Estimation of Place of Articulation During Stop Closures of Vowel–Consonant–Vowel Utterances

Prem C. Pandey, *Member, IEEE*, and Milind S. Shah

***Abstract*—Production of vowel–oral stop consonant–vowel utterances involves movement of articulators from the articulatory position of the initial vowel towards that of the oral stop closure, and then to that of the final vowel. As the closure segments have zero or low signal energy, linear predictive coding (LPC)-based estimation of vocal tract shape fails during stop closure. This paper reports a technique for estimation of place of articulation during stop closures by performing bivariate polynomial modeling on vocal tract area values during transition segments preceding and following the closure. The technique with second-degree polynomial modeling was found to be suitable for estimating the place of maximum constriction during stop closure segments of vowel–consonant–vowel utterances with bilabial, alveolar, and velar stops. The estimated places compared well with the actual constriction locations observed from the articulatory data. The technique may be useful for improving effectiveness of speech-training aids for production of stop consonants by providing visual feedback of the estimated place of articulation.**

***Index Terms*—Estimation, place of articulation, polynomial approximation, speech training aids.**

I. Introduction

THE acoustic-to-articulatory mapping problem, or articulatory inversion, consists of recovering the sequence of vocal tract shapes that produce a given acoustic speech signal [1]. A solution to this problem has important applications in speech recognition, synthesis and coding, and speech-training aids for the hearing-impaired persons. The shape of the vocal tract can be specified by its cross-sectional area as a function of position along the vocal tract length. Estimation of vocal tract shape from the speech signal can be carried out using several techniques, including linear predictive coding (LPC) analysis [2], use of formants [3]–[5], and mapping via articulatory codebook [1]. Most of these techniques are reported to work satisfactorily for vowels. However, vocal tract shape estimation fails if spectral information is not available, for example during the closure segments in oral stops [1], [2], [6]. Hence, it is important to investigate a technique for the estimation of vocal tract shape during such segments.

Manuscript received July 02, 2007; revised October 31, 2008. Current version published February 11, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Hasegawa-Johnson.

P. C. Pandey is with the Department of Electrical Engineering, Indian Institute of Technology Bombay, Powai Mumbai 400076, India (e-mail: pcpandey@ee.iitb.ac.in).

M. S. Shah is with the Department of Electronics and Telecommunication Engineering, Fr. C. Rodrigues Institute of Technology, Vashi, Navi Mumbai 400703, India (e-mail: milind05in@yahoo.co.in).

Digital Object Identifier 10.1109/TASL.2008.2010285

A child with profound prelinguistic hearing impairment has great difficulty in acquiring speech because he cannot use hearing as the principal sense in speech correction. The signals received through hearing must be supplemented by other cues. Speech-training aids extract important acoustic parameters (such as speech intensity, voicing and pitch, spectral features) and articulatory parameters (such as nasality, lip movement, tongue movement) and provide visual or tactile feedback of these parameters [7]–[9]. Electromyography data of the hearing-impaired and those of the normal hearing persons are almost similar with respect to lip movement, but they differ with respect to tongue movement [8]. Thus, bilabial consonants produced by hearing-impaired persons tend to be more intelligible than lingual consonants and vowels. This emphasizes the importance of the relative visibility of articulatory gestures in determining the ease with which hearing-impaired persons learn to produce specific sounds. Speech-training aids providing visual feedback of vocal tract shape have been found to be useful for improvement in vowel articulation by the hearing-impaired [10]–[17]. Estimation and visual feedback of place of articulation of stop consonants can be used for improving the effectiveness of these aids.

Crichton and Fallside [10] developed a speech-training aid for deaf persons which displays vocal tract shape for sustained vowels. They used Wakita's linear prediction-based inverse filtering method [2] for vocal tract shape estimation, performed parabolic interpolation between discrete area values, and displayed log area values as a function of the distance along the vocal tract length. In the system reported by Pardo [11], the relative area values, obtained by Wakita's method, were subjected to constant volume normalization for reducing the dispersion in the values. Its use for speech training for the production of a set of five vowels over four months showed that error with respect to the target shapes decreased progressively with training.

Black [14] reported a system for displaying real-time tongue positions during vowel production by means of a midsagittal plot of the human head. Input speech was processed for extraction of formant frequencies every 10 ms, based on filter bank analysis and use of the peak-picking algorithm. A procedure proposed by Ladefoged *et al.*. [3], involving an empirical relationship between the formant frequencies and the degree of back and front raising of the tongue was used for the estimation of tongue shape related data. Training with the system improved accuracy and consistency in vowel production of the hearing-impaired persons. Park *et al.* [15] reported a speech-training system for displaying the vocal tract shape, intensity, fundamental frequency, nasality, and frequency spectra, in real-time. The vocal tract area function was estimated using Wakita's method [2] and the area values

were mapped to the corresponding section heights. The first two section heights were determined from the three formant frequencies. The estimated vocal tract shapes for five Korean vowels were remarkably similar to the X-ray data. Using this system, hearing-impaired children mastered the syllables /jɑ/ and /pɑ/ in 5–6 days. Rossiter *et al.* [17] reported a real-time LPC-based vocal tract display system for use as an aid for voice development. The vocal tract was modeled as concatenated tubes and the estimated area values were interpolated with a cardinal spline function to produce smooth contours.

LPC-based vocal tract shape estimation, not involving automated tracking of formants, is suitable for real-time processing, and hence despite its limitation in modeling the spectral zeros, it is used for developing speech training aids. Other indirect methods for shape estimation are based on measurement of acoustic impedance or impulse response at the lips using an impedance tube [18], [19]. Sondhi [20] has critically compared these procedures. Some of the difficulties in estimating the vocal tract transfer function from the speech signal, as pointed out by him, involve uncertainty due to the presence of glottal source characteristics, non-uniqueness of the estimated shape, and lack of high-frequency information. Estimation of the vocal tract area by measuring the impedance at the lips overcomes some of these difficulties. However, as the techniques based on acoustic measurements involve the use of an impedance tube and the speaker has to articulate silently, they cannot be used in speech-training aids. Direct imaging methods including cinefluography, X-ray microbeam, electropalatography (EPG), electromagnetic articulography (EMA), optopalatography, ultrasound imaging, and real-time magnetic resonance imaging have been used for obtaining speech production data for the vocal tract posture and movement [21]–[26]. EPG provides information on the location and timing of the tongue–palate contact. X-ray microbeam and EMA have been used for recording the movement of midsagittal points on the articulators including the tongue, lips, and points on the jaw. Real-time magnetic resonance imaging is an emerging technique for acquiring speech production data in the form of complete views of the vocal tract including the pharyngeal structures in a safe and noninvasive manner.

Qin and Carreira-Perpiñán [27] investigated non-uniqueness in the acoustic-to-articulatory mapping using articulatory data for normal speech from the Wisconsin X-ray microbeam database [21]. They analyzed simultaneously recorded articulatory and acoustic data using statistical machine learning techniques. The database was searched for all the articulatory vectors that approximately mapped to a specific acoustic vector, across utterances, and contexts. Non-uniqueness of the articulatory vectors was observed for some of the sounds (/θ/, /ɹ/, /l/, /w/), but only for about 5% of the acoustic vectors, indicating that non-uniqueness of the vocal tract is an infrequent situation. In an empirical study for finding the best acoustic features for articulatory inversion [28], they investigated the use of acoustic features (LPC, LSF, filter banks, MFCC, short-time cepstral representation, perceptual linear prediction (PLP), and RASTA-PLP) and a multilayer perceptron to map from the acoustic features to the articulatory ones. The articulatory representation was taken from multichannel articulatory database MOCHA [29] which provides speech signal along with synchronously recorded laryngograph, electropalatograph, and electromagnetic articulograph (EMA) data. Variation in the results for different features was not very large, and best results were obtained with LSF and PLP.

Our investigation for vocal tract shape estimation is based on LPC analysis of the speech signal. From the analysis carried out for various vowel–stop consonant–vowel (VCV) utterances, it was observed that estimated vocal tract area values were random and unrelated to places of articulation during stop closures. However, the area values during the transition segments preceding and following the stop closures, plotted as a function of time and position along the vocal tract length, showed different patterns of variation for different places of closure. The present paper reports investigations for vocal tract shape estimation during stop closures by performing interpolation of bivariate polynomials based on estimated area values during transition segments in VCV utterances. The technique permits automated processing of acquired utterances, and may be useful for generating slow motion visual feedback for improving the articulation of oral stops.

II. Investigation of LPC-Based Vocal Tract Shape Estimation

Vocal tract shape was estimated from reflection coefficients obtained using Wakita's method for direct estimation of the vocal tract shape by inverse filtering of the speech signal [2], [30]–[32]. In this method, the vocal tract of length l is modeled as a rigid and lossless acoustic tube with M sections of equal length and varying cross-sectional areas S_m. For plane wave propagation along the length of the tube, reflections occur at the section interfaces due to different areas on the two sides. The reflection coefficients r_m are calculated from the autocorrelation coefficients of the speech signal using LPC analysis [2], [30]–[32]. The lip end is assumed to be connected to a tube of infinite area, resulting in reflection coefficient $r_0 = 1$. The back end is assumed to be terminated with a tube of normalized cross-sectional area S_{M+1} of 1. The areas are calculated as $S_m = S_{m+1}(1+r_m)/(1-r_m)$. The section position m can be related to the physical distance from the lips as $mc/2F_s$, where $c =$ speed of sound (340 m/s) and F_s is the sampling frequency.

The speech signal for analysis was digitized with F_s of 11.025 kHz and 16-bit quantization. First difference of the signal was taken for an approximate 6-dB/octave pre-emphasis. Hamming window was applied on analysis frames with duration equal to twice the average pitch period. The inter-frame interval of 5 ms was used for tracking the variation in vocal tract shape. Autocorrelation coefficients of the windowed frame were used for computing the reflection coefficients for LPC order $M = 12$, and these coefficients were used for calculating the area values. The processing was carried out with floating-point arithmetic, in order to avoid fixed-point arithmetic related recursive errors and dynamic range limitation.

In order to study the dynamics of shape estimation during transitions, we have used the "areagram," a spectrogram-like

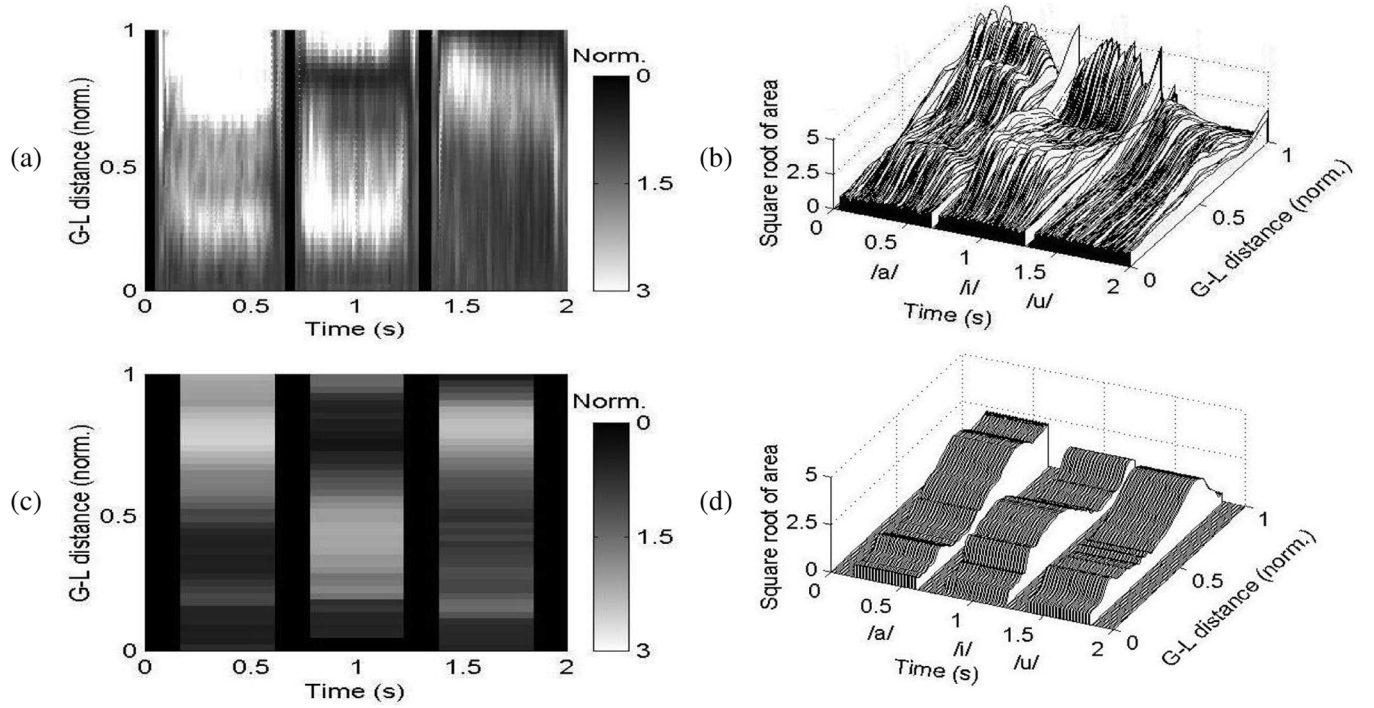

Fig. 1. Comparison of the estimated vocal tract shapes based on LPC analysis with those reported using MRI for the vowel sequence /-a-i-u-/: (a) areagram and (b) waterfall diagram from LPC analysis of the recorded sequence for a male speaker with $F_0 = 100$ Hz; (c) areagram and (d) waterfall diagram based on MRI values reported in [35].

two-dimensional display of square-root of cubic-spline interpolated vocal tract area values plotted as grey levels as a function of time along x—axis and glottis-to-lips (G-L) distance along y—axis. This display provides a visualization of the variation in vocal tract shape. For speech training, appropriate displays involving cartoons or games based on dynamically varying vocal tract shape need to be devised and tested. A software package was developed in Matlab for processing the speech signal and graphical display of the recorded and selected speech segments, pitch and energy contours, spectrogram, and areagram. The pitch (F_0) was estimated by the short-time autocorrelation method [33], with energy as zeroth autocorrelation coefficient. The package was tested by analyzing sustained vowel and VCV utterances to check consistency and validity of the estimated shapes. The vowels and VCV utterances, from five speakers (three male and two female), were recorded using a PC sound card and an electret microphone in an acoustically treated room. A Klatt synthesizer [34] was used for synthesizing vowels with varying pitch for studying the effect of pitch on shape estimation.

Fig. 1 compares the estimated vocal tract shapes for the vowel sequence /-a-i-u-/ with those reported using MRI. Shapes are displayed as areagram as well as waterfall diagram. Parts (a) and (b) show the areagram and waterfall diagram based on LPC analysis of the recorded vowel sequence for a male speaker with F_0 of approximately 100 Hz. For every analysis window, 12 discrete area values along glottis-to-lips distance, estimated by LPC analysis, are converted to 40 values using cubic-spline interpolation and then the square-root of these values are displayed in the waterfall diagram and areagram. Story *et al.* [35] have reported MRI-based vocal tract shapes as area values for 40 discrete segments. The square-root of these area values for the vowels /a/, /i/, and /u/ was used for generating the areagram and waterfall diagram shown in parts (c) and (d), respectively. It is observed that the vocal tract shapes and place of tongue elevation estimated using LPC analysis are reasonably similar to those based on MRI, for each of the three vowels.

Sustained vowel utterances were analyzed to identify the minimum signal energy above which estimation of the vocal tract shape remains consistent. The waveforms with 16–bit quantization were amplitude scaled, and the lowest scaling factor that resulted in qualitatively reasonable vocal tract shapes was noted. Analysis for various vowel waveforms showed that vocal tract shape estimation works properly for 40 dB of the available dynamic range [36]. Analysis of vowels synthesized with step and ramp variation in pitch showed that the estimation of the vocal tract shape did not get affected by pitch variation.

In speech training aids for improving the articulation of consonant sounds, vowel–consonant–vowel utterances can be used, as the resulting short duration of vocal tract shape display will be easier for the hearing-impaired person to monitor and mimic. For observing the shape tracking ability of the algorithm, vowel–semivowel–vowel utterances, involving semivowels /j/ and /w/, were analyzed. Areagram results [37] showed qualitatively reasonable transition in vocal tract shapes and places of articulation for all the speakers. Next, investigations were carried out into vocal tract shape estimation for VCV utterances involving oral stops, which have low signal energy during their closure portion. Fig. 2 shows analysis results for VCV utterance /aka/. The areagram and waterfall diagram show reasonable place of articulation and tongue height for the vowel segments, but area estimates are random and unrelated

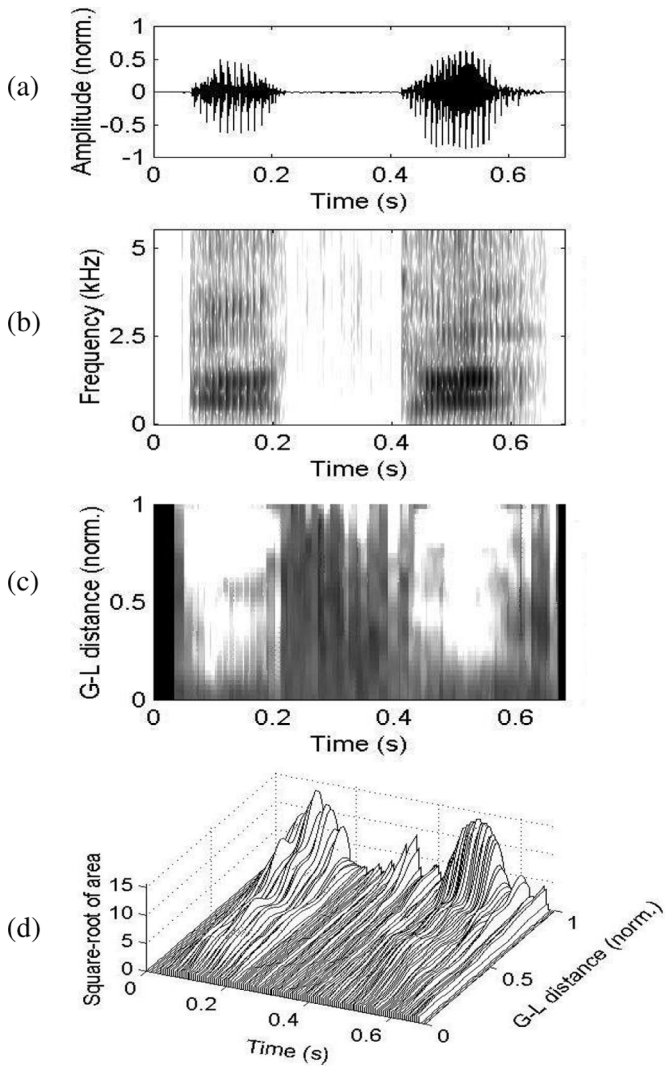

Fig. 2. Analysis results for /aka/ from a male speaker: (a) waveform for 0.7 s; (b) wideband spectrogram; (c) areagram; (d) waterfall diagram.

to the place of maximum constriction during stop closure, as no spectral information related to the articulation is available.

III. Estimation of Vocal Tract Shape During Stop Closures

In a VCV utterance involving an oral stop, the dynamic movement of articulators during VC and CV transitions is related to the articulatory positions for the vowels preceding and following the stop and the place of the stop closure. This movement is acoustically characterized by formant transitions, and the loci of the second and third formants are related to the place of closure [31], [38]. However, there are several difficulties in tracking the formants during short transition segments and in estimating the vocal tract shape from the formants. An inspection of areagrams for VCV utterances with different oral stops showed that the area values during transition segments on either side of stop closures had different two-dimensional patterns for different places of closure. Estimation of area values during the closure may be possible by surface modeling of the area values

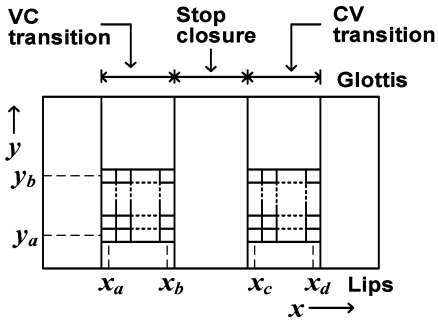

Fig. 3. Selection of area values during transition segments for interpolation.

during the preceding and the following transition segments. As an empirical approach to surface modeling, least-squares based bivariate polynomial approximation [36], [37], [39]–[44] on the area values is investigated. The polynomial function is used for interpolating the area values during the closure and for estimating the place of maximum constriction.

In Sections III-A–III-C, we describe the method of bivariate polynomial modeling and interpolation, estimation of boundary locations of the stop closure, and investigations for validation of the proposed technique.

A. Bivariate Polynomial Modeling and Interpolation

The estimated vocal tract area values $g(x, y)$ during VC and CV transition segments can be modeled by a bivariate polynomial $f(x, y)$, with an approximation error $r(x, y)$. Here "x" represents analysis frame number along time axis and "y" represents the section number from the lip end in the M-section acoustic tube model. Fig. 3 shows the transition segments along with a possible way for selecting area values for the surface modeling. Value x_a along x-axis corresponds to the starting position of the transition segment along x-direction, x_b and x_c mark the segment over which area values cannot be estimated, and x_d marks the end of the transition. Thus, the estimated area values which are used for surface approximation correspond to $x_a \leq x \leq x_b$, $x_c \leq x \leq x_d$, and $y_a \leq y \leq y_b$. The number of frames to the left and right of stop closure used for surface modeling are $L_{\text{col}} = x_b - x_a + 1$ and $R_{\text{col}} = x_d - x_c + 1$, respectively. The number of sections used is $j = y_b - y_a + 1$.

A second-degree bivariate polynomial is given by

$$f(x, y) = c_0 + c_1 x + c_2 y + c_3 xy + c_4 x^2 + c_5 y^2 \quad (1)$$

and the coefficients $c_0 - c_5$ can be obtained to minimize the approximation error. Similarly, the coefficients can be obtained for the third degree bivariate polynomial

$$f(x, y) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 y + c_5 y^2 + c_6 y^3 + c_7 xy + c_8 x^2 y + c_9 xy^2. \quad (2)$$

In order to evaluate the coefficients, we need to have $j \geq 3$ and $j \geq 4$ for second- and third-degree polynomials, respectively.

In matrix notation, the bivariate polynomial approximation may be written as

$$\mathbf{Az} = \mathbf{b} + \mathbf{r} \quad (3)$$

where $\mathbf{b}^T$ and $\mathbf{r}^T$ are shown by the first equation at the bottom of the page. For the second degree polynomial approximation, $\mathbf{A}$ and $\mathbf{z}^T$ are given by

$$\mathbf{A} = \begin{bmatrix} 1 & x_a & y_a & x_a y_a & x_a^2 & y_a^2 \\ 1 & x_a & y_{a+1} & x_a y_{a+1} & x_a^2 & y_{a+1}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_a & y_b & x_a y_b & x_a^2 & y_b^2 \\ 1 & x_{a+1} & y_a & x_{a+1} y_a & x_{a+1}^2 & y_a^2 \\ 1 & x_{a+1} & y_{a+1} & x_{a+1} y_{a+1} & x_{a+1}^2 & y_{a+1}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{a+1} & y_b & x_{a+1} y_b & x_{a+1}^2 & y_b^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_d & y_b & x_d y_b & x_d^2 & y_b^2 \end{bmatrix}$$

$$\mathbf{z}^T = [c_0 \quad c_1 \quad \cdots \quad c_5].$$

For the third degree polynomial approximation, $\mathbf{A}$ and $\mathbf{z}^T$ are given by the second equation shown at the bottom of the page. As the number of coefficients p in the bivariate polynomial is smaller than the number of area values q, (3) corresponds to an overdetermined set of simultaneous linear equations. The coefficients can be obtained for minimizing the sum of the squared errors. A necessary condition for minimization of the error

$$E(c_0, \ldots, c_{p-1}) = \sum_{n=0}^{q-1} r_n^2$$

is that

$$\frac{\partial E}{\partial c_i} = 0, \quad i = 0, 1, \ldots, p-1 \tag{4}$$

which leads to the polynomial coefficient vector $\mathbf{z}$ given as

$$\mathbf{z} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}. \tag{5}$$

The matrix $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is known as the pseudo-inverse of $\mathbf{A}$ [43]–[45]. If the degree of the polynomial is large, the resulting matrix $\mathbf{A}$ is typically ill-conditioned [46]. Hence, the investigations were restricted to the use of second- and third-degree bivariate polynomials.

With $\mathbf{z}$ as obtained from (5), the interpolation for area values in the closure portion $(x_b < x < x_c)$ is performed by

$$\hat{\mathbf{b}} = \hat{\mathbf{A}}\mathbf{z} \tag{6}$$

where $\hat{\mathbf{b}}^T$ and $\hat{\mathbf{A}}$ for the second- and third-degree polynomial approximation are as shown by the equations at the bottom of the next page. The interpolation for $1 \le y_a \le M - j$ is carried out using the coefficients of the bivariate polynomial obtained for modeling the area values with $y_a \le y < y_{a+j}$. The coefficients obtained for modeling the area values with $M - j < y \le M$ are used for interpolation for $M - j < y_a \le M$.

As mentioned earlier, analysis of the speech signal with $F_s = 11.025$ kHz was carried out with $M = 12$. For each frame, cubic-spline interpolation was applied on the 12 area values to get a 40-point smooth vocal tract area function and the square root of the values was used for displaying the vocal tract shape.

$$\mathbf{b}^T = [g(x_a, y_a) \quad g(x_a, y_{a+1}) \quad \ldots \quad g(x_a, y_b) \quad g(x_{a+1}, y_a) \quad g(x_{a+1}, y_{a+1}) \quad \ldots \quad g(x_{a+1}, y_b) \quad \ldots \quad g(x_d, y_b)]$$

and

$$\mathbf{r}^T = [r(x_a, y_a) \quad r(x_a, y_{a+1}) \quad \ldots \quad r(x_a, y_b) \quad r(x_{a+1}, y_a) \quad r(x_{a+1}, y_{a+1}) \quad \ldots \quad r(x_{a+1}, y_b) \quad \ldots \quad r(x_d, y_b)]$$

$$\mathbf{A} = \begin{bmatrix} 1 & x_a & x_a^2 & x_a^3 & y_a & y_a^2 & y_a^3 & x_a y_a & x_a^2 y_a & x_a y_a^2 \\ 1 & x_a & x_a^2 & x_a^3 & y_{a+1} & y_{a+1}^2 & y_{a+1}^3 & x_a y_{a+1} & x_a^2 y_{a+1} & x_a y_{a+1}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_a & x_a^2 & x_a^3 & y_b & y_b^2 & y_b^3 & x_a y_b & x_a^2 y_b & x_a y_b^2 \\ 1 & x_{a+1} & x_{a+1}^2 & x_{a+1}^3 & y_a & y_a^2 & y_a^3 & x_{a+1} y_a & x_{a+1}^2 y_a & x_{a+1} y_a^2 \\ 1 & x_{a+1} & x_{a+1}^2 & x_{a+1}^3 & y_{a+1} & y_{a+1}^2 & y_{a+1}^3 & x_{a+1} y_{a+1} & x_{a+1}^2 y_{a+1} & x_{a+1} y_{a+1}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{a+1} & x_{a+1}^2 & x_{a+1}^3 & y_b & y_b^2 & y_b^3 & x_{a+1} y_b & x_{a+1}^2 y_b & x_{a+1} y_b^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_d & x_d^2 & x_d^3 & y_b & y_b^2 & y_b^3 & x_d y_b & x_d^2 y_b & x_d y_b^2 \end{bmatrix}$$

$$\mathbf{z}^T = [c_0 \quad c_1 \quad \cdots \quad c_9]$$

B. Estimation of Stop Closure Boundary Locations

For bivariate polynomial generation, the stop closure boundary locations (i.e., x_b and x_c in Fig. 3) need to be known. These locations were estimated using a two-step process: estimation of the beginning and ending points of the VCV utterance followed by the estimation of stop closure boundary locations within the utterance. The end-points of the VCV utterance were estimated based on short-time average magnitude of the signal, as reported in [47]. A measure of zero-crossing rate is important if the end-points are to be estimated for the utterances which may involve a consonant at the initial or final position. For the VCV utterances, estimation of end-points could be carried out using only the short-time average magnitude.

For estimating the stop closure boundary locations, the root-mean-square (rms) value σ_s of the signal waveform between the two end-points of the VCV utterance was computed, and a threshold value of $0.2\sigma_s$ was selected empirically. The short-time average magnitude of successive frames was compared with this threshold, from left-to-right (i.e., initial vowel to closure) and right-to-left (i.e., end vowel to closure), for marking the closure boundary associated with VC and CV transitions, respectively. The region within the marked frames, showing average magnitude consistently below the threshold, was defined as the stop closure region. The end-point location of the stop closure, estimated by using short-time average magnitude, was delayed to exclude the closure release burst.

C. Validation of the Proposed Technique

As a first step in the investigation, the technique was applied for estimating the place of articulation during artificially silenced transition segments of different durations in vowel–semivowel–vowel utterances. For this purpose, the sample values in the central portion of the semivowel part of the utterance were set to zero. The analysis was carried out on /aja/ and /awa/ utterances from the three male and two female speakers, with the objective of comparing the results from the two types of polynomial approximations, finding the appropriate parameter values for surface modeling and interpolation, and finding the minimum duration of vowel-to-semivowel and semivowel-to-vowel transition segments required for the recovery of the place of articulation [36], [37]. It was observed that modeling based on the second-degree polynomial for all the five speakers could successfully estimate the place of articulation. The mean value of L_{col} and R_{col} (i.e., number of frames to the left and the right of the silence interval) required for the surface modeling of area values for /aja/ and /awa/ was 5.2 and 5.6, respectively. For estimation of reasonably consistent place of articulation, the mean value of the minimum required duration of the transition segments to be present on either side of the artificially introduced silence gap for the utterances /aja/ and /awa/ was about 30 ms and 28 ms, respectively. For recovering a reasonably consistent vocal tract shape during the silenced interval, the technique needed more than 90 ms of transition segment on both sides. Most VCV utterances have short transition segments and hence the technique can be used for estimating the place of maximum constriction during the closure and not the actual vocal tract shape.

After establishing the minimum frames needed, the technique was applied for estimating the place of stop closure in the VCV utterances of the type /aCa/, /iCa/, /aCi/, /iCi/, and /uCu/ involving consonants /p/, /b/, /t/, /d/, /k/, and /g/. These utterances were recorded for English stops with the three places of articulation: bilabial (/p/ and /b/), alveolar (/t/ and /d/), and velar (/k/ and /g/). Vowel-stop consonant-vowel utterances of the type /aCa/ and /iCa/ were recorded and analyzed for three male and two female speakers, while utterances of the type /aCi/, /iCi/, and /uCu/ were recorded and analyzed for one male speaker. The estimated places were compared with those reported earlier using MRI and X-ray images. The technique was also applied for utterances with dental and retroflex stops, from a male speaker of Marathi (a language spoken in the western part of India).

For a direct validation, the utterances of the type $/\text{ʌCa}/$, involving voiced stop consonants /b/, /d/, and /g/, spoken by 20 male and 20 female speakers, and acquired along with the articulatory data using XRMB technique at the University of Wisconsin [21] were analyzed for estimation of the places of maximum constriction. The speech signals in the XRMB database, originally acquired at 21.379 kHz, were down-sampled to 11.025 kHz. For each selected time location in an utterance, the articulatory plot in the database shows the back pharyngeal wall, the palate, markers for upper and lower lips, and the four pellet markers on the tongue joined by a smooth curve. The position of pellets gives a point-parameterized representation of lingual,

$$\hat{\mathbf{b}}^T = [\, g(x_{b+1}, y) \quad g(x_{b+2}, y) \quad \ldots \quad g(x_{c-1}, y) \,]$$

$$\hat{\mathbf{A}} = \begin{bmatrix} 1 & x_{b+1} & y_a & x_{b+1}y_a & x_{b+1}^2 & y_a^2 \\ 1 & x_{b+2} & y_a & x_{b+2}y_a & x_{b+2}^2 & y_a^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{c-1} & y_a & x_{c-1}y_a & x_{c-1}^2 & y_a^2 \end{bmatrix}$$

and

$$\hat{\mathbf{A}} = \begin{bmatrix} 1 & x_{b+1} & x_{b+1}^2 & x_{b+1}^3 & y_a & y_a^2 & y_a^3 & x_{b+1}y_a & x_{b+1}^2y_a & x_{b+1}y_a^2 \\ 1 & x_{b+2} & x_{b+2}^2 & x_{b+2}^3 & y_a & y_a^2 & y_a^3 & x_{b+2}y_a & x_{b+2}^2y_a & x_{b+2}y_a^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{c-1} & x_{c-1}^2 & x_{c-1}^3 & y_a & y_a^2 & y_a^3 & x_{c-1}y_a & x_{c-1}^2y_a & x_{c-1}y_a^2 \end{bmatrix}$$

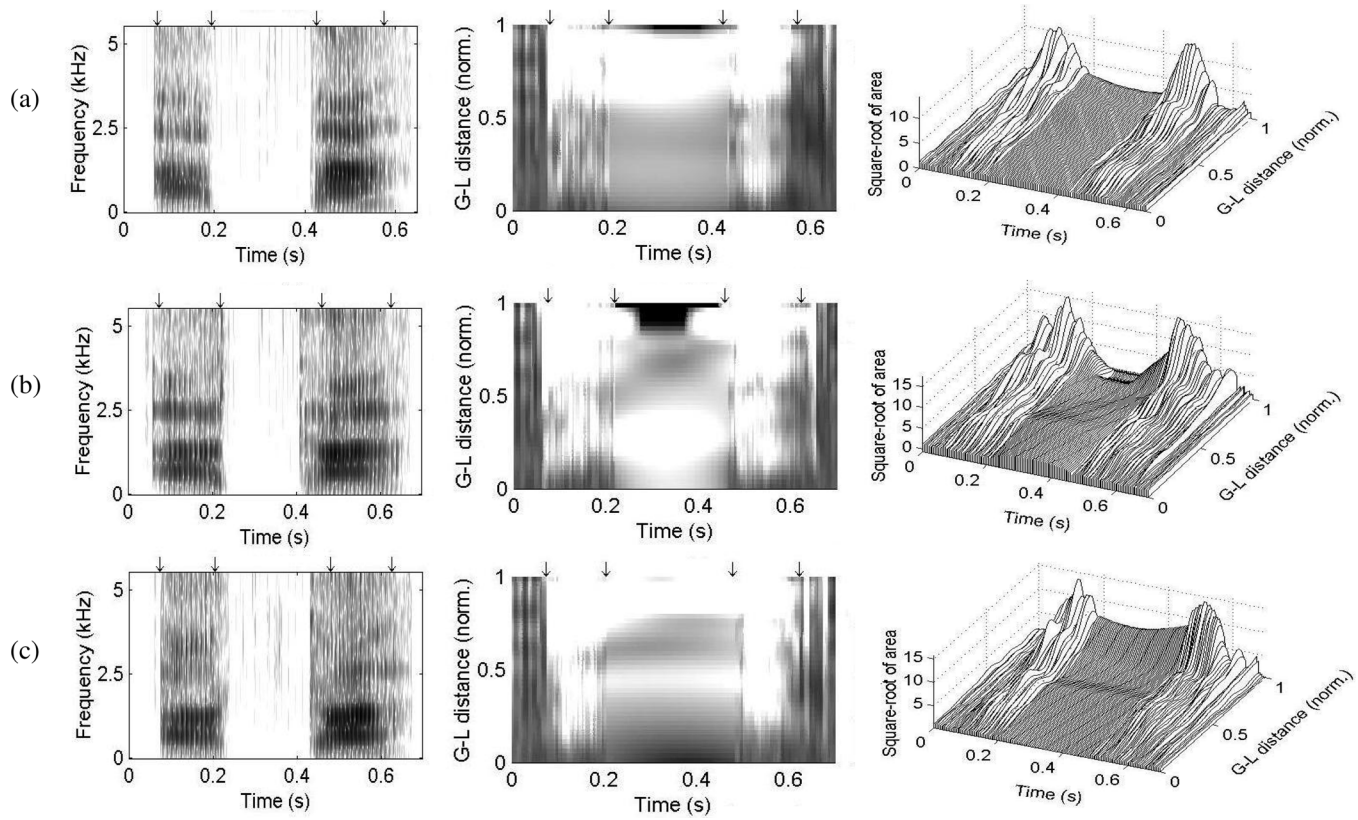

Fig. 4. Interpolation results obtained with the second degree bivariate polynomial interpolation of area values for the VCV utterances with unvoiced bilabial, alveolar, and velar stops: (a) /apa/ (surface generation parameters: $j = 5$, $L_{\rm col} = 3$, $R_{\rm col} = 3$); (b) /ata/ ($j = 4$, $L_{\rm col} = 5$, $R_{\rm col} = 4$); and (c) /aka/ ($j = 6$, $L_{\rm col} = 4$, $R_{\rm col} = 4$). Left side: wideband spectrograms, middle: areagrams, right side: waterfall diagrams.

labial, and mandibular movements. The plot also displays the x and y positions of the pellets in millimeters. The position of maximum constriction from the lips along the length of the vocal tract was calculated in millimeters by adding the straight line distances along the curve joining the lower lip marker and the pellet markers.

IV. Results and Discussion

We first present and discuss the results obtained by application of the technique for estimation of the place of articulation on the VCV utterances with voiced and unvoiced stops. This is followed by the results for the utterances with voiced stops from the XRMB database.

A. VCV Utterances With Voiced and Unvoiced Stops

From the data based on MRI [35] and X-ray images [48], typical values for the place of articulation, on the normalized distance of 0 to 1 (0 corresponds to glottis position and 1 corresponds to lips position), for the velar, alveolar, and bilabial oral stops are 0.47 to 0.70, 0.75 to 0.89 and 1.0, respectively. Estimated places of maximum constriction for the stop consonants in all the recorded utterances were compared with these earlier reported places of articulation.

Interpolation results for the English VCV utterance involving unvoiced stops /p/, /t/, and /k/ spoken by an adult male speaker, based on the second degree bivariate polynomial modeling of area values are shown in Fig. 4. Parts (a), (b), and (c) show wideband spectrograms (on the left side), areagrams (in the middle), and waterfall diagrams (on the right side) for /apa/, /ata/, and /aka/, respectively. The two outer arrows on the spectrogram indicate the detected end points and the two inner arrows indicate the boundaries of the stop closure. The figure caption gives the number of frames ($L_{\rm col}$ and $R_{\rm col}$) and number of rows (j) required for surface modeling and interpolation for estimation of place of articulation. From the analysis results for /apa/ in part (a) of Fig. 4, it is observed that second degree surface based interpolation shows lower area values around the normalized distance of 1.0, thus estimating the bilabial place of articulation. Results in part (b) for /ata/ show lower area values near the lips as well as at the normalized distance of 0.8, which corresponds to the expected place of articulation for alveolar /t/. The interpolation results in part (c) for /aka/ show lower area values at a normalized distance of 0.6, the velar region. The analysis results showed that the estimated place of articulation for all the utterances, involving unvoiced as well as voiced oral stops, from all the speakers matched with place of articulation reported from MRI and X-ray data.

From the analysis of results for English utterances of the type /aCa/, it was observed that the results with the third-degree surface modeling were qualitatively better in place estimation of velar stops than that of bilabial or alveolar stops. Compared to the results obtained from third-degree surface modeling, those

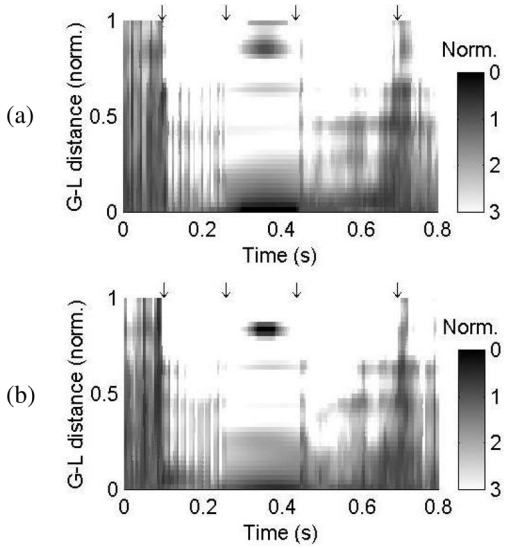

Fig. 5. Areagrams obtained with the second degree bivariate polynomial interpolation of area values for the VCV utterances with unvoiced dental and retroflex stops: (a) /a t̪ a/ ($j = 3, L_{\rm col} = 7, R_{\rm col} = 7$); (b) /a ʈ a/ ($j = 3, L_{\rm col} = 7, R_{\rm col} = 7$).

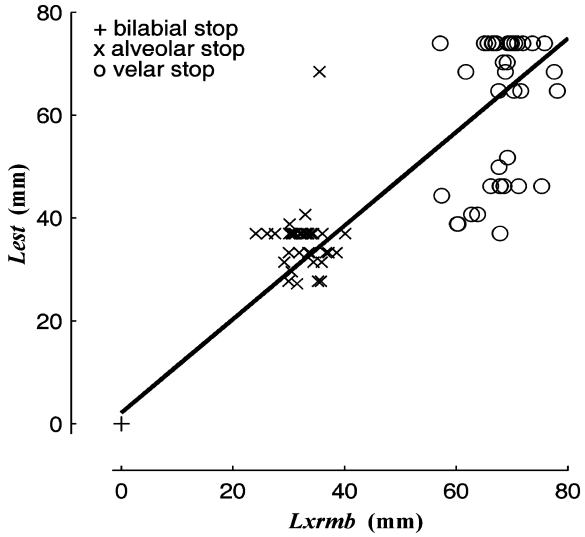

Fig. 6. Scatter plot of estimated places of articulation ($L_{\rm est}$) versus the actual ones from the XRMB database ($L_{\rm xrmb}$) for the 120 /ʌCa/ utterances involving stop consonants /b/, /d/, and /g/. Linear regression: $L_{\rm est} = 2.179 + 0.909 L_{\rm xrmb}$.

from second-degree surface modeling gave a better approximation of the place for all the stops. This indicates that the second-degree polynomial is better suited for modeling of articulatory movements. This is in conformity with observations during the application of the technique on vowel–semivowel–vowel utterances with artificially introduced silence gaps.

Interpolation results, based on second- and third-degree surfaces, for the VCV utterances of the type /uCu/ showed consistent estimation of all the three places of articulation. It was observed that estimation of velar place of articulation for utterances of the type /iCa/, /aCi/, and /iCi/ were less consistent across speakers. Thus, the proposed technique is less effective for articulatory movement involving transition of place of articulation from front (as for vowel /i/) to back (for velar /k/ and /g/).

Most languages in India have dental and retroflex stops, distinct from the alveolar stops of English [38]. Utterances /a t̪ a/ and /a ʈ a/, involving dental and retroflex stops respectively, from a male speaker of Marathi, were analyzed. The areagrams for the two utterances, with area values during the closure obtained by second degree surface interpolation, are shown in Fig. 5. The estimated places are in the correct order with respect to the bilabial, alveolar, and velar places.

B. VCV Utterances From the XRMB Database

A total of 120 vowel-consonant-vowel utterances of the type /ʌCa/ involving voiced stop consonants /b/, /d/, and /g/, spoken by 20 male and 20 female speakers, from the XRMB database [21] were analyzed for the estimation of the places of maximum constriction. The estimated places were compared with those obtained from the x-y articulatory plots.

Fig. 6 gives a scatter plot of estimated places versus the actual ones for utterances from all the 40 speakers. All the utterances involving the bilabial stop get indicated by a single point. For the alveolar stop, the spread in the estimated values is comparable to the spread in the actual values. The spread in the estimated values is larger for the velar stop, forming two clusters (each having about similar number of male and female speakers). The root-mean-square error from the best fit line obtained by linear regression (as shown in Fig. 6) was 9.4 mm. The values of the correlation coefficient between the estimated and actual places for the male speakers, the female speakers, and all the speakers were 0.95, 0.93, and 0.94 respectively, and the difference between the values for male and female speakers was not statistically significant. The means and standard deviations, for the utterances from the 40 speakers for the three stops, of the actual and estimated places of maximum constriction are given in Table I. This table also gives the root-mean-square error in the estimated values. It is observed that for all three stops, the errors in estimation were small and comparable to the standard deviation in the actual values. These results show a good correspondence between the estimated and actual places of maximum constriction.

TABLE I
COMPARISON OF ESTIMATED PLACE OF MAXIMUM CONSTRICTION (DISTANCE FROM THE LIPS) WITH THOSE OBTAINED FROM XRMB DATA: MEAN VALUES (S.D. IN PARENTHESIS) FOR UTTERANCES FROM 40 SPEAKERS

Utterance	Place of constriction from XRMB data, in mm	Estimated place of constriction, in mm	RMS value of estimation error, in mm
/ʌba/	0.0 (0.0)	0.0 (0.0)	0.0
/ʌda/	32.5 (3.2)	35.7 (6.3)	7.6
/ʌga/	68.6 (5.5)	62.9 (15.0)	15.0

V. Conclusion

It may be concluded that for utterances of the type /aCa/ and /ʌCa/, bivariate second degree polynomial modeling of the area values during the transition segments preceding and following the stop closure, and interpolation of the area values during the stop closure can be used for estimating the place of articulation. The technique may be useful for improving the effectiveness of speech-training aids for the production of oral stops by providing visual feedback of the vocal tract shape and specifically the place of maximum constriction during stop closures. Use of bivariate polynomial modeling and interpolation may be investigated with some of the other methods for vocal tract shape estimation.

Acknowledgment

The authors would like to thank Prof. S. D. Agashe and Prof. P. Rao of IIT Bombay, Dr. V. K. Madan of BARC, Mumbai, Prof. S. Umesh of IIT Kanpur, and Prof. C. Giguère of the University of Ottawa, for the input and suggestions received from them. They would also like to thank Dr. J. R. Westbury of the University of Wisconsin-Madison for making available the XRMB speech production database.

References

[1] 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, pt. 2, pp. 133–150, Apr. 1994.

[2] 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, Oct. 1973.

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

[4] Z. Yu and P. C. Ching, "Determination of vocal-tract shapes from formant frequencies based on perturbation theory and interpolation method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 369–372.

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

[6] M. S. Shah and P. C. Pandey, "Estimation of vocal tract shape for VCV syllables for a speech training aid," in *Proc. 27th Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2005, pp. 6642–6645.

[7] J. F. Curtis, *Processes and Disorders of Human Communication*. New York: Harper and Row, 1978.

[8] R. S. Nickerson, "Characteristics of the speech of deaf persons," *Volta Rev*, vol. 77, pp. 342–362, 1975.

[9] , H. Levitt, J. M. Pickett, and R. A. Houde, Eds., *Sensory Aids for the Hearing Impaired*. New York: IEEE Press, 1980.

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

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

[12] S. Aguilera, A. Borrajo, J. M. Pardo, and E. Munoz, "Speech-analysis-based devices for diagnosis and education of speech and hearing impaired people," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1986, pp. 641–644.

[13] M. Shigenaga and H. Kubo, "Speech training system for handicapped children using vocal tract lateral shapes," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1986, pp. 637–640.

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

[15] 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.

[16] 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 Eng. in Med. and Biol. Soc.*, 1997, pp. 1757–1759.

[17] D. Rossiter, D. M. Howard, and M. Downes, "A real-time LPC-based vocal tract area display for voice development," *J. Voice*, vol. 8, no. 4, pp. 314–319, 1994.

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

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

[20] M. M. Sondhi, "Estimation of vocal-tract areas: the need for acoustical measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 3, pp. 268–273, Jun. 1979.

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

[22] C. Medizinelektronik, "3D recording of speech-movement inside the mouth," 2008 [Online]. Available: http://www.articulograph.de

[23] 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.

[24] A. Rouco and D. Recasens, "Reliability of electromagnetic midsagittal articulometry and electropalatography data acquired simultaneously," *J. Acoust. Soc. Amer.*, vol. 100, no. 5, pp. 3384–3389, 1996.

[25] A. A. Wrench, A. D. McIntosh, and W. J. Hardcastle, "Optopalatography: Development of a device for measuring tongue movement in 3D," in *Proc. Int. Conf. Eurospeech*, 1997, pp. 1055–1058.

[26] D. Byrd, C. P. Browman, L. Goldstein, and D. N. Honorof, "Magnetometer and X-ray microbeam comparison," in *Proc. 14th Int. Congr. Phonetic Sci.*, J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, and A. C. Bailey, Eds. New York: American Institute of Physics, 1999, pp. 627–630.

[27] C. Qin and M. Á. Carreira-Perpinñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping," in *Proc. Interspeech*, 2007, pp. 74–77.

[28] C. Qin and M. Á. Carreira-Perpinñán, "A comparison of acoustic features for articulatory inversion," in *Proc. Interspeech*, 2007, pp. 2469–2472.

[29] A. A. Wrench, "MOCHA multichannel articulatory database," 2008 [Online]. Available: http://data.cstr.ed.ac.uk/mocha/README_v1.2.txt.

[30] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[31] D. O'Shaughnessy, *Speech Communications: Human and Machines*, 2nd ed. Piscataway, NJ: IEEE Press, 2000.

[32] A. M. Kondoz, *Digital Speech*, 2nd ed. Chichester, West Sussex, U.K.: Wiley, 2004.

[33] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 1, pp. 2–8, Jan. 1976.

[34] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoustic Soc. Amer.*, vol. 67, pt. 3, pp. 971–995, 1980.

[35] 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.

[36] 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.

[37] 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.

[38] P. Ladefoged, *A Course in Phonetics*, 2nd ed. New York: Harcourt Brace Jovanovich, 1982.

[39] G. M. Philips, *Interpolation and Approximation by Polynomials*. New York: Springer-Verlag, 2003.

[40] V. Pratt, "Direct least-squares fitting of algebraic surfaces," *Comput. Graphics*, vol. 21, no. 4, pp. 145–152, 1987.

[41] M. S. Shah and P. C. Pandey, "Surface modeling of vocal tract shapes in transition segments of vowel-consonant-vowel syllables for estimation of place of closure," in *Proc. 19th Int. Congr. Acoust.*, 2007, Paper CAS-03-011..

[42] M. S. Shah and P. C. Pandey, "Estimation of place of constriction during stop closures by bivariate surface modeling," in *Proc. Int. Symp. Frontiers Res. Speech and Music*, Kolkata, India, 2008, pp. 161–166.

[43] J. M. D. Pereira, P. M. B. S. Girão, and O. Postolache, "Fitting transducer characteristics to measured data," *IEEE Instrum. Meas. Mag.*, vol. 4, no. 4, pp. 26–39, Dec. 2001.
[44] R. L. Branham, Jr, *Scientific Data Analysis: An Introduction to Overdetermined Systems*. New York: Springer-Verlag, 1990.
[45] H. W. Brinkmann and E. A. Klotz, *Linear Algebra and Analytic Geometry. Reading*. Boston, MA: Addison-Wesley, 1971.
[46] S. D. Stearns and R. A. David, *Signal Processing Algorithms*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
[47] 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.
[48] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*, 2nd ed. New York: Springer-Verlag, 1975.

Prem C. Pandey (M'88) received the B.Tech. degree in electronics engineering from Banaras Hindu University, Varanasi, India, in 1979, the M.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1981, and the Ph.D. degree in biomedical engineering from the University of Toronto, Toronto, ON, Canada, in 1987.

In 1987, he joined the University of Wyoming, Laramie, as an Assistant Professor in electrical engineering and later joined the Indian Institute of Technology Bombay, Powai Mumbai, India, in 1989, where he is a Professor in electrical engineering, with a concurrent association with the biomedical engineering program. His research interests include speech and signal processing, biomedical signal processing, and electronic instrumentation.

Milind S. Shah received the B.E. degree in electronics engineering from the University of Mumbai, Mumbai, India, in 1990, the M.Tech. degree in electronics engineering from the University of Nagpur, Nagpur, India, in 1993, and the Ph.D. degree in electrical engineering from the Indian Institute of Technology Bombay, Powai Mumbai, India, in 2008.

He is a Professor in electronics and telecommunication engineering at the Fr. C. Rodrigues Institute of Technology, Navi Mumbai, affiliated to the University of Mumbai. His research interests are in speech and signal processing.