Estimation of Vocal Tract Shape for Speech Training Aids

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

Master of Technology

by

K. S. Nataraj
(Roll No. 09307612)

under the supervision of

Prof. P. C. Pandey

Department of Electrical Engineering
Indian Institute of Technology Bombay
June 2012

Indian Institute of Technology, Bombay

M. Tech. Dissertation Approval

This dissertation entitled **"Estimation of Vocal Tract Shape for Speech Training Aids"** by **K. S. Nataraj** (Roll No. **09307612**) is approved, after the successful completion of *viva voce* examination, for the award of the degree of **Master of Technology** in **Electrical Engineering.**

Supervisor	.....	(Prof. P. C. Pandey)
Examiners	.....	(Prof. Preeti Rao)
	.....	(Prof. M. S. Shah)
Chairperson	.....	(Prof. B. Ravi)

Date: 22 June 2012

Place: Mumbai

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Natg. K S

(K. S. Nataraj)

Date: June 2012

Place: Mumbai

K. S. Nataraj / Prof. P. C. Pandey (Supervisor), "Estimation of vocal tract shape for speech training aids", *M.Tech. dissertation*, Department of Electrical Engineering, Indian Institute of Technology Bombay, June 2012.

Abstract

Speech-training aids providing a visual feedback of articulatory efforts can be used for improving articulation by the hearing-impaired persons. LPC-based estimation of vocal tract shape works satisfactorily for vowels but fails during stop closure due to very low signal energy and lack of spectral information. The vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterances can be estimated by bivariate surface modelling of the vocal tract area function during the vowel-consonant (VC) and consonant-vowel (CV) transition segments. The accuracy of this method depends on the accurate location of the transition segments and accuracy of the estimation of dynamically changing shapes. This thesis presents investigations for improving the estimation of vocal tract shape during the steady state and transition segments as well as during the closure segments.

A windowed energy index is calculated as the ratio of the energy of the windowed signal to the frame energy, and it is shown that the shapes in the frames corresponding to the valleys in this index have a reduced variability. Thus the selection of the frames based on this index can be used for improving the consistency of vocal tract shape estimation during VC and CV transition segments of VCV utterances and these shapes can be used to more accurately estimate the place of articulation during stop closures of VCV utterances. Different LPC analysis techniques were investigated for estimation of the vocal tract shape of vowels. As compared to autocorrelation method, covariance and lattice methods showed reduced variability. However, covariance analysis resulted in unstable predictor polynomial for some utterances. A technique for detecting the VC and CV transitions in VCV utterances based on a measure of the rate of change of vocal tract area function is presented. The automatically marked start and end points of transitions showed a good match with the manually marked ones and resulted in a consistent estimation of the place of closure of velar, alveolar, and bilabial stops.

CONTENTS

List of Abbreviations and symbols

Abbreviation	Explanation
LPC	linear predictive coding
VCV	vowel consonant vowel
VC	vowel consonant
CV	consonant vowel
XRMB	X-ray microbeam
2-D	two-dimaneional
EMA	electromagnetic articulography
RMS	root mean square
GCI	glottal closure instant
MRI	magnetic resonance imaging
MFCC	mel frequency cepstrum coefficients
LSF	line spectrum frequencies
PLP	perceptual linear prediction

Symbol	Explanation
α_k	linear predictor coefficient
μ_m	reflection coefficient
A_m	area value of a section
$s_n(m)$	speech signal
$E_w(n)$	windowed energy index
$e_n(m)$	prediction error

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Problem overview

In children with normal hearing, speech correction during the learning process of speech production is aided by auditory feedback of the sounds produced by themselves and others. However, speech development in hearing-impaired children is disrupted due to lack of such auditory feedback. They have great difficulty in acquiring the ability to control position and movements of various articulators involved in speech production. Thus there is a need for speech training aids which can provide appropriate non-auditory feedback to help the process of speech correction in hearing-impaired children.

Several speech training aids have been developed to dynamically display important acoustic parameters, such as speech intensity, voicing and pitch, spectral features etc. [1]-[5]. It was found that in the speech of the hearing-impaired children, the phonemes produced in the front of the mouth are more intelligible than the phonemes produced in the back of the mouth [6]. This highlights the importance of the relative visibility of articulatory gestures in speech development of hearing-impaired children. Speech-training aids providing a visual feedback of articulatory efforts have been found to be useful in improving vowel articulation by the hearing-impaired children [7]-[9].

Wakita's LPC-based method [10] is commonly used for direct estimation of vocal tract shape in these speech training aids due to its suitability for real-time processing [7]. The method can be used for estimating fixed as well as transitional vocal tract configurations during speech segments with glottal excitation which can be modeled as produced by an all-pole vocal tract filter. Some of the limitations of the LPC-based estimation of vocal tract shape, as outlined by Wakita [11], are (i) related to the errors in estimation of vocal tract transfer function from the band-limited speech signal, (ii) due to uncertainty in glottal source characteristics, and (iii) due to lack of a method for estimating the scaling factor for converting the area ratios into

area values during dynamically varying vocal tract configuration. Despite its limitations, Wakita's method works satisfactorily for vowels. However, vocal tract shapes estimated are random and unrelated to place of articulation during stop closure due to low signal energy and lack of spectral information [10], [12]. Hence, a technique for estimating the place of articulation of stop consonants is needed for improving the effectiveness of speech-training aids.

Pandey and Shah [12] reported a method to interpolate the vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterances by using a bivariate surface model fitted on the vocal tract shapes during the transition segments preceding and following the stop closure. Use of least-squares based bivariate quadratic approximation resulted in satisfactory estimation of place of closure for different unvoiced and voiced oral stops. The accuracy of this method depends on the accuracy in locating the transition segments and the accuracy in the estimation of dynamically changing shapes. A detailed investigation showed that the accuracy of the estimation was highest if the segments used for modeling corresponded closely to the VC transition segment preceding the closure and the CV transition segment following the closure [13]. Inclusion of a part of the closure or the release burst in the segments used in the modeling can introduce errors, while that of a part of the vowel on either side can decrease the sensitivity of the fitted model in interpolating the place of closure. Investigations also showed that the vocal tract shapes estimated for steady-state vowel vary with the position of the LPC analysis frame. Low-pass filtering of the shapes across the frames for improving the consistency cannot be used during transition segments.

1.2 Project objective

The project objective is to improve the accuracy of the vocal tract shape estimation for improving the visual feedback of articulatory efforts in speech training aid. Investigations are carried out for (i) improving the consistency of the vocal tract shapes obtained from LPC analysis, and (ii) developing a method for automatically detecting the VC and CV transitions as needed for improving the accuracy of the estimated place of closure of VCV utterances.

1.3 Dissertation outline

Chapter 2 reviews the various methods for estimation of vocal tract shapes. Investigations for improving the consistency of the vocal tract shape estimation using a method for selecting the analysis frames based on the windowed energy, is presented in the third chapter. Chapter 4 presents the investigations of different LPC analysis methods for estimation of vocal tract shape. Investigations for marking the VC and CV transition segments based on the rate of change measure in the area values is presented in the next chapter. The last chapter provides summary and conclusions along with suggestions for future work.

Chapter 2

VISUAL FEEDBACK FOR SPEECH-TRAINING AIDS

In children with normal hearing, the process of speech correction while acquiring the speech is aided by auditory feedback. The hearing impaired children lack this feedback and therefore experience difficulty in acquiring normal speech characteristics. The objective of the speech-training aid is to provide information using visual or tactile feedback as a substitute for the cues normally available through the auditory channel [1]-[5]. This chapter presents a review of visual speech-training aids, followed by a review of techniques for vocal tract shape estimation.

Estimation of the vocal tract shape consists of recovering the sequence of vocal tract cross-sectional area values that produce a given acoustic speech signal. Vocal tract shapes can be estimated directly using imaging techniques and indirectly from acoustic measurements or speech signal. Vocal tract shapes can be estimated using direct imaging techniques like X-ray [20]-[23], electromagnetic articulography (EMA) [24], electropalatography (EPG) [17], and real-time magnetic resonance imaging [25], [26]. Vocal tract shapes can be estimated indirectly from acoustic measurements based on measurement of input impedance or impulse response at the lips using an impedance tube [27]-[29]. Estimation of vocal tract shapes can be carried out from the speech signal using several methods: formants and factor analysis [28], [30], [31], linear predictive coding (LPC) based analysis [10]-[12], articulatory codebook mapping [32], etc.

2.1 Visual speech-training aids

Computer aided speech-training systems generally provide visual feedback of several features of speech production. Several speech training aids have been developed to dynamically display important acoustic parameters such as speech intensity, voicing, pitch, and spectral features. Nickerson and Stevens [2] developed a computer-aided speech-training system, for the hearing-impaired, providing visual feedback of

loudness, fundamental frequency, nasalization and voicing. The loudness correction was aided by visual display of a ball expanding or contracting as the loudness increases or decreases respectively. Voicing information was displayed using a facial expression of a cartoon and the voice pitch control was assisted by the process of moving a ball through a hoop. All the acoustic parameters were also displayed as a function of time. Its use resulted in varying degrees of improvements along different aspects of speech development. The authors concluded that computer-based speech training could be highly helpful for persons with hearing-impairment and suggested the need for an integrated speech training system to focus on the entire problem instead of focusing on isolated aspects of speech training. Mahshie [3] used the visual feedback of airflow to teach production of voicing distinction. He reported that improvement was greater for specifically trained utterances than for spontaneous or untrained speech. Zahorian and Venkat [4] used spectral features to train the vowel articulation in hearing-impaired persons and the results indicated that the display was useful in improving vowel articulation. The Video Voice Training System, from Micro Video Corporation [14], provides various games for training voicing, intensity, and pitch training. Vowel triangle plot is used for vowel training, in which the corresponding vowel in the plot is shaded for the utterance made. However, Neri *et al.* [15] reviewed existing computer assisted speech training systems and concluded that the visualization of the acoustic parameters is not adequate for unsupervised speech training. Hence, it is essential to focus on providing visual feedback of articulatory efforts for effective speech-training of hearing-impaired.

Fletcher [6] developed a speech-training system called orometer, which provides visual feedback of articulatory actions along with acoustic parameters. The articulatory information included position, configuration and movement of tongue (from optical sensors), pattern of tongue contact with upper palate and teeth (by using thin acrylic plate called pseudopalate), and position and movement of lips and lower jaws (from video camera). Experiments with 10 children with severe-to-profound hearing impairment and 10 children with normal hearing showed that the lip position targets were achieved with high accuracy. They concluded that the feedback was useful for the hearing-impaired group in mastering the production of sounds which are articulated inside the mouth. Javkin *et al.* [16] reported a speech training system in which several types of instrumentally measured articulatory parameters (namely palatography, nasal vibration, airflow and presence or absence of voicing) and

acoustic parameters (such as amplitude, fundamental frequency, and spectral shape) were integrated and displayed in both technical and motivating game formats. They reported another system [17] displaying the articulatory model movements required to produce an utterance entered by the learner. The system also gave a feedback of the similarity between the standard parameters and the learner parameters of the typed utterance. It was reported to be highly motivating to the children and encouraging them to participate in speech-training. However, the palatography apparatus for speech-training is somewhat inconvenient to use.

Crichton and Fallside [7] developed a speech-training aid for teaching the articulation of vowels to hearing-impaired children. The system provided visual feedback of vocal tract shape obtained using Wakita's linear prediction based inverse filtering method [10]. Vocal tract shape was displayed as a plot of log area values against linear distance along the vocal tract. The experiments with the hearing impaired children in the 6–9 years age group showed the aid to be useful for improving vowel articulation. Pardo [8] reported a speech-training system providing visual feedback of articulatory motion for teaching vowels and that the errors with respect to the target shapes decreased progressively with training. Park *et al.* [9] developed an integrated speech training system displaying the vocal tract shapes and other speech parameters such as intensity, fundamental frequency, nasality, and frequency spectra, in real-time. Using this system, hearing-impaired children mastered the syllables /ja/ and /pa/ in 5–6 days. "Dr. Speech", from Tiger DRS, Inc. [18], provides the vocal tract shape and lip movement for the predefined set of utterances. However, it does not provide the visual feedback of the vocal tract shape made by the hearing-impaired children. This speech therapy software also provides training for controlling the pitch, intonation, stress, and loudness using graphical displays. A computer speech-training aid named ARTUR developed by Engwall *et. al* [19] uses a 3D animated head to give visual feedback of vocal tract shape. In this method, vocal tract shapes are estimated from the video images of the face by exploiting the correlation between face and tongue positions. It combines speech recognition and visioacoustic-to-articulatory inversion to give feedback in the form of clear instructions and animations. For evaluating it, interviews were conducted with 3 children aged 9-14, 3 children aged 6, and adults accompanying the children. The interview results showed that the system to be helpful in acquiring accurate articulation, easy to use, and helpful in practicing alone, although a few children found it difficult to imitate the pronunciation.

A review of speech-training aids for the hearing-impaired has shown that systems with visual feedback of articulatory efforts were useful for acquiring the speech. Hence, it is essential to investigate techniques for estimation of vocal tract shape.

2.2 Estimation of vocal tract shape by direct methods

MacMilan and Kelemen [21] described a technique to extract the X-ray images of vocal tract. In this technique, the outline of the tongue, the roof of the mouth, and the pharyngeal wall were made visible by giving the subject a mixture of barium and water. Speaker was asked to hold the primary position of the articulation till the X-ray picture was taken during phonation. In case of stops, two pictures were taken, the first showing the constriction and the second the position of the articulators after the release. These vocal tract shapes are commonly used for validation of the vocal tract shapes obtained from indirect methods. To obtain the time-varying properties of the vocal tract, high speed X-ray films were used [22]. An X-ray film database created by Munhall *et. al.* [23] provides twenty-five high quality X-ray films, totaling 55 minutes of footage. This database contains a series of X-ray movies of mid-sagittal views of vocal tracts in motion, in synchrony with the resulting acoustic signal. However, X-ray imaging cannot be used extensively to estimate the vocal tract shape due to the harmful effects of radiation exposure. Another limitation of X-ray films includes difficulty in accurately deducing the cross-sectional morphology from mid-sagittal profiles.

X-ray microbeam (XRMB) is an alternate approach to vocal tract imaging during speech [20]. This uses less toxic narrow X-ray beams for tracking the movement of articulators in the mid-sagittal plane (side–view) using pellets glued at the tongue, lips, and on the jaw. These measurements provide good information about the articulatory characteristics of tongue movements. But the images of the tongue root and pharynx are not available. Other limitations of this technique include that it is expensive and cannot be used for speech-training aids.

Electromagnetic articulography (EMA) [24] provides information about the movement of various articulators (tongue, jaw, lips, and teeth) using a plastic helmet which the speaker wears during data recordings. Three transmitter coils are mounted equidistant from one another on the helmet. Small receiver coils (sensors) are placed on the upper and lower lip, on the tongue, and on the base of low front incisors. The

magnetic field, generated at different frequencies by transmitter coils, induces an alternating signal in the receiver coils. The alternating voltage induced is inversely related to the distance between the transmitter and the receiver coil. A computer algorithm provides the XY co-ordinates of each of the receiver coil as they move in x-y space over time during phonation. The equipment being expensive and complex to use is not suitable as a speech-training aid.

Electropalatography [17] provides the information about the timing and location of tongue contact with the hard palate during continuous speech. The technique uses an artificial palate that is moulded to fit the roof of the mouth of the speaker. Tongue-palate contact is recorded by a number of silver or gold contacts located on the surface of the pseudopalate. The electropalatography is not suitable for use in speech training as pseudopalates are expensive and have to be manufactured for each child separately.

Real-time magnetic resonance imaging can be used to acquire complete views of the vocal tract during speech production in a safe and noninvasive manner [25], but it is costly and time consuming [26].

2.3 Estimation of vocal tract shape from acoustic measurements

Mermelestein [27] showed that the admittance function of the vocal tract measured at the lips has a unique relationship to the resonant frequencies of the vocal tract. It was empirically shown that the resonant frequencies under closed lip condition correspond to admittance zeros and the resonant frequencies during normal speech correspond to admittance poles. The vocal tract area function was estimated from the admittance function measured at the lips. The area values compared well with the earlier reported X-ray data for 6 Russian vowels [33]. Schroeder [28] reported a method for measuring the acoustic admittance of the vocal tract by attaching an acoustic tube apparatus to the lips of the subject. An electrodynamic driver providing an acoustic output consisting of periodic pulses was coupled to one end of the tube, while its other end of the tube was connected to the mouth of the subject who articulated silently. Two closely spaced microphones were used to record the sound pressure resulting from the incoming and reflected sound waves. Acoustic impedance and hence vocal tract area function were calculated from the recorded signals, assuming the vocal tract to be of

constant length and closed at the glottis. The area values obtained from the admittance function matched well with the reference values.

Sondhi and Gopinath [29] reported a method, for estimation of vocal tract shape from the measurement of the impulse response at the lips, which did not involve any assumptions about the length of vocal tract and boundary condition at the glottis. The impulse response at the lips was obtained by measuring the pressure developed at the lips in response to a unit impulse of volume velocity applied at lips of a vocal tract. The major advantage of this method is that it can extract the area functions for nasalized vowels also, but it is not practical for a speech-training aid.

2.4 Vocal tract shape estimation from the speech signal using formant and factor analysis

Schroeder [28] showed that for a uniform pipe of given length closed at one end and open at the other, each odd term in the Fourier series expansion of the area function affects only one of the resonant frequencies. He computed the anti-symmetric logarithmic area function from the measured formant frequencies. The area functions for the vowels /a/ and /i/ matched with the area functions calculated from X-ray images. But for the vowel /u/, the area function had a strong symmetric component i.e., even Fourier coefficients and hence did not match with the X-ray images.

Ladefoged *et al.* [30] estimated vocal tract shapes of vowels from the first three formant frequencies. The tongue shapes were represented as displacements of the tongue surface from a reference line. Analysis of 50 tongue shapes acquired from X-ray tracings made during the phonation of 10 English vowels by 5 speakers was carried out to develop a simple description of all the tongue shapes in the data set in terms of minimum number of underlying components. The analysis results showed that two components corresponding to front and back raising of the tongue were adequate to describe the entire set of data. The vocal tract shape for each vowel was estimated by taking weighted sum of the two components. The weights were calculated using the formant frequencies. There was a close resemblance between the original and generated shapes for vowels produced by five subjects.

Iskarous [31] showed that it is possible to recover the significant articulatory information, such as the constriction location, constriction degree, and rounding, from the formant frequencies. The first two formant frequencies were used to obtain the

Fourier components of the area function. Instead of measuring the vocal tract length for obtaining the neutral vocal tract formants for normalization step, the formants were standardized for individual speaker. The normalization was carried out by subtracting the mean values from each of the formant raw data and dividing by the standard deviation within all the vowels produced by the individual speaker. The recovered constriction location, constriction degree, and lip aperture showed a good match with those obtained from direct X-ray imaging from XRMB database for 39 participants. The constriction information for the 11 vowels from 6 speakers of American English also showed a good match with the ones obtained from the area functions measured from MRI data.

2.5 Vocal tract shape estimation from the speech signal using linear predictive coding (LPC) based methods

Wakita [10] estimated the vocal tract shape directly from the acoustic speech waveform using LPC analysis by employing a speech analysis model. As shown in Fig. 2.1, the speech was modelled as being generated by an excitation source followed by a filter which takes into consideration the effects of the glottis, vocal tract and radiation at the lips. For non-nasalized voiced sounds, excitation was assumed to be an impulse train. The inverse filter coefficients are obtained by least mean square error technique. An acoustic tube filter that is equivalent to the inverse filter was obtained so that the vocal tract configuration can be related to the frequency domain behavior of the speech signal. The vocal tract configuration was estimated by relating the two inverse filter models.

In the first part of the analysis, the power spectral envelope of the speech was assumed to be approximated by poles only. The transfer function $H(z)$ was expressed as

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.1}$$

where G is a gain term. The inverse filter $I(z)$ was assumed to be a linear filter with only zeros in its transfer function, given by,

$$I(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{2.2}$$

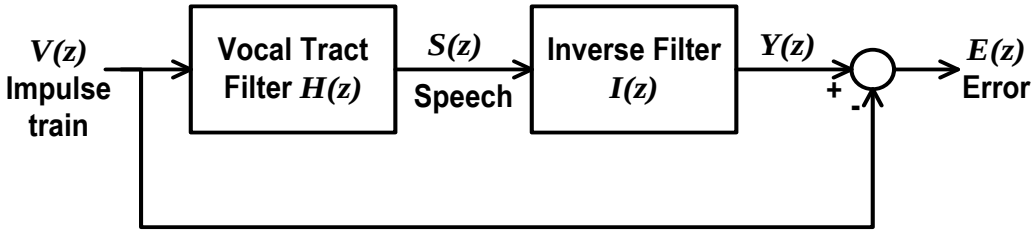

Figure 2.1 Speech analysis model used by Wakita [10]

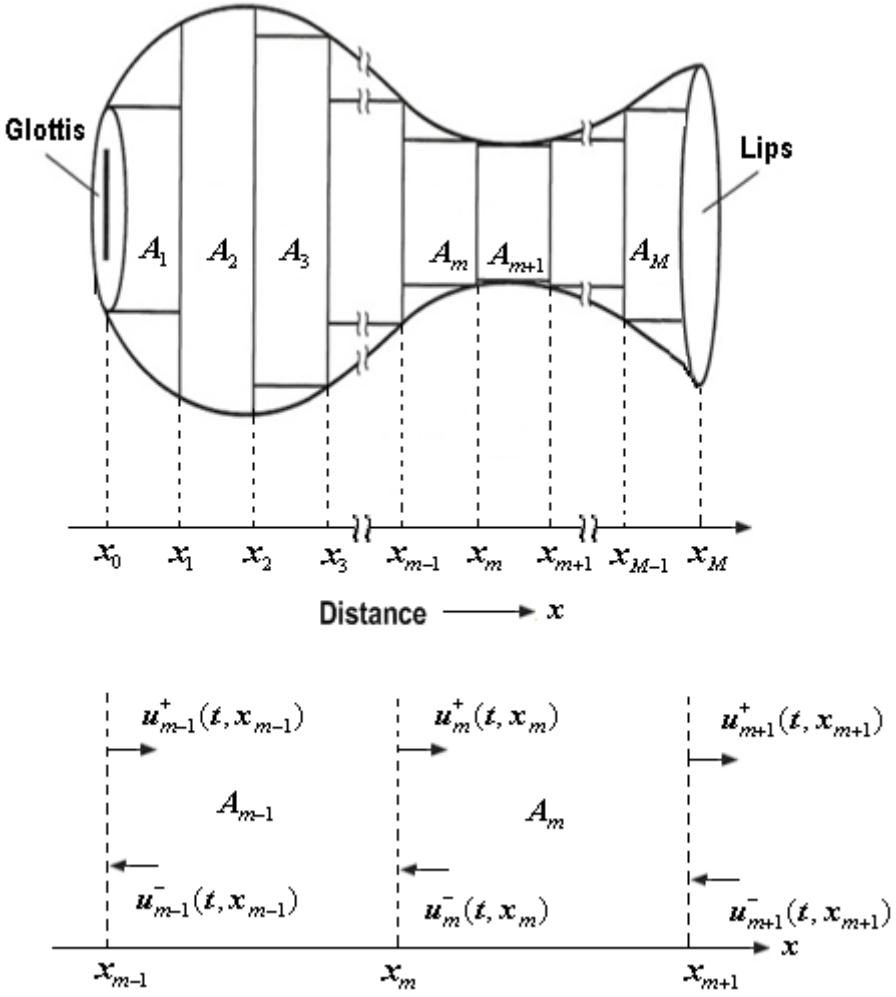

Figure 2.2 Acoustic tube model of the vocal tract

where α_k are predictor coefficients which were obtained from previous samples of speech signal by minimizing the sum of the squared approximation error between the output of the inverse filter and the input impulse train. For the optimum inverse filter, α_k estimated from speech signal match with a_k of the inverse transfer function $H(z)$ in (2.1). The optimum inverse filter coefficients were obtained by Robinson's method [10] using the recursive relation given by,

$$\begin{bmatrix} I_{m+1}(z) \\ J_{m+1}(z) \end{bmatrix} = \begin{bmatrix} 1 & -k_m \\ -k_m z^{-k} & z^{-1} \end{bmatrix} \begin{bmatrix} I_m(z) \\ J_m(z) \end{bmatrix}, \; m=0,1,\ldots,M\text{-}1 \tag{2.3}$$

The coefficients k_m were calculated using the relation

$$k_m = -\frac{\sum_{i=0}^{m} a_i^{(m)} r_{m+1-i}}{\sum_{i=0}^{m} a_i^{(m)} r_i} \tag{2.4}$$

where r_i was a short–term autocorrelation of the speech signal for lag i, $I_0(z) = 1$, $J_0(z) = -z^{-1}$, and $k_0 = r_1 / r_0$ were the initial values, and $I_M(z)$ was defined as the inverse transfer function.

In the second part of the analysis, vocal tract was modeled as a lossless acoustic tube with M sections of equal length and varying cross-sectional area to obtain an acoustic inverse transfer function as shown in Fig. 2.2. The volume velocity in section m was represented by $u_m(t, x_m)$, where t is the time variable and x is the distance variable. For plane wave propagation through the acoustic tube, reflections occur at the section interfaces due to different areas on the two sides. The inverse transfer function for the acoustic tube model was defined as

$$D_M(z) = \frac{\text{forward-going volume velocity component at the glottis end}}{\text{volume velocity component at the lip end}} \tag{2.5}$$

and was obtained using the following recursive relation

$$\begin{bmatrix} D_{m+1}^{+}(z) \\ D_{m+1}^{-}(z) \end{bmatrix} = \begin{bmatrix} 1 & -\mu_{m+1} \\ -\mu_{m+1} z^{-k} & z^{-1} \end{bmatrix} \begin{bmatrix} D_m^{+}(z) \\ D_m^{-}(z) \end{bmatrix} \tag{2.6}$$

This inverse transfer function was compared with the one obtained from LPC analysis in (2.3) to calculate the reflection coefficients μ_m at the junction between sections m and $m+1$, using

$$\mu_m = k_{m-1} \tag{2.7}$$

The area ratios A_m of the mth section are calculated from the reflection coefficients μ_m using the relation

$$A_m = \frac{1+\mu_m}{1-\mu_m} A_{m+1} \tag{2.8}$$

Thus the area values were directly obtained from the acoustic speech waveform. Wakita [10] used +6 dB/octave pre-emphasis to account for the –12 dB/octave slope

of glottal spectral envelope and +6 dB/octave slope for the radiation impedance. The spectrum of the resulting speech was assumed to be the spectrum of the transfer function of a lossless tract, with a zero load at the lips and a resistive load at the glottis. The speech signal was digitized with a sampling frequency of 7 kHz. The method can be used for estimating fixed as well as transitional vocal tract configurations during speech segments with glottal excitation which can be modeled as produced by an all-pole vocal tract filter. Five vowels uttered by a male American speaker, and a diphthong /ai/ were analyzed. It was found that fairly reliable area function for voiced sounds could be extracted.

Wakita [11] compared the LPC-based estimation of vocal tract shape with the lip impulse response method suggested by Sondhi [29]. He discussed some of the limitations of the LPC-based estimation and proposed solutions for some of the problems. It was reported that theoretically it is not possible to obtain a unique continuous area function from the speech signal. However, the information about the finer structure of the vocal tract shape can be obtained from the discrete area function obtained by LPC analysis. Use of pre-emphasis does not eliminate the glottal and radiation characteristics introducing an uncertainty in the estimated area functions. He suggested that the effect of unknown source characteristics can be removed by estimating the vocal tract shape from the speech segment corresponding to the closed-glottis condition. One of the error sources in the estimated vocal tract shapes is the assumption of lossless acoustic tube model for vocal tract. It was reported that energy loss due to vocal tract wall vibrations affected lower formant frequencies. A suitable transformation of the measured formant frequencies using a conversion chart can be used to reduce the effect of losses on shape estimation. In the LPC-based estimation, vocal tract area function was obtained as ratios of the areas on both sides of the section interfaces and these area ratios were converted into areas by assuming a constant normalized area of unity at the glottis end. But during dynamically varying vocal tract configuration, area at the glottis end varies and hence introduces error in the estimated area values. He suggested that this problem can be alleviated if simultaneous measurement of the lip opening is available.

Despite its limitations, Wakita's LPC based method for direct estimation of the vocal tract shape has been widely used as it is suitable for real-time processing [7]-[9]. It is reported to work satisfactorily for vowels, but the estimated shapes tend to be random during the closure segments of oral stops due to low energy and lack of

spectral information [10], [12]. Hence, a technique for estimating the place of articulation of stop consonants is needed for improving the effectiveness of speech-training aids. Pandey and Shah [12] proposed a method to estimate the vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterances by bivariate surface modeling of the shapes during transition segments. The accuracy of this method depends on the accurate location of the transition segments and accuracy of the estimation of dynamically changing shapes.

Nayak *et al*. [34] proposed a method to estimate the scaling factor for converting the area ratios into area values during dynamically varying vocal tract configuration. The scaling factor was estimated from the lip opening area, obtained from a video recording of the speaker's face by using template matching approach. Use of the lip opening area for scaling the area ratios resulted in improvement of the area values at the places of maximum opening without significantly affecting the places of constriction.

One of the major challenges of LPC based method is modelling of the source and radiation characteristics accurately to obtain the accurate transfer function. In this direction, Akande *et al*. [35] proposed a method for overcoming the problem of varying glottal characteristics by suppressing the glottal wave contribution using a dynamic, multi-pole high-pass filter, prior to analysis. The first formant was estimated accurately but the effect of glottal wave over the entire spectrum could not be suppressed.

Deng *et al.* [36] obtained the glottal waveform estimate and the vocal tract filter coefficients from speech signal based on known speech samples during the closed glottal phase. The vocal tract area values were obtained from the autocorrelation coefficients of the filter impulse response over the closed glottal phases. An unbiased vocal tract shape for two vowel sounds were similar to that measured from the magnetic resonance image from some other speaker.

Kang and Lee [37] developed a generalized vocal tract model for pole zero type linear prediction which can be used to obtain the vocal tract shape for the nasalized sounds. They described a method to extract the reflection coefficients of the generalized vocal tract model. But the procedure to obtain the reflection coefficients and vocal tract transfer function was inefficient. Lim and Lee [38] obtained the transfer function for the generalized vocal tract model by assuming lossless condition. They used the concept of line spectral pairs to obtain a neat expression of the transfer

function along with meaningful interpretations. However, they reported that the pole-zero modeling is not simple, computationally expensive, and not well understood.

2.6 Vocal tract shape estimation from the speech signal using analysis-by-synthesis methods

Schroeter and Sondhi [32] reviewed the various techniques for estimating the vocal tract shape from the speech signal including mapping via articulatory codebooks, mapping by basis functions, and mapping by neural networks. Analysis-by-synthesis approach involved using an initial estimate of the articulatory parameters to synthesize the speech and the new set of articulatory parameters is estimated by an optimization procedure to minimize the acoustic distance between the original speech signal and the synthesized speech signal. The area function was calculated from the estimated articulatory parameters. They used Coker's articulatory model [39] for the articulatory speech synthesis. The startup parameters for the optimization process were obtained using articulatory codebook in which a parameter of the speech was used as a key to the look-up for the associated vocal tract shape. However, the codebook look-up failed in many cases due to the non-uniqueness of the inverse problem. In order to optimize the codebook search, dynamic programming involving possible sequence of shapes was used. It resulted in a smoother sequence of shapes for words like /wa/. It failed during the stop closure due to the lack of spectral information during the stop closure. To solve this problem, they obtained the intermediate shapes by extrapolating the articulatory parameters of the Coker articulatory model from left and right contexts as the spectral information is contained in the few frames before and after silence gap and the burst.

Richard *et al*. [40] used the articulatory codebook for estimating the shapes of the unvoiced sounds. But it required a large search through the huge articulatory codebook. Neural networks have been used for the acoustic-to-articulatory mapping with reduced computation time. Soquet *et al.* [41] used a 1-hidden-layer MLP (Multi layer perceptrons) with ten units in the hidden layer to estimate 30 tract areas given target values for the lowest three formant frequencies of 11 French vowels. As the formant frequencies were not sufficient to estimate the shape, the results were not accurate for the vowel /u/. It was concluded that the acoustic features used for

articulatory inversion play an important role in estimating the vocal tract shapes accurately.

Qin and Carreira-Perpinán [42] compared commonly used acoustic features such as LPC, short-time cepstral representation (LPCC), mel frequency cepstrum coefficients (MFCC), line spectrum frequencies (LSF), perceptual linear prediction (PLP), and relative spectral PLP (RASTA-PLP), for finding the best acoustic features for articulatory inversion. The acoustic features were mapped to the articulatory features by a multilayer perceptron with a single layer of 55 hidden units. They found that best mapping resulted in by use of acoustic features such as LSF which are closely related to vocal tract. Laprie and Mathieu [43] used a variational calculus method for estimating the articulatory trajectories from the speech signal using Maeda's articulatory model [44]. Formant trajectories were extracted from the speech signal and from the acoustic simulation of the model. Articulatory parameters were estimated by minimizing a cost function which combined acoustic distance between the two formant trajectories and rate of change of articulatory parameters. For vowel sequence /iai/ and /iui/, the trajectories of three articulatory parameters (jaw position, tongue dorsum position, and tongue shape) were found to be as expected.

2.7 Summary

A review of speech-training aids for the hearing impaired has shown that systems with visual feedback of articulatory efforts were useful for improving the vowel articulation. Most of these systems were based on LPC analysis of speech for the estimation of vocal tract area values. However, speech training systems for training the consonant articulation are not available. Hence, for improving the effectiveness of speech-training systems, it is important to investigate techniques for estimation of vocal tract shape of stop consonants.

Estimation of vocal tract shape based on direct methods and indirect methods based on acoustic measurements are not suitable for development of speech-training aid, because the use of the apparatus interferes with normal speech production. Analysis-by-synthesis methods are not useful due to the difficulty in obtaining the start-up parameters. Vocal tract shape estimated by LPC analysis is reported to work satisfactorily for vowels, but the estimated shapes tend to be random during the closure segments of oral stops. Hence, a technique for accurately estimating the place

of articulation of stop consonants is needed for improving the effectiveness of speech-training aids.

Chapter 3

IMPROVING THE CONSISTENCY OF VOCAL TRACT SHAPE ESTIMATION

3.1 Introduction

LPC based vocal tract shape estimation involves applying Hamming window on the analysis frames of the speech signal, with a short inter-frame interval to track the vocal tract shape changes. The estimated area values are found to vary even for the fixed tract configurations, thus showing an inconsistency. During vowels with fixed tract configurations, low-pass filtering of the estimated area values across the frames can be used for improving the consistency. The variability in the estimation can also be reduced by increasing the length of the analysis frame and by using frame length equal to a multiple of the pitch period. But these methods cannot be used during transitional configurations, e.g. diphthongs, semivowels, and vowel-consonant and consonant-vowel transitions. Hence other means for improving the consistency of the estimated shapes without smearing the transitions are needed.

Rabiner *et al.* [45] observed that LPC prediction error varied substantially with the position of the analysis frames, independent of the analysis method. They proposed two pre-processing methods to reduce the variation in the prediction error: all-pass filtering and pre-emphasis of the speech signal. Although these methods reduced the variability, they increased the prediction error. They also reported that speech synthesized with LPC coefficients obtained from speech frame with maximum prediction error was more nasal-like than the one synthesized from the coefficients obtained from speech frame with the minimum prediction error. It has been reported that the peaks of the prediction error are not always prominent [46]. Hence they cannot be used for selecting frames for improving the estimation of the coefficients. Mizoguchi *et al.* [47] showed that by selecting samples in time domain with prediction error less than a threshold, the variation in the prediction coefficients across the frames was reduced for steady state vowels. Ma *et al.* [48] showed that short time energy based selection of samples was more robust than the LPC prediction

error based selection. Mezzalama [49] reported that large errors were introduced in the LPC-based formant estimation as the window position shifted from the glottal pulse and the error could be reduced by repeatedly concatenating the segment in the frame before applying the window. In this method, repeating the frames introduces error in the analysis if the frame length is not equal to the multiple of the pitch period.

An examination of the LPC analysis results for steady-state voiced segments showed that the variations in the vocal tract shape and in the prediction error both were related to the windowed signal energy. A method for selecting the frames, based on the windowed energy, for improving the consistency of the estimation without smearing the variations during transitional vocal tract configuration is described in the following section.

3.2 Variation in estimated vocal tract shape

For estimation of vocal tract shape using Wakita's method [10], the speech signal was sampled at 10 kHz with 16-bit quantization and a pre-emphasis was applied by taking its first difference. The LPC analysis of order 12 was carried out on analysis frames of duration equal to twice the average pitch period after applying Hamming window. In this analysis, the vocal tract from glottis to lips was modeled as 12 cylindrical sections of equal length and the area values of the sections were calculated from the reflection coefficients obtained using the autocorrelation coefficients of the windowed frame. Analysis of speech signals of steady-state vowels with an inter-frame interval of 5 ms showed a significant variability in the area values for most of the sections. The variation in the area values may be attributed to the natural variation during phonation or to the errors related to the frame position with respect to the instants of glottal closure.

To eliminate the variations contributed by natural changes in phonation, the analysis was carried out on the vowel sequence /-a-i-u-/ synthesized using Klatt synthesizer [50] with a constant pitch frequency of 90 Hz and a constant amplitude. To study the effect of the position of the analysis frame, analysis was carried out for inter-frame interval of one sample. The vocal tract shape as a function of time was displayed using areagram, a two-dimensional display of square root of vocal tract area values plotted as gray levels as a function of frame position along x-axis and glottis-to-lips (G-L) distance along y-axis. The area values of the twelve sections were

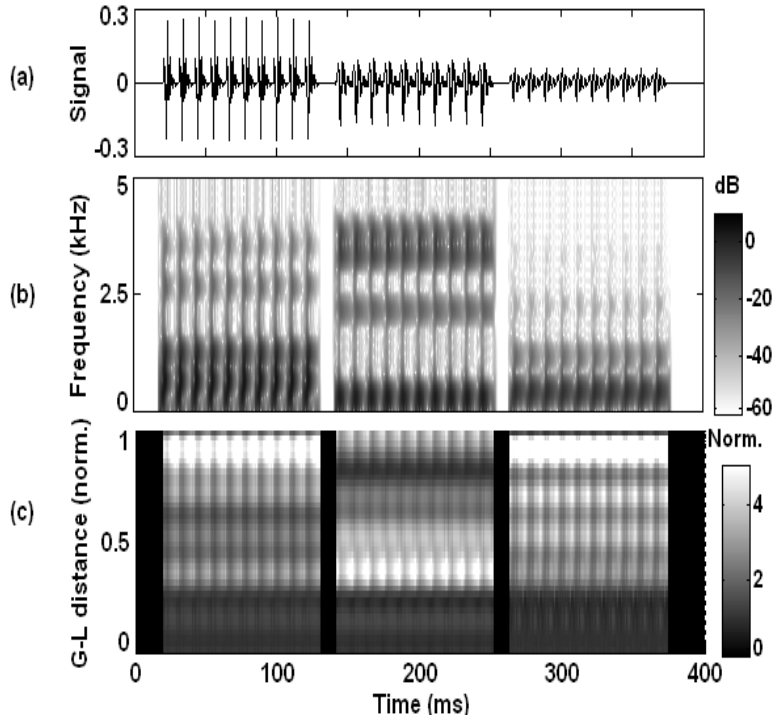

Figure 3.1. Vocal tract shape estimation of the synthesized vowel sequence /-a-i-u-/: (a) speech waveform, (b) wideband spectrogram, (c) areagram.

converted to 40 values using cubic-spline interpolation and square root of these values were used for plotting the areagram [51]. The square root of area values are used in subsequent analysis and display, although they are referred to as area values. Fig. 3.1 shows, for the vowel sequence /-a-i-u-/, the speech waveform, the wide-band spectrogram, and the areagram. The estimated area values are in accordance with the respective places of articulation for the three vowels: mid constriction (nearly neutral) for /a/, front constriction for /i/, and back constriction for /u/. However, there is a large variation in the area values as a function of time and the variations are related to the position of the analysis frame.

3.3 Windowed energy index for improving the consistency of the estimated area values

A plot of RMS value of the LPC prediction error showed a large variation with the analysis frame position, with the peaks occurring for the analysis frames coinciding with the starting of the glottal pulses. The frame positions corresponding to the minimum in the prediction error are likely to be related to the least estimation error in the vocal tract parameters. However, locating the peaks or the valleys of the LPC

prediction error consistently was found to be difficult. In the speech signal, a glottal closure instant (GCI) marks the beginning of the glottal excitation. The GCIs were obtained using Childers and Hu's algorithm [52] and it was found that the variation in the prediction error was related to the GCIs. However, the location of the frame positions for minimum error with respect to the GCIs was found to be different for the three vowels, making it difficult to use the GCI locations for automated selection of frame positions for minimizing the estimation error.

The LPC analysis using autocorrelation method involves application of Hamming window on the analysis frame. The prediction error was found to be related to the energy of the windowed frame, with the lowest values of the prediction error coinciding with the minima of the windowed energy for all the three vowels. For automating the selection of the frames with the minimum energy, a function called "windowed energy index" was calculated as the ratio of the energy of the windowed signal to the frame energy, and given as

$$E_w(n) = \sum_{m=0}^{N-1} [s_n(m)w(m)]^2 \Big/ \sum_{m=0}^{N-1} {s_n}^2(m) \qquad (3.1)$$

where $w(m)$ is the Hamming window function of length N samples and $s_n(m)$ is the speech segment for the frame position n. For steady state segments and frame length equal to an exact multiple of the pitch period, the function is periodic with period equal to the pitch period. It varies smoothly with easily detectable maxima and minima.

Figure 3.2 shows the synthesized vowel sequence /-a-i-u-/, the windowed energy index, and the RMS prediction error, plotted as a function of the analysis frame position. The GCI positions are marked on the speech waveform. Natural speech has significant jitter in pitch period. To examine the effect of the window length with respect to the pitch period, windowed energy index function was calculated for the window length equal to two pitch periods, and for window lengths decreased and increased by 10%, as shown in Fig. 3.2. The minima of the function, located using valley detection, are also marked. The function obtained with the window length of two pitch period has distinct minima and these correspond to the low values of the prediction error for all the three vowels. The function with the decreased window length also has distinct minima corresponding to the low values of prediction error, but its shape is different for the three vowels. Increased window

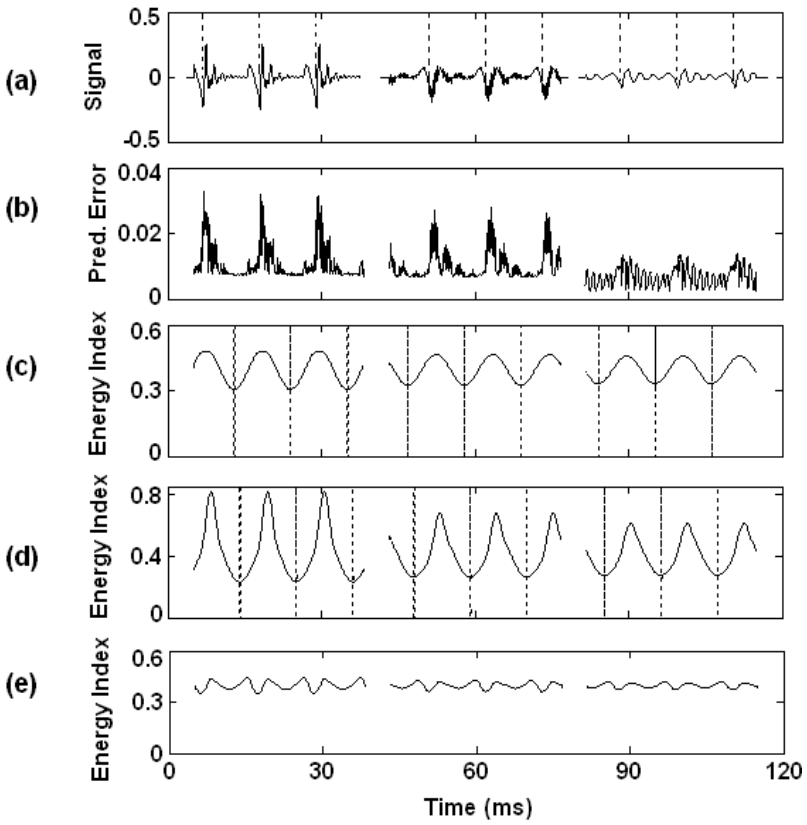

Figure 3.2. Speech waveform, windowed energy index, and RMS prediction error for three glottal cycles each of synthesized vowels /a/, /i/, /u/: (a) speech waveform with GCI marked by dotted lines, (b) RMS prediction error, (c) windowed energy index for window length of two pitch periods, with the minima marked as dotted lines, (d) windowed energy index with window length decreased by 10 %, (e) windowed energy index with window length increased by 10 %.

length results in a function with relatively indistinct minima. Thus the function can be used for locating the frame positions corresponding to low prediction error by using a window length equal to two pitch periods or slightly shorter. The technique was applied for estimating the vocal tract shape from speech signals of synthesized and natural vowels and vowel-semivowel-vowel utterances.

3.4 Results and discussion

In order to reduce the variability in the estimated area values, they were calculated at the frame positions corresponding to the E_w minima. Fig. 3.3 shows the resulting areagram for the synthesized vowel sequence /-a-i-u-/. The estimated vocal tract shape for all the three vowels shows much smaller variation as compared to that in the areagram in Fig. 3.1. For a visual comparison of the spread in the area values, the

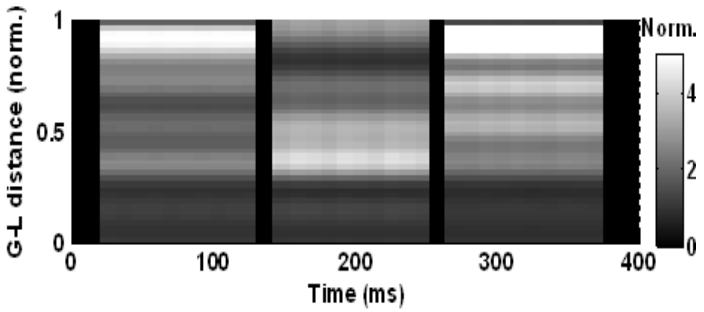

Figure 3.3. Areagram of the synthesized vowel sequence /-a-i-u-/ (same as in Fig. 3.1) as obtained for E_w–minima selected frames

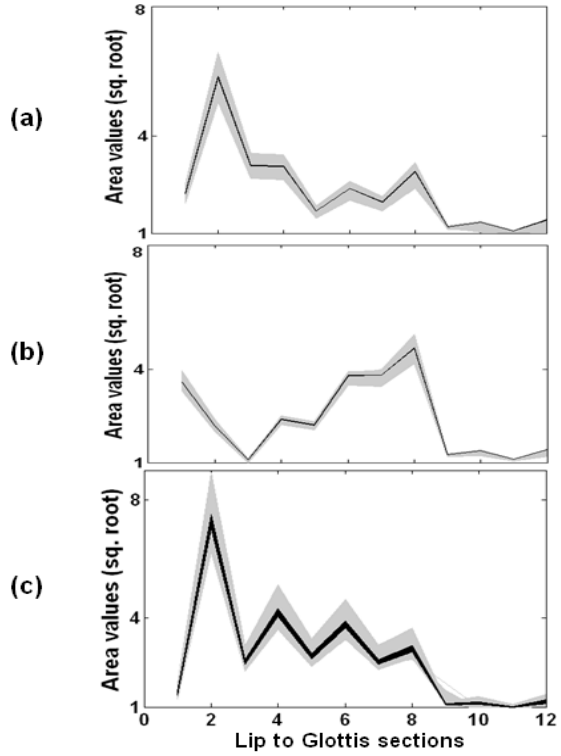

Figure 3.4. Plot of area values for synthesized vowels: (a) /a/ (b) /i/ (c) /u/ (shaded lines: original area values, solid dark lines: area values for E_w–minima selected frames).

plots of the estimated values for the twelve sections for all the frame positions are superimposed in Fig. 3.4. The plots of the area values for the frames selected at the E_W–minima are also superimposed in the same figure. The shaded lines show a large spread in the values, while the dark lines corresponding to the E_W–minima selected frames show a much smaller spread.

The variation in the estimated square root of the area values of each section was quantified by the max-min deviation (the difference between maximum and

Table 3.1 Mean value and max-min (m-m) deviation of square-root of vocal tract area values of 12 sections for three synthesized vowels.

(A) Inter-frame interval of 1 sample

Vowel	Parameter	Lips-to-glottis section number											
		1	2	3	4	5	6	7	8	9	10	11	12
/a/	mean	2.26	5.94	3.14	3.12	1.70	2.40	1.97	2.90	1.20	1.33	1.07	1.38
	m-m	0.61	1.56	0.79	0.78	0.40	0.60	0.46	0.82	0.12	0.37	0.08	0.41
/i/	mean	3.63	2.20	1.09	2.41	2.23	3.81	3.80	4.68	1.26	1.36	1.10	1.40
	m-m	0.66	0.38	0.14	0.29	0.27	0.45	0.54	0.92	0.09	0.20	0.07	0.25
/u/	mean	1.49	7.33	2.57	4.24	2.76	3.83	2.60	3.04	1.18	1.13	1.02	1.20
	m-m	0.53	2.71	0.87	1.52	0.95	1.40	0.84	1.07	0.74	0.43	0.15	0.44

(B) E_w-minima selected frames

Vowel	Parameter	Lips-to-glottis section number											
		1	2	3	4	5	6	7	8	9	10	11	12
/a/	mean	2.22	5.82	3.08	3.06	1.68	2.38	1.97	2.90	1.19	1.33	1.07	1.41
	m-m	0.02	0.05	0.02	0.02	0.01	0.02	0.01	0.02	0.00	0.00	0.00	0.01
/i/	mean	3.59	2.19	1.09	2.39	2.21	3.80	3.81	4.68	1.25	1.38	1.11	1.42
	m-m	0.01	0.00	0.00	0.01	0.01	0.02	0.01	0.01	0.00	0.01	0.00	0.01
/u/	mean	1.45	7.23	2.51	4.17	2.70	3.76	2.51	2.95	1.09	1.12	1.00	1.18
	m-m	0.07	0.35	0.12	0.21	0.13	0.18	0.12	0.15	0.03	0.07	0.02	0.09

[m-m : max-min deviation]

minimum values). The mean values and the max-min deviations for inter-frame interval of one sample are given in Table 3.1(A). The deviations for the three vowels are significant for all the sections and they are very large for some of the sections. The mean values and the max-min deviations for the E_w–minima selected frames are given in Table 3.1(B). As compared to the corresponding values in Table 3.1(A), there are no significant changes in the mean values. The max-min deviations have significantly decreased for all the sections. They have decreased by more than an order of magnitude, except for a few front sections in case of /u/.

To study the effect of E_w–minima selected frames on the vocal tract shape estimation during transitional tract configurations, the method was used for obtaining areagrams of synthesized vowel-semivowel-vowel utterances. Fig. 3.5 shows the speech waveform, wideband spectrogram, and areagrams for the utterance /aja/. It is

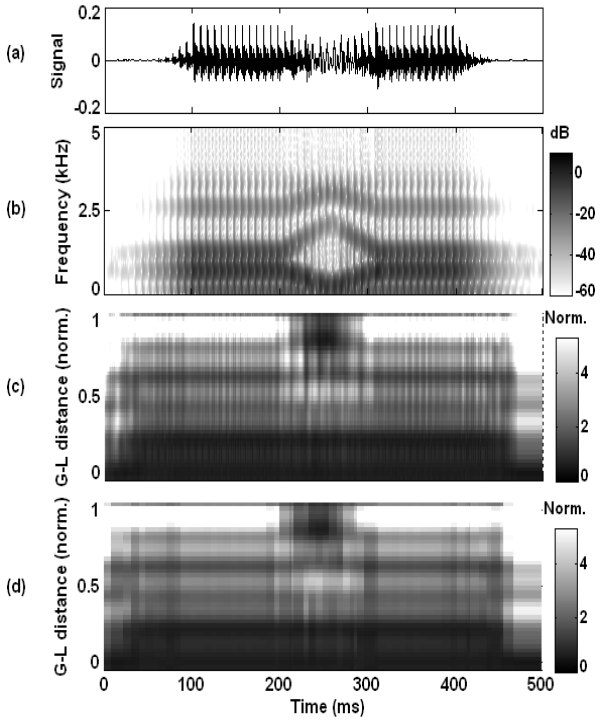

Figure 3.5. Estimated vocal tract shape variation of synthesized vowel-semivowel-vowel sequence /aja/. (a) speech waveform (b) wideband spectrogram (c) original areagram (d) areagram for E_w–minima selected frames.

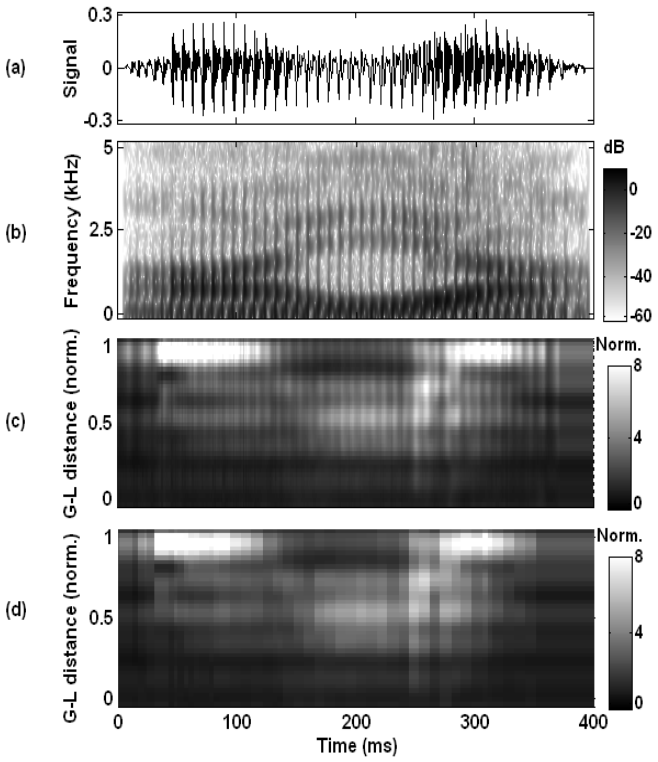

Figure 3.6. Estimated vocal tract shape variation of natural vowel-semivowel-vowel sequence /aja/. (a) speech waveform (b) wideband spectrogram (c) original areagram (d) areagram for E_w–minima selected frames.

observed that E_w–minima selected frames resulted in areagram with reduced variability during fixed tract configuration without smearing the changes in the area values during the transitional tract configuration

To examine the effectiveness of the method in estimating the vocal tract shape from natural speech signal, areagrams were obtained for vowel and vowel-semivowel-vowel utterances from several speakers. As an example, Fig. 3.6 shows the speech waveform, wideband spectrogram, and areagrams for the natural utterance /aja/ from a male speaker. It is observed that E_w–minima selected frames resulted in reduced variability in the area values during the fixed vocal-tract configuration without smearing the changes in the area values during the transitional configuration.

Chapter 4

LPC ANALYSIS METHODS FOR VOCAL TRACT SHAPE ESTIMATION

4.1. Introduction

Vocal tract shape during vowels can be estimated using Wakita's LPC based inverse filtering method [10]. In this analysis, the vocal tract along the length from glottis to lips was modeled as 12 cylindrical sections of equal length and the area values of the sections were calculated from the reflection coefficients obtained using the autocorrelation coefficients of the windowed frame. The all-pole filter is guaranteed to be stable in the autocorrelation method, but the windowing of the signal causes a reduction in spectral resolution [53]. However, stability is not always guaranteed in finite word length computations [54]. LPC analysis equation can also be solved using covariance method [55] and lattice method [53].

Krishnamurthy and Childers [56] compared the different methods for solving the linear prediction analysis equations and found that the closed phase covariance method provides better formant tracking as compared to autocorrelation and pitch synchronous covariance method. It was also observed that pitch synchronous covariance analysis gave good spectral estimates with comparatively fewer speech samples. The covariance analysis generally performs better as it considers the samples outside the frame interval. However, covariance analysis is computationally more intense and does not necessarily give a stable filter [57]. Atal [59] reported a method to modify the covariance analysis so that the estimated synthesis filter is always stable. However, this is done at the expense of introducing additional error in the estimation of LPC coefficients. Barnwell [53] studied the lattice method for LPC analysis using spectral estimation method suggested by Burg [58]. Lattice method provided good quality spectral estimate using smaller analysis frames than those used by autocorrelation method and the filter was always stable. But the algorithm increased the computation [57]. Since the speech-training aid is a non-real time application, certain delays in estimation and display of vocal tract shape are acceptable. Hence it is useful to investigate the different LPC analysis techniques for

estimating the vocal tract shape, even if they are somewhat computationally intensive. This chapter describes the methods for estimating the vocal tract shapes using LPC covariance and lattice based analysis.

4.2. LPC covariance analysis

LPC analysis involves estimating a set of predictor coefficients directly from speech signal in such a manner so as to obtain a good spectral estimate of the speech signal. The basic approach involves finding a set of predictor coefficients that minimize the mean-squared prediction error over a short segment of the speech waveform. The mean-squared prediction error sequence is defined as

$$\begin{aligned} E_n &= \sum_m e_n^2(m) \\ &= \sum_m \left[s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \right]^2 \end{aligned} \tag{4.1}$$

where $s_n(m)$ is a speech segment selected in the vicinity of sample n, α_k are the predictor coefficients. The set of equations used to obtain the values of α_k that minimize E_n is given by

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0) \qquad i=1, 2, ...p \tag{4.2}$$

where $\phi_n(i,k)$ is a correlation matrix given by

$$\phi_n(i,k) = \sum_m s_n(m-i) s_n(m-k) \quad i=1, 2, ...p \tag{4.3}$$

LPC analysis by autocorrelation is carried out by applying Hamming window on the analysis frames of the speech signal, with a short inter-frame interval to track the vocal tract shape changes. In this method, speech frame of length N is assumed to be zero outside the frame interval $0 \leq m \leq N\text{-}1$ and hence the prediction error $e_n(m)$, for a p^{th} order predictor, will be non-zero over the interval $0 \leq m \leq N\text{-}1+p$. The short-time average prediction error is defined as

$$E_n = \sum_{m=0}^{N+p-1} e_n^2(m) \tag{4.4}$$

The prediction error is large at the beginning of frame interval because signal samples are predicted from samples which are assumed to be zero. Similarly prediction error is

large at the end of the frame interval as the zero samples are predicted from samples that are non-zero. To alleviate this problem, Hamming window is used to taper the ends of the speech segment. In this approach, $\phi_n(i,k)$ of (4.2) is autocorrelation matrix and is a Toeplitz matrix, which can be solved efficiently using Robinson's recursive algorithm [10].

In the covariance analysis, the interval over which the prediction error is computed is fixed between 0 and *N*-1. The short-time average prediction error is defined as

$$E_n = \sum_{m=0}^{N-1} e_n^2(m) \tag{4.5}$$

For predicting the samples at the beginning of the frame interval, *p* samples before the interval $0 \leq m \leq N\text{-}1$ are used. Thus, there is no need of applying a tapering window on the speech segment in the covariance analysis. In this method, $\phi_n(i,k)$ is called covariance matrix. The covariance matrix is symmetric but not Toeplitz. Hence, Cholesky decomposition is used to solve the linear equations to compute the LPC coefficients. For this method, the (4.2) is expressed in matrix notation as

$$\mathbf{\Phi\alpha = \psi} \tag{4.6}$$

where $\mathbf{\Phi}$ is the covariance matrix with $(i, k)^{th}$ element $\phi_n(i,k)$, and $\boldsymbol{\alpha}$ and $\boldsymbol{\psi}$ are column vectors with elements α_k and $\phi_n(i,0)$ respectively. The covariance matrix is expressed in the form

$$\mathbf{\Phi = VDV^t} \tag{4.7}$$

where $\mathbf{V}$ is a lower triangular matrix, and $\mathbf{D}$ is diagnoal matrix. Once the $\mathbf{V}$ and $\mathbf{D}$ have been determined by solving for $(i, k)^{th}$ element on both sides of (4.7), the predictor coefficient vector $\boldsymbol{\alpha}$ can be obtained in a two-step procedure using the following equations

$$\mathbf{VDV^t\alpha = \psi} \tag{4.8}$$

which can be written as

$$\mathbf{VY = \psi} \tag{4.9}$$

Thus $\boldsymbol{\alpha}$ can be calculated if we know the $\mathbf{V}$ and $\mathbf{D}$ matrices.

Wakita has demonstrated the process of obtaining the reflection coefficients from the autocorrelation coefficients using Robinson's algorithm. However, it cannot be used to obtain the reflection coefficients in the covariance method. Hence, initially

LPC coefficients are obtained from the covariance matrix using Cholesky decomposition. Reflection coefficients are then obtained from the LPC coefficients using Robinson's recursive relation [10].

4.3. LPC lattice based analysis

Lattice method involves minimizing the sum of mean-squared forward and backward prediction errors. In this method, the predictor coefficients are directly obtained from the samples of the speech signal instead of solving the set of linear equations represented by matrix of correlation values. The all-pole filter is guaranteed to be stable even when the computation is performed using the finite word length [54]. The system function of i^{th} order prediction error filter at the i^{th} stage of recursion according to Robinson's algorithm [10] can be defined as,

$$A^{(i)}(z) = 1 - \sum_{k=1}^{i} \alpha_k^{(i)} z^{-k} \tag{4.10}$$

where $\alpha_j^{(i)}$, $j = 1, 2, ..., i$ are the LPC coefficients. The prediction error at i^{th} stage is given by

$$e^{(i)}(m) = s(m) - \sum_{k=1}^{i} \alpha_k^{(i)} s(m-k) \tag{4.11}$$

This equation suggests that we are attempting to predict $s(m)$ from the past i samples of the input $s(m)$. Thus $e^{(i)}(m)$ is called as forward prediction error sequence. Applying z transform on (4.11) gives

$$E^{(i)}(z) = A^{(i)}(z) S(z) \tag{4.12}$$

The Robinson's recursion for the LPC coefficients at the i^{th} stage is given by

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \tag{4.13}$$

By using (4.13) into (4.10), $A^{(i)}(z)$ can be expressed in terms of $A^{(i-1)}(z)$ as

$$A^{(i)}(z) = A^{(i-1)}(z) - k_i z^{-i} A^{(i-1)}(z^{-1}) \tag{4.14}$$

Substituting (4.14) into (4.12) we obtain

$$E^{(i)}(z) = A^{(i-1)}(z)S(z) - k_i z^{-i} A^{(i-1)}(z^{-1})S(z) \tag{4.15}$$

The first term in the above equation is the z-transform of the prediction error for an (i-1)$^{\text{th}}$ order predictor. For interpretation of second term, $B^{(i)}(z)$ is defined as

$$B^{(i)}(z) = z^{-i} A^{(i)}(z^{-1})S(z) \tag{4.16}$$

The inverse transform of $B^{(i)}(z)$ can be obtained as

$$b^{(i)}(m) = s(m-i) - \sum_{k=1}^{i} \alpha_k^{(i)} s(m+k-1) \tag{4.17}$$

This equation suggests that we are attempting to predict $s(m-i)$ from the i samples that follow the input $s(m-i)$. Thus $b^{(i)}(m)$ is called as backward prediction error sequence. Now using (4.15) and (4.17), the forward prediction error sequence $e^{(i)}(m)$ can be expressed as

$$e^{(i)}(m) = e^{(i-1)}(m) - k_i b^{(i-1)}(m-1) \tag{4.18}$$

Similarly backward prediction error sequence can be expressed as

$$b^{(i)}(m) = b^{(i-1)}(m-1) - k_i e^{(i-1)}(m) \tag{4.19}$$

Thus using the recursion (4.18) and (4.19) we can obtain the forward and backward prediction error sequences for an i^{th} order predictor in terms of the corresponding prediction errors of an (i-1)$^{\text{th}}$ order predictor provided zeroth order predictor values are assumed as

$$e^{(0)}(m) = b^{(0)}(m) = s(m) \tag{4.20}$$

The k parameters can be obtained by minimizing the sum of the mean-squared forward and backward prediction errors as described by Burg [58]. The expression for k_i is given by

$$k_i = \frac{2\sum_{m=0}^{N-1}\left[e^{(i-1)}(m)b^{(i-1)}(m-1)\right]}{\sum_{m=0}^{N-1}\left[e^{(i-1)}(m)\right]^2 + \sum_{m=0}^{N-1}\left[b^{(i-1)}(m-1)\right]^2} \tag{4.21}$$

The magnitude of k_i parameters obtained using (4.21) will always be less than unity which is a necessary and sufficient condition for stability of predictor polynomial i.e., roots of the $A(z)$ lie inside the unit circle. Thus lattice method always yields stable predictor polynomial. The predictor coefficients can be obtained from the k_i parameters using (4.13).

The reflection coefficients μ_m at the junction between sections m and $m+1$ were directly calculated from the k_i parameters using the relation obtained by Wakita [10] given by (2.7). Vocal tract area functions are obtained from the reflection coefficients using (2.8).

4.4. Results and discussion

The speech signal was sampled at 10 kHz with 16-bit quantization and first difference of the signal was taken for providing an approximate 6-dB/octave pre-emphasis. The LPC analysis of order 12 was carried out on analysis frames of duration equal to twice the average pitch period, and the successive analysis frames were shifted by 5 ms. In this analysis, the vocal tract from glottis to lips was modeled as 12 cylindrical sections of equal length and the area values of the sections were calculated from the reflection coefficients obtained from LPC analysis of the analysis frame. To eliminate the variations contributed by natural changes in phonation, the analysis was carried out on the vowel sequence /-a-i-u-/ synthesized using Klatt synthesizer [50] with a constant pitch frequency of 90 Hz and constant amplitude. The area values of the twelve sections were converted to 40 values using cubic-spline interpolation and square root of these values were used for plotting the areagram. Fig. 4.1 shows, for the vowel sequence /-a-i-u-/, the speech waveform, the original areagram obtained using autocorrelation analysis, covariance analysis, and lattice based analysis. The estimated area values are in accordance with the respective places of articulation for the three vowels: mid constriction (nearly neutral) for /a/, front constriction for /i/, and back constriction for /u/. However, there is a large variation in the area values obtained using autocorrelation as a function of time. The estimated vocal tract shape obtained by covariance analysis shows much smaller variation as compared to that estimated using autocorrelation analysis for all the three vowels. The estimated vocal tract shape obtained by lattice based analysis shows much smaller variation as

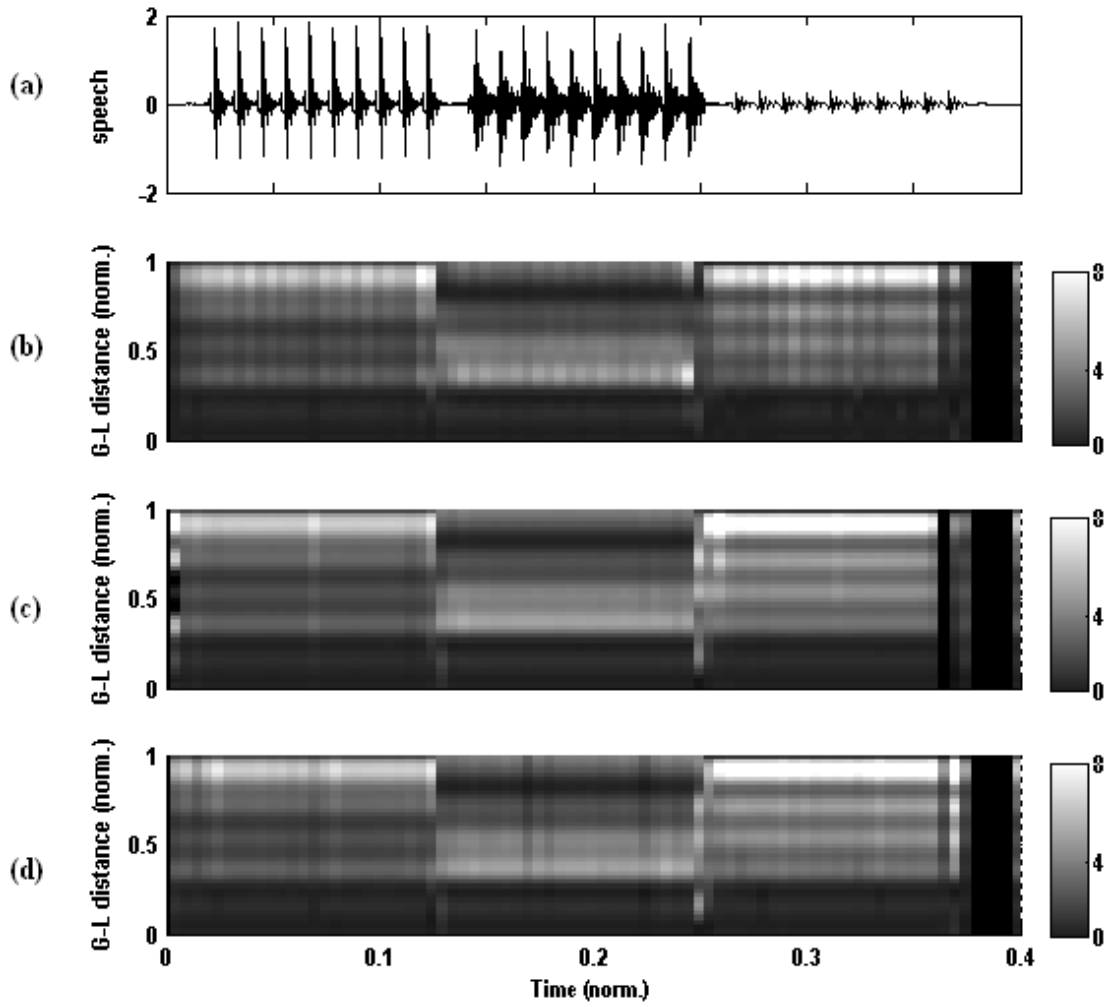

Figure 4.1. Vocal tract shape estimation of the synthesized vowel sequence /-a-i-u-/: (a) speech waveform, (b) areagram obtained by autocorrelation analysis, (c) areagram obtained using covariance analysis, (d) areagram obtained using lattice based analysis.

compared to that estimated using autocorrelation analysis for all the three vowels.

To examine the effectiveness of the method in estimating the vocal tract shape from natural speech signal, areagrams were obtained for vowel and vowel-semivowel-vowel utterances from several speakers. As an example, Fig. 4.2 shows the speech waveform and areagrams for the natural utterance /aja/ from a male speaker for covariance and lattice analysis. It is observed that covariance analysis resulted in reduced variability in the area values during the fixed vocal-tract configuration without smearing the changes in the area values during the transitional configuration. Similarly, it is observed that lattice analysis resulted in reduced variability in the area values during the fixed vocal-tract configuration without smearing the changes in the area values during the transitional configuration.

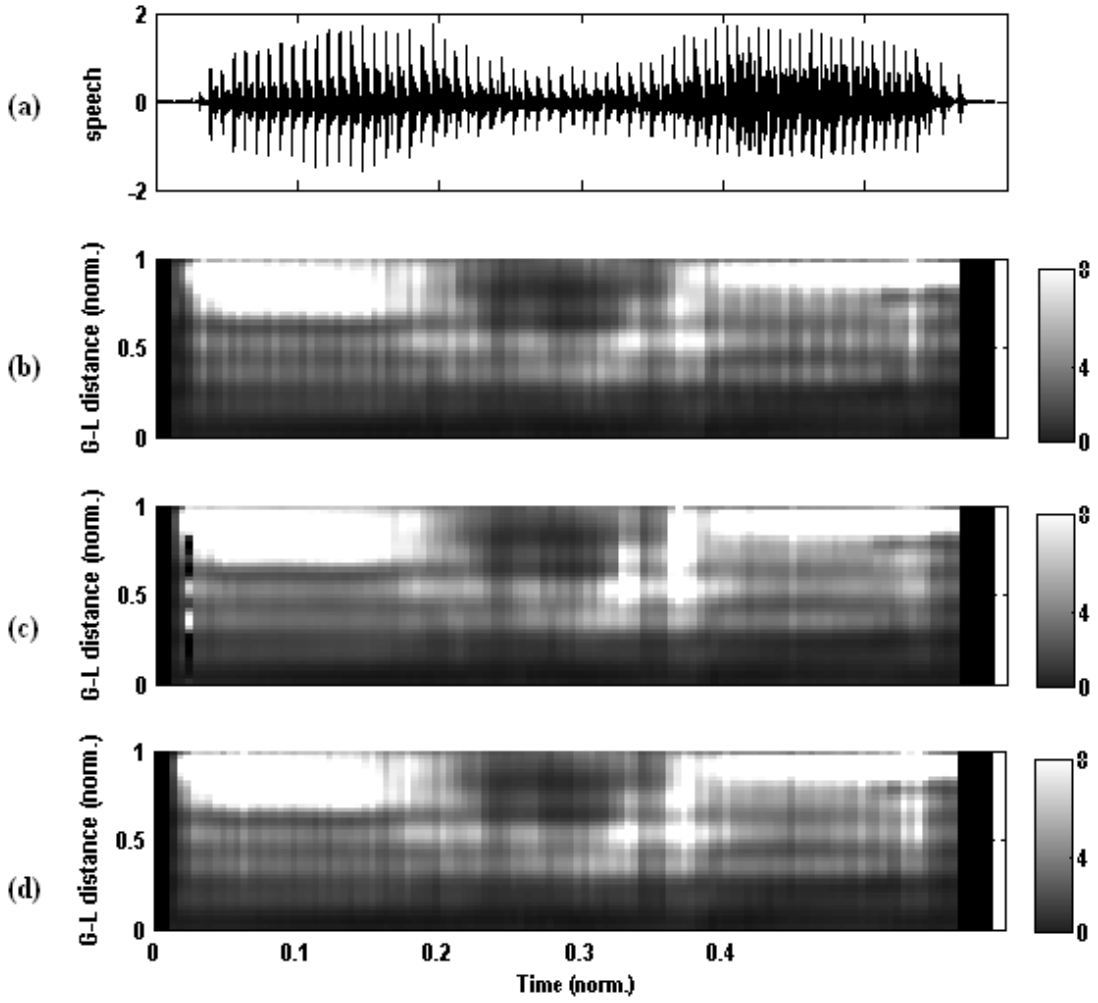

Figure 4.2 Estimated vocal tract shape variation of natural vowel-semivowel-vowel sequence /aja/: (a) speech waveform, (b) areagram obtained by autocorrelation analysis, (c) areagram obtained using covariance analysis, (d) areagram obtained using lattice based analysis.

However, for some of the utterances, it was observed that covariance analysis resulted in the roots of the predictor polynomial lying outside the unit circle leading to an unstable filter. The LPC lattice always resulted in stable predictor polynomial. Thus LPC lattice based analysis can be used for reducing the variability in the estimated values and improving the consistency of vocal tract shape estimation, without smearing the variations in the shape during transitional tract configuration.

Chapter 5

TRANSITION SEGMENT DETECTION FOR ESTIMATING PLACE OF ARTICULATION OF ORAL STOPS

5.1. Introduction

Vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterances can be estimated by using a bivariate surface model fitted on the vocal tract shapes during the transition segments preceding and following the stop closure [12]. Use of least-squares based bivariate quadratic surface modeling resulted in satisfactory estimation of place of closure for different unvoiced and voiced oral stops. The estimated place of closure for /ʌCa/ utterances, involving stop consonants /b/, /d/ and /g/ from 20 male and 20 female speakers , showed a good match (with a correlation coefficient of 0.94) with those obtained from direct X-ray imaging in XRMB database [20]. The accuracy of the estimation was highest if the segments used for the bivariate modeling corresponded to the VC transition segment preceding the closure and the CV transition segment following the closure [13]. An inclusion of the closure segment or the closure release burst can introduce errors in the surface modeling. Inclusion of any significant part of the vowel on either side can decrease the sensitivity of the fitted model in correctly interpolating the place of closure. Hence a method for automatically and accurately detecting the VC and CV transition boundaries is needed for improving the accuracy of the estimated place of closure and improving the feedback given by the visual aids for speech training.

Several methods for detecting the acoustic landmarks associated with stop consonants have been reported [60]-[62], but they do not mark the start and end points of the transitions. Onset of voicing after the stop closure is generally detected as the onset of periodicity in the acoustic waveform. Francis *et al.* [63] compared the commonly used methods and concluded that presence of aspiration decreased the detection accuracy. Further, these methods are not useful in detecting the CV transition. Closure release burst can be located with a good temporal accuracy using

rate of change of spectral moments [61] , but it cannot be used to estimate the end of the CV transition. Formant transitions during VC and CV transitions cannot be used to detect the transition segment boundaries because accurate formant tracking using spectral analysis is still a challenging task [64].

It is observed that the plots of area values as a function of time and position along the vocal tract length, for VCV syllables with different oral stops, have a distinct two-dimensional pattern during the transition segments as compared to the steady state segments. A method based on a measure of the rate of change in the area values for marking the VC and CV transition segments is described in the following section.

5.2. Bivariate surface modelling for estimation of place of articulation in oral stops

Vocal tract shapes were estimated using Wakita's method [10]. The speech signal was sampled at 10 kHz with 16-bit quantization and a first difference of the signal was taken for providing an approximate 6-dB/octave pre-emphasis. Hamming window was applied on analysis frames with duration equal to twice the average pitch period, and the successive analysis frames were shifted by 5 ms. The LPC analysis of order 12 was used to compute the reflection coefficients from the autocorrelation coefficients of the windowed frame. In this method, the vocal tract from glottis to lips was modeled as 12 cylindrical sections of equal length and the vocal tract area function was obtained as ratios of the areas on both sides of the section interfaces and these area ratios were converted into areas by assuming a constant normalized area of unity at the glottis end. The amount of opening in a section of the vocal tract is obtained as the square root of the area value. These values are used in subsequent analysis and display, although they are referred to as area values.

LPC based vocal tract shape estimation fails during stop closure due to unavailability of relevant spectral information. However, the vocal tract shape during the stop closures of VCV utterances can be estimated by bivariate polynomial modeling of the shapes during transition segments preceding and following the stop closure as they tend to show different two-dimensional patterns for different places of closure [12].

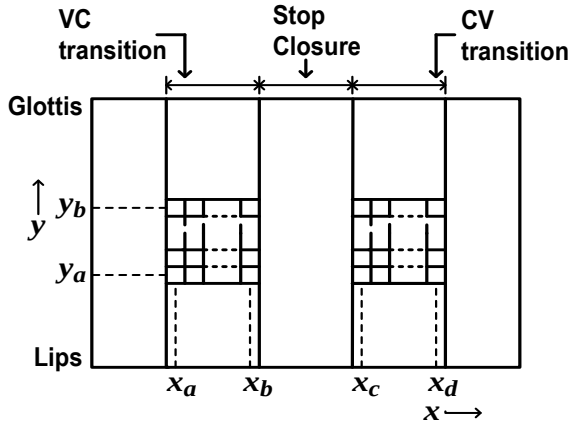

Figure 5.1. Selection of transition area values for 2D modeling and interpolation [12].

The estimated vocal tract area values, given as $g(x, y)$ at analysis frame x (along the time axis) and the area section number y from lip end, during the VC and CV transition segments were modeled as bivariate quadratic function. The polynomial coefficients were obtained for minimizing the sum of the squared approximation error. The area section numbers $y_a \le y \le y_b$ for the analysis frames $x_a \le x \le x_b$ and $x_c \le x \le x_d$ corresponding to VC and CV transition segments, respectively, are used for surface modeling as shown in Fig. 5.1. The number of frames used for surface modeling in VC and CV transition segments are $L_{col} = x_b - x_a + 1$ and $R_{col} = x_d - x_c + 1$, respectively. The functions with the estimated coefficients were used for interpolating $g(x, y)$ during the closure segment ($x_b \le x \le x_c$).

In [12], the boundary locations (x_a , x_b , x_c , x_d) were estimated using a two-step process. The beginning and ending points of the VCV utterance were estimated using the short-time average magnitude of the signal [65]. The stop closure boundaries (x_b , x_c) were marked using an empirical threshold of 0.2 times the RMS value of the signal waveform between the two end-points of the VCV utterance. The end-location of the stop closure was delayed to exclude the closure burst. The start of the VC transition and the end of the CV transition segments (x_a , x_d) were marked by empirically set values.

The transition segment lengths vary across speakers and it is particularly difficult to empirically set the values for utterances with speech impairment, as needed in speech training aids.

5.3. Transition Segment Detection

Investigations were carried out on vocal tract shape estimates for VCV utterances involving oral stops. It has been shown, earlier in Chapter 3, that the area values obtained at the minima of the windowed energy index (the ratio of the energy of the windowed signal to the energy of the signal within the frame) can be used for improving the consistency of vocal tract shape estimation. This method was used to interpolate the values for frames at 5 ms intervals. Figure 5.2 shows analysis results for isolated VCV utterance /ata/. The areas for first 6 sections are relatively steady as the vocal tract shape does not change significantly during the vowel. The estimated values show transition during the VC and CV transitions, and the transitions extend over approximately the same interval as the formant transitions. Thus a measure of the rate of change of the vocal tract shape may be useful in marking the transition segments during VCV utterances.

5.3.1 Computation of rate of change

A rate of change measure of the two-dimensional pattern of the area values may be computed by combining the rate of change of individual section area values or by bivariate surface modeling of the area values.

Rate of change of the section area values based on first difference was found to be noisy and hence slope of moving multi-point linear regression was used. Moving 7-point linear regression (with the segment used for regression centered on the current frame) resulted in a good compromise between the requirements of ripple rejection and transition detection. The combined rate of change is computed using the root-mean-square of slopes of the first 6 sections. Figure 5.2(i) shows this combined rate of change. It has distinct minima at the beginning of VC transition and at the end of the CV transition. Its peaks are seen to be associated with the end of the VC transition and beginning of the CV transition. However, an examination of the plots of this measure for different utterances from several speakers showed that the minima and the peaks were not consistently associated with the transition end points. Hence this measure is not suitable for automated marking of the transitions.

Time-slope of the moving bivariate surface fitted on the area values was used as another measure of rate of change. We used linear and quadratic approximations.

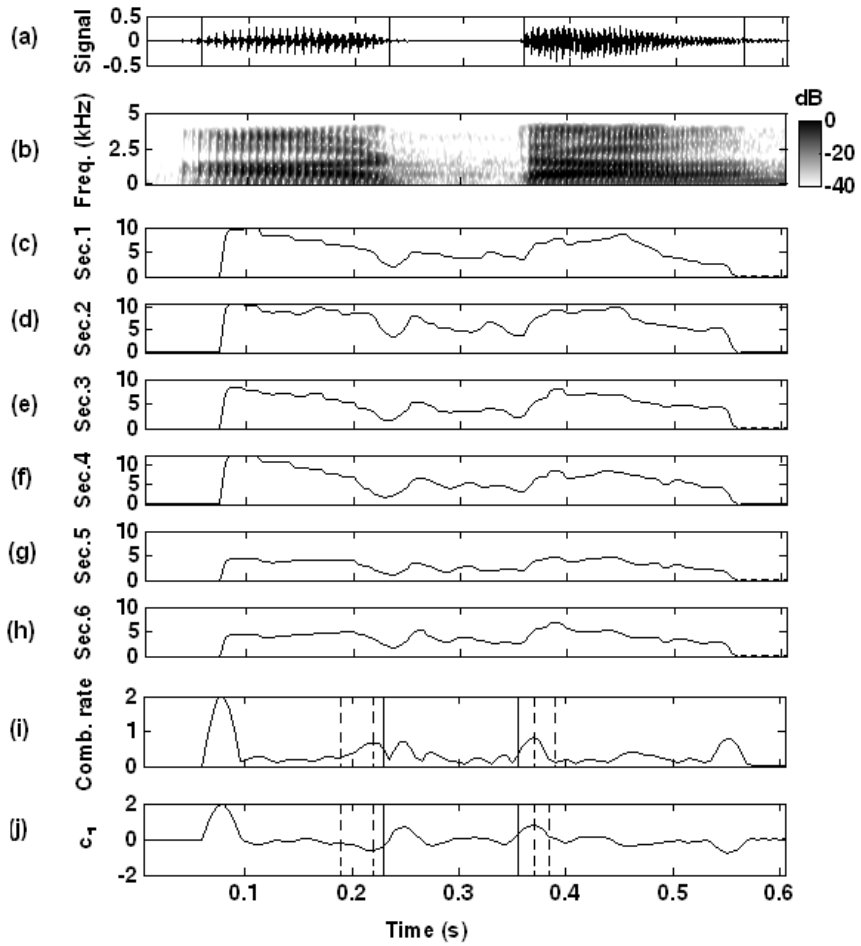

Figure 5.2. Analysis results for /ata/ from a male speaker: (a) waveform for 0.6 s; (b) wideband spectrogram; (c)-(h) area ratios of first 6 sections starting from lips; (i) combined rate of change of area ratios; (j) time-slope from bivariate linear approximation.

The bivariate linear approximation is given by

$$g(x,y) = c_0 + c_1 x + c_2 y + \varepsilon(x,y) \tag{5.1}$$

where $g(x,y)$ is the estimated area value at analysis frame x (along time axis) and the area section number y from lip end, and ε is the approximation error. Modeling is carried out for first six sections, i.e. $1 \le y \le 6$ and for moving 7-frame segment centered at the current frame x_n, i.e. $x_n - 3 \le x \le x_n + 3$. The coefficients $c_0 - c_2$ are obtained for minimizing the sum of the squared approximation error. In matrix notation, the bivariate linear polynomial can be expressed as,

$$\mathbf{Az} = \mathbf{b} + \mathbf{r} \tag{5.2}$$

where **r** represents the approximation error, and **A**, **z**, and **b** matrices are given by

$$\mathbf{A}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ x_n-3 & x_n-3 & \cdots & x_n-3 & x_n-2 & \cdots & x_n+3 \\ 1 & 2 & \cdots & 6 & 1 & \cdots & 6 \end{bmatrix}$$

$$\mathbf{z}^T = \begin{bmatrix} c_0 & c_1 & c_2 \end{bmatrix}$$

$$\mathbf{b}^T = \begin{bmatrix} g(x_n-3,1) & g(x_n-3,2) & \cdots & g(x_n+3,6) \end{bmatrix}$$

The polynomial coefficient vector **z** is obtained for minimizing the sum of the squared errors as,

$$\mathbf{z} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{5.3}$$

The coefficient c_1 is the time-slope and is taken as the combined rate of change of area values along time axis x. Figure 5.2(j) gives a plot of c_1. It has a small value during the steady state vowel segments. At the start of the VC transition, the constriction begins and hence c_1 becomes negative, reaching a negative peak almost near the closure. After the closure release, the opening starts increasing. This is indicated by positive value of c_1. Its value reaches a peak and then falls to almost zero near the end of CV transition. This pattern was observed across the utterances examined from 6 consonants across 4 speakers, indicating the suitability of c_1 for marking the VC and CV transitions. It may be noted that the peaks and valleys in c_1 may occur at the start and end of the utterance and also during the closure, but being outside the search region they do not affect the transition detection.

Similar investigation was carried out using bivariate quadratic approximation given by

$$g(x,y) = c_0 + c_1x + c_2y + c_3xy + c_4x^2 + c_5y^2 + \varepsilon(x,y) \tag{5.4}$$

Plots of the coefficients showed that the change in vocal tract shape during the transitions was indicated by c_1 and c_3 both, but neither of them captured the change as effectively and consistently as c_1 from bivariate linear approximation.

5.3.2 Detection of transition segment boundaries

An example of markings for locating the boundaries of the transition segments is shown in Fig. 5.2. The start and end of the stop closure are estimated using short time average magnitude with a threshold value of 0.2 times the RMS value of the signal [12]. These points marked as solid vertical lines on the waveform in Figure 5.2(a), are used as first estimates of the VC transition end and the CV transition start, respectively. Moving left from the closure, a negative peak c_{1n} is searched to mark a refined estimate of the VC transition end. The values in the frames further left are compared with a threshold to mark the VC transition start. To make the detection adaptive to different rates of transition in the utterances, the threshold is kept as $0.2c_{1n}$ for the first 5 frames and then changed by $0.1c_{1n}$ for each frame. Similarly, a positive peak c_{1p} is searched on the right side of the closure to mark the CV transition start. The positive peak tends to enter into the CV transition during aspirated stops, leading to an error in marking the CV transition start. In such stops, the time delay between the short time average magnitude based estimate and the positive peak was observed to be consistently more than 15 ms. Hence if the time delay between the initial estimate and the positive peak is more than 15 ms, a point corresponding to $0.5c_{1p}$ is located before the positive peak and is marked as the CV transition start. The values in the frames further right are compared with a threshold to mark the CV transition end. The threshold is kept as $0.2c_{1p}$ for the first 5 frames and then changed by $0.1c_{1p}$ for each frame. The detected transition start and end points are marked by dotted lines in Fig. 5.2(j).

5.4. Results

Utterances of the type /aCa/ with stop consonants /p,b,t,d,k,g/ from two male and two female speakers were analyzed for estimation of place of maximum constriction during the stop closures. The start and end points of VC and CV transitions were marked by the method presented in the previous section. The accuracy of the automated markings was evaluated by comparing them with the manual markings obtained by visual inspection of the wideband spectrograms using PRAAT [66]. The errors in the marking of the four transition points were calculated. There was one

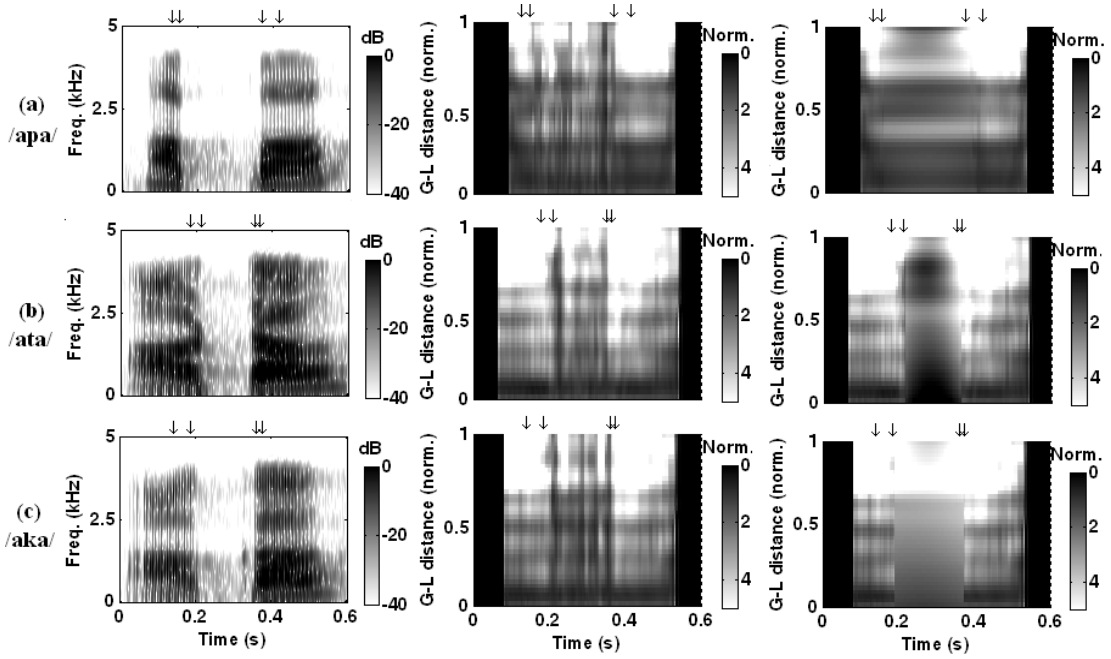

Figure 5.3 Interpolation results for the VCV utterance with unvoiced bilabial, alveolar and velar stops: (a)/apa/; (b) /ata/; and (c) /aka/. Left side: wideband spectrogram, middle: original areagram, right side: areagram obtained using bivariate quadratic surface modeling of the automatically detected transition segments.

large error, of 30 ms in marking VC transition start for /ada/ from one speaker. As the transition in this utterance was much longer than the error, the error did not affect the bivariate modeling. All the other errors were within ±15 ms. The range, mean and standard deviation of the errors are given in Table 5.1.The standard deviations are less than 10 ms, indicating a match with the manually marked locations.

The marked transition segments were used for estimating the vocal tract shape during the stop closure using the bivariate quadratic modeling and interpolation. The results for VCV utterances with stops /p, t, k/ are shown in Fig. 5.3. The vocal tract area values are displayed as areagram, a two-dimensional plot of the square root of area values as gray levels as a function of time along x-axis and normalized distance from glottis towards lips (G-L) along y-axis. The four arrows indicate the detected transitions. The largest constrictions during the closure in the areagrams are at the normalized distance of 1, 0.8, and 0.6 for /p/, /t/, and /k/, respectively, matching with the values of 1, 0.75 – 0.89, and 0.47 – 0.70, for the bilabial, alveolar, and velar oral stops, respectively, as estimated from MRI [66] and X-ray images [68].

In order to investigate the effect of lattice-based LPC analysis and windowed energy index (E_w) based frame selection (E_w method) on the estimation of transition segment detection, transition segments were detected using vocal tract area values

Table 5.1: Errors in transition segment estimation

Transition point	Error (ms)			
	Range		Mean	Std. dev.
VC transition start	-10	30	8.1	8.3
VC transition end	-15	15	1.2	7.4
CV transition start	-5	10	2.7	3.9
CV transition end	-15	10	-1.0	7.7

Table 5.2: Comparison of errors in transition segment estimation by various methods

Method	Values	Error in transition points (ms)			
		VC_start	VC_end	CV_start	CV_end
AR	Mean	8.1	1.3	2.7	-1.0
	Std. dev.	8.3	7.4	3.9	7.7
	+ve Max	30	15	10	10
	-ve Max	-10	-15	-5	-15
AR_E_w	Mean	7.1	2.7	1.0	-3.1
	Std. dev.	8.5	6.3	4.9	10.0
	+ve Max	25	15	10	25
	-ve Max	-15	-15	-10	-25
Lattice	Mean	5.6	2.1	1.3	-1.7
	Std. dev.	8.6	6.2	8.0	9.7
	+ve Max	15	15	10	25
	-ve Max	-15	-10	-30	-20
Lattice_E_w	Mean	5.6	1.7	2.7	-2.7
	Std. dev.	7.8	9.4	6.8	10.7
	+ve Max	20	15	25	25
	-ve Max	-5	-30	-10	-25

AR - Autocorrelation analysis, AR_E_w- Energy index analysis with area values estimated using autocorrelation, Lattice - lattice based analysis, Lattice_E_w - Energy index analysis with area values estimated using lattice analysis.

estimated from autocorrelation (AR), and lattice based analysis, and by their combination with E_w method. Table 5.2 shows the errors in the marking of the four transition points for the different methods of analysis. The mean and standard deviations by all the methods are comparable. The maximum positive error and maximum negative error are low for the AR-based method compared to other methods. Thus even though lattice and E_w methods provide consistent vocal tract shape estimates, they do not help in improving the automated marking of transition segments. It may be noted that coefficient c_1 used in the transition detection is obtained by fitting a bivariate linear model on the area values. This model rejects short time ripples in the values, and therefore improved consistency in the values obtained by E_w method does not improve the estimation. Hence it is concluded that the

values obtained by AR or lattice method, without smoothening by E_w method, should be used for detection of transition points. The values obtained by E_w method should be used only in bivariate surface modelling and interpolation.

Chapter 6

SUMMARY AND CONCLUSIONS

Speech-training aids providing visual feedback of articulatory motion are found to be useful in teaching lingual consonants and vowels. It is essential to improve the consistency and accuracy of the vocal tract shape for improving the feedback given by the visual aids for speech training. The investigations showed that vocal tract shape estimated for steady-state vowel varied with the position of the LPC analysis frame. Vocal tract shape estimated from the analysis frames positioned at the frames with minimum energy resulted in low prediction error and significantly reduced variability in the estimated area values. It is shown that minima of the windowed energy index (ratio of the windowed energy to the frame energy), detected by valley picking, can be used for selecting the frame positions for reducing the variability in the estimated values and improving the consistency of vocal tract shape estimation, without smearing the variations in the shape during transitional tract configuration.

Different LPC analysis techniques were investigated for estimation of the vocal tract shape of vowels. LPC lattice method and covariance method showed reduced variability in vocal tract shapes of steady state vowels. However, LPC covariance analysis resulted in unstable predictor polynomial for some utterances. The predictor polynomial obtained by lattice based analysis was guaranteed to be stable.

A technique using the slope of the bivariate linear approximation fitted on the vocal tract area function for automatically marking the VC and CV transitions during VCV utterances has been investigated. The automatically marked points have a good match with the manually marked ones, and resulted in satisfactory estimates of place of closure of oral stops.

The techniques emerging from these investigations have to be incorporated as part of visual speech-training aid. The windowed energy based method has to be used to estimate the VC and CV transition area values during VCV utterances and these values can be used to more accurately estimate the place of articulation during stop

closures of VCV utterances. Technique for automatically marking the VC and CV transition segments based on bivariate linear approximation has to be investigated for its effectiveness in marking the transition segments in impaired speech. LPC lattice analysis based estimation of vocal tract shapes for VCV utterances has to be investigated.

REFERENCES

[1] H. Levitt, J. M. Pickett, and R. A. Houde, (Eds.), *"Speech training aids,"* part VII in Sensory Aids for the Hearing Impaired. New York: IEEE Press, 1980, pp. 349-419.

[2] R. S. Nickerson and K. N. Stevans, "Teaching to a deaf: can a computer help?," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, no. 5, pp. 445-455, 1973.

[3] J. J. Mahshie, "Feedback considerations for speech training systems," in *Proc. IEEE Int. Conf. Spoke., language*, 1996, Philadelphia, pp. 153-156.

[4] S. A. Zahorian and S. Venkat, "Vowel Articulation Training Aid for the Deaf," in *Proc. lnt. Conf. on Acoust., Speech, and Signal Process.,*1990, New York, pp. 1121-1124.

[5] S. G. Fletcher, "Seeing speech in real time," *IEEE Spectrum*, vol. 19, no. 4, pp. 42–45, 1982.

[6] E. H. Nober, "Articulation of the deaf," *Exceptional Children*, vol. 33, no. 9, pp. 611-621, 1967.

[7] R. G. Crichton and F. Fallside, "Linear prediction model of speech production with applications to deaf speech training," *Proc. IEE Control Sci.,* vol. 121, pp. 865–873, 1974.

[8] J. M. Pardo, "Vocal tract shape analysis for children," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* 1982, pp. 763–766.

[9] S. H. Park, D. J. Kim, J. H. Lee, and T. S. Yoon, "Integrated speech training system for hearing impaired", *IEEE Trans. Rehab. Eng.,* vol. 2, no. 4, pp. 189–196, Dec. 1994.

[10] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.,* vol. AE-21, no. 5, pp. 417–427, 1973.

[11] H. Wakita, “Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art,” *IEEE Trans. Acoust., Speech, Signal Process.,* vol. ASSP-27, no. 3, pp. 281–285, 1979.

[12] P. C. Pandey and M. S. Shah, “Estimation of place of articulation during stop closures of vowel-consonant-vowel utterances,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 2, pp. 277-286, 2009.

[13] M. S. Shah, “Estimation of place of articulation during stop closures of vowel-consonant-vowel syllables,” *Ph.D. dissertation*, Dept. Elect. Eng., Indian Inst. of Technology Bombay, India, 2008.

[14] Video Voice Speech Training System, (Micro Video Corp., Ann Arbor, Michigan, 2003). [Online]. Available: http://www.videovoice.com (Last accessed in May, 2012).

[15] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, “The pedagogy-technology interface in computer assisted pronunciation training,” *Comput. Assisted Language Learn.*, vol. 15, pp. 441–467, 2002.

[16] H. R. Javkin, N. A. Barroso, A. Das, D. Zerkle, Y. Yamada, N. Murata, H. Levitt, and K. Youdelman, “A motivation-sustaining articulatory/acoustic speech training system for profoundly deaf children,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* 1993, pp. 145–148.

[17] H. R. Javkin, E. G. Keate, N. Antonanzas-Barroso, and B. A. Hanson, "Synthesis-based speech training system," US Patent 5,340,316, Aug. 23, 1994.

[18] Dr. Speech Software Group, Software demo on Dr. Speech 4, and Speech Therapy, (Tiger DRS, Inc., Seatle, Wa, 2003). [Online]. Available: http://www.drspeech.com (Last accessed in May, 2012).

[19] O. Engwall, O. Balter, A. M. Oster, and H. Kjellstrom, "Designing the user interface of the computer-based speech training system ARTUR based on early user tests," *Behaviour and Information Technology*, vol. 25, pp. 353-365, 2006.

[20] J. R. Westbury, “X-ray microbeam speech production database user’s handbook (version 1.0),” 1994. [Online]. Available: http://www.medsch. wisc.edu/ubeam/.

[21] A. S. MacMillan and G. Kelemen, “Radiography of the supraglottic speech organs,” *A. M. A. Archives of otolaryngology*, vol. 55, no. 6, pp. 671-688, 1952.

[22] J. S. Perkell, Physiology of Speech Production. M.I.T. Press, Cambridge, MA, 1969.

[23] K. G. Munhall, E. Vatikiotis-Bateson, and Y. Tohkura, "X-ray Film database for speech research," *J. Acoust. Soc. Am.*, vol. 98, pp. 1222-1224, 1995.

[24] J. Z. summer, "Articulograph AG100 electromagnetic articulation analyzer," 1997. [Online]. Available:http://www.linguistics.ucla.edu/faciliti/facilities/ physiology/Emamual .html#1 (Last accessed in May, 2012).

[25] E. Bresch, Y. C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shape using real-time magnetic resonance imaging," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 123–132, Mar. 2008.

[26] J. W. Devaney and C. C. Goodyear, "A comparison of acoustic and magnetic resonance imaging techniques in the estimation of vocal tract area functions", in *Proc. International Symposium on Speech, Image Processing and Neural Networks*, 13-16 April 1994, Hong Kong, Vol 2, pp. 575-578.

[27] P. Mermelstein, "Determination of vocal tract shapes from measured formant frequencies," *J. Acoust. Soc. Am.*, vol. 41, no. 5, pp. 1283-1294, 1967.

[28] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *J. Acoust. Soc. Am.*, vol. 41, no. 4, pt. 2, pp. 1002–1010, 1967.

[29] M. M. Sondhi and B. Gopinath, "Determination of vocal-tract shape from impulse response at the lips," *J. Acoust. Soc. Am.*, vol. 49, no. 6, pt. 2, pp. 1867–1873, 1971.

[30] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *J. Acoust. Soc. Am.*, vol. 64, no. 4, pp. 1027–1035, 1978.

[31] K. Iskarous, "Vowel constrictions are recoverable from formants," *J. Phonetics*, vol. 38, no. 3, pp. 375–387, 2010.

[32] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 133–150, 1994.

[33] G. Fant, Acoustic Theory of Speech Production. The Hague: Mouton, 1960.

[34] N. S. Nayak, R. Velmurugan, P. C. Pandey, and S. Saha, “Estimation of lip opening for scaling of vocal tract area function for speech training aids,” in *Proc. National Conference on Communications* (NCC 2012), 2012, Kharagpur India, pp. 521-525.

[35] O. O. Akande and P. J. Murphy, “Estimation of the vocal tract transfer function with application to glottal wave analysis” *Speech Commun.*, vol.46, pp. 15–36, 2005.

[36] H. Deng, M. P. Beddoes, R. K. Ward, and M. Hodgson, “Estimating the glottal waveform and the vocal-tract filter from a vowel sound signal,” in Proc. *IEEE Pacific Rim Conf. Commun., Comp., Signal Process.*, 2003, Victoria Canada, pp. 297–300.

[37] M. G. Kang and B. G. Lee, “A generalized vocal tract model for pole zero type linear prediction”, in *Proc. Acoustics, Speech, and Signal Process., 1988. ICASSP-88*., 1988, vol.1, pp. 687 – 690.

[38] L. T. Lim and B. G. Lee, “Lossless pole-zero modeling of speech signals,” *IEEE Trans. Speech Audio Process.*, vol. 1, no. 3, pp. 269-276, 1993.

[39] C. H. Coker, “A model of articulatory dynamics and control,” *Proc. IEEE,* vol. 64, no. 4, pp. 452-460, 1976.

[40] G. Richard, M. Goirand, D. Sinder, and J. Flanagan, "Simulation and visualization of articulatory trajectories estimated from speech signals", in *Proc. of the International Symposium on Simulation, Visualization and Auralization for Acoustic Research and Education* (ASVA97), (Tokyo, Japan), Apr. 1997.

[41] A. Soquet, M. Saerens, and P. Jospa, “Acoustic-articulatory inversion based on a neural controller of a vocal tract model: further results,” in *Proc. Artificial Neural Networks*, North Holland, 1991, pp. 371-376.

[42] C. Qin and M. Á. Carreira-Perpinán, "A comparison of acoustic features for articulatory inversion", in *Proc. Int. Conf. Interspeech 2007-Eurospeech*, (Antwerp, Belgium), 2007, pp. 2469–2472.

[43] Y. Laprie and B. Mathieu, “A variational approach for estimating vocal tract shapes from the speech signal,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.,* 1998, pp. 929–932.

[44] S. Maeda, “A digital simulation method of the vocal-tract system,” *Speech Commun.,* vol. 1, pp. 199–229, 1982.

[45] L. R. Rabiner, B. S. Atal, and M. R. Sambur, “LPC prediction error-analysis of its variation with the position of the analysis frame,” *IEEE Trans. Acoustics, Speech, Signal Process.,* vol. ASSP-25, no.5, pp. 434-442, 1977.

[46] T. V. Ananthapadmanabha and B. Yegnanarayana, “Epoch extraction from linear prediction residual for identification of closed glottis interval,” *IEEE Trans. Acoust., Speech, Signal Process.,* vol. ASSP-27, pp. 309–319, 1979.

[47] R. Mizoguchi, M. Yanagida, and O. Kakusho, “Speech analysis by selective linear prediction in the time domain,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* 1982**,** vol. 3, pp**.** 1573-1576.

[48] C. Ma, Y. Kemp, and L. F. Willems, “ Robust signal selection for linear prediction analysis of voiced speech,” *Speech Commun.,* vol. 12, no. 2, pp. 69-81, 1993.

[49] M. Mezzalama, “Influence of the position of the analysis frame in LPC pitch synchronous analysis,” *Signal Process.,* vol. 1, pp.191-204, 1979.

[50] D. H. Klatt, “Software for a cascade/parallel formant synthesizer*,” J. Acoust. Soc. Am.,* vol. 67, pt. 3, pp. 971–995, 1980.

[51] M. S. Shah and P. C. Pandey, “Estimation of place of articulation in stop consonants for visual feedback,” in *Proc. Interspeech,* 2007, pp. 2477–2480.

[52] D. G. Childers and T. H. Hu, “Speech synthesis by glottal excited linear prediction,” *J. Acoust. Soc. Am.,* vol. 96, no. 4, pp. 2026-2036, 1994.

[53] T. Barnwell III, “Windowless technqiues for LPC analysis,” *IEEE Trans. Acoust. Speech, and Signal Process.,* vol. 28, pp. 421–427, 1980.

[54] J. Markel and A. H. Gray, Jr., “Fixed-point truncation arithmetic implementation of a linear prediction autocorrelation vocoder,” IEEE *Trans. Acoust., Speech, Signal process.,* vol. ASSP-22, pp. 273-281, 1974.

[55] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” *J. Acoust. Soc. Amer.,* vol. 50, pp. 637-655, 1971.

[56] A. K. Krishnamurty and D. G. Childers, “Two-channel speech analysis,” *IEEE Trans. Acoust., Speech, Signal Process.,* vol. ASSP-34, pp.730–743, 1986.

[57] J. Makhoul, “Stable and efficient lattice methods for linear prediction,” *IEEE Trans. Acoust. Speech Signal Process.,* vol. ASSP- 25, pp. 423-428, 1977.

[58] J. Burg, “Maximum entropy spectral analysis,” *Ph. D. dissertation,* Stanford Univ., Stanford, CA, 1975.

[59] B. S. Atal, “Predictive coding of speech signals at low bit rates,” *IEEE Trans. Comm.,*vol. COM-30, no. 4, pp. 600-614, 1982.

[60] S. A. Liu, “Landmark detection for distinctive feature based speech recognition,” *J. Acoust. Soc. Am.,* vol. 100, pp. 3417-3430, 1996.

[61] A. R. Jayan, P. S. Rajath Bhat, and P. C. Pandey, “ Detection of burst onset landmarks in speech using rate of change of spectral moments,” in *Proc. National Conf. on Commun. (NCC),* 2011, Bangalore, India, paper SpPrI.3.

[62] C. Park, “Consonant landmark detection for speech recognition,” *Ph.D. dissertation,* Massachusetts Institute of Technology, 2008.

[63] A. L. Francis, V. Ciocca, and J. M. C. Yu, “Accuracy and variability of acoustic measures of voicing onset,” *J. Acoust. Soc. Am.,* vol. 113, no. 2, pp.1025–1032, 2003.

[64] C. Glaser, M. Heckmann, F. Joublin, and C. Goerick, “Combining auditory preprocessing and Bayesian estimation for robust formant tracking,” *IEEE Trans. on Audio, Speech, and Language Process.,* vol. 18, no. 2, pp. 224–236, 2010.

[65] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” *Bell Syst. Tech. J.,* vol. 54, no. 2, pp. 297–315, 1975.

[66] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program],” Version 5.3.17, 2012. [Online] Available: http://www.praat.org/ (Last accessed in June 2012).

[67] B. H. Story, I. R. Titze, and E. A. Hoffman, “Vocal tract area functions from magnetic resonance imaging*,” J. Acoust. Soc. Amer.,* vol. 100, no. 1, pp. 537–554, 1996.

[68] J. L. Flanagan, Speech Analysis, Synthesis, and Perception, 2nd ed. New York: Springer-Verlag, 1975.

Acknowledgements

I am deeply grateful to my guide Prof. P.C. Pandey, for his guidance and encouragement throughout the course of this project. He has given me an opportunity to pursue my higher studies at IIT and provided me lot of opportunities to build the confidence in handling different challenges during the project. I am also thankful to him for sparing his invaluable time in correcting my reports and research papers.

I am thankful to Prof. M. S. Shah for his support during various stages of the project. I am deeply thankful to Jagbandhu and Rajath for their love and support during the entire period of the project. I would like to thank Jagbandhu, Nitya, and Sudipan for sparing their time in correcting my reports. I am thankful to Jayan, Parveen Lehana, Khadar, Santosh, Badri, Vinod for sharing interesting discussions with me and creating a joyful environment to work. I would like to thank Vidyadhar Kamble for helping me in all the lab related issues. I am also thankful to all my friends especially in the SPI lab and EI Lab for their lovely support during the tenure of this project.

Nataraj K S

June 2012

Author's Resume

K. S. Nataraj: He received the B. E. degree in electronics and communication engineering from the Bapuji Institute of Engineering and technology, Davangere, Karnataka in 2005. He worked as a software engineer at Robert Bosch Engineering and Business Solutions, Bangalore from September 2005 to August 2008. Presently, he is pursuing the M.Tech. degree in electrical engineering at the Indian Institute of Technology Bombay. His research interests include digital signal processing, speech processing and embedded system design.

Thesis related publication

K. S. Nataraj, Jagbandhu, P. C. Pandey, and M. S. Shah, "Improving the consistency of vocal tract shape estimation," in *Proc. National Conf. Commun. (NCC)*, 2011, Bangalore, India, paper SpPrII.4.

Jagbandhu, K. S. Nataraj, and P. C. Pandey, "Detection of transition segments in VCV utterances for estimation of the place of closure of oral stops for speech training," accepted for publication in *Proc. Interspeech 2012*, Portland, Oregon, USA, Sept. 9-13, 2012.