A Speech Training Aid for Hearing Impaired

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Master of Technology

by

Sumedha A. Kshirsagar

(96307023)

Guide: **Dr. P. C. Pandey**

Department of Electrical Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai 400 076
January 1998

INDIAN INSTITUTE OF TECHNOLOGY, BOMBAY

M.Tech. Dissertation Approval

Dissertation entitled **"A Speech Training Aid for Hearing Impaired"** by Sumedha A. Kshirsagar, is approved for the award of the degree of Master of Technology in Electrical Engineering.

Guide	:	PCpandey 22.1.98	(Dr. P. C. Pandey)
Internal Examiner	:		(Dr. V. M. Gadre)
External Examiner	:		(Dr. Arun Pande)
Chairman	:		(Prof. C Chattopadhyay)

Abstract

Sumedha Kshirsagar, "A Speech Training Aid for Hearing Impaired", M.Tech. dissertation, Dept. of Elec. Engg., Indian Institute of Technology, Bombay, January '98, Guide: Prof. P. C. Pandey.

Lack of auditory feedback in the hearing impaired results in failure to produce intelligible speech. Providing visual feedback of efforts involved in speech production process would help such a person in learning to speak. A system for analysing speech for extracting pitch, intensity, and vocal tract area, and displaying these in real time has been earlier developed at IIT Bombay. This system is based on TI/TMS320C25 (16-bit fixed point processor) DSP-board having on-board memory sharable with PC. The implementation of the algorithms for estimation of vocal tract area, pitch, and intensity has been done on the DSP-board and user interfacing is handled by the PC. The variation of these parameters with time can also be displayed for a duration of 1.28 second. "Areagram" is a display for variation of vocal tract area with time. The main limitation of the system is that, the vocal tract shape estimation cannot be made during stop closures as a result of very low energy in the speech signal and there are inconsistencies in the shape display even during the vowel segments.

In this project, the areagram display software is modified to obtain more realistic and consistent results. The modification is done for normalization of area values and their mapping to 16 grey levels to be displayed on the areagram. Care has been taken that erroneous peaks in area distribution do not affect this normalization. This gave more meaningful results when areagrams for individual vowels and vowel sequences were compared. An attempt was also made to improve this system for estimation over weak energy phonemes. For extraction of information during the stop closures, based on the area variation just before and after the closure, off-line processing of areagram matrix was done. Spatial low-pass filtering for 2-D interpolation did not help in extracting the information regarding place of closure. Normalizing speech segments with respect to average magnitude improved the estimation over semi-vowels. The minimum signal strength required for meaningful estimation has also come down by a factor of 2. The estimation algorithm was also implemented using floating point arithmatic to investigate the improvement in estimation over weak energy signal, but this was not much better than that achieved by normalization of input segments and real-time analysis in fixed point arithmatic.

Acknowledgments

I express my deep sense of gratitude towards my guide, Dr. P. C. Pandey, for his guidance and support. This work would not have been possible without his valuable suggestions and encouragement. I also thank my lab-mates, Mandar, Sachin, Mr. Chaudhari, and Harshal. Their time to time help in various matters made this work smooth and enjoyable. I also take this apportunity to thank Mr. A. D. Apte of the Standards Lab. Thanks to all my friends who helped me indirectly during this project work.

Sumedha Kshirsagar.

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Problem Overview

In children with normal hearing, the process of learning to speak is aided by auditory feedback. The hearing impaired children lack this feedback and therefore experience difficulty in acquiring the normal speech characteristics. Thus, in spite of proper speech production mechanism, these persons may not be able to produce intelligible speech. It is possible to teach such persons to speak by use of appropriate feedback.

Visual feedback can be provided by display of certain speech parameters which are easily controllable by the person undergoing speech training. These parameters should provide necessary cues for uttering a specific phoneme. Vocal tract area, pitch, and energy variations can be used for this, since these are directly linked with speech production. A PC based speech training aid could be developed for this. This should help in indicating how to produce a particular speech segment, and should also bring about the difference between actual and desired parameters. A deaf person can see changes in his or her own vocal tract shape in real-time on the PC monitor, and try to modify it in the learning process, by comparing with the variation pattern for teacher uttering the same segment.

1.2 Project Objective

This project is a continuation of the ongoing project at IIT Bombay [1, 2, 3, 4, 5, 6, 7]. The system as developed by Baragi [7], "STS-3", estimates the vocal tract shape and pitch from the speech signal, and displays the same along with intensity in real time. The vocal tract shape can also be displayed in the form of an "areagram" in review mode for a duration of 100 frames (1.28 second). Areagram is a two dimensional representation of varying area of vocal tract from the glottis to the lips with time. In "Animation mode", the captured frames can be seen on the display in slow motion. There is a facility for "Manual select" mode in which the captured frames can be

viewed one by one.

In case of stop consonants, during the closure, the energy in the speech signal is so low that the estimate of the vocal tract shape becomes erroneous. In the speech training aid developed so far, the vocal tract shape during the closure was frozen to the shape obtained just before this period. A better method would be to identify the duration for which the estimation is not reliable, and interpolate the vocal tract shape during this period based on the shape obtained just before and just after the closure. This is not possible in real time and is to be achieved by using the captured data. Any other possible method to solve this problem has to be investigated and implemented.

During the testing of the system, it was observed that comparative area variation for different vowels is not reflected in the areagram properly. The examination of the scaling procedure suggested the need for a robust method to map area values to 16 grey levels. This method should work independent of the phoneme uttered, retain all the necessary information that is obtained from processing of the speech signal, eliminating redundant information, and also lead to the display of the areagram which is realistic and more perceivable.

In this project, system STS-3 has been tested for consistency and appropriateness of its results. Various likely modifications have been invstigated, and appropriate ones have been incorporated as part of a new set-up "STS-4" which retains the hardware of STS-3.

1.3 Organization of Report

The following chapter reviews different speech training aids reported in various technical literature. It also explains the vocal tract area, pitch, and energy estimation from the speech signal. Chapter 3 explains the speech training aid STS-3 developed earlier by Baragi [7] in detail. It also elaborates on the need for modification of the software developed for the aid. The modifications made in the software are explained in Chapter 4. Testing and results are given in the fifth chapter. Summary and suggestions for further work are covered in the last chapter. The operation of the overall system developed STS-4 is given at the end as an appendix.

Chapter 2

SPEECH TRAINING AIDS

2.1 Introduction

Efforts have been made to develop a training aid for deaf in various forms. These include various visual speech training aids to provide visual feedback of speech parameters. Use of spectrograms, display of pitch using a color scale, monitoring of opening and closing of vocal cords *etc.* are included in these efforts. In these methods, the information displayed does not have a direct relationship to the actual physical efforts involved in speech production. Speech training aids were also developed in the form of games in order to make learning process more interesting. However, after realizing inadequacy of these techniques, several researchers have been developing visual aids that present information related to vocal tract shape (reviewd in [4]). The vocal tract shape display is often supplemented by voicing/pitch and energy tracks. This chapter reviews literature that has been published to report development of various speech training aids. Later sections of this chapter explains use of linear predictive coding (LPC) to solve the problem of estimation of vocal tract shape from speech signal. Estimation of pitch and energy of the speech signal is also described.

2.2 Vocal Tract Shape Display Aids

Estimation of vocal tract shape is the important part of any visual display aid for the hearing impaired. X-ray and MRI techniques have been used for determination of vocal tract shape. There are certain difficulties found in these techniques regarding the measurement of lateral dimension and strong limitations on safe dosage [8]. A simpler method is extraction of vocal tract shape from speech signal itself. There are two main techniques to determine vocal tract area from the acoustic signal: LPC method and lip impulse response method. In the first method, vocal tract area along the tract length is estimated by LPC analysis of speech during normal utterances. In the second method, an acoustic tube must be coupled to the mouth and the subject

must silently phonate [9], and therefore, the first method is preferred.

Formant frequencies have been used to generate vocal tract shape by Ladefoged *et.al.* [10]. Authors have stated that, since the acoustic structure of a vowel is fairly completely determined by the first three formants; it should also be possible to recover a plausible vocal tract shape for a vowel knowing the formants. They have specified the vocal tract in terms of distances of different points along the lower surface of the vocal tract assuming the upper surface to be fixed. To solve the problem, a set of x-ray images of tongue were analyzed for different vowel utterances. It was found that the tongue shapes could be described fairly precisely by simple proportions of a front raising component and a back raising component. A series of stepwise multiple regression analysis was then conducted in order to correlate these proportions and the formant frequencies. A relation between the distance between the lips and formants was found out in a similar manner. The validity of such a relation was tested by obtaining vocal tract shapes for various subjects and comparing them with x-ray images. The authors however comment that the method is useful for a limited set of vowel utterances.

Park, Kim, Lee, and Yoon [11] used the relation developed by Ladefoged *et.al.* in developing a speech training system for hearing impaired. This system displays intensity, fundamental frequency, and nasality along with vocal tract shape. Fundamental frequency and nasality are detected using separate vibration sensors. Vocal tract area function from lips to the glottis are found out using Wakita's method [8] supported by lip-to-lip distance found from first three formant frequencies as given by [10]. The vocal tract shape is displayed in lateral form where each section along the tract is represented by its height. A self training program is developed which displays correction messages to the trainee after comparing the speech parameters of the trainee with the reference ones. Tests were carried out to train deaf children successfully, for five Korean vowels.

David Rossiter *et.al.* [9] have reported design and implementation of graphical display that presents approximation to vocal tract area in real time for vowel articulation. Reflection coefficients are derived for a model of the vocal tract consisting of concatenation of uniform area acoustic tube sections, from the autocorrelation coefficients of the sampled speech signal. These are used for calculating the area function and spline interpolation is used to produce smooth contours. The cycle of analysis and display is repeated 20 times/s. There is a provision to store the vocal tract shape data, for use as a reference against which the user can match vocal performance. The authors have not reported the performance of the system for segments other than sustained vowels.

2.3 Estimation of Vocal Tract Shape

Estimation of vocal tract shape involves estimation of area along the vocal tract, from the glottis to the lips. To understand the use of LPC to vocal tract shape estimation,

it is necessary to understand model of speech production system. The discussion in this section is based on [8, 12].

2.3.1 Acoustic Tube Model for Speech Production System

The vocal tract can be regarded as an acoustic tube with a varying cross-sectional area. The tube is divided into M number of sections with equal length Δl. Each tube is assumed to be rigid and lossless [12].

Let $u_m(t,d)$ and $p_m(t,d)$ be the volume velocity and pressure respectively, in section m where t is the time and d is the distance variable. The subscript m is taken from lips to the glottis in increasing order as shown in the *Fig.* 2.1.

The velocity and pressure are related to each other by continuity and force equations [12, 13].

$$-\frac{\partial u_m(t,d)}{\partial d} = \frac{S_m}{\rho c^2}\frac{\partial p_m(t,d)}{\partial t} \tag{2.1}$$

$$-\frac{\partial p_m(t,d)}{\partial d} = \frac{\rho}{S_m}\frac{\partial u_m(t,d)}{\partial t} \tag{2.2}$$

where c is velocity of sound, S_m is the cross-sectional area of the m^{th} section, and ρ is the density of air. The volume velocity can be related to velocity potential Φ as

$$u_m(t,d) = -S_m\frac{\partial}{\partial d}\phi_m(t,d) \tag{2.3}$$

From these three equations, we get the wave equation for the velocity potential

$$\frac{\partial^2\phi_m(t,d)}{\partial d^2} - \frac{1}{c^2}\frac{\partial^2\phi_m(t,d)}{\partial t^2} = 0 \tag{2.4}$$

A solution to *Eqn.* 2.4 is

$$\phi_m(t,d) = Ae^{j\omega(t-d/c)} + Be^{j\omega(t+d/c)} \tag{2.5}$$

where A and B are constants. $u_m(t,d)$ is given by the difference of a component $u_m^+(t,d)$ of volume velocity of the sound wave traveling from the glottis to the lips and a component $u_m^-(t,d)$ due to wave traveling towards the glottis in the section m. Then $u_m(t,d)$ and $p_m(t,d)$ are given by,

$$u_m(t,d) = u_m^+(t,d) - u_m^-(t,d) \tag{2.6}$$

$$p_m(t,d) = \frac{\rho c}{S_m}\{u_m^+(t,d) + u_m^-(t,d)\} \tag{2.7}$$

where

$$u_m^+(t,d) = \frac{j\omega S_m A}{c}e^{j\omega(t-d/c)} \tag{2.8}$$

and

$$u_m^-(t,d) = \frac{j\omega S_m B}{c} e^{j\omega(t+d/c)} \tag{2.9}$$

As volume velocity and the pressure are continuous at the junction between section m and $m+1$, we can write

$$u_{m+1}(t,d_m) = u_m(t,d_m) \tag{2.10}$$

$$p_{m+1}(t,d_m) = p_m(t,d_m) \tag{2.11}$$

Here d_m represents the distance from the glottis to the junction between m and $(m+1)^{th}$ tubes, as shown in *Fig.* 2.1. Since it is assumed that the tubes are lossless, $u_{m+1}^+(t,d_m)$ is equivalent to the component of volume velocity that started at d_{m+1} at time $\Delta l/c$ earlier. Similarly $u_{m+1}^-(t,d_m)$ is equivalent to volume velocity that will arrive at d_{m+1} at time $\Delta l/c$ later. Hence the notation can be simplified by eliminating the distance variable, *i.e.*, $u_m^+(t,d_m) = u_m^+(t)$. Then *Eqn.* 2.10 and *Eqn.* 2.11 can be expressed, with using substitution form *Eqn.* 2.6 and *Eqn.* 2.7 respectively, as

$$u_{m+1}^+(t-\Delta t) - u_{m+1}^-(t+\Delta t) = u_m^+(t) - u_m^-(t) \tag{2.12}$$

$$\frac{\rho c}{S_{m+1}}\{u_{m+1}^+(t-\Delta t) + u_{m+1}^-(t+\Delta t)\} = \frac{\rho c}{S_m}\{u_m^+(t) + u_m^-(t)\} \tag{2.13}$$

where $\Delta t = \Delta l/c$. By rearranging *Eqn.* 2.12 and *Eqn.* 2.13 we get,

$$u_{m+1}^+(t-\Delta t) = \frac{1}{1+r_m}\{u_m^+(t) - r_m u_m^-(t)\} \tag{2.14}$$

and

$$u_{m+1}^-(t+\Delta t) = \frac{1}{1+r_m}\{-r_m u_m^+(t) + u_m^-(t)\} \tag{2.15}$$

where

$$r_m = \frac{S_m - S_{m+1}}{S_m + S_{m+1}} \tag{2.16}$$

defines the reflection coefficient at the junction between sections m and $m+1$.

By taking z transforms , with respect to sample delay of $2\Delta t$, of both sides of *Eqn.* 2.14 and *Eqn.* 2.15, and combining them we can write,

$$\begin{bmatrix} U_{m+1}^+(z) \\ U_{m+1}^-(z) \end{bmatrix} = \frac{z^{1/2}}{1+r_m} \begin{bmatrix} 1 & -r_m \\ -r_m z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} U_m^+(z) \\ U_m^-(z) \end{bmatrix} \tag{2.17}$$

Here we have chosen sample delay as $2\Delta t$, resulting in

$$Z[u_{m+1}^-(t-\Delta t)] = z^{-\frac{1}{2}} U_{m+1}^-(z)$$

in order to obtain arecursive relationship that would be consistent to that of the inverse filter to be derived in the next subsection.

Acoustic tube inverse transfer function can be derived from this by applying boundary condition. Let us assume that the front end (lip end) is connected to a tube of infinite area (*i.e.*, $r_0 = 1$) and that the back end (glottis end) is connected to a tube of area S_{M+1} that is terminated with its characteristic impedance, $\rho c/S_{M+1}$. Also let us assume that the tube is excited at the back end by a source through this characteristic impedance with forward going volume velocity of $U^+_{M+1}(z)$. We now can write using *Eqn.* 2.17 as,

$$\begin{bmatrix} U^+_{m+1}(z) \\ U^-_{m+1}(z) \end{bmatrix} = z^{(m+1)/2} K_m \begin{bmatrix} D^+_m(z) \\ D^-_m(z) \end{bmatrix} \{U^+_0(z) - U^-_0(z)\} \tag{2.18}$$

where

$$K_m = \prod_{i=0}^{m} \frac{1}{1 + r_i} \tag{2.19}$$

and

$$\begin{bmatrix} D^+_m(z) \\ D^-_m(z) \end{bmatrix} = \begin{bmatrix} 1 & -r_m \\ -r_m z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 & -r_{m-1} \\ -r_{m-1} z^{-1} & z^{-1} \end{bmatrix} \cdots \cdots \begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ -z^{-1} \end{bmatrix} \tag{2.20}$$

This leads to the following recursive equation

$$\begin{bmatrix} D^+_{m+1}(z) \\ D^-_{m+1}(z) \end{bmatrix} = \begin{bmatrix} 1 & -r_{m+1} \\ -r_{m+1} z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} D^+_m(z) \\ D^-_m(z) \end{bmatrix} \tag{2.21}$$

If $U^+_{m+1}(z)$ and $U^-_{m+1}(z)$ are normalized by K_m,

$$\begin{bmatrix} \hat{U}^+_{m+1}(z) \\ \hat{U}^-_{m+1}(z) \end{bmatrix} = z^{(m+1)/2} \begin{bmatrix} D^+_m(z) \\ D^-_m(z) \end{bmatrix} \{U^+_0(z) - U^-_0(z)\} \tag{2.22}$$

where

$$\hat{U}^+_{m+1}(z) = U^+_{m+1}(z)/K_m$$

and

$$\hat{U}^-_{m+1}(z) = U^-_{m+1}(z)/K_m$$

The inverse transfer function between the section 'm' and the front end is defined as,

$$\begin{aligned} C_m^+ &= \frac{\text{forward volume velocity component at the end of section '}m\text{'}}{\text{volume velocity at the front end}} \\ &= \frac{\hat{U}_{m+1}^+(z)}{U_0^+(z) - U_0^-(z)} \end{aligned} \tag{2.23}$$

Similarly we can define

$$C_m^- = \frac{\hat{U}_{m+1}^-(z)}{U_0^+(z) - U_0^-(z)} \tag{2.24}$$

Hence, from eqs. 2.22, 2.23, and 2.24 we can arrive at

$$\begin{bmatrix} C_{m+1}^+(z) \\ C_{m+1}^-(z) \end{bmatrix} = z^{m/2+1} \begin{bmatrix} D_{m+1}^+(z) \\ D_{m+1}^-(z) \end{bmatrix} \tag{2.25}$$

2.3.2 Linear Predictive Coding

The speech analysis model is shown in *Fig.* 2.2. It is assumed that the speech is non-nasalized and voiced. In this model , impulse train at the input is analogous to the glottal excitation and filter is analogous to the vocal tract. The filter is assumed to have only poles. This implies that the inverse filter is having only zeros. The purpose of the analysis is to determine the inverse filter transfer function so that the difference between the output of the inverse filer and the input pulse train attains a minimum for a certain error criterion. It will eventually be seen that the inverse filter model so obtained is equivalent to the acoustic tube model derived before.

The transfer function of the inverse filter of order 'M' can be expressed as,

$$A(z) = Y(z)/X(z) = \sum_{i=0}^{M} a_i z^{-i}, \qquad a_0 = 1 \tag{2.26}$$

where $X(z)$ and $Y(z)$ are the z transforms of the input and the output of the inverse filter respectively. The ordered set $\{a_0, a_1, \ldots, a_M\}$ define the inverse filter coefficients. The error between the output $y(n)$ and an impulse train of amplitude η and repetetion period η_0 is given by

$$\in (n) = \sum_{i=0}^{M} a_i x(n-i) - \eta \delta_{n_0}(n) \tag{2.27}$$

Since this impulse train is an unknown quantity, some kind of assumption has to be made in order to perform the comparison. The power spectral enevelope of speech sound is assumed to have poles only. Hence, the speech signal has the form

$$X(z) = \eta / A(z)$$

This assumption implies that the input impulse train can be replaced by an impuse of strength η. Hence, impuse train in *Eqn.* 2.27 can be replaced by $\eta\delta(n)$.

Let us consider a frame of speech signal over interval $0 \leq n \leq N-1$, corresponding output $y(n)$ will be nonzero over $0 \leq n \leq N+M-1$. For this frame, we apply least mean square error technique to the *Eqn.* 2.27 i.e.,

$$\frac{\partial E}{\partial a_j} = 0 \qquad \text{for } j = 1, 2, \ldots, M \tag{2.28}$$

where E is the sum of $\in_n^2$ over $0 \leq n \leq N+M-1$

The above equation is equivalent to,

$$2\sum_{i=0}^{M} a_i r_{i-j} - 2\eta \sum_{n=0}^{N+M-1} x(n-j)\delta(n) = 0, \qquad \text{for } j = 1, 2, \ldots, M \tag{2.29}$$

where r_k represents the k^{th} auto correlation of the sequence $x(n)$. As it is assumed that $x(n)$ is non zero for $0 \leq n \leq N-1$, the above equation reduces to,

$$\sum_{i=0}^{M} a_i r_{i-j} = 0 \qquad j = 1, 2, \ldots, M \tag{2.30}$$

where $a_0 = 1$.

The set of M simultaneous linear equations in 2.30 can be solved by using Robinson's method. The problem can be solved in a recursive manner. The steps of recursion to obtain the set $\{a_i\}$ are given by

$$\begin{aligned}
a_0^{(m+1)} &= a_0^{(m)} = 1 \\
a_1^{(m+1)} &= a_l^{(m)} + k_m a_{(m)}^{(m)} \\
a_2^{(m+1)} &= a_2^{(m)} + k_m a_{(m-1)}^{(m)} \\
&\vdots \\
a_m^{(m+1)} &= a_m^{(m)} + k_m a_{(1)}^{(m)} \\
a_{m+1}^{(m+1)} &= k_m
\end{aligned} \tag{2.31}$$

Here, superscripts of the filter coefficients denote recursion step and k_m is given by,

$$k_m = -\frac{\sum_{i=0}^{m} a_i^{(m)} r_{m+1-i}}{\sum_{i=0}^{m} a_i^{(m)} r_i} \tag{2.32}$$

Eqn. 2.31 can be transformed into recursive matrix form. By multiplying by $z^{-1}, i = 0, 1, \ldots, m$ on both sides of *Eqn.* 2.31, (i increasing from top to bottom), and summing up over all i we get,

$$A_{m+1}(z) = A_m(z) - k_m B_m(z) \tag{2.33}$$

where

$$A_m(z) = \sum_{i=0}^{m} a_i^{(m)} z^{-i} \tag{2.34}$$

and

$$B_m(z) = -z^{-(m+1)} A_m(1/z) \tag{2.35}$$

Similarly, multiplying by $z^{-(m+1-i)}$ on both sides of *Eqn.* 2.31 and summing over all i gives

$$B_{m+1}(z) = z^{-1}\{B_m(z) - k_m A_m(z)\} \tag{2.36}$$

Using matrix notation for *Eqn.* 2.33 and *Eqn.* 2.36 we get the following:

$$\begin{bmatrix} A_{m+1}(z) \\ B_{m+1}(z) \end{bmatrix} = \begin{bmatrix} 1 & -k_m \\ -k_m z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} A_m(z) \\ B_m(z) \end{bmatrix} \tag{2.37}$$

Initial values of $A_0(z)$ and $B_0(z)$ are 1 and $-z^{-1}$ respectively. From *Eqn.* 2.32, $k_0 = r_1/r_0$. The quantities at steps $m = 1$ to $m = M$ can be found out from eqs 2.37, 2.32, and 2.31 in a recursive manner. We see that recursive relation for the inverse transfer function for the acoustic tube model as given in *Eqn.* 2.21 and the relationship for the optimal inverse filter as given in *Eqn.* 2.37 are equivalent. By setting $A_M(z)$ equal to $D_M(z)$, we get the reflection coefficient r_m as

$$r_m = k_{m-1} \tag{2.38}$$

Thus it becomes possible to obtain the reflection coefficients from the speech waveform.

In order to achieve speedy computation, the calculation of reflection coefficients is to be done on the DSP processor TMS320C25. This is a fixed point processor. Hence, appropriate modification for the above described recursive algorithm was needed, so

that it could be implemented on the fixed point processor. For this purpose, we will be using Le Roux and Gueguen algorithm [14].

2.4 Pitch Estimation

Information about pitch is useful in a training aid along with the vocal tract shape. Short time auto-correlation method can be used for estimation of pitch in real time. The short time auto-correlation is defined as,

$$r_n(k) = \sum_{m=0}^{N-1-k} [x(n+m)w(-m)][x(n+m+k)w(-k-m)] \tag{2.39}$$

where $x(.)$ is the speech signal, $w(.)$ is the window placed at sample n and N is the window length. The periodicity of the auto-correlation function can be used to extract the pitch. The estimate however can be erroneous, because of the peaky nature of the function. These peaks can be attributed to damped oscillations of the vocal tract response which are responsible for each period of the speech wave. Hence, prior to the computation of auto-correlation coefficients, center clipping of the signal should be done [12]. The output is made zero if the input is below a certain threshold and is made equal to input otherwise. The threshold can be chosen as 30 to 50% of the peak encountered in that window. The window should be of the duration to contain at least 3 to 4 pitch periods.

2.5 Energy Estimation

The estimation of energy is done using the relation,

$$E_n = \sum_{m=0}^{N-1} (x(m)w(n-m))^2 \tag{2.40}$$

Here, if the window is too small, then the energy function will fluctuate rapidly. If on the other hand, N is too large, as compared to the pitch period, E_n will vary slowly and will not reflect the changing properties of the speech signal. The problem is further complicated by large variation in the pitch of males, females, and children. The suitable choice of N could be between 100-200, for a sampling frequency of 10 kHz [12].

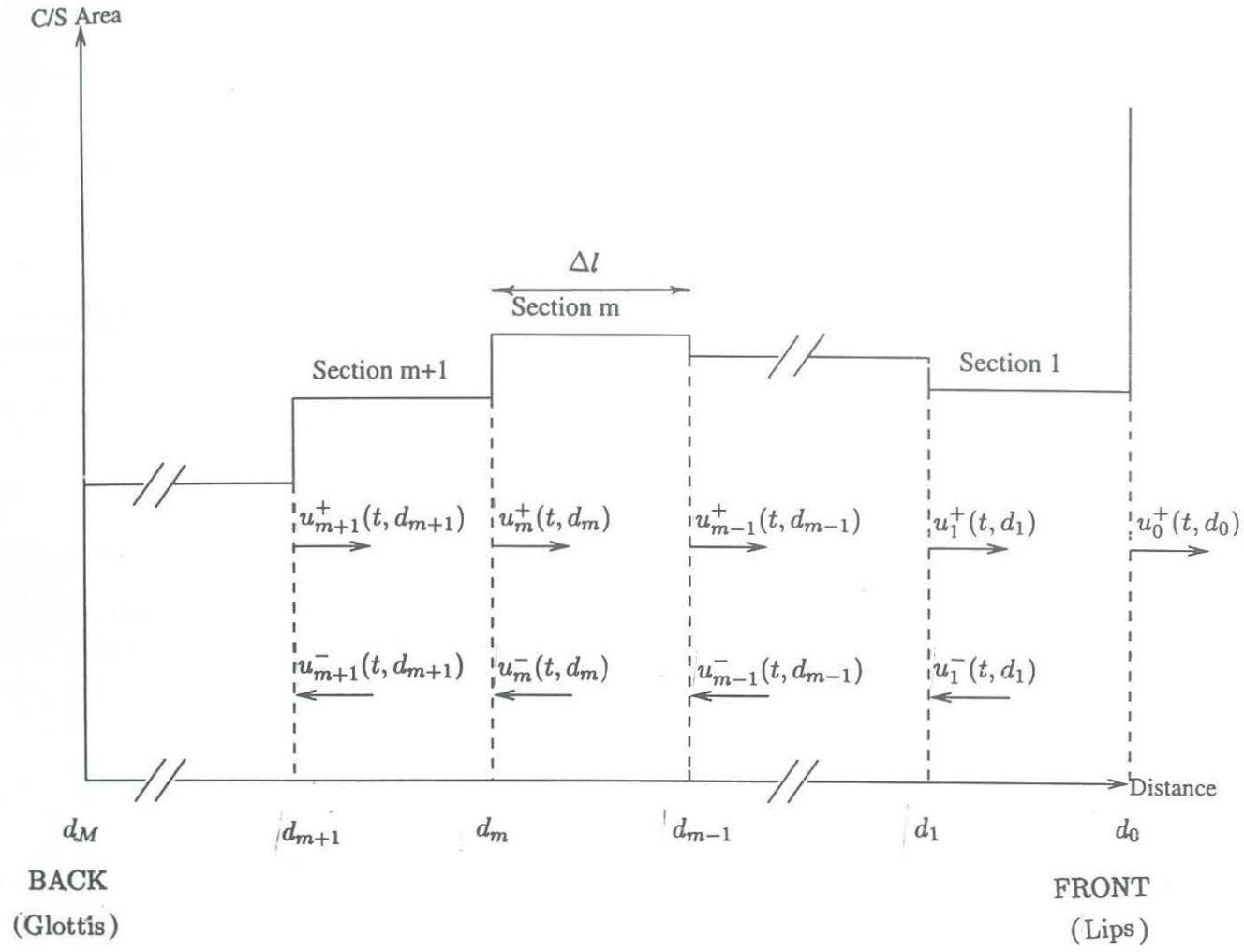

Figure 2.1: Acoustic tube model of the vocal tract

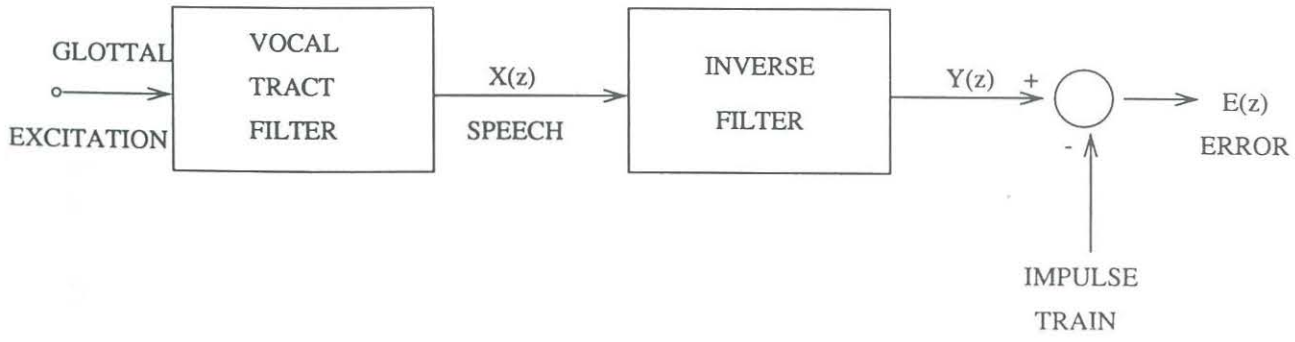

Figure 2.2: Speech analysis model

Chapter 3

SPEECH TRAINING AID "STS-3"

3.1 Introduction

In the previous chapter, the theory behind the system implementation has been presented. Continuing efforts at IIT Bombay [1, 2, 3, 4, 5, 6] to develop a speech training aid using this approach have resulted in the development of a prototype system STS-3, developed as an M.Tech. project by Ashok Baragi [7]. The need for real-time display of the vocal tract shape demands for fast computation. Also, user friendly display of the speech training aid is required. To accomplish both these tasks, the system was implemented using a DSP board interfaced to PC-bus. Data acquisition and the computationally intensive part is implemented on the DSP board and the user interface and display are handled by the PC. In this chapter, the system implementation of STS-3 is described, followed by test results with this system, and finally the modifications required are outlined.

3.2 System Implementation

In order to accomplish the desired aim of real-time display of vocal tract shape, pitch, and energy, the tasks can be divided into signal acquisition and preprocessing, implementing estimation algorithms, and presenting results in a form understandable by the user. This section explains overall algorithmic steps involved. The hardware setup and implementation steps in the software are explained in later sections.

3.2.1 Signal Acquisition and Preprocessing

The speech signal is to be acquired through a microphone and signal conditioning circuit. It is digitized at a rate of 10 k Sa/s. The properties of the speech signal

change relatively slowly with time. During certain sustained vowel segments, the vocal tract shape may remain stationary for durations upto 200 ms. In most situations, however, vowel durations average about 80 ms, and "short time" analysis can be applied [15]. Thus, the signal is processed in discrete blocks, assuming the parameters to be unchanged for the duration of that block or window. The choice for the length of analysis window is also influenced by pitch and energy estimation as explained in the sections 2.4 and 2.5, and it is desirable to have it about 20 ms. In the present system, it is selected to be 256 samples, *i.e.* 25.6 ms.

The analysis window is pitch asynchronous. This results in discontinuities in the paramaters estimated by successive analysis windows. Therefore, it was decided to use overlapping positioning of windows. It was examined and found by Gavankar [6] that 50% overlap improves the estimation. Thus, in each analysis window of 256 samples, 128 samples are from the previous window.

Pre-emphasis is to be applied to the speech signal before processing. In the derivation of the estimation algorithm, the vocal tract tube model was assumed to be lossless. Losses due to the glottal source and radiation impedance are not considered. The effect of these can be canceled by using 6 dB/octave equalization which takes into account -12 dB/octave slope for the glottal wave spectral envelope, and 6 dB/octave slope for the radiation impedance [8]. The signal used for processing $x(n)$ is then given by

$$x(n) = s(n+1) - s(n) \tag{3.1}$$

where s(n) is the input speech signal.

Hamming window is then applied to data block of the samples in the analysis window. This is in order to reduce the effects of large prediction errors at the boundaries of the blocks. These errors occur due to the fact that processing over the blocks assumes that the signal is zero beyond the boundaries of the block. The error is large at the beginning, since we are trying to predict the signal from samples that have been arbitrarily set to zero. At the end of the block, the error is large because we are trying to predict zero from non-zero samples. Hence, a window that tapers the speech segment to zero towards the boundaries is generally used [12].

3.2.2 Parameter Estimation

The steps of pre-emphasis, Hamming windowing, and parameter estimation are carried out block-by-block. As described in subsection 2.3.1, vocal tract has been modelled as a concatenation of 12 tube sections of different cross-sectional area. In this acoustic tube model, reflections of the acoustic wave occur at the tube interfaces due to different areas on the two sides. It has been further shown that the reflection coefficients can be calculated from auto-correlation coefficients of the speech signal. The recursive algorithm is given by *Eqn.* 2.31, and *Eqn.* 2.38. As said before, Le Roux and Gueguen algorithm [14] is used for fixed point implementation of this. The reflection coefficients are related to the acoustic tube areas as given by *Eqn.* 2.16, and therefore

we get,

$$S_m = \frac{1 + r_m}{1 - r_m} S_{m+1} \tag{3.2}$$

The area function, S_m for $m = 1, \ldots, M$, where M is the predictor order, can be obtained by specifying $S_{M+1} = 1$. Thus, all the other area values are in relation with the area at the glottis end.

In the system developed by Gavankar [6], these 12 area values were displayed on the monitor using 12 straight lines. Distances of these lines from a reference line were proportional to the areas of the corresponding sections. In order to make the vocal tract shape look more realistic, it was thought necessary to interpolate these 12 values to a larger number using some suitable curve fitting algorithm. This involves finding an equation of a curve that passes through these available 12 points, and using this equation to obtain more points. This may not produce good results with an 11th degree polynomial since, it can have large peaks. Hence, these 12 area values can be divided into suitable groups, and interpolation can be done seperately on each group. The interpolation algorithm should be computationally efficient so that real time implementation is possible. Baragi [7] assessed the possibility of application of Lagrange's polynomial for interpolation. However, this was found unsuitable for fixed point implementation. Also, interpolated values are not guaranteed to be positive even if all the control points are positive. Bezier form algorithm [16] was found suitable and is now used for interpolation.

Bezier form algorithm finds an equation involving a third degree polynomial of the curve, given four control points. This curve passes through the two extreme points and the remaining two points are used to define tangents to the curve. The algorithm ensures that the curve obtained is bound by the quadrilateral defined by the four control points. (Refer Appendix A).

The 12 area values are interpolated to 176 values. This number was decided by considering the computation time required and perceptivity of the resulting vocal tract shape. For this, all the 12 area values are normalized with respect to the maximum amongst them, for suitability of fixed point computation. The 12 area values are then divided into 4 groups, each group consisting of 4 points:

Group 1 = $A0, A1, A2, A3$
Group 2 = $A3, A4, A5, A6$
Group 2 = $A6, A7, A8, A9$
Group 3 = $A8, A9, A10, A11$

The curve fitting is done for all the four groups, extracting 48 points from each group.

Short time energy function is defined by *Eqn.* 2.40. It can be seen that the zeroth auto-correlation coefficient can be used as an estimate of energy. No additional computation is necessary to be done since finding auto-correlation coefficient is a part of vocal tract area estimation algorithm. It is to be noted here that, we get energy calculated over the data block after pre-emphasis and application of Hamming window.

The pitch is extracted using the method explained in section 2.3. The speech signal is center clipped using a threshold of 50% of the peak encountered in the frame. Then auto-correlation function is calculated. The position of the first peak which is above 50% of the zeroth value gives the pitch period.

3.2.3 Presentation of Results

After the estimation of all the parameters is done block-by-block, these are to be presented to the user. Parameters obtained for one block are considered as a frame for the display purpose. The display refresh rate should be adequate for real-time operation. Pitch, energy, and vocal tract shape contour consisting of 176 points are to be displayed for every frame. In addition to real time display, it is necessary to store the information extracted so that it can be reviewed later on. It should be possible to display a selected frame from the stored frames and to compare it with the one being displayed currently. Also, some tool should be available which will enable study of a number of frames at a time.

Speech parameters for the most recent 100 frames are kept available for use in review mode. In review mode, the software running on PC generates "areagram" from this captured data for 100 frames (1.28 seconds). The areagram feature was developed in order to study the consistency in the vocal tract shape estimation, and to study the transitions at vowel-nonvowel boundaries. Areagram is a two dimensional representation of vocal tract area with time. In the areagram display, x-axis shows the number of frames. It is to be noted that, each frame corresponds to a 25.6 ms analysis block, and the signal blocks have 50% overlap. Therefore, successive frames are positioned 12.8 ms apart. The distance from glottis to lips is plotted along y-axis. The grey level at any point on the areagram represents the area as a function of time and distance. The brightest shade corresponds to maximum area or largest opening. The darkest region indicates closure or minimum area. Each pixel is plotted five times in the x direction, in order to get better perception of the display. It is also possible to store the areagram in the form of an image in a file. The image file can be viewed on the monitor, can be printed on a laser printer, or can be converted "post script" format (by using XView). *Fig.* 3.1 shows the display in review mode showing the areagram, pitch, and energy variation for the synthetic vowel sequence /uai/, with $f_0 = 125Hz$.

In real-time as well as review mode, the display should be realistic, and easily understandable by the user. The system is to be implemented in such a manner that while estimation is done over current frame, signal can be acquired for the next frame and display can be refreshed for the previous frame. Next section explains how this is implemented with help of the hardware and software to support that.

3.3 Hardware Setup

The hardware setup of the system STS-3 can be represented in the *Fig.* 3.2.

1. Speech signal acquired from a microphone is amplified and filtered using the analog signal conditioning circuit developed by Gavankar [6]. It has a pre-amplifier with variable gain. It is followed by a seventh order elliptic low-pass filter which acts as anti-aliasing filter. It has a pass band up to 4.6 kHz, with a pass band ripple of 0.3 dB. The stop band starts at 5 kHz with minimum attenuation of 40 dB.

2. The filtered analog signal is fed to the TI/TMS320C25 based DSP-board [17]. This can be used as add-on board on the PC mother board, and has on-board ADC, DAC, and timer. The ADC has a resolution of 16 bits and its conversion rate can be set by configuring the timer. It has been set to 10 kHz. The ADC has an input range of ± 10 V. The processor operates at 40 MHz. The DSP board is interfaced with the PC using the PC-bus, and the data memory area of the processor can be directly accessed by the PC which enables fast data transfer between the PC and the DSP processor. Tasks handled by the DSP board include acquisition of data, calculation of reflection coefficients, pitch, and energy, and curve fitting of area function.

3. The PC communicates with the DSP board, acquires various parameters computed by the processor, and handles user interface. Calculation of area from reflection coefficients and display of various speech parameters are the main tasks handled by the PC. The parameters are displayed by directly writing the image on to the VGA RAM.

3.4 Software Implementation

The software developed can be divided into the one running on the DSP board and the other on the PC. Allocation of the tasks is shown in the *Fig.* 3.3 and can be described as the following:

1. The program running on the PC

   - initializes graphics display
   - initializes the DSP board
   - writes necessary information (Hamming window coefficients and coefficients of the curve fitting polynomial) on to the on-board data memory, and loads the object file on to the program memory of the DSP processor.

2. The program running on the DSP board:

- acquires analog data in blocks of 256 samples with 50% overlap, at the rate of 10 k Sa/s
- applies pre-emphasis and Hamming window
- calculates reflection coefficients by use of Le Roux and Gueguen algorithm [14]. The predictor order is 12, thus 12 reflection coefficients are obtained. The energy is estimated as the zeroth auto-correlation coefficient.

All this information can be stored in the on-board memory from which it can directly be read by the PC.

3. The calculation of area from reflection coefficients is done by the PC, using *Eqn.* 3.2 and the area values are written back to the DSP board for curve fitting.

4. Simultaneously, pitch is calculated by using short time auto-correlation method on the DSP-board.

5. 12 area values are interpolated to 176 values using Bezier form algorithm. The interpolated vocal tract area values, along with the pitch are read back by the PC.

6. In real-time mode, this process continues and the current vocal tract shape is displayed on the monitor by directly writing to the VGA RAM. This enables fast display refresh. If energy is below a certain threshold, then vocal tract shape is frozen to the present value, till energy again goes back above this threshold. The energy threshold was selected after observing the areagram of different speech utterances [7].

7. Vocal tract area, pitch, and energy for 100 frames can be stored when real-time operation is in progress. These can be used in review mode as explained before.

3.5 Testing of STS-3 and Modifications Required

It was reported by Baragi [7] that the estimated vocal tract shapes for vowels were comparable with those given by Rabiner & Schafer [12]. During this project, STS-3 was further tested for consistency, by obtaining areagrams for different vowel utterances by various subjects. Though individual areagrams showed consistency, comparison between areagrams for different vowels showed discrepancy. Hence, it was found necessary to work in the direction of devising a more robust method for scaling the area values and displaying the areagram.

As explained in section 1.2, the estimation of area becomes erroneous during utterance of the stop consonants because of very low energy in the speech signal. Interpolation from the area values obtained just before and just after the stop was

thought of as a possible option to obtain the missing area variation. This has to be done off-line, and can be used in the review mode. The areagram can be looked at as an image or two dimensional matrix. Then spatial averaging can be used for interpolation.

The implementation of the aid is done using fixed point arithmatic. Care has been taken to avoid overflows, especially, in the computation of auto-correlation coefficients. This leads to an upper limit on the minimum strength of speech signal necessary for satisfactory estimation. This also results in unreliable estimates during low energy frames in vowel-consonant-vowel sequences. It was experimentally found out that the minimum signal strength for satisfactory estimation of area as well as pitch for the vowels is approximately 1.3 V rms. Thus, it was thought necessary to investigate the results of estimation by:

1. implementing the estimation algorithm in floating point arithmatic

2. selectively scaling the digitized speech signal, when its strength is low.

The modifications incorporated in the software are described in the following chapter.

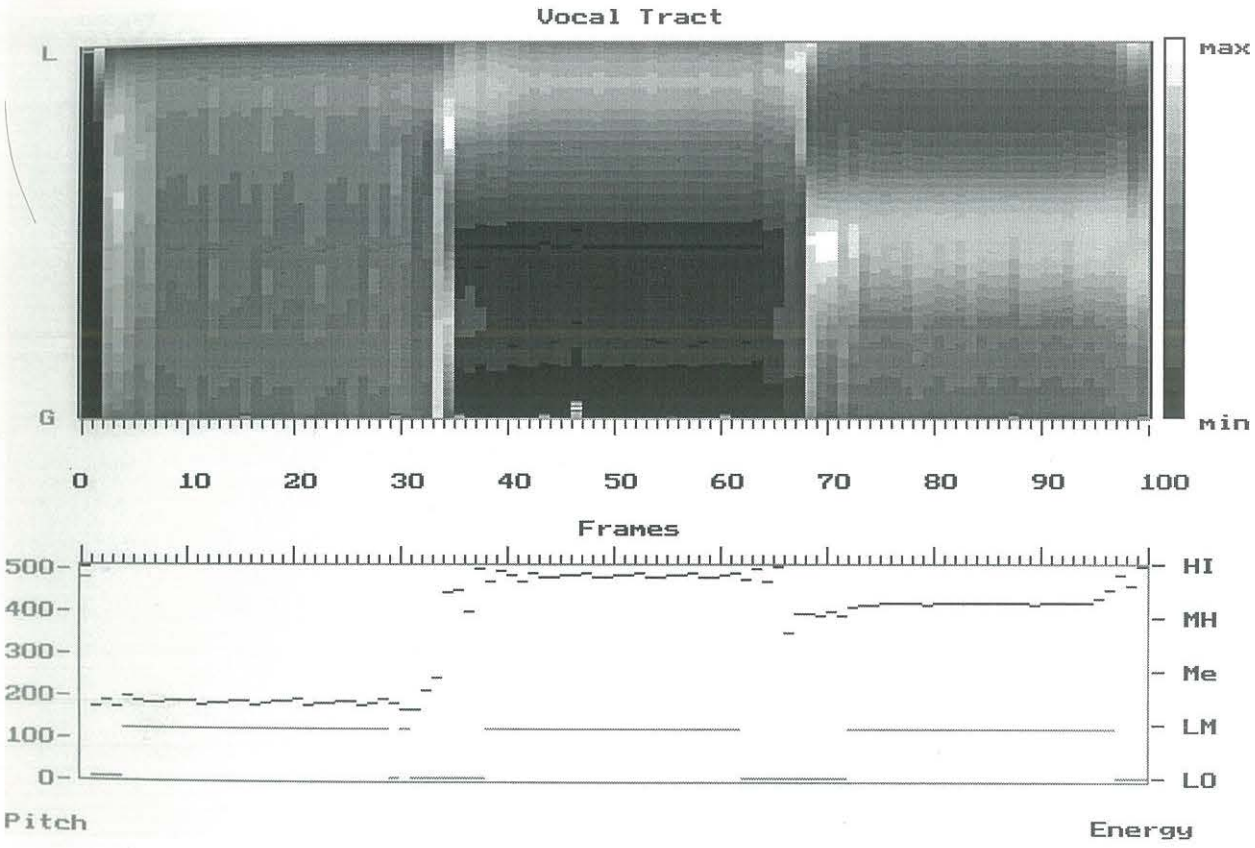

Figure 3.1: Areagram, pitch, and energy display for synthetic vowel sequence /uai/

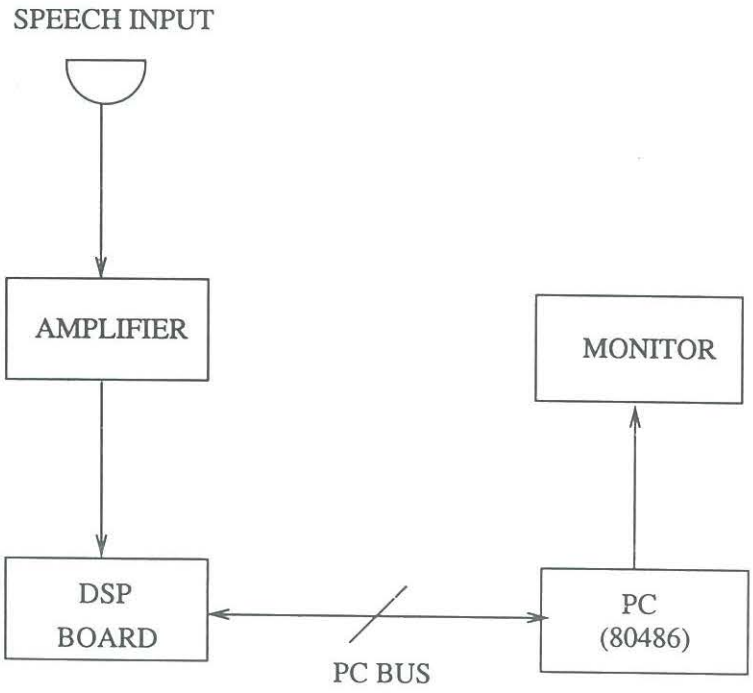

Figure 3.2: Block diagram of the aid

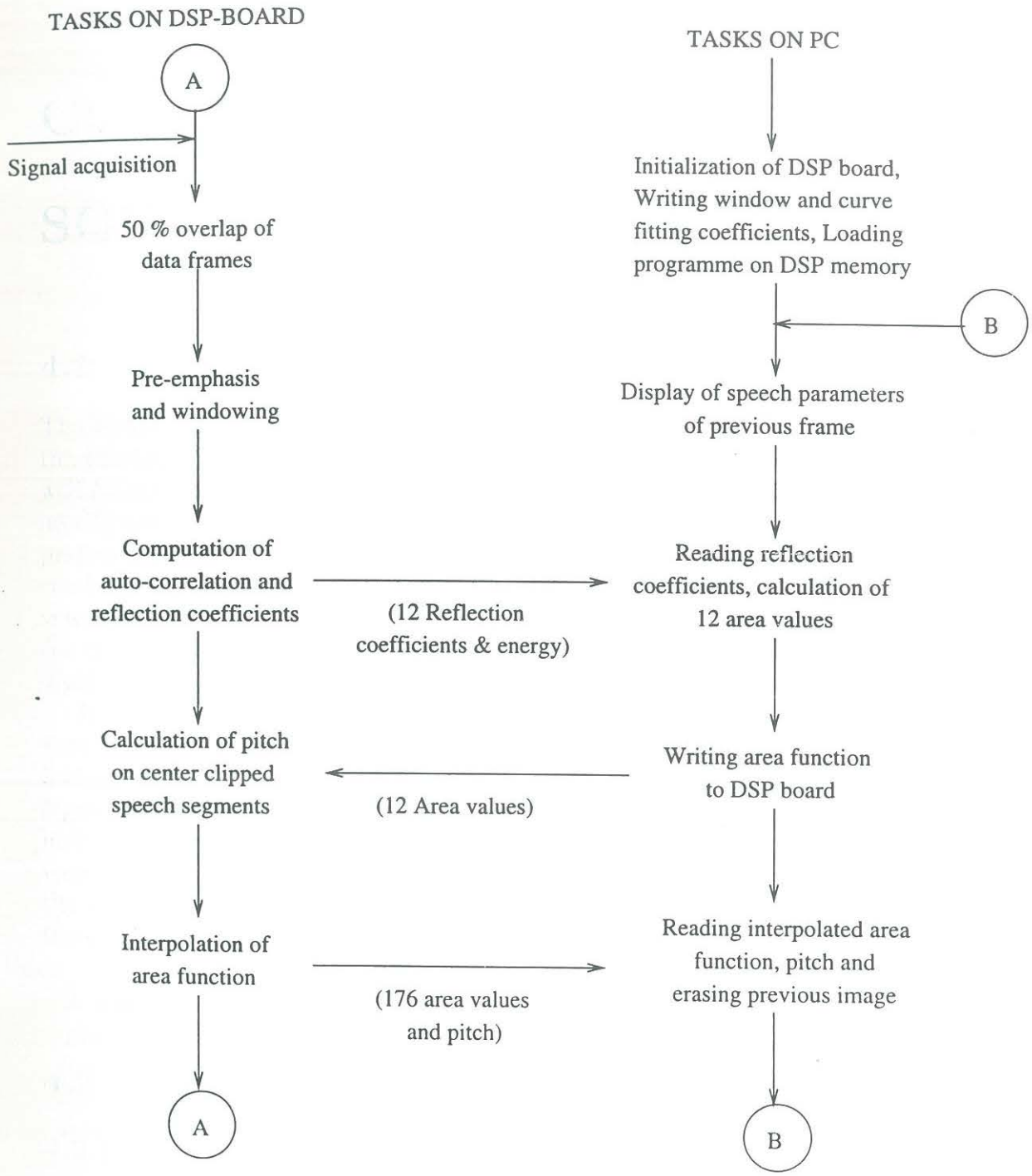

Figure 3.3: Allocation of tasks between PC and DSP-board

Chapter 4

SOFTWARE DEVELOPMENTS

4.1 Introduction

The limitations of the speech training aid STS-3 have been explained in brief in the previous chapter. It has been mentioned before that the system STS-3 works well for sustained vowels. However, testing of the system brought about the need to modify the software for areagram display even for vowels. The primary aim of this project was to devise and implement modifications in the existing system, so that it can be used for vocal tract area estimation for speech segments other than sustained vowels also. The modified software incorporates a scheme for normalizing and scaling the area values for areagram display in such a manner that erroneous peaks in area distribution do not cause severe error. This chapter explains the modification.

In order to be able to obtain areagrams for consonants, two different approaches were taken. The low pass filtering of the areagram image for interpolation was the first one. The second approach involved frame-by-frame normalization of the speech segments to improve estimation over weak energy phonemes. These are explained in the later sections of this chapter. The number of points after curve fitting of the vocal tract shape using the Bezier form algorithm is also increased after discarding the sections towards the glottis which do not show variation in the area values, and therefore are not likely to be useful for estimating the changes in vocal tract shape.

The estimation algorithm was implemented using floating point operations on the PC. The results were compared with the previously obtained results.

4.2 Normalization and Mapping of Area Values

4.2.1 Need for modification

It has already been explained that 12 area values are obtained from the reflection coefficients calculated by the DSP processor. These area values are interpolated to 176 points by curve fitting to obtain a more realistic display. The calculation of

area is done by the PC and the values are written back to the DSP board for curve fitting. For fixed point implementation of the curve fitting algorithm, all the 12 area values are normalized with respect to the maximum of them and then multiplied by a suitable constant before curve fitting. Secondly, area values are to be mapped to 0-15 grey levels for PC monitor display. This requires appropriate scaling after curve fitting.

When a number of frames are to be observed at a time in the form of an areagram, this may give a wrong picture. A larger value in one frame may get normalized to the same value as a smaller one in the next frame, because of possible difference in the maximum. Hence while plotting the areagram, maximum over all the frames is to be considered for normalization. However, the maximum area obtained cannot be relied upon for normalization. The reasons for the maximum area to be in error can be explained as follows:

1. The computation is done on a fixed point processor, and adequate care has been taken to avoid overflows. But scaling of the signal values at various stages may result in a loss of precision in the representation of values and even underflows. These result in errors in the final estimation of area values.

2. The model of speech production assumes that there are only poles in the vocal tract transfer function. This causes errors particularly during segments other than sustained vowels.

3. Analysis is pitch-asynchronous, *i.e.* positioning of analysis windows is not synchronous with respect to pitch periods. Large errors may be introduced if the number of pitch periods in a window is not nearly an integer or if the start of a pitch period falls close to one of the edges of the analysis window.

Since the maximum area obtained may be in error, it was necessary to devise a method to scale the area values before mapping to 16 grey levels, so that the erroneous maximum does not cause further error.

Another requirement was to be able to bring out the difference between the relative area distributions for vowels as seen in individual areagrams as well as in vowel sequences. If maximum area obtained for /a/ is more than that obtained for, say /i/, the gradation of the areagram should distinctly show this. For this, it is necessary to determine possible maximum area for a particular subject, which can be used as a reference for normalization of the area values obtained for any phoneme uttered by that subject. The maximum area and variation of area along the vocal tract changes from subject to subject. As stated before, because of possibility of presence of erroneous peaks in the area values, the absolute maximum value cannot be relied upon. Hence, a method needs to be found out, which for a given subject, irrespective of the phoneme uttered, will scale the area values without losing any information, will take care of erroneous peaks, if any, and give improvement in the areagram display.

4.2.2 Modification implemented

In order to solve the above mentioned problem, area values obtained for different vowel utterances by different subjects were studied for the range, maximum, and mean values for one frame and also over a number of frames. It was observed that though the range of the area values is quite large, most of the information is present in the low area region, i.e., the mean value of the area is very small as compared to the maximum area value. This observation holds good even after ignoring the erroneous peaks which may be present and are very small in number. Square-root and logarithmic mapping was tried in order to expand the small area region and compress the large area one, out of which square root mapping showed improvement.

The scaling constant used for an areagram for vowel sequence /aiu/ can be used for individual vowels also. The area distributions for 100 captured frames for vowels and vowel sequence utterances by several subjects were studied and a method is suggested for normalization and mapping of area values. In the modified method following steps need to be done:

- The subject to be trained would be asked to utter a sequence /aiu/. The trainer would need to capture this suitably in 1.28 seconds and carry the histographic analysis, which is incorporated in the software.
- The scaling procedure involves looking at that area value below which 99 % of the total area values lie. This area value can be safely assumed to be the maximum for that subject and can be used to normalize all the area values irrespective of the vowel uttered. The 99 % criterion was decided after looking at distribution of area values for various vowel utterances by different subjects.
- The areagram will be plotted in review mode using a software which takes square root of the area values and then scales all the values appropriately before mapping to 0-15 grey levels.

4.3 2-D Spatial Interpolation of Areagram

For obtaining the areagrams for stop consonants, it was decided to make use of the area variation available just before and after the closure duration. Some kind of interpolation was necessary for extracting the missing information during closure. Low pass filtering is the simplest form of interpolation that can be applied to two dimensional data. Hence it was decided to use spatial low pass filtering on the areagrams for the vowel-consonant-vowel (V-C-V) sequences to bring about interpolation. If a square region of $n \times n$ (n odd) values is considered, the central value is replaced by the average of all the values in that region [18]. This was tried on a number of areagrams for different V-C-V sequences like /apa/, /ata/, /aka/ *etc.* with various mask sizes. In order to obtain considerable amount of interpolation, it was necessary to use a

mask of much large order *i.e.* 9×9 or 11×11. It was observed that, this kind of interpolation does not give the desired result. The information that is desired to be extracted from the areagram is the place and manner of articulation. For example, if we consider the utterances /apa/ and /ata/, then after low pass filtering the place of articulation should get distinctly reflected in the filtered areagrams. But this method of interpolation fails to do so. This indicates that the information which is missing because of weak speech signal, cannot be easily recovered using the data available just before and after the stop. Low pass filtering with a mask of size 5×5 was found suitable to smooth out the image and to suppress any possible erroneous peaks in the calculated values of area.

4.4 Computation Over Normalized Speech Segments

If it is possible to amplify the signal during low energy speech sequences, then some improvement could be possible for estimation. However, this projects the necessity of selective amplification, as for a V-C-V sequence, amplification during vowel section is not needed, but is essential during consonant section. This can be implemented by selectively scaling the digitized sample values.

As explained earlier, the processing is done on the speech data block by block, each block being of length 256 samples. Hence, the scaling for any particular block must be done by such a factor that very low valued samples are amplified to sufficiently large levels, and negligibly small number of samples get saturated. For this, rms value of the frame can be considered as the criterion. However, calculation of rms value involves taking square root. This is a time consuming process and should be avoided in real-time operation. Hence, mean magnitude over a frame is used for scaling. It was empirically found out that after normalization, average magnitude should equal half the integer value representable by 16 bit signed number, in order to get desired result.

4.5 Modification in Curve Fitting

The vocal tract area function is obtained from the reflection coefficients as area values for 12 sections. For getting a more realistic display of the vocal tract shape, Baragi [7] tested and implemented Bezier curve fitting for interpolation area contour. By this, 12 area values were interpolated to 176 values before displaying on the monitor. After studying area distribution for various vowels, vowel sequences, and V-C-V sequences, almost no change was observed in area values near the glottis end. Displaying the vocal tract length with no changes in its shape is not likely to contribute in speech training. Hence, it was decided to discard these sections and increase the number of points in curve fitting for the remaining part. It was noticed that 2 sections at the glottis end can be safely discarded, as there is no change present in that region. The

total number of interpolated points after curve fitting is still kept to be 176, thus achieving a stretch for the rest of the region.

4.6 Computation Using Floating Point Arithmatic

It has been explained that the estimation of pitch and vocal tract area becomes unreliable during low energy speech segments like semi-vowels and consonants. As it was observed that, simple spatial interpolation of the areagram image is not sufficient for obtaining correct area distributions for stop consonants, a different approach is needed to solve the problem. In STS-3, the computation is done using fixed point arithmatic on the DSP processor. In the implementation, care has been taken to avoid overflows especially in the computation of autocorrelation coefficients. This leads to loss of data during computation over weak energy segments. It was suggested that use of floating point arithmatic would improve the resolution and reduce the loss of data. A C programme was written for estimation of vocal tract shape and was tested off-line on various speech data files for semi-vowels and consonants. The areagrams were obtained for the same and compared against the ones obtained using the previous software. The analysis results will be presented and discussed in the following chapter.

Chapter 5

RESULTS

5.1 Introduction

The system STS-3 was tested for consistency and the need for modification was felt as explained in Chapter 3. The previous chapter described various modifications implemented in the software. The hardware setup was not modified. The modified software was tested by acquiring areagrams for different vowels, semi-vowels, and V-C-V sequences uttered by various subjects. The results of the comparisons with the relevant areagrams are included in this chapter. The system with the final modification is called STS-4 and its operational details are provided in Appendix B.

5.2 Modification in Areagram Display

As explained in Chapter 4, the software for areagram display was modified for the following:

1. taking square root of area values before mapping to grey levels
2. normalizing all the area values with respect to the maximum value obtained over 100 frames to be displayed
3. using user defined maximum which can be treated as the maximum possible area for the subject under test

While interpreting areagram, a look at the corresponding spectrogram may be of interest. In the wide band spectrogram, tracks of formant frequencies can be seen. The first two formant frequencies have relatively direct relationship with articulatory parameters [15].

- F1 decreases as the opening at the place of maximum constriction decreases.

- F2 decreases as the place of maximum constriction shifts backwards, thus indicating the position of the constriction.

Fig. 5.1 shows the areagram for the natural vowel sequence /aiu/ (author's voice), and the corresponding spectrogram is shown in (*Fig.* 5.2).

Here, /i/ which has higher F2 than /a/, indicates constriction more towards lips. While uttering /u/, tongue is shifted backwards, thus resulting in a lower F2. Similarly, lower F1 indicates lower area at the constriction, in case of /i/. These observations can be verified by looking at the corresponding areagram.

For comparison, two areagrams can be seen in *Fig.* 5.3 and 5.4. The areagrams are obtained for vowel sequence /aiu/, uttered by a male speaker. The acquired signal has been processed for obtaining the areagrams separately. *Fig.* 5.3 shows areagram obtained using STS-3. Because of the normalization procedure used before, the maximum area displayed during the utterance of /a/, /i/, and /u/ appear the same, while we know that, these are different for the three vowels. The areagram as obtained by using modified normalization and square root mapping is shown in *Fig.* 5.4, and we notice an improvement in representation of area variations.

For overall comparison, areagrams and spectrogram for synthetic vowel sequence /uai/ generated using a cascade formant synthesizer [19] with pitch of 125 Hz are shown in *Fig.* 5.5. Thie figure shows the areagram obtained using STS-3, the one using modified software, and the spectrogram from top to bottom respectively.

5.3 Spatial Filtering of Areagrams

As explained before, 2-D spatial interpolation of the areagram matrix was suggested as a possible solution to the extraction of missing information during stop closures. In the areagram image, the frame spacing is 12.8 ms, and the length of the vocal tract is represented as 176 points (after interpolation from 12 area values). Along the x-axis, each frame is represented by 5 pixels, but the 2-D interpolation was done treating data for each frame as one point. Though, certain amount of interpolation was possible, this failed to give desired results. *Fig.* 5.6 and 5.7 show filtered areagrams for /apa/ and /ata/ respectively. To obtain these, area frames for which energy is below the threshold were suppressed to zero. Then 2-D spatial filtering was carried out using a mask size of 9×9. After comparison between the two areagrams, it is not possible to extract much information regarding the place of articulation for the two stop consonants.

5.4 Modification in Curve Fitting

As explained in Section 4.5, there is almost no change in the area values near glottis end. Hence, these are not important for training purpose, and can be safely discarded.

The curve fitting is modified in such a way that two area values near glottis end are discarded, keeping total number of points for interpolation the same. *Fig.* 5.8 and *Fig.* 5.9 show the areagrams obtained using previous and modified curve fitting respectively, for the synthetic vowel sequence /uai/.

5.5 Normalizing Speech Segments

Another modification in the software was done by normalizing the speech segments before any processing. This not only reduced the minimum signal level required for satisfactory estimation, but also made it possible to obtain areagrams for semi-vowels which was not possible before. The minimum level of speech signal necessary for satisfactory estimation has come down to 0.6-0.7 V from 1.3 V for vowels. Since the normalization involves finding out the mean magnitude of any given frame, and dividing each and every signal sample by twice this value, it was necessary to investigate the real time operation of entire system after incorporating this. It was ensured that the system still operates in real time.

Areagrams for semi-vowels /aya/ and /awa/ as obtained after normalization are given in the figures. *Fig.* 5.10 shows the areagram, pitch, and energy display for /aya/ as obtained using STS-3. *Fig.* 5.11 shows the areagram as obtained using modified software. *Fig.* 5.12 and 5.13 show the corresponding areagrams for semi-vowel /awa/. In case of V-C-V sequence also, improvement was investigated. The criterion used for this is the number of frames for which the estimation is erratic. Table 5.1 summarizes the observation. The results obtained with the estimation using off-line implementation using floating point arithmatic are also included here for comparison. Number of frames for which the estimation was unreliable, during the V-C-V utterance of a male and a female speaker for

(a) Earlier real-time implementation STS-3
(b) Off-line floating point computation FPC
(c) Modified real-time implementation STS-4

5.6 Testing of STS-4 on Variable Pitch Segments

In order to study the operation of STS-4 for pitch estimation, speech segments with variable pitch were used. *Fig.* 5.14 and *Fig.* 5.15 show areagrams obtained for vowel segment /a/ uttered by a male and a female speaker respectively. Corresponding spectrograms are also provided. It can be concluded that, variations in pitch of the input speech are approrpiately estimated, and estimation of vocal tract area is not affected by change in pitch.

Table 5.1: Comparison between different schemes

V-C-V by female speaker	STS-3	FPC	STS-4
/apa/	17	13	15
/ata/	18	13	14
/aka/	18	11	13
V-C-V by male speaker			
/apa/	12	8	8
/ata/	17	13	14
/aka/	12	8	10

5.7 Testing of STS-4 for Nasals

Since the vocal tract model is assumed to be an all-pole model, it is valid for vowels. For modelling the vocal tract during the utterance of nasals, resonances associated with nasal cavity are also to be considered, and hence all-pole model is not sufficient. Areagrams for nasals were obtained using STS-4, and the results are shown in *Fig.* 5.16, *Fig.* 5.17, *Fig.* 5.18, and *Fig.* 5.19, which show areagrams and spectrograms for sequence /ama/ and /ana/ for a male and a female speaker. The areagrams do not give information directly regarding the place of closure, or manner of articulation. However, these can be useful for extraction of this information.

5.8 Testing of STS-4 for Unvoiced Segments

The estimation was also made for whispers, sounds with unvoiced excitation at the glottis. This means, the vocal tract shape is maintained the same as for sustained vowel, but the excitation is unvoiced as in the sequences /aha/, /ihi/, and /uhu/. *Fig.* 5.20 and *Fig.* 5.21 show these for /aha/ uttered by a male and a female speaker respectively. Corresponding areagrams and spectrograms for sequences /ihi/ and /uhu/ are shown in *Fig.* 5.22 and *Fig.* 5.23, and *Fig.* 5.24 and *Fig.* 5.25. It can be seen that, the estimation can be done over unvoiced speech segments also. It can further be observed that estimation is less meaningful at the voiced-unvoiced boundaries of the sequences. Since the estimation is done after frame-by-frame normalization of the sampled speech signal, the frames at the voiced-unvoiced boundaries have more mean magnitude value, and hence are scaled upwards by a smaller factor. On the other hand, the frames with unvoiced excitation result in lower mean magnitude, and are scaled by a larger factor, resulting in more meaningful estimates.

5.9 Discussion

The modifications implemented in STS-3 and the results obtained thereby have been presented in the previous sections. Modifications leading to definite improvement in the overall performance of the system are incorporated in STS-4. These are summarized here.

1. Modification in normalization/scaling of area values for areagram display has resulted in improvement for presentation of results for vowels. All the area values are now normalized with respect to the maximum over entire set of frames to be displayed, against frame-by-frame normalization in STS-3. Square root mapping is used after normalization, in order to represent area by 16 grey levels on the PC monitor.

2. After observing a number of areagrams for different vowel utterances, very less variation in the area values near glottis end was found. The software for curve fitting was modified to discard two area values near glottis end. The total number of points to represent the vocal tract is kept the same, and thus stretching is achieved, eliminating the information which is not useful for speech training.

3. Normalization of speech signal samples with respect to mean magnitude over that data block has brought down the minimum signal level requirement of the system. This has also resulted in meaningful estimation for semi-vowels. The system STS-4 is ensured to operate in real time after incorporating this modification.

4. 2-D low pass filtering of the areagram image using an averaging mask, though not helpful in improving aregram display for consonants, has been incorporated to impart smoothness to the areagram display.

Implementing algorithm in floating point arithmatic is not possible for real-time operation with the existing hardware. It has been seen that normalization of the speech segments for real-time processing with fixed point arithmatic gives analysis results comparable to those obtained with the off-line processing with floating point arithmatic. Therefore floating point implementation may not be considered, unless it becomes necessary for extracting the information regarding the place of closure.

Testing of STS-4 was also carried out using vowels with varying pitch, whispers, and nasals. For vowels with variable pitch and whispers, the system provides satisfactory estimates of pitch, intensity, and vocal tract area. For nasals, useful information cannot be extracted directly from the areagrams, but may be useful by further analysis.

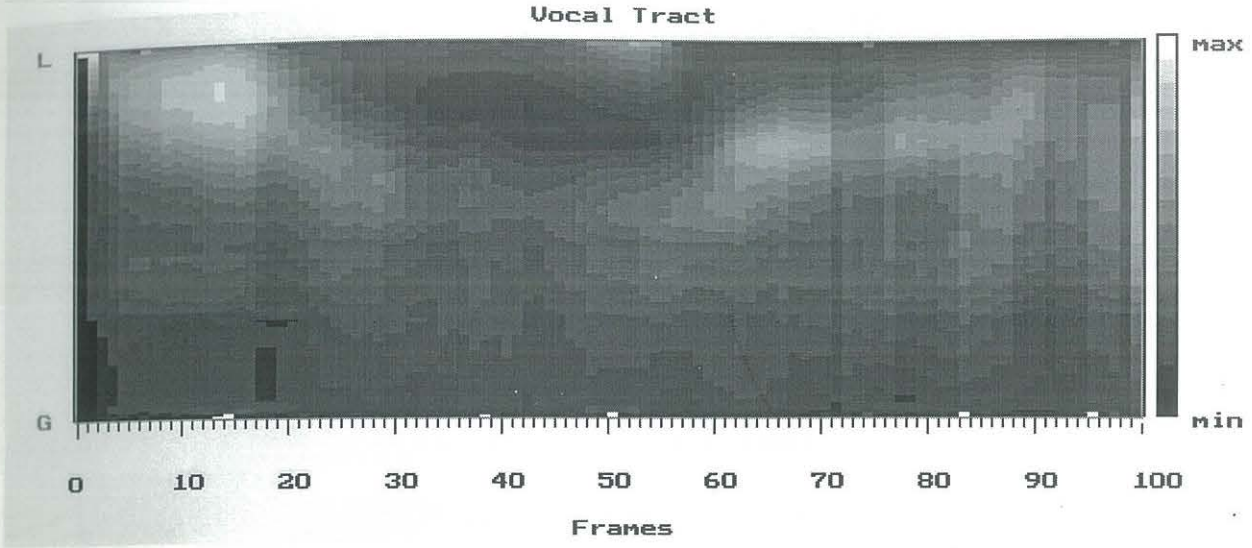

Figure 5.1: Areagram for /aiu/ uttered by female speaker

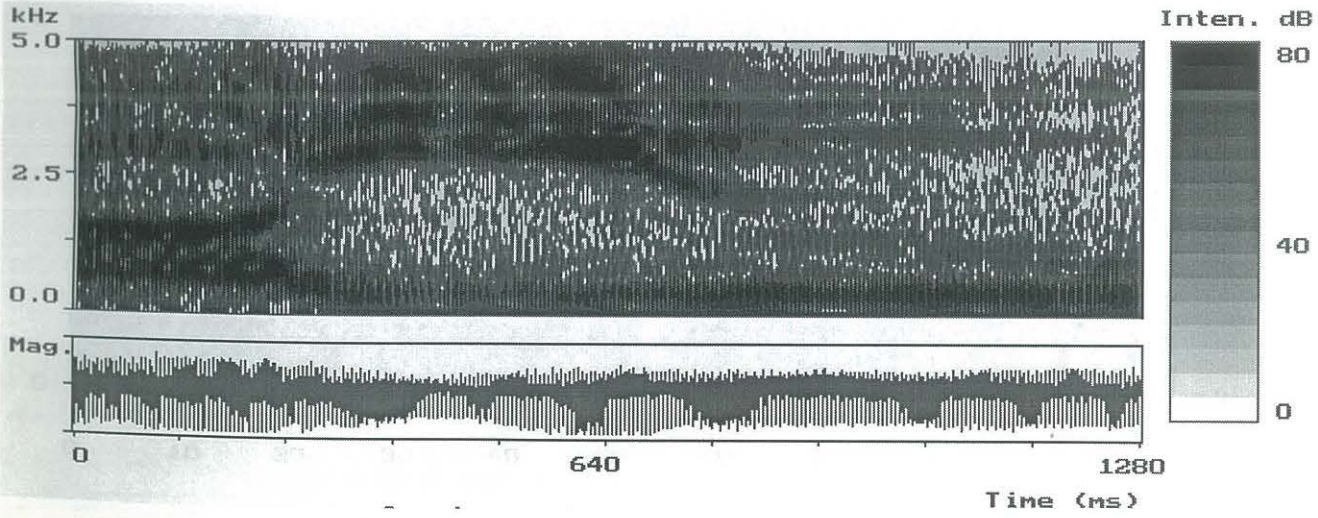

Figure 5.2: Spectrogram for /aiu/, same as that in *Fig.* 5.1

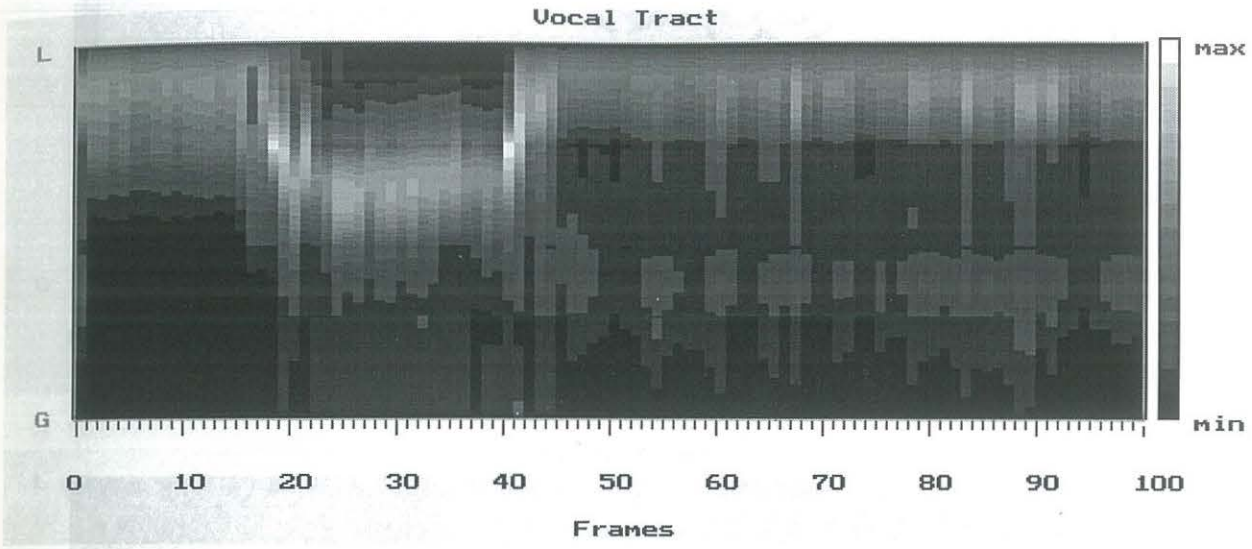

Figure 5.3: Areagram for /aiu/ uttered by male speaker as obtained using STS-3

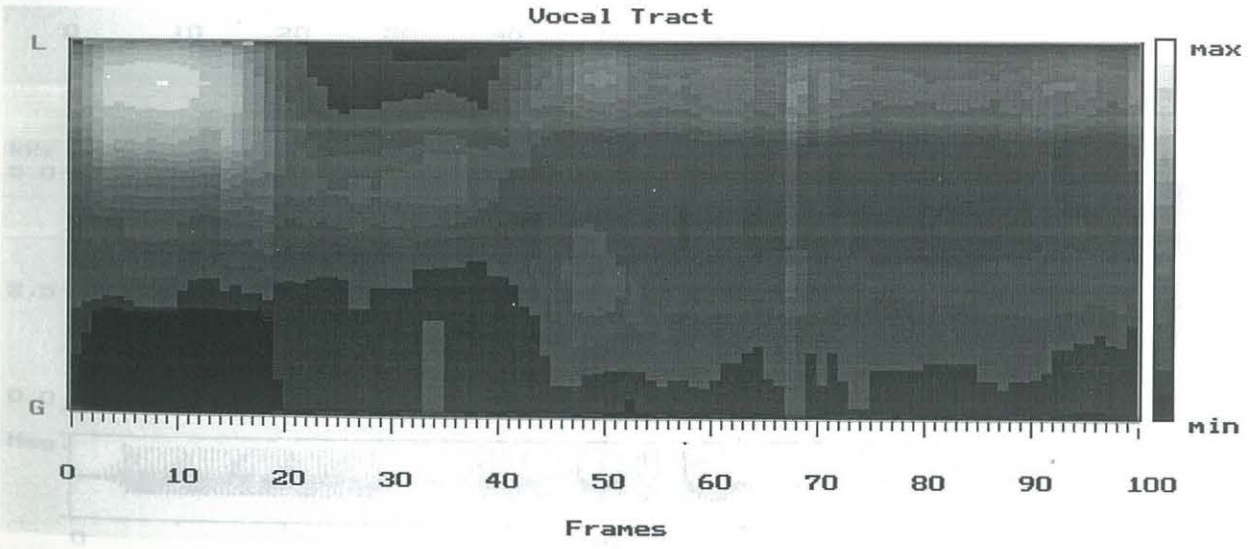

Figure 5.4: Areagram for /aiu/, same as in *Fig.* 5.3, obtained using modified software

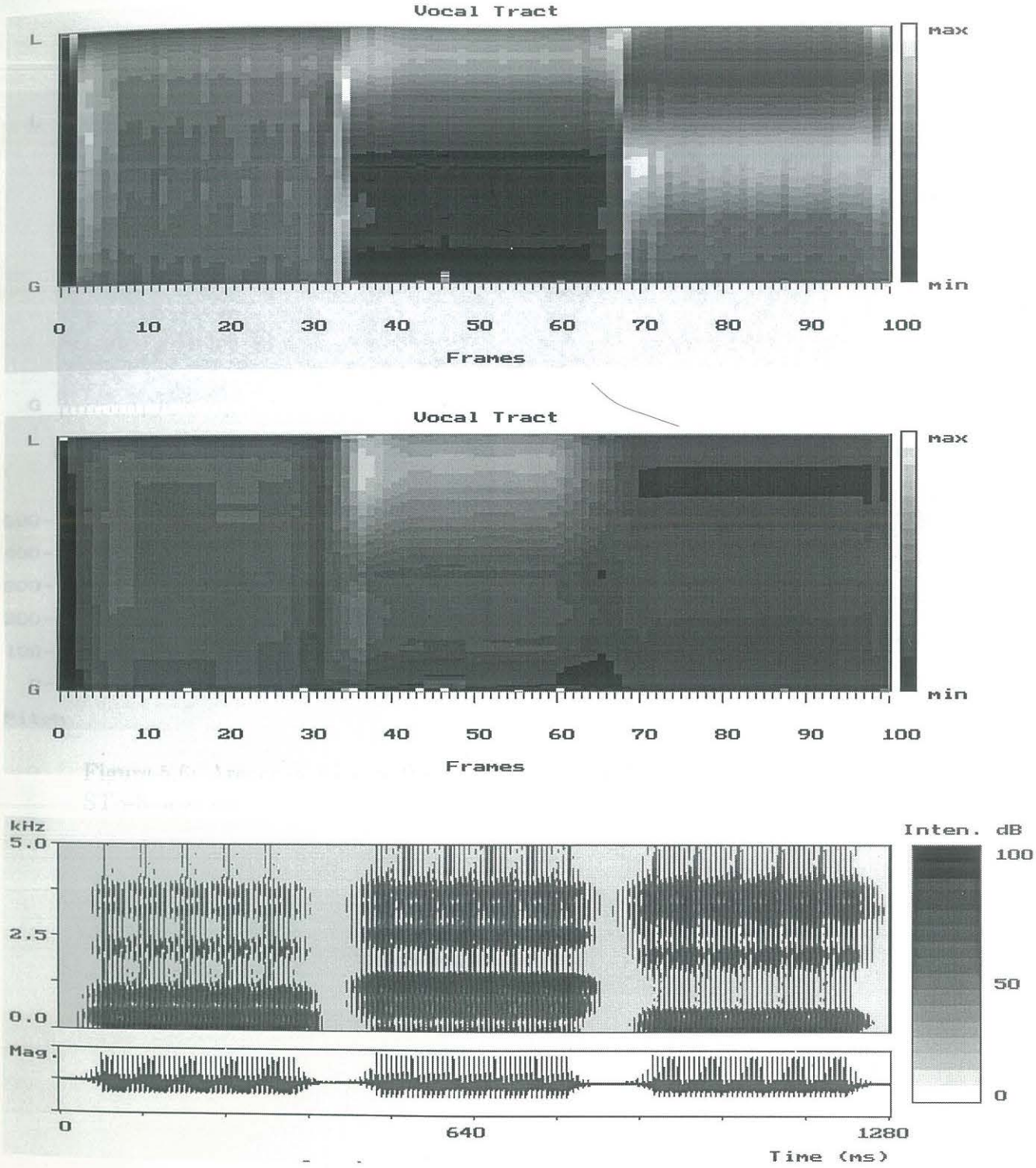

Figure 5.5: Areagram obtained using STS-3, using modified software, and spectrogram (from top to bottom) for synthetic vowel sequence /uai/ with $f_0 = 125Hz$

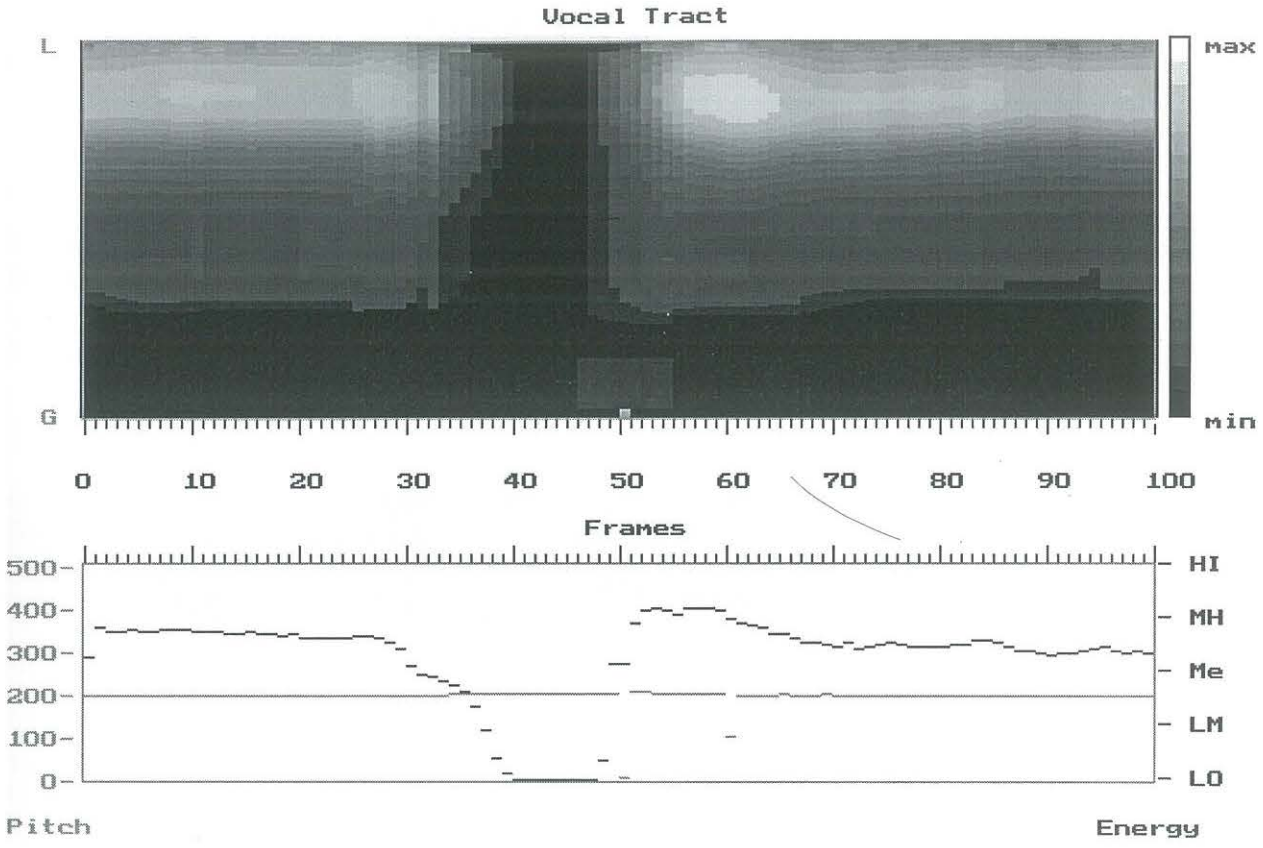

Figure 5.6: Areagram display for /apa/ uttered by female speaker, as obtained using STS-3, and filtered with a mask of size 9×9

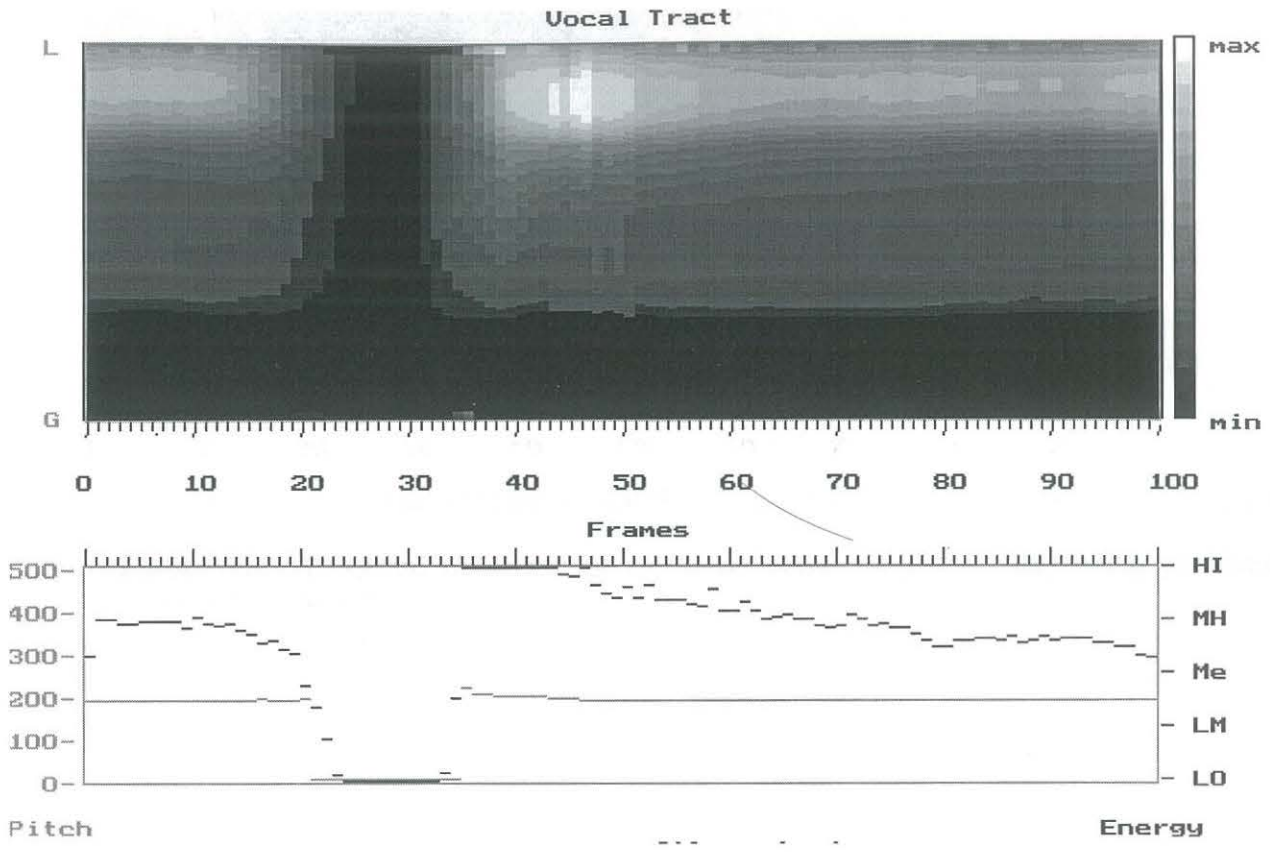

Figure 5.7: Areagram display for /ata/ uttered by female speaker, as obtained using STS-3, and filtered with a mask of size 9×9

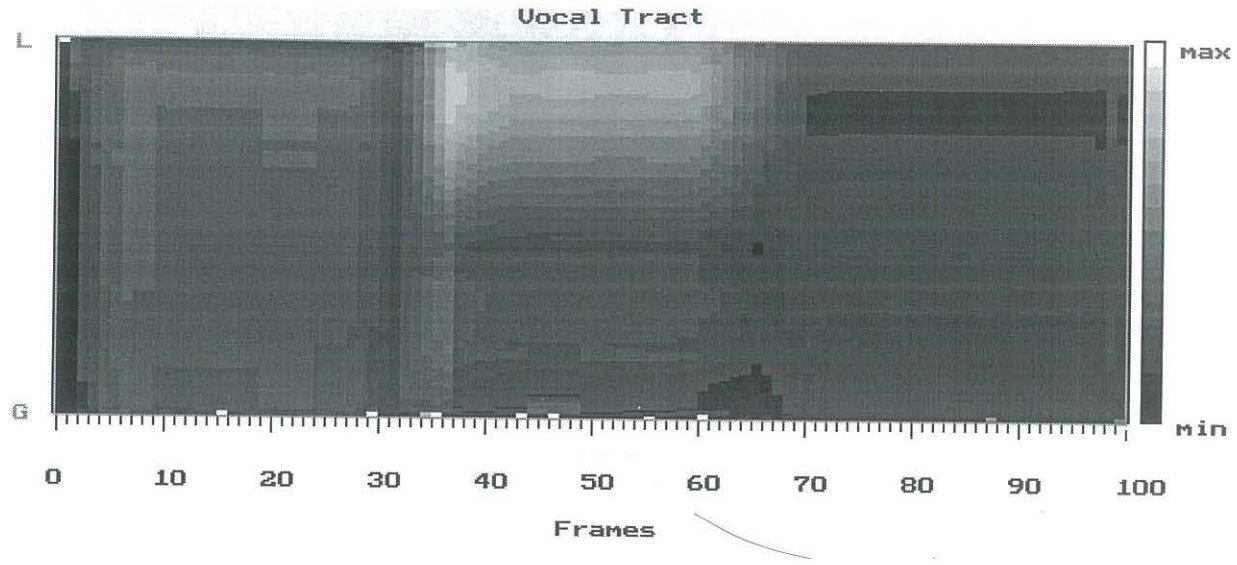

Figure 5.8: Areagram obtained for synthetic vowel sequence /uai/ using previous curve fitting in STS-3

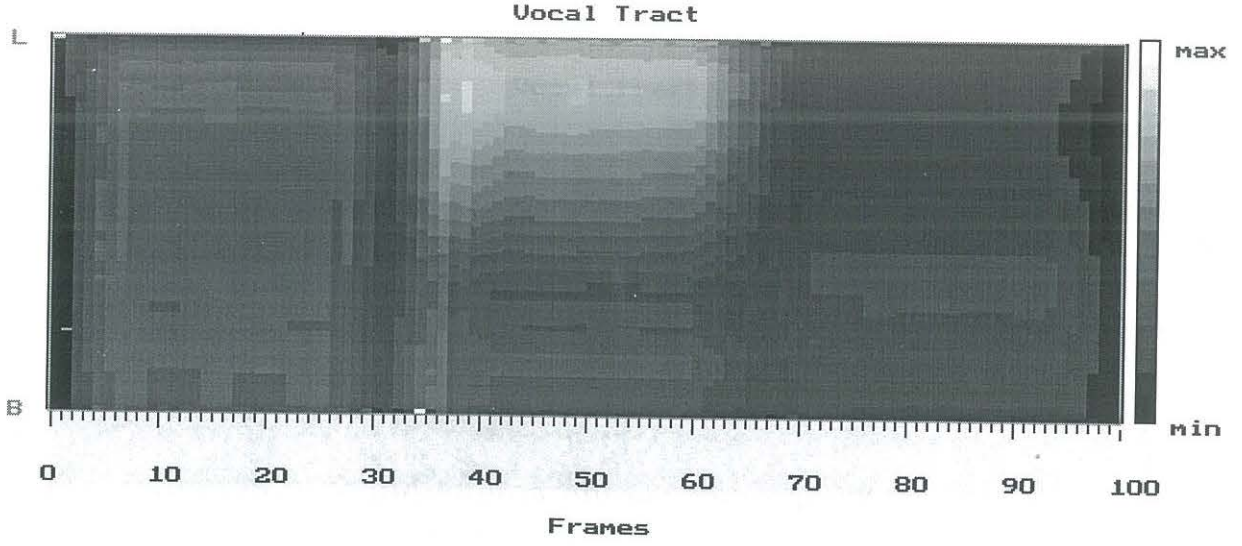

Figure 5.9: Areagram obtained for synthetic vowel sequence /uai/, same as in *Fig.* 5.8, using modified curve fitting

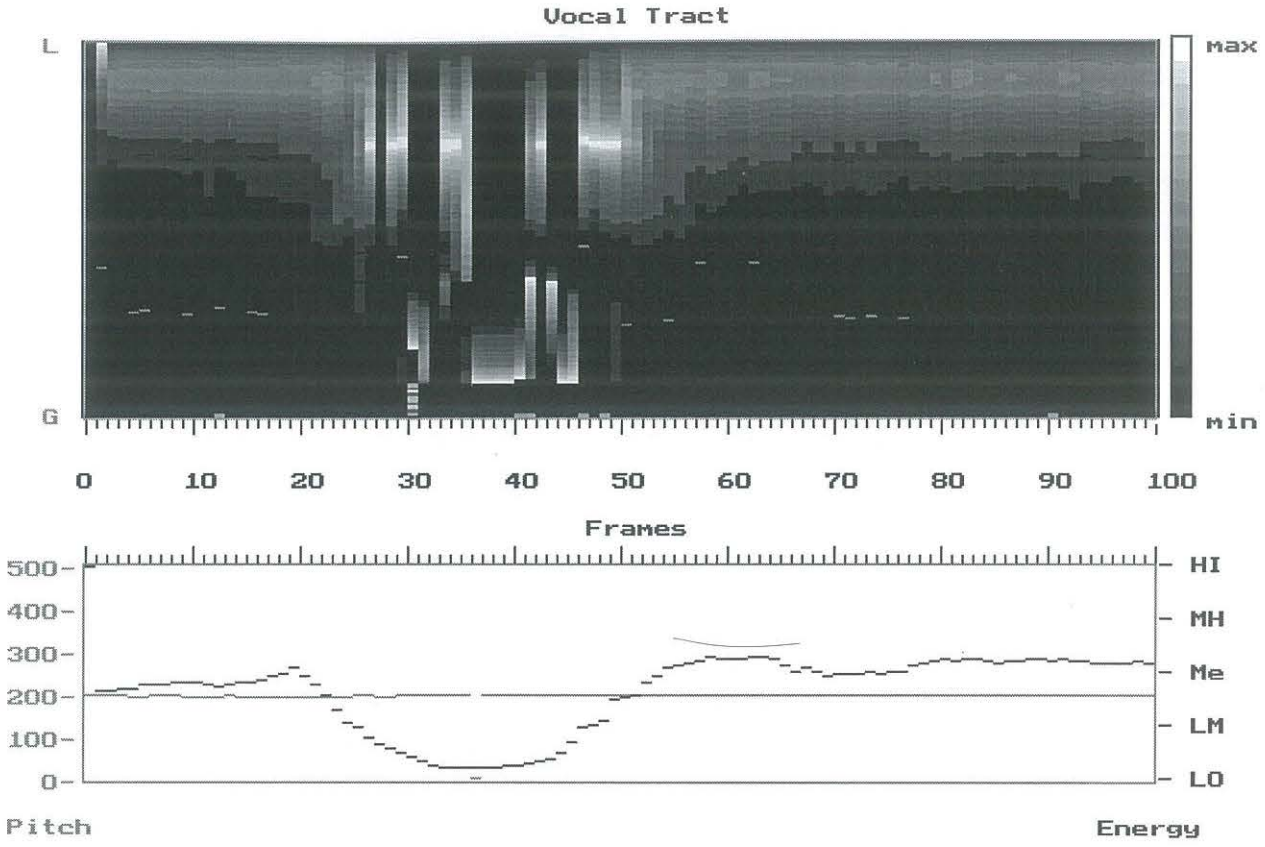

Figure 5.10: Areagram, pitch and energy display for /aya/ uttered by male speaker, as obtained using STS-3

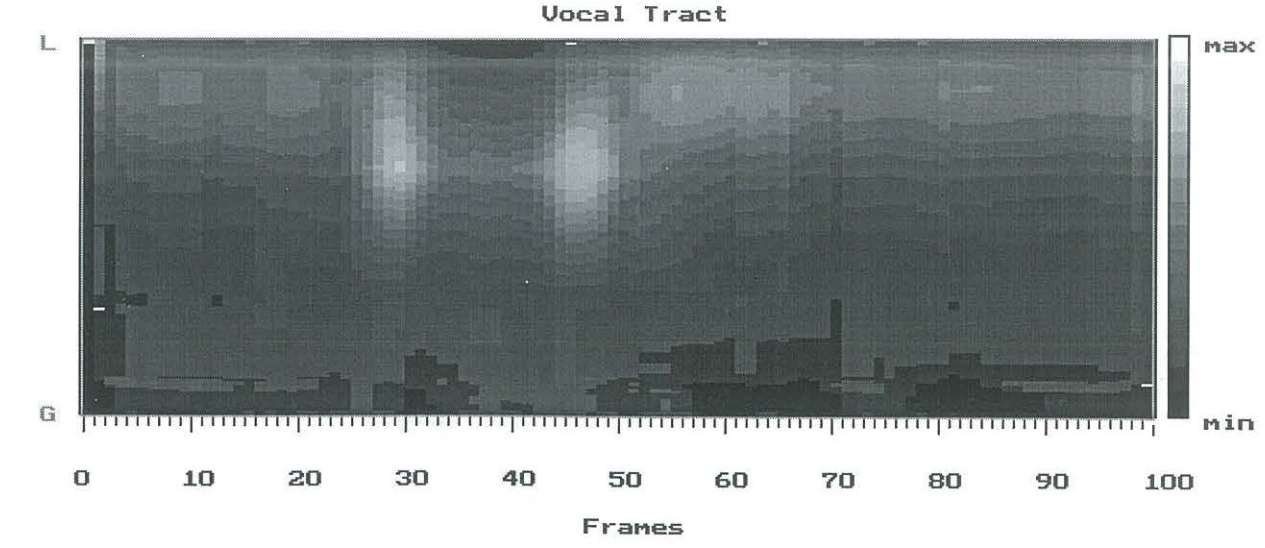

Figure 5.11: Areagram for /aya/ uttered by male speacker, as obtained using software modified for normalization of speech segments

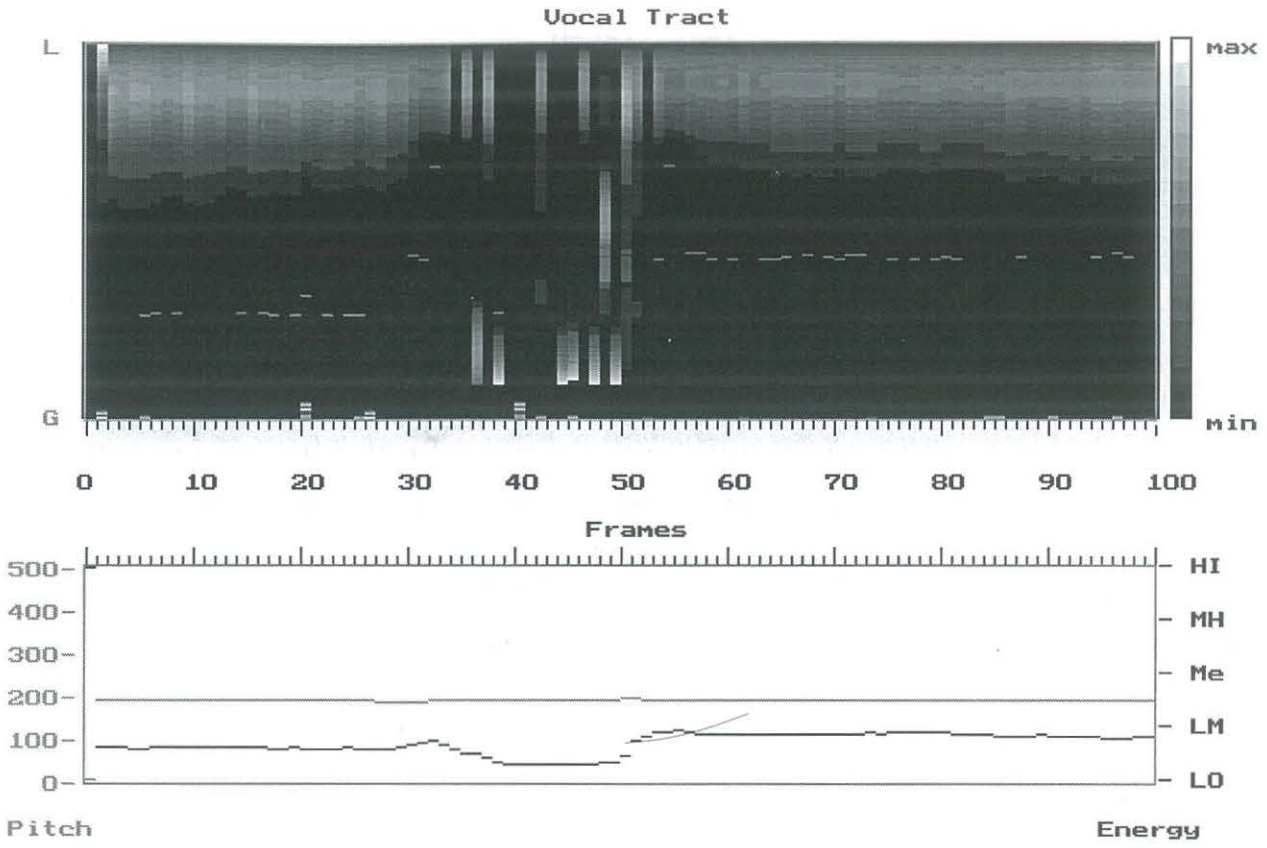

Figure 5.12: Areagram, pitch and energy display for /awa/ uttered by male speaker as obtained using STS-3

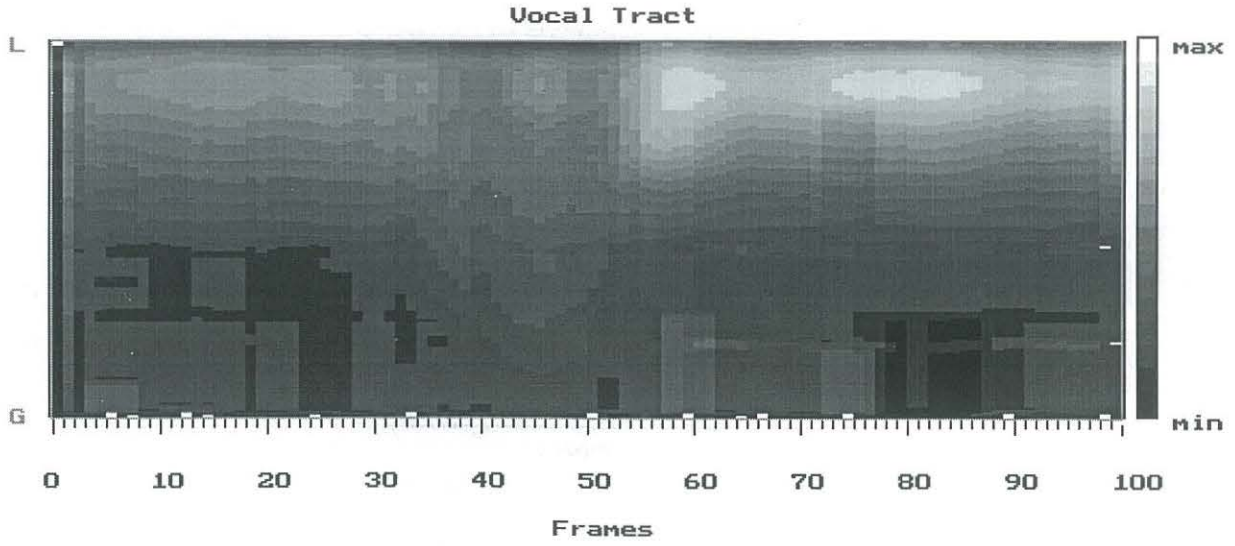

Figure 5.13: Areagram for /awa/ uttered by male speaker, as obtained using software modified for normalization of speech segments

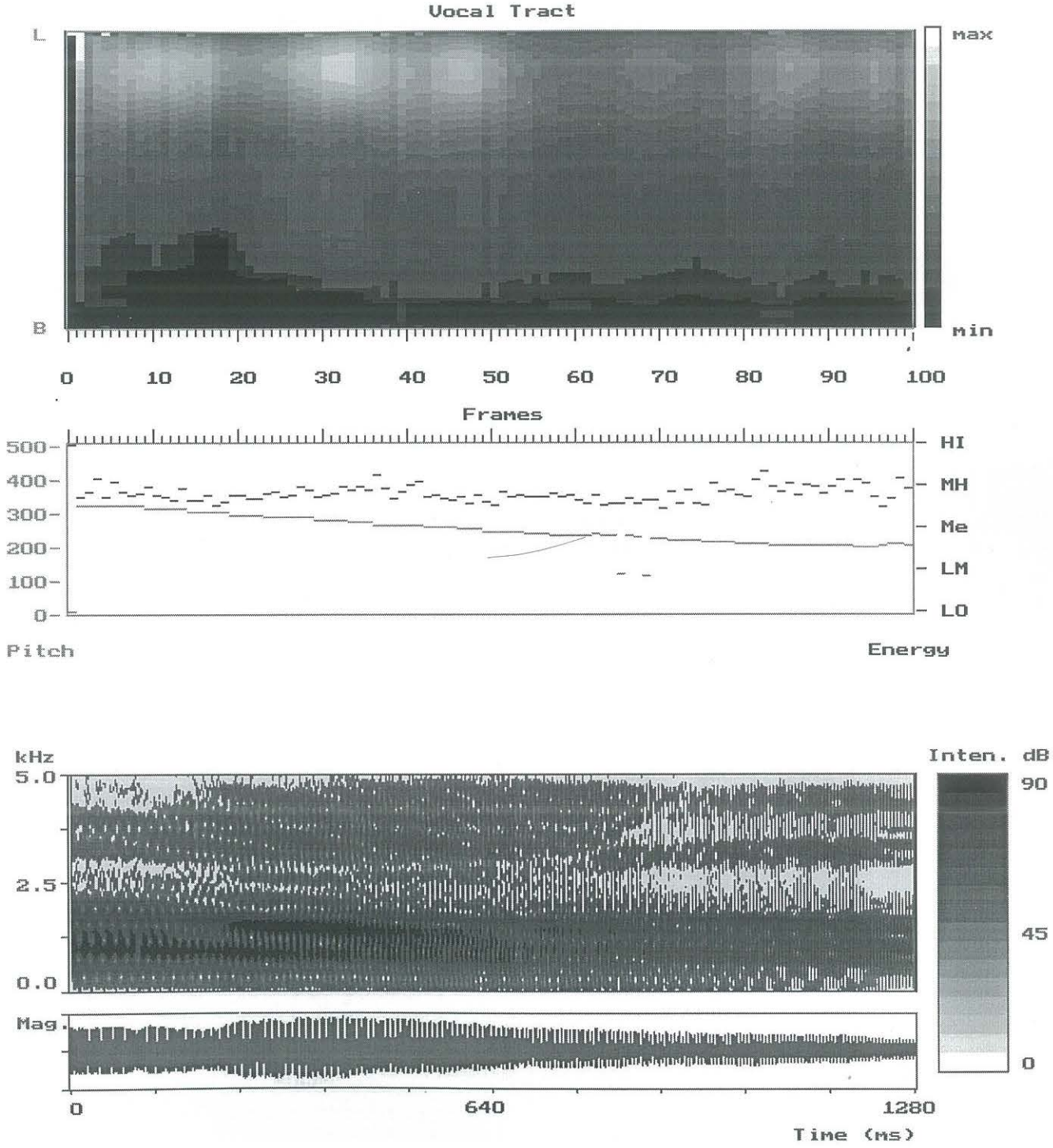

Figure 5.14: Areagram obtained using STS-4, and spectrogram for vowel /a/ with variable pitch (male speaker)

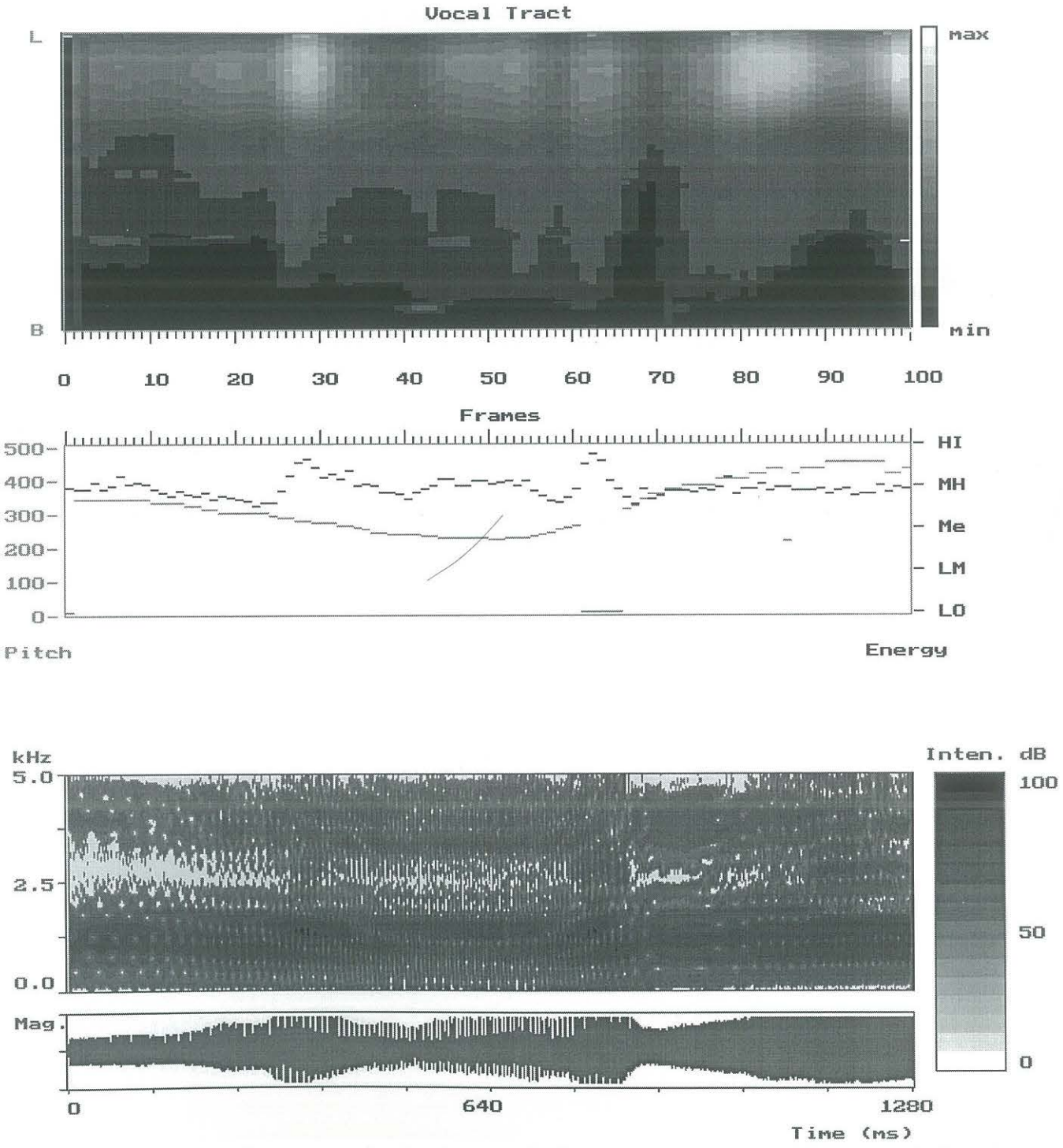

Figure 5.15: Areagram obtained using STS-4, and spectrogram for vowel /a/ with variable pitch (female speaker)

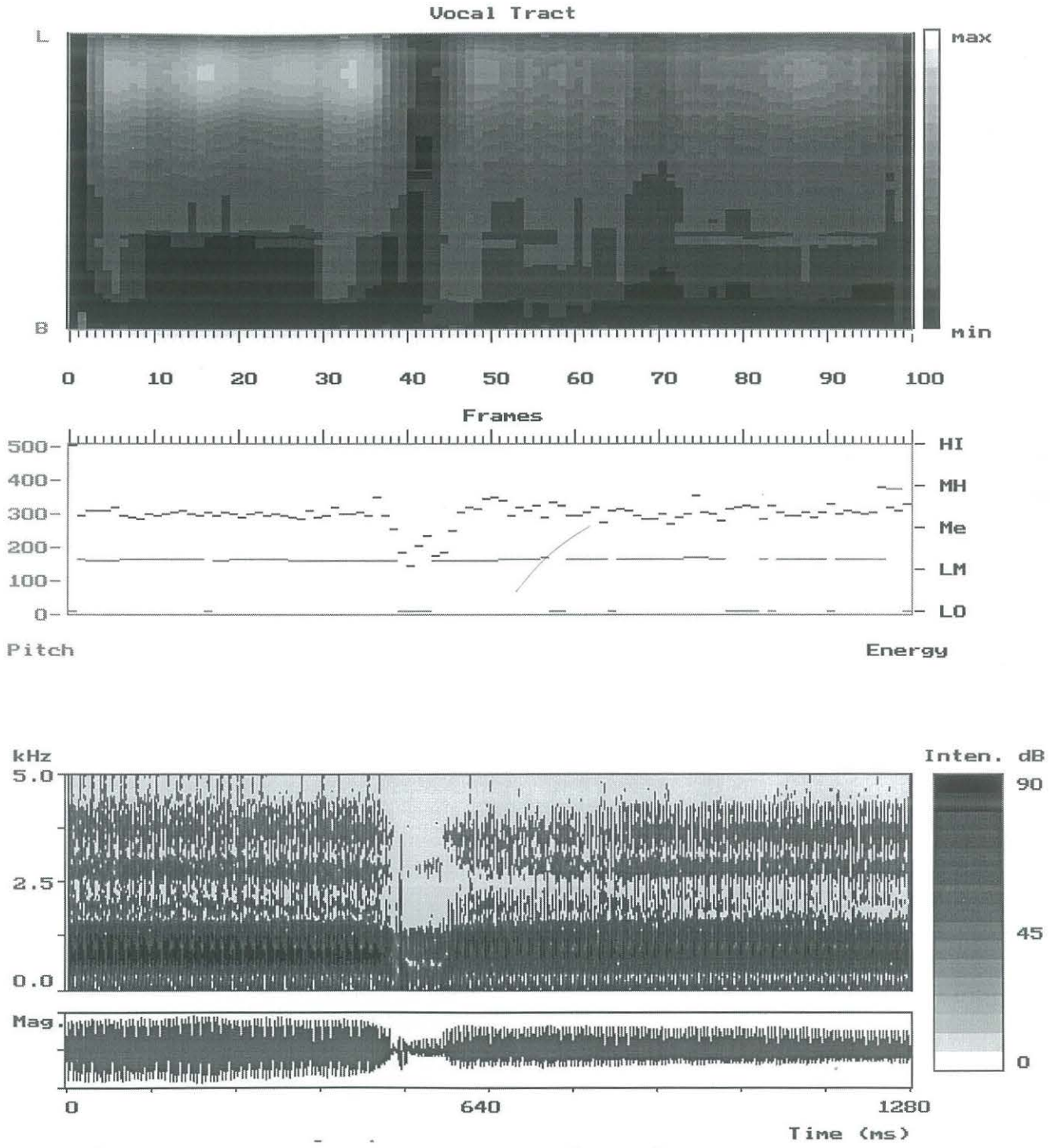

Figure 5.16: Areagram obtained using STS-4, and spectrogram for /ama/ (male speaker)

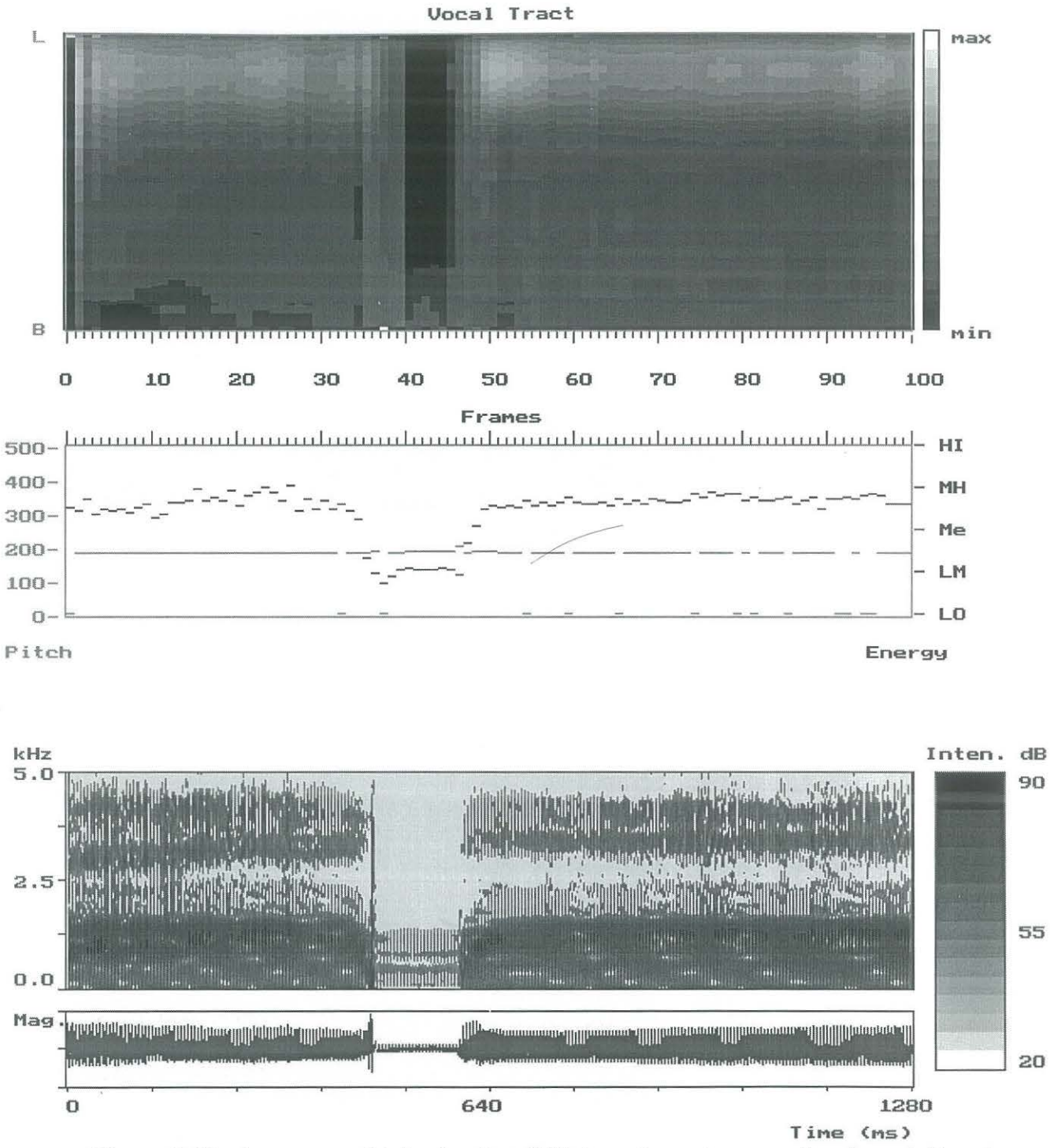

Figure 5.17: Areagram obtained using STS-4, and spectrogram for /ama/ (female speaker)

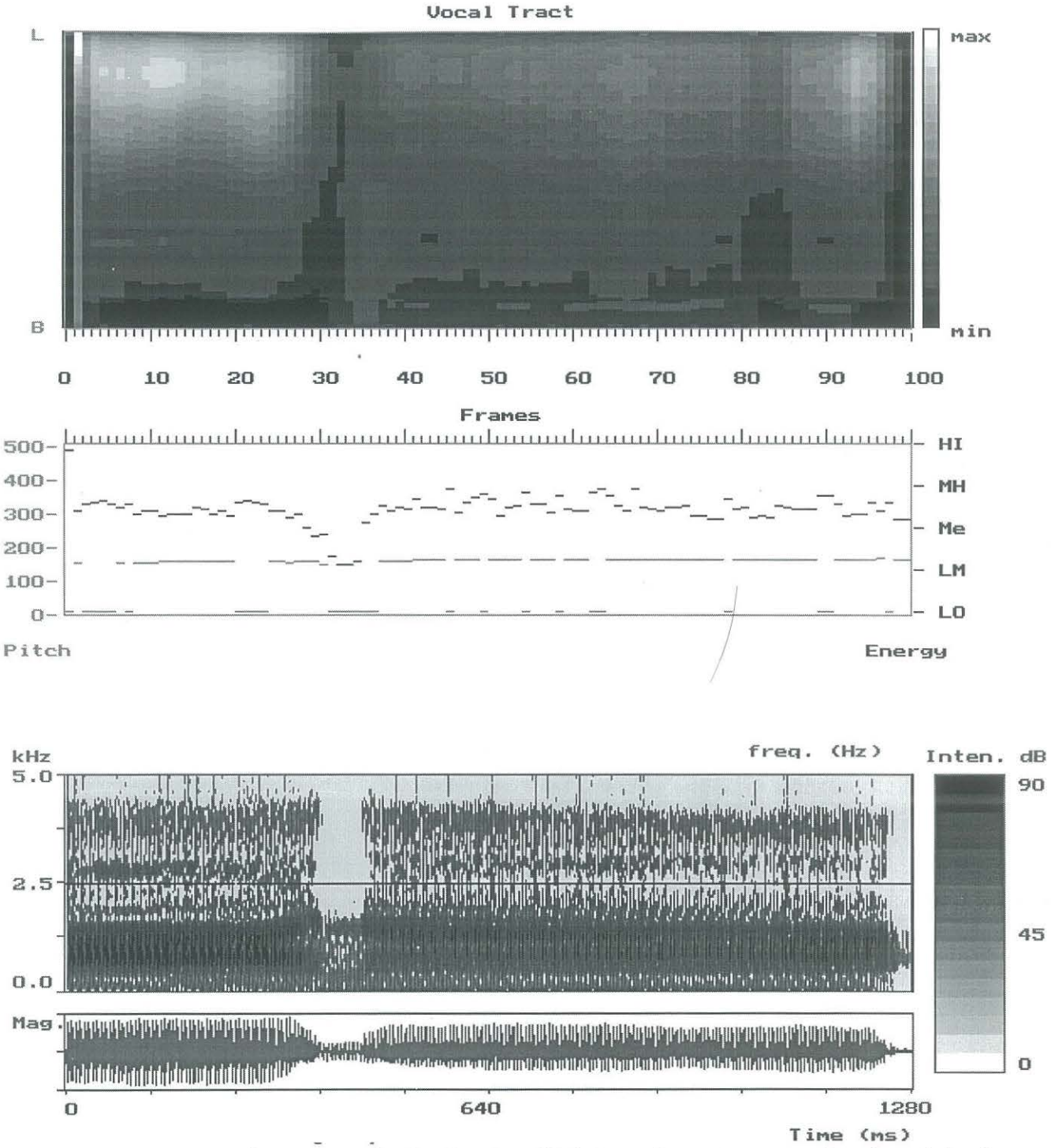

Figure 5.18: Areagram obtained using STS-4, and spectrogram for /ana/ (male speaker)

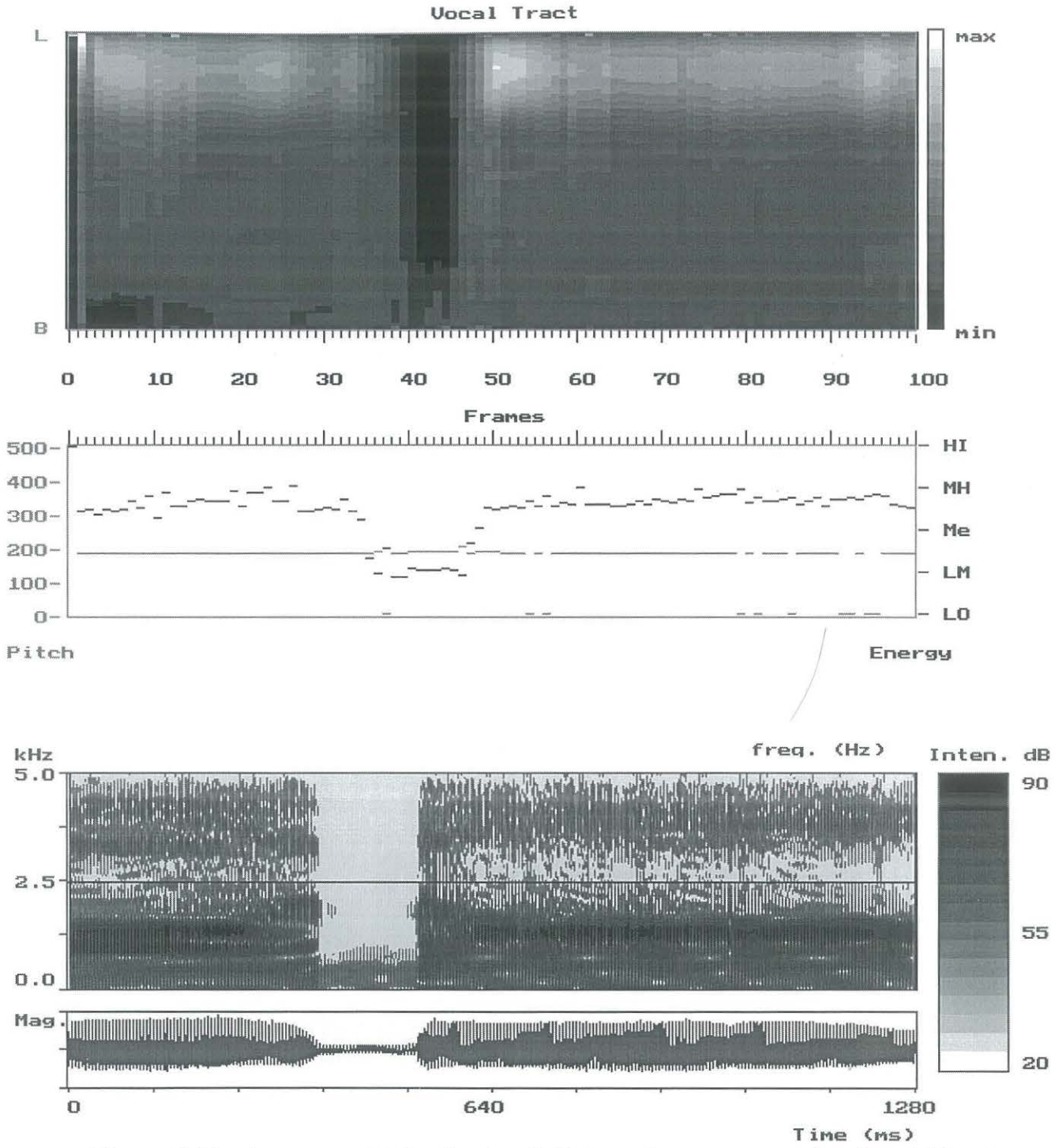

Figure 5.19: Areagram obtained using STS-4, and spectrogram for /ana/ (female speaker)

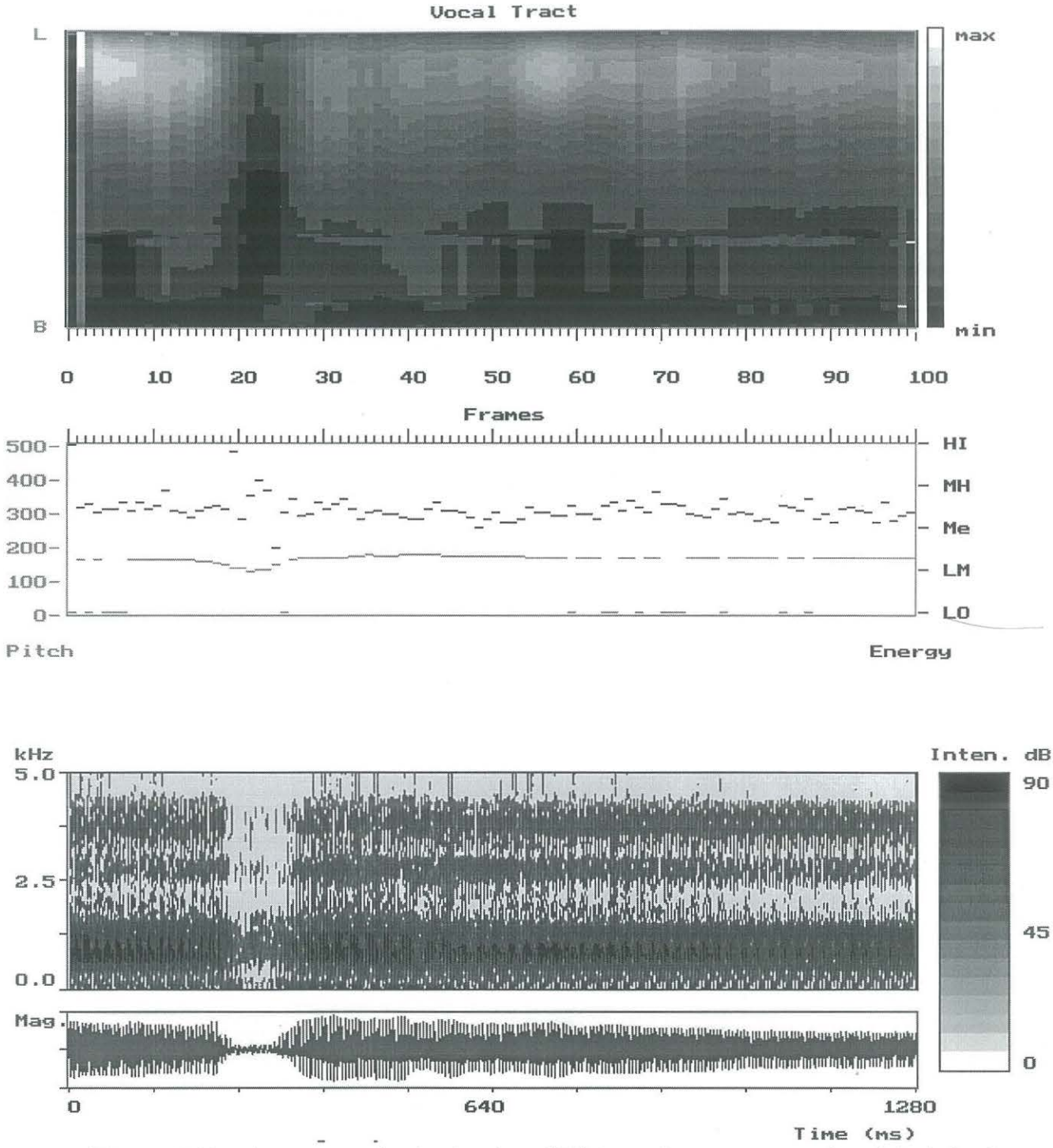

Figure 5.20: Areagram obtained using STS-4, and spectrogram for /aha/ (male speaker)

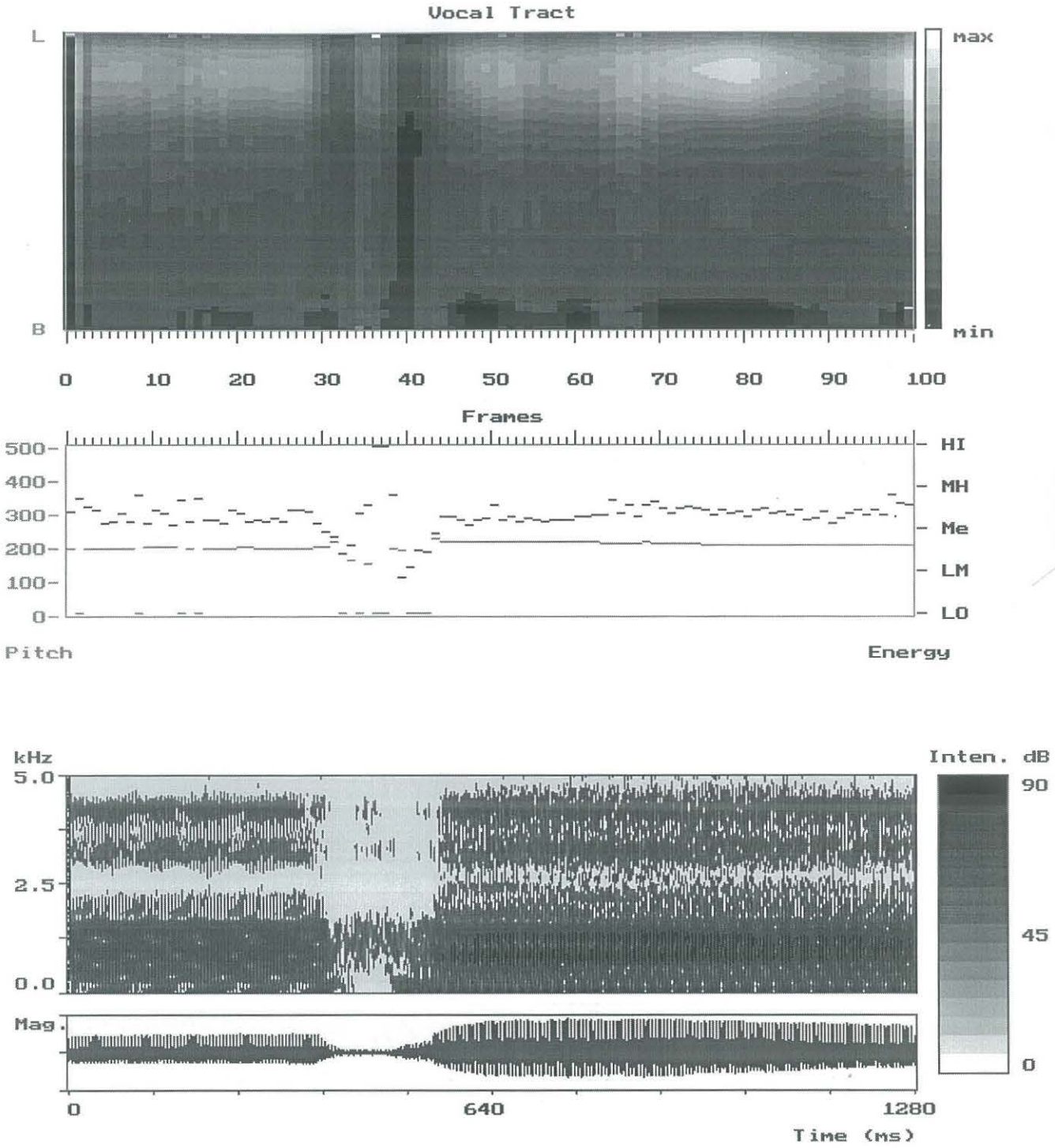

Figure 5.21: Areagram obtained using STS-4, and spectrogram for /aha/ (female speaker)

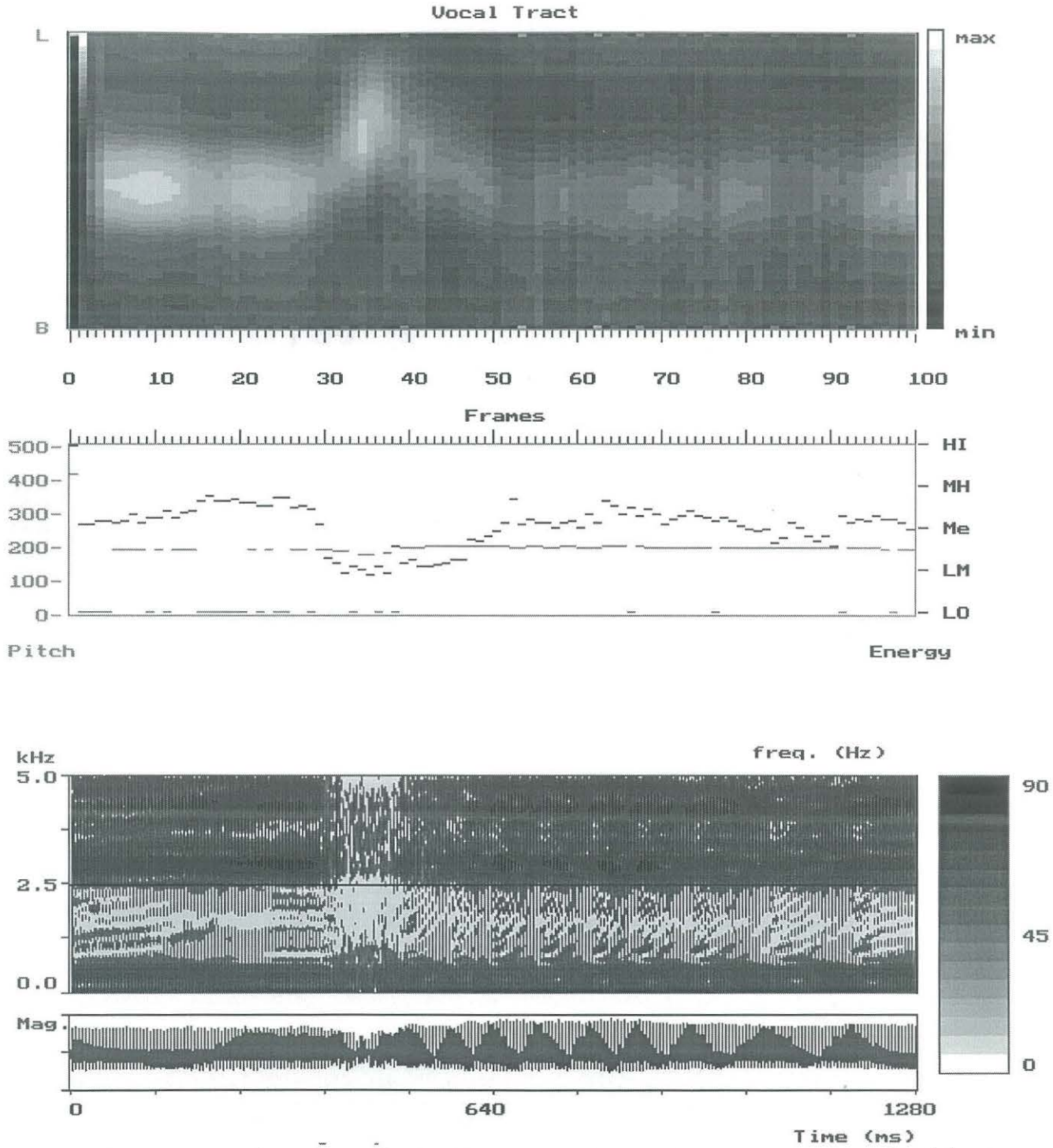

Figure 5.22: Areagram obtained using STS-4, and spectrogram for /ihi/ (male speaker)

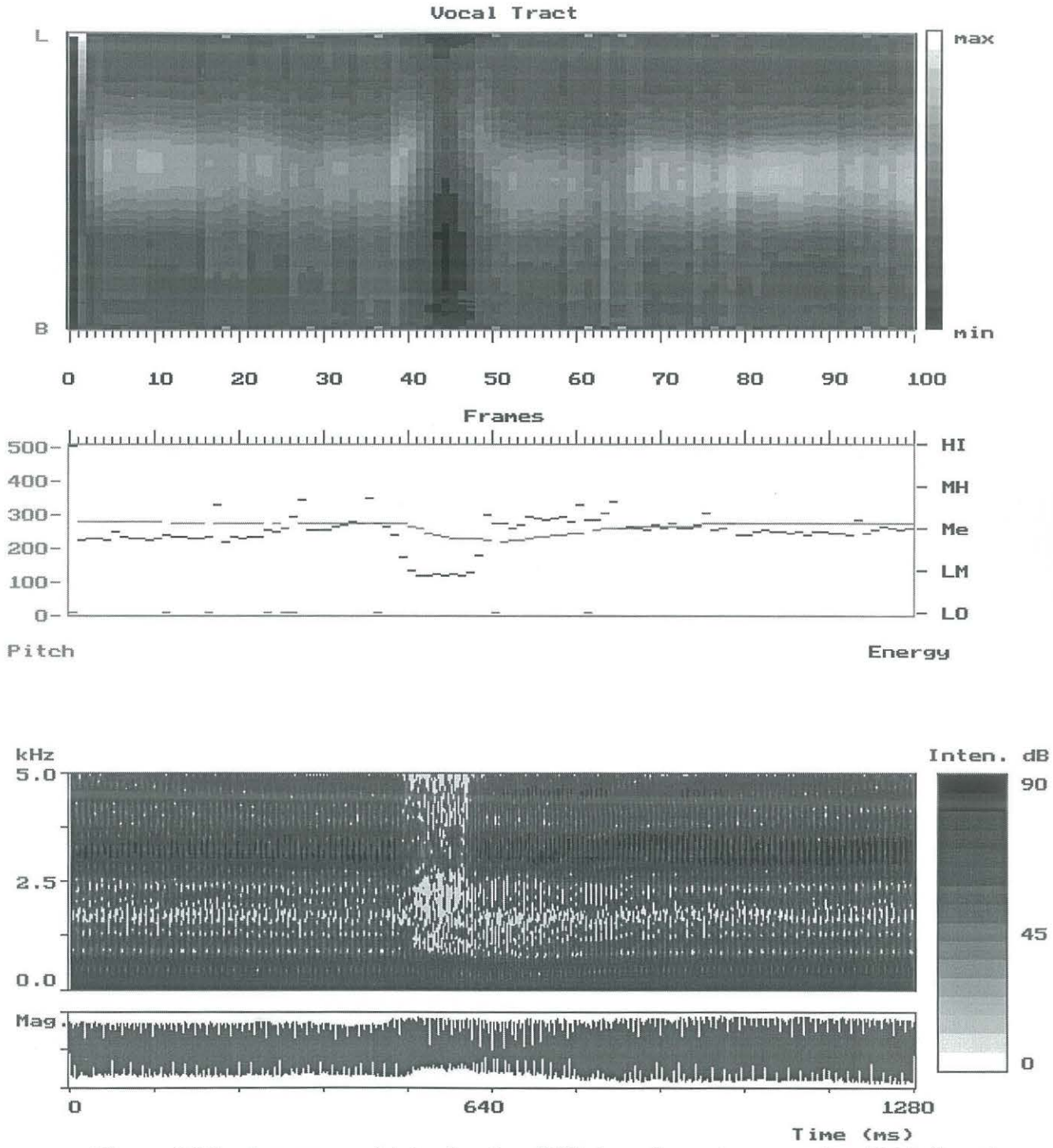

Figure 5.23: Areagram obtained using STS-4, and spectrogram for /ihi/ (female speaker)

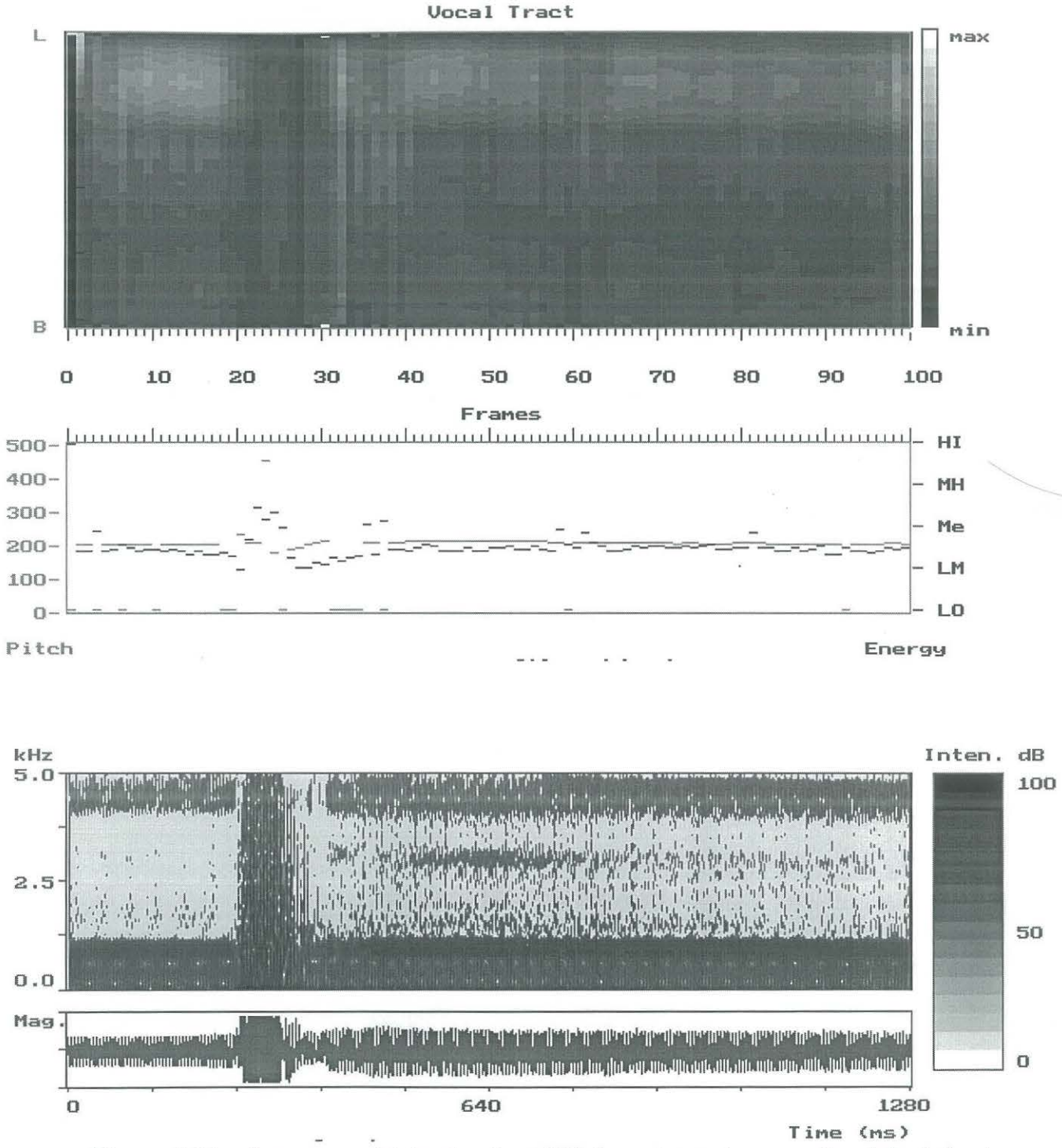

Figure 5.24: Areagram obtained using STS-4, and spectrogram for /uhu/ (male speaker)

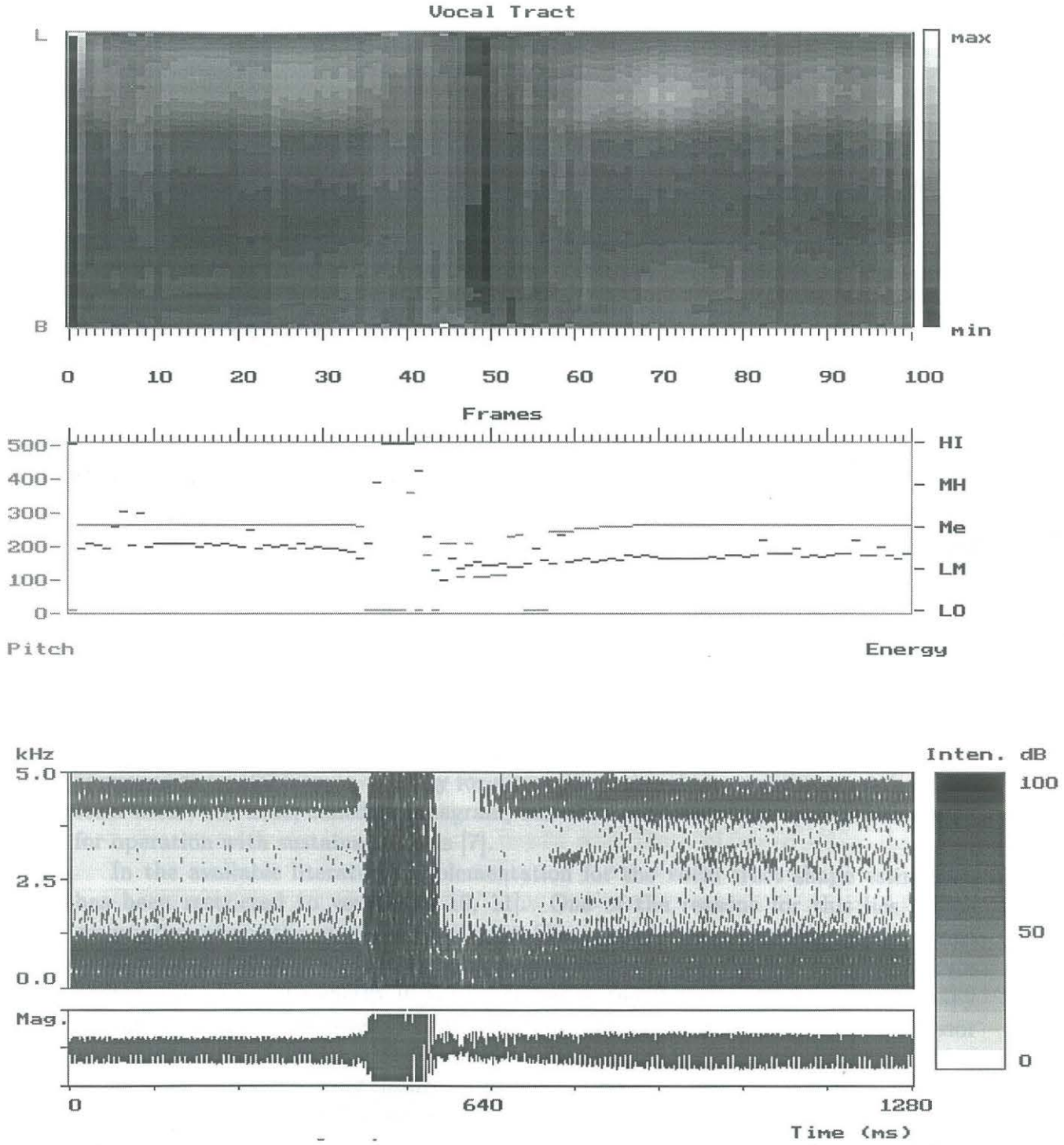

Figure 5.25: Areagram obtained using STS-4, and spectrogram for /uhu/ (female speaker)

Chapter 6

SUMMARY AND SUGGESTIONS

6.1 Summary

The objective of this project is to develop a speech training aid for the hearing impaired which will help them in learning to produce intelligible speech. A real-time visual feedback of vocal tract area, pitch, and energy for the speech has been selected for this purpose. It should be possible for the trainee to use these parameters for improvement in speech production. The system STS-3 developed earlier at IIT Bombay, uses an add-on DSP board and a PC for estimating these parameters directly from speech signal and displaying in real-time. There is a facility for review mode in which these parameters can be displayed for a duration of 1.28 seconds which corresponds to 100 frames. Facility for a two dimensional representation of the vocal tract area with time, called as areagram, has been provided. The system was tested for operation with sustained vowels [7].

In the available literature, implementation for the vocal tract shape estimation has been restricted to vowels [9, 10, 11]. One of the reasons for this lies in the assumption about the model of vocal tract filter, that it has only poles, and no zeros. This, though justified for sustained vowels, gives rise to errors for estimation over consonants. Also, the energy in speech signal is very low during utterance of many of the consonants, and therefore, the computation done on a fixed point processor results in a loss of precision and underflows. It is not possible to obtain information about area variation during closure durations of stop consonants. Thus, STS-3 cannot operate satisfactorily with consonants.

For speech training with utterances other than sustained vowels, the important information is the place of articulation. Because of the limitaions of the system, this critical information is not available to the user. It is not possible to obtain this information by processing the speech signal directly within the framework of the present system. Hence, it was decided to use the information about area variation

obtained just before and just after the stop. Since, areagram gives the area variation with time, it can be used for this. It was decided to extract the required information by 2-D interpolation of the areagram matrix.

For this purpose, various areagrams obtained for vowels, vowel sequences, semi-vowels, and vowel-consonant-vowel sequences uttered by various users were studied. After a close examination, and after studying actual area values, frame to frame inconsistancies in area variation became apparent. The reason for this was found to be the normalization procedure for areas before these are mapped on to areagram display. After removing some minor errors in the software, the normalization procedure was modified. This normalization now takes care of any erroneous peaks in the area, and the error introduced in the entire area distribution thereby, and also helps in bringing out the difference between relative area distributions for different vowels more clearly. The area values are mapped to 16 grey levels after taking square root, which resulted in improved areagram display. Also, the area values near glottis which show very less change, were discarded and curve fitting is done keeping the total number of points constant.

As explained before, an effort was made to extract the missing information of area variation during stop consonants, by 2-D spatial interpolation of the areagrams for V-C-V sequences. After applying this to a number of areagrams, it was found that it was not possible to interpolate the vocal tract shape during the closure based on the information preceding and following the closure. It was further found necessary to investigate the results as obtained using floating point arithmatic. This would bring down the minimum rms value of the signal necessary for satisfactory estimation and reduce the duration for which the estimation is unreliable. The interpolation thereafter would give better results. For this, the estimation algorithm was implemented on a PC using floating point arithmatic. Off-line processing of previously recorded speech segments was done, and areagrams obtained for V-C-V sequences were low pass filtered to compare the results with those obtained using the fixed point implementation. However, this did not give any significant improvement.

In order to improve the estimation over weak energy segments, upward scaling of the signal samples prior to estimation, especially when signal strength is low, can be helpful. The estimation over the speech segments normalized with respect to mean magnitude showed improvement in the estimation of vocal tract shape for semi-vowels. It brought down the minimum required signal strength necessary for satisfactory estimation by a factor of 2. This could be implemeneted on fixed point arithmatic, and the results obtained are comparable to those obtained with the off-line floating point implementation.

It was tested and verified that the system STS-4 works satisfactorily for vowel utterances with variable pitch. The estimated pitch follows the changes in the input pitch, and vocal tract area estimation is unaffected by changes in the pitch. Vocal tract area estimation can also be made for unvoiced speech segments, however pitch estimation is not reliable.

6.2 Suggestions for Further Work

It was concluded that it is not possible to extract the information regarding the area variation during stop closures by simple interpolation. Other techniques for estimating the place of closure, on the basis of transition in the area function preceding and following the closure need to be investigated. Though this project resulted in a more meaningful display for areagram, the display of the vocal tract shape in real time could be made more realistic by showing outline of human face on the monitor. Its vocal tract shape could be made to vary according to speech input.

Appendix A

BEZIER FORM ALGORITHM

This algorithm finds a third degree polynomial which passes through the two out of four given points. The remaining two points are used to define the tangents to the curve. Consider four points shown in the *Fig.* A.1 [16]. Let the curve to be fitted have the equation $x(t)$ as function of t. Let $x(t_1)$, $x(t_2)$, $x(t_3)$, and $x(t_4)$ be represented by $P1$, $P2$, $P3$, and $P4$ respectively. A smooth curve is to be fitted through them. $x(t)$ is assumed to have the following form:

$$x(t) = at^3 + bt^2 + ct + d \tag{A.1}$$

where a, b, c, and d are the constants to be found out. For this, four equations would be needed. It is further assumed that the curve passes through end two points *i.e.* $P1$ and $P4$. Hence,

$$x(t_1) = at_1^3 + bt_1^2 + ct_1 + d \tag{A.2}$$

and

$$x(t_2) = at_2^3 + bt_2^2 + ct_2 + d \tag{A.3}$$

t_1 can be put to 0, and hence *Eqn.* A.2 reduces to

$$x(0) = d \tag{A.4}$$

Also, the straight line joining $P1$ and $P2$ is a tangent to the curve at $P1$, and the one joining $P3$ and $P4$ is a tangent at $P4$. Differentiating A.1 gives

$$\frac{dx}{dt} = 3at^2 + 2bt + c \tag{A.5}$$

For the tangent at $P1$ ($t_1 = 0$), this reduces to

$$\frac{x(t_2) - x(0)}{t_2} = c \tag{A.6}$$

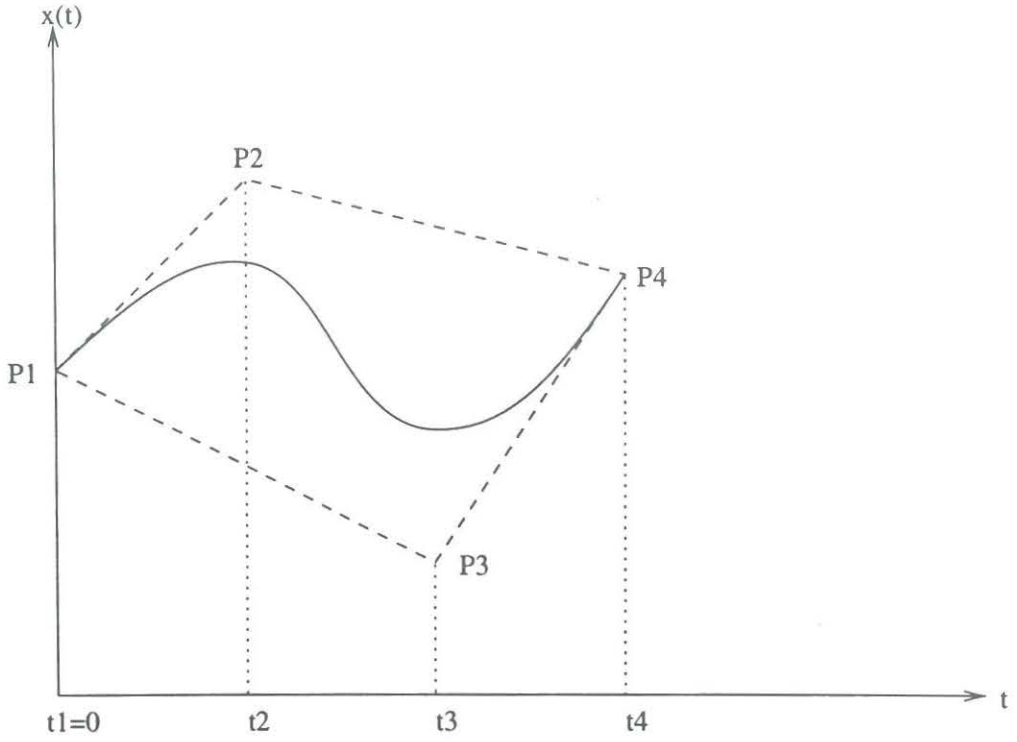

Figure A.1: Bezier curve fitting through four points

whereas for tangent at $P4$, it gives

$$\frac{x(t_4) - x(t_3)}{t_4 - t_3} = 3at_4{}^2 + 2bt_4 + c \tag{A.7}$$

Eqn. A.3, A.4, A.6, and A.7 can be solved simultaneously for a, b, c, and d. After solving and substituting the values in *Eqn.* A.1, the following final expression is obtained in terms of known quantities.

$$x(t) = \frac{1}{t_4{}^3}\{t^3x(t_4) + t^2(t_4 - 3t)x(t_3) + 3t(t_4 - t)^2x(t_2) + (t_4 - t)^3x(0)\} \tag{A.8}$$

Examining the above polynomial coefficients, it is clear that each ranges between 0 and 1, and their sum is 1 for $0 \leq t \leq t_4$. Thus, the expression is a weighted average of the four control points. It can further be shown that, the weighted average of n points falls within the convex hull defined by those n points [16].

Appendix B

OPERATION OF "STS-4"

The requirement of the Speech Training System "STS-4" are as follows:

- PC 386 with VGA monitor
- PCL-DSP25 (TMS 320C25 Digital Signal Processor Board, from M/S Dynalog, Mumbai)

The DSP processor TMS 320C25 is a 16 bit fixed point processor. The PCL-DSP25 board can be used as an add-on card on the PC slot. It has on-board ADC, DAC, and timer. These are mapped on the I/O ports of the DSP chip. The 16 bit timer is clocked at 5 MHz, and can be loaded with appropriate count to set the conversion rate of the ADC. It has been set to 10 kHz. Both ADC and DAC are 16 bit devices. The on-board memory can be shared by the PC. All the on-board memory access is through a block of 8 I/O port addresses on the PC. This enables fast data transfer between the PC and the DSP board. These I/O ports are used for communicating control and status information, and for exchanging memory addresses and data between the two processors [17].

The speech signal is acquired through a microphone, amplifier, and filter; as described earlier in section 3.3. The signal presented to the A/D input of the DSP board should be in the range of ± 10 V and must have been low pass filtered to avoid components above 5 kHz. The acquisition and analysis are done in real time, and results can be displayed either in real time or in review mode. It is also possible to use 'off-line' mode for analysis and display of previously digitized and stored signal. Separate executable files are provided for both modes.

B.1 Real-time Mode

For real time mode, "v_ear.exe" is to be run. "v_ear.mpo" file also must be available in the same working directory. Options are provided to the user at every step. The opening screen allows entering real-time mode. In real time-mode, three boxes for

pitch, energy, and vocal tract area can be seen on the monitor. The corresponding parameters are displayed in real time as the user speaks into the microphone. At this stage, by use of function keys, speech parameters for maximum 100 frames (1.28 second) can be stored.

1. Press 'F2' to start the capture of data. Area, pitch, and energy information is saved temporarily in a buffer for immediate use. The number of the frame being captured will be displayed. After 100 frames have been captured, the buffer is over-written to contain data for most recent 100 frames.

2. Press 'F4' to end the capturing process. At this point, the captured data is ready for use in review mode. At the same time, real time operation also continues till the user does not press 'F3'. Pressing 'F2' again will restart capturing.

Once the data are captured, the stored information can be displayed in following three formats:

1. Pressing 'F1' enters "Manual select" mode where any frame out of captured 100 frames can be selected by use of arrow keys and displayed at a time

2. Pressing 'F2' enters "Animate" mode where the captured frames can be reviewed in slow motion. This mode provides facility to pause at any frame, by pressing 'F1' (Toggle Pause).

   During the operation in both these modes, real-time operation still continues, and the current vocal tract shape is superimposed over the one being displayed for the selected frame. The two can be distinguished by the brightness of the contour. The current area is displayed with the brighest shade, whereas, the area from captured frames appears dull.

3. Pressing 'F3' enters "Areagram" mode. In this mode, two dimensional representation of the area with time and distance from glottis to lips can be seen. Area values are square-root mapped before displaying. Pitch and energy displays are also provided. For display of areagram, maximum area of the vocal tract for a particular user can be input to the software which is used for normalization. Otherwise, the software finds this value, and displays it for further reference. This value can subsequently be used for normalization of area values for areagram display of other speech utterances by the same user. The option is given to the user to save this information permanently in a file, to be used for further study/reference.

4. Pressing 'Esc' at any of the steps above returns the control to the real-time mode.

B.2 Off-line Mode

The signal can be acquired and stored in a file by using the programme "adc.exe". The duration for which signal is to be acquired is user defined. This acquired signal can be used further for off-line processing. In order to obtain speech parameters for speech data captured previously, "off_line.exe" can be used. The working directory must contain the file "off.mpo". The speech data should be in the binary form, for a duration of 1.28 seconds (12800 samples), at the sampling rate of 10 kSa/s. The programme "f_load.exe" has to be used to load this speech data on to the DSP board. Upon running "off_line.exe", speech parameters are computed and can be stored in a file. Further, programme "display.exe" can be used to display these parameters in the manner same as that used in areagram mode in "v_ear.exe".

In both real-time as well as off-line mode, facility is provided to store the parameters in the form of an image which can be viewed using XView, and printed on a laser printer.

B.3 Floating Point Computation

In order to study the results obtained by running the estimation algorithm using floating point arithmatic, the programme "fpc.exe" can be used. This uses the previously captured speech data files, obtained using "adc.exe", similar to "off_line.exe". The length of the speech segment must be 12800 samples at the sampling rate of 10 kSa/s. The stored speech data file must be in the binary format. The programme "fpc.exe" estimates the vocal tract area and stores the results in the user defined file. The corresponding areagram can be viewed later using the programme "display.exe".

B.4 Spectrographic Analysis

In order to compare the areagrams obtained using STS-4 with corresponding spectrograms, software developed by Baragi [7] is useful. The programme "spectro.exe" developed by Baragi, and further modified by Mr. D. S. Choudhari, uses the stored speech signal file, obtained using "adc.exe". Spectrogram can be obtained and printed in a manner similar to the the areagram obtained using STS-4.

References

[1] Milind S. Gupte, "A Speech Processor and Display for the Speech Training of the Hearing Impaired", M.Tech. Dissertation, Dept. of Electrical Engg., IIT Bombay, January 1990, Dr. P. C. Pandey & Prof. T. Anjaneyulu.

[2] Kishori S. Taklikar, "A Speech Training Aid for the Deaf", M.Tech. Dissertation, Dept. of Electrical Engg., IIT Bombay, January ~~1990~~ 1991, Guides: Dr. P. C. Pandey & Prof. T. Anjaneyulu.

[3] Sebactian Gracias, "A Speech Training Aid for the Hearing Impaired", ~~M.Tech.~~ B.Tech., Dissertation, Dept. of Electrical Engg., IIT Bombay, January 1991, Guide: Dr. P. C. Pandey.

[4] Niranjan D. Khambete, "A Speech Training Aid for the Deaf", M.Tech. Dissertation, School of Biomedical Engg., IIT Bombay, June 1992, Guides: Prof. S. R. Devasahayam & Dr. P. C. Pandey.

[5] Yogesh A. Bhagwat, "Real-time Vocal Tract Shape and Pitch Estimator", M.Tech. Dissertation, Dept. of Electrical Engg., IIT Bombay, January 1993, Guide: Dr. P. C. Pandey & Prof. S. D. Agashe.

[6] Prashant S. Gavankar, "A speech training aid for the deaf", M.Tech. Dissertation, Dept. Electrical Engg., IIT Bombay, 1995, Guides: Dr. P. C. Pandey & Prof. S. D. Agashe.

[7] Ashok Baragi B. N., "A Speech Training Aid for the Deaf", M.Tech. Dissertation, Dept. of Electrical Engg., IIT Bombay, January 1996, Guides: Dr. P. C. Pandey & Prof. S. D. Agashe.

[8] Hisashi Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", *IEEE Trans. Audio & Electro-acoustics*, Vol. 21, pp. 417-427, October 1973.

[9] David Rossiter, David M. Howard, & Marcus Downes, "A real time LPC-based vocal tract area display for voice development", *Journal of Voice*, Vol. 8 (4), pp. 314-319, 1994.

[10] Peter Ladefoged, Richard Harshman, Louis Goldstein, & Lloyd Rice, "Generating vocal tract shapes from formant frequencies", *J. Acoustic Soc. Am.*, Vol. 64 (4), pp. 1027-1035, October 1978.

[11] Sang Park, Dong Kim, Jae Lee, & Tae Yoon, "Integrated speech training system for hearing impaired", *IEEE Trans. Rehab. Engg.*, Vol. 2 (4), pp. 189-196, December 1994.

[12] Lawrence R. Rabiner & Ronald W. Schafer, *Digital Processing of Speech Signal*, Englewood Cliffs, New Jersey : Prentice Hall, 1978.

[13] Lawrence E. Kinsler, Austin R. Frey, Alan B. Coppens, & James V. Sanders, *Fundamentals of Acoustics*, New York : John Wiley & Sons, 1982.

[14] L.Roux & C. Gueguen, "A fixed point computation of PARCOR coefficients", *IEEE Trans.Acoustics, Speech & Signal Processing*, Vol. 25, pp. 257-259, June 1977.

[15] Douglas O'Shaughnessy, *Speech Communication:Human and Machine* , New York : Addison-Wesley, 1987.

[16] J. D. Foley & A. Vandam, *Fundamentals of Interactive Computer Graphics*, pp. 514-536, New York : Addison-Wesley, 1983.

[17] *PCL-DSP25, User's Manual*, Dynalog Microsystems, Mumbai.

[18] Anil K. Jain, *Fundamentals of Digital Image Processing*, Englewood Cliffs, New Jersy : Prentice Hall, 1989.

[19] D. H. Klatt, "Software for a cascade/parallel formant synthesizer", *J. Acoust. Soc. Am.*, Vol. 67 (3), pp. 971-995, March 1980.