. Rabiner, B. S. Atal, M. R. Sambur, "LPC Prediction Error - Analysis of its Variation with Position of the Analysis Frame," IEEE Trans. on ASSP, Vol. ASSP-25, No. 5, October 1977, pp. 434-442.

31. H. Strube, "Determination of the Instant of Glottal Closure from the Speech Wave," J. Acoust. Soc. Am., Vol. 56, No. 5, November 1974, pp. 1625-1629.

32. Paul Milenkovic, B. S. Atal, "Improved Estimate of Vocal-Tract Areas from the Speech Wave," Paper EE4, 98th Meeting Acoustical Society of America, Salt Lake City, Utah, 26-30 November, 1979.

33. B. S. Atal, "Linear Prediction of Speech - Recent Advances with Applications to Speech Analysis," Speech Recognition, Academic Press, New York, 1975, pp. 221-230.

34. Peter Ladefoged, Richard Harshman, Louis Goldstein, Lloyd Rice, "Generating Vocal Tract Shapes from Formant Frequencies," J. Acoust. Soc. Am., Vol. 64, No. 4, October 1978, pp. 1027-1035.

35. Richard Harshman, Peter Ladefoged, Louis Goldstein, "Factor Analysis of Tongue Shapes," J. Acoust. Soc. Am., Vol. 62, No. 3, September 1977, pp. 693-707.

36. C. J. Riordan, "Control of Vocal-Tract Length in Speech," J. Acoust. Soc. Am., Vol. 62, No.4, October 1977, pp. 998-1002.

37. B. E. F. Lindblom, J. E. F. Sundberg, "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movements," J. Acoust. Soc. Am., Vol. 50, 1971, pp. 1166-1179.

38. G. Fant, "Glottal Source and Excitation Analysis," Speech Transmission Laboratory Quarterly Progress and Status Report, Vol. STL-QPSR 11, January-March 1979, pp. 85-107.

39. R. Viswanathan, J. Makhoul, "Efficient Lattice Methods for Linear Prediction," Programs for Digital Signal Processing, IEEE Press, New York, 1979, pp. 4.2-1 thru 4.2-14.

40. Carl-Erik Froberg, Introduction to Numerical Analysis, Second

Edition, Addison-Wesley Publishing Company, Reading, Massachusetts, 1969.

41. A. H. Gray, J. D. Markel, "A Spectral-Flattness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis," IEEE Trans. on ASSP, Vol. ASSP-22, June 1974, pp. 207-217.

42. M. M. Sondhi, "Measurement of the Glottal Waveform," J. Acoust. Soc. Am., Vol. 57, 1975, pp. 228-232.

43. R. L. Miller, "Nature of the Vocal Cord Wave," J. Acoust. Soc. Am., Vol.31, June 1959, pp. 667-677.

44. J. N. Holmes, "An Investigation of the Volume Velocity Waveform at the Larynx During Speech by Means of an Inverse Filter," Proceedings of the Speech Communication Seminar, Stockholm, August 1962, pp. 1-4.

45. M. Rothenberg, "A New Inverse-Filtering Technique for Deriving the Glottal Air Flow During Voicing," J. Acoust. Soc. Am., Vol. 53, 1973, pp. 1632-1645.

46. M. V. Mathews, Joan E. Miller, E. E. David, Jr., "Pitch Synchronous Analysis of Voiced Sounds," J. Acoust. Soc. Am., Vol. 33, No. 2, February 1961, pp. 179-186.

47. Joan E. Miller, "Decapitation and Recapitation, A Study in Voice Quality," Paper J8, ASA Conference in October 1964, Austin, Texas.

48. M. R. Sumbur, A. E. Rosenberg, L. R. Rabiner, C. A. McGonegal, "On Reducing the Buzz in LPC Synthesis," J. Acoust. Soc. Am., Vol. 63, No. 3, March 1978, pp. 918-924.

49. A. E. Rosenberg, "Effect of Glottal Pulse on the Quality of Natural Vowels," J. Acoust. Soc. Am., Vol. 49, 1971, pp. 583-588.

50. Joan E. Miller, "PISA - Computer Programs for Pitch Synchronous Analysis," Bell Telephone Labaratory Technical Memorandum, February 23, 1967.

51. Lloyd J. Griffiths, "Rapid Measurement of Digital Instantaneous Frequency," IEEE Trans. on ASSP, Vol. ASSP-23, April 1975, pp. 207-222.

52. D. G. Nichol, R. E. Bogner, "Quasi-Periodic Instability in a Linear Prediction Analysis of Speech," IEEE Trans. on ASSP, Vol. ASSP-26, No. 3, June 1978, pp. 210-216.

53. J. L. Flanagan, K. Ishizaka, K. Shipley, "Synthesis of Speech from a Dynamic Model of the Vocal Cords and Vocal Tract," Bell

System Technical Journal, Vol. 54, 1975, pp. 485-506.

54. M. V. Mathews, Joan E. Miller, E. E. David, Jr., "Strategies for Automatic Pole-Zero Analysis of Speech," Paper B10, Proceedings of the Speech Communication Seminar, Stockholm, August 1962.

A1. Ref. 7.

A2. Ref. 8.

A3. Ref. 4.

A4. Ref. 9.

A5. A. V. Oppenheim, R. W. Schafer, Digital Signal Processing, Prentice Hall, Englewood Cliffs, New Jersey, 1975.

B1. J. D. Markel, A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method," IEEE Trans. on ASSP, Vol. ASSP-22, April 1974, pp. 124-134.

B2. Ref. 9.

B3. Ref. 11.

B4. Ref. 12.

B5. Ref. 7.

B6. Ref. 3.

B7. E. T. Whittaker, G. N. Watson, A Course of Modern Analysis, Cambridge University Press, London, 1973.

B8. Ref. 25.

B9. Ref. 36.

B10. Ref. 37.

B11. H. W. Strube, "Sampled-Data Representation of a Nonuniform Lossless Tube of Continuously Variable Length," J. Acoust. Soc. Am., Vol. 57, pp. 256-257.

B12. Johan Sundberg, "The Acoustics of the Singing Voice," Scientific American, March 1977.

B13. Ref. 23.

C1. Ref. 9.

C2. Ref. 13.

C3. Ref. 7.

C4. Ref. 33.

C5. J. G. Ables, "Maximum Entropy Spectral Analysis," Astron. Astrophys. Suppl. Series, Vol. 15, 1974, pp. 383-393.

C6. William I. Newman, "Extension to the Maximum Entropy Method," IEE Transactions on Information Theory, Vol. IT-23, January 1977, pp. 89-93.

C7. Ref. 51.

C8. Jae S. Lim, "Spectral Root Homomorphic Deconvolution System," IEEE Trans. On ASSP, Vol. ASSP-27, No. 3, June 1979, pp. 223-233.

D1. Ref. 38.

D2. Ref. 1.

E1. Ref. 32.

E2. Ref. 1.