A Visual Feedback of Vocal Tract Shape for Speech Training

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

Master of Technology
by

Jagbandhu
(09307603)

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 **"A Visual Feedback of Vocal Tract Shape for Speech Training"** by **Jagbandhu** (Roll No. **09307603**) is approved, after the successful completion of *viva voce* examination, for the award of the degree of **Master of Technology** in **Electrical Engineering.**

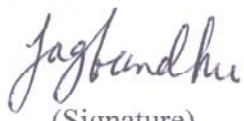
Supervisor	PCPandey	(Prof. P. C. Pandey)
Examiners	Preeti Rao	(Prof. Preeti Rao)
	SEL	(Prof. M. S. Shah)
Chairperson		(Prof. B. Ravi)

Date: 22 June 2012

Place: Mumbai

Declaration

I declare that this dissertation 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 I 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.



(Signature)

Jagbandhu

09307603

Date: 22 June 2012

Place: Mumbai

Jagbandhu / Prof. P.C. Pandey (Supervisor), "A visual feedback of vocal tract shape for speech training", *M.Tech. dissertation*, Department of Electrical Engineering, Indian Institute of Technology Bombay, June 2012.

Abstract

The absence of auditory feedback in hearing-impaired children severely affects the accuracy of articulation. For providing a visual feedback of articulatory efforts for speech training, the vocal tract shape can be estimated by LPC analysis of the speech signal. Since the signal energy is very low during stop closures segments of vowel-consonant-vowel (VCV) utterances, the estimated area values are unrelated to the place of closure. It has been reported earlier that vocal tract shape during stop closures can be estimated by performing bivariate polynomial interpolation of the estimated area during the transition segments preceding and following the stop closure. The project involves investigations for improving the visual feedback of vocal tract shape.

For validation of the vocal tract shape estimated from speech signal with reference to the X-Ray Microbeam database, an automated method to find the place of maximum constriction from the X-Ray Microbeam images is developed. It involves estimation of axial curve of the vocal tract from the pellet points given in the database and measuring the vocal tract opening as a distance between the palatal and the tongue outlines measured along the normal to the axial curve. For a realistic display of the vocal tract shape, it is shown that cubic B-spline interpolation of the estimated area values results in a smooth vocal tract shape and a distinct place of constriction. In the LPC-based estimation, area ratios on both sides of the section interfaces are converted into areas by assuming a constant normalized area of unity at the glottis end. This assumption fails during VC and CV transition segments as the vocal tract configuration is dynamically varying during these segments. Lip area value estimated from the video images can be used as a reference area for scaling the area ratios. It is shown that use of lip scaled area values for bivariate surface modeling results in consistent estimation of the place of closure for alveolar and bilabial stops. An animation for providing the visual feedback of the vocal tract shape for vowels and VCV utterances for the hearing-impaired is also developed.

CONTENTS

List of abbreviations

Abbreviation	Explanation
2D	two-dimensional
B	blue
CMI	central maxillary incisors
CV	consonant-vowel
G	green
GUI	graphical user interface
MaxOP	maxillary occlusal plane
MRI	magnetic resonance imaging
LPC	linear predictive coding
R	red
VC	vowel-consonant
VCV	vowel-consonant-vowel
XRMB	X-Ray Microbeam

List of Figures

Chapter 1

INTRODUCTION

1.1 Problem overview

The process of speech acquisition in children with normal hearing is aided by the auditory feedback. With the feedback of their own sound, they acquire the ability to control positions and movements of various articulators (tongue, velum, lips, and jaw) involved in speech production [1], [2]. Most hearing impaired persons have a proper speech production mechanism but they have a great difficulty in acquiring speech because of lack of feedback of the auditory cues. The hearing impaired persons often use lip reading as a visual cue for perceiving speech. This method is effective after special training, but it cannot help in differentiating between utterances involving internal articulators. Several alternate means in the form of tactile and visual feedback have been developed [3]-[8]. It has been reported that use of computers in speech training increases student motivation for working on speech skills [1], [9], [10]. Acoustic parameters (such as speech intensity, voicing, pitch, and spectral features) and articulatory parameters (such as nasality, lip movement, and tongue movement) have been used as feedback parameters in speech-training aids. Many systems incorporate interactive game formats and other graphics that children find highly motivating.

Several speech training aids providing visual feedback of vocal tract shape have been developed. Most of these speech training aids uses Wakita's [11] linear predictive coding (LPC) based method for estimation of the vocal tract shape directly from the speech signal. This method works well for vowels, semivowels, and diphthongs but fails during the closure of the vowel-consonant-vowel (VCV) utterance due to low signal energy. Pandey and Shah [12] proposed a method for estimating the place of articulation during the closure of the VCV utterance by performing bivariate polynomial modeling on area values in the VC and CV transition segments. The place of articulation estimated during the closure segment of VCV utterances was validated with the ones obtained manually from the X–Ray

Microbeam (XRMB) database. Vocal tract shape estimated from the LPC analysis is interpolated from 12 to 40 sections for smooth display of vocal tract shape, using cubic spline interpolation. However, it is difficult to obtain the place of maximum constriction precisely from the cubic spline interpolated area values during stop closure of the VCV utterances due to the variability in the vocal tract shape across the sections. Hence, a method is needed to estimate the accurate place of constriction during stop closure of VCV utterances. For validation of the results, the process of finding the place of maximum constriction from the XRMB images also needs to be automated.

One of the limitation of the LPC based method as mentioned by Wakita [13] is the lack of method for obtaining reference area value for scaling the area ratios during dynamically changing vocal tract configuration. Hence, area values during the VC and CV transition segments need to be scaled by an appropriate scaling factor to improve the accuracy of the estimated shape during the closure of VCV utterances. Area at the lips can be used as a reference area for scaling the area ratios during the transitional tract configuration.

1.2 Project objective

The project objective is to develop techniques to facilitate the development of a speech training aid for visual feedback of vocal tract shape to the hearing impaired children. Towards this end investigations have been carried out for following:

(i) Obtaining reference shapes from XRMB database for validating the estimated shapes of the vocal tract during the closure of the VCV utterances.

(ii) Estimation of the place of constriction during stop closure of VCV utterance.

(iii) Studying the effect of lip scaling on vocal tract shape estimation for VCV utterances.

(iv) Developing the graphics for displaying the vocal tract shape.

1.3 Dissertation outline

Chapter 2 reviews the literature related to the speech training aids and the vocal tract shape estimation techniques. Chapter 3 presents the method for obtaining the place of maximum constriction from XRMB database for validation purpose. In Chapter 4, interpolation methods for a smooth display of estimated area values are presented.

Chapter 5 presents the use of lip opening for scaling of the vocal tract area functions in VCV utterances. Development of the animation for visual display of the vocal tract shape is presented in Chapter 6. The last chapter presents a summary of the investigations carried out and some suggestions for future work.

Chapter 2

SPEECH TRAINING AIDS

2.1 Introduction

Children with prelingual hearing impairment have great difficulty in acquiring speech because they cannot use auditory feedback for speech correction. It is very difficult for a hearing-impaired person to compare their own articulation with that of other speakers. Other modes of feedback may help the hearing impaired children in learning to speak [1]-[8]. Generally, visual feedback of the articulatory efforts has been reported to be most effective [3], [4]. This chapter gives a review of visual feedback based speech training aids and the direct and indirect methods for obtaining the vocal tract shape and their suitability and limitations for use in speech training aid.

2.2 Visual speech training aids

Bernstein *et al.* [14] have reported a PC based system for training of intensity and voicing control with an option of selecting functions, games, and setting of parameters. Monitoring of the speech production is done by the system consisting of microphone, electroglottograph (for monitoring opening and closing of the vocal folds), and pneumotachograph (for monitoring volume velocity of expiration). It has been reported that these games have received good response from the children aged 3-5 and 9-11 years.

Watanabe *et al.* [5] reported a system in which speech is represented as a continuous color plot against time. In this system, four neural networks, trained with 62 /VCV/ syllables, uttered by 20 males and 20 females were used for identifying the source, manner, and place of articulation. According to a specific weighted combination of first 3 formants of the voiced sound, outputs from the neural nets are represented as R, G and B colors in the plot. The brightness of the plot changes with sound intensity and silence gets represented by a black section in the plot. Different

patterns are used for indicating the type of articulation in the plots. Nasals, plosives, and fricatives are represented by a mesh, horizontal bar, and random dots, respectively. The place of articulation is represented by the position of the color in the plot. The glottis to lip position in the vocal tract spans from the left to the right of the screen, respectively. Accordingly, velar, alveolar, and bilabial utterances are displayed on the left, middle, and right of the screen, respectively. The performance of the system was evaluated in a preliminary test in which three students read the visual patterns for 300 items, consisting of 75 words uttered by four males. The learning curves showed a steep rise and the correct answer rate reached 96-99% after eight trials. The response times shortened as the test proceeded and reached 1.3 s/word at the final trial.

Zahorian and Venkat [33] reported a speech training system to improve vowel articulation for the hearing impaired children. Processing of the speech was based on the non-linear/linear artificial neural network transformation of 16-channel filter bank data to a two-dimensional space which approximates a first and second formant (F1/F2) space. Spatial location and color were used as cues for vowel identification for speaker-independent vowel training displays. Experiments showed that the display can be easily interpreted and the information provided can be used to improve the vowel articulation of hearing impaired speakers.

Indiana Speech Training Aid (ISTRA) developed by Indiana University, is a speaker-dependent visual speech training system [32]. In this system visual feedback was given by means of template matching between the speaker produced sound and the stored template of the same utterance. The template is made from the best utterances a hearing impaired child can make. Feedback is based on the goodness metric which is derived from the match between the new utterance and the stored template. Testing of the system was performed on both hearing impaired as well as normal hearing children. Results indicate speech improvements for both trained and untrained words for both the hearing-impaired and normal-hearing children.

"Video Voice Training System", from Micro Video Corporation [6], provides a number of activities for speech training using different games. For vowel training, the F2/F1 formant plot is displayed. The model and trial pattern appear on the screen simultaneously, for analysis of similarities and differences, with instant replay for review. The goal is to match the model pattern. Models can be changed by clicking mouse in the plot for individual practice against their own voice patterns. There are

games for voicing, intensity, and pitch training. For voicing duration, onset, and articulation, a picture slowly appears by continuous voicing, vocal onset, maintaining a desired volume level, or repeated production of a target word or sound. For intensity and pitch training, level indication or a hot air balloon game is used. The level indication or balloon moves up or down based on volume or frequency content in the speech signal.

"Speech Viewer III" by IBM is a computer based training aid which has various exercises and games to acquire the control on the multiple aspects of voice like pitch, loudness, voicing etc. [7]. The programs are divided into three sections: awareness, skill building, and speech patterning. Basic awareness section displays various speech parameters such as pitch, loudness, timing, and the presence or absence of voicing. The skill building module has different games to strengthen the ability in pitch, timing, voicing, and speech sound production. There are phonology exercises for the exact production of the phonemes of a specific language. The spectrum of the phrases, words and single utterances produced by the learner are compared with the spectra of the target model. Learner's own best production is stored as a target to establish a new and more correct behavior through frequent training. The goal is accomplished when the learner's best production becomes his or her most common production.

"Dr. Speech", from Tiger DRS, Inc. [8], uses over 70 voice-activated games to provide training in pitch, loudness, voiced and unvoiced phonation, voicing onset, and maximum phonation time. It also displays the lip movement associated with predefined set of utterances along with the vocal tract shape animation showing articulatory effort required for that utterance.

2.3 Vocal tract shape as a feedback

Visual display of the vocal tract shape in speech-training aids, can serve an important role in providing feedback of articulatory efforts required in speech production. In these systems, vocal tract shape is estimated from the speech signal and displayed on the screen along with the target shape.

Speech training aid developed by Crichton and Fallside [15], displays the vocal tract shape for sustained vowels estimated using Wakita's LPC based method [11]. Smoothened display of vocal tract shape was generated by performing a

parabolic interpolation between discrete areas and displayed as log area values as a function of the distance along the vocal tract length. The left-hand end of the plot corresponds to the position of the glottis, and the right-hand end to the lips. The system was used for 9 months by the hearing impaired children in the 6–9 years age group and was found useful for improving vowel articulation.

In the system reported by Pardo [4], the relative area values were obtained by Wakita's method and displayed using a smoothened log area function for the teacher and student separately. Along with the two curves, an animation of puppet was used which smiles when a correct sound has been uttered and frowns when it has not. Use of the system in speech training for the production of a set of five vowels over four months showed that error with respect to target shapes decreased progressively with training.

Black [16] reported a system for displaying real-time tongue positions for vowel production by means of a mid-sagittal plot of the human head. In this system the formants were calculated from a filter bank analysis of the input speech waveform using peak picking algorithm. Tongue shape is estimated by the procedure proposed by Ladefoged *et al*. [17], involving a relationship between the formant frequencies and the degree of back and front raising of the tongue. Training with the system improved accuracy and consistency in vowel production of the hearing impaired persons.

Oliveira and Souza [18] reported a PC based speech-training aid for vowel articulation, providing real-time visual feedback of vocal tract shape of hearing impaired person. This system provides an image of two sagittal cuts of the vocal tract, one representing the reference shape and other the estimated shape. The speech signal was sampled at 11.025 kHz in 30 ms blocks. The formants were obtained from 12th order LPC spectrum estimation using Durbin's recursive algorithm. Vocal tract shape was obtained from the experimental observation, relating first formant to the jaw opening angle and second formant to the average positions of the tongue. It was reported that system was able to display well the positions of the speaker's tongue. For adult hearing impaired people, the system gave good results as a biofeedback tool.

Park *et al.* [3] developed an integrated speech training system which can display the vocal tract shape and other speech parameters like intensity, fundamental frequency, nasality, and log spectra. For estimation of the vocal tract shape, Wakita's LPC based method [11] was used. The vocal tract graphics developed shows the

cross-section of the vocal organs from the glottis to the lips. Initially, the axes for the movement of the articulators are defined and then vocal tract shapes were displayed using these axes. The motion axes were developed based on the following principles, (i) vocal tract was modeled as an acoustic tube with 15 equal tube sections, (ii) the shortest path from the upper to the lower part of each section was selected, and (iii) the upper palate was assumed to be fixed and the tongue was assumed to be movable. Correction messages were displayed during intensity and pitch training where the user was asked to lower or increase the corresponding parameter to match the required shape. It was reported that hearing impaired children, using this system, mastered two syllables /ja/ and /pa/ in 5~6 days.

Massaro and Light [34] developed a computer-animated head known as Baldi, as a language tutor for speech perception and production for children with hearing loss. It uses four different views (back view, sagittal view, side view, and front view) with transparent skin to show the articulator activity within the mouth as well as facial movement during speech. The system has facility for slowing down Baldi's speaking rate to highlight speech distinctions. It also makes the Baldi produce a strong air stream for voiceless and a limited air stream for voiced sounds. Seven children with hearing loss were trained on both perception and production for 6 hours over a period of 21 weeks. The proportion of correct identifications measured through a forced choice reference test increased from 0.64 at pre-test to 0.86 at post-test. One of the disadvantages of this system is that it cannot analyze user's speech and provide feedback for speech production.

Mahdi [19] reported a computer-based system for visualization of vocal tract shape by means of a mid-sagittal view of the human head. The system also displays sound intensity, pitch, and first three formants. The vocal tract area values and the formants are extracted from the speech signal using an auto-regressive speech production model and linear-predictive analysis. Estimated area functions are mapped to corresponding mid-sagittal distances using a numerical algorithm. To compensate for possible errors in the estimated area due to variation in vocal tract length, the first two sectional areas are estimated from the first three formants. The estimated 18 sections of the vocal tract for various vowels were compared with the reference shapes and the deviation was expressed in terms of mean-square error. The results showed good correlation with the reference X-ray data.

Engwall *et al.* [20] reported a speech training system known as ARticulation TUtor (ARTUR) developed at KTH (Royal Institute of Technology), Sweden. AURTUR uses speech recognition along with the facial features to detect the mispronounced speech. It gives the feedback of the difference between the user's and correct pronunciation in the form of written and oral instructions and three-dimensional animations of the face and internal parts of the mouth (tongue, palate, jaw etc.). Tongue shape of the 3D vocal tract model is based on the statistical analysis of Magnetic Resonance Imaging (MRI) data to make it articulatorily correct and the articulatory movements are modeled from Electromagnetic Articulography (EMA) measurements, both for the same subject. Testing of the system was performed on two user groups: three children aged 9-14 with experience of speech training systems, and three children aged 6 with limited or no experience of speech training systems. It was reported that the children aged 9-14 showed positive responses about the ARTUR's oral and written instructions and had no problem in understanding the animation except the black line representation of the hard palate. All the three children aged 6 appreciated the animation and were able to correct their pronunciation with the help of animation.

A review of speech-training aids for the hearing impaired has shown that systems with visual feedback of articulatory efforts, not visible from outside, were useful for improving vowel articulation. Most of these systems were based on visual feedback of vocal tract shape.

2.4 Direct methods for vocal tract shape estimation

The display of vocal tract shape is the most popular among the visualization methods since it gives direct feedback about the articulatory efforts required for making an utterance. The vocal tract shape can be estimated by direct and indirect methods. Direct methods involve exposing the human body to different electromagnetic waves and extracting its different features.

In cinefluorography, X-ray images are taken using X-rays on a fluorescent screen, and these images are photographed with a cine camera. The X-Ray Film Database for Speech Research [21] created by Queen's University/ATR Labs has total of 55 minutes of footage of 25 X-ray films with its original audio recordings. However, this method has been discontinued due to the radiation hazard which also

limits the amount of data that can be obtained and is mainly limited for validation purpose. X–Ray Microbeam (XRMB) database has been developed by the University of Wisconsin [22]. The XRMB system generates a narrow beam of high energy X-Rays which tracks and records the motion of 2-3 mm diameter gold pellets attached to the tongue, jaw, lips, and soft palate during speech production. The database consists of the XRMB data in synchrony with the speech signal for VCV utterances, isolated vowels, and sentences from the 48 speakers.

Magnetic resonance imaging (MRI) is an imaging technique that can be used for vocal tract area measurements without involving radiation exposure. Story *et al.* [23] used MRI technique to acquire speaker-specific three-dimensional vocal tract shapes that correspond to a particular set of vowels and consonants. They obtained a set of 18 shapes for 12 vowels, 3 nasals, and 3 plosives of one male subject who vocalized while being scanned. However, MRI technique has a number of disadvantages. The scanning time required for acquisition of a full image set is of the order of several minutes. When this is coupled with the pauses required for subject respiration, complete image set may require double or triple the actual scan time. Also, the air-tissue interface, teeth, and bone are poorly imaged by MRI.

Electropalatography (EPG) requires the subject to wear an artificial plate against the hard palate and records the tongue's contact with the plate through electrodes on its surface. EPG uses a computer display to show the user's tongue position. It uses 62 contacts which are distributed along 8 rows over the upper palate. Contact made by the tongue on any one of the contact causes that particular contact to be shaded in the display. This method has traditionally been used to improve consonant production. Electromagnetic articulography (EMA) works by creating a magnetic field. It comprises of small connector coils (sensors) that are positioned on the tongue, lips, jaw and on the nose of the subject. Three transmitter coils attached to a head mount produce an alternating magnetic field. The magnetic fields induce an alternating current in the sensors attached to the articulators (tongue, lips, jaw, etc). The size of the induced current is inversely proportional to the X (horizontal) and Y (vertical) distance of each sensor from the transmitter coils. Signals from the sensors are amplified and given to a computer via analog unit. A computer program provides the location of the receiver coil in x-y distance over time. The MOCHA database created by the University of Edinburgh [35] provides both the EMA and EPG data along with the audio recordings for a set of 460 sentences for 10 speakers.

2.5 Indirect methods for vocal tract shape estimation

In indirect methods, the vocal tract shape is estimated by using acoustic measurements or from the analysis of the speech signal. Schroeder [24] proposed a method for vocal tract shape estimation by acoustic measurements of speech formants and acoustic impedance at lips. The impedance value at the lips was obtained with an experimental setup consisting of impedance tube, electrodynamic driver, and microphones. On the left end of the tube, electrodynamic driver provides the acoustic output consisting of periodic pulses with a bandwidth of 4 kHz and repetition period 10 ms. The right end of the tube was connected, via a specially designed seal, to the mouth of the subject who articulated without phonation (i.e., silently). Two closely spaced microphones pick the sound pressure resulting from the incoming and reflected sound waves. The signals were processed to obtain acoustic impedance at the lips and hence vocal tract area function.

Ladefoged *et al.* [16] proposed a method, that uses first three formant frequencies for estimating the vocal tract shapes for vowels. The vocal tract is specified in terms of position of 18 equally spaced points from the reference line along the lower surface of the vocal tract. It has been shown that the tongue shape can be estimated precisely in terms of certain proportions of front raising and back raising components. The proportion is calculated from the first three formant frequencies. Two different vocal tract shapes may result for the same set of formant frequencies but constraints are imposed on the possible vocal tract shapes. The estimated shape and the original shape show good correlation for the vowels produced by five subjects.

Wakita [11] described a method for estimation of vocal tract shape directly from the speech signal. In this method, vocal tract was assumed to be an all pole filter for periodic non nasalized voiced sound. Filter was driven by an impulse train. The excitation is estimated by passing the speech signal through an inverse of the all-pole filter. The vocal tract was also modeled as a lossless acoustic tube having a set of discrete and equal length sections with variable cross-section area. The acoustic tube was made equivalent to the inverse filter for the condition that the speech sampling frequency should be equal to $Mc/2l$ where, c is the speed of sound and l is the length of the vocal tract assumed. Using this equivalent relation, the reflection coefficients at the section interfaces of the acoustic tube were estimated from the speech waveform.

The reflection coefficients are then converted to the ratios of the areas on the two sides of section interfaces. The area values were obtained by scaling the ratios with respect to the area of a reference section. Scaling is carried out by assuming the glottis end of the tract to be having a normalized area of unity. Most of the visual speech training aids having vocal tract shape as a feedback parameter use Wakita's LPC-based method [11] for direct estimation of vocal tract shape as it is suitable for real-time processing. The method can be used for estimating fixed vocal tract configurations like vowels as well as transitional tract configurations like semivowels and diphthongs.

Estimation of vocal tract shape, based on LPC analysis of speech, showed that estimated shapes for vowels, semivowels and diphthongs are satisfactory and related to place of articulation but the method fails during the closure segments of stops due to low energy and lack of spectral information [11]. Pandey and Shah [12] reported a technique for estimating the vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterance by using a bivariate surface model fitted on the vocal tract shapes during the transition segments preceding and following the stop closure. The estimated place of closure for 120 /ʌCa/ utterances, involving stop consonants /b/, /d/, and /g/ showed a good match (with a correlation coefficient of 0.94) with those obtained from direct X-ray imaging from XRMB database [22].

One of the main shortcomings of the LPC based method, as mentioned by Wakita [13] is lack of a method for reference section area estimation from the speech signal for converting the area ratios into area values during varying configuration of the vocal tract. Usually for a static vowel, the scaling is carried out by assuming a reference section at the glottis end with a normalized area of unity. But this method introduces errors during dynamic variation of the vocal tract shape due to variation in the area at the glottis end. The choice of reference area for scaling the area ratios is rather arbitrary and in general may vary from frame to frame in an analysis. Nayak *et al*. [25] proposed a technique for estimation of the lip opening area from the video images of the speaker's face to be used as the reference for scaling of the area ratios for vocal tract shape estimation. The lip opening areas were calculated using template matching method from the video recordings of the vowels /a/, /i/, and /u/ from eight male speakers. It was reported that the area values obtained for all the eight speakers after scaling using the lip area are consistent with those obtained using MRI.

The vocal tract shape estimated using Wakita's method shows variation in the area values for steady-state, voiced segments of speech which is dependent on the position of the analysis window. Nataraj *et al.* [26] reported a method for improving the consistency of the vocal tract shape estimation by selecting the frame positions having the minimum windowed energy index which is the ratio of the windowed signal energy to the frame energy. The area values show less variation for the frames having the minimum energy index.

2.6 Summary

Speech training aids provide the visual feedback of acoustic parameters and articulatory parameters. Visual feedback of vocal tract shape has been reported to be useful in developing correct articulatory efforts. Most of these systems use mid-sagittal plot of the human head for visualization of the vocal tract shape and are based on improving vowel articulation. Hence, a speech training aid providing the visual feedback of the vocal tract shape for the VCV utterances needs to be developed.

Vocal tract shape can be estimated by direct and indirect methods. Impracticality of the use of direct methods in speech training limits their scope to validation of vocal tract shapes obtained using the indirect methods. Most of the speech training aids providing visual feedback of vocal tract shape use LPC based method for vocal tract shape estimation. It works satisfactorily for vowels, semivowels, and diphthongs but it fails during the closure segments of stops due to low signal energy. Vocal tract shape during the stop closure can be estimated by bivariate surface modeling of the VC and CV transition segments. Area ratios obtained by LPC analysis are scaled by the glottis area value of unity. This assumption does not hold true for the transitional tract configuration. Lip area can be used as a reference area for scaling the area ratios during the VC and CV transition segments. Hence, there is a need to investigate the effect of scaling by the lip area on the estimation of the place of articulation for VCV syllables.

Chapter 3

PLACE OF ARTICULATION FROM XRMB DATABASE

3.1 Introduction

Indirect methods used for the estimation of vocal tract shape need to be validated with the shape estimated from the direct imaging methods. Place of articulation during the stop closure of the VCV utterance can be obtained from the bivariate surface modeling of the transition segments preceding and following the stop closure as proposed by Pandey and Shah [12]. For validation, the technique was applied to the speech signal from the XRMB database [12], and the place of constriction obtained from the bivariate surface modeling was compared with the actual place of constriction obtained manually from the x-y articulatory plot of XRMB [12]. The position of maximum constriction from the lips, along the length of the vocal tract, was calculated by adding the straight line distances along the curve joining the lower lip marker and the tongue pellet markers to the point of constriction. As manual detection of constriction is time consuming, a method to automatically determine place of maximum constriction is developed.

3.2 XRMB database

The XRMB database has been developed by the University of Wisconsin-Madison [21]. It contains recordings of motions of articulators during speech production. It is done by generating a very narrow beam of high-energy X-rays, and rapidly directing this beam, to track the motions of 2-3 mm diameter gold pellets glued to the tongue, jaw, lips, and soft palate. This provides details of articulators in the mid-sagittal plane. A large number of speakers (48 speakers) and their utterances (VCV utterances, isolated vowels, sentences) are stored in the database. The database is accompanied by software, which allows the user to control the rate of display of the articulograph. It also displays pitch, RMS trace of the audio signal, and spectrogram.

XRMB x-y articulatory plot consist of four pellet points (T1-T4) on the tongue

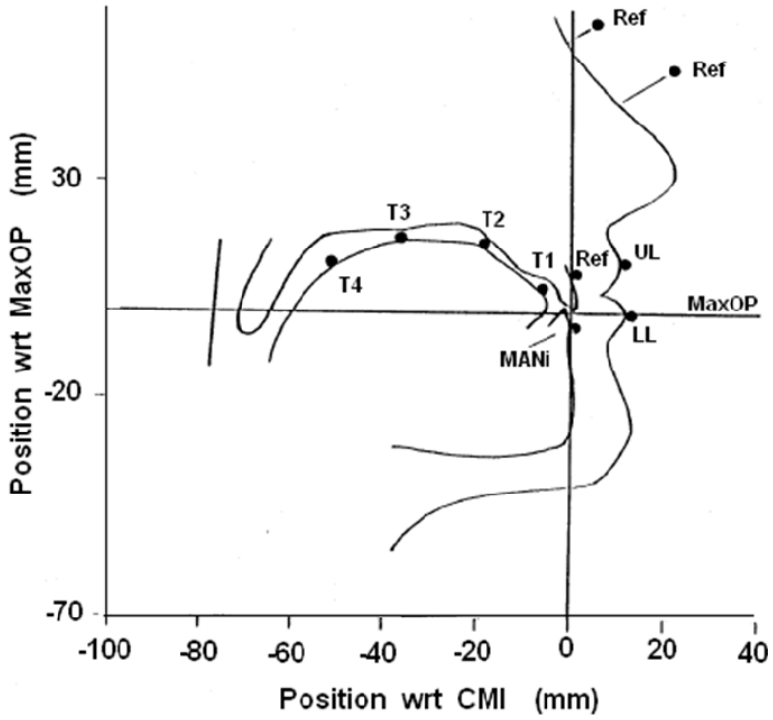


Figure 3.1 Position of the pellet points [21]

and one each on the upper lip (UL), lower lip (LL) and incisor (MNi) as shown in Fig. 3.1. Origin of the plot is centered at the diastema of the central maxillary incisors (CMI). Its x-axis corresponds to the intersection of the mid-sagittal plane and a second plane, known as the maxillary occlusal plane (MaxOP), given by the tips of the central incisors and two other maxillary teeth on opposite sides of the mouth. The y axis was normal to the MaxOP, and intersected that plane at the origin, with the +x and +y directions out of the mouth, and toward the roof of the mouth, respectively. Along with the position of the pellets, the plot includes the palatal outline and posterior pharyngeal wall.

3.3 Vocal tract shape from XRMB data

For determining the actual shape of the vocal tract from XRMB x-y plot, vocal tract is assumed to be a tube like structure. Vocal tract opening is the distance between the palatal outline and the tongue outline measured perpendicular to the axis of the vocal tract. While determining the vocal tract opening at the lips, an offset needs to be applied to the positions of the upper lip and lower lip pellet points as these points are placed outside the vocal tract. The offset is calculated as half of the distance between the upper lip pellet point and the lower lip pellet point, for the utterance /aba/ during

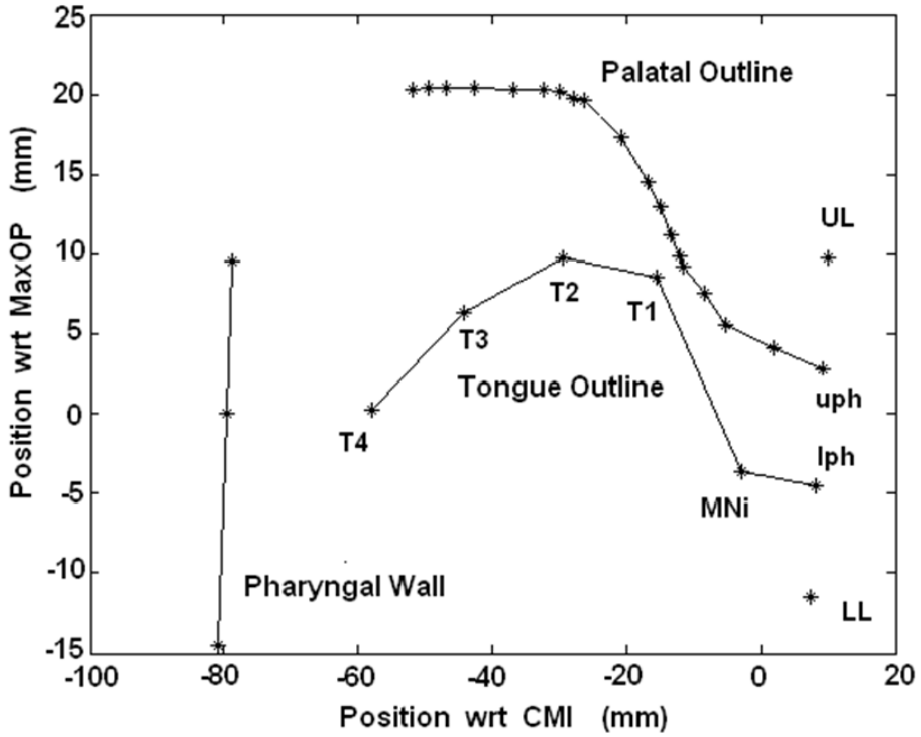


Figure 3.2 Palatal and tongue outlines for the VCV utterance /ada/

the closure. The offset is added to the position of the lower pellet and subtracted from the position of the upper pellet, along the straight line joining them. The resulting points are called as the upper phantom point (uph) and the lower phantom point (lph) and represent the upper lip and the lower lip on the vocal tract [27]. Curve joining the palatal points and the upper phantom point forms the new palatal outline. Curve joining the tongue pellet points, mandibular pellet point (MNi), and lower phantom point forms the tongue outline as shown in Fig. 3.2. Both the outlines are divided into 50 equal segments as shown in Fig. 3.3. The mid points of straight lines joining the corresponding points on the palatal outline and the tongue outline are obtained. Curve joining these mid points using straight line segments is taken as the axial curve of the vocal tract as shown in Fig. 3.4.

3.4 Estimation of place of articulation from XRMB data

Vocal tract opening at a particular point on the axial curve is taken as the distance between the palatal and the tongue outlines measured along the normal to the axial curve as shown in Fig. 3.5. For drawing the normal to the axial curve at a given point, the mean of the slopes of the line segments on the left and right sides of the point is

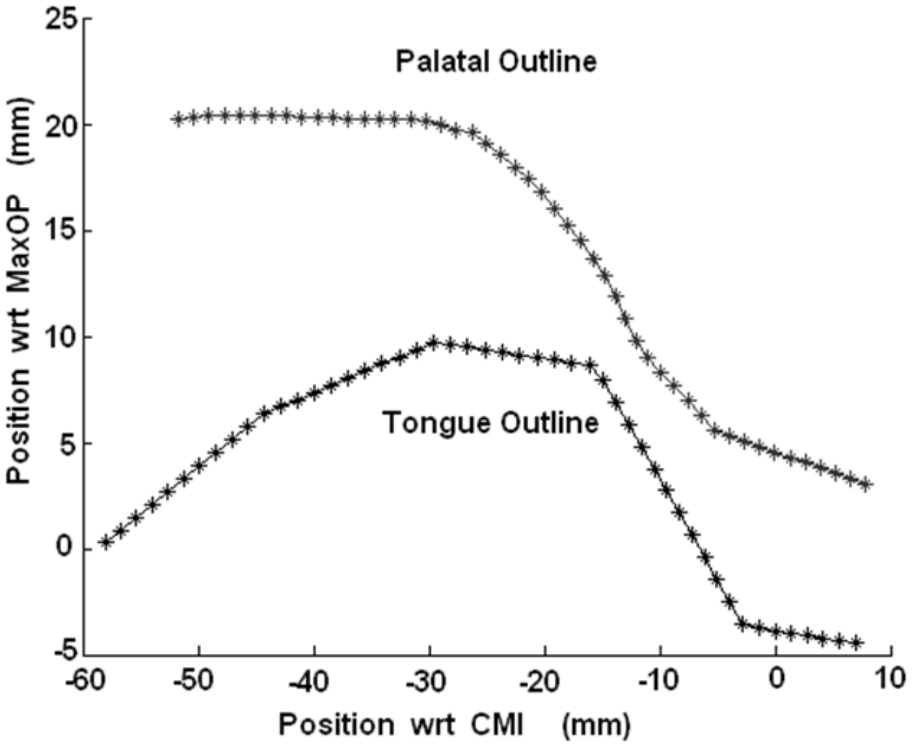


Figure 3.3 Division of palatal and tongue outlines in to equal sections

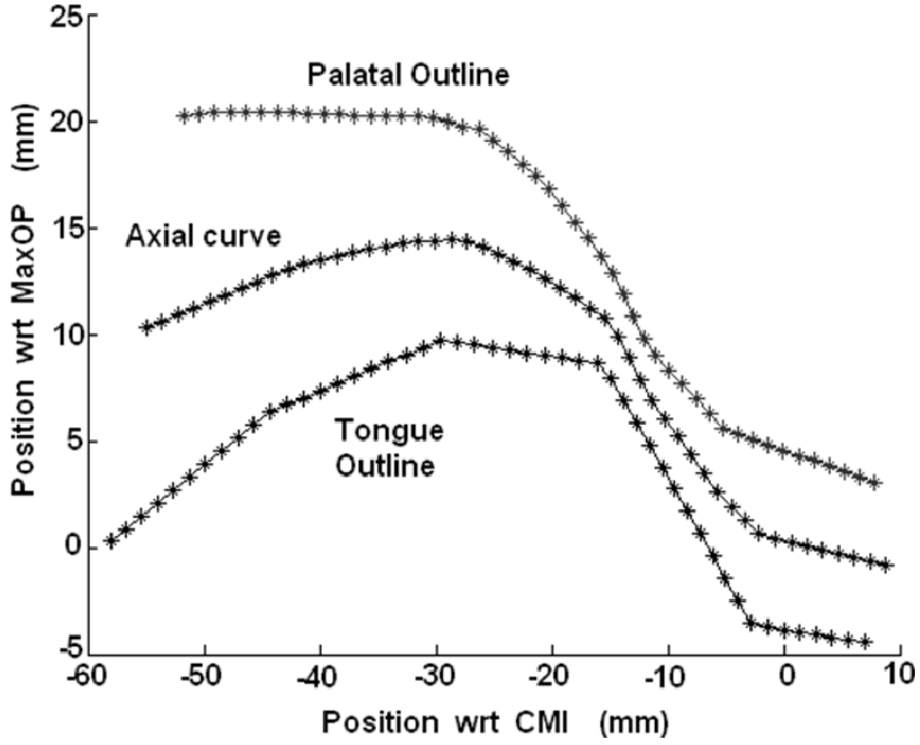


Figure 3.4 Axial curve of the vocal tract

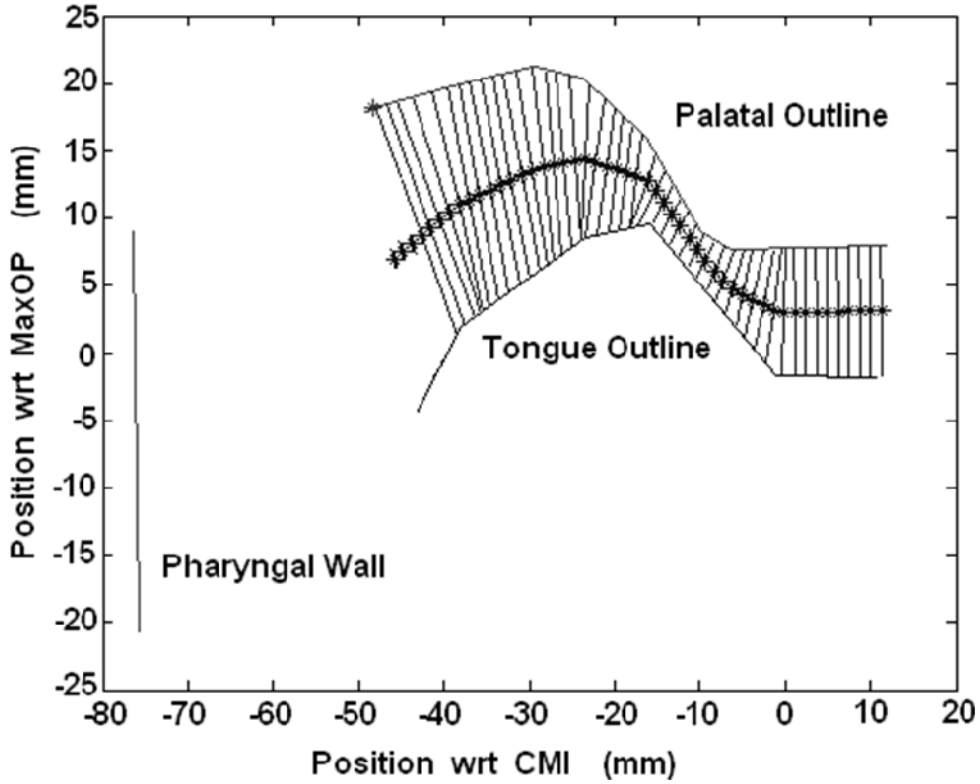


Figure 3.5 Vocal tract opening along the axial curve

taken as the slope of the curve. For the first and the last points, the slope of the first and the last line segments are used. Vocal tract openings are obtained at the points on the axial curve, having their x axis value in between the upper phantom point (uph) and the last palatal point. The place of articulation is obtained by finding the distance of the place of minimum opening from the lip position, along the axial curve.

3.5 Results

Place of constriction was obtained using the automated axial curve method as explained in Section 3.3 and 3.4 for /ʌCa/ utterances, involving stop consonants /b/, /d/ and /g/ from 35 speakers of XRMB database. The results obtained by the automated method were validated by comparing them with those obtained manually. Place of constriction was obtained manually along the length of the vocal tract by tongue profile method, which involves adding the straight line distances along the curve joining the lower lip marker and the pellet markers, as reported by Shah [28]. Fig. 3.6 shows the scatter plot of the places of constriction obtained from the automated axial curve method versus those obtained by the manual tongue profile

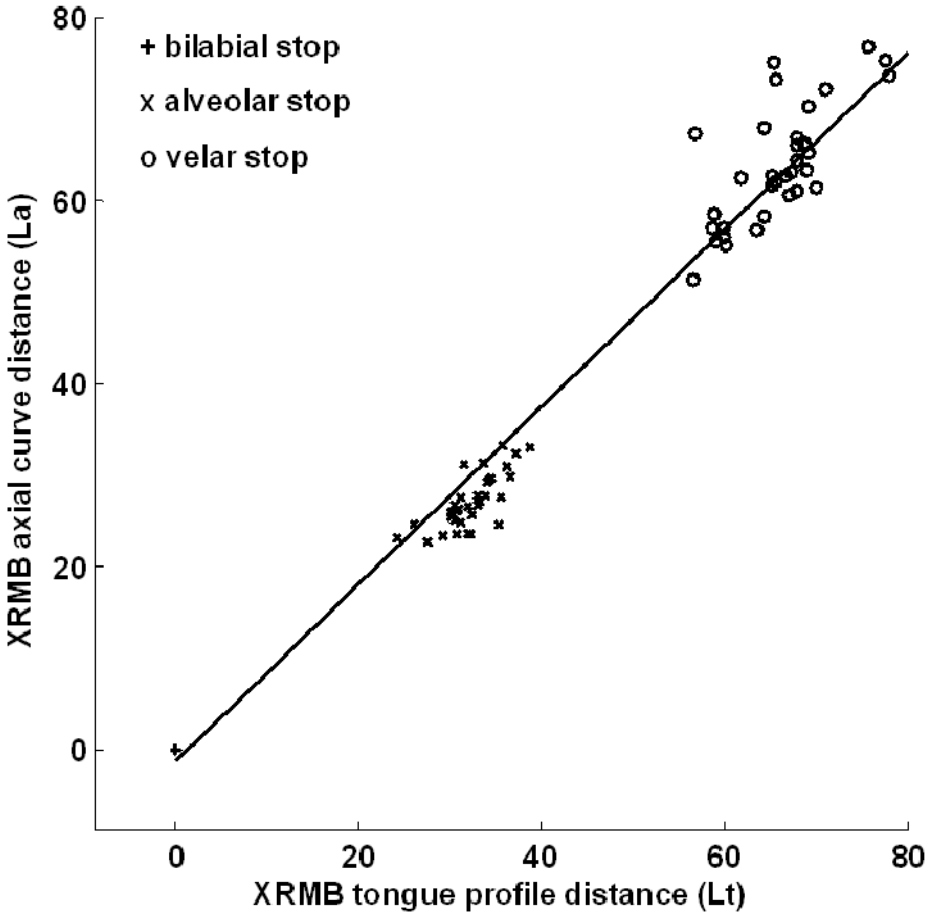


Figure 3.6 Scatter plot of place of constriction obtained from the tongue profile (*Lt*) versus those obtained from XRMB along axial curve (*La*) for 105 /ʌCa/ utterances involving stop consonants /b/, /d/ and /g/. Linear regression: *La*=-1.366+0.96795**Lt*

method. There is an excellent match between the two sets of values, with a correlation coefficient of 0.99.

Chapter 4

PLACE OF ARTICULATION FROM ESTIMATED VOCAL TRACT SHAPE

Speech training aids providing visual display of vocal tract shape are reported to be helpful in acquiring speech for hearing impaired children. Most of these training aids, use Wakita's LPC based method [11] for vocal tract shape estimation. Estimated vocal tract shape from the LPC based method needs to be interpolated before mapping it into the vocal tract graphics for smooth display of vocal tract shape. However, the vocal tract shape can show multiple constrictions or it can mask the original constriction when it is interpolated for smoother displays. This chapter investigates various polynomial curves suitable for the vocal tract graphics.

4.1 Curve fitting for vocal tract area function

For estimation of the vocal tract shape using Wakita's method [11], the LPC analysis of order 12 was carried out to model the vocal tract from lips to glottis as 12 cylindrical sections of equal length. The area values of the sections were calculated from the reflection coefficients obtained using the autocorrelation coefficients of the windowed frame. However, visual display of vocal tract shape requires a more realistic and smooth display. Baragi [29] used Bezier curve interpolation to convert 12 discrete area values along glottis-lips distance to 176 values before displaying on the screen. Bezier curve interpolation was computationally efficient and gave more realistic display of articulatory motion. Kshirsagar [30] displayed the vocal tract shape, by discarding the 2 sections at the glottis end as there was no significant change in the area values near glottis end. The total number of interpolated points after Bezier interpolation was still kept to be 176 and thus stretching the display of the rest of 10 sections. Shah [28] compared cubic spline interpolation with the Bezier curve interpolation for converting the 12 discrete values to 40 values and concluded that cubic spline function preserves the original shape much better than Bezier form interpolation, while a better smoothening of shape is observed with Bezier form

interpolation. The vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterances can be estimated by bivariate surface modeling of the shapes during transition segments [12]. However, it is difficult to obtain the place of maximum constriction precisely from the cubic spline interpolated area values of the stop closure during VCV utterances due to the variability in the vocal tract shape across the sections. Bezier curve provides smoother display but the actual place of constriction gets masked due to global control. Hence, investigation of different interpolation method for precisely locating the place of maximum constriction is needed.

4.2 Analysis of estimation of place of maximum constriction

Investigations were carried out for locating the place of maximum constriction in vowel /i/ using Bezier and cubic spline interpolation. Bezier curve uses four control points and produces an approximating curve which passes through 2 extreme points and the interior points are used to define tangents to the curve [31]. The Bezier curve has global control, i.e. change in any one of the point in the curve leads to the change in the modeling of all Bezier segments in the curve. Cubic-spline curve passes through all given points. Cubic-spline fits (n-1) individual segments of degree 3 to n given control points [31]. At each data point, the two polynomials connect and their first and second derivatives have the same values. Similar to Bezier curve, cubic spline also has no local control. Fig. 4.1 shows the mean of the area values obtained for the vowel /i/ at an inter frame interval of 5 ms by LPC analysis of the speech segment of length equal to twice the average pitch period. Maximum constriction can be observed at the front for the vowel /i/. Cubic-spline follows the original area values and it is difficult to locate the section with maximum constriction. Bezier curve provides smoother shape near the constriction but fails to identify the place of maximum constriction. This can be attributed to the lack of local control in Bezier interpolation. Cubic B-spline was investigated for interpolation as it provides a local control as well as a smoother display.

4.3 B-spline curve interpolation

The B-spline curve [31] offers local control at the cost of not necessarily passing through all the control points, and provides desired continuity between the individual

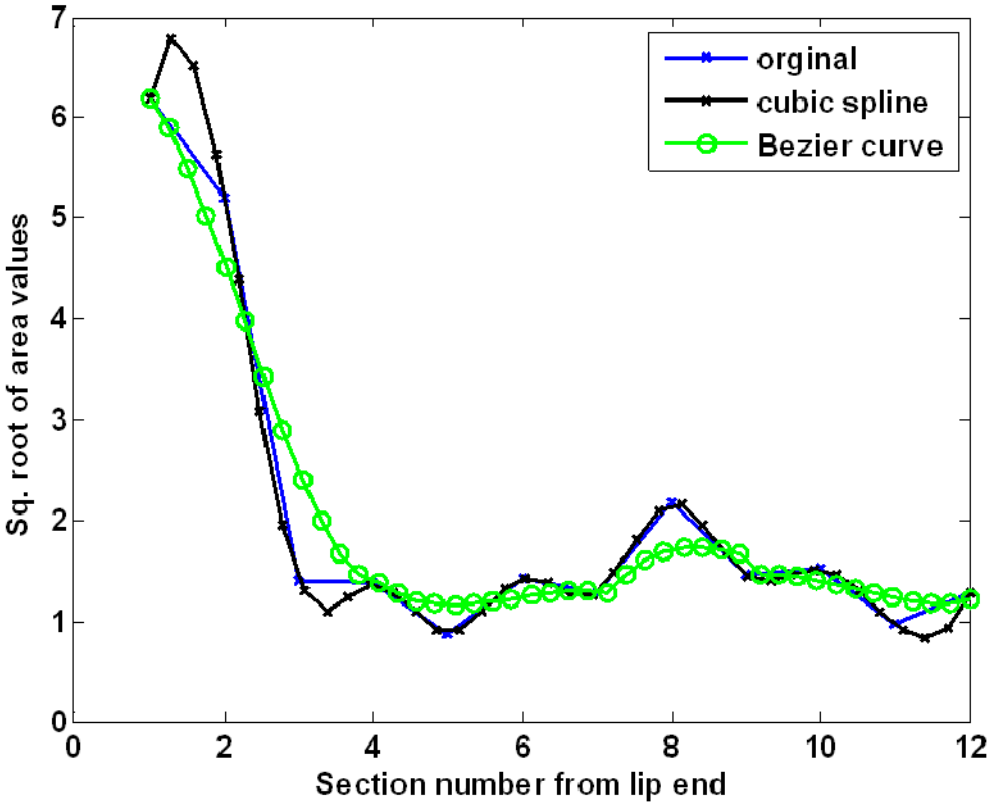


Figure 4.1 Plot of area values for vowel /i/ obtained using linear interpolation, cubic spline interpolation, and Bezier interpolation.

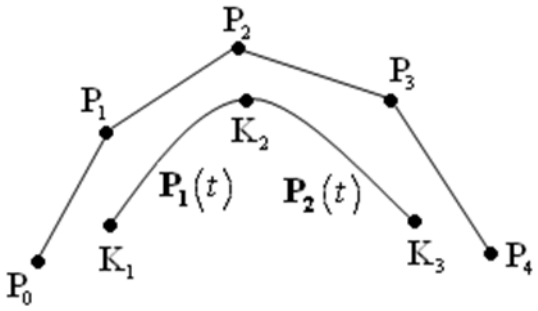


Figure 4.2 Cubic B-spline

spline segments. The B-spline curve provides a smooth curve approximating the given curve. The curve is defined by $n + 1$ control points, P_0, $P_{1,\ldots,}$, P_n and we need to calculate a set of cubic B-spline curves $\mathbf{P}_i(t)$, each defined by 4 control points P_{i-1}, P_i, P_{i+1}, and P_{i+2} as shown in Fig. 4.2. The first segment, defined by P_0, P_1, P_2, and P_3, starts at K_1 and ends at K_2. Any changes in the points after P_3 will not affect this segment. Thus, the interpolation provides the local control. Similarly, each B-spline segments are defined for the entire curve considering four control points at a time. The coefficients of the cubic B-spline polynomial are obtained using the C^2 continuity

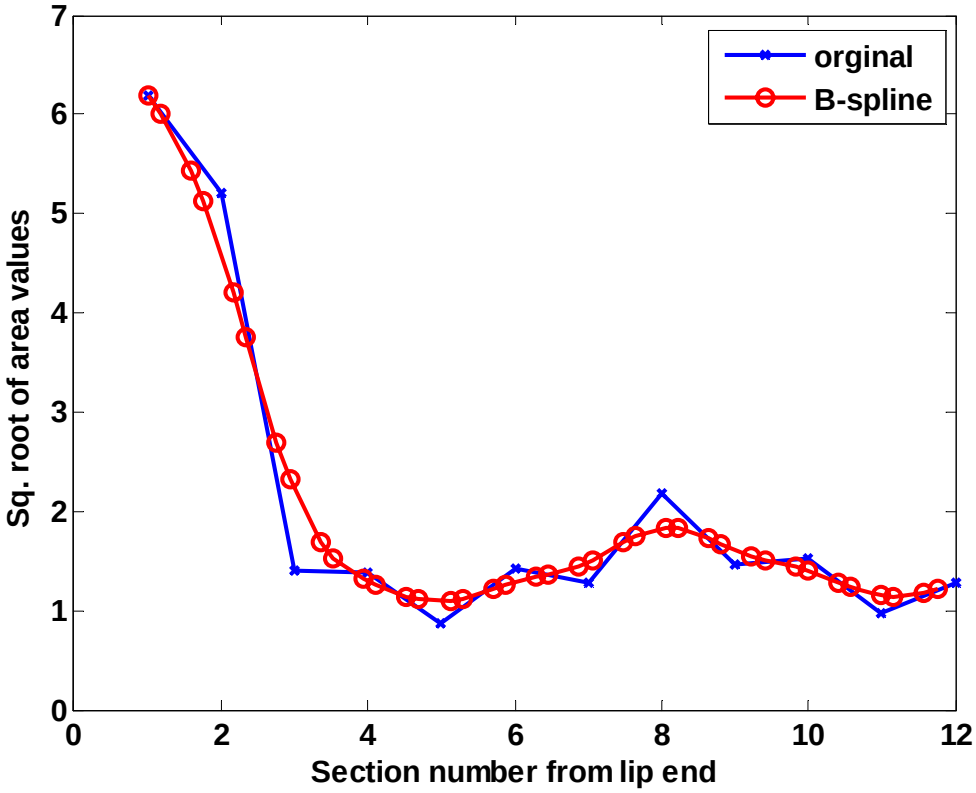


Figure 4.3 Plot of area values for vowel /i/ obtained using B-spline interpolation.

condition between the two segments. The cubic B-spline segment equation for segment *i* can be written as,

$$\mathbf{P}_i(t)=\frac{1}{6}(-t^3+3t^2-3t+1)\mathbf{P}_{i-1}+\frac{1}{6}(3t^3-6t^2+4)\mathbf{P}_i+\frac{1}{6}(-3t^3+3t^2+3t+1)\mathbf{P}_{i+1}+\frac{t^3}{6}\mathbf{P}_{i+2} \quad (4.1)$$

Each segment of curve lies between 2 points K_i and K_{i+1} given by,

$$\mathbf{K}_i=\mathbf{P}_i(0)=\frac{1}{6}\left(\mathbf{P}_{i-1}+4\mathbf{P}_i+\mathbf{P}_{i+1}\right),$$

$$\mathbf{K}_{i+1}=\mathbf{P}_i(1)=\frac{1}{6}\left(\mathbf{P}_i+4\mathbf{P}_{i+1}+\mathbf{P}_{i+2}\right)$$

The B-spline curve interpolation was used to convert 12 discrete area values of the vowel /i/ to 40 discrete area values. In order to force the curve to pass through the first and last control points, two phantom points are defined at these points. Fig. 4.3 shows the area values obtained from linear, Bezier, cubic spline, and cubic B-spline interpolation. It can be observed that B-spline curve provides distinct minimum at the front constriction compared to other interpolation methods. The place of maximum

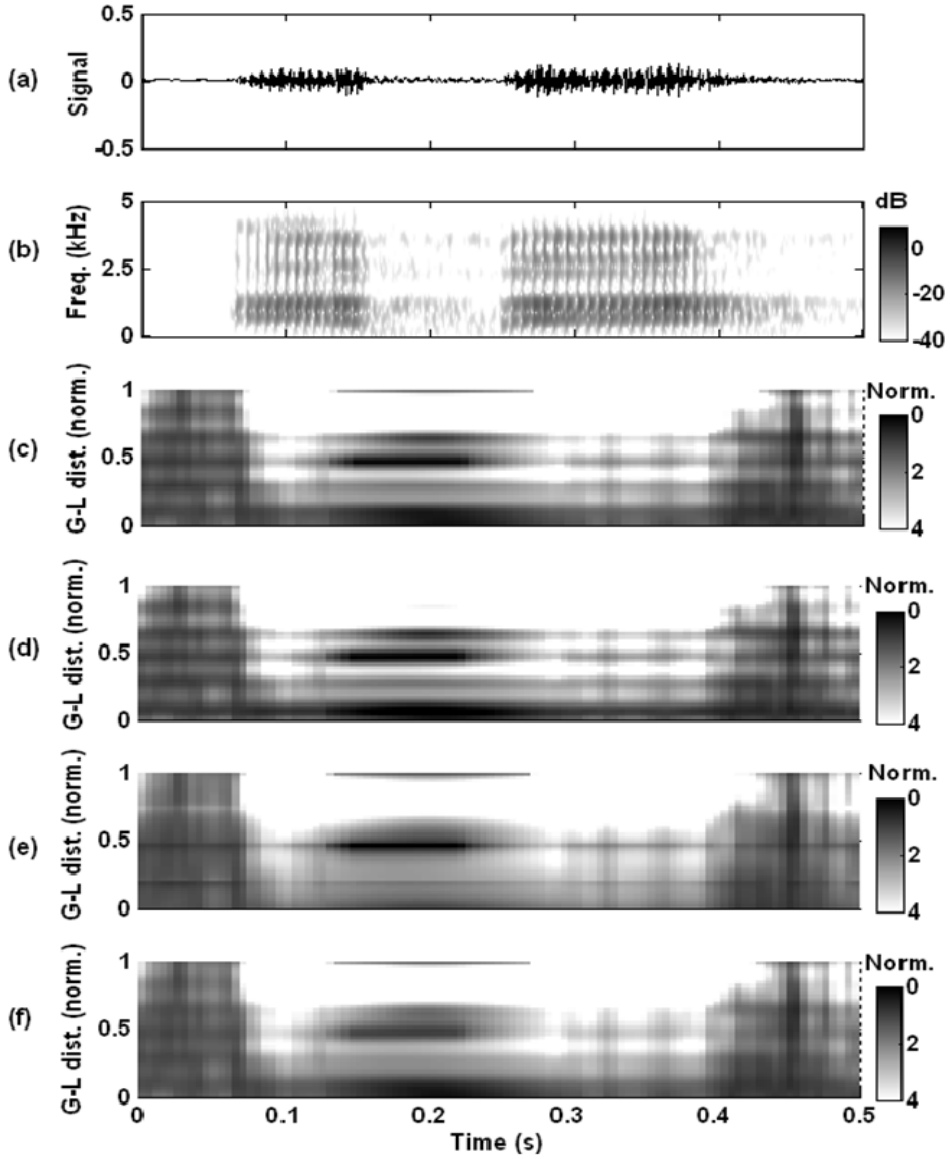


Figure 4.4 Interpolation results for the VCV utterance /aba/: (a) speech waveform; (b) wideband spectrogram; areagrams obtained using (c) linear interpolation, (d) cubic spline interpolation; (e) Bezier interpolation, and (f) B-spline interpolation.

constriction can be obtained by finding minimum in the B-spline interpolated curve.

4.4 Estimation of place of maximum constriction in stop closure

To examine the effectiveness of the cubic B-spline interpolation for bringing out the place of maximum constriction in stop closure during VCV utterances, vocal tract shapes were obtained for VCV utterances from several speakers. Vocal tract shape was estimated using Wakita's method [11], the speech signal was sampled at 10 kHz and

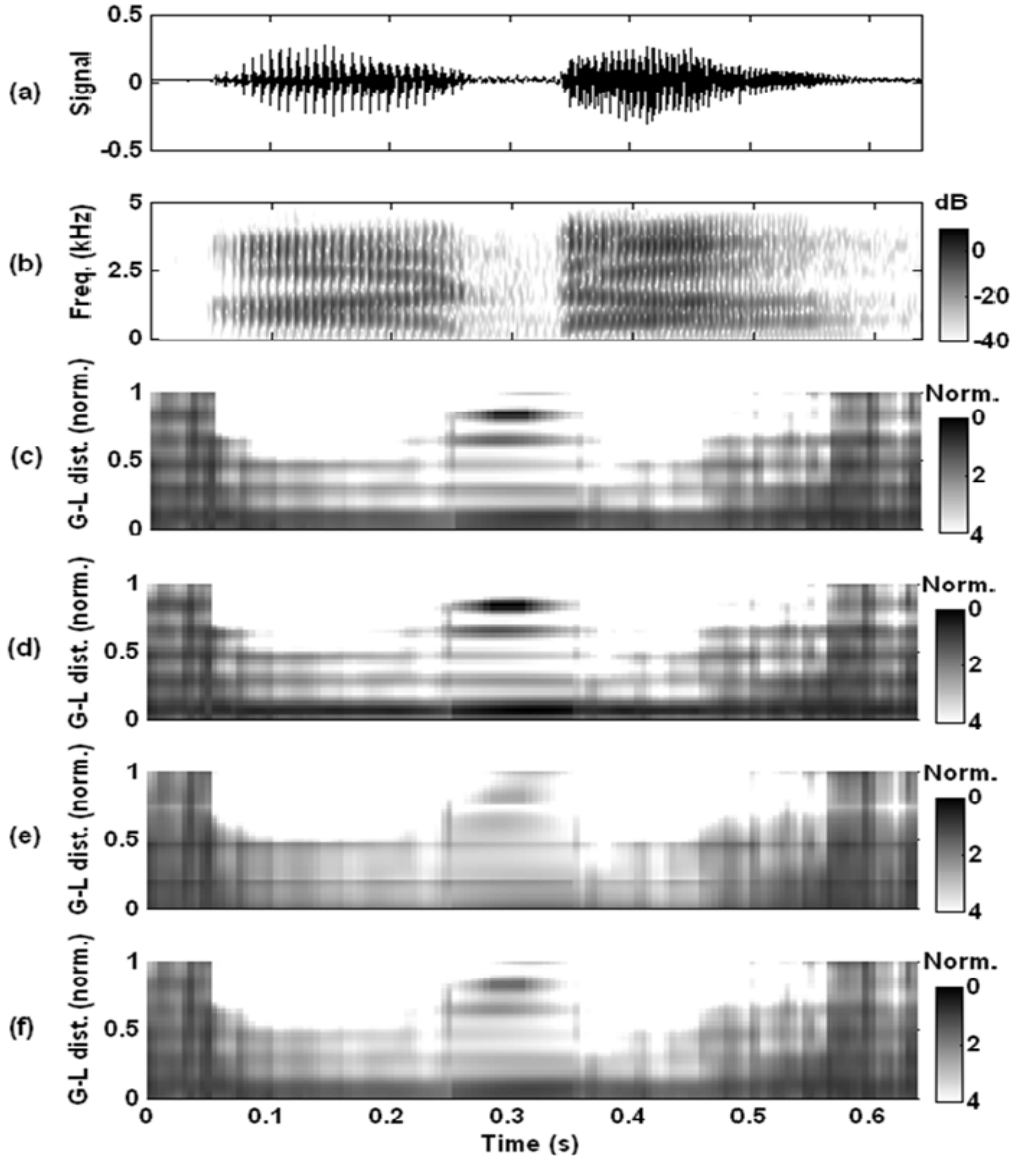


Figure 4.5 Interpolation results for the VCV utterance /ada/: (a) speech waveform; (b) wideband spectrogram; areagrams obtained using (c) linear interpolation, (d) cubic spline interpolation; (e) Bezier interpolation, and (f) B-spline interpolation.

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 window shift of 5 ms. LPC analysis of order 12 was carried out to obtain the vocal tract of 12 sections from glottis to lips. The vocal tract shape during the closure is estimated by bivariate surface modeling of the VC and CV transition area values. 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 along glottis to lips were

converted to 40 values using linear, cubic spline, Bezier, and cubic B-spline interpolation and square root of these values were used for plotting the areagram. Fig. 4.4 shows the areagrams obtained using linear, cubic spline, Bezier, and cubic B-spline interpolation for the natural utterance /aba/ from a male speaker. Multiple constrictions can be observed in the linear and cubic spline interpolated areagrams. Bezier interpolation provides better smoothened shape but shows constriction at lips and at velar position at normalized distance of 0.5. Cubic B-spline interpolation preserves the constriction at lip and masks the other dark bands corresponding to alveolar and velar positions. Fig. 4.5 shows the areagrams obtained using linear, cubic spline, Bezier, and cubic B-spline interpolation for the natural utterance /ada/ from a male speaker. Linear and cubic spline interpolation shows distinct constriction at the normalized distance of 0.8 indicating alveolar stop. Better smoothening of the shape can be observed in the areagram obtained by Bezier interpolation but the actual constriction is masked. Cubic B-spline interpolation provides better smoothened shape while preserving the original constriction.

Chapter 5

EFFECT OF LIP SCALING DURING CLOSURE OF THE VCV UTTERANCES

5.1 Introduction

Most of the speech training aids providing visual feedback of the articulatory efforts use Wakita's inverse filtering method [11] for vocal tract shape estimation. In this method, the vocal tract is modeled as a lossless acoustic tube with sections of equal length and varying cross-sectional area. The ratios of the areas on the two sides of section interfaces are calculated from the reflection coefficients obtained from the autocorrelation coefficients of the windowed speech signal. The ratios are converted into area values by scaling them with respect to the area of a reference section. For the steady state vowel sounds, constant normalized area of unity at the glottis end is used as a reference area value. Wakita [13] reported that the glottis area changes during varying vocal tract configuration leading to improper area estimation. Nayak *et al*. [25] used the lip opening area estimated from the video images, as a reference area for scaling the area ratios, to obtain the area values during speech utterances with transitional tract configuration. The scaling improved the estimation of area values of vowels at the places of maximum opening without significantly affecting the places of constriction.

The vocal tract shape during the stop closures of VCV utterances can be estimated by bivariate surface modeling of the vocal tract area function during the VC and CV transition segments [12]. The accuracy of the estimated shape during the closure depends on the accuracy of the estimation of dynamically changing shapes during VC and CV transition segments [28]. The method of scaling of area values using estimated lip area values as proposed by Nayak *et al*. [25], may be useful in improving the vocal tract shape estimates during VC and CV transition segments of VCV utterances. Thus, it is useful to investigate the effect of scaling of the area values during VC and CV transition segments of VCV utterances using lip area values.

The investigation was carried out to estimate the vocal tract area function during VCV utterances of type /aCa/, involving both voiced and unvoiced stop consonants. Wakita's LPC based method [11] was used for estimating the area ratios from the speech signal. The area ratios were scaled by two methods, using unity area at the glottis end and by using lip opening area estimated from video images of speaker. Vocal tract shape during the stop closures of VCV utterances was estimated by 2D interpolation of the VC and CV transition area values.

5.2 Scaling of the vocal tract area function using lip area values

For estimation of vocal tract shape using Wakita's method [11], the speech signal was sampled at 10 kHz with 16-bit quantization. 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. An inter-frame interval of 5 ms was used for capturing the variation in the vocal tract shape during transition segments. LPC analysis of order 12 was carried out to compute the reflection coefficients from the autocorrelation coefficients of the windowed frame. The vocal tract from lips to glottis 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. Minimum windowed energy index method [26] was then applied on the estimated area values for improving the consistency of the vocal tract shape. The amount of opening in a section of the vocal tract is obtained as the square root of the area value.

Audio-visual signals were recorded using a Sony HDR-CX130E camcorder having a resolution of 1080p (1920x1080 pixels) at 50 frame/s. Initially Viola–Jones' method [36] was used for face detection, followed by localization of mouth region using the method proposed by Hsu *et al*. [37]. Lip opening areas were obtained manually using the method proposed by Nayak [38]. The points corresponding to the lip opening were marked manually using the mouse clicks in the detected mouth sub-image. Lip opening was represented by joining all the marked points by straight lines to form the polygon. Number of pixels within the polygon was used as the lip area. Lip area values were calculated for each video frame captured at an interval of 20 ms. The lip opening area values were obtained at an interval of 5 ms by using cubic spline interpolation to match the inter-frame interval of the area values estimated using LPC

analysis of speech signal. These area values were normalized with respect to the area of the largest lip opening.

The area ratios of the 12 sections estimated from the LPC analysis of the speech signal were scaled by the scaling factor obtained from the lip area values for corresponding analysis frame. Scaling factor S_f is given by,

$$S_f = L_v / L_s \tag{5.1}$$

where, L_v is the actual lip area value obtained from the video image, and L_s is the estimated lip area value from the LPC analysis.

5.3 Results

Video along with the audio were recorded for 6 male speakers for the VCV utterances corresponding to the bilabial stops (/aba/ and /apa/), alveolar stops (/ada/ and /ata/), and velar stops (/aga/ and /aka/). During the stop closure of the VCV utterances, vocal tract shape was estimated by 2D interpolation of the transition area values preceding and following the stop closure [12]. To study the effect of lip scaling, glottis section area values after 2D interpolation, for both lip scaled area values and the area values scaled by unity area at the glottis end were plotted. Fig. 5.1, 5.2, and 5.3 shows the glottis area values for the VCV utterances /aba/, /ada/, and /aga/, respectively. Dotted line and solid line in the glottis area value plot corresponds to the original glottis area value and lip scaled glottis area value, respectively. Inner vertical dashed lines indicate the start and stop of the closure. For the utterance /aba/, lip scaled glottis area value decreases gradually during the VC transition, reaches to zero at the closure, and increases during the CV transition as shown in Fig. 5.1. Similar pattern was observed for unvoiced bilabial stop /apa/. For the utterance /ada/, lip scaled glottis area value increases gradually during the VC transition, reaches maximum at the closure, and decreases during the CV transition as shown in Fig. 5.2. Similar pattern was observed for the unvoiced alveolar stop /ata/. For the utterance /aga/, lip scaled glottis area value increases gradually during the VC transition, reaches maximum at the closure, and decreases during the CV transition as shown in Fig. 5.3. Similar pattern of the glottis area values can be observed across all the speakers for both the voiced and unvoiced stops. It can be seen that the lip scaled glottis area values are constant in the steady state part of the vowel and change during the transition part preceding and

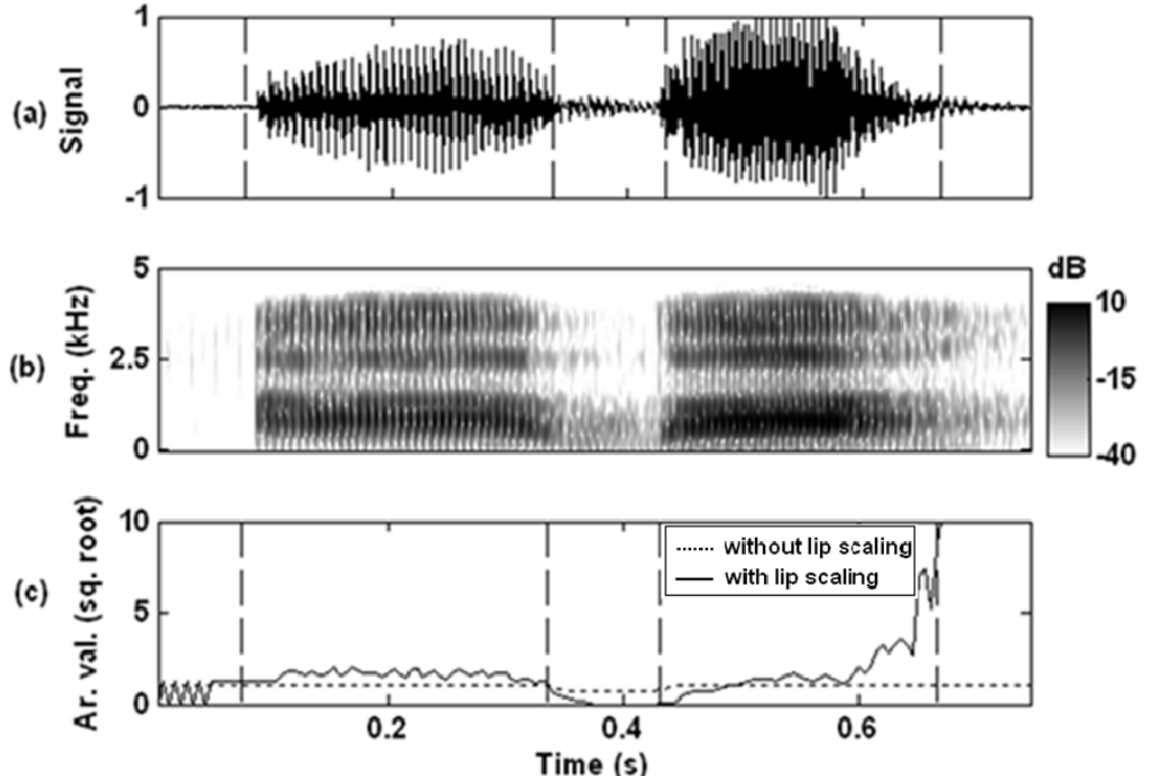


Figure 5.1 Comparison of the glottis area without lip scaling and after lip scaling for VCV utterance /aba/ (a) speech waveform, (b) wideband spectrogram, and (c) glottis area values

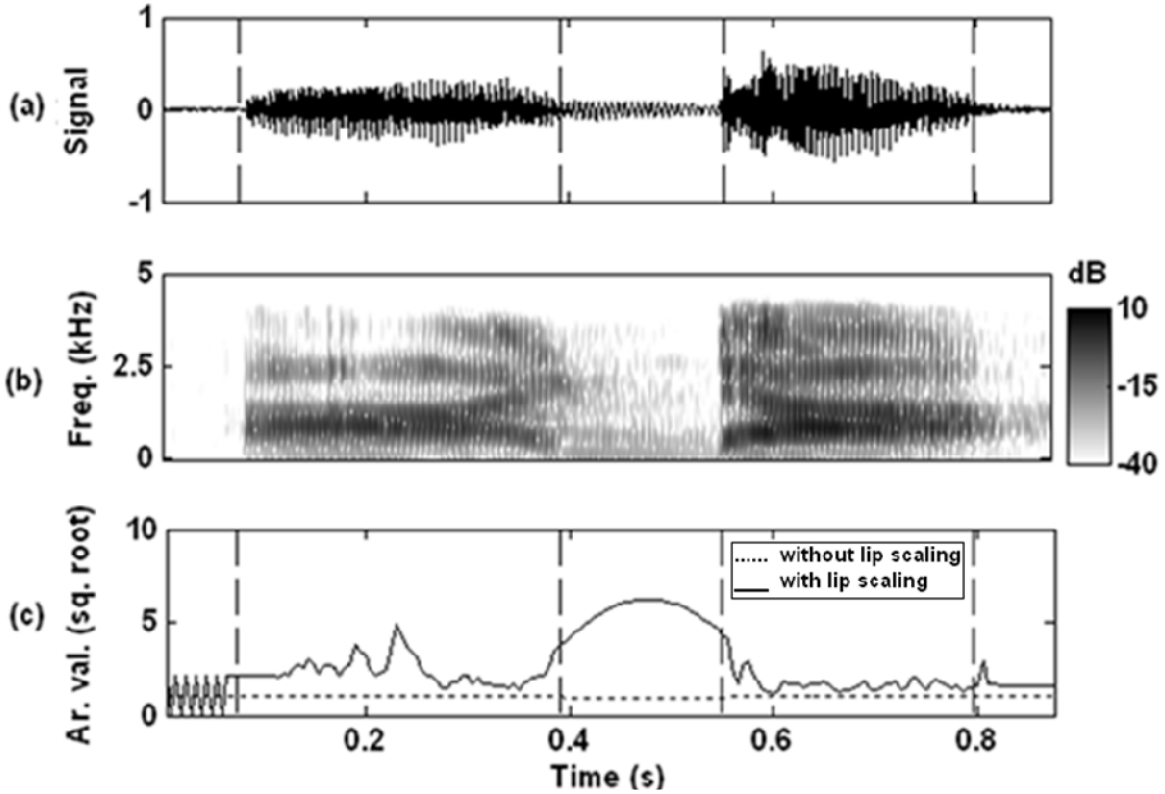


Figure 5.2 Comparison of the glottis area without lip scaling and after lip scaling for VCV utterance /ada/ (a) speech (b) wideband spectrogram, and (c) glottis area values

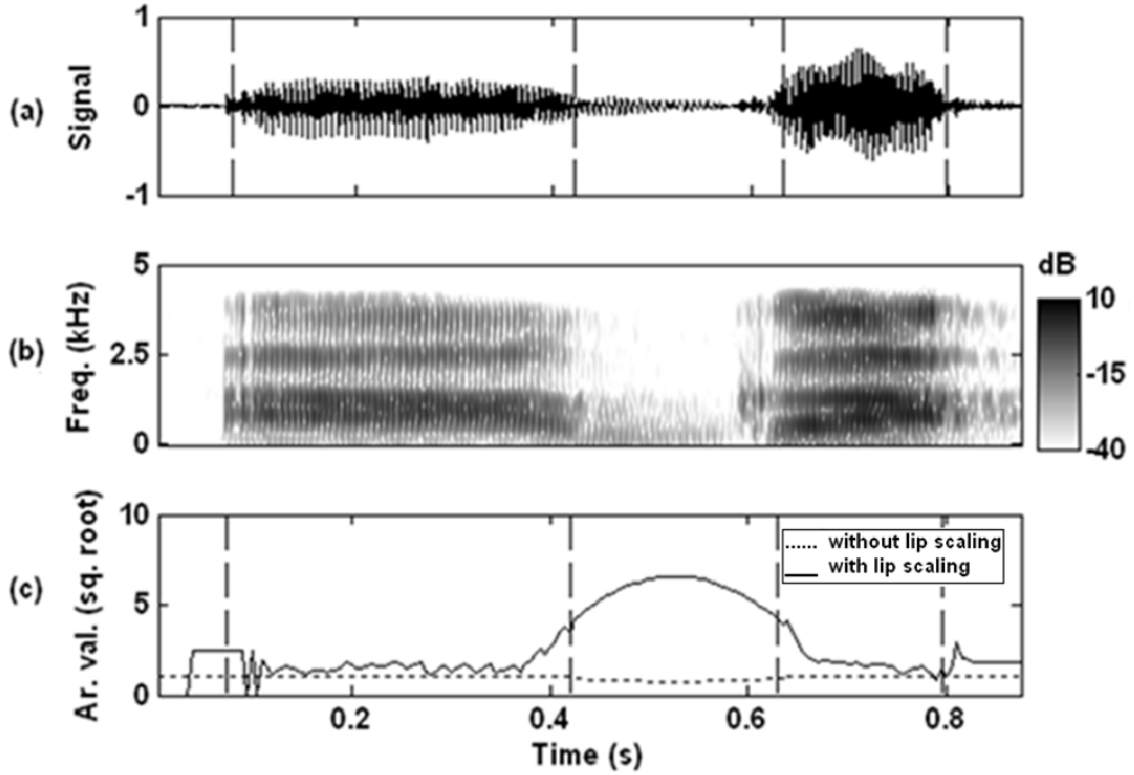


Figure 5.3 Comparison of the glottis area without lip scaling and after lip scaling for VCV utterance /aga/ (a) speech waveform, (b) wideband spectrogram, and (c) glottis area values

following the closure, whereas, the original glottis area values remains constant for the whole duration. This indicates that the assumption of constant area at the glottis is invalid.

Areagrams were obtained both for glottis scaled and lip scaled area values. For smooth display, area values are interpolated from 12 to 40 sections along glottis to lips using cubic B-spline interpolation. Fig. 5.4 shows the speech waveform, wideband spectrogram, areagram displaying area values scaled by unity area at the glottis end, and areagram displaying the area values scaled by the lip opening area, for bilabial stops /aba/ and /apa/. For the utterance /aba/, lip scaled areagram shows very low area value at the lips during the stop closure as compared to the glottis scaled areagram. For the utterance /apa/, both areagrams show constriction at the lips during the closure, but the lip scaling has increased the contrast between place of constriction and other sections. For the alveolar stops /ada/ and /ata/, lip scaled areagram shows significant improvement in the estimated place of constriction. For the utterance /ada/, glottis scaled areagram shows constriction for the region extending from the normalized distance of 1.0 to 0.75 as shown in the left side of the Fig. 5.5(c). Lip

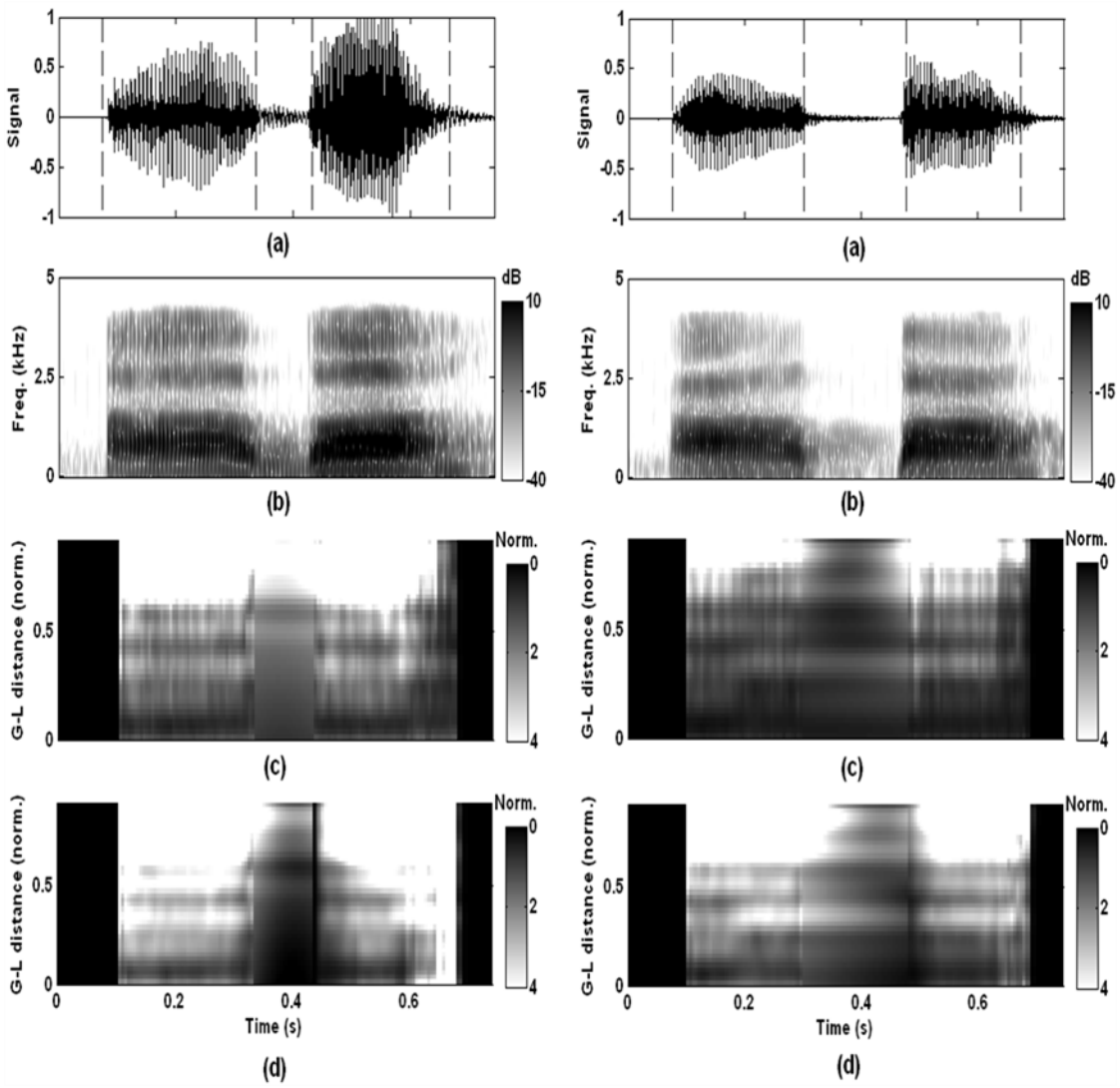


Figure 5.4 Interpolation results for the VCV utterance with bilabial stops: (a) speech signal, (b) wideband spectrogram, (c) glottis-scaled areagram, and (d) lip scaled areagram. Left: /aba/, right: /apa/

scaled areagram in the Fig. 5.5(d) shows distinct constriction at the normalized distance of 0.8 matching with the value of 0.75 – 0.89, for the alveolar stops, as estimated from MRI [23]. Similarly, for the utterance /ata/, the place of constriction estimated for lip scaled area value was at the normalized distance of 0.8. Fig. 5.6 shows a comparison of the glottis scaled and lip scaled areagram for the velar stops /aga/ and /aka/. Lip scaled area values based estimated shapes were not able to locate the precise place of constriction during the closure for the velar stops. However, wide opening at the front for both the /aga/ and /aka/ sounds was observed as compared to glottis scaled ones, which shows constriction at the front for both the utterances.

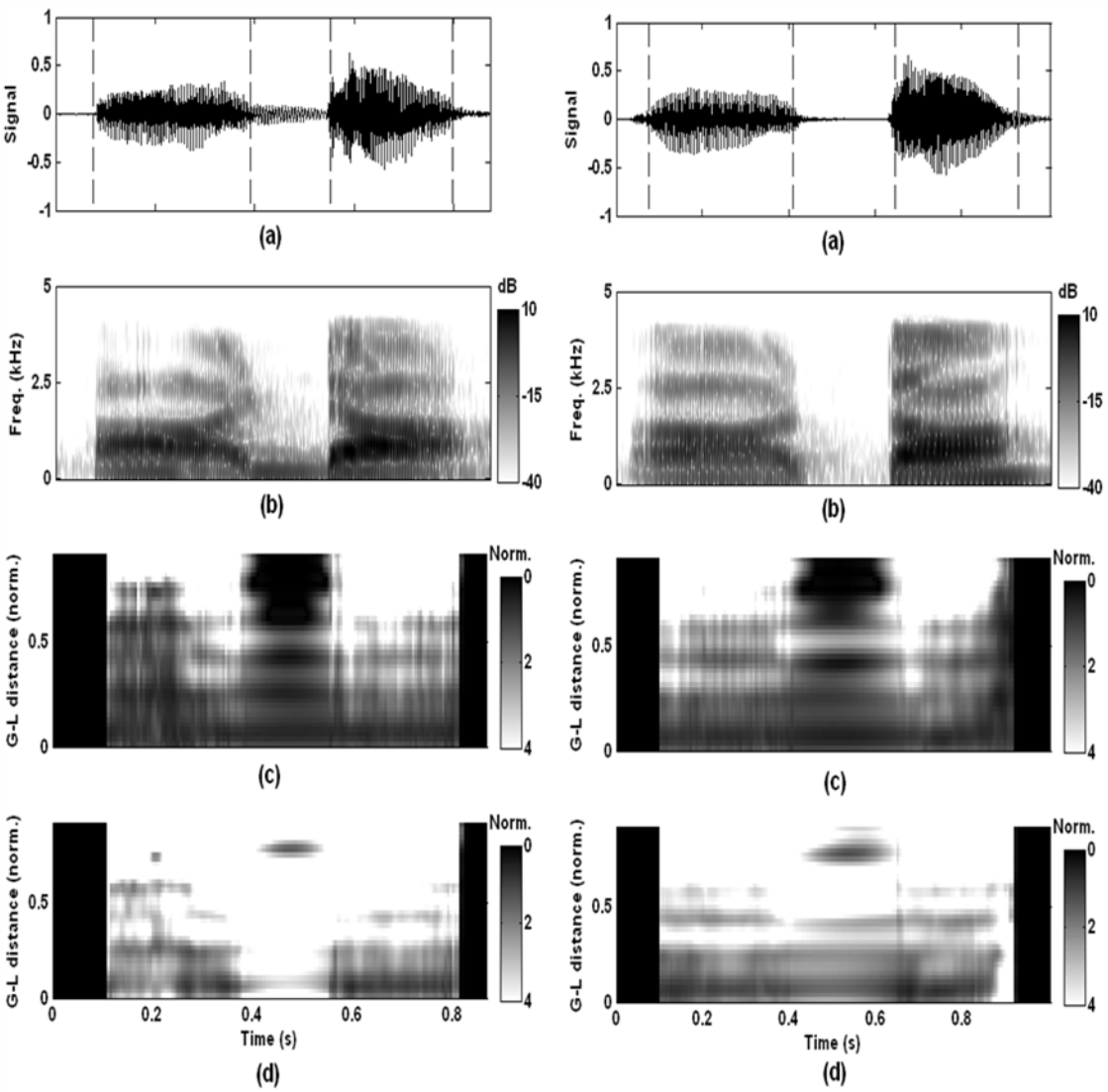


Figure 5.5 Interpolation results for the VCV utterance with alveolar stops: (a) speech signal, (b) wideband spectrogram, (c) glottis-scaled areagram, and (d) lip scaled areagram. Left: /ada/, right: /ata/

Similar behavior was observed across all the 6 speakers for their corresponding VCV utterances.

5.4 Discussion

Analysis of the lip scaled estimated shapes during the stop closure of the VCV utterances, shows that scaling of the area ratios by lip values significantly improves the estimation of the place of articulation for the bilabial and alveolar stops. It shows marginal improvement in the case of velar stops. However, use of a high resolution camera with a higher frame rate is needed for tracking the lip opening accurately. The

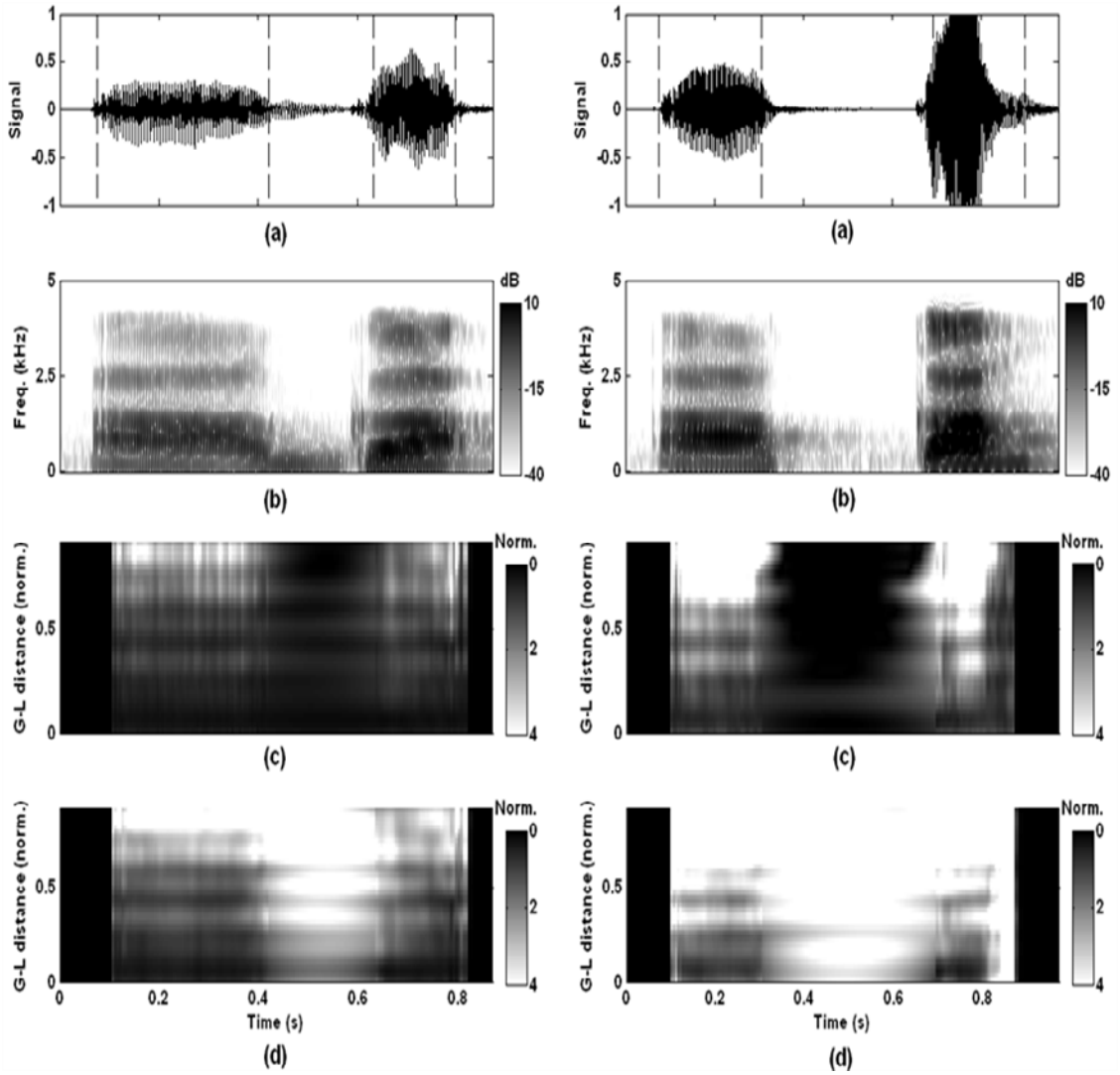


Figure 5.6 Interpolation results for the VCV utterance with velar stops: (a) speech signal, (b) wideband spectrogram, (c) glottis-scaled areagram, and (d) lip scaled areagram. Left: /aga/, right: /aka

duration of the transition segment can be as low as 20 ms. Camcorder used in the current project has a frame rate of 50 frames/s providing a video frame every 20 ms. Thus, it is possible to capture at most one lip image during a transition segment. In certain cases the captured lip image during transition was towards one end of the transition, resulting in an improper scaling of the area values. For example, Fig. 5.4 shows a dark vertical strip at the end of the closure duration in the lip scaled areagram for the utterance /aba/. This was due to the absence of video frame during the CV transition segment leading to a very low lip area value. Therefore, camera with a higher frame rate is needed for proper estimation of the vocal tract shape for the VCV utterances by using lip opening as reference area.

Chapter 6

DYNAMIC DISPLAY OF VOCAL TRACT SHAPE

6.1 Vocal tract graphics for speech training aids

For estimation of vocal tract shape using Wakita's method [12], the LPC analysis of order 12 was carried out to model the vocal tract from lips to glottis as 12 cylindrical sections of equal length. Vocal tract shapes estimated from Wakita's LPC based method for all the frames are displayed using a slow moving animation. Cross sectional areas of 12 sections are converted to 20 sections using cubic spline interpolation. Square root of the estimated area values gives the height of the vocal tract at each section. Mid-sagittal view of the vocal tract with the tongue in neutral position is shown in Fig. 6.1. Vocal tract is divided into 20 equal sections. The section lines are drawn considering the shortest distance between the upper palate and tongue [3]. The slopes of these lines are pre-calculated. Animation was developed, assuming the upper jaw to be fixed and lower jaw to be movable. A base image is created consisting only of upper jaw and neck portion. Estimated height *h* obtained by LPC

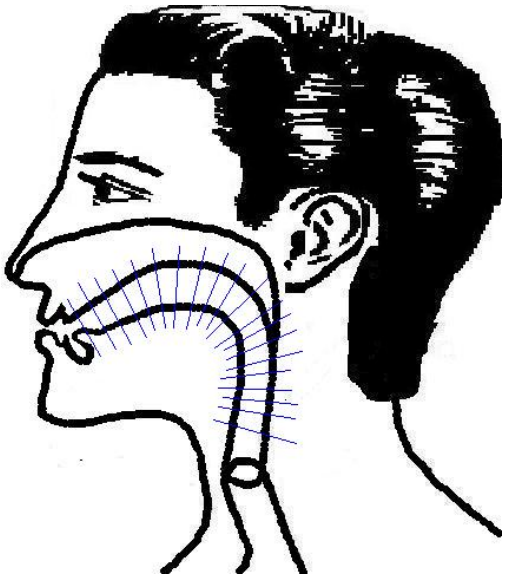


Figure 6.1 Mid sagittal view of vocal tract

analysis is mapped to the height d in the animation graphics as the following,

$$d = \left(h/h_{\max}\right) d_{\max} \tag{6.1}$$

where, $d_{\max}$ is the maximum possible height in the animation graphics and $h_{\max}$ is the maximum possible height estimated from the LPC analysis. When the estimated heights change, the corresponding animation height will also change and new points of the tongue on its corresponding section in animation are obtained by assuming the slopes of section lines to remain unchanged. A line is drawn passing through all the tongue points using linear interpolation.

6.2 Calculation of the tongue points

The graphics image is assumed to be in 4th quadrant of rectangular coordinate system. For animation, (x_1, y_1) is used as the point of the neutral tongue position and (x_2, y_2) as the point of required new tongue position. The slope m of a section is given by

$$m = \tan(t) = y/x \tag{6.2}$$

A scenario of section line having negative slope is shown in Fig. 6.2. Initially, the tongue point is at (x_1, y_1) and the new point (x_2, y_2) is at a distance d from the (x_1, y_1). Then, the change in the x and y coordinates are calculated using,

$$x = \frac{d}{\sqrt{1+m^2}} \ , \ y = m\frac{d}{\sqrt{1+m^2}} \tag{6.3}$$

New coordinate points of tongue are then obtained as,

$$y_2 = y_1 + y, \ x_2 = x_1 + x \tag{6.4}$$

Similarly for section with positive slopes as shown in Fig. 6.3, new coordinate points of tongue are given by,

$$y_2 = y_1 + y, \ x_2 = x_1 - x \tag{6.5}$$

After obtaining all the tongue points, a line is drawn passing through all these points using linear interpolation.

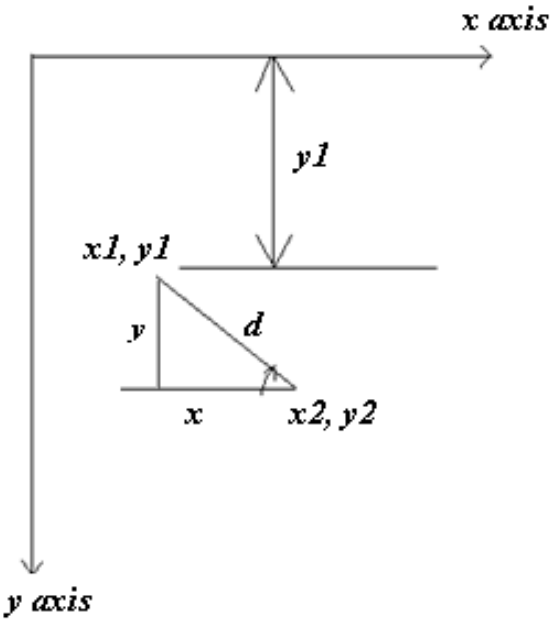


Figure 6.2 Calculation of *x* and *y* co-ordinates, for the line with negative slope

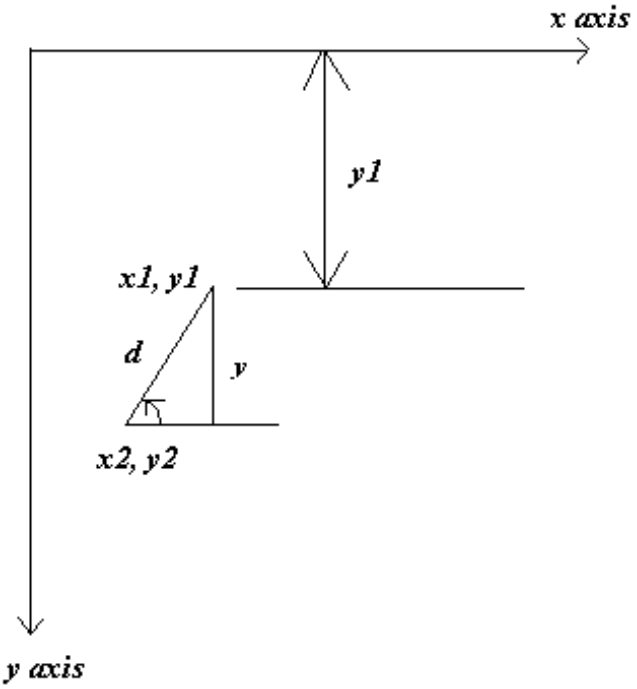


Figure 6.3 Calculation of *x* and *y* co-ordinates, for the line with positive slope

6.3 GUI for speech training aid

A GUI with speech waveform, spectrogram, areagram, options for selecting the speech signals, and graphics of vocal tract shape developed for vowel and VCV utterances is shown in the Fig. 6.4. Separate displays are used for teacher and student.

The estimated vocal tract shapes of each of the analysis frame are mapped to the graphics using the method as described in Sections 6.1 and 6.2. The delay between the displays of frames can be controlled using scroll bar with the left end providing the smallest delay and right end providing the largest delay. The graphics of the student and teacher can be displayed separately as well as in a combined manner.

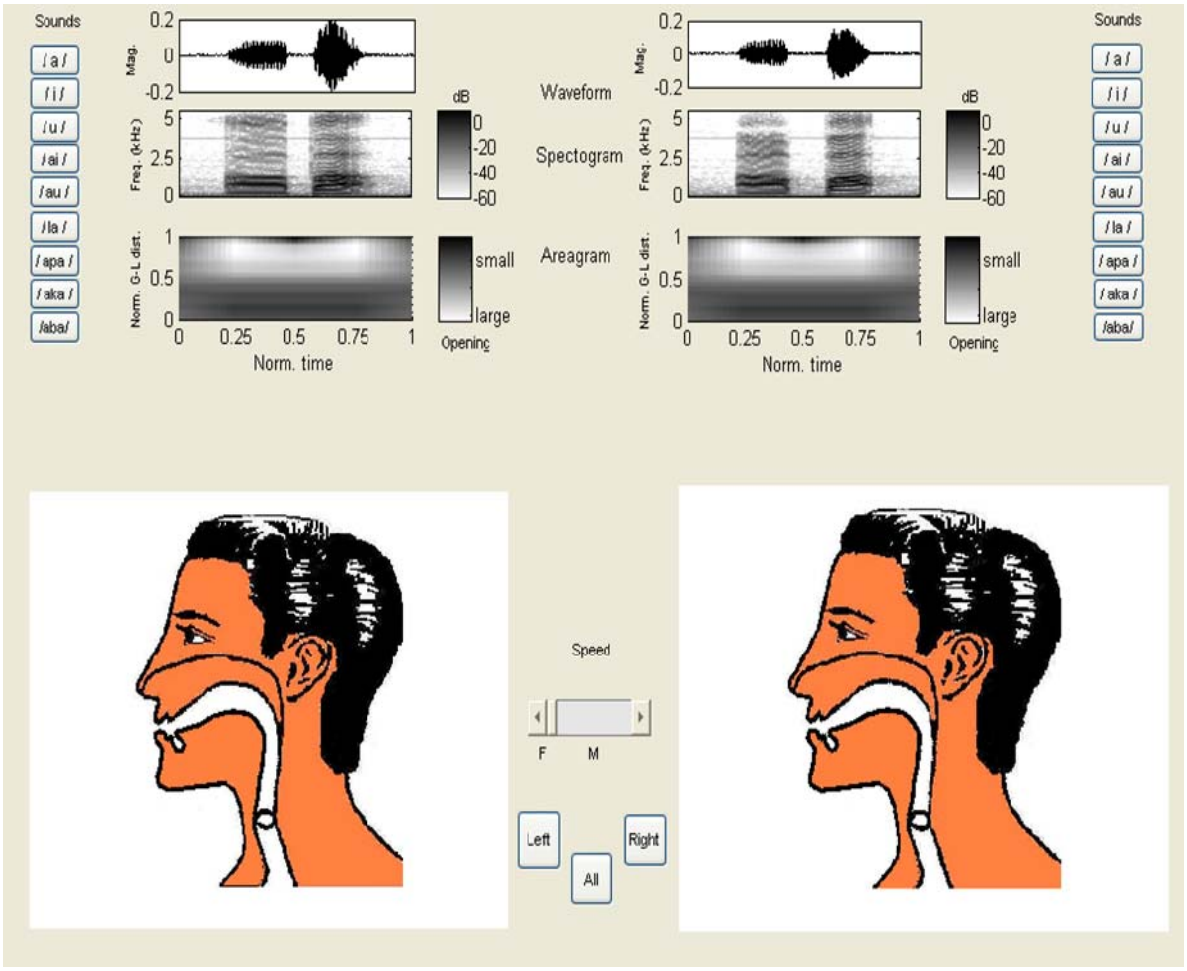


Figure 6.4 GUI for displaying the speech parameters and vocal tract shape for speech training aid

Chapter 7

SUMMARY AND CONCLUSIONS

Visual feedback of the vocal tract shape in the speech training systems are reported to be useful for improving vowel articulation. These systems use LPC analysis of speech for vocal tract shape estimation. The vocal tract shape during the stop closures of VCV utterances can be estimated by bivariate surface modeling of the vocal tract area on the VC and CV transition segments. The place of constriction estimated during the closure segment of VCV utterances was validated with the ones obtained manually from the XRMB database. A method to automatically obtain actual place of constriction from XRMB database during stop closure of stop consonants was proposed. It showed a good match with the manually obtained place of constriction.

The area values estimated from the LPC analysis needs to be interpolated for smoother display. Cubic spline, Bezier, and cubic B-spline interpolation methods were investigated. Cubic spline interpolation preserves the original shape. Bezier interpolation gives smoothened shape but masks the constrictions. Cubic B-spline interpolation provides better smoothening compared to cubic spline interpolation along with preserving the original shape. Cubic B-spline interpolated area values showed distinct dark band in the areagram at the place of maximum constriction during stop closure of the VCV utterances while masking other erroneous constrictions.

Effect of scaling of area ratios by the lip area on the estimation of the place of articulation for VCV utterances has been investigated. Lip scaled area values showed varying glottis area in the VC, CV transition segments and during the stop closure of the VCV utterances. Scaling by the lip area also significantly improved the estimation of the place of articulation for the bilabial and alveolar stops. It has also been observed that the low frame rate of camera used for lip area estimation can affect the estimated vocal tract shape during the stop closure of VCV utterances.

A graphics for displaying the vocal tract shape in a speech training aid was developed. The estimated vocal tract shapes are mapped to the graphics in a mid-

sagittal view using scaling factor. A GUI for displaying the speech parameters and vocal tract shapes in a speech training aid was developed.

Vocal tract shape obtained from cubic B-spline method has to be mapped with the vocal tract graphics. Video camera having the frame rate of 100-200 frame/s is needed for lip area estimation. Lip estimation needs to be automated and interfaced with the speech training system module along with the video camera for capturing lip image and estimating lip area for lip scaling. Further, a realistic display of vocal tract shape showing 3D view of the articulatory organs is needed.

REFERENCES

[1] R. S. Nickerson and K. N. Stevens, "Teaching speech to the deaf: can a computer help?," *IEEE Trans. Audio Electroacoust*., vol. 21, no. 5, pp. 445–455, 1973.

[2] L. S. Liben, (Ed.), *Deaf Children: Development Perspectives*. New York: Academic Press, 1978.

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

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

[5] A. Watanabe, S. Tomishige, and M. Nakatake, "Speech visualization by integrating features for the hearing impaired," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 454–466, 2000.

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

[7] Speech Viewer III, (Synapse Adaptive, San Rafael, CA, 2000). [Online]. Available: http://www.synapseadaptive.com/edmark/prod/sv3. (Last accessed in May, 2012).

[8] 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).

[9] Y. Yamada, H. Javkin, and K. Youdelman, "Assistive speech technology for persons with speech impairments," *Speech Communication*, vol. 30, pp. 179–187, 2000.

[10] H. 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.

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

[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] 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.

[14] L. E. Bernstein, J. B. Ferguson, and M. H. Goldstein Jr. "Speech training devices for profoundly deaf children," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1986, pp. 633–636.

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

[16] N. D. Black, "Application of vocal tract shapes to vowel production," in *Proc. 10th Int. Conf. IEEE Engg. Med. Biol. Soc.*, 1988, pp. 1535–1536.

[17] 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.

[18] P. M. T. de Oliveira and M. N. Souza, "Speech aid for the deaf based on a representation of the vocal tract: the vowel module," in *Proc. 19th Int. Conf. IEEE Engg. Med. Biol. Soc.*, 1997, pp. 1757–1759.

[19] A. E. Mahdi, "Visualisation of the vocal-tract shape for a computer-based speech training system for the hearing-impaired," *The Open Electrical and Electronic Engineering Journal*, vol. 2, pp. 27-32, 2008.

[20] 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.

[21] K.G. Munhall, E. Vatikiotis-Bateson, and Y. Tohkura, "X-Ray Film Database for Speech Research," *J. Acoust. Soc. Am.*, vol. 98, no. 2, pp. 1222-1224, 1995.

[22] J. R. Westbury. X-ray Microbeam Speech Production Database User's Handbook (Version 1.0). June 1994. [Online]. Available: http://www.medsch.wisc.edu/ubeam/.

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

[24] 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.

[25] 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.

[26] 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.

[27] B. H. Story, "Time dependence of vocal tract modes during production of vowels and vowel sequences, " *J. Acoust. Soc. Am.*, vol. 121, no. 6, pp. 3770-3789, 2007.

[28] M. S. Shah, "Estimation of place of articulation during stop closures of vowel-consonant-vowel syllables," *Ph.D. thesis*, Dept. of Elect. Engg., IIT Bombay, India, 2008.

[29] B. N. A. Baragi, "A speech training aid for the deaf," *M.Tech. dissertation*, Dept. of Elect. Engg., IIT Bombay, India, 1996.

[30] S. A. Kshirsagar, "A speech training aid for hearing impaired," *M.Tech. dissertation*, Dept. of Elect. Engg., IIT Bombay, India, 1998.

[31] D. Salomon, *Computer Graphics and Geometric Modeling*, New York: Springer-Verlag, 1999.

[32] D. Kewley-Port, C.S. Watson, M. Elbert, K. Maki and D. Reed, “The Indiana Speech Training Aid (ISTRA) II: Training curriculum and selected case studies,” *Clinical Linguistics and Phonetics*, vol. 5, pp. 13-38, 1991.

[33] 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.

[34] D. W. Massaro and J. Light, “Using visible speech to train perception and production of speech for individuals with hearing loss,” *Journal of Speech, Language, and Hearng Research*, vol. 47, pp. 304-320, 2004.

[35] A. A. Wrench. “MOCHA multichannel articulatory database,” 2008. [Online]. Available: http://data.cstr.ed.ac.uk/mocha/README_v1.2.txt. (Last accessed in May, 2012).

[36] P. Viola and M. Jones, “Robust real-time face detection,” *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[37] R. Hsu, M. Abdel-Mottaleb, and A. K. Jain, “Face detection in color images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 696– 706, 2002.

[38] N. S. Nayak, “Estimation and display of vocal tract shape for speech training aids,” *M.Tech. dissertation*, Dept. of Elect. Engg., IIT Bombay, India, 2011.

Acknowledgements

I express my sincere gratitude towards Prof. P. C. Pandey for the guidance and support he gave me during this project. The regular discussions with him on every aspect of the research work helped me refine my approach towards the problem and motivated me to give my best. Working with him was a very pleasant learning experience and my interest in the subject has considerably grown. I am also thankful to him for sparing his invaluable time in correcting my reports.

I would like to thank Dr. M. S. Shah for his valuable suggestions and help during various stages of my project. Technical discussions with him helped me in the understanding of various issues related to this project. I would like to thank Vidyadhar Kamble for helping me in all the lab related issues, and Nataraj, Rajath, and Sudipan Saha for their helpful advices in the different issues of the project and sparing their time in correcting my reports. I would like to thank Jayan, Parveen Lehana, Nitya, Manas, and Santosh for their help and sharing interesting discussions with me. I am also thankful to all my friends for their whole-hearted support during the tenure of this project.

Jagbandhu

June 2012

Author's Resume

Jagbandhu: The author completed his under-graduation and received B.E degree from Bhilai Institute of technology, Durg in Electronics and Telecommunication in 2009. Presently, he is pursuing the M.Tech. degree in electrical engineering (specialization in electronic systems) at the Indian Institute of Technology Bombay, Mumbai, India. His research interests include speech processing, digital signal processing, and embedded system design.

Publications

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, 2012.