# ESTIMATION OF THE VOCAL TRACT SHAPE FROM THE ACOUSTIC WAVEFORM

by

Douglas Baker Paul

B.E.S. The Johns Hopkins University 1971

S.M. Massachusetts Institute of Technology 1973

E.E. Massachusetts Institute of Technology 1976

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 1976

Signature redacted

Signature of Author .....
Department of Electrical Engineering and
Computer Science, August 9, 1976

Signature redacted

Certified by .....
Thesis Supervisor

Signature redacted

Accepted by .....
Chairman, Departmental Committee on Graduate Students

Estimation of the vocal tract shape

from the acoustic waveform

by

Douglas Baker Paul

Submitted to the Department of Electrical Engineering and Computer Science on August 9, 1976 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

## Abstract

X-ray cineradiographs have been the traditional method for obtaining vocal tract shapes used in speech production. Due to the dangers of X-ray exposure, several acoustic methods have been devised. Unfortunately, only half of the required information appears in the speech waveform. The proposed algorithm is an attempt to overcome this difficulty.

This algorithm is a sequence of operations on the formants of a non-nasal sonorant. The formants are first perturbed to remove yielding wall effects. Artificial bandwidths are applied to provide constraints and the tract length and area normalization are estimated. A cross-sectional area function is then generated by an LPC (Levinson recursion) conversion to an area function.

The accuracy of the algorithm is first tested by comparing the X-ray derived area functions of six Russian vowels with their acoustically derived area functions. It is then tested on two sets of English vowels. The algorithm is lastly tested on a set of continuous utterances containing English vowels and non-nasal sonorant consonants. Its results are reasonably accurate for almost all of the test utterances.

Table of Contents

## Illustrations and Tables

# Estimation of the vocal tract shape from the acoustic waveform

## I. Introduction

An aspect of interest for speech study is the vocal tract shape. This information about the vocal tract allows the phonetician to describe the phones of a language in terms of articulatory parameters. It, with a description of the sources and a knowledge of the properties of the vocal tract, enables one to calculate the acoustic outputs corresponding to these phones. Thus one is able to represent an utterance as a sequence of vocal tract descriptions.

This sequence of descriptions is useful in a number of ways. For example, it facilitates the study of human speech production by showing the movements of the articulators and provides clues to the processes which generate the commands to the articulators. The results of these studies can be used to diagnose and treat speech pathologies. They can be used to teach the deaf to speak. The shapes themselves, plus source information, can be transmitted as a substitute for the acoustic waveform in speech communication (as in articulatory vocoders). These results also provide for the use of vocal tract models as the acoustic signal generators in speech generation systems.

## II. The model

The vocal tract is a complex acoustic signal-generating apparatus. During speech production, it has two basic signal sources--semi-periodic vibration of the vocal cords and noise generated by turbulent airflow at a constriction. The vocal cords have adjustable spacing and tension which allow control of the frequency, mode and amplitude of the voicing. The trachea is usually assumed to be decoupled so that the vocal cords provide the rear boundary for the acoustically active portion of the tract [4, 5]. For fricatives, the vocal cords are abducted to allow airflow, thereby weakening this assumption. For sonorants, however, the vocal cords vary from slightly open to closed making the approximation justifiable. The nearly adducted vocal cords are generally approximated for analytical purposes as a total closure and a volume velocity source.

Turbulence sources occur any place in the tract where the flow velocity is sufficiently high--generally at a constriction--and result in broadband noise produced at the constriction. Usually the front portion of the tract is acoustically isolated at the constriction from the back portion, thus allowing the approximation of a shortened linear propagating tract with a broadband noise pressure source next to a total closure at the point of constriction.

The outputs of these sources are modified by the filtering effects of the tract configuration and shape. The configuration, a single tube with or without a sidebranch, is controlled by connecting or disconnecting the nasal passages with the velum and by providing closures at selected places. Given suitable assumptions, the effects of the tract may be analyzed. First, with the exception of regions of turbulence as discussed earlier, the sound levels in the tract are low enough that linearity of propagation can be assumed. Second, if one ignores frequencies above a certain limit dictated by the cross dimensions, transverse modes need not be considered. This is the plane wave assumption, which is valid below about 5 kHz and which makes the cross-sectional area a sufficient parameter to describe the acoustic effects (except for the losses) of a particular section of the vocal tract.

The vocal tract has several loss mechanisms [4, 5, 26]. As the vocal cords are not decoupled from the vocal tract, losses occur due to the backward wave flow out from the cords and in the soft tissues of the cords. As the air pressure at each point changes due to the passing sound waves, adiabatic heating of the air takes place. But the heating is not truly adiabatic as heat is lost to the vocal tract walls. The air flow corresponding to the sound waves also results in viscous friction against the tract walls.

The tract walls themselves are not rigid structures and move in response to the instantaneous pressure placed on them. (The wall losses are dependent on the local tract cross-sectional perimeter.) Additional sound energy is lost to the vocal tract from the nose and lips. The last two of these loss mechanisms result in direct radiation from the orifices and radiation from the skin overlying the yielding walls. All of these loss mechanisms introduce reactive components as well as resistive components into the propagation constant so that the formants (resonances of the vocal tract transfer function) may be perturbed.

Thus the sound generation system consists of two types of sources and an acoustic tube with a variable configuration, cross-sectional area function, a perimeter function, and five loss mechanisms. Given the source, the configuration, the area function, the perimeter function, and the characteristics of the losses, it is possible to calculate the acoustic output by treating the system as a current or voltage source, a lossy nonuniform transmission line with a sidebranch, and the output as the sum of the radiation losses. Frequently, however, it is desirable to further simplify the model for ease of analysis.

This simplification can be accomplished by idealizing any combination of the tract parameters. One common simplified case is the non-nasal sonorant (non-nasalized

vowels and some consonants), which is characterized by a vibrating vocal cord source at the back end, no sidebranches, no turbulence noise, and radiation only from the lips. A second common case is the lossless non-nasal sonorant with idealized boundary conditions (zero glottal area and infinite area tube just outside of the lips). This latter case is frequently used as an analytical model or where one desires to use Webster's horn equation, which is limited to lossless tracts with lossy terminations.

## III. History

### X-ray

The traditional method for measuring the area function of the vocal tract is the lateral X-ray photograph, from which the midsagittal plane shape is derived. The method, which is very attractive as it is a direct measurement on the articulators, has many difficulties. The spatial resolution is poor. The dynamic range is too great for unaided visual observation and thus picture processing is required. Spatial distortions occur in the X-ray camera. Image intensification is required if frame rates high enough for motion are desired.

And when one finally has useable X-ray photographs, quantitative measurements are also difficult. The landmark structures--typically the cervical vertebrae and maxilla (hard palate)--are not stationary. The images of the soft tissue articulators may be difficult to discern. To estimate the cross-sectional area, additional information about the shape--usually obtained by molds, palatograms, and guesswork--is required. (Perhaps the accuracy of cross-sectional area and shape estimation could be improved by the use of the recently developed computerized three dimensional X-ray systems at a cost of exposure and frame rate.)

Finally, two more difficulties exist--subject X-ray exposure limits--i.e. data quantity limits--and the noise of the camera which frequently degrades any simultaneous audio recording of the utterance. In spite of these difficulties, the method has been used successfully by a number of investigators [2, 4, 11, 21]. One approach to avoiding some of these difficulties is the use of a computer controlled X-ray beam which may be used to track a lead pellet that has been placed on the surface of an articulator [6].

The method, with all of its uncertainties, is still the best method in use today. It is the only method which directly gives the shapes and positions of the articulators. The alternative methods of vocal tract shape estimation that have been proposed since the development of the basic X-ray method all rely on the X-ray methods to check their performance.

## Acoustic methods

Due to the difficulties and dangers of the X-ray methods a number of methods have been devised which use only acoustic parameters which can be measured externally to the vocal tract. To date, all of these methods are limited to non-nasal sonorants.

## Perturbation methods

The first analytical approaches to relating the acoustic properties of an acoustic tube to its area function used first order perturbation of the area function of a uniform lossless tube with idealized boundary conditions (i.e. lossless non-nasal sonorant with idealized boundary conditions). The modes (poles and zeros) of the acoustic admittance of the open (lip) end of a uniform lossless acoustic tube with the other (glottal) end closed are [17]:

$$f_{On}=(c/4L)n \qquad (1)$$

where $f_{On}$=frequency of nth mode
c=velocity of sound
L=tube length

The $f_{On}$ for n odd are the poles of the lip admittance function, which correspond to the resonances of the tube with the lip end open which appear as the formants in speech. The $f_{On}$ for n even are the zeros of the lip admittance function, which correspond to the resonances of the tube with the lip end closed and which do not appear in the speech waveform.

The wave functions ($\cos n\pi x/2L$) of these modes are well known. By any of several methods, the perturbation of $f_n$ by the area function of the tube can be shown to be [9, 16, 24]:

$$\partial A(x)/A_0 = -2\sum_{n=1}^{\infty} (\partial f / f_{0n}) \cos n\pi x/L \qquad (2)$$

where $A_0$=area of uniform tube
$A(x)=A_0+\partial A(x)$=tube area function
$f_n=f_{0n}+\partial f_n$=nth mode frequency
x=position in tube (=0 glottis, =L lips)

for small perturbations. For somewhat larger perturbations--$1/7<A(x)/A_0<7$ [25]--$\ln(\partial A(x)/A_0)$ may be substituted for $\partial A(x)/A_0$ [16]. Thus, the area function can be computed from only the length and modes of the acoustic tube. Basically, each non-DC term of a normalized discrete Fourier expansion of the length normalized acoustic tube (ln) area function relates only to the perturbation of the corresponding mode.

Neglecting inaccuracies for large perturbations, the theoretical difficulties with the first order perturbation results are threefold. If one assumes $\partial f = 0$ for $n>N$, which just limits spatial resolution and corresponds to the plane wave assumption for N~9, one can achieve any set of critical frequencies for any value of L although this might require large perturbations of the area function A(x). The second difficulty is (area) normalization of the area function. The only place where $A_0$ enters, other than in the form $\partial A(x)/A_0$, is in the cross dimensions which determine the frequency limit of the plane wave assumption. The third difficulty is the measurement of the even modes, as they do

not appear in the waveform of an idealized non-nasal sonorant. A further practical difficulty is that the first five formants (limited by the plane wave assumption) are frequently difficult to measure [5].

Schroeder attempted to reconstruct the area functions of non-nasal sonorants and stay within the above difficulties [24]. To do this, he measured the first three formants of a sound, assumed $\delta f_2$, $\delta f_4$, and $\delta f_n = 0$ for $n \geq 6$, guessed a tract length and area normalization and generated an area function by the above perturbation method. This assumption corresponds to reconstructing only the odd components of the tract log area function. When this method proved inadequate to describe certain phonemes (such as /u/) which contain sizeable even components, he devised a new method for measuring the acoustic parameters.

The new method required attaching an acoustic tube apparatus to the lips of the subject. The apparatus contained both a source and transducers such that the acoustic admittance of the vocal tract could be measured. The subject was required to articulate with adducted vocal cords so that no sound was produced within the tract. The modes $f_n$ for $n \leq 6$ could then be calculated and used to compute an area function of assumed length and area normalization.

This new method removed the theoretical difficulty of the loss of half of the required modes, but introduced new practical difficulties. The apparatus may be subject to leakage at the lips and may interfere with some lip and mandible gestures. The articulations themselves may be unnatural as a side effect of the adducted vocal cords. The problem of length measurement, area normalization, and the effects of losses were still unsolved.

Mermelstein investigated the properties of the perturbation formula in the $\ln(\partial A(x)/A_0)$ form [16]. He empirically determined that the transform between the area function and the critical frequencies of the admittance function for $n \leq 6$ was nonsingular and therefore unique. He also compared Fant's X-ray area functions of six Russian vowels [4] with the corresponding acoustic tube of the same $f_n$ for $n \leq 6$ as determined by Webster's horn equation and found the errors to be within reasonable limits.

## Other methods using endpoint acoustic parameters

Paige and Zue present an algorithm for generating area functions from the modes and length of the tract [19]. They treat the vocal tract as a cascade of equal length uniform lossless sections of differing cross-sectional area and derive a procedure which produces an n+1 section tube (where n is the number of modes). To lessen the roughness in the

area function, a method, based only on the length of the tract, is supplied for generating more modes which results in more sections in the area function. (This does not, however, increase the amount of information in the area function--it only allows a more attractive output format.) The algorithm's chief attractions are that it is computationally efficient and requires no assumptions of small variations from the uniform tract. It does not avoid the primary difficulties of the perturbation methods. The same sets of modes--the poles and zeros of the lip input admittance--are required as input. (The paper does not treat measurement of these modes.)

Gopinath and Sondhi also present algorithms for converting endpoint parameters into continuous area functions [7]). In addition, they offer a conversion of modes of a tube with no discontinuities to a tube with a finite number of prespecified discontinuities in an attempt to include the larynx-pharynx junction into the model. This paper, like the preceding one, presents an algorithm which requires data which are difficult to obtain reliably. (Both compute their data from Fant's X-ray derived area functions [4] and attempt to regenerate those area functions.)

## LPC methods

More recently, two methods for area function extraction have been devised which depend only upon the radiated acoustic waveform as measured by a single external microphone. Wakita has examined a method suggested by Atal which consists of an acoustic tube implementation of the filter generated by a linear predictive estimation (LPC--see Appendix) analysis of prefiltered (fixed approximation of 6db/octave by a digital difference to remove glottal waveform and radiation factors) speech [28]. He analyzed American vowels into eight section tubes (7 kHz sample rate), compared his results with Fant's published data on Russian vowels [4] and found a degree of agreement which was better at the mouth than at the larynx. These results, however, deteriorate rapidly as the number of tube sections (and the sampling rate) is increased [20]. (The length is not an input--each section is cT/2 long where T is the sampling interval and the n+1st section does not affect the previous n sections. The number of sections corresponding to the vocal tract length is about equal to the sampling rate in kilohertz.) Wakita also attempted to show that his method could be used to generate the modes necessary for a perturbation theory area function by analyzing his own area functions for the modes, which yields nothing more than a smoothed approximation to his original area functions.

Wakita does not mention that his implied model is inaccurate. The LPC area function realization has certain assumptions--total reflection at the mouth and all of the losses absorbed into the backward wave flowing out of the glottis. His fixed prefilter does not realistically model the source and radiation characteristics--which are not constant--and the error is absorbed into the generated area function. (As LPC is an autocorrelation domain analysis, only the frequency magnitude characteristics of the prefilter are important.) His results are surprisingly good considering the modelling errors involved.

Nakajima et al. have attempted to remove one of the weaknesses of the Wakita scheme [18]. Based on the assumption that the source spectra and radiation characteristics have two degrees of freedom, they have designed a two section adaptive prefilter composed only of real zeros (in the z plane) which attempts to level the spectrum by removing its gross characteristics. Presumably, what is left is the filtering effect of the vocal tract on a white or comb filtered white (i.e. voiced) source. They then assume that the glottal losses dominate all of the loss mechanisms, which allows the use of the LPC method to generate an area function.

Their results do appear to be superior to those of Wakita in spite of a 15 kHz sample rate which violates the plane wave assumption. The method appears capable of meaningful results over the entire tract although the front of the tract is resolved better than the back. The glottal constriction can be seen in some of the area functions. As the LPC methods do not require length as an input and the adaptive prefilter is general enough to compensate for frication, the method is theoretically limited only to non-nasals. Examination of some of their published results shows degradation in at least some instances of frication. Some of their stops also do not show the correct point of articulation. (The stop itself cannot be analyzed, but its formation and release can be.)

## Analysis-by-synthesis

At least two analysis-by-synthesis attempts at area function extraction have been tried. Both estimate the controlling parameters of an articulatory synthesis model, calculate its output, and compare the output to the speech acoustic signal. The process continues refining the estimates until the comparison indicates sufficient similarity. The process may not converge or may converge to a nonunique solution.

Hafer used Coker's articulatory model, which does not include losses (except for yielding wall perturbations of the first formant) or nasals [3, 8]. He then used the method of steepest descent on the mean square error of the formant frequencies while varying only the tongue body position parameters. (The model also includes the other articulators.) He indicates convergence to a single minimum for several test cases. He does not, however, explore the method in the full generality of the model, which could indicate more about the performance of the method as well as make some suggestions regarding alternate articulations.

Rice, using his own line analog synthesizer which includes some loss mechanisms, attempts to match his model to the articulatory data by minimizing the mean square difference of the first derivatives of the LPC spectrum of the speech and the spectrum generated by his model [23]. In an examination of several vowels, he notices relatively consistent errors for the first formant, which he suggests may be due to a lack of matching between the model and the speaker. (He does not indicate awareness of the yielding wall effects on the first formant, which could also explain the systematic error.)

## Linear regression

Atal attempted to reconstruct area functions using the techniques of linear regression [1]. He started with a vocal tract area function model with seven control parameters. He generated the acoustic output from the tract using a lossy transmission line implementation with a voiced excitation for a large set of parameter values. Several different techniques (one at a time) were used to analyze the acoustic output into acoustic parameters. Linear regressive techniques were then used to generate an optimal weighted sum (of the acoustic parameters) estimate of the vocal tract area function model control parameters.

The method yielded fairly accurate control parameter estimates for the output from the vocal tract model. It has not, however, been tested on real speech. This would require an accurate vocal tract model and would require a test for accuracy such as a comparison of X-ray data with the estimated model shape for a prohibitively large data (and training) set. The method appears to have potential, but its performance on real speech is unknown.

## Summary

Each of these methods has certain advantages and disadvantages (Table 1). One of the chief advantages of the perturbation methods is the simplicity of the resulting

relation between between the modes and the area function. In this form, it yields insight above and beyond the Helmholtz resonator analyses of the years prior to the perturbation theory result [2, 4]. But its disadvantages--difficulty of obtaining even the modes, length degeneracy, and the lossless assumption--limit its practical usefulness.

Methods such as those of Paige and Zue can allow wide variation in area and provide conveniences such as computational speed, but do not solve the measurement problems inherent in the perturbation methods. These provide an alternate implementation but provide no insight.

The LPC methods offer a fast, easily implemented method which does not require length as an input. They require, however, an accurate separation of the formants from all other factors affecting the output spectrum. The bandwidths must also be measured accurately. The assumption that the losses are dominated by the glottal losses is incorrect with the possible exception of the first formant [5].

The analysis-by-synthesis methods perhaps offer the best chance for successful incorporation of the various loss mechanisms. But if this is so, the cost in terms of slow execution and little direct insight will be high. Performance of the method requires more accurate knowledge

of the losses than presently available and a vocal tract model which may have to be matched to the individual subject. Such a matching, however, might provide sufficient constraints to limit an articulatory analysis to one or a few vocal tract shapes.

The linear regression technique is computationally efficient and offers potential insight into the relation between a set of acoustic parameters and the shape of a lossy vocal tract. Its major problems are the necessity of a good vocal tract model, the requirement of a large and difficult to obtain training set, and its current lack of testing on real speech.

Each method has advantages and each has disadvantages. To date, no method, including midsagittal plane X-ray photography, is free from serious difficulties. The following describes an attempt to devise an acoustic method of area function extraction which provides a partial solution to some of the difficulties encountered in the earlier efforts.

Table 1

Acoustic area function measurement techniques

| Method | Disadvantages | Advantages |
|---|---|---|
| First order perturbation | Measurement of lip admittance zeros | Insightful |
| | Errors for large deviations from uniform tube | Computationally simple |
| | Idealized boundary conditions | |
| | Lossless | |
| | Needs length | |
| Paige and Zue | Measurement of lip admittance zeros | Computationally fast |
| | Idealized boundary conditions | Accurate for large deviations from uniform |
| | Lossless | |
| | Needs length | |
| Gopinath and Sondhi | Measurement of lip admittance zeros | Accurate for large deviations from uniform |
| | Lossless | |
| | Needs length | |
| | Assumes knowledge of pharynx-larynx boundary | |

Table 1 (cont)

| Method | Disadvantages | Advantages |
|---|---|---|
| LPC (Wakita) | Wrong boundary conditions | Easily measured parameters |
| | Wrong loss model | Length output |
| | Glottal spectrum sensitivity | Easily computed |
| | Radiation effect sensitivity | |
| | Low sample rates only | |
| Adaptive prefiltered LPC (Nakajima et al.) | Wrong boundary conditions | Easily measured parameters |
| | Wrong loss model | Easily computed |
| | | Length output |
| Analysis-by-synthesis | Requires vocal tract model | Physical constraints |
| | May need subject matching | Gives area normalization |
| | May not converge | |
| | Slow | |
| Linear regression | Requires large training set | Physical constraints |
| | Requires vocal tract model | Gives area normalization |
| | May need subject matching | Potentially insightful |
| | Performance on real speech unknown | |

## IV. The algorithm

The algorithm is designed with several constraints in mind. The first is that there should be no risk to the subject which rules out X-rays or any form of intrusive instrumentation which might require medical supervision. It is also desired that no special apparatus be required to measure the input parameters. These constraints limit the algorithm to the acoustic output produced by normal speech and minimize the probability that the subject's speech will be perturbed by the measuring apparatus itself.

Use of only the acoustic output signal as input places its own set of limitations on the algorithm. First, it suggests that the set of allowable phonemes be limited to the set of non-nasal sonorants, as sufficient theoretical support for nasals does not yet exist. Disallowance of fricatives eliminates phonemes which radically alter the effective tract length. The zeros of the lip acoustic input admittance are also absent from the output signal.

The available information is now the convolution of the glottal (voiced) excitation, the vocal tract impulse response, and the radiation factors, from which only the vocal tract impulse response is desired. In theory, this impulse response is composed only of poles [5]. Extraction of four or five (limited by the plane wave assumption)

formants and their bandwidths (i.e. eight or ten poles) from the speech waveform is well known to be nontrivial [5, 15]. To further complicate matters, a set of formants and bandwidths is not always an adequate description of the tract effects as changes in the tract shape (such as /flap d/) can occur at rates approaching the bandwidths of some of the formants. The formants are, however, a compact representation of the tract effects which is adequate the vast majority of the time.

At present, the nature of the vocal tract losses is reasonably well known. Sufficient data, such as the vocal tract perimeter function, for quantitative application of this knowledge are lacking. The effects of interaction between the loss mechanisms are also not known. Theoretical support for the reverse relations--transfer function to the physical tract shape--is just beginning to appear. Thus, knowledge of the bandwidths of the formants is not currently useful in area function extraction from the acoustic output. The nature of the relation between the formants and the area function in the lossless case is, however, well known.

This, then, outlines the basic strategy of the following algorithm--use the lossless case with corrections to account for the effects of the loss mechanisms. In addition, the algorithm is implemented in such a way that an additional set of empirical constraints is provided to

recover some of the information lost when measurement of the lip acoustic input admittance zeros is abandoned.

## General description

The algorithm starts with the frequencies of the first N formants of a given non-nasal sonorant, where N is chosen by the plane wave assumption. Substitution of the appropriate constants (L~17 cm.) into Equation 1 will show the formants to have an average density of one per kilohertz. Thus N was chosen to be five. This causes occasional difficulty in measurement of the fifth formant but pays off in better length estimation and better area function estimation by the algorithm.

With one exception, the loss mechanisms do not perturb the formant frequencies to any great extent [5]. Sondhi has published a theoretical model for wave propagation in the lossy vocal tract [26]. In it, he postulates and empirically verifies a transform between the resonant frequencies of a lossy vocal tract and a lossless vocal tract. Due to the yielding walls:

$$F^2_{1,\text{rigid}} = F^2_{1,\text{yielding}} - 200^2 \text{Hz}^2 \qquad (3)$$

where $F_1$ = first formant frequency

This transform now allows the formants of a lossless tract to be calculated from the formants of a lossy vocal tract.

The Levinson recursion, which is frequently used in LPC to compute the predictor polynomial and the reflection coefficients from the autocorrelation coefficients, can be used to convert between any of these three representations (see Appendix A). (The reflection coefficients convert easily into a cascade of equal length uniform acoustic-tube sections of differing cross-sectional area.) This recursion, however, becomes singular if the modes are lossless. Thus some form of loss model must be supplied.

This loss model serves several functions. First it provides control of the glottal boundary condition. The LPC area function inherently contains the lip boundary condition corresponding to an opening into free space--the standard idealized lip boundary condition for the lossless vocal tract models. As all loss in the LPC model is in the form of a backward wave flowing out of the glottis, the standard lossless glottal boundary condition--complete closure--requires that all formants have zero bandwidths. This can be approximated by assigning narrow bandwidths to the formants. (Note that the reverse is not rigorously true--no loss implies total reflection at the glottal boundary, which implies either total closure or opening into free space--i.e. an infinite area section of acoustic tube--at the glottis. The form of the boundary depends on the formant frequencies. The open glottal boundary

condition has never appeared in the testing of the algorithm.) The relative values of these bandwidths provide a mechanism for applying shape constraints to the generated area functions. Varying these bandwidths by a constant factor--so long as all bandwidths are kept small enough to maintain an "almost lossless" case--varies only the area of the tube section beyond the glottis, which is quite small, in keeping with the boundary condition approximation.

These formants and their bandwidths can be mapped into poles in the z-plane and expressed as the roots of second order polynomials which can be multiplied to form a predictor polynomial. The Levinson recursion can now be used to generate reflection coefficients which specify an area function, the length of which was set by the s-plane to z-plane mapping. As with all other acoustic methods, only the relative areas can be determined.

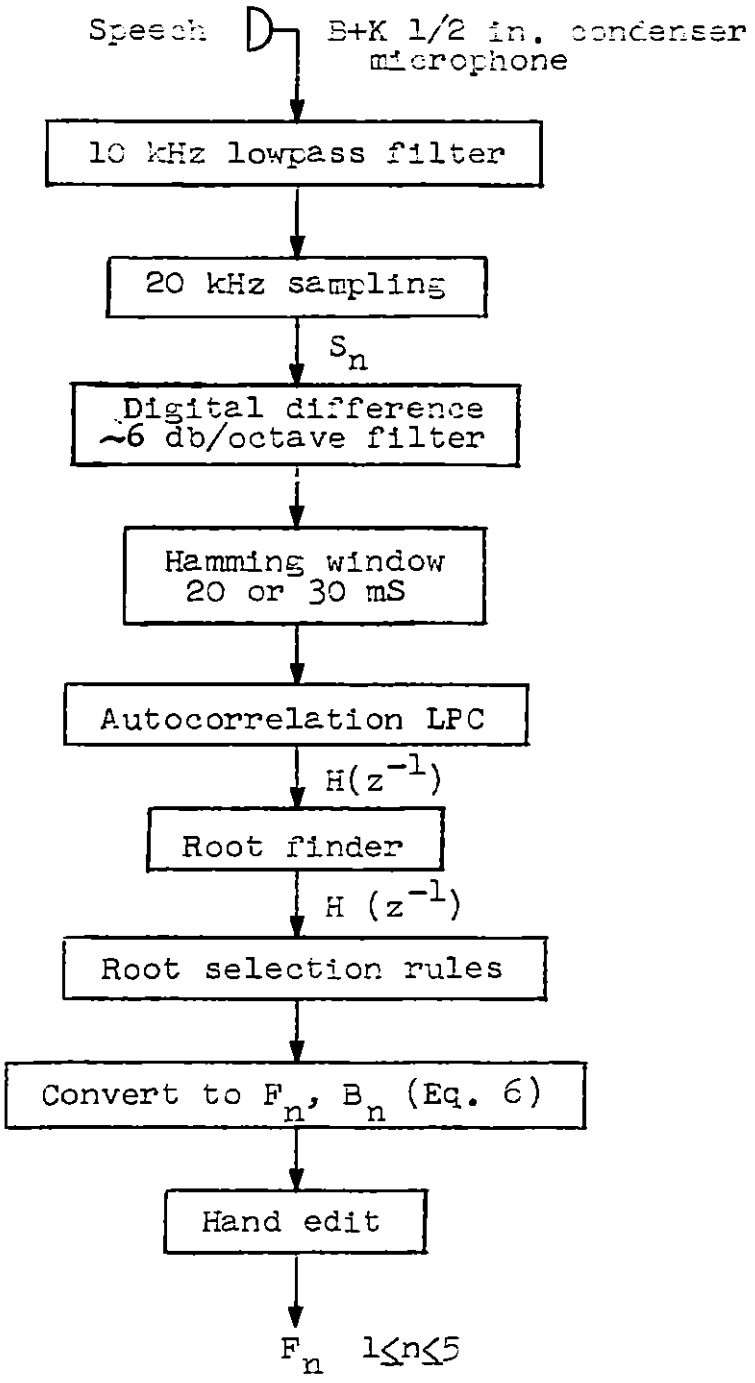
The problem of length estimation has two possible solutions. Wakita has suggested a method based on minimization of the mean square deviation from uniform of the log area function as a function of length [29]. A second method is suggested by the performance of the algorithm itself. When the generated area function is too long, the lip end tends to open up to excessive areas as if the algorithm is attempting to move the lip boundary condition to a point closer to the glottis.

The area normalization requires a vocal tract model. X-ray data [4] indicates that the area of the glottis just forward of the vocal cords is relatively invariant. This method would exhibit a strong dependence on the length and the accuracy of the measurement of the first formant as a result of Sondhi's transform. In addition, the local errors caused by the truncation of the set of formants would cause difficulty. A more global vocal tract model is required.

## The implementation

The fundamental algorithm deals only with stationary time frames and has no memory between these frames. For continuous speech, the formants can be measured at short (interframe) intervals, the corresponding area functions computed, and these area functions displayed sequentially to create a movie of the vocal tract.

To analyze speech, the acoustic signal is sensed with a B&K 1/2 inch condenser microphone about 30 cm. from the speaker's lips in a quiet room environment, lowpassed at 10 kHz, and sampled (12 bits) at 20 kHz. The formant tracking is accomplished by a Markel formant tracker [15] (Figure 1). First the signal is filtered with a digital difference to approximatly remove the voiced source and radiation spectral effects. Second, this signal is LPC analyzed [15] at intervals of 10 mS with the number of poles selected by

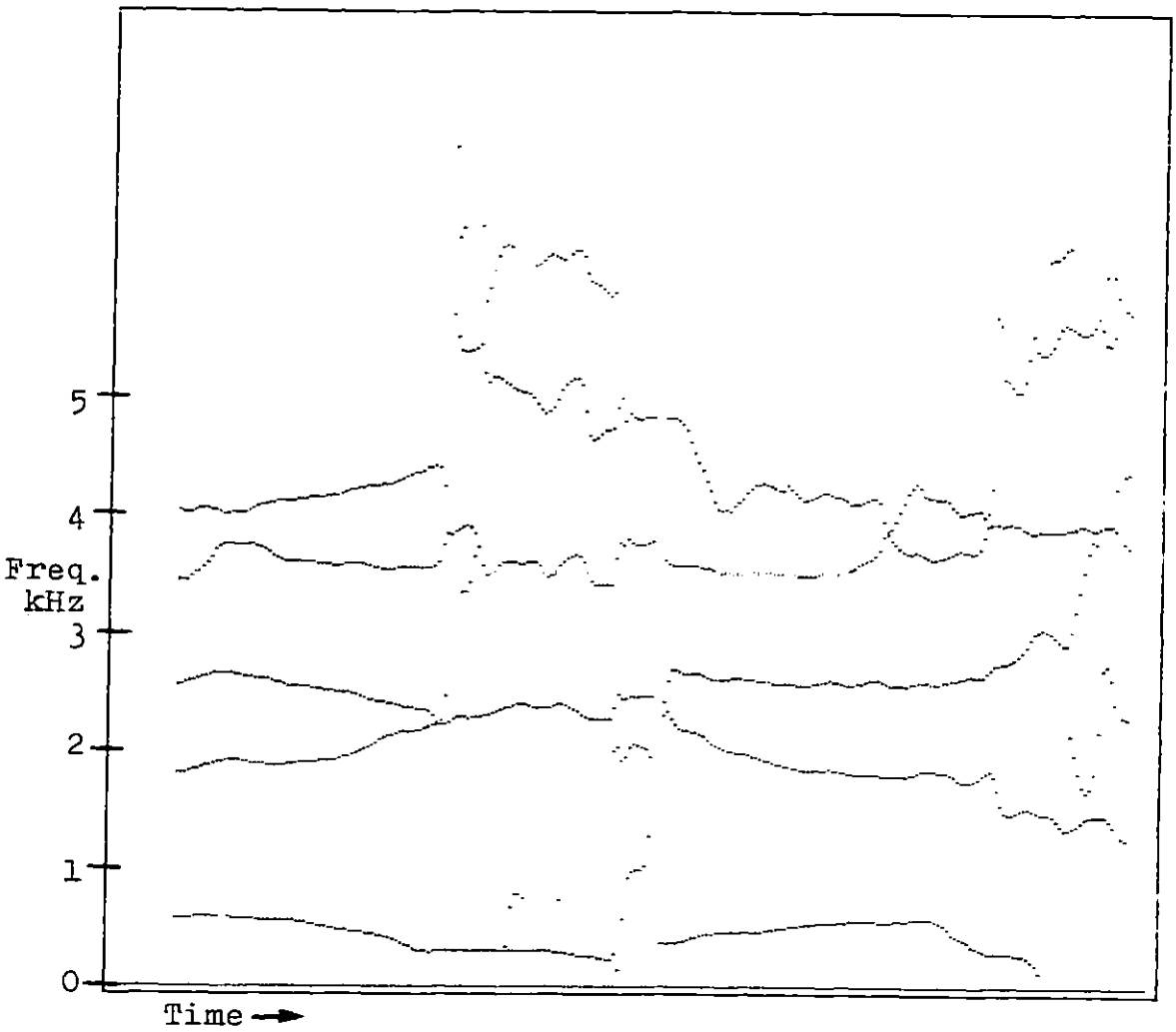



The formant tracker

Figure 1.

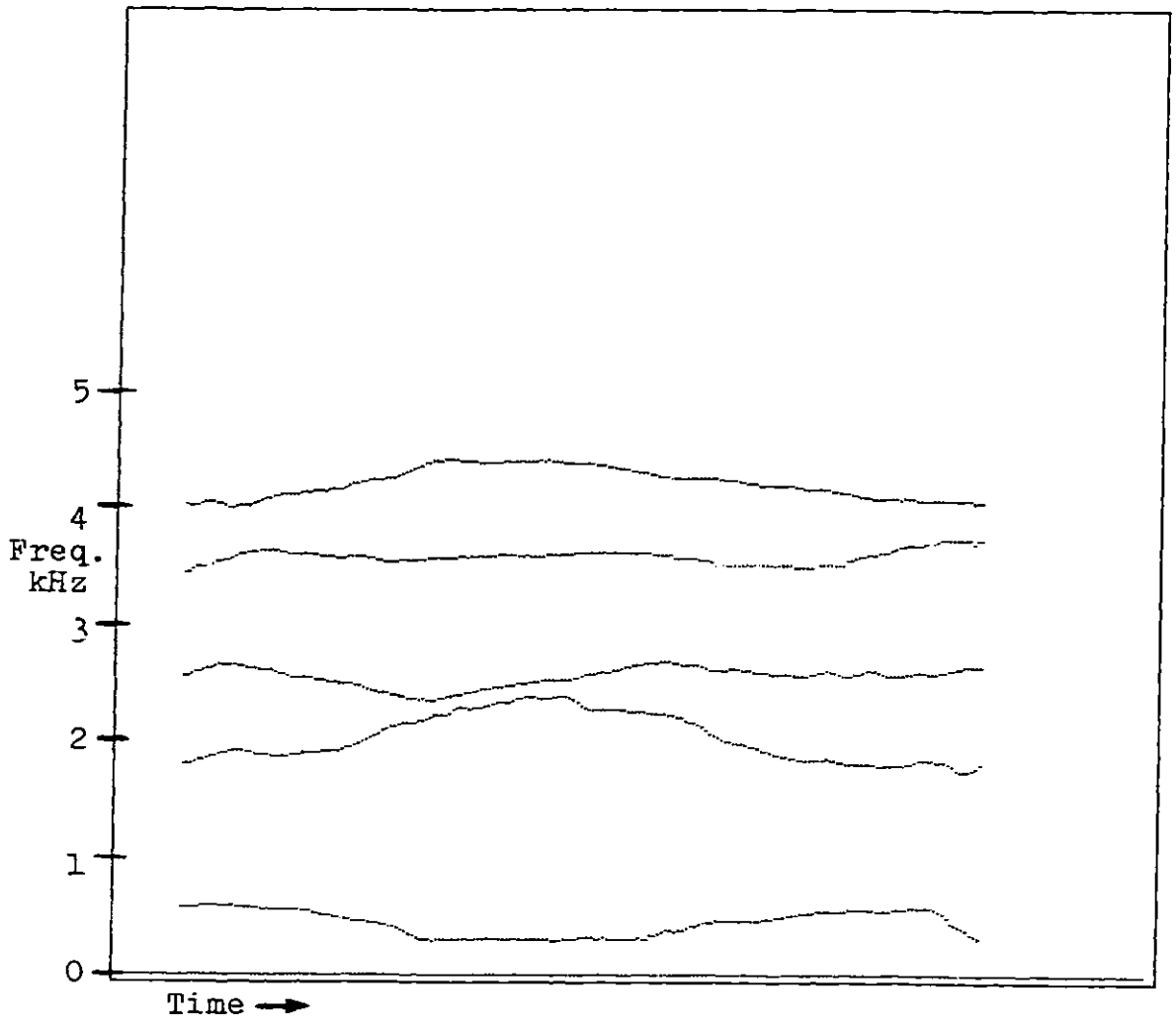
trial and error to give the best results for the entire utterance (generally 16 to 20 poles). Third, a root finder is applied to the predictor polynomial to find its roots. Fourth, a set of selection rules--the first five conjugate pole pairs of sufficient radius in the z-plane--is applied to select the pole pairs corresponding to the first five formants. Fifth, the frequencies and bandwidths of the formants are stored in a file. Sixth, as formant tracking itself is not a completely solved problem [5, 15], the frequencies of these formants are hand edited to correct errors in the tracking procedure (Figure 2). (This hand editing also allows the operator to draw formants through periods of closure by invoking the continuity of the tracks and their first derivatives which is a result of similar continuity in the movements of the articulators [3].) The bandwidths of the formants, which were saved only as an aid to the operator for the editing operation, are discarded and a file containing the formant frequencies at each time frame is saved.

The input to the area function estimation algorithm is just the N formants (N=5 here), whether they have been measured by the above scheme or have come from a published source. (Length can be treated as either an input to a single instance of an area function estimation or the output of an iterative sequence of area function estimations.)




Raw formant tracks
(/ɛgɛ/)

Figure 2a.

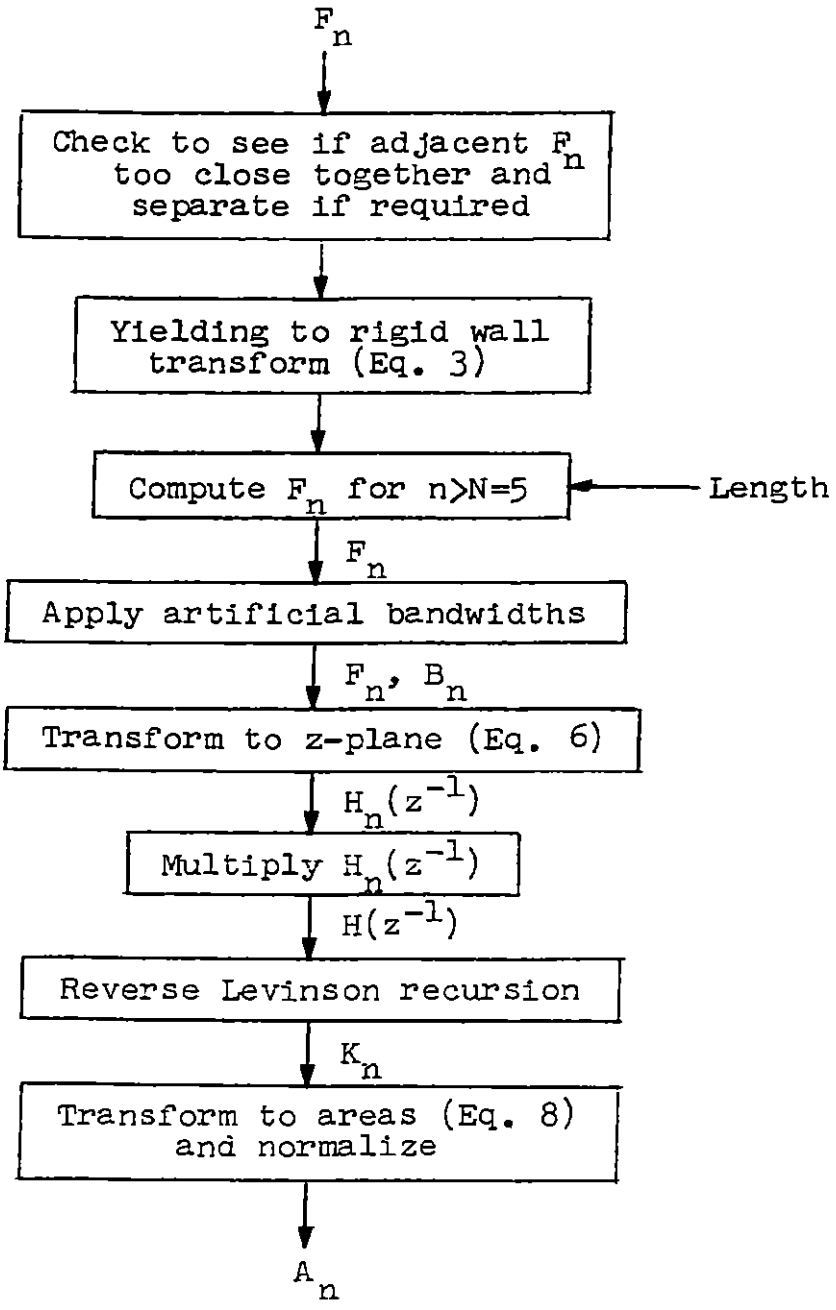



Edited formant tracks
(/ɛgɛ/)

Figure 2b.

The algorithm (Figure 3) first applies a check on the formant frequencies, as adjacent formants have a minimum frequency separation. (Some formant trackers can violate this condition.) If any two of these adjacent formants lie too close together, both are perturbed away from each other until they are separated by 10 percent of their average frequency to approximate the mode splitting which would occur in the real tract. Next, Sondhi's transform (Equation 3) is applied to remove the effects of the yielding walls. As this can yield an imaginary first formant if the measured first formant is too low, the transformed formant is assigned to have a value of 25 Hz if it would otherwise be imaginary or less than 25 Hz. The set of N formants is now the first five resonances of the hardwalled almost-lossless acoustic tube whose area function is to be estimated.

The LPC area function generation technique results in a p+1 section acoustic tube where p is the number of poles which is so far equal to 2N. An eleven section area function with one section (the glottal end section) discarded to provide the glottal boundary condition is not an attractive output format. To improve this format, the acoustic tube is generated in half centimeter sections.




The area function estimation algorithm

Figure 3.

$$L=(cT/2)p \qquad (4)$$

where $cT/2=1/2$ cm
i.e. $T=1/34000$ Sec.

To generate such an acoustic tube of length (L) of 17 cm. would require that p=34 and N=17. As it is not practical to measure 17 formants in speech, the formants above the fifth are supplied as the resonances of a uniform tube of the standard boundary conditions and length L, i.e. the odd modes of Equation 1.

$$F=(c/4L)(2n-1) \qquad (5a)$$

$$F=(1/2Tp)(2n-1) \qquad N<n\leq p \qquad (5b)$$

This shorter section length does not increase the amount of information extracted from the speech but does allow a much smoother sampling of the curves in the area function and permits a better set of constraints in the bandwidth model. Also, it suggests that the length be varied in half centimeter steps. (The length could also be varied by varying T and keeping p constant, but this would yield varying length sections.)

The bandwidths from the bandwidth function B(f) described in the next section are now applied to the formants and mapped into the z-plane

$$H_n(z^{-1})=(1-z^{-1}e^{2\pi TS_n})(1-z^{-1}e^{2\pi TS_n^*}) \qquad (6)$$

where $S_n=-(B_n+jF_n)$

and combined into the predictor polynomial.

$$H(z^{-1})=\prod_{n=1}^{p/2} H_n(z^{-1}) \qquad \text{p even} \quad (7a)$$

$$=(1-z^{-1}e^{2\pi TB_N})\prod_{n=1}^{(p-1)/2} H_n(z^{-1}) \qquad \text{p odd} \quad (7b)$$

The reverse Levinson recursion now transforms this polynomial into reflection coefficients which convert via Equation 8 into area ratios at the uniform tube section junctions.

$$\frac{A_{n+1}}{A_n}=\frac{1-K_n}{1+K_n} \qquad (8)$$

where $A_n$ = area of the nth uniform tube section
$K_n$ = nth reflection coefficient

So far, the length of the area function has been treated as an input to the algorithm. If the algorithm is tested on the acoustic parameters corresponding to a known X-ray area function, the length can be treated as a known value. When only the acoustic parameters are known, however, the length is an unknown. One possible method for estimating the length is a version of Wakita's method [29] tailored to this algorithm by using this area function

estimator rather than Wakita's estimator.

$$p \text{ s.t. } \min(1/p)\sum_{n=1}^{p}(\ln A_n - \overline{\ln A_n})^2 \qquad (9)$$

While this method was implemented, its performance was adequate on stationary vowels, but was inadequate for use on continuous utterances. (See Chapter VI.) The other method, which is based on observation of lip area throughout the utterance, is the one used where the modified Wakita scheme is unsuitable. (The characteristics of this method will also be discussed in Chapter VI.)

The vocal tract model used to normalize the areas is an extremely simple one--constant vocal tract volume [27]. Examination of Fants's X-ray data [4] shows that the six Russian vowels have an average volume of 89 cc. and a standard deviation of 11 cc. for his subject. This yields a fairly reasonable estimate of the correct normalization and avoids the problems of models which contain representations of the articulators or limits on the areas as a function of position.

The polynomial manipulations of the above algorithm (including the Levinson recursion) require highly accurate computation. In this implementation, 36 bit mantissas proved adequate for all manipulations except for Equation 7. (This representation is adequate here too if the terms are

multiplied in the order: first, last, second, next to last, etc.) Non-polynomial portions of the algorithm can be computed at much lower accuracies without compromising the area functions.

## The bandwidth function

As stated earlier, non-vanishing bandwidths are required to prevent the Levinson recursion from becoming singular, to set up the glottal boundary condition, and to provide area function constraints. Fant's X-ray area functions of six Russian vowels [4] were converted into a formant and bandwidth representation by a reverse of the portion of the algorithm following the bandwidth application (Figure 4) and searched for regularities. Note that the bandwidths so derived are artificial and bear no relation to the acoustic bandwidths of the formants in the original speech waveform.

The X-ray area functions are given in areas at each half centimeter along the vocal tract, the desired format. An additional section of small area (2 sq. mm.) is added to the glottal end of each area function to provide the desired boundary condition. This area function is converted to reflection coefficients by Equation 8 and then converted by the Levinson recursion (Equation A7 in the Appendix) to a polynomial describing its z-plane transfer function. The

roots of this polynomial (resonances of the acoustic tube) are computed using a root finder and transformed into frequencies and bandwidths by reversing Equation 6. The results for all of the six vowels are plotted in Figure 5.

Simple inspection of Figure 5 suggests strong regularities in the bandwidths of these resonances. There is no theory to suggest a means of modelling these data. Thus the choice of basis functions is purely arbitrary. Inspection of the data strongly suggests an all conjugate pole pair magnitude response representation of bandwidth as a function of frequency. Indeed, a trial and error fit by three s-plane resonances fits the data fairly well (Figure 5).

$$B(f)=B_0\prod_{n=1}^{3}\left(\frac{W_n^4}{(Bw_n f)^2+(W_n^2-f^2)^2}\right)^{1/2} \qquad (10)$$

where $B_0 \sim 30$ hz
$W_1=4.0$ kHz
$Bw_1=.7$ kHz
$W_2=13.0$ kHz
$Bw_2=.7$ kHz
$W_3=20.0$ kHz
$Bw_3=4.0$ kHz

Other models, such as constant bandwidth and $B_n=f(F_n,n)$ using straight line sections for each n, gave poor results. Better bandwidth functions may exist but would probably be

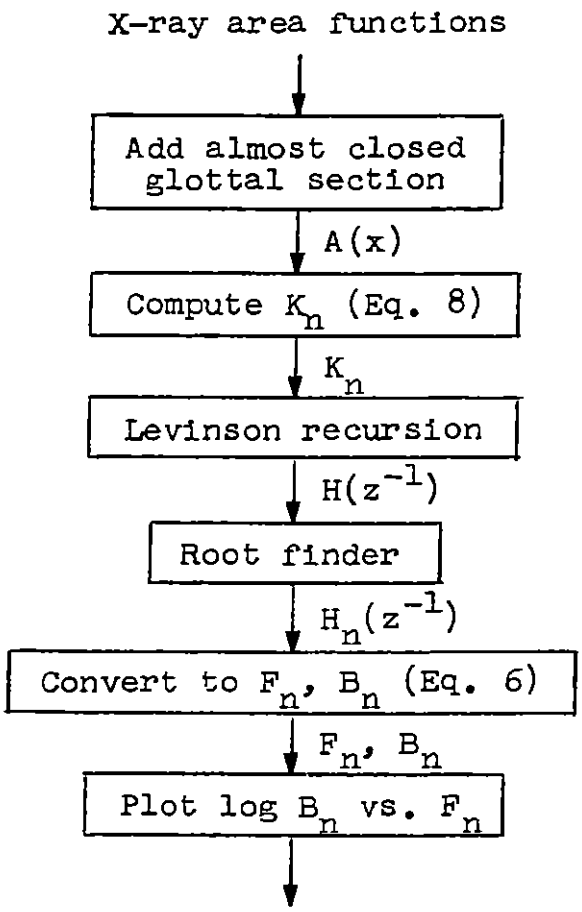
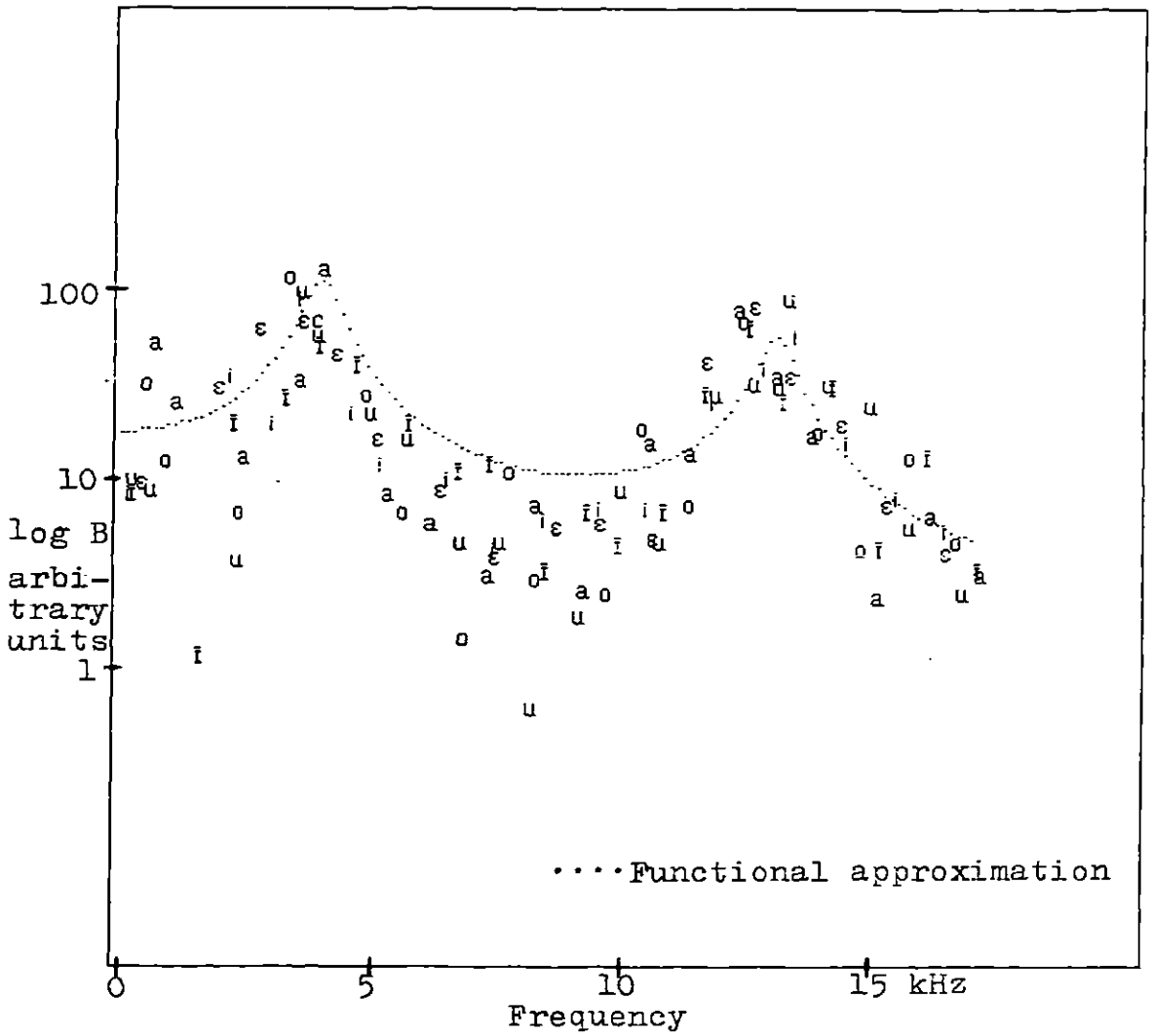



Figure 5

Artificial bandwidth examination system

Figure 4.




Artificial bandwidths of Fant's vowels as computed by Figure 4 and the bandwidth function approximation of Equation 10.

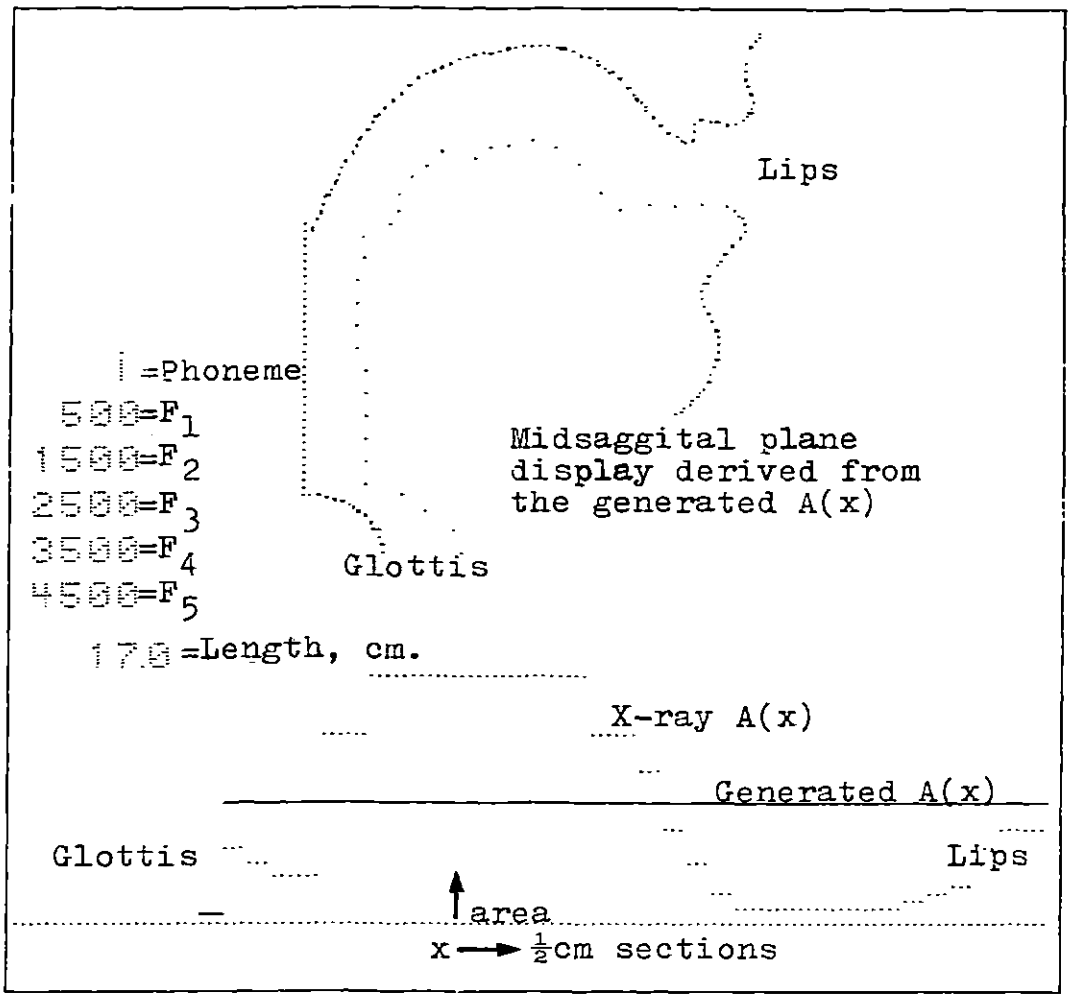
Figure 5.

much more complex.

## The midsagittal plane display

Direct area or log area function displays are not a very good way to examine the output of an area function estimator unless a corresponding X-ray area function exists for comparison. In articulatory domain vocal tract modelling, one thinks in terms of the physical structures present. When one watches a motion area function, bumps roll back and forth and closures occur at various places. It requires a good imagination (and knowledge of articulation) to fit vocal tract shapes to these area functions as the movie flashes by. In order to remove much of the imagination factor from the interpretation, a simple procedure was devised to produce a display similar to a midsagittal plane X-ray which would allow a viewer familiar with articulatory phonetics to examine the results.

The immovable structures of the display (Figure 6)—the upper lips, upper teeth, palate and velum—are a tracing from the same X-ray films used by Perkell [21]. The back of the pharynx is a straight line of variable length as determined by the length of the area function. (This is the only way in which length variations are incorporated into the display.) Starting at the lips, a dot is shown every half centimeter to show the position of the moveable structures (bottom and front) of the display. The first dot




The midsaggital plane display

Figure 6.

is connected to a tracing of the lower lip, which moves along with the dot. While these movable dots do not explicitly model any particular structure, it is generally quite easy to see the positions and shapes of the actual structures from their outlines.

The input to this display is the area function in half centimeter sections and the length. The transform between these areas and the tract height (separation between the top and bottom or back and front of the tract—i.e. that which is seen in a midsagittal plane X-ray) is not simple since the width of the tract varies with both position and height as well as the positions of the articulators. The conversion is accomplished with a table which is a modification of one due to Lagefoged [12]. The table is a list of cross-sectional areas corresponding to tract heights at seventeen points in the vocal tract. The line for 8 cm. in the original table is not strictly monotonic so that the conversion from area to height is multivalued. To correct this problem, three of the values were altered. The table (Table 2) is then extended to include greater heights than in the original table by adding entries to the original table.

The table is defined at one centimeter intervals along the tract. To use the table with the half centimeter section, variable length area functions, an interpolation is

used. The table is scaled so that its length corresponds to the length of the area function. Then, for each area function section, the nearest table row is chosen and used to convert to the corresponding tract height. These tract heights can then be used to determine the separation of the moveable dots from the fixed structures.

This display is good enough to allow even an untrained observer to understand its meaning. It is not, however, without shortcomings. One point of inaccuracy is at the lips. The real structures have three degrees of freedom--extension (rounding), width, and height--but the display has only one--height. Length is another parameter which is handled incorrectly by the display. The only way to change the length of the display is to drop or raise the larynx--i.e. the insertion or removal of pharyngeal sections, which also alters the interpolation of the rows of the height to area function table. In reality, the length of the vocal tract is a function of lip extension, tongue position and laryngeal height. The display is correctly proportioned for a length of about 17 cm. and therefore for phonemes with larger tract lengths, such as the back vowels, shows the tongue body too far back in the tract (Figure 8e). Finally, the display assumes that the cross-sectional area at a given point in the tract is a function only of the height. This is also particularly in error forward of the

velum as the width is a function of mandible (lower jaw) height and lip width, and the height is a function of mandible height, tongue position, lip extension and lip height.

This midsagittal plane display is intentionally simple so that the approximate vocal tract shape can be displayed without using iterative matching to a complex model. It is therefore a better display for the evaluation of an area function estimator as it makes no attempt to correct errors in the area function as would a model which contains specific representations of the articulators. If displayed simultaneously with the area function, it makes a valuable tool, not only for evaluation of the area function estimator, but also for observing the vocal tract in a realistic manner.

## Table 2

### Vocal tract cross-sectional area as a function of position and height

Height in mm. at:

Approximate position (in tract)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| 31 | 32 | | | | | | | | |

Areas in sq. mm. corresponding to the heights:

0 cm. (lips)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| - | - | - | - | - | - | | 294 | 326 | 356 |
| 384 | 405 | 425 | 443 | 459 | 475 | 500 | 535 | 580 | 635 |
| 690 | 750 | 825 | 910 | 1005 | - | | | | |

1 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| - | - | - | | | | | | | 386 |
| 422 | 458 | 493 | 527 | 560 | 592 | 623 | 653 | 682 | 711 |
| 731 | 751 | 771 | 791 | 811 | 831 | 851 | 871 | 891 | 911 |
| 931 | 951 | | | | | | | | |

2 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| - | - | | | - | | | | | 392 |
| 437 | 482 | 527 | 573 | 619 | 665 | 711 | 756 | 801 | 844 |
| 887 | 930 | 970 | 1010 | 1050 | 1090 | 1130 | 1170 | 1210 | 1250 |
| 1290 | 1330 | | | | | | | | |

3 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| - | | - | | - | | - | 309 | 356 | 404 |
| 452 | 500 | 549 | 599 | 650 | 702 | 754 | 808 | 862 | 916 |
| 970 | 1024 | 1088 | 1142 | 1196 | 1250 | 1304 | 1358 | 1412 | 1466 |
| 1520 | 1574 | | | | | | | | |

4 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| - | - | 90 | 120 | 152 | 186 | 222 | 260 | 306 | 360 |
| 418 | 480 | 542 | 602 | 660 | 716 | 773 | 831 | 890 | 950 |
| 1009 | 1068 | 1127 | 1186 | 1244 | 1300 | 1352 | 1402 | 1452 | 1502 |
| 1552 | 1602 | | | | | | | | |

5 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 30 | 64 | 100 | 138 | 186 | 240 | 288 | 340 | 400 | 465 |
| 525 | 585 | 640 | 700 | 762 | 826 | 890 | 965 | 1130 | 1194 |
| 1256 | 1316 | 1376 | 1436 | 1496 | 1556 | 1614 | 1672 | 1730 | 1788 |
| 1848 | 1906 | | | | | | | | |

6 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 32 | 68 | 106 | 146 | 196 | 252 | 302 | 356 | 418 | 485 |
| 547 | 609 | 666 | 728 | 792 | 856 | 920 | 995 | 1160 | 1224 |
| 1286 | 1346 | 1406 | 1466 | 1526 | 1586 | 1644 | 1702 | 1740 | 1818 |
| 1878 | 1936 | | | | | | | | |

7 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 34 | 72 | 112 | 154 | 206 | 264 | 316 | 372 | 436 | 505 |
| 569 | 633 | 692 | 756 | 822 | 886 | 950 | 1025 | 1190 | 1254 |
| 1316 | 1376 | 1436 | 1496 | 1556 | 1616 | 1674 | 1732 | 1770 | 1848 |
| 1908 | 1966 | | | | | | | | |

8 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 28 | 58 | 92 | 122 | 154 | 188 | 210 | 248 | 295 | 354 |
| 398 | 436 | 468 | 503 | 544 | 598 | 655 | 710 | 764 | 818 |
| 870 | 931 | 981 | 1031 | 1087 | 1142 | 1196 | 1248 | 1270 | 1293 |
| 1313 | 1336 | | | | | | | | |

9 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20 | 44 | 70 | 96 | 121 | 160 | 200 | 244 | 295 | 342 |
| 385 | 437 | 488 | 545 | 618 | 682 | 742 | 800 | 868 | 940 |
| 1110 | 1180 | 1250 | 1320 | 1390 | 1460 | 1530 | 1600 | 1670 | 1740 |
| 1810 | 1880 | | | | | | | | |

10 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 28 | 52 | 74 | 96 | 126 | 162 | 195 | 232 | 268 | 314 |
| 355 | 400 | 446 | 498 | 538 | 592 | 650 | 705 | 762 | 816 |
| 876 | 936 | 996 | 1056 | 1116 | 1176 | 1236 | 1296 | 1356 | 1416 |
| 1476 | 1536 | | | | | | | | |

11 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 9 | 20 | 33 | 50 | 72 | 94 | 116 | 140 | 171 | 204 |
| 231 | 253 | 292 | 326 | 362 | 392 | 428 | 468 | 505 | 550 |
| 595 | 640 | 684 | 728 | 772 | 816 | 860 | 904 | 948 | 992 |
| 1036 | 1080 | | | | | | | | |

12 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 22 | 36 | 55 | 78 | 96 | 120 | 144 | 176 | 208 |
| 233 | 254 | 287 | 318 | 352 | 380 | 412 | 450 | 482 | 514 |
| 547 | 580 | 614 | 654 | 694 | 734 | 774 | 814 | 854 | 894 |
| 934 | 974 | | | | | | | | |

13 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 20 | 32 | 50 | 68 | 83 | 100 | 121 | 142 | 165 |
| 188 | 210 | 233 | 260 | 290 | 313 | 340 | 378 | 412 | 448 |
| 484 | 520 | 556 | 596 | 636 | 676 | 716 | 756 | 796 | 836 |
| 876 | 916 | | | | | | | | |

14 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 9 | 18 | 28 | 44 | 60 | 78 | 94 | 114 | 134 | 156 |
| 178 | 198 | 219 | 245 | 274 | 297 | 324 | 361 | 395 | 430 |
| 466 | 502 | 538 | 578 | 618 | 658 | 698 | 738 | 778 | 818 |
| 858 | 898 | | | | | | | | |

15 cm.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 26 | 41 | 62 | 88 | 102 | 146 | 178 | 210 | 240 |
| 275 | 315 | 355 | 395 | 435 | 475 | 515 | 555 | 595 | 635 |
| 675 | 715 | 755 | 895 | 935 | 975 | 1015 | 1055 | 1095 | 1135 |
| 1175 | 1215 | | | | | | | | |

16 cm. (larynx)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 15 | 30 | 40 | 60 | 86 | 118 | 143 | 182 | 216 | 245 |
| 270 | 310 | 350 | 390 | 430 | 470 | 510 | 550 | 590 | 630 |
| 670 | 710 | 750 | 890 | 930 | 970 | 1010 | 1050 | 1090 | 1130 |
| 1170 | 1210 | | | | | | | | |

## V. Testing of the algorithm

How does one say that one area function extraction algorithm is better than another? This is again an area with little theory to guide an evaluator. Two basic methods of evaluation are apparent--some sort of error function or detailed testing and observation of the qualitative and quantitative performance.

A quantitative error function should have several characteristics. First, it is desirable that such a mapping $d(A,A')$ between two area functions and a scalar be a metric--i.e. $d(A,A')\geq 0$, $d(A,A')=0$ iff $A=A'$, $d(A,A')=d(A',A)$ and $d(A,A'')\leq d(A,A')+d(A',A'')$--if the comparisons are going to be meaningful. Second, perturbation theory results suggest that, if the lengths and volumes of the area functions are equal, the metric should be based on the area ratios (preferably the log area ratios) of the corresponding points of each area function. Third, some means is required for dealing with the case where the lengths of the area functions differ if the length is an output of either of the competing algorithms.

A metric satisfying the above conditions for equal lengths is:

$$d(A,A') = \left( \int_0^L (\ln A(x) - \ln A'(x))^2 dx \right)^{1/2} \qquad (11)$$

where $\overline{\ln A(x)} = \overline{\ln A'(x)}$

This ignores any area normalization, which is desirable if one considers the constraints of acoustic methods but ignores the existence of the normalization problem itself. It is also incapable of insights such as the tendency of Wakita's area function extractor to give better results at the front of the tract than at the back of the tract. The error function tests will probably be more useful when area function estimators that have less systematic error are developed.

Intelligent visual comparison based on extensive testing is probably the best method available for the present area function estimators. It lacks the ability of a metric to give exact answers but yields genuine insights into the nature of errors and the general performance of an algorithm. It appears that several other authors agree as they use superimposed graphs to show their results [7, 19].

As testing of the algorithm over a large set of speakers for all non-nasal sonorants with simultaneous X-ray photography is prohibitive, a more limited scheme is used. First, the algorithm is tested by superimposed plots of the area functions (with the length and volume set equal to

those of X-ray area functions) for Fant's Russian vowels [4]--the training set for the bandwidth function. Attempts were made to test the algorithm on the audio recordings made simultaneously with the X-ray (motion) photographs used by Perkell [21], but the mechanical noise of the camera and the utterance set rendered the results uninteresting. Instead, several sets of data were used without an X-ray standard and displayed as both area functions and midsagittal plane outlines. This unfortunately requires that the reader be familiar with articulatory phonetics to be able to evaluate the figures.

Several sets of data are tested in this way. First area functions generated from published vowel formants [22] are supplied. Second, the area functions of a reduced set of vowels for one subject are supplied. Finally, key frames from "motion area functions" of utterances including non-nasal sonorant consonants are shown (same subject as the vowels). The general intent is to give the reader a knowledge of the strong and weak points of this algorithm rather than to try to determine which algorithm is best.

## VI. Performance

### Length estimation

The method of length estimation based on Wakita's method [28] is tested on Fant's six Russian vowels [4]. It performed fairly well, except that for the phoneme /u/, it produced a length that is far too long--probably as a result of the low first and second formants. If the allowable set of lengths is limited to 16 to 20 cm, then the results improve where the algorithm attempts to exceed these bounds (Table 3).

| Phoneme | L-Fant | L-est (bdd.) | L-est (unbdd.) |
|---|---|---|---|
| i | 17.0 | 17.5 cm | |
| ɛ | 17.0 | 17.5 | |
| a | 17.5 | 18.0 | |
| o | 19.0 | 20.0 | 20.5 |
| u | 20.0 | 20.0 | 23.5 |
| ɨ | 19.5 | 20.0 | 20.5 |

Lengths of vocal tract for Fant's Russian vowels [4] as estimated by the bounded modified Wakita method and the unbounded modified Wakita method.

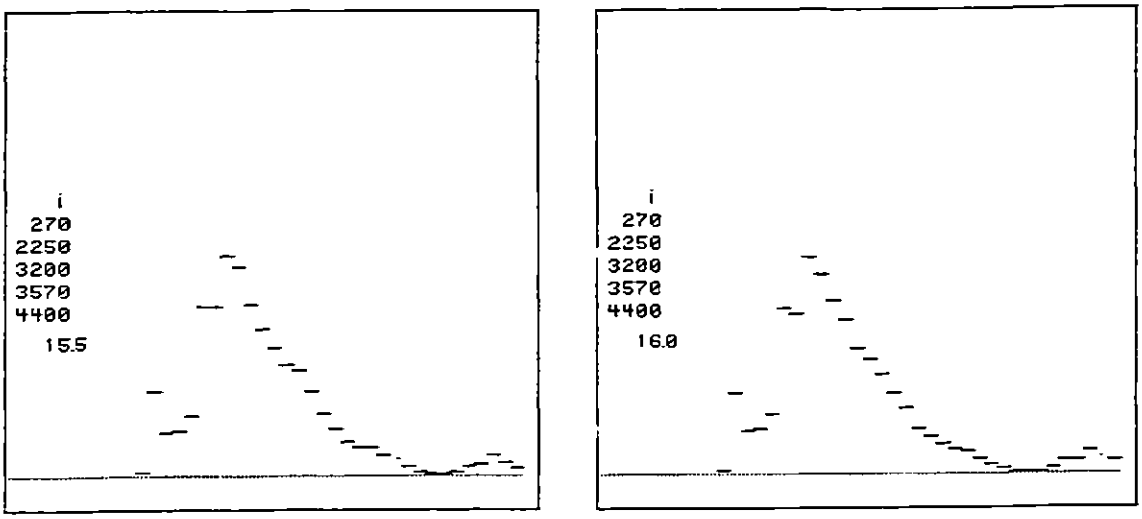
Table 3

When used on a frame by frame basis (no continuity constraints) on continuous speech the method performs erratically. The lengths of the vowels are generally estimated fairly accurately at their centers but the lengths of the consonants are estimated quite poorly. As most consonants involve a constriction, the frequency of the

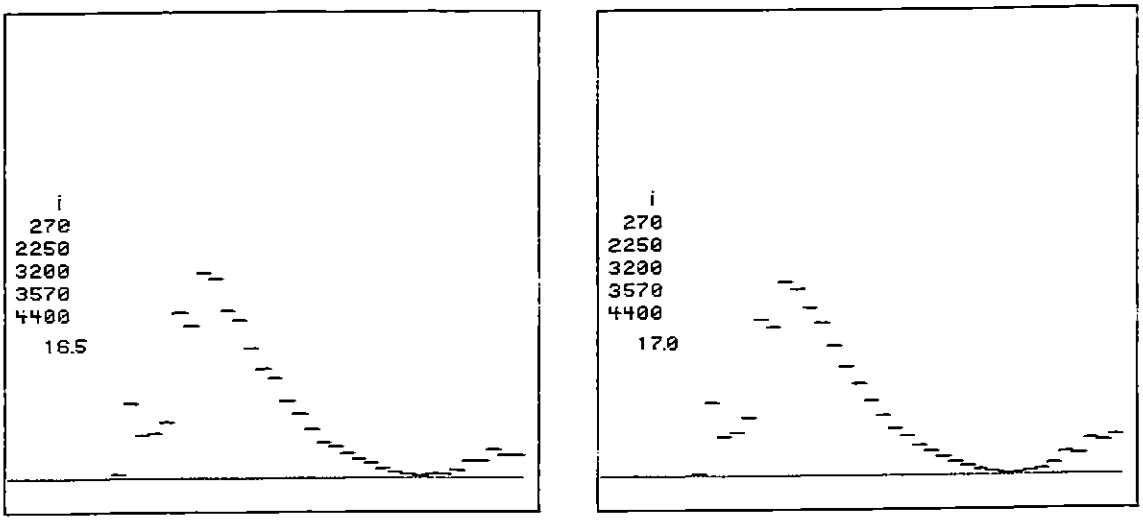
first formant drops, which increases the estimated length. This makes the length of the generated area function increase in an unrealistic manner.

As the length of the vocal tract changes fairly slowly, a method based on more than just the immediate frame of the speech is required. When used to analyze continuous speech, the area function extraction algorithm provides such a method. As illustrated in Figure 7, the lips open too wide when the length is too long. When the length is several centimeters too long, a discontinuity appears at about the proper length in what appears to be an attempt by the algorithm to shorten the tube. The polarity of this discontinuity is such that it approximates the proper lip boundary condition.

This method is primarily useful when analyzing complete utterances containing vowels. Estimates of the length can be made during the vowels and continuity assumed to provide an estimate during the adjacent consonants. While this implementation just assumed length as a constant input for the entire utterance and the iterative length estimation loop was closed by the operator, there is no reason that the length estimation could not be included in an implementation.




a b

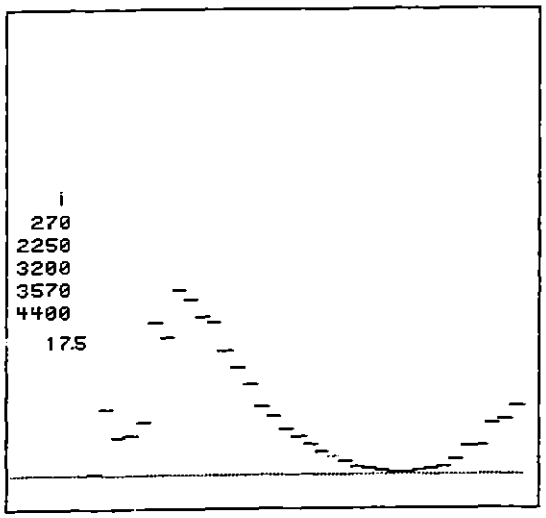



c d

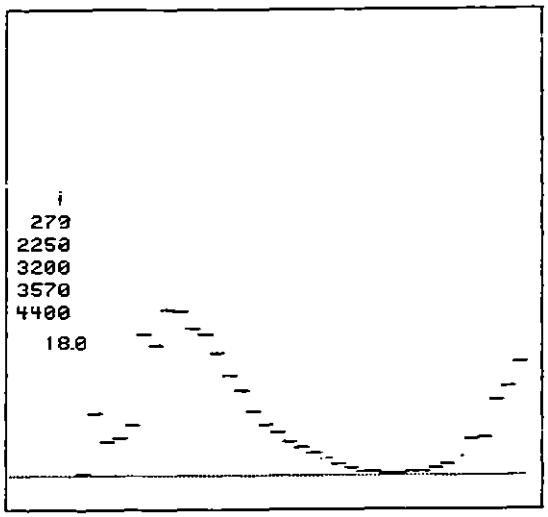
Area function of /i/ as a function of length

Formant data from Fant
Correct L=17 cm.

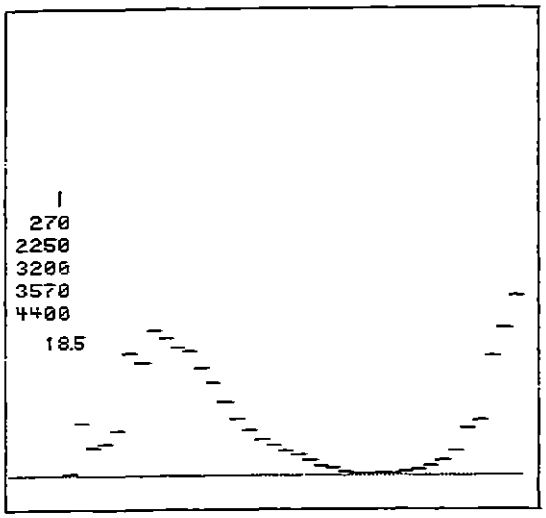
Figure 7.



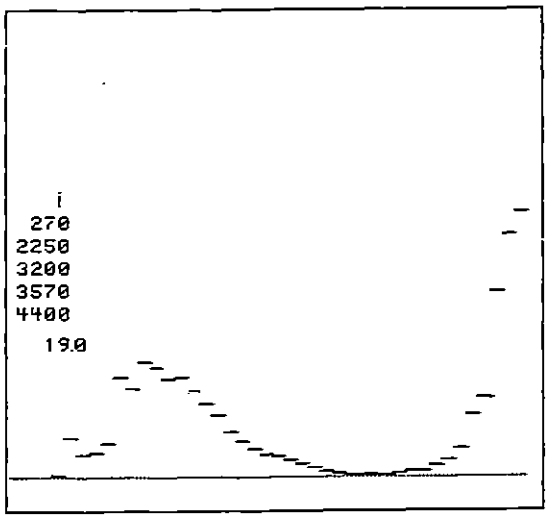

e




f




g




h

Figure 7 cont.

## The area normalization

Little more can be quantitatively said about the constant volume assumption than the earlier stated statistics on Fant's X-ray data [4]. The model does force all areas to be related--i.e. an increase in one section must be accompanied by a decrease elsewhere. This is not generally obvious in the generated area functions except in extreme cases such as the earlier stated one where too long a length causes excessively large areas at the lip end which noticeably shrink the rest of the area function. Qualitatively, the area normalization appears to be fairly realistic in both the area function display and the midsagittal plane outline display. (As the midsagittal plane outline display involves a non-linear position dependent transform from area to distance from the top and back walls of the tract, it is sensitive to the area normalization.)

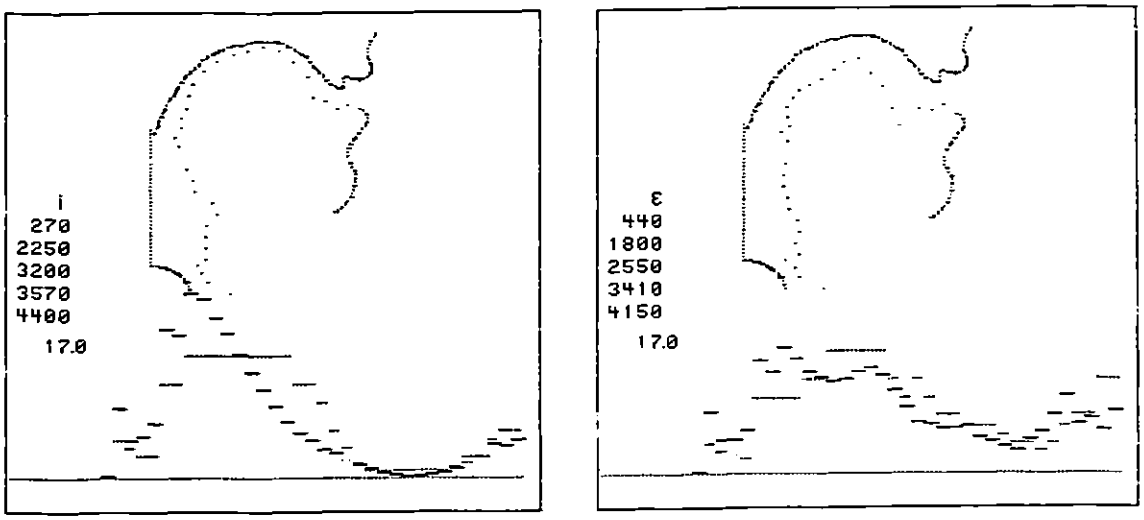
## The area functions

Fant's published formants corresponding to the X-ray area functions [4] were used as data for the algorithm in its first test. As he only measures the first three formants from the acoustic waveform, the fourth and fifth were taken from his table of the formants measured with his transmission line analog LEA set to the same area functions.

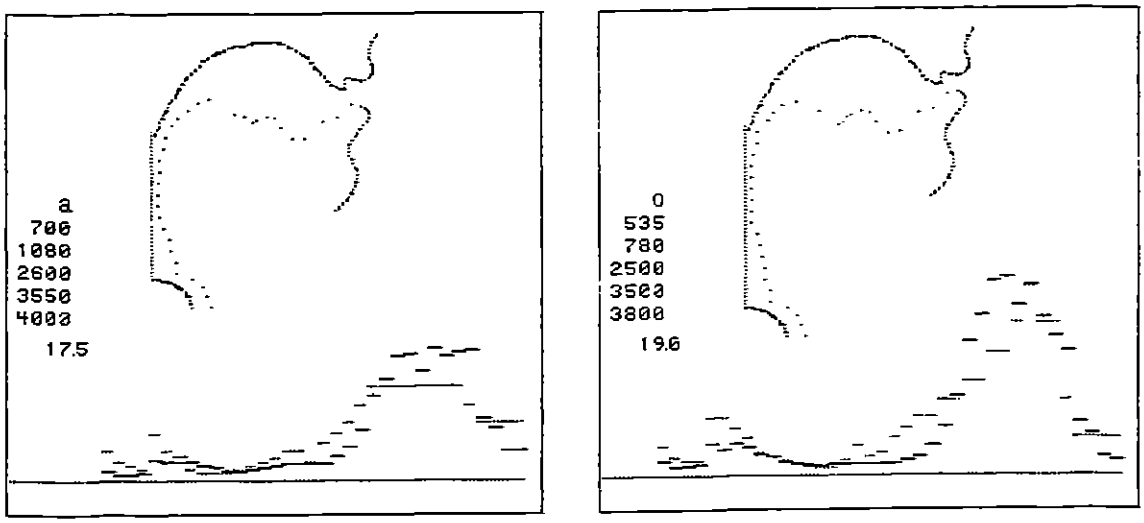
(These, however, should be fairly accurate as the higher modes are not significantly perturbed from those of simple models.) The length and the tract volume for each vowel are set equal to those of the X-ray area functions in order to test only the fundamental algorithm (Figure 8).

The results vary from fair to good depending on the phoneme. /u/, /ɛ/, /a/, and /o/ give the best results. (Remember that the ratio between the correct and the estimated areas is more important than the difference.) The frequency of the first formant in /i/ as given by Fant is too low, probably due to the spectrographic methods used to measure its value. Therefore, its value is increased to 270 Hz which is more likely. The remaining error can be much reduced by hand tailoring one of the bandwidths, which indicates that most of the error occurs in the bandwidth function. /ɨ/ gives the poorest results of the vowels, but it does preserve the gross shape of the desired area function.

One point common to all of the area functions is the preservation of the shape of the glottal zone. As this shape is preserved in an area function with all of the formants set equal to those of a lossless uniform tube (Figure 9), this piece of information must be embodied in the bandwidth function. Additionally, the area functions for Fant's vowels under conditions identical to Figure 8,



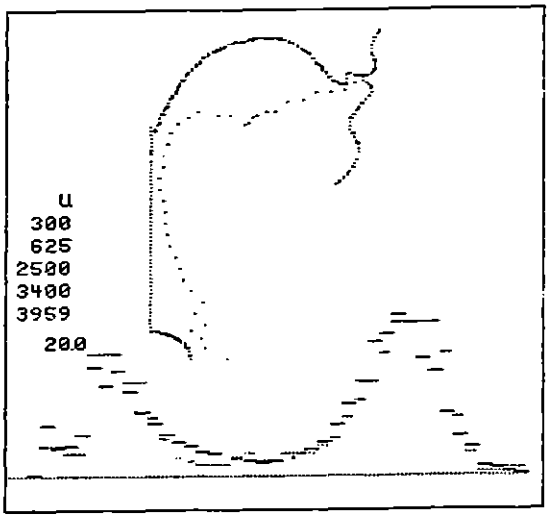

a b



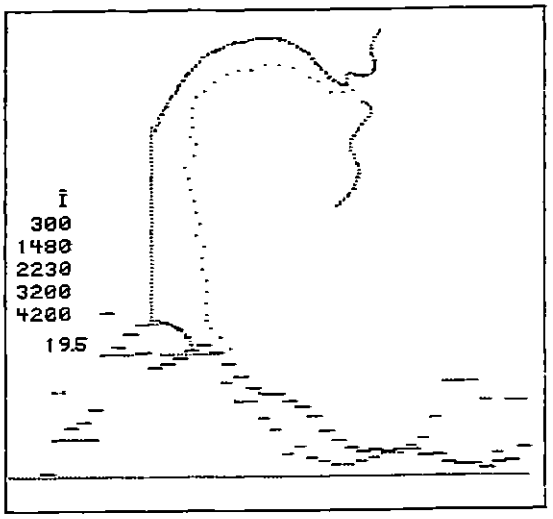

c d

Comparison of Fant's X-ray area functions
with the output of the algorithm
for equal tract lengths and volumes.

Figure 8.

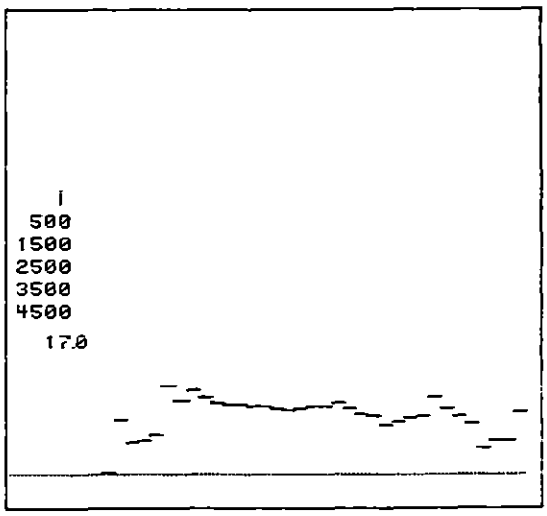



e

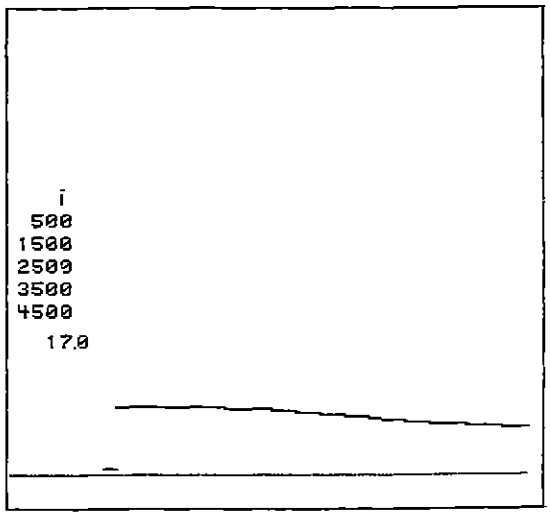



f

Figure 8 cont.




a
W... bandwidth function



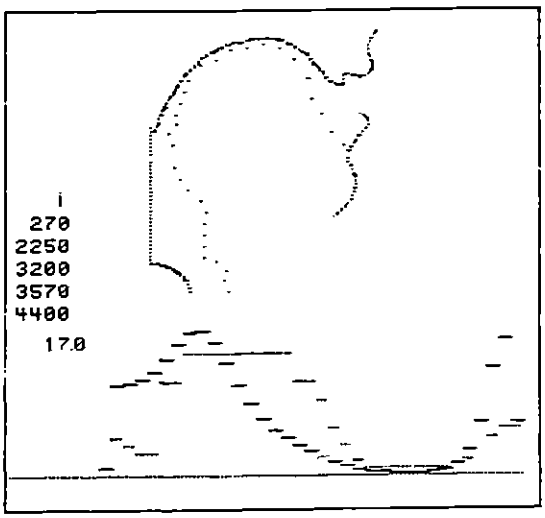

b
With constant bandwidth function
$B_n(F_n)=B_0$

Area functions generated from the formants of a uniform tube.

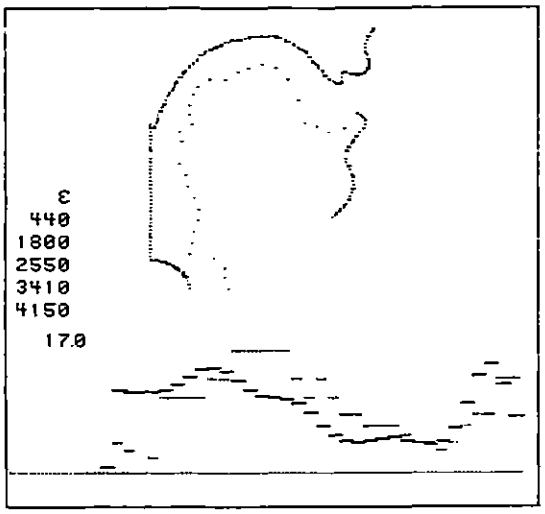
Figure 9.

except that the bandwidth function is changed to $B(f)=B_0$, are displayed in Figure 10. Note that the resolution of the glottal zone and glottal-pharyngeal discontinuity is totally eliminated. The general accuracy of the area functions also decreases both in section by section areas and in the general shapes--some of which no human could produce.

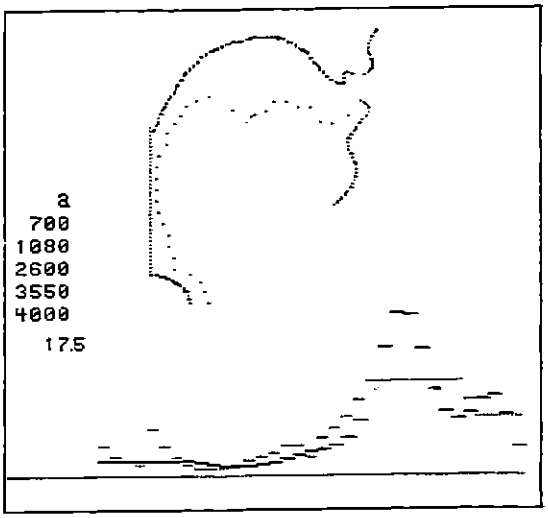
Next the area function estimator is tested on the average (adult male) formants of native American speakers as measured by Peterson and Barney [22]. As only three formants are given (i.e. N=3), length estimation by the modified Wakita scheme loses accuracy over the desired N=5 case. Therefore the lengths are chosen by hand for the best results. The area functions (Figure 11a-j) are fairly reasonable. A couple of general errors are obvious--The glottal region varies in area by a factor of about three to one (Fant's data indicate that it varies far less) and discontinuities--probably artifacts of the bandwidth function--frequently appear in the tract at about two centimeters from the lip end. While no claims can be made for the quantitative accuracy of these area functions, their qualitative accuracy is good, especially if one considers that the input has only four degrees of freedom (three formants and the length). The midsagittal plane displays reveal these vowels all to be recognizable from the estimated area functions although the discriminations



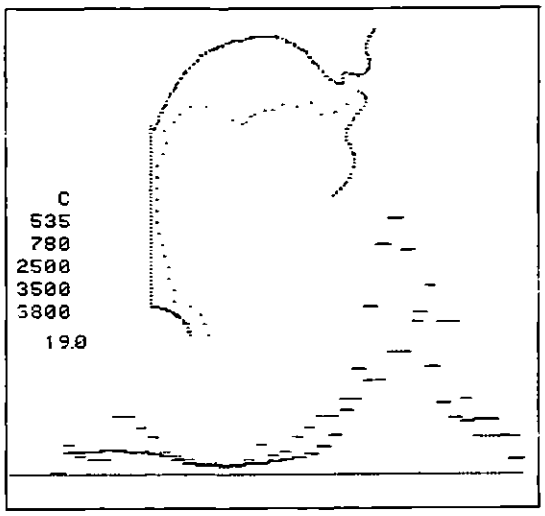

a




b



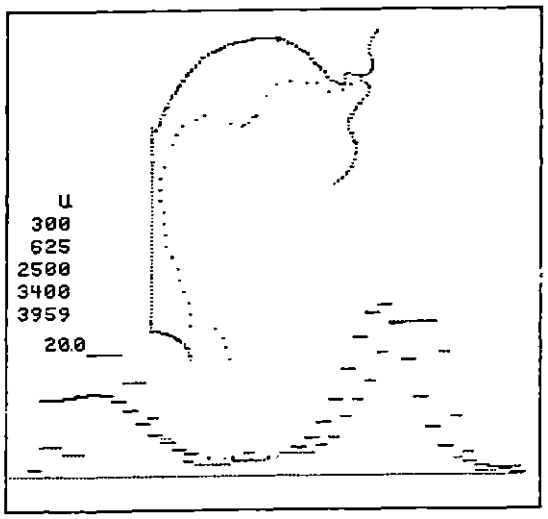

c



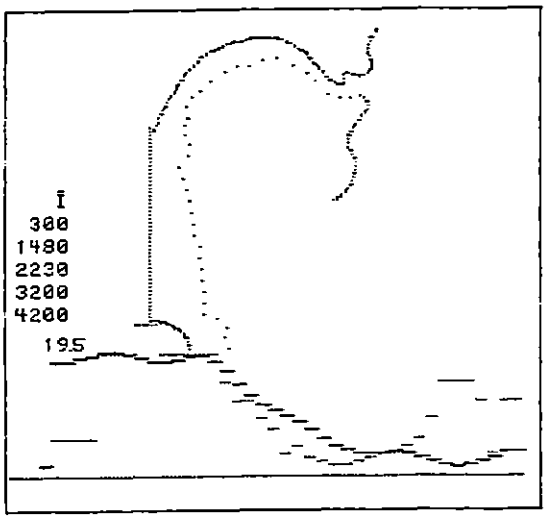

d

Comparison of Fant's X-ray area functions with the output of the algorithm for equal tract lengths and volumes and with constant bandwidth function $B_n(F_n)=B_0$.

Figure 10.

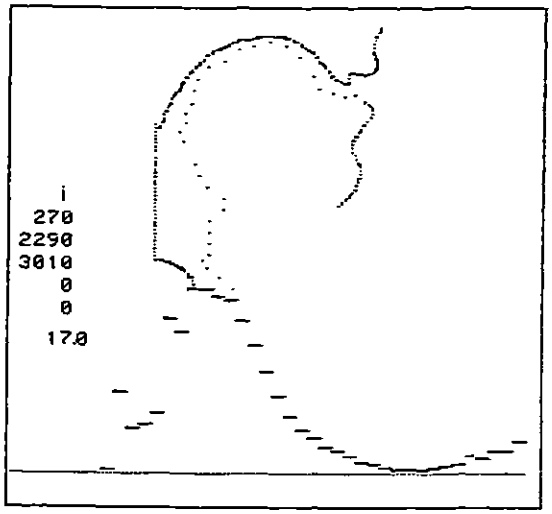



e

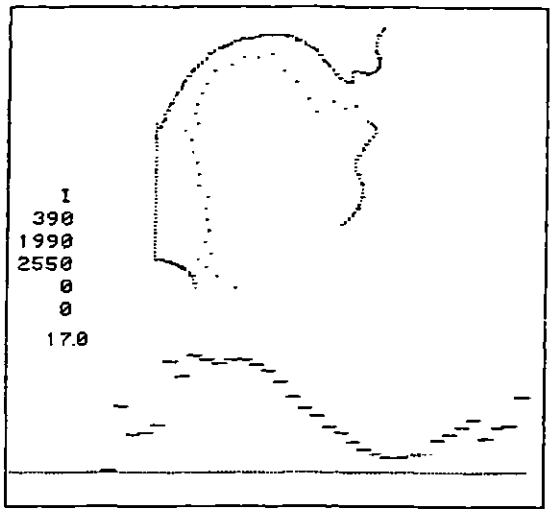



f

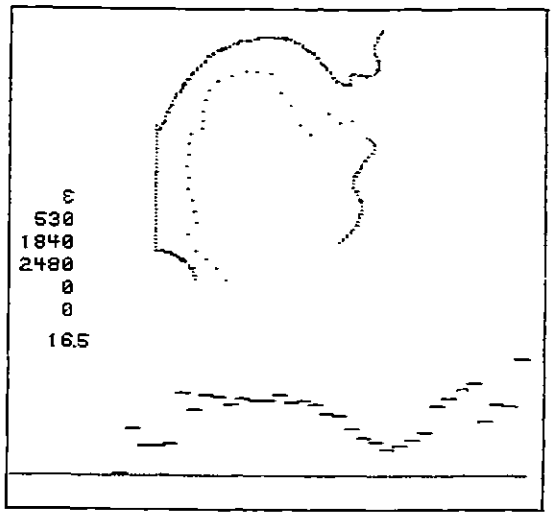
Figure 10 cont.



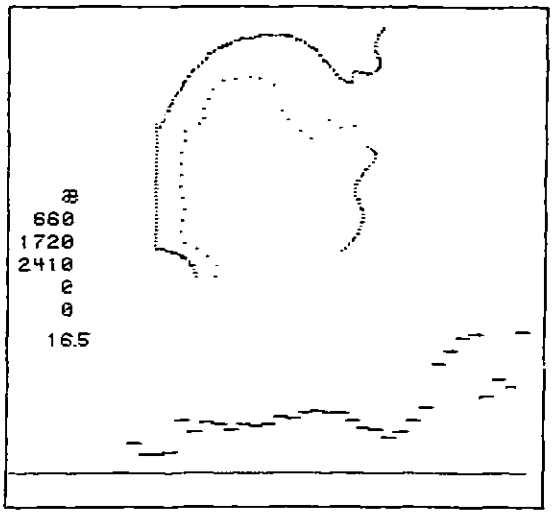

a




b



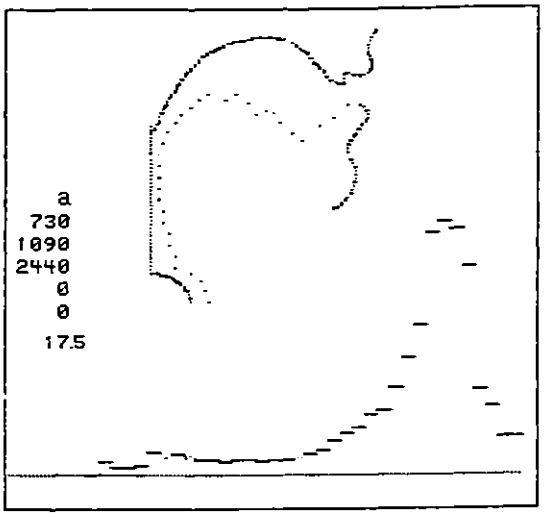

c



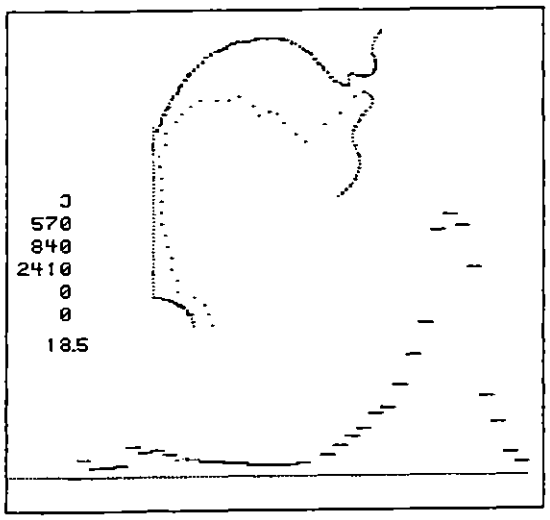

d

Area functions generated from
Peterson and Barney's vowel formants
with length estimated by hand.

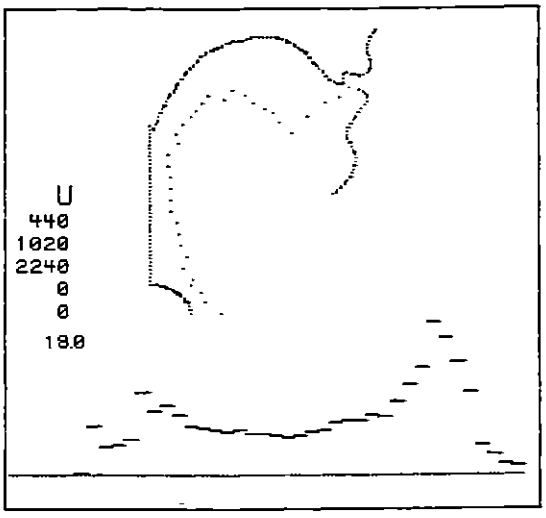
Figure 11.



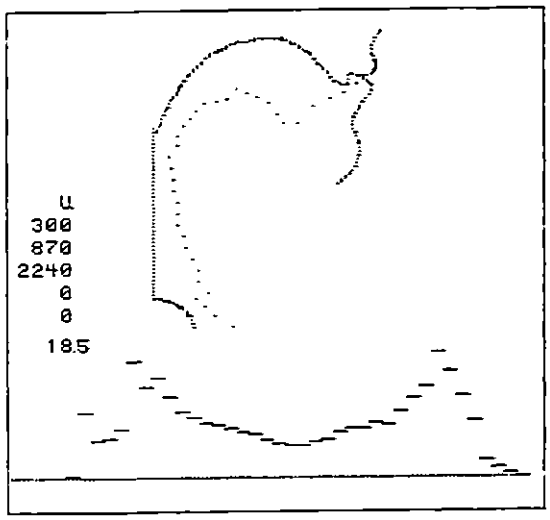

e




f

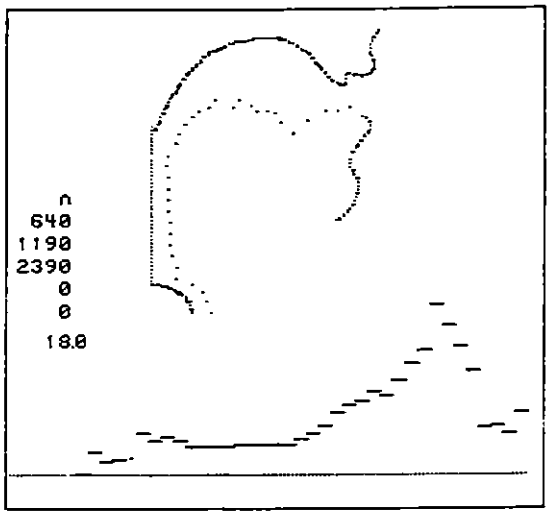



g

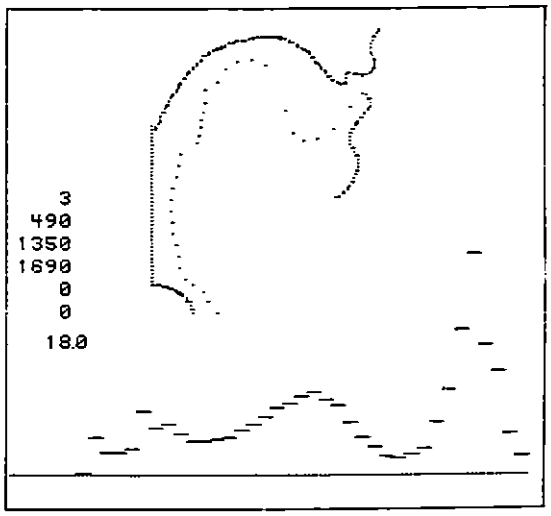



h

Figure 11 cont.




i



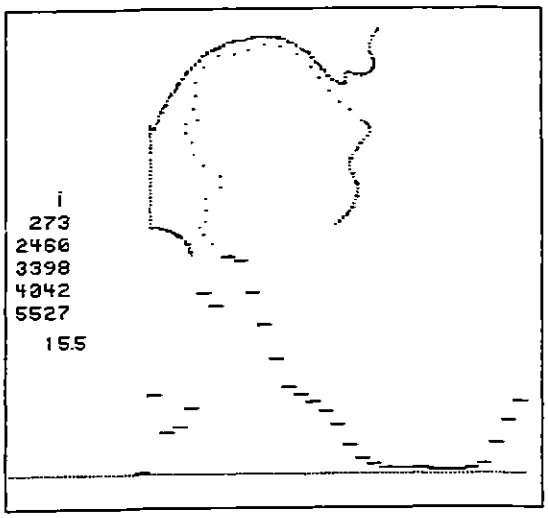

j

Figure 11 cont.

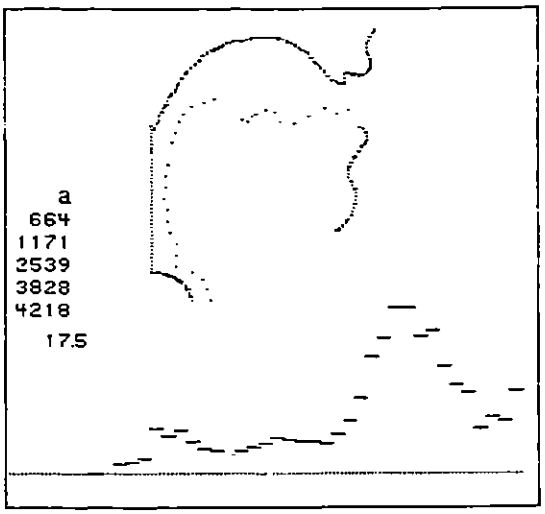
between some minimal pairs (/I/-/ɛ/, /ɛ/-/æ/, /a/-/ɔ/ and /U/-/u/) are subtle.

One subject is tested on the vowels /i/, /a/, and /u/. As all five formants are available here, the modified Wakita length finder is used to determine the lengths of the generated area functions (Figure 12). The results here are similar to the just preceding set of results. In /a/ a slight discontinuity appears near the lips. The pharynx shape for /i/ is not properly estimated. The tongue constriction for /u/ should be smaller. Generally, however, the area functions are fairly good. In the midsagittal plane display, only /u/ appears in error--the constriction is too far back in the tract, but this is the fault of the display, not the area function. The lengths exceed a hand chosen optimum by 0, 1/2, and 1 cm. for /i/, /a/, and /u/ respectively, which indicates adequate performance of the length estimation algorithm. Generally, while not without errors, the system performed quite well on these stationary vowel test cases.

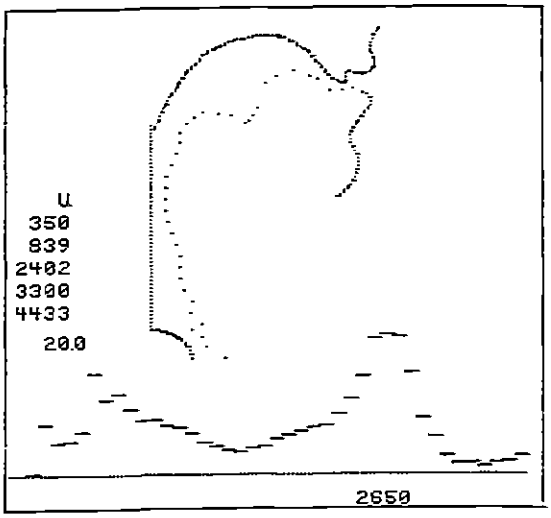
Finally the system is tested on continuous speech from a non-nasal sonorant phoneme set. This method allows the testing of massive quantities of data (thousands of frames were analyzed), which enables one to observe the results of non-target (transitional) sections of speech within their contexts. Unfortunately, it is impossible to present these




a




b




c

Area functions generated from
a native English speaker's vowels.
Length by modified Wakita method.

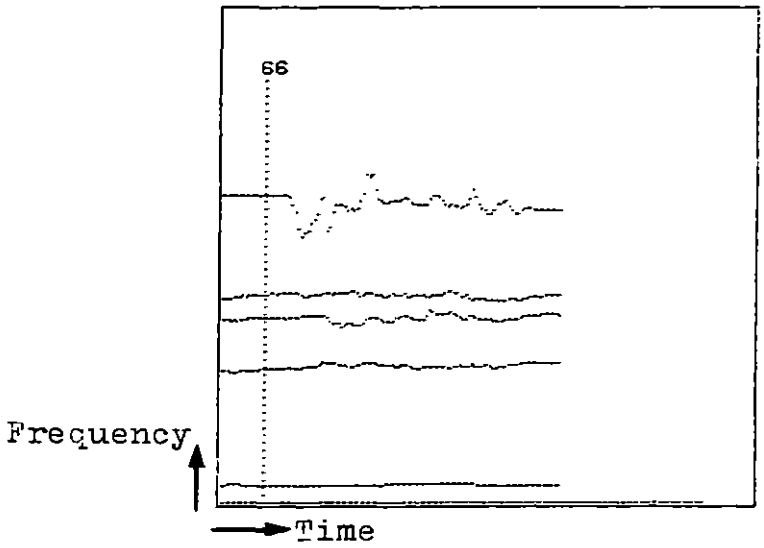
Figure 12.

results to the reader. Therefore, selected key frames will be shown with the intent of demonstrating both the strong and weak points of the algorithm.

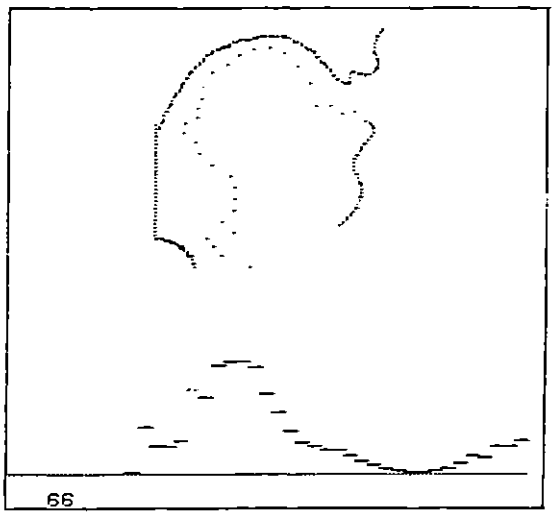
In general, the formant tracks for these utterances are hand edited. This has been done only with frequency domain knowledge of the sort gained by looking at spectrograms. Other than smoothing of the data, the editing is used to continue the formant tracks into regions where they cannot be tracked--i.e. during the closures of (voiced) stops. The editing, with the exception of bad frames missed in the first edit, was done exclusively before the generation of any area functions so that feedback could not affect the results. The lengths, which are constant over each analysis, are supplied by hand due to the earlier mentioned difficulties in the modified Wakita method when applied to continuous speech. The area normalizations are by the constant volume assumption.

The following utterances were analyzed--/i/, /a/, /u/, /bid bad bud/, /aga/, /ɛgɛ/, /ala ili/, /ara iri/, /aya iyi/, /awa iwi/, and /aǰa iǰi uǰu/.

The results for /i/, /u/, and /a/ are similar to the earlier presented results for these vowels (different utterances--same speaker). /i/ (Figure 13) shows its characteristic improperly shaped pharynx (compare with




a
Formant tracks with displayed
frames marked (12.8 mS/frame nr.)



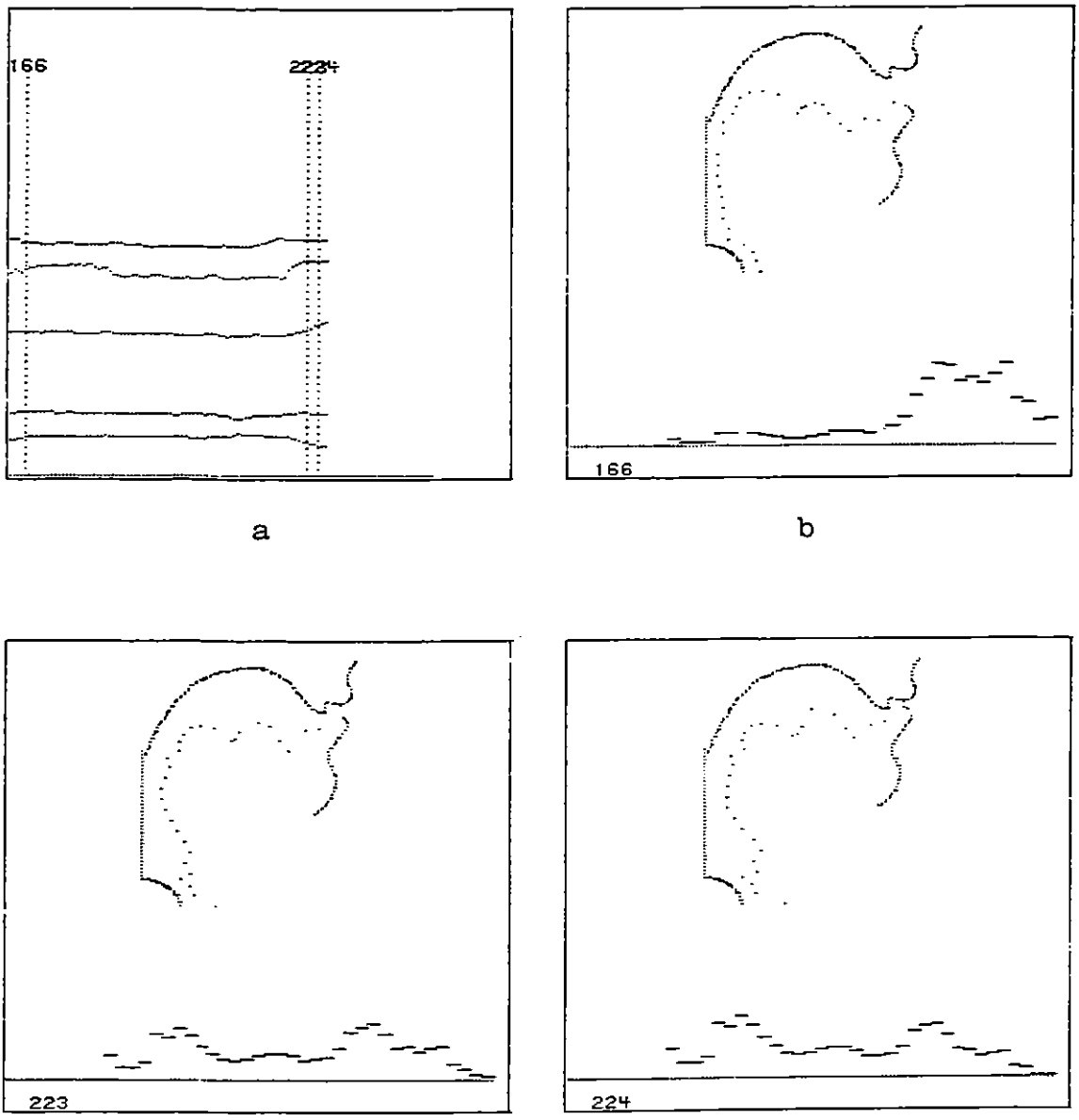

b
Midsaggital plane display
and area function

/i/

Figure 13.

Figure 8a) due to an improper bandwidth. The utterance as a whole is quite stationary and contains little more of interest. /a/ (Figure 10) is quite well shaped--compare it with Figure 8c. This utterance is also mostly stationary with the exception of the last few frames. Figures 14c and 14d show the speaker's relaxation toward the end of the utterance. The tract can be seen neutralizing and the lips closing. The phenomenon occurs in many of the utterances analyzed. In a number of cases, formation of an initial vowel can also be seen. Phonation, however, starts only after the bulk of the formation is completed so that this effect is rarely as pronounced as the termination of final vowels. /u/ (Figure 15) has litte to distinguish it from the stationary case analyzed earlier.

Next a set of three /bVd/ utterances is analyzed. At the start of /bid/ (Figure 16b), the closure occurs at the lips and the shape of the tongue and pharynx anticipate the following vowel. The lips open rapidly as shown in Figure 16c to reach the steady state of Figure 16d. Figures 16e and 16f show the formation of the /d/ closure. The start of /bad/ (Figure 17b) shows the proper initial closure, but less anticipation of the vowel. The following three frames (Figures 17c-17e) capture the rapid release of the /b/ as well as the formation of the vowel /a/. The target (Figure 17f) is somewhat misshapen as the same length was used for

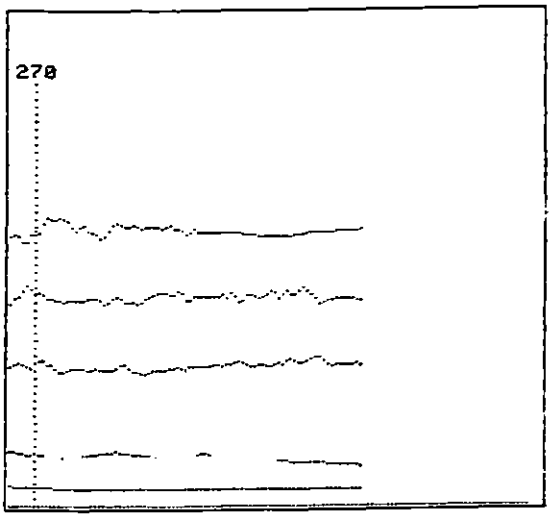



a b

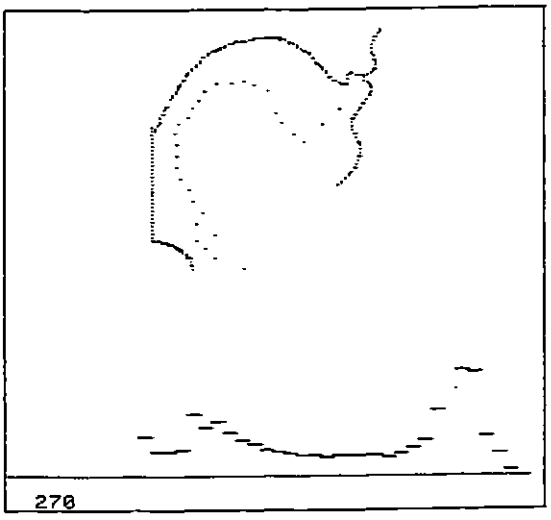
c d

/a/

Figure 14.



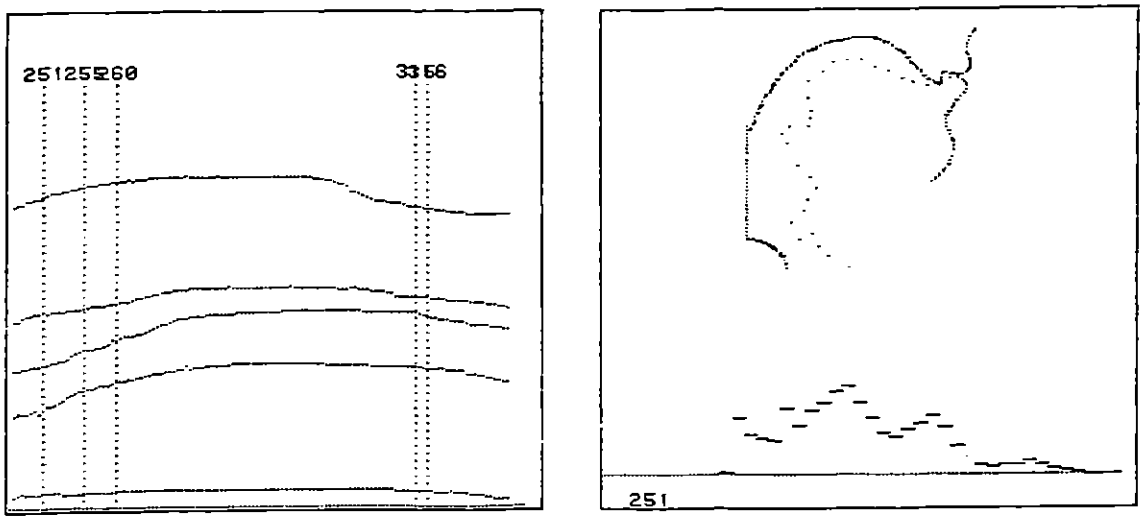

a



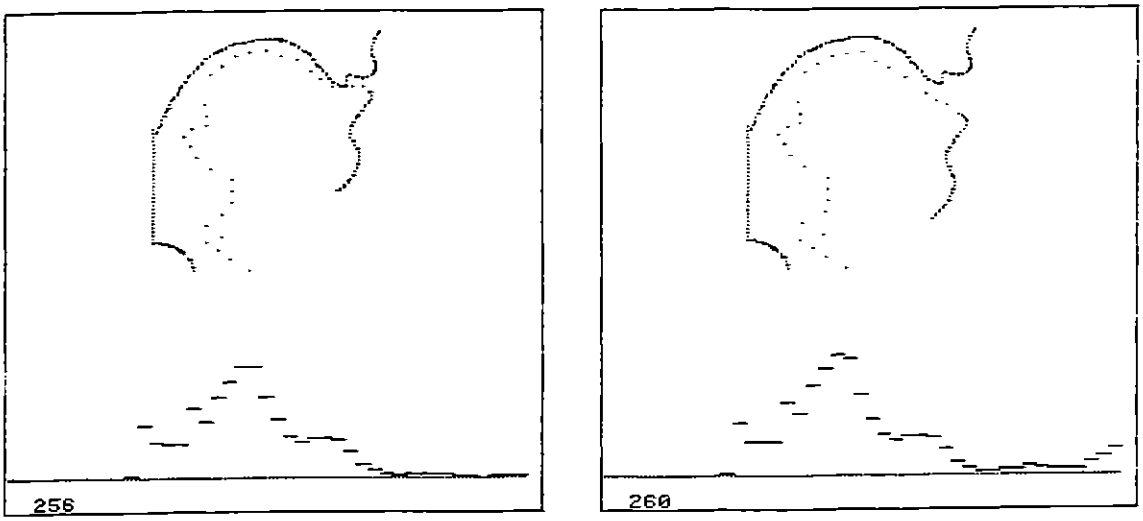

b

/u/

Figure 15.



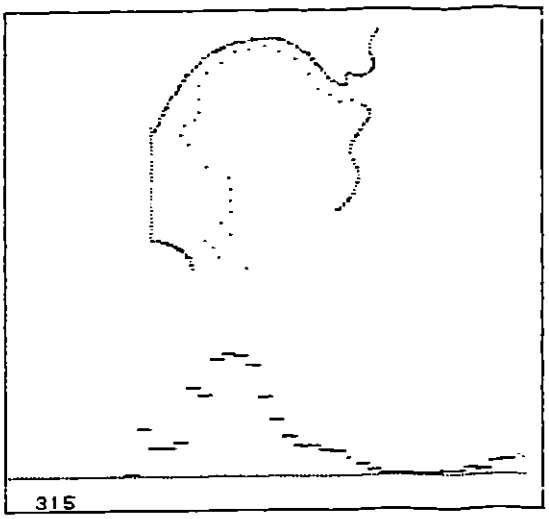

a b



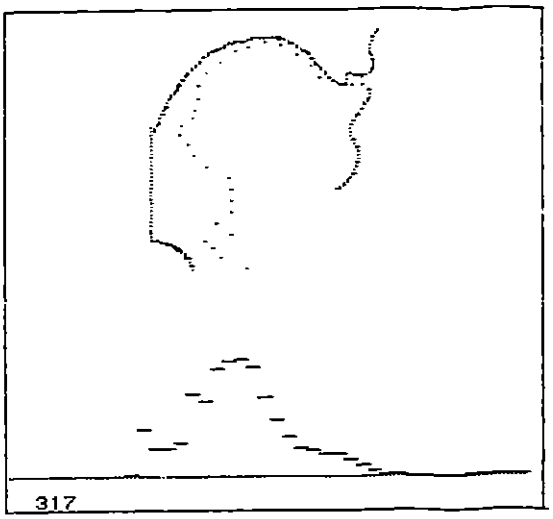

c d

/bid/

Figure 16.



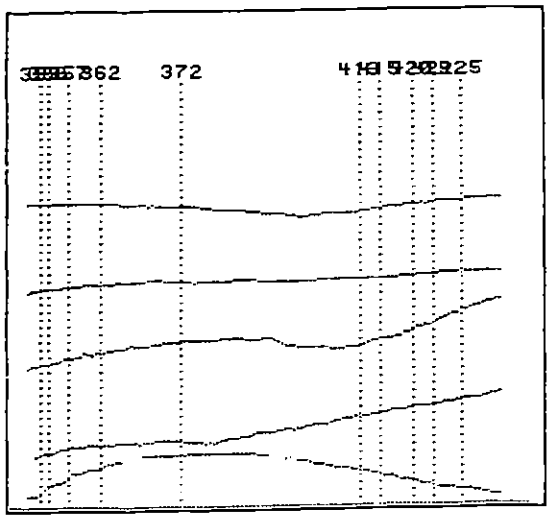

e



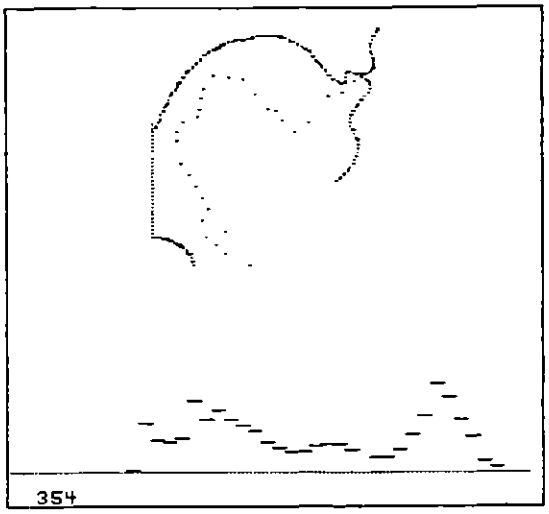

f

/bid/

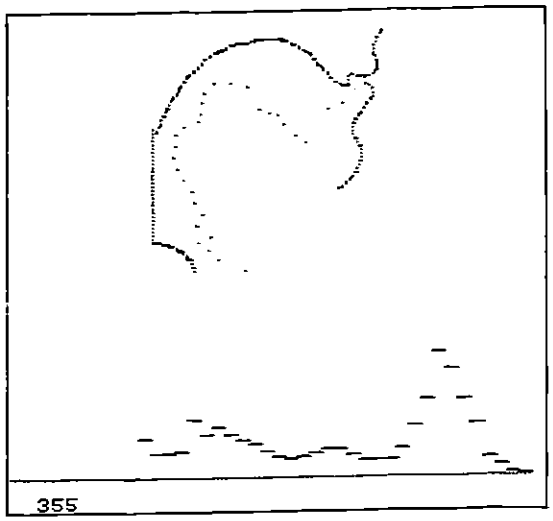
Figure 16 cont.



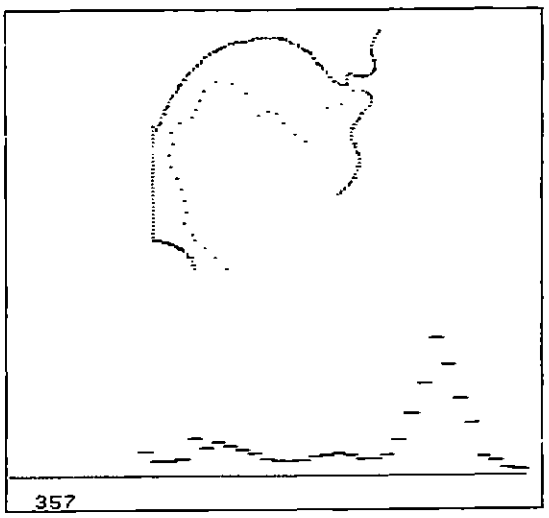

a




b



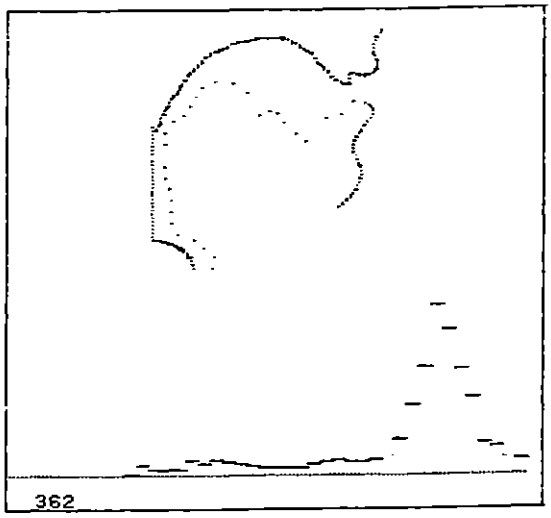

c



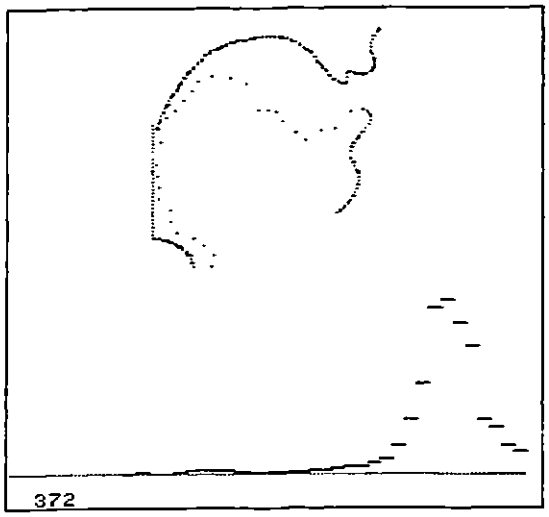

d

/bad/

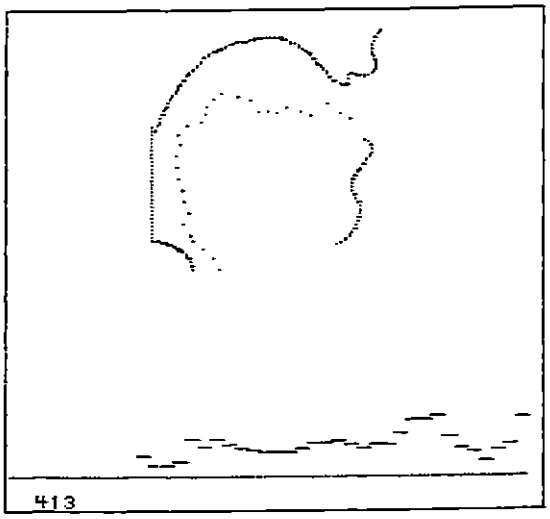
Figure 17.



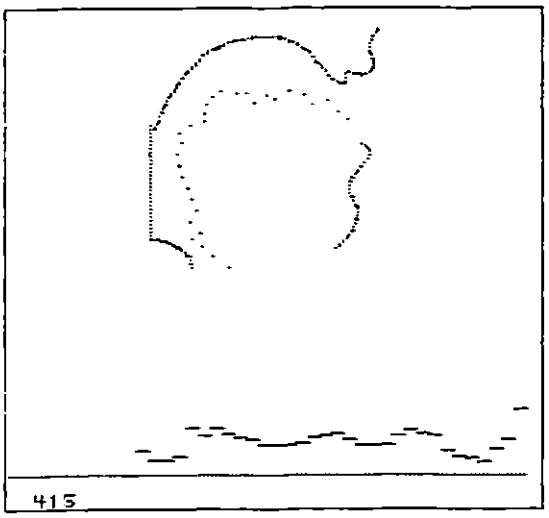

e




f



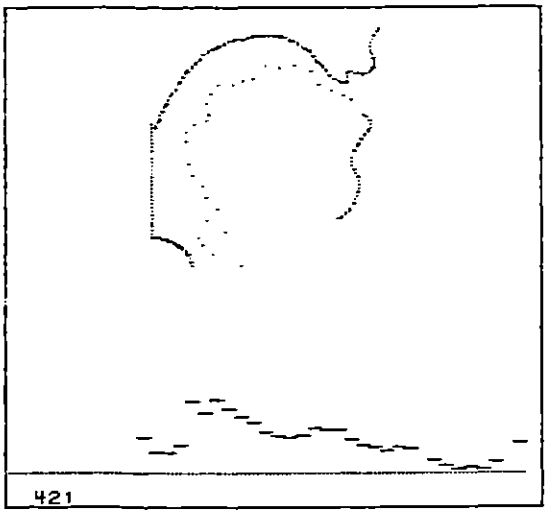

g



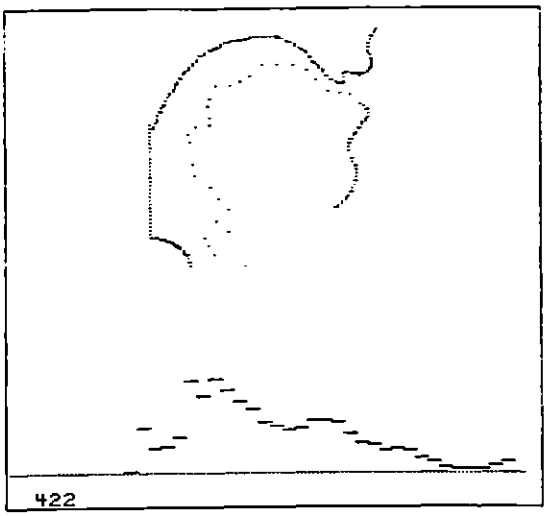

h

/bad/

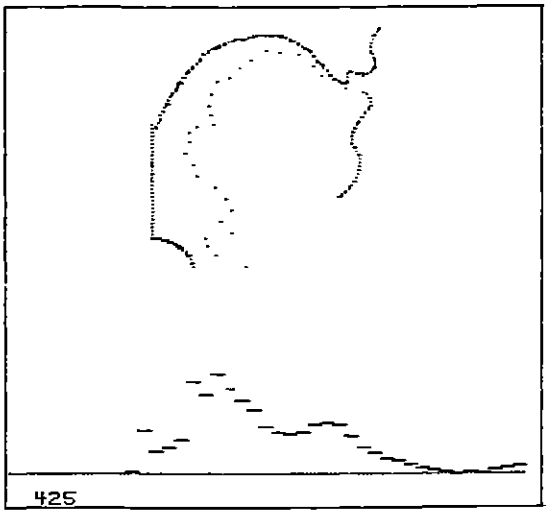
Figure 17 cont.




i




j



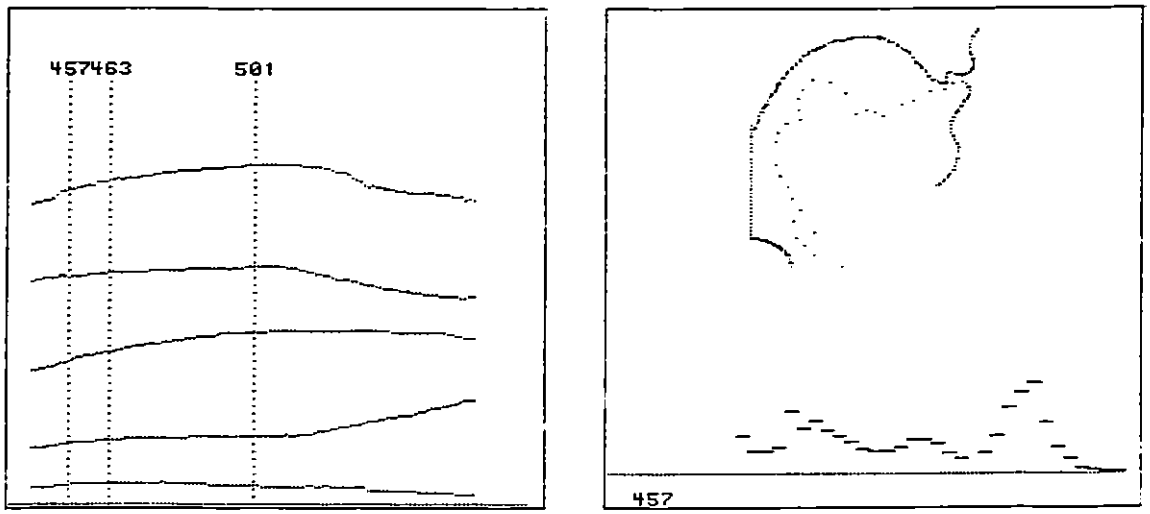

k

/bad/

Figure 17 cont.

all three utterances and the compromise length chosen is too short for /a/. Figures 17g through 17k show the /d/ closure just posterior to the teeth. The beginning of /bud/ (Figure 18b) shows the stop closure just after the opening and the anticipation of the following /u/, which is realized in Figure 18c. The impending /d/ closure is clearly visible in Figure 18d. These three utterances all show several basic features--the point of closure for the initial /b/, the anticipation of and transition into the vowel, and the formation of the final /d/ closure.

Next two /VgV/ utterances are displayed to show the algorithm's performance on the stop /g/. The utterance /aga/ starts with a well formed /a/, this time with the proper length. Figures 19c through 19f indicate that, while the tongue body does move somewhat toward the proper velar closure, the closure occurs incorrectly at the lips. By Figure 19h, the indicated closure releases back to the normal /a/ configuration. Analysis of the utterance /ɛgɛ/, however, does properly analyze the /g/. The analysis (Figure 20) starts with a well formed /ɛ/ and shows the proper formation of the /g/ constriction (Figure 20e) and returns to a much more relaxed /ɛ/ at the end. Why these conflicting results? Several possibilities exist--the former case is a velarized /g/ and the latter is a palatalized /g/, errors due to the approximations inherent

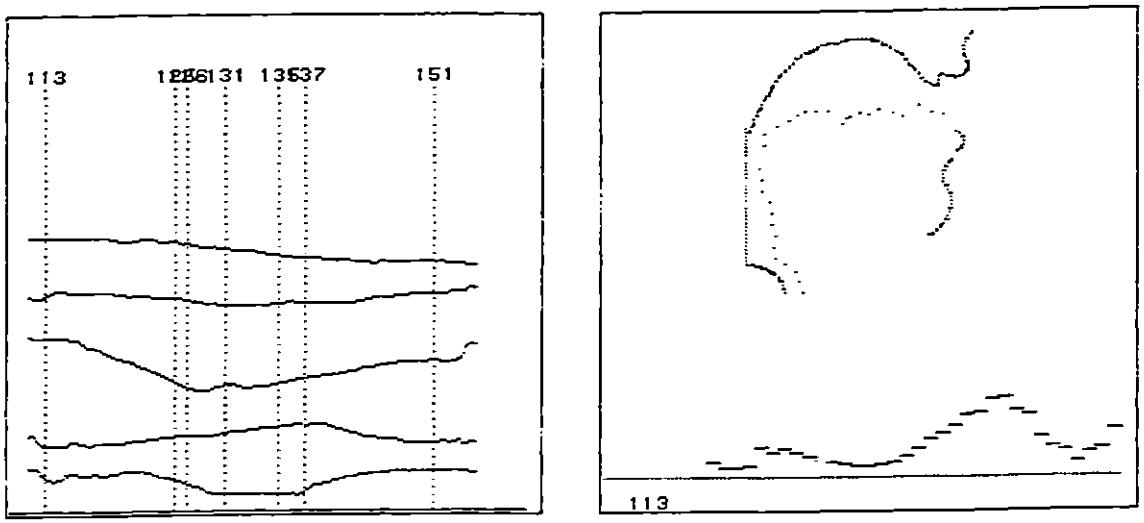



a b

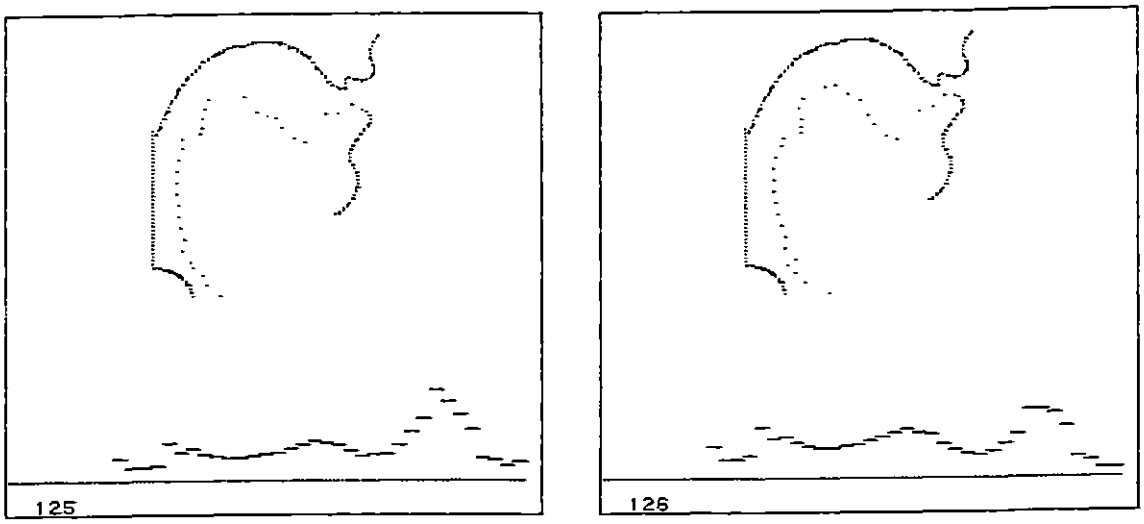
c d

/bud/

Figure 18.



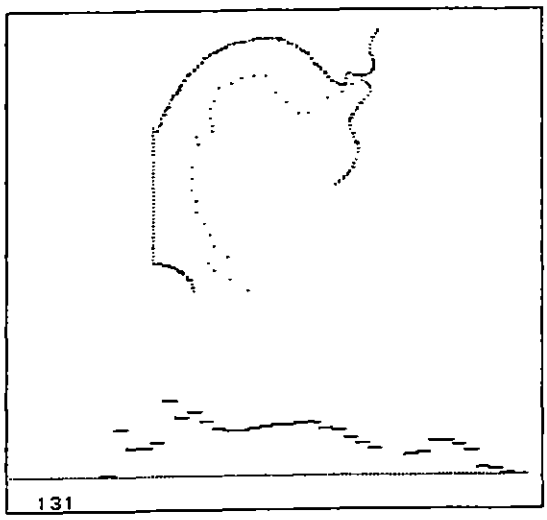

a b



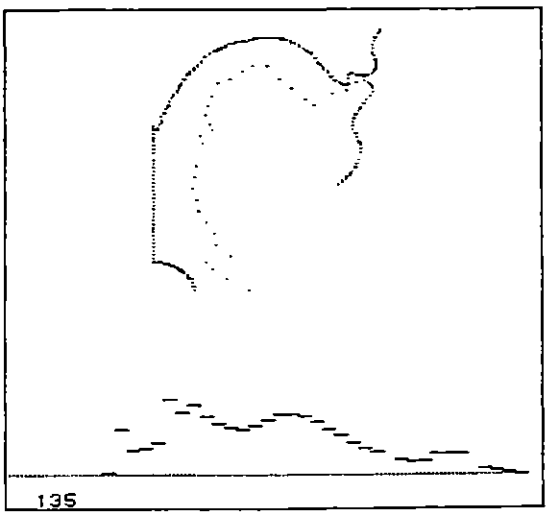

c d

/aga/

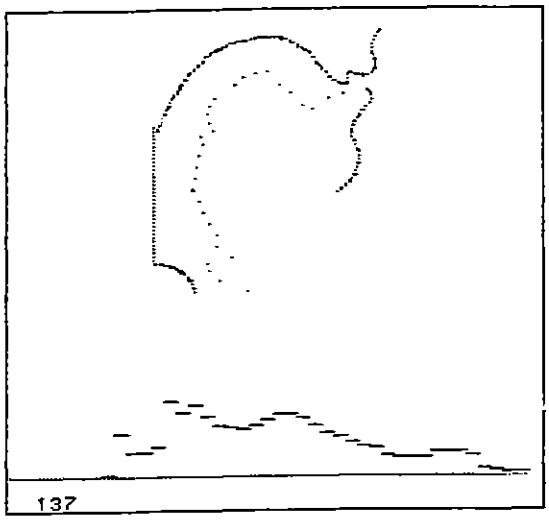
Figure 19.



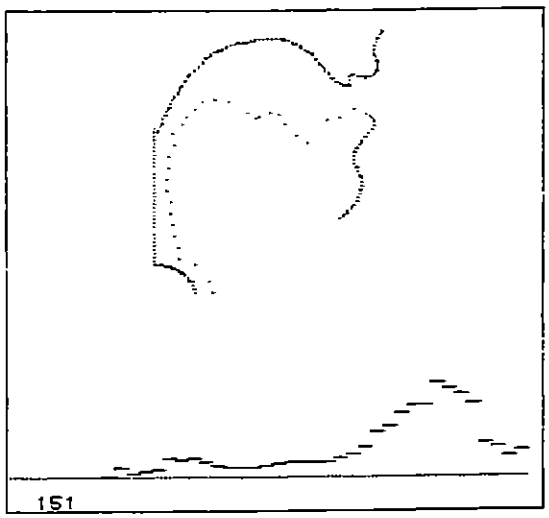

e




f




g




h

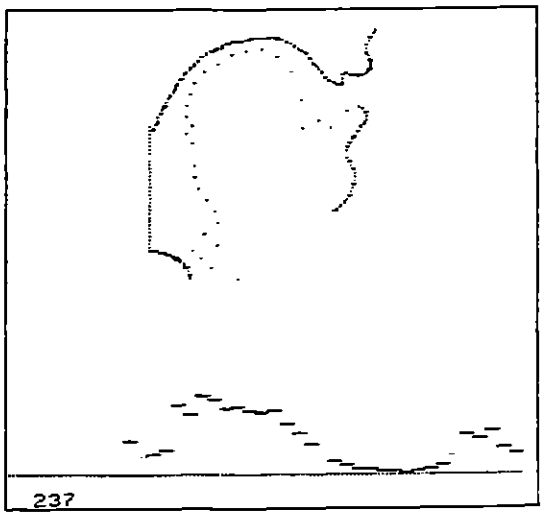
/aga/

Figure 19 cont.

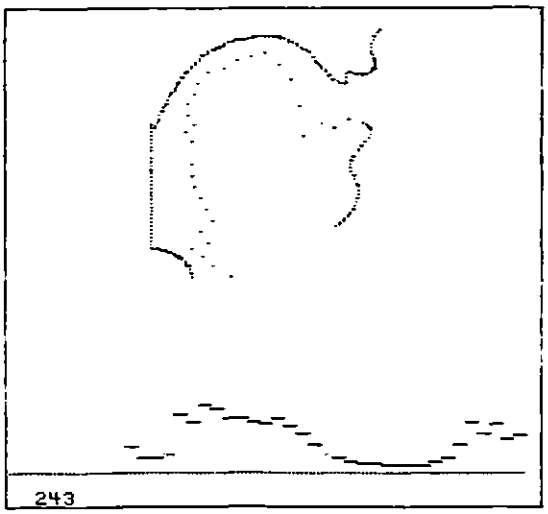
a b c d

/ɛ gɛ/

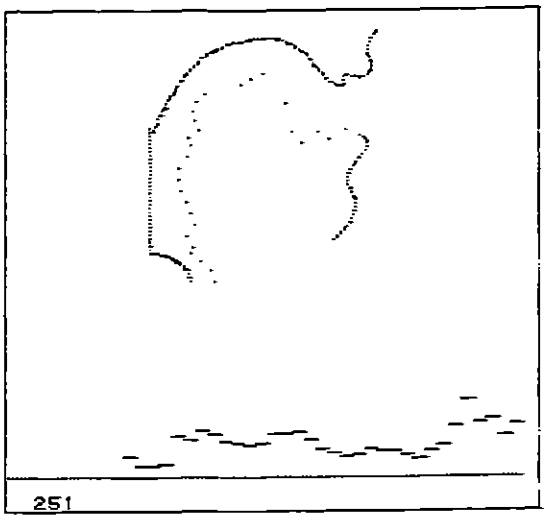
Figure 20.



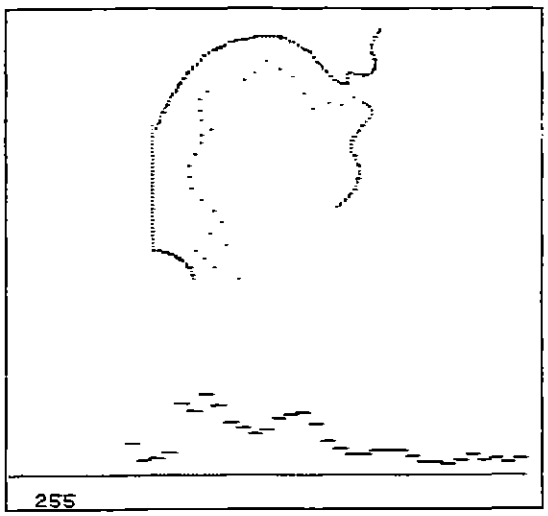

e




f




g



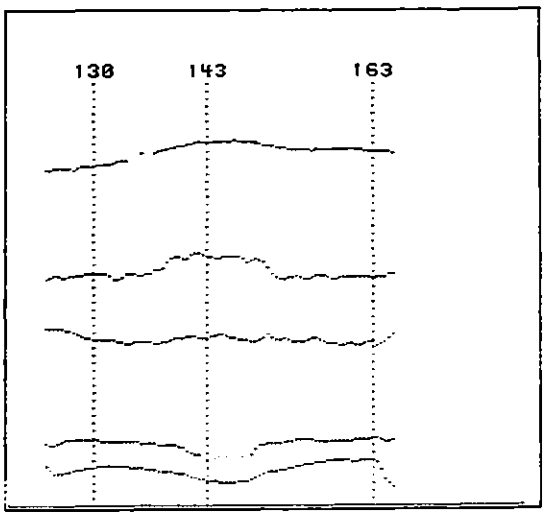

h

/ɛgɛ/

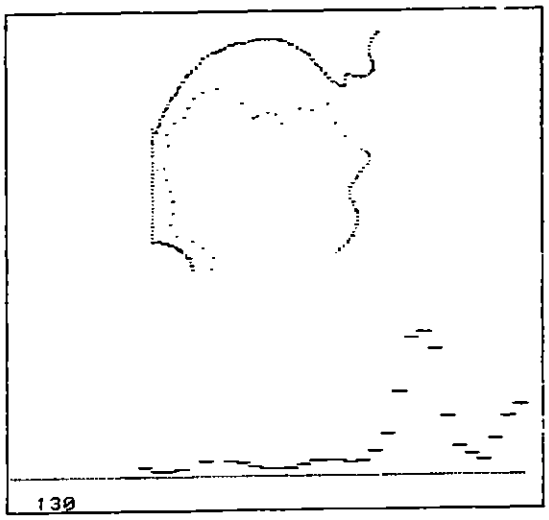
Figure 20 cont.

in the simple bandwidth function as it dominantly embodies the even Fourier components of the (log) area function (which are strong for /g/), or just formant tracking errors.

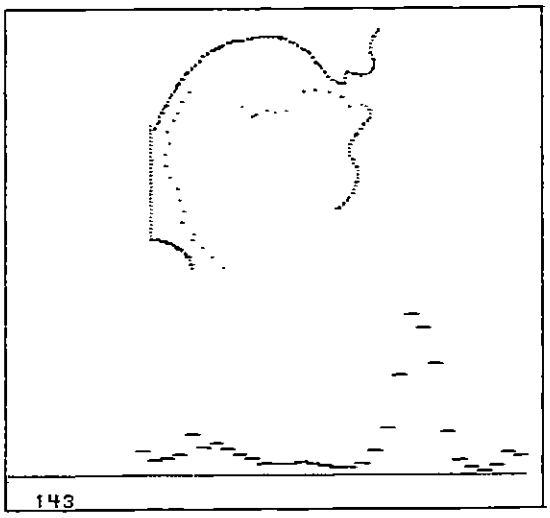
The next utterance analyzed is /ala/. As /l/ is produced by a lateral constriction at the side of the upraised tongue blade, the midsagittal plane display cannot show it as such. It can only show the constriction as if formed between the top of the tongue and the alveolar ridge. Figure 21 shows the algorithm performing as desired--well formed /a/'s at the beginning and end, and a constriction just posterior to the (upper) teeth for the /l/. It also indicates that the tongue root moves slightly forward for the /l/--a necessary consequence of moving the tongue tip upward and forward. The companion utterance, /ili/ (Figure 22), shows essentially similar performance. It begins and ends with the normally formed /i/ and shows the same point of constriction as does /ala/. Additionally, it shows several details--the tongue blade shape is controlled by the consonant--not the vowel--as required by /l/, but the tongue root position is dominantly set by the vowel. The lower lip position during the /l/ is also of interest as it too is strongly influenced by the vowel, being more open for /ala/ as the /a/ sets a lower mandible position than does /i/, which exerts a strong influence on the lower lip.



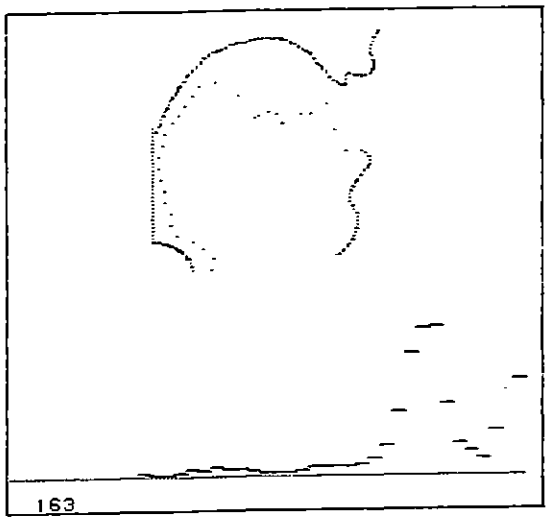

a




b



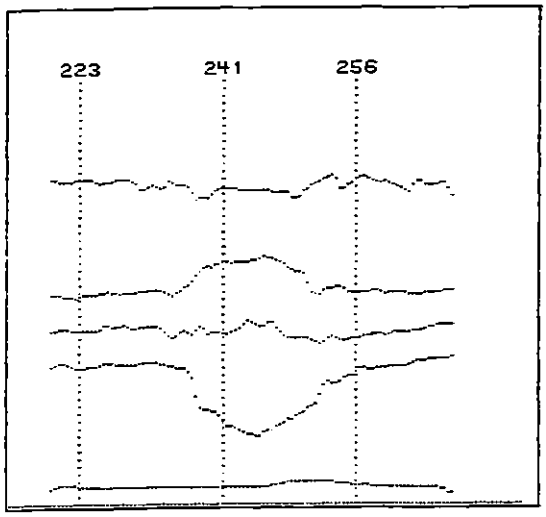

c



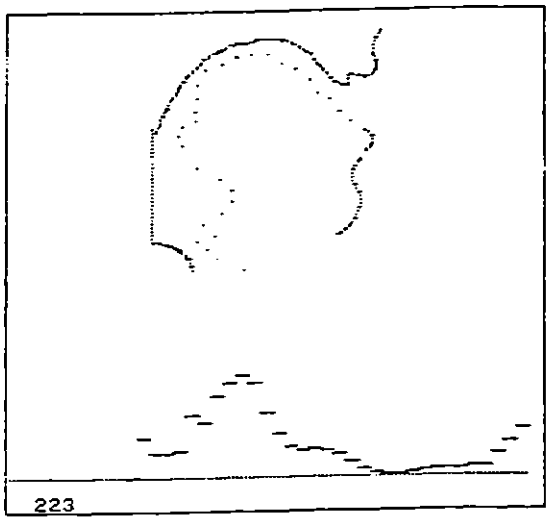

d

/ala/

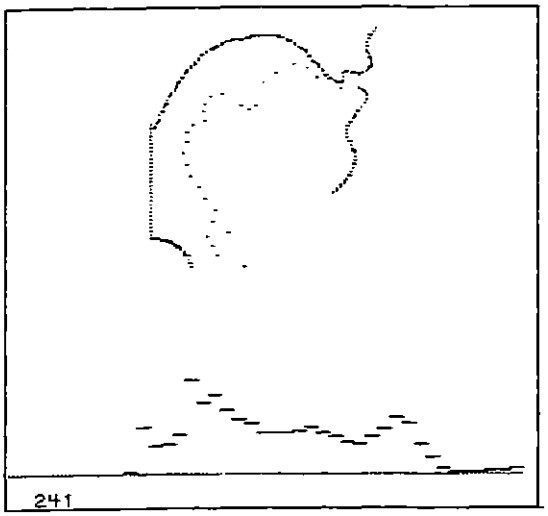
Figure 21.



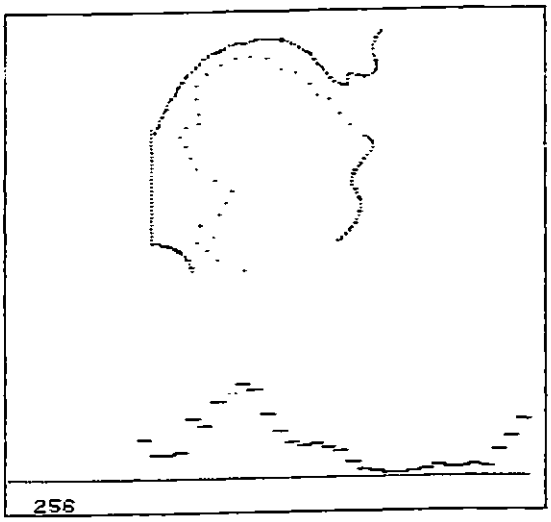

a




b




c



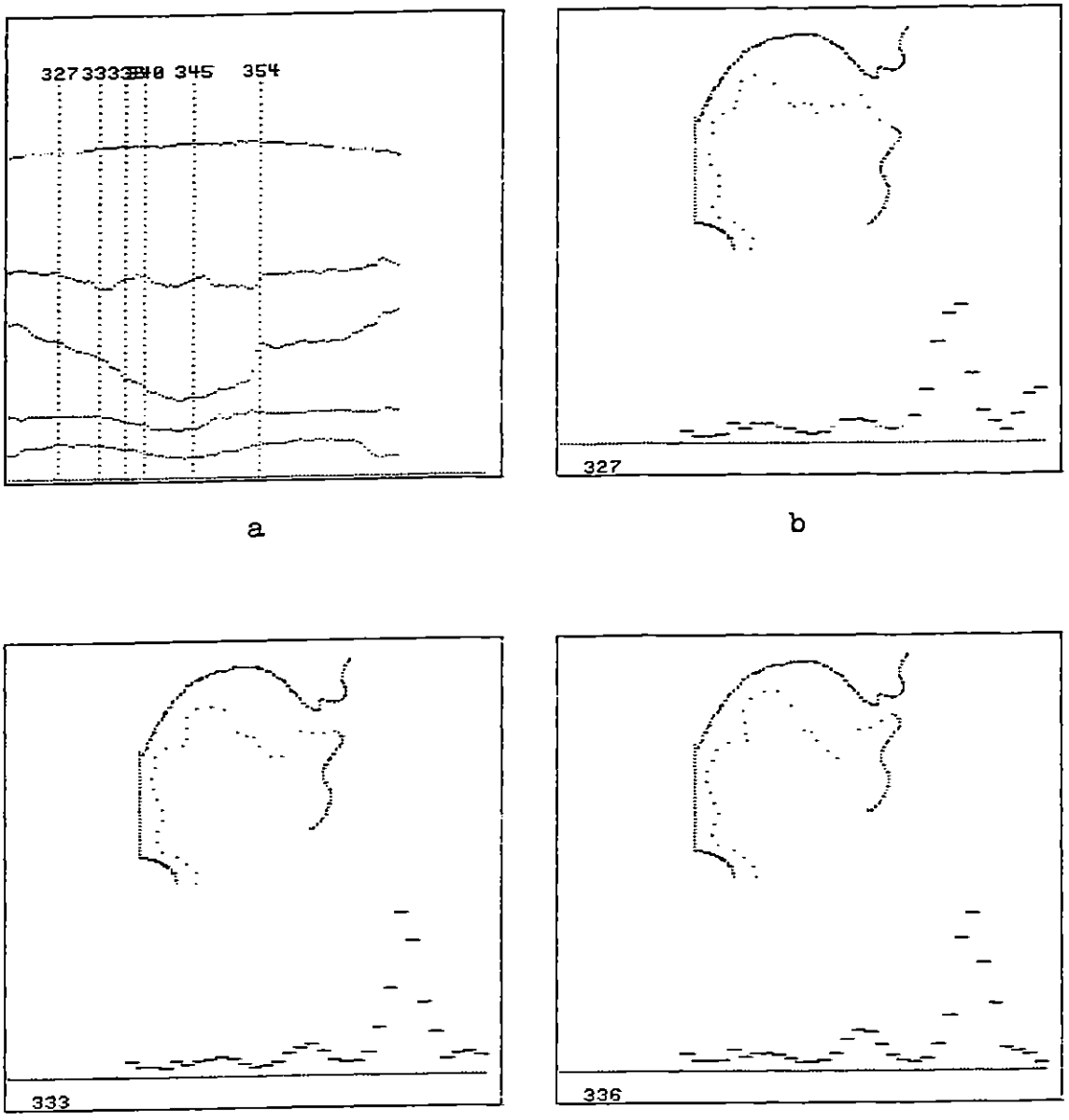

d

/ili/

Figure 22.

/ara/ (Figure 23) and /iri/ (Figure 24) demonstrate both the target postions of the phonemes as well as the transitions. Both utterances are shown to start with a well formed vowel. Each transforms smoothly into the configuration for /r/ (Figures 23f and 24e) with constrictions at both the (hard) palate (slightly too far back in the tract) and the lips. The /r/ tongue root is influenced by the adjacent vowels—/ara/ has less pharynx volume in response to the lesser pharynx volume of /a/. The smaller lip area shown for the /r/ in /ara/ is in error and is probably the result of a single length chosen for both utterances as they were analyzed simultaneously. (The tract for /a/ should be slightly longer than for /i/.)

Two glides in the VCV format are also analyzed. /y/ can be approximated as the transition between /i/ and the following vowel. Analyses of /aya/ and /iyi/ are displayed in Figures 25 and 26. /aya/ starts with a reasonably formed /a/ (the length is slightly short). The lip drops and the tongue body moves up and forward, which advances the tongue root, to form a good looking /i/ (Figure 25e). The tongue body then moves slightly back (Figure 25f) and down (Figure 25g) toward the /a/ position and by Figure 25h, the tongue body and root have returned to the /a/ configuration and the lower lip has also risen to its initial position. /iyi/ (Figure 26) shows that, in order to maintain contrast with

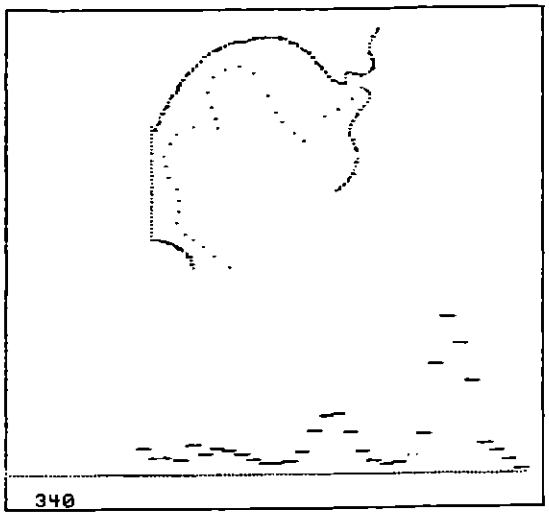



a b

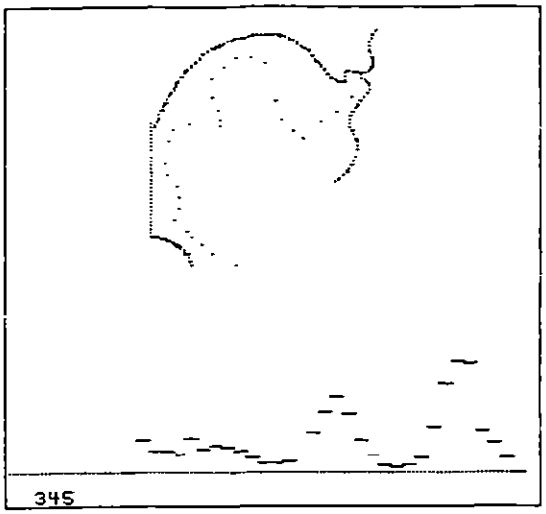
c d

/ara/

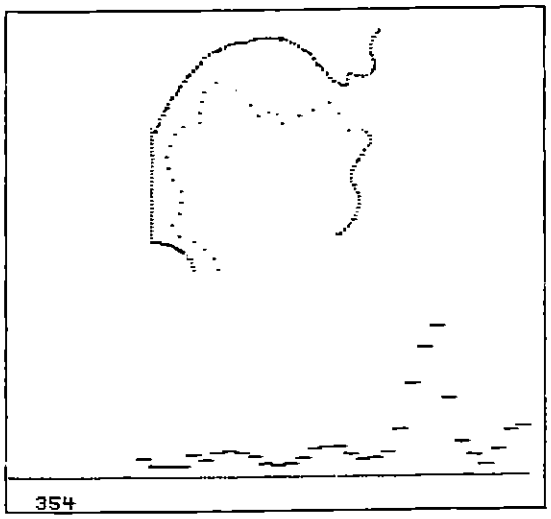
Figure 23.




e



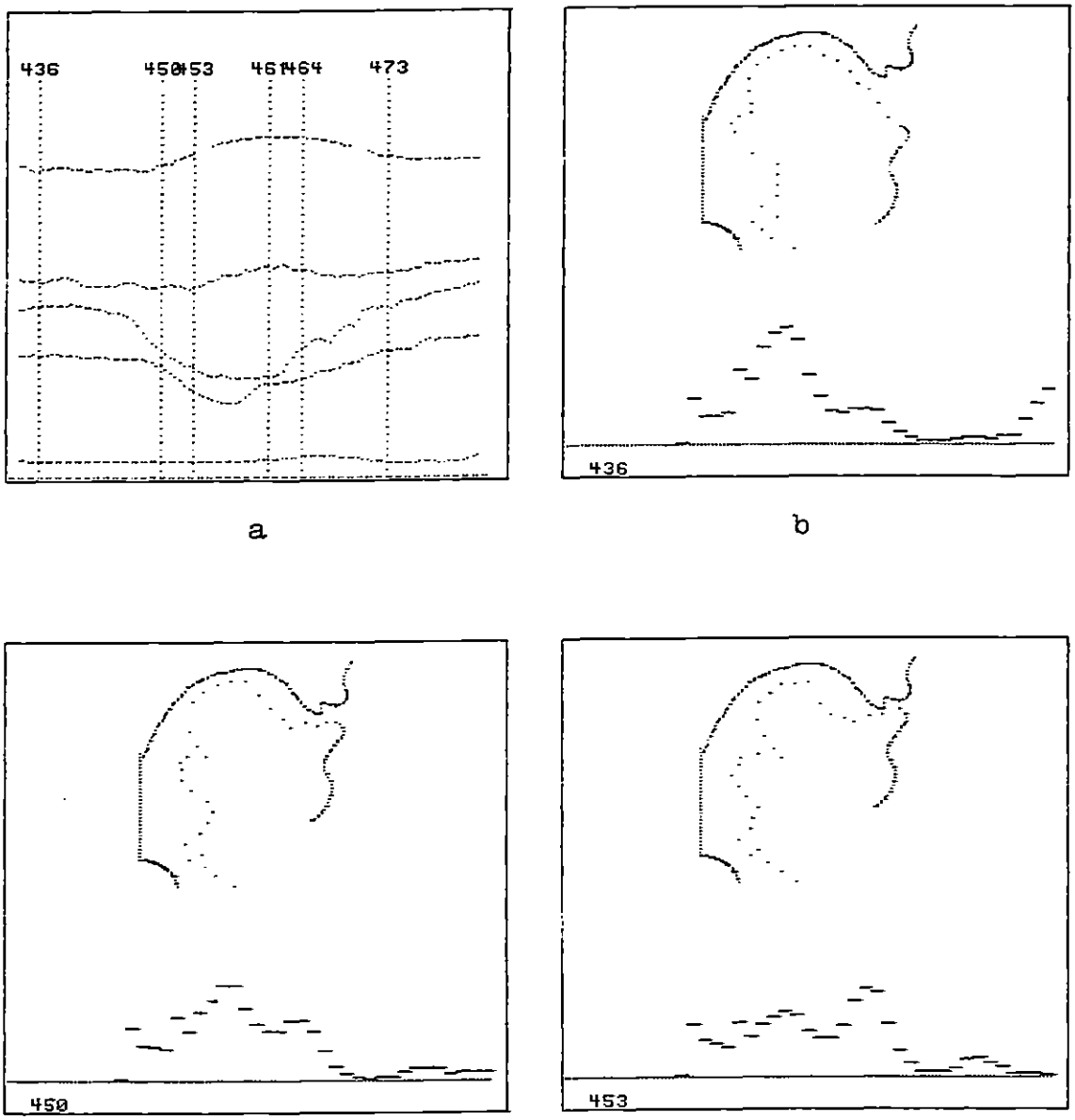

f




g

/ara/

Figure 23 cont.

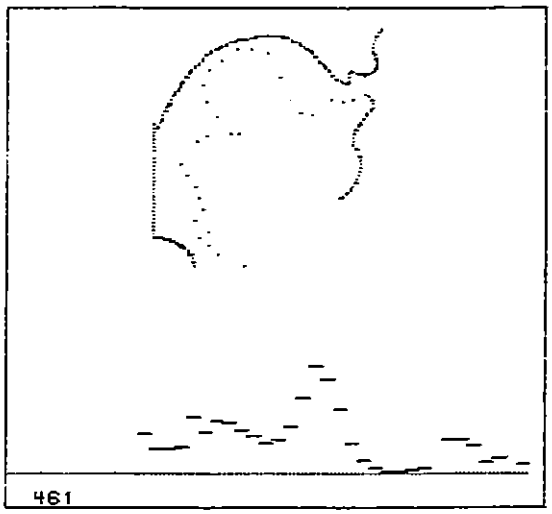



a b

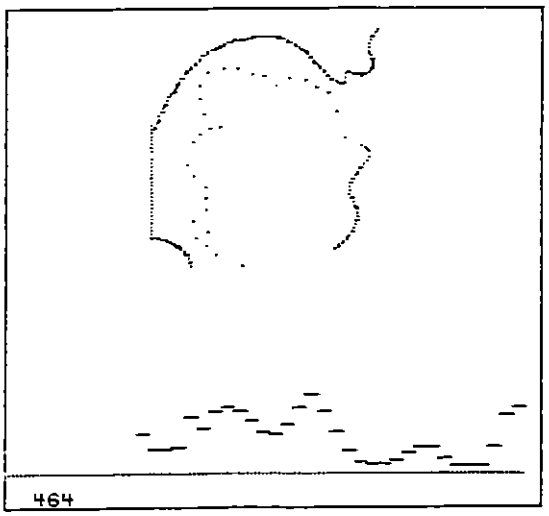
c d

/iri/

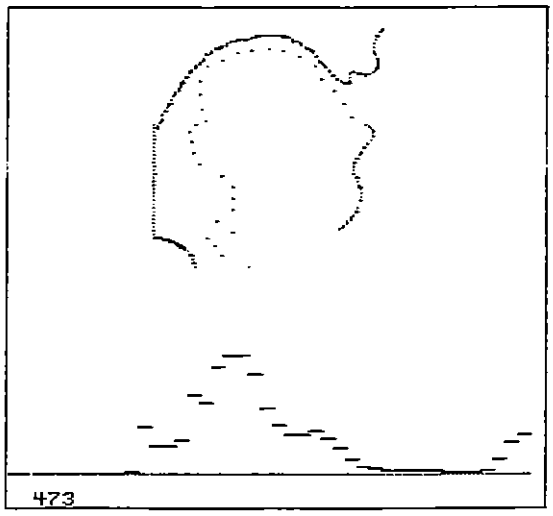
Figure 24.




e



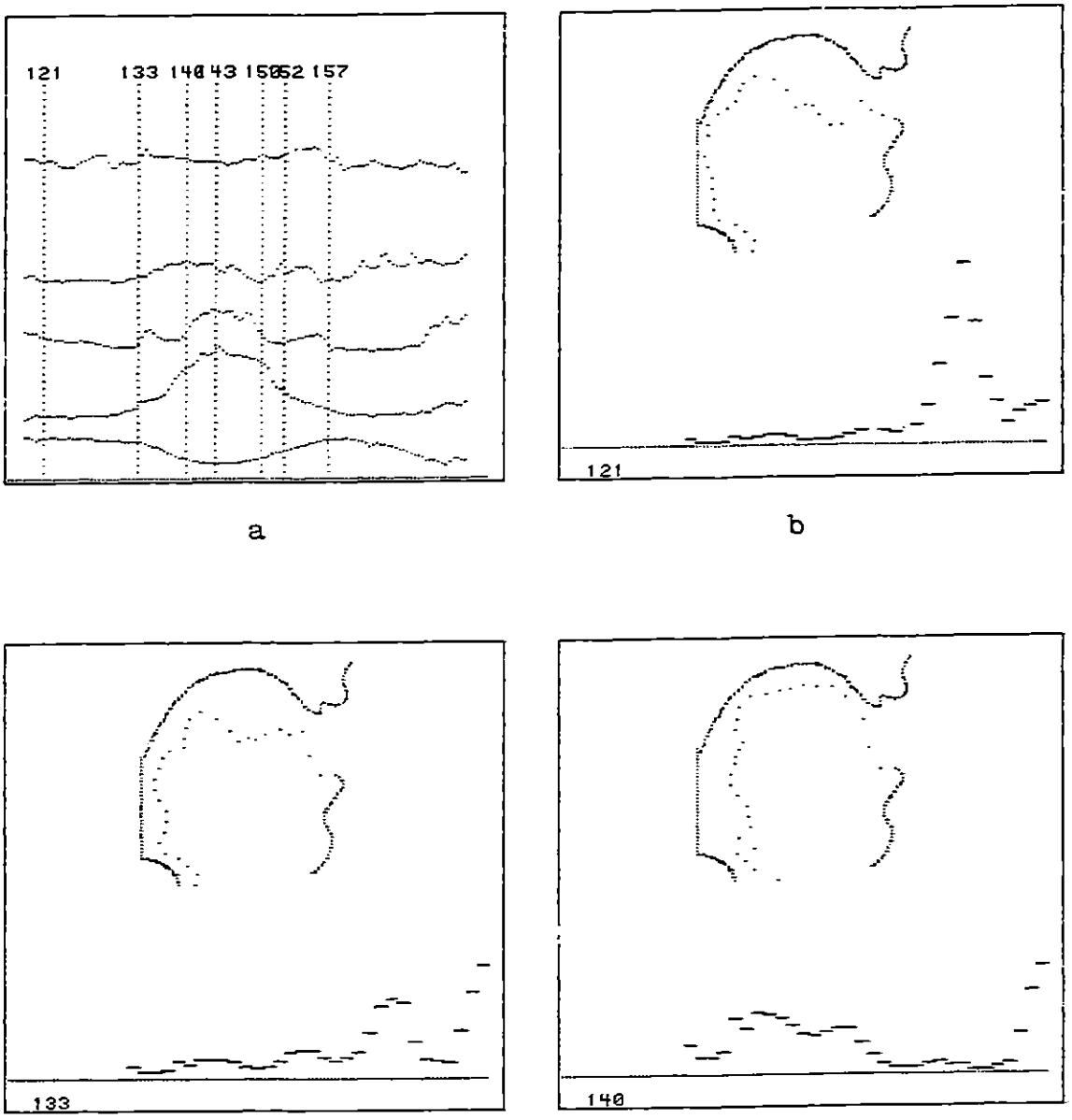

f




g

/iri/

Figure 24 cont.

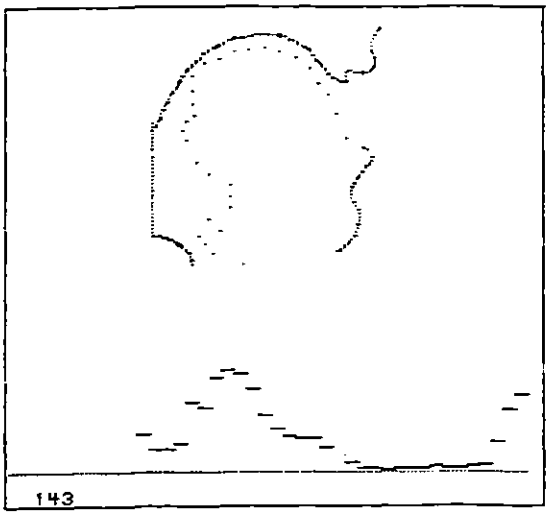



a b

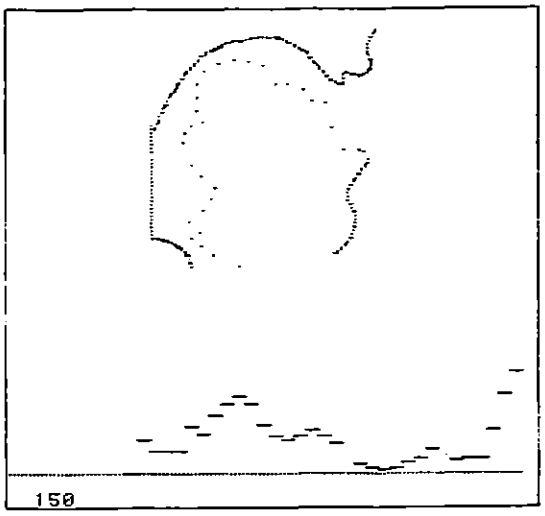
c d

*/aya/*

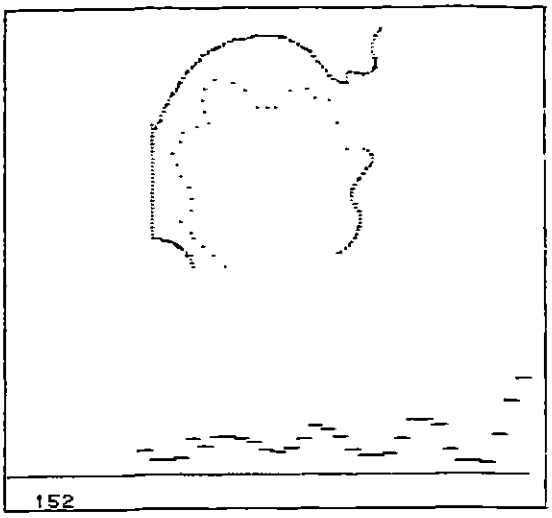
Figure 25.



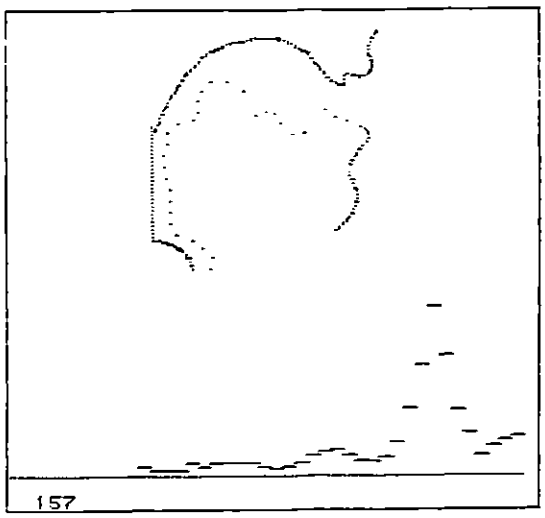

e




f



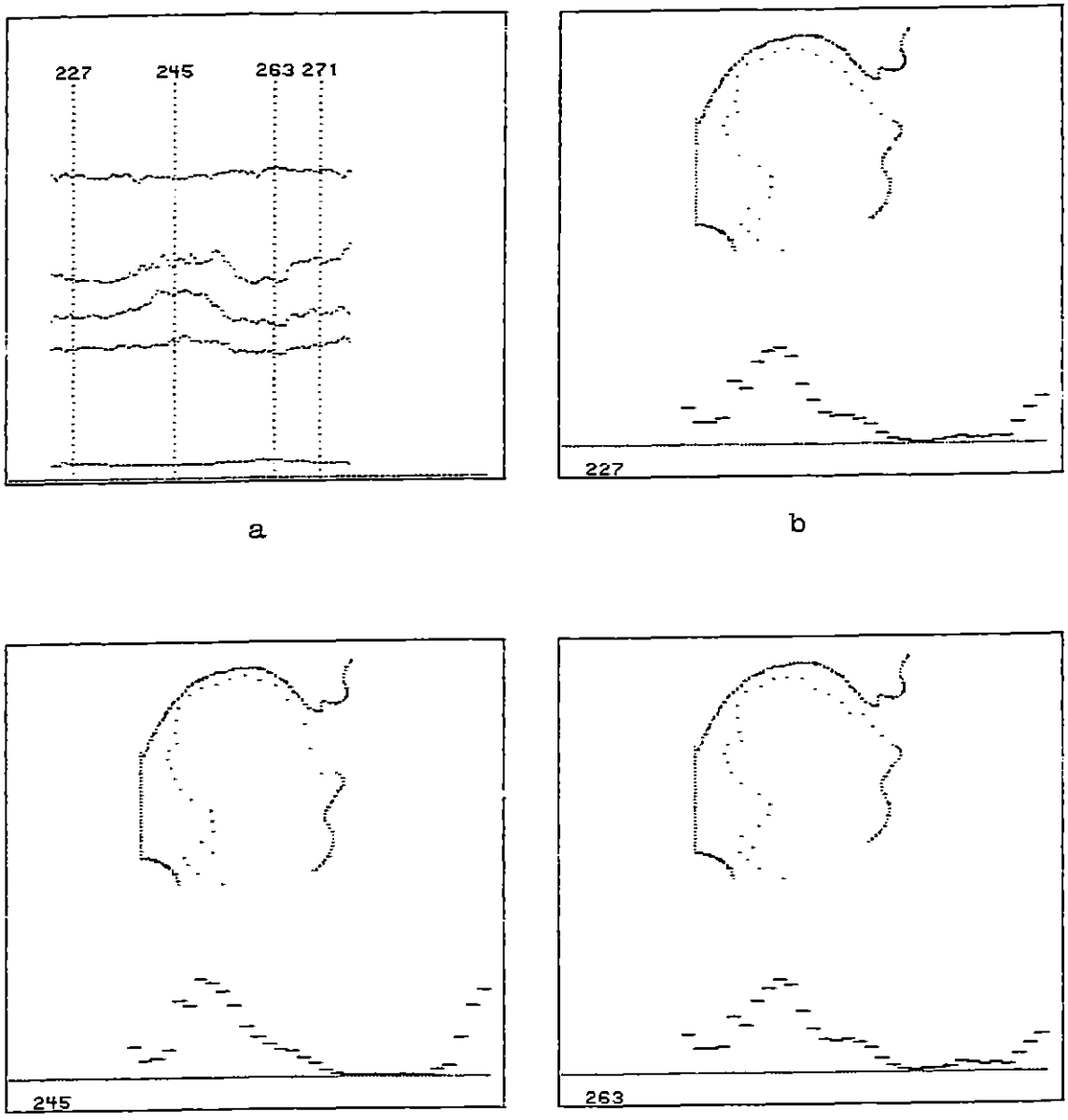

g




h

/aya/

Figure 25 cont.

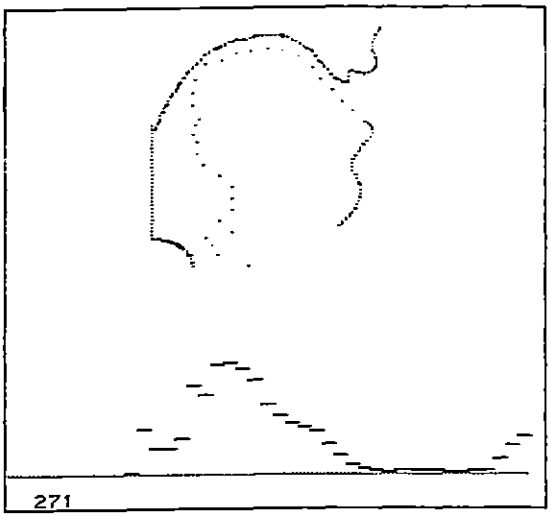



a b

c d

/iyi/

Figure 26.




e

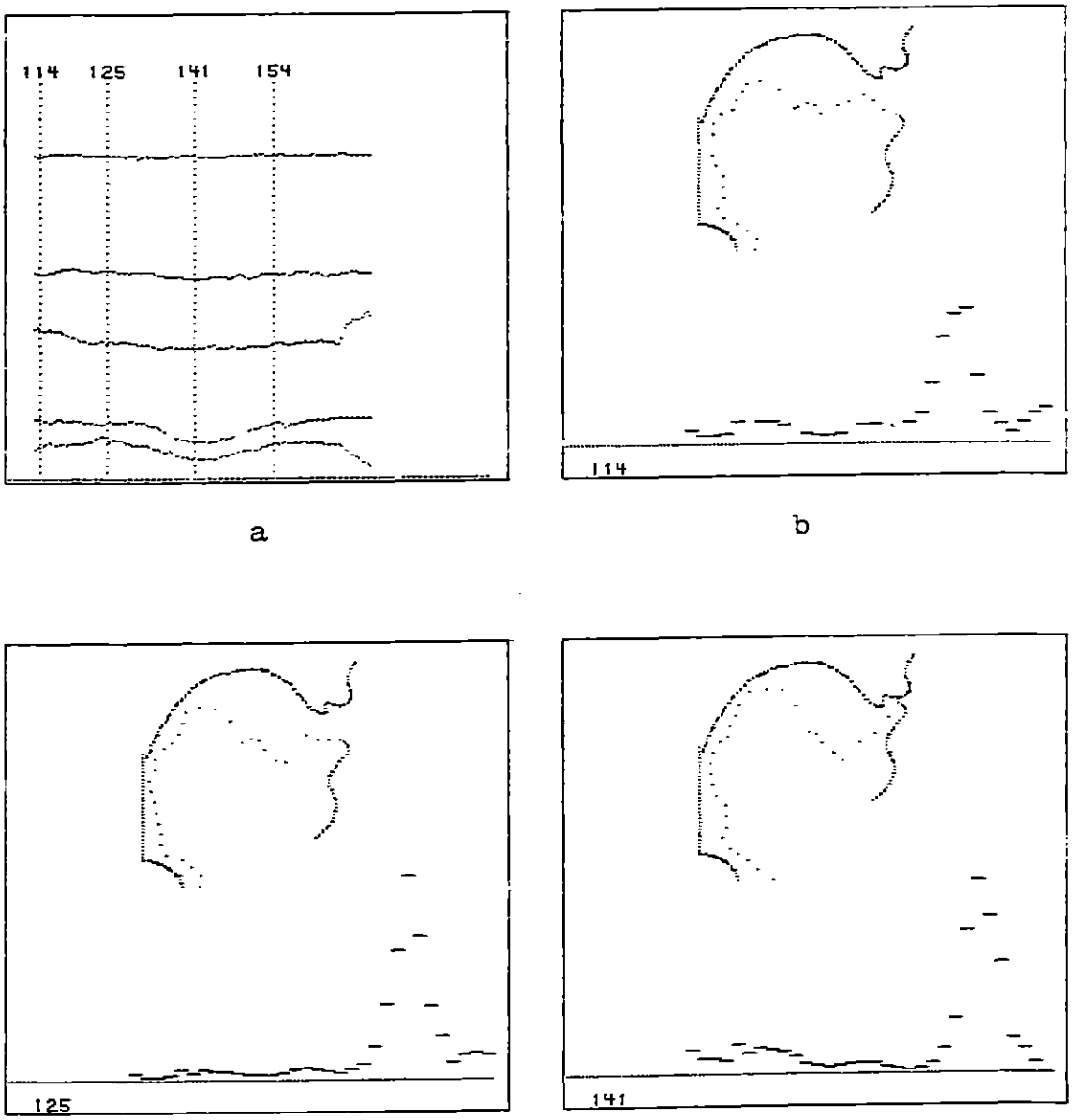
/iyi/

Figure 26 cont.

the vowels, the /y/ causes a deviation from the /i/ configuration. The utterance begins and ends with the usual area function for /i/. The two intervening figures show a slight drop of the lower lip (Figure 26c) followed by a slight drop of the tongue tip (Figure 26d) before returning to the vowel. In each case, the analysis is consistent with the predicted articulation.

/w/ can similarly be approximated as the transition between /u/ and the following vowel. /awa/ (Figure 27) indeed does this as the vocal tract moves from /a/ (Figure 27b) to a configuration similar to /u/ (Figure 27d) and back to /a/ (Figure 27e). The /u/ configuration, however, shows the coarticulatory effects of the adjacent vowels in its small pharynx volume. /iwi/ (Figure 28) starts out in the usual /i/ configuration (Figure 28b), the lips close (Figures 28c and 28d), the tongue body moves back to form an /u/ configuration (Figure 28e), and the process reverses to form the final /i/ in Figure 28g. The results again are consistent with the predictions.

Finally, a set of utterances with the affricate /ʃ/ are analyzed: /aʃa/ in Figure 29, /iʃi/ in Figure 30, and /uʃu/ in Figure 31. Generally all of the utterances follow a similar pattern. Each starts with the properly formed vowel, forms a closure just behind the teeth, while, if necessary, raising (but not closing) the lower lip, and

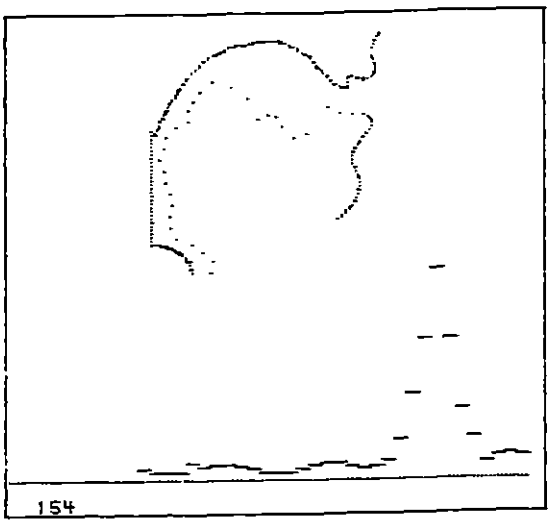



a b

c d

/awa/

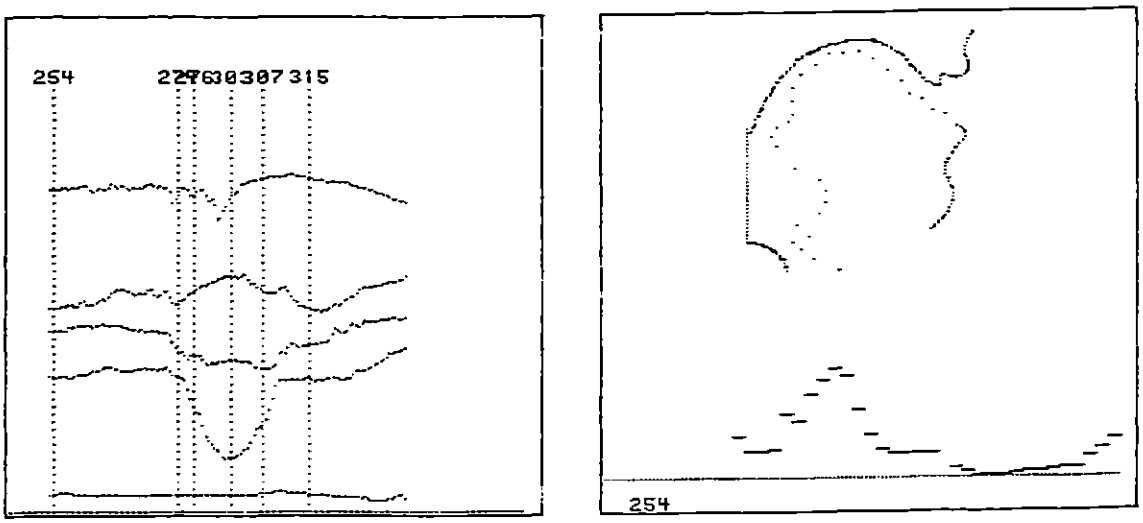
Figure 27.



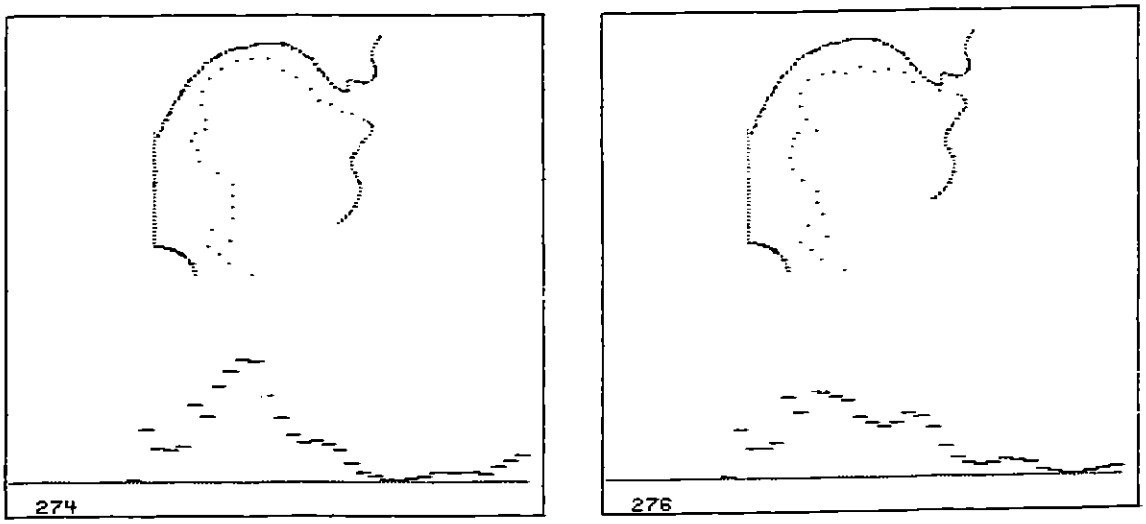

e

/awa/

Figure 27 cont.



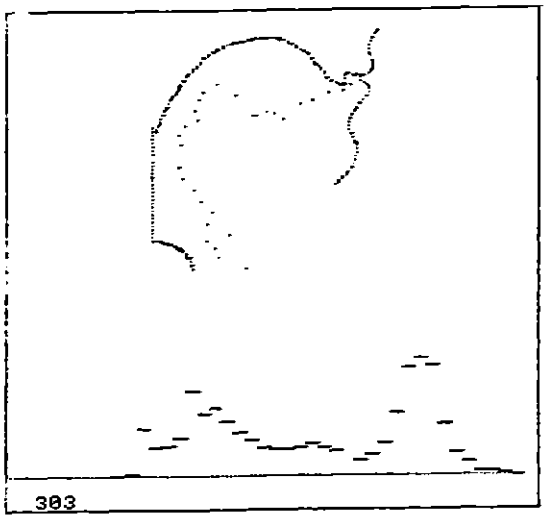

a b



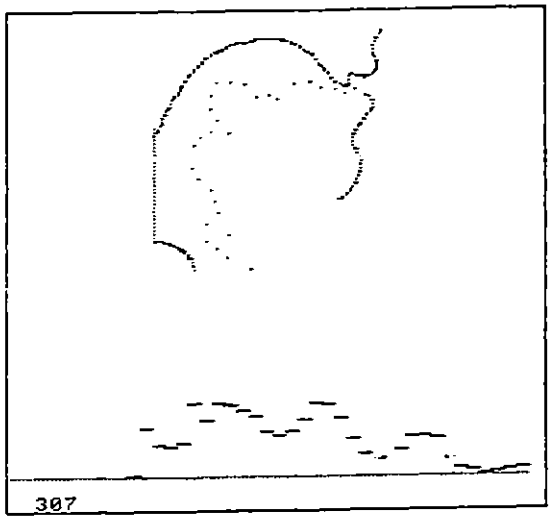

c d

/iwi/

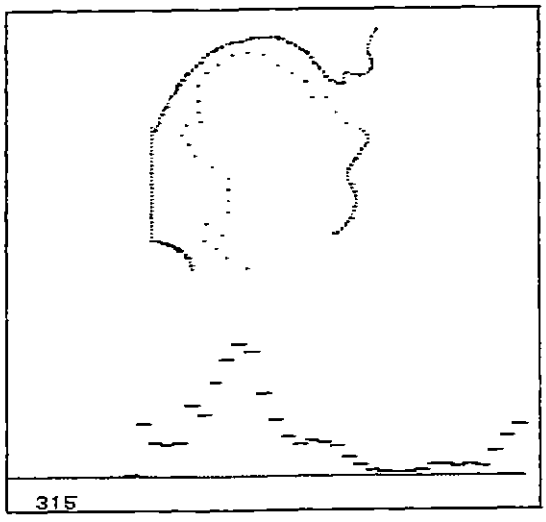
Figure 28.




e



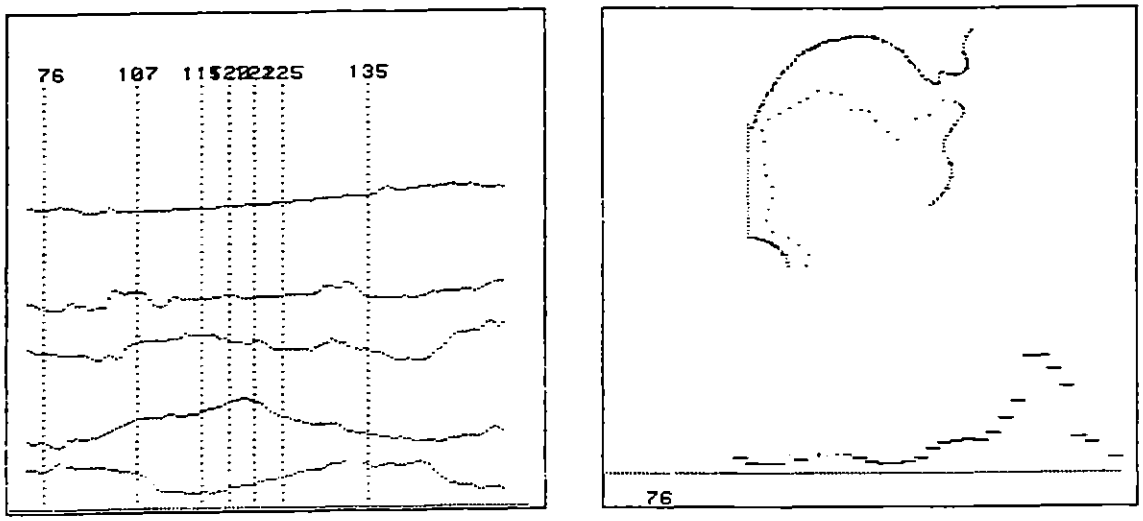

f



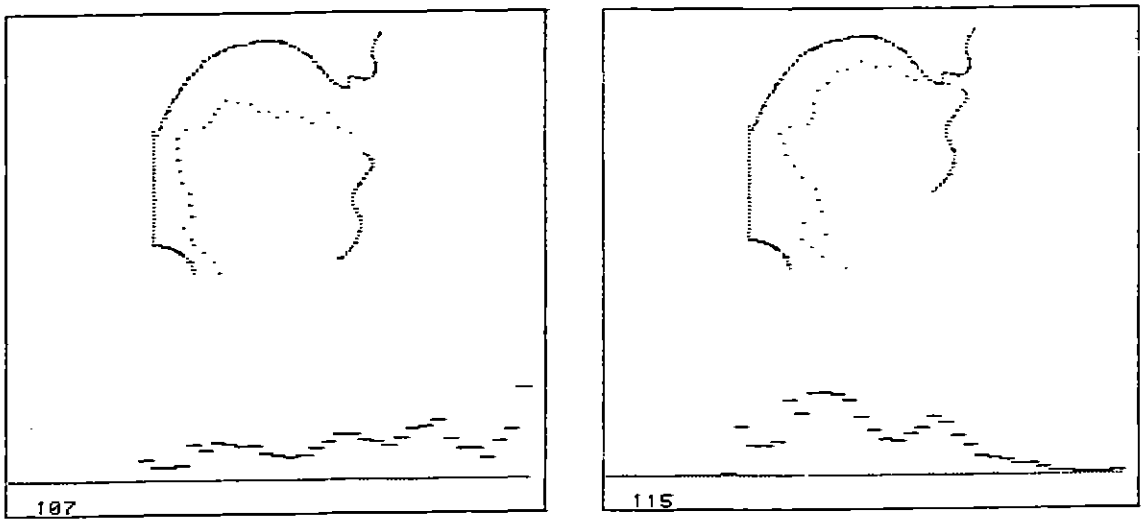

g

/iwi/

Figure 28 cont.



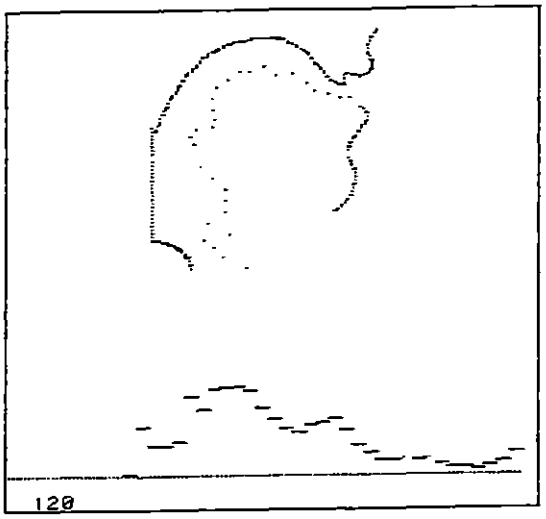

a b



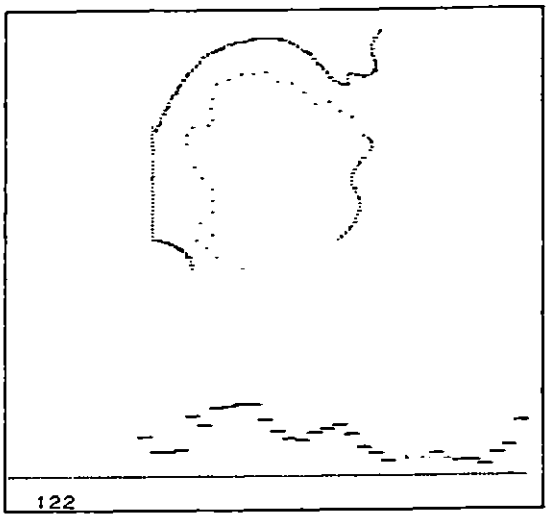

c d

/aǰa/

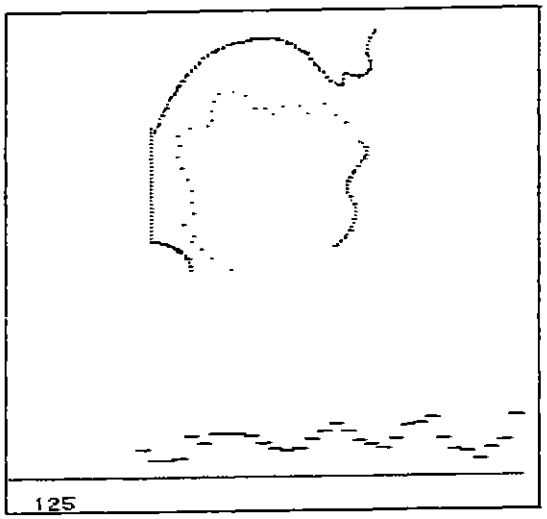
Figure 29.



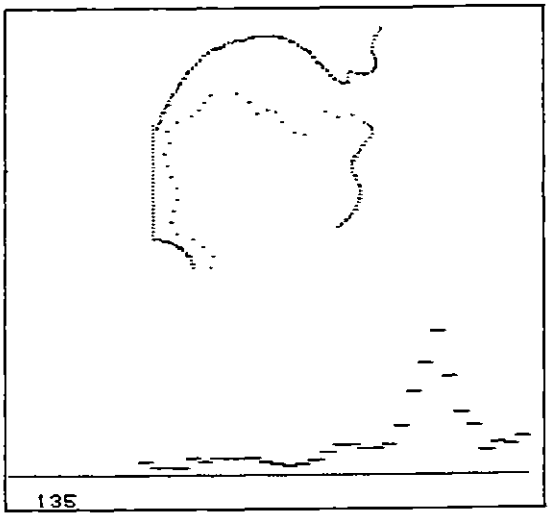

e




f



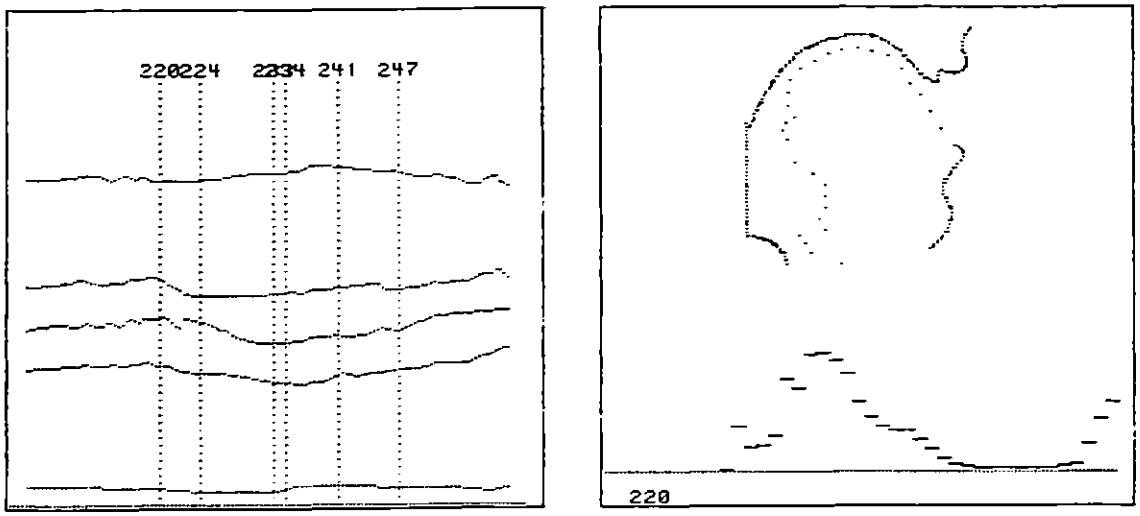

g




h

/aǰa/

Figure 29 cont.

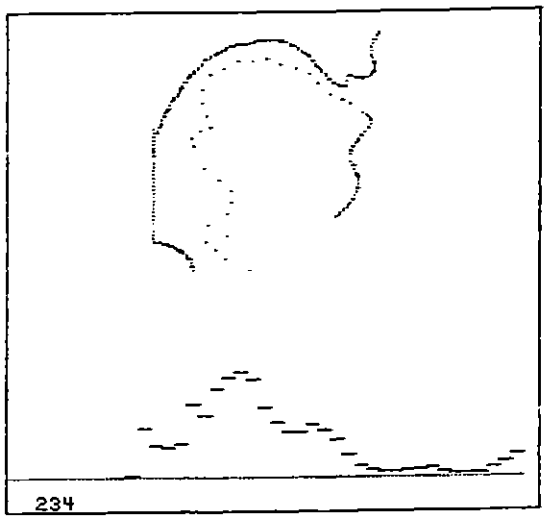



a b

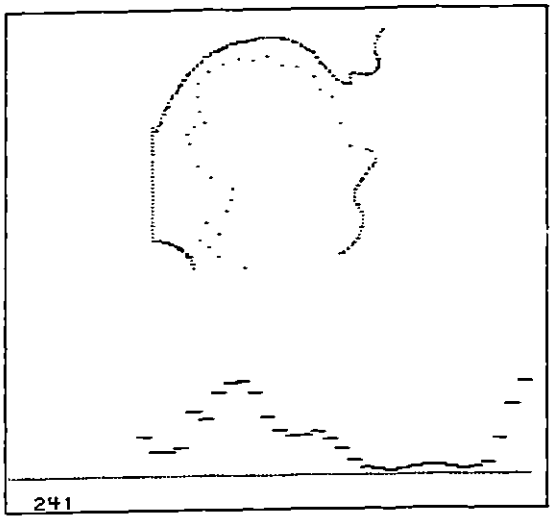
c d

/iǰi/

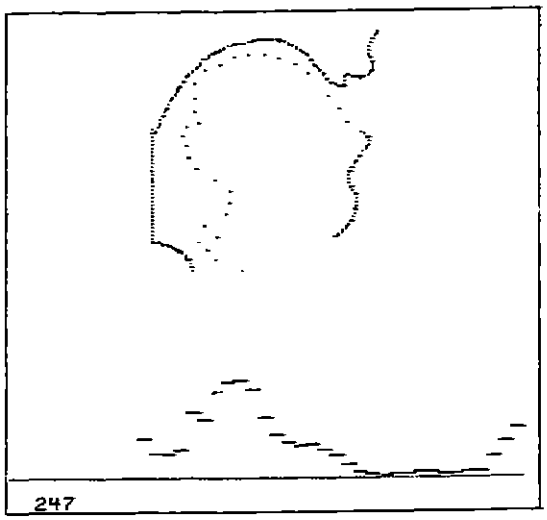
Figure 30.




e



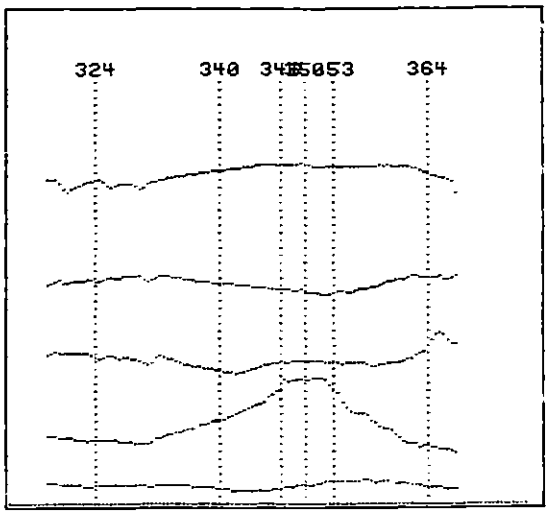

f



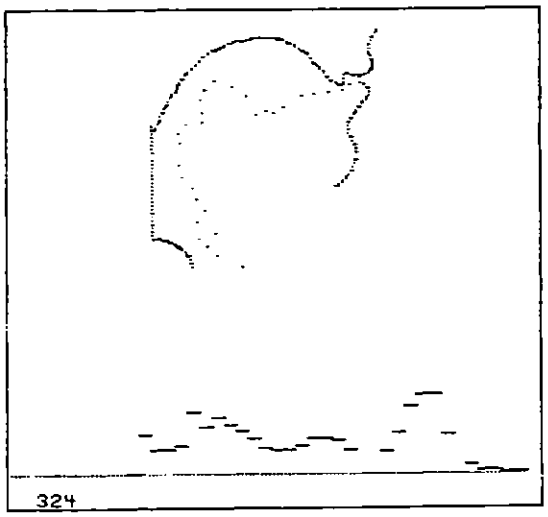

g

/iǰi/

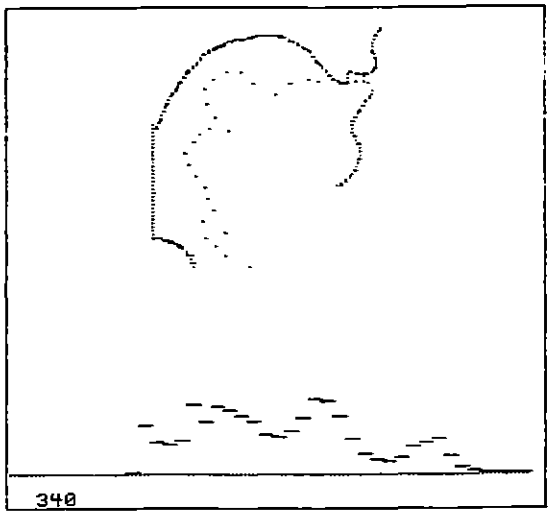
Figure 30 cont.



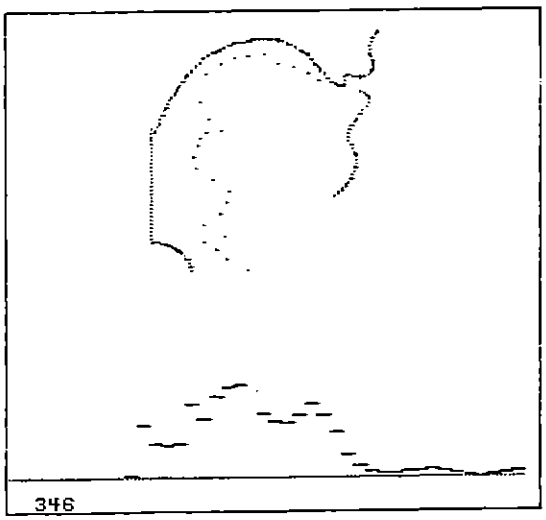

a




b



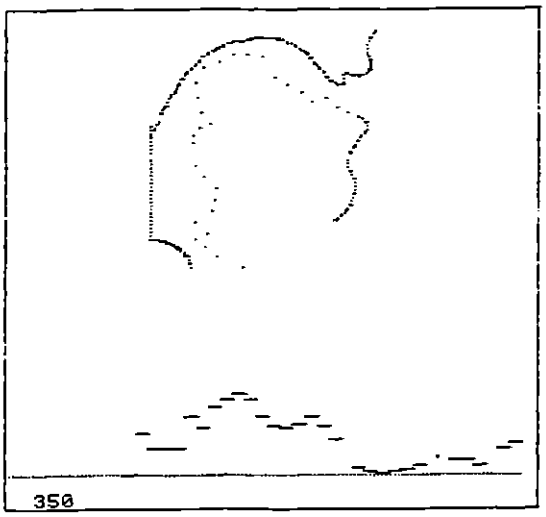

c



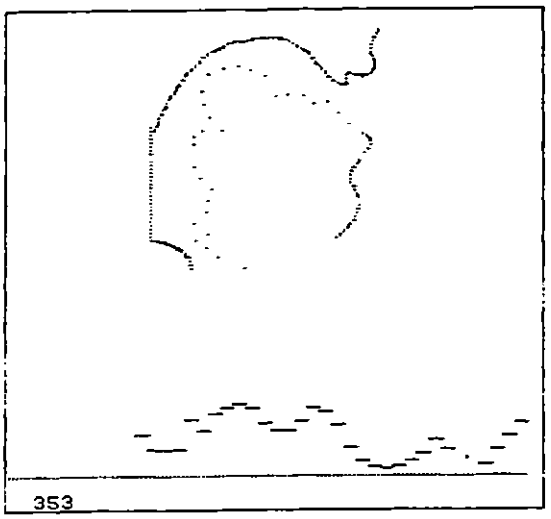

d

/uǰu/

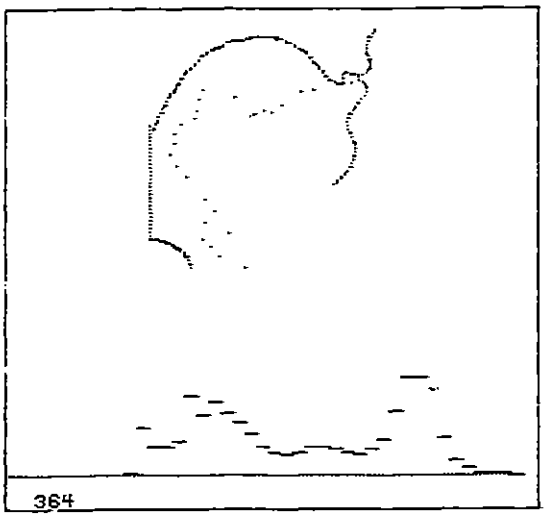
Figure 31.




e




f




g

/uǰu/

Figure 31 cont.

finally returns to the vowel shape. The tongue body position at the time of closure is influenced by coarticulation from the vowel contexts. The only departure from the pattern is in Figure 31c (/iǰi/), where the impending point of closure is indefinite.

Several utterances other than those shown here have been tried, but are not shown due to the difficulty of formant tracking or an insufficient quantity of utterances. An attempt was made to analyze some Arabic pharyngeal stops, but all showed closures at the lips rather than at the pharynx. A few utterances from another speaker were also tried and no systematic differences were noted. The utterance set, however, was too small to be conclusive.

## VII. Comparative performance

A number of algorithms for calculation of the area function from the acoustic properties of the tract have been devised. Each has its specific set of assumptions and requirements. (See Chapter III for a description of the methods.) As these sets vary and universal data sets containing the simultaneously taken acoustic data for all the algorithms along with lateral X-ray photographs to provide a standard do not exist, rigorous comparison of the results of the algorithms is impossible. The best comparison that can be achieved using published results is to choose a few relatively invariant phonemes which all algorithms can analyze (and were published) and compare the results.

All of the methods reviewed in Chapter III are capable of analyzing non-nasal sonorants. Thus the set of phonemes must be chosen from this class of utterances. The non-nasal sonorants include all of the vowels and some consonants. Consonant production tends to be heavily influenced by coarticulation from the nearby vowels--such as in /b/ where the tongue body position is set not by the consonant but by the adajacent vowels. Vowels, while not totally free from coarticulation, are much less subject to such effects. /i/, /a/, and /u/ have the advantages of bounding the $F_1$--$F_2$ space for most vowel sets, appearing in most area function

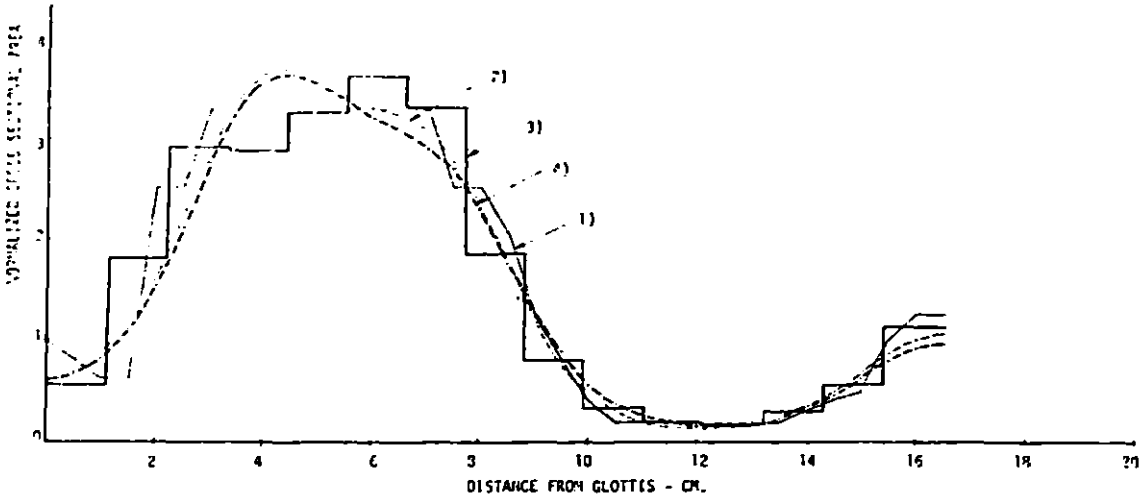
test sets and being in Fant's [4] set of X-ray derived area functions. In addition, the acoustic data used in many of the test sets is derived by using Webster's horn equation to compute the desired acoustic parameters from the X-ray derived area functions themselves without reference to any real speech. (All such acoustic data sets are subject to the errors inherent in the assumptions imposed by Webster's horn equation--rigid walls and no loss within the tract.)

The methods compared have the following inputs:

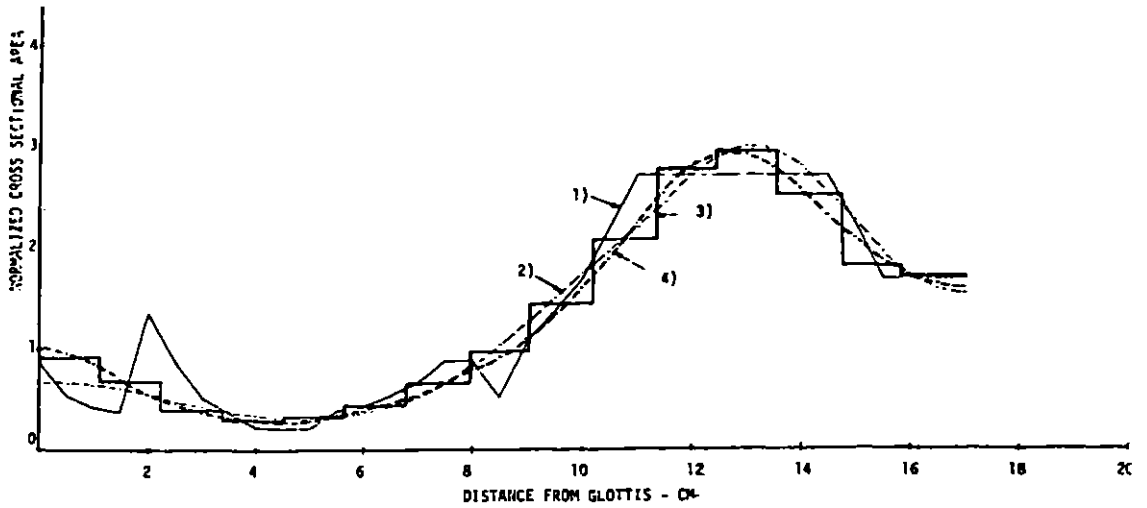
Paige and Zue: 3 poles and 3 zeros of lip admittance computationally derived from X-ray area functions
length from X-ray area function
Adaptive LPC--Nakajima: speech waveform--15 khz sampled
LPC--Wakita: speech waveform--7 khz sampled
Perturbation: 3 poles and 3 zeros of lip admittance computationally derived from X-ray area functions
length from X-ray area function
Perturbation from real speech: 3 formants
Paul: 5 formants acoustically measured from speech with simultaneous X-ray photography
length from X-ray area function
Gopinath and Sondhi: 4 poles and residues of the lip admittance computationally derived from X-ray area functions

Inputs to the compared algorithms
Table 4

Paige and Zue present superimposed graphs (Figure 32) of Fant's X-ray area functions, bandlimited area function, and their results. (The bandlimited area functions are limited to the Fourier coefficients of the log area function corresponding to the first six modes of the lip admittance. This process is analogous to analysis and synthesis by first



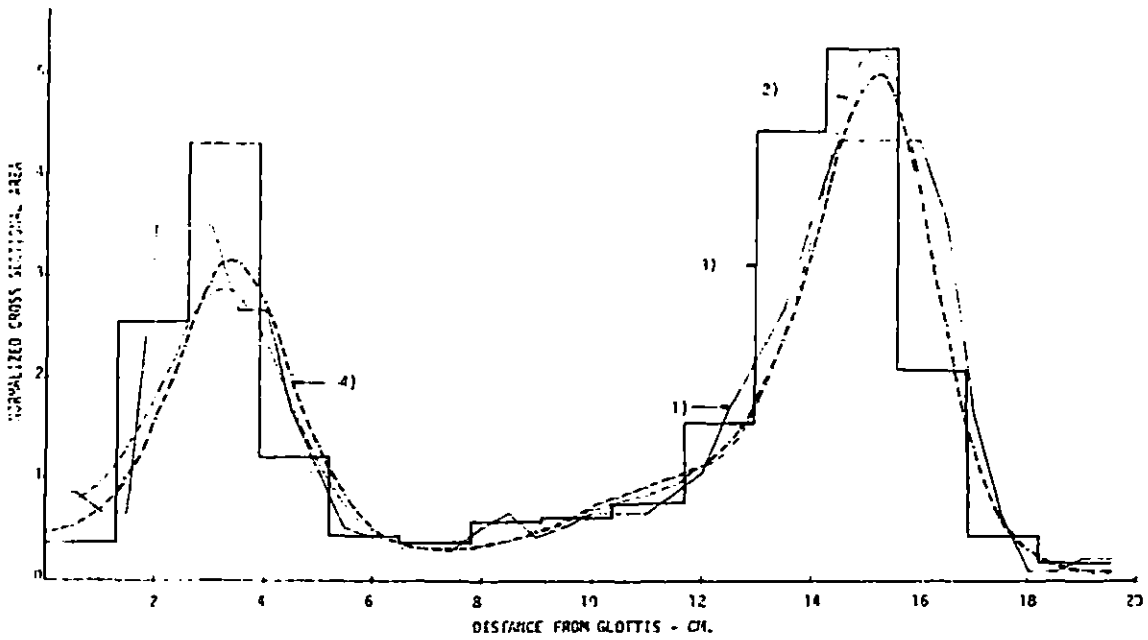

A 15-section approximation to the vocal tract for the vowel /i/: 1) Fant's data, 2) 6-term Fourier expansion of 1), 3) 15-section approximation, 4) 6-term Fourier expansion of 3).




A 15-section approximation to the vocal tract for the vowel /a/: 1) Fant's data, 2) 6-term Fourier expansion of 1), 3) 15-section approximation, 4) 6-term Fourier expansion of 3).

Area functions due to Paige and Zue [19]

Figure 32.



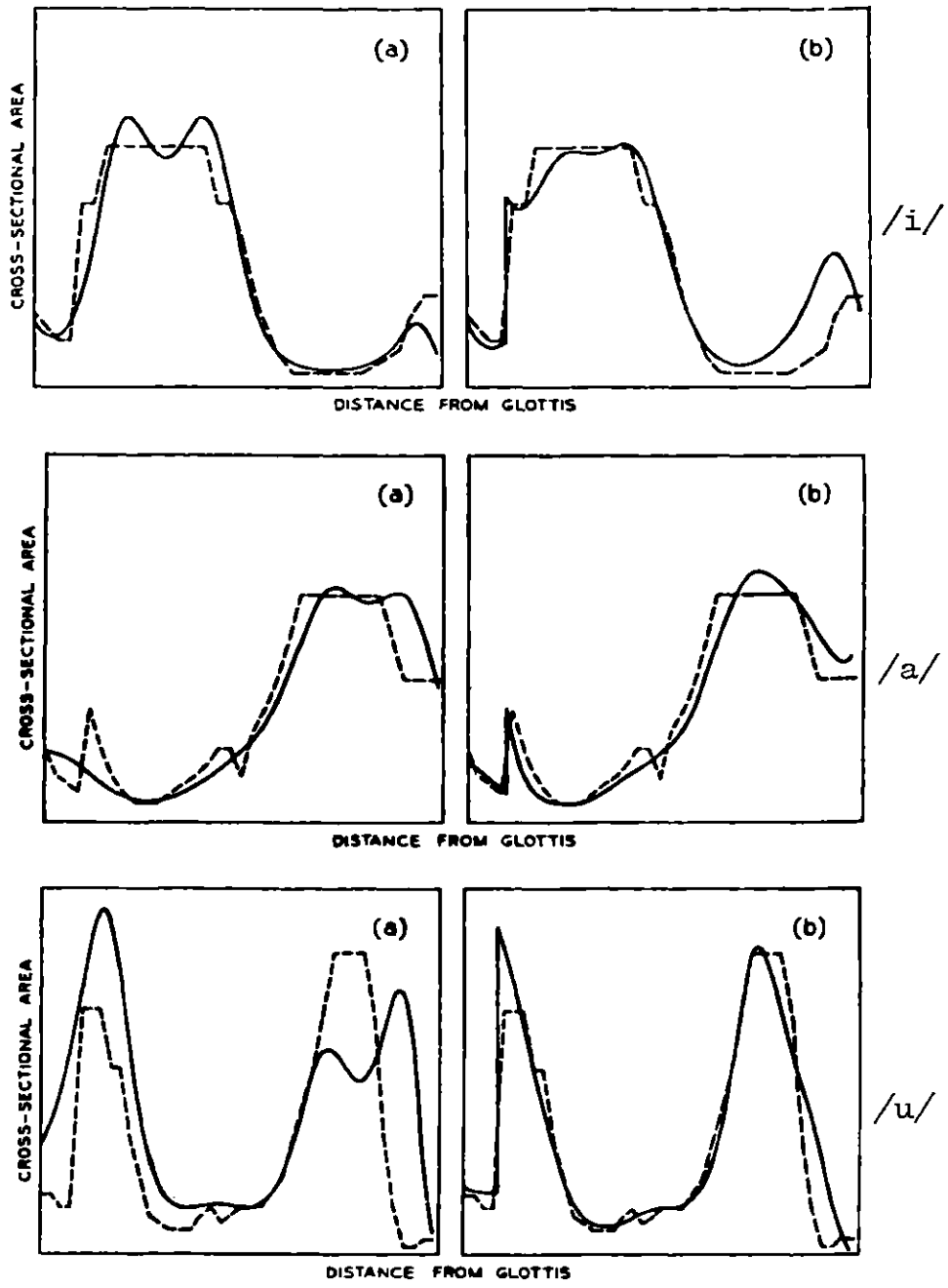

A 15-section approximation to the vocal tract for the vowel /u/: 1) Fant's data, 2) 6-term Fourier expansion of 1), 3) 15-section approximation, 4) 6-term Fourier expansion of 3).

Figure 32 cont.

order perturbation.) The bandlimited area functions approximate the original area functions fairly well with the exception of local details--thus indicating that six modes can provide a good general description of the area function. The area functions generated by their area function generation algorithm for these six modes are all reasonably accurate.

The method of Gopinath and Sondhi also shows fairly good results, which improve if the canonical tube is chosen to contain the laryngical-pharyngeal discontinuity (Figure 33). The perturbation method using all six modes, however, shows a poorly estimated pharyngeal region for /i/. Otherwise, its results appear to be fairly accurate (Figure 34).

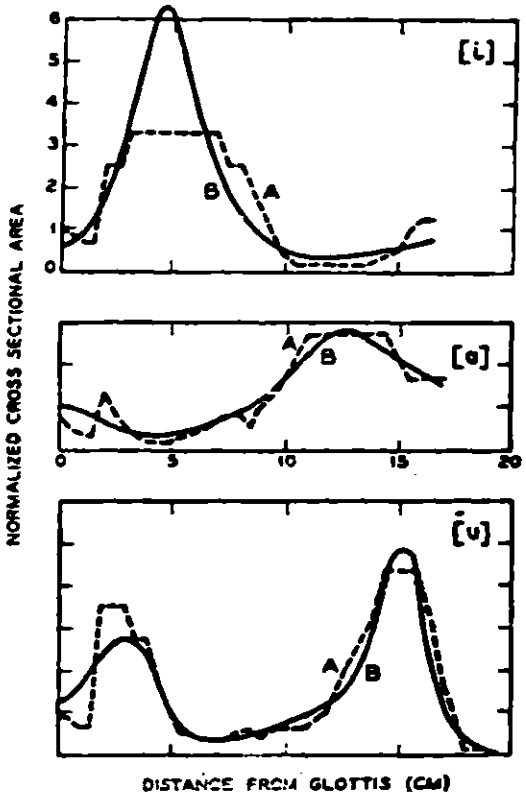
All of the above tests of methods are unrealistic. All use acoustic data which has been computed from given (X-ray) area functions and all use data (zeros of the lip admittance or residues of the driving point impedance) which are impossible to measure without complex and intrusive apparatus. Toward this end, Schroeder and Mermelstein reconstructed X-ray area functions using only the parameters--the poles of the lip admittance--which could be measured from the speech signal (Figure 35). Again, however, these "acoustic" parameters were computed from the X-ray area function.




Area functions reconstructed from the first *four* poles and residues: (a) the reconstruction using a uniform canonical tube, and (b) the reconstruction with a discontinuous canonical tube as in Section V. Dashed curves are X-ray measurements.

Area functions due to Gopinath and Sondhi [7]

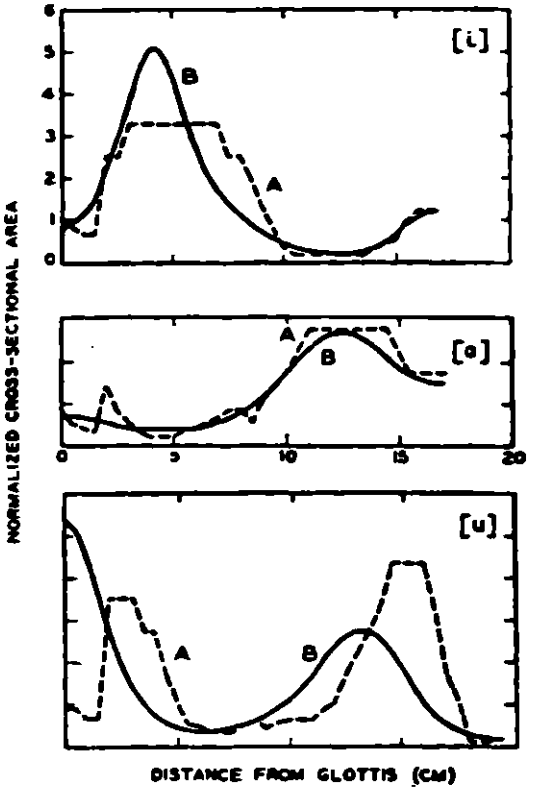
Figure 33.




Vocal-tract area functions band limited to six components determined from the first six admittance poles/zeros of the x-ray-derived area functions of six Russian vowels. A: X-ray derived area function (after Fant). B: Computed band-limited approximation.

Area functions by perturbation with six modes due to Mermelstein [16]

Figure 34.




Area functions antisymmetric in their logarithms approximating x-ray-derived area functions for six Russian vowels and matching their first three formant frequencies. A: X-ray derived area function (after Fant.) B: Computed antisymmetric approximation.

Area functions by perturbation with formants only due to Mermelstein [16] and Schroeder [24]

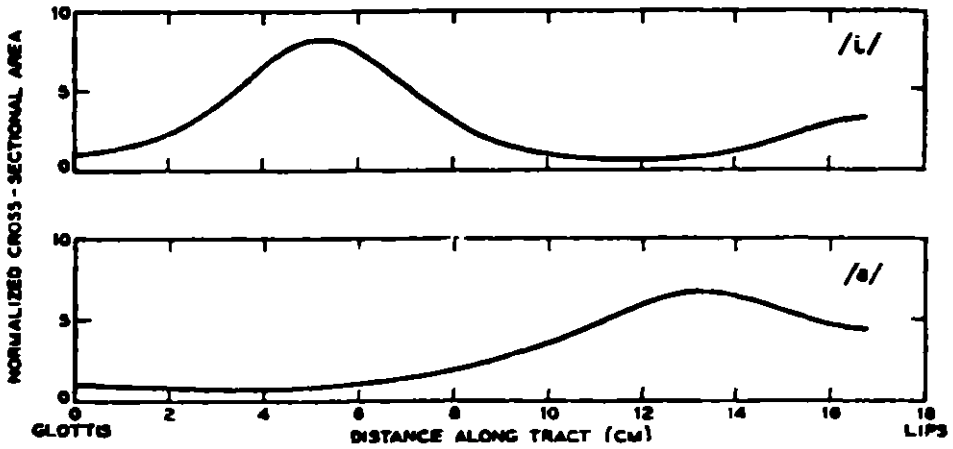
Figure 35.

This set of constraints limits the generated area functions to antisymmetric Fourier components only. The results degrade only slightly for /a/, more for /i/ (which was signifigantly degraded with the six modes), and considerably for /u/. Clearly, antisymmetric Fourier component area functions are inadequate for good area function estimation.

To be useful, an algorithm using acoustic data as input must be evaluated on real acoustic data--not data calculated from the correct answer using idealized models. Of the methods compared, only the antisymmetric Fourier component perturbation, LPC, analysis-by-synthesis (Hafer) and Paul (postulated in this paper) have published results of tests on acoustically measured data.

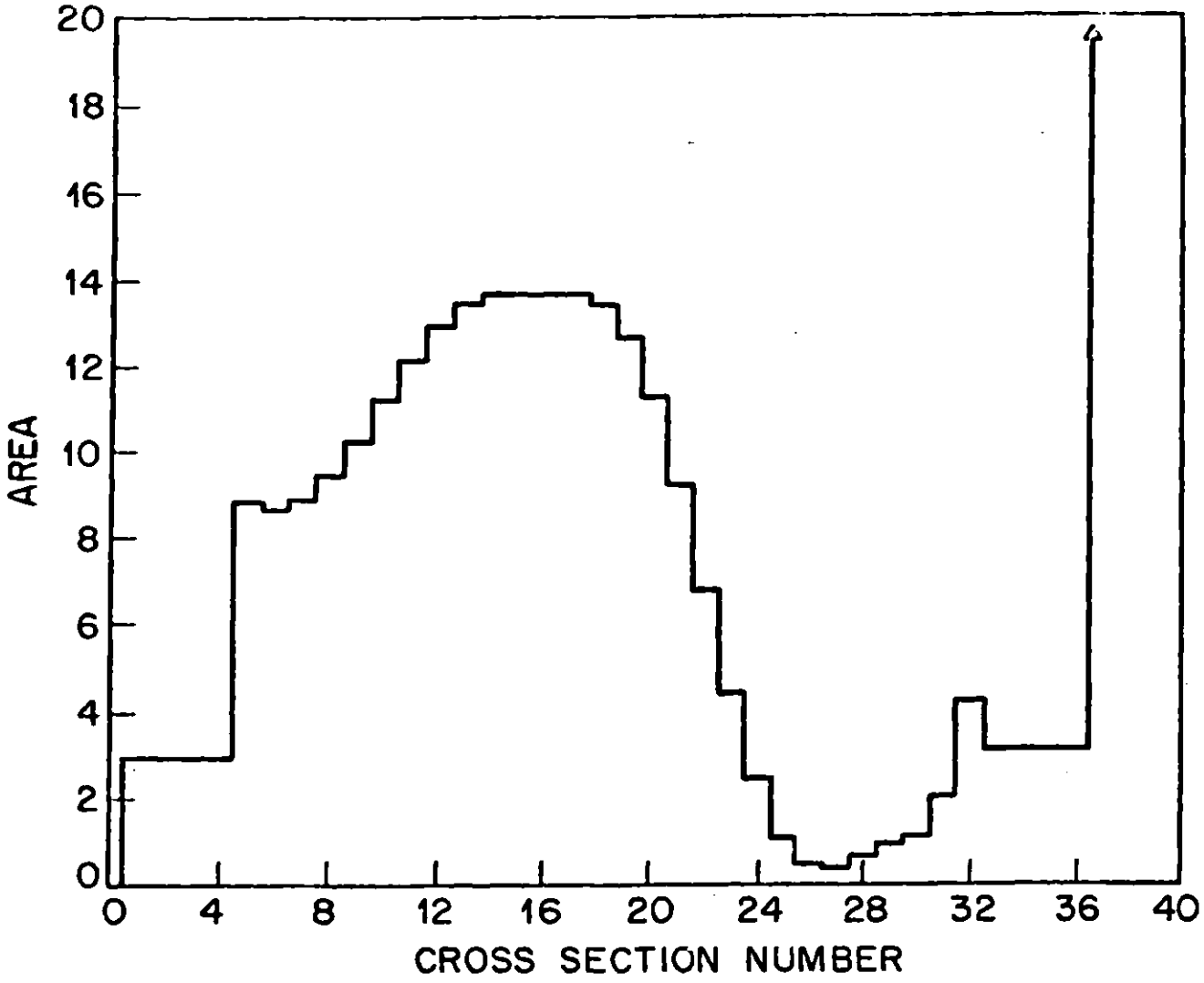
For antisymmetric Fourier component perturbation, only the area functions for /i/ and /a/ for the utterance /iba/ are available (Figure 36). These results appear to be similar to those of the same method as applied to artificially generated acoustic data--some error on /a/ and more error on /i/ with poor resolution on both.

Analysis-by-synthesis (Hafer) removes resolution problems, but in very incomplete tests--only the tongue body was allowed to move--on the utterance /agi/, the area function for /i/ is estimated fairly well (Figure 37). /a/,




Area functions from real speech by perturbation using three formants due to Mermelstein [16] and Schroeder [24]

Figure 36.

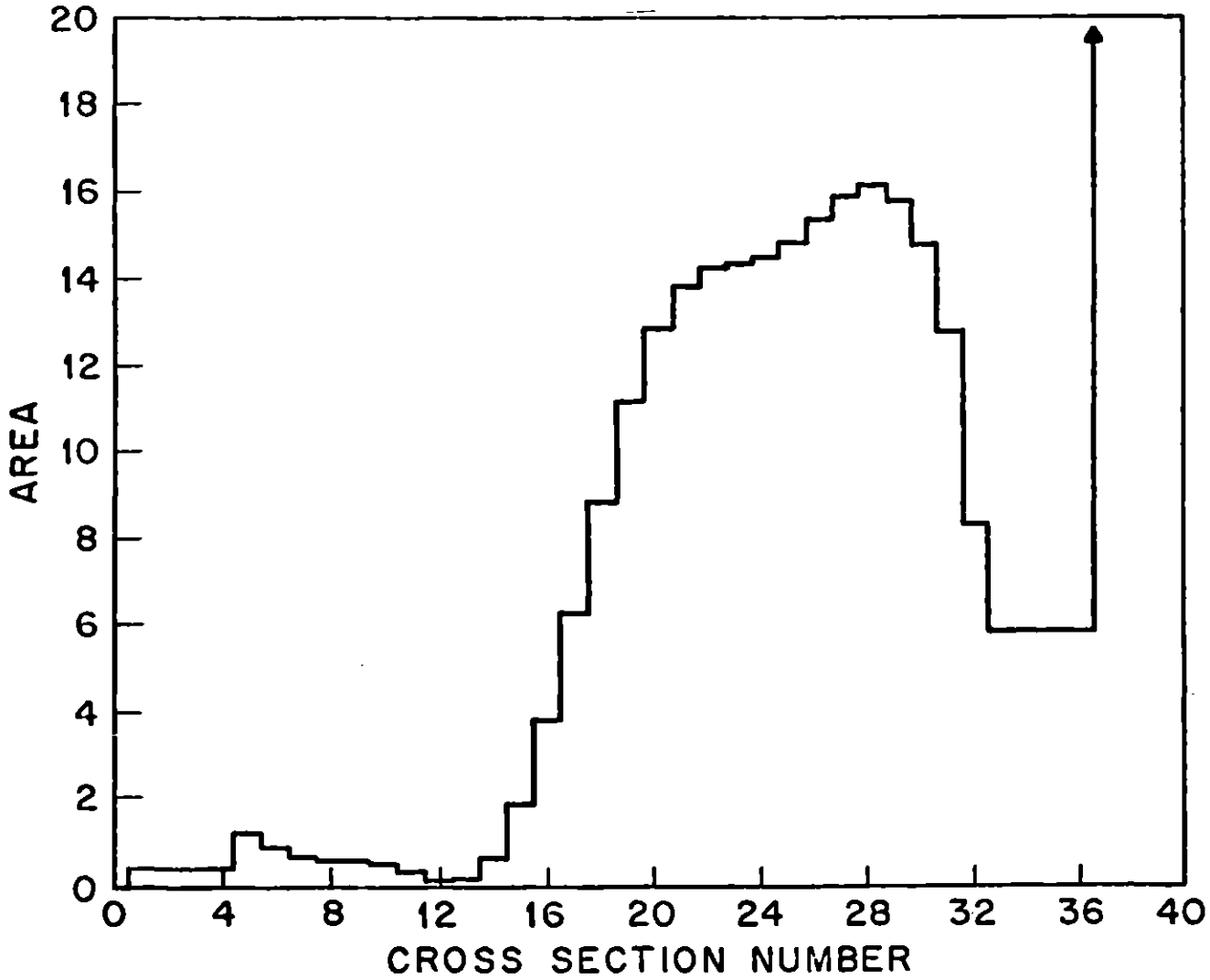



/i/

Area functions by analysis-by synthesis

due to Hafer [8]

Figure 37.



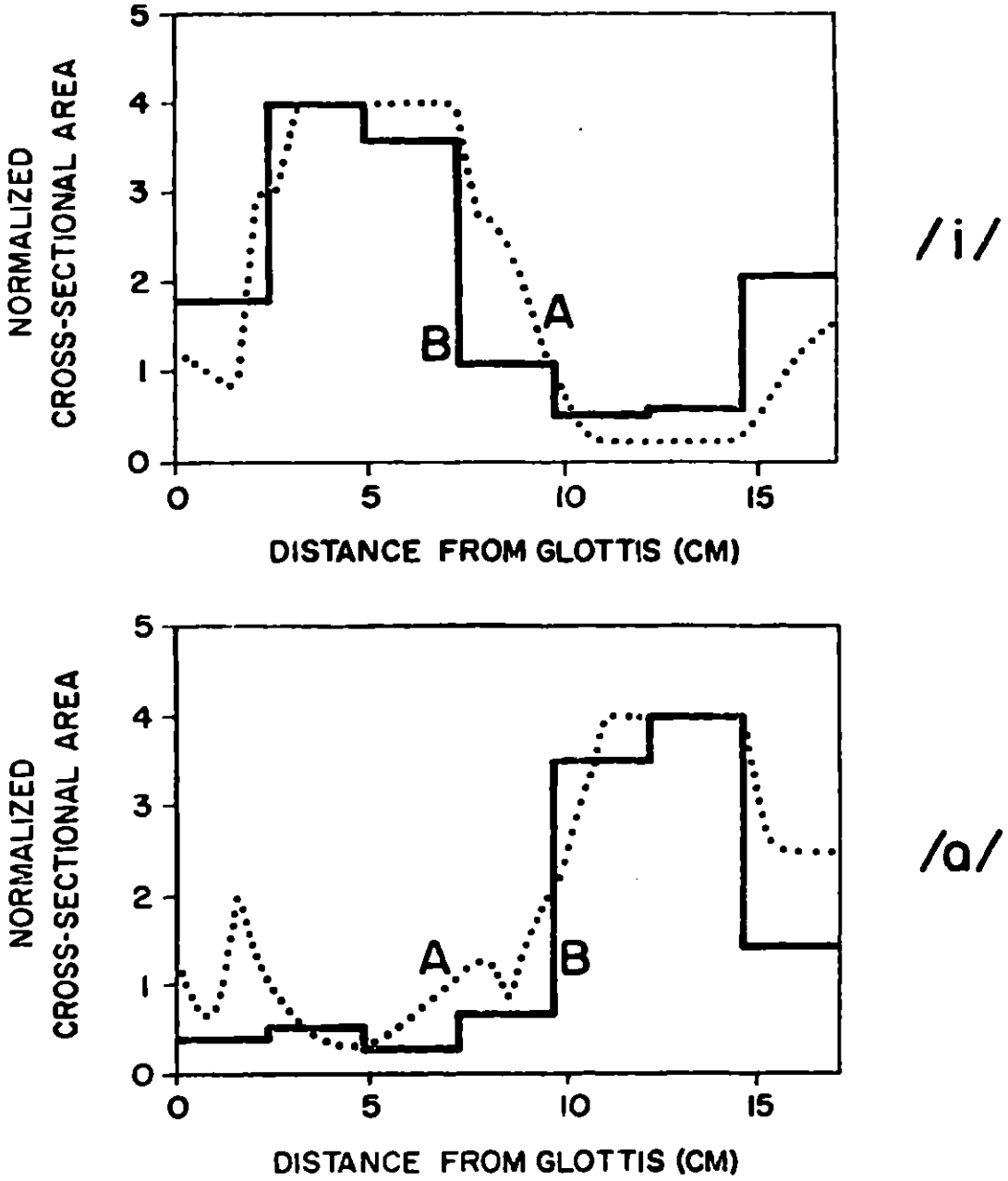

/a/

Figure 37 cont.

however, shows much too small a pharyngeal region. The constraint of only tongue body movement reduces this test almost to the point of uselessness for a less limited utterance set.

The LPC techniques suffer from problems which are unique to these methods. The Wakita scheme (Figure 38) suffers from a very coarse output format and, for /a/ and /i/, appears to discern the basic shape of the vowels. In /a/, however, the larynx is completely lost. For /u/, the pharyngeal shape is completely lost from the area function. These area functions appear to be more accurate at the lip end than at the laryngeal end.

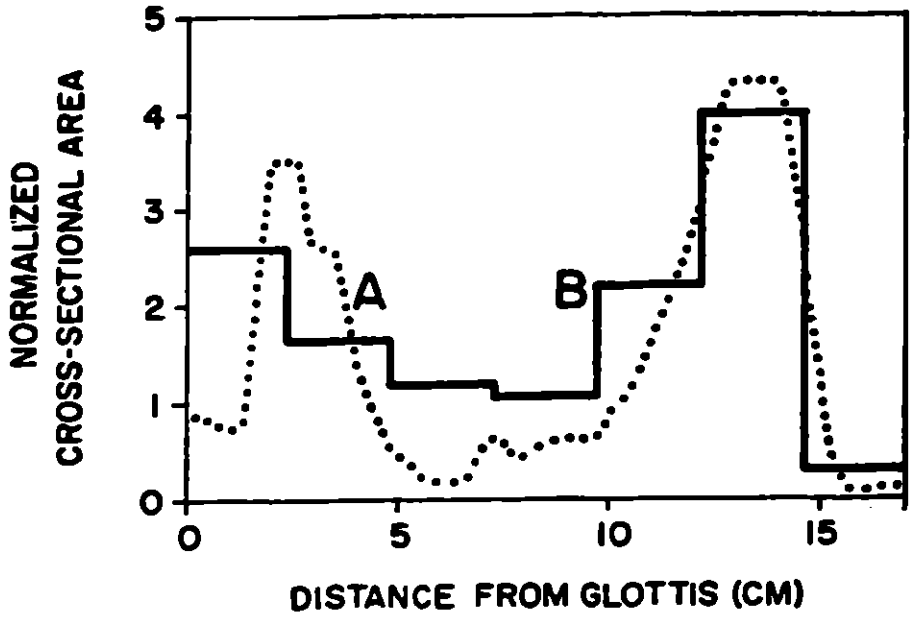
The Nakajima et al. LPC scheme (Figure 39) fares better than the Wakita formulation. The area functions are not as coarse and are somewhat more accurate. The area function for /i/ locates the pharyngeal cavity too far forward and shows the back of the pharynx unrealistically closed. The area function for /a/ does not show the slight constriction at the lips and for /u/ shows the front cavity too small and the back cavity too large. Of the three, only /u/ properly discerns the larynx. These area functions also appear to be more accurate at the lip end than at the laryngeal end.




A-X-ray area funtion due to Fant
B-American vowel

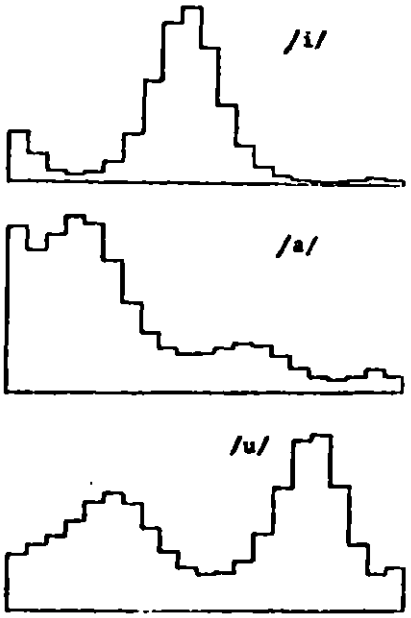
Area functions by LPC due to Wakita [28]

Figure 38.




/u/

Figure 38 cont.

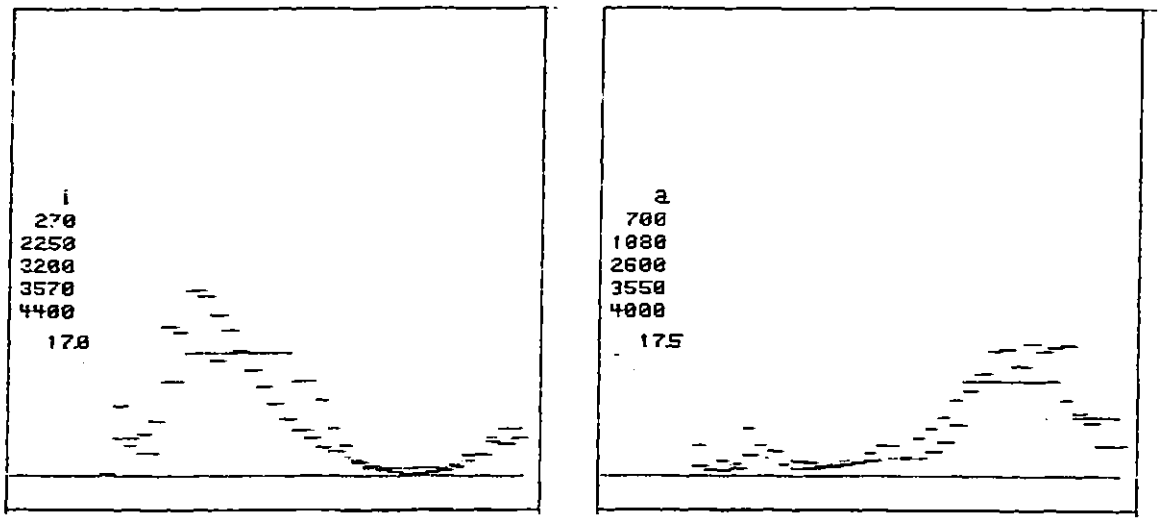


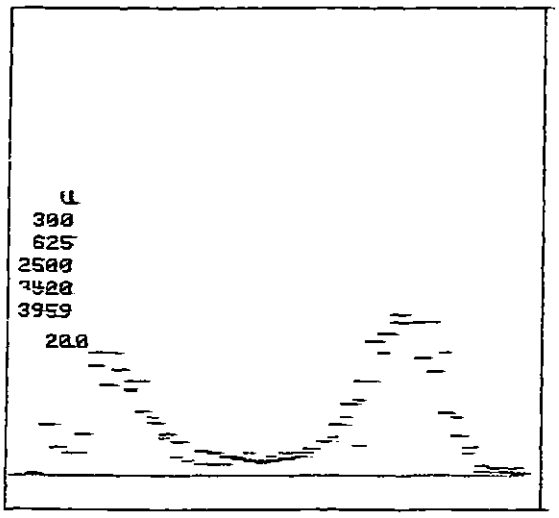

Example of extraction of vocal tract area functions. (left end: lips, 1 section = 1.1 cm)

Area functions by adaptive LPC

Due to Nakajima et al. [18]

Figure 39.




Area functions due to estimation

algorithm of Paul

(from Figure 8)

Figure 40.

Of all of the methods tested only the Paul scheme (Figure 40) is examined with acoustic input and the output compared to the corresponding X-rays. The shape of pharynx cavity is distorted for /i/ and the larynx in /a/ is shown slightly undersize. Otherwise, the accuracy of the three area functions approaches that of Paige and Zue's bandlimited area functions. (It does not, however, perform as well on all of the vowels with X-ray area functions--see Figure 8f.) In all cases, the larynx is clearly shown.

These results are inherently not subject to rigorous comparison as assumptions, data sets and data set sources vary widely. A weak conclusion can be drawn in the hope that sufficient consistency in the data sets does exist so that comparison of the results is somewhat valid.

For the test set, Paige and Zue's algorithm appears to give the most accurate area functions if one is allowed to compute the "acoustic" data from the target area functions using an idealized model and to have modes which cannot be measured without disturbing the subject. Again for the test set, the Paul algorithm appears to give the most accurate area functions when one is limited to real acoustic data. These conclusions, however, apply only to published results for the test set of /i/, /a/ and /u/ and could vary for other data sets.

## VIII. Discussion

This algorithm implies that the formants contain all of the necessary information about non-nasal sonorants for area function reconstruction (neglecting the area normalization problem). Its performance suggests that this implication is at least partially true. The maximum meaningful formant set (five formants) is sufficient to estimate the tract length of vowels to a reasonable accuracy. The length of consonants, however, is probably better estimated by continuity from the adjacent vowels. The area normalization appears to be adequately handled by the constant volume assumption. Given a length, the formants alone are shown by this algorithm to be sufficient to estimate the area function of most non-nasal sonorants. The pharyngeal consonants appear to be treated incorrectly by this algorithm. /g/, in inconclusive testing, appears subject to some difficulties, possibly as a result of a formant tracking error. (A larger set of vowels in /VgV/ utterances should be able to clear this up.) If one assumes that the formants, as measured, and the length are correct, then the fault must lie in the bandwidth function.

The bandwidth model as chosen here is a single-valued function $B=f(F)$. This suggests that only one area function can be used to generate a particular sound by the speaker. (Note, however, that the formants of the estimated area

function will be correct for any legal bandwidth function.) A multivalued bandwidth function would allow multiple area functions for a given set of formants. Coker's vocal tract model has eight explicit degrees of freedom in tract shape for non-nasal sonorants (with at least one more handled in the acoustic domain) [3], which is greater than the six (N+length) degrees allowed here. Possibly, a more complex bandwidth model would handle the pharyngeal and velar consonants correctly.

But what does the bandwidth function actually model? It is derived only from generalities observed in some X-ray area functions of vowels. The peaks of the function occur at 4 and 13 kHz--about the resonances of a 2.1 cm. uniform tube closed at one end and open at the other (see Equation 5a). The larynx, a fixed structure with a closure at the vocal cords and an opening into the far larger pharynx, is the only part of the vocal tract which approximates such a structure. Thus the bandwidth function appears to contain the information describing the larynx and its boundaries. Indeed, when the bandwidth function is made constant, the larynx vanishes from the area function and the general accuracy of the area function is degraded.

So far, the algorithm has been tested only on male speakers as their lower vocal cord vibration frequencies allow less difficult formant tracking than do the higher

female excitation frequencies. The algorithm would require, at most, only changes in a few parameters to analyze female voices. The constant in the yielding wall to rigid wall transform (Equation 3) might have to be changed as a result of slightly different dimensions and/or tissue impedances in the female vocal tract. The bandwidth function is likely to require modification as the female larynx is smaller than the male. This change, however, might be a simple scaling of the input frequencies or equivalently a scaling of the W and Bw of Equation 10. The limits on the length finder algorithm might have to be changed. Finally, the volume used in the area normalization will vary with the subject. (The data used here for normalization was from a subject with a rather large vocal tract.) These comments also apply if one attempts to "tune" the algorithm to any individual, regardless of the sex of the subject.

## IX. Conclusion

The algorithm provides a method for area function extraction from only the formants of the acoustic signal. As theory predicts that additional modes (the zeros of the lip input admittance) and the length are required to compute the area function, the method cannot be fully theoretically justified. The missing information must be obtained by the algorithm or contained within the algorithm by methods which are not theoretically provable. The limited available deformation and general smoothness of the tract suggest Wakita's method as a possible way to estimate the length. The volume statistics of some X-ray area functions suggest a simple method for area normalization without recourse to complex vocal tract models. Finally, placement of the area functions into the LPC framework allows a simple description of a set of constraints to compensate for the missing modes. Nothing, however, can fully compensate for the loss of the formants above the plane wave limit. All of these theoretically unprovable methods require empirical testing to evaluate their effectiveness.

Such testing shows the algorithm postulated here to perform adequately for most non-nasal sonorants. It requires no special equipment outside of that in a computer equipped speech laboratory, is easy to set up, fairly computationally efficient--the slowest part of this

implementation is the formant tracking operation,--poses no undue hazard to the subject, and compares favorably with the other currently known acoustic methods. Until exposureless, high frame rate, and mechanically silent three dimensional X-ray systems become available, the basic approach outlined in this paper should remain useful for vocal tract area function estimation from the acoustic waveform.

# X. Appendix

## LPC

LPC (linear predictive coding) is an algorithm which produces an all pole optimal estimate to an arbitrary signal. It can be developed in several ways, but the derivation favored by Makhoul and Wolf [13, 14] is both straightforward and clear.

Postulate that the nth sample of a signal may be estimated by a linear combination of the preceeding p samples:

$$\hat{S}_n = -\sum_{k=1}^{p} a_k S_{n-k} \tag{A1}$$

Define an error function:

$$E = \sum_{n} (S_n - \hat{S}_n)^2 \tag{A2}$$

Minimize the error:

$$0 = \frac{\partial E}{\partial a_k} = \frac{\partial}{\partial a_k} \left[ \sum_{n} \left( S_n + \sum_{k=1}^{p} a_k S_{n-k} \right) \right]^2 \tag{A3}$$

$$\sum_{k=1}^{p} a_k \sum_{n} S_{n-k} S_{n-i} = -\sum_{n} S_n S_{n-i} \qquad 1 \leq i \leq p \tag{A4}$$

If $S_n$ is assumed stationary, the summations over n can be simplified:

$$\sum_{k=1}^{p} a_k R_{|k-i|} = -R_i \qquad 1 \leq i \leq p \qquad \text{(A5)}$$

$$\text{where} \qquad R_i = \sum_{n=-\infty}^{\infty} S_n S_{n-i} \qquad \text{(A6)}$$

The set of simultaneous equations described by Equation A5 can be solved in several ways, the most efficient of which is the Levinson recursion [10, 13, 14].

$$i=1 \qquad \text{(A7.1)}$$

$$E_0 = R_0 \qquad \text{(A7.2)}$$

$$K_i = -\left[R_i + \sum_{j=1}^{i-1} a_j^{(i-1)} R_{i-j}\right] \Big/ E_{i-1} \qquad \text{(A7.3)}$$

$$a_i^{(i)} = K_i \qquad \text{(A7.4)}$$

$$a_j^{(i)} = a_j^{(i-1)} + K_i a_{i-j}^{(i-1)} \qquad 1 \leq j \leq i-1 \qquad \text{(A7.5)}$$

$$E_i = (1-K_i^2) E_{i-1} \qquad \text{(A7.6)}$$

$$i = i+1 \qquad \text{(A7.7)}$$

$$\text{if } (i \leq p) \text{ go to A7.3} \qquad \text{(A7.8)}$$

This procedure gives three useful outputs: the predictor coefficients $a_i$ $(1 \leq i \leq p)$, the error energy E and the reflection (or partial correlation) coefficients $K_i$ $(1 \leq i \leq p)$.

These reflecton coefficients provide an alternate description of the predictor coefficients in the form of a ladder network, which is formally equivalent to a nonuniform

transmission line composed of p+1 equal length uniform lossless sections of differing impedance with a (locally) matched source impedance and a zero impedance termination. This transmission line is also equivalent to a lossless acoustic tube composed of equal length uniform plane-wave propagating sections of differing cross-sectional area with an infinitely long input section (i.e. a locally matched source) and terminating into an infinite area tube (free space). These equivalent formulations (Figure A1) have only one mechanism for incorporating losses--the backward wave flowing out of the source (glottal) end.

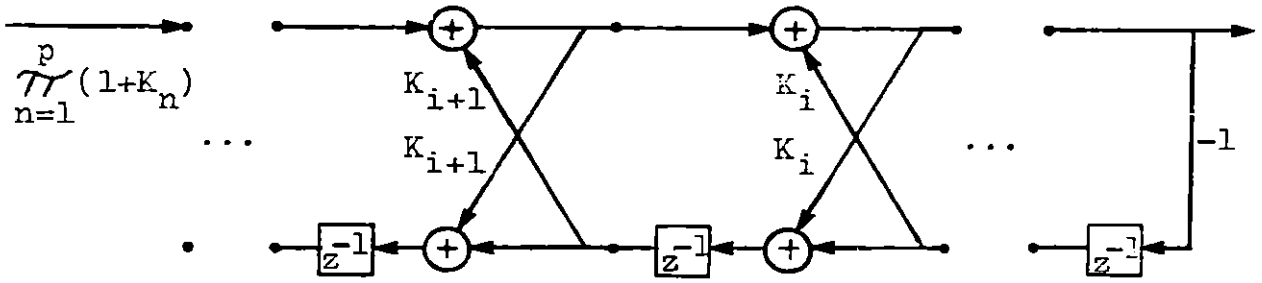
A reformulation of LPC into the frequency domain shows its error function to be of the form [13, 14]:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{P(\omega)}{\hat{P}(\omega)}\,d\omega=1$$

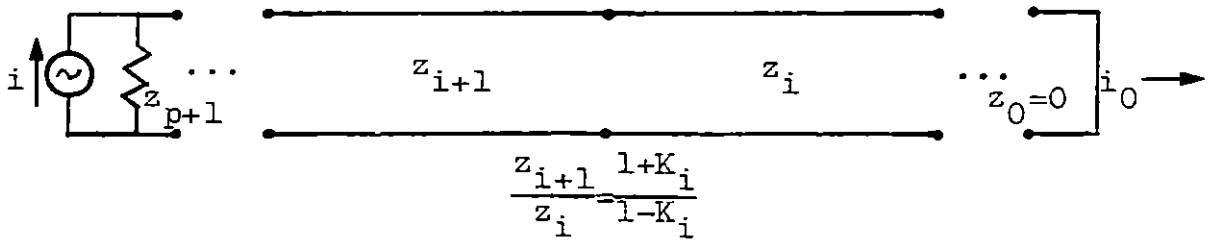
where $P(\omega)$=signal power spectrum
$\hat{P}(\omega)$=LPC estimate of $P(\omega)$

This suggests that regions where $P(\omega)>\hat{P}(\omega)$ make a greater local contribution to the error than the other regions. Thus LPC can estimate the generating filter of a signal produced by a periodic pulse train feeding an all pole filter because it pays more attention to the peaks of the lines in the line spectrum than to the regions in between the lines which contain little energy. This quality is very useful for estimating the vocal tract filtering effects for

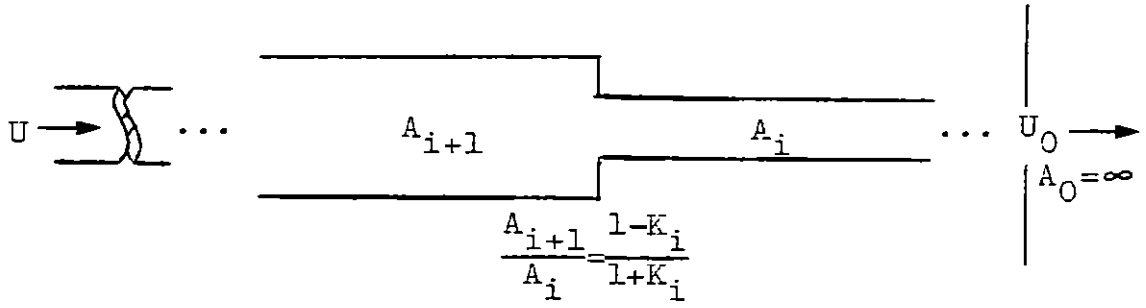
voiced (periodic source) phonemes (Figure A2).




LPC Ladder network



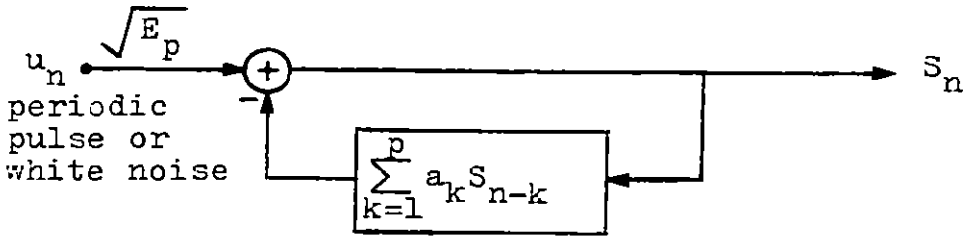

Transmission line



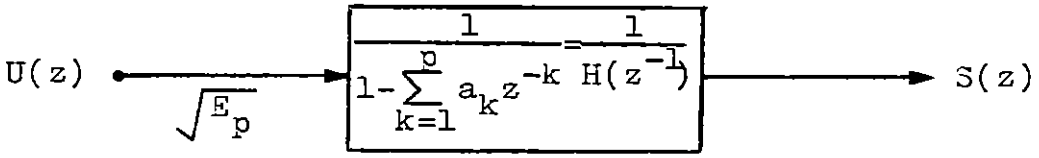

Acoustic tube

Three equivalent filters.

Figure A1.




Time domain




Frequency domain

LPC speech generation model

Figure A2.

## XI. Bibliography


1. B. S. Atal: Towards determining the articulator positions from the speech signals, Stockholm Speech Communication Seminar, August 1974.

2. T. Chiba, M. Kajiyama: The vowel, its nature and structure. Tokyo-Kaiseikan Publishing Co., Tokyo, 1941.

3. C. H. Coker: A model of articulatory dynamics and control. Proc. IEEE, vol. 64, no. 4, pp. 452-460, 1976.

4. G. Fant: Acoustic theory of speech production. Mouton and Co., 's-Gravenhage, 1960.

5. J. L. Flanagan: Speech analysis synthesis and perception. Second edition, Springer-Verlag, New York, 1972.

6. O. Fujimura, H. Ishida, S. Kiritani: Computer controlled dynamic cineradiography. Annual Bulletin (Research Inst of Logopedics and Phoniatrics), Univ. of Tokyo, no. 2, 6-10, 1968.

7. B. Gopinath, M. M. Sondhi: Determination of the shape of the human vocal tract from acoustical measurements. Bell System Technical Journal, vol. 49, no. 6, pp. 1195-1214, 1970.

8. E.H. Hafer: Speech analysis by articulatory synthesis. MS. Thesis, Northwestern Univ., Illinois, June 1974.

9. J. M. Heinz: Reduction of speech spectra to descriptions in terms of vocal tract area functions. ScD. Thesis, MIT, August 1962.

10. E. M. Hofstetter: An introduction to the mathematics of linear predictive filtering as applied to speech analysis and synthesis. Technical note 1973-36 Rev. 1, Lincoln Labratory, MIT, Massachusetts, April 1974.

11. R. A. Houde: A study of tongue body motion during selected speech sounds. PhD. Thesis, Univ. of Michigan, 1967.

12. P. Ladefoged, J. Anthony, C. Riley: Direct Measurement of the vocal tract. ASA Meeting, Houston, Nov. 1970.

13. J. Makhoul: Linear prediction: a tutorial review. Proc. IEEE, vol. 63, no. 4, pp. 561-580, April 1975.

14. J. Makhoul, J. J. Wolf: Linear prediction and the spectral analysis of speech. Bolt Beranek and Newman Inc., Cambridge, Mass., Report 2304, August 1972.

15. J. D. Markel: Digital inverse filtering--a new tool for formant trajectory estimation. IEEE Trans. Audio and Electroacoust. AU-20, June 1972.

16. P. Mermelstein: Determination of the vocal tract shape from measured formant frequencies. JASA, vol. 41, pp. 1283-1294, 1967.

17. P. M. Morse: Vibration and sound. McGraw-Hill Book Co., New York, 1948.

18. T. Nakajima, H. Omura, K. Tanaka, S. Ishizaki: Estimation of vocal tract area functions by adaptive inverse filtering methods and identification of articulatory model. Stockholm Speech Communication Seminar, August 1974.

19. A. Paige, V. Zue: Computation of vocal tract area functions, IEEE Trans. on Audio and Electroacoust., vol. AU-18, no. 1, March 1970.

20. D. B. Paul: unpublished observations at sampling rates of 6.67, 10, and 20 kHz.

21. J. S. Perkell: Physiology of speech production: results and implications of a quantitative cineradiographic study. MIT Press, Cambridge, Mass., 1969.

22. G. E. Peterson, H. L. Barney: Control methods used in a study of the vowels. JASA, vol. 24, pp. 175-184, 1952.

23. D. L. Rice: Articulatory tracking of the acoustic speech signal. Speech Communication Seminar, Stockholm, pp. 21-26, August 1974.

24. M. R. Schroeder: Description of the geometry of the human vocal tract by acoustic measurements. JASA, vol. 41, pp. 1002-1010, 1967.

25. M. R. Schroeder: personal communication.

26. M. M. Sondhi: Model for wave propagation in a lossy vocal tract. JASA, vol. 55, no. 5, pp. 1070-1076, May 1974.

27. K. N. Stevens: personal communication.

28. H. Wakita: Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. IEEE Trans. Audio and Electroacoust., vol. AU-21, pp. 417-427, 1973.

29. H. Wakita: An approach to vowel normalization. JASA, vol. 57, supplement no. 1, pp. 53, Spring 1975.

## XII. Biographical note

The author was born in Princeton, New Jersey and brought up in Middlesex, New Jersey. He attended The Johns Hopkins University and graduated in 1971 with a B.E.S. in Electrical Engineering and General and Departmental Honors. He received his S.M. and E.E. in Electrical Engineering from Massachusetts Institute of Technology in September 1973 and June 1976 respectively.

He is a member of Phi Beta Kappa, Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.