Estimation of Place of Articulation in Stop Consonants for Visual Feedback

Milind S. Shah and Prem C. Pandey

Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India

{milind, pcpandey}@ee.iitb.ac.in

Abstract

Speech-training systems providing visual feedback of vocal tract shape are found to be useful for improving vowel articulation. Estimation of vocal tract shape, based on LPC and other analysis techniques, generally fails during stop closures, due to very low signal energy and unavailability of spectral information. Based on estimated area values and line spectrum pair (LSP) coefficients before and after stop closure in vowel-consonant-vowel (VCV) syllables, least-squares bivariate conic and cubic polynomial surfaces were generated and used for shape estimation during stop closure by performing 2D interpolation. Implementation of the technique with automated processing of /aCa/ syllables successfully estimated the place of articulation during closure segments of velar, alveolar, and bilabial stops. The technique can be used to provide visual feedback to improve production of stop consonants.

Index Terms: estimation of place of constriction, vocal tract shape estimation, speech training aids

1. Introduction

Normal-hearing children use auditory cues in acquiring speech. A prelingually deafened child, having hearing loss above 80 dB in the frequency region of speech, 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 (e.g., speech intensity, fundamental frequency, spectral features, etc.) and articulatory (e.g., voicing, nasality, lip movement, tongue movement, etc.) speech parameters and provide a sensory feedback of these parameters with visual or tactile feedback [1]-[3].

Electromyography data show that articulatory behaviors of speakers with hearing impairment and those with normal hearing are almost similar in case of lip movement, but they differ in case of tongue movement [2]. Labial consonants produced by hearing impaired persons tend to be more intelligible than lingual consonants and vowels. This emphasizes the importance of relative visibility of articulatory gestures in determining ease with which hearing impaired persons learn to produce specific sounds. Visual feedback of vocal tract shape can serve an important role in developing correct articulatory efforts. Speech-training systems providing visual feedback of vocal tract shape have been found to be useful for improvement in vowel articulation by the hearing impaired [4]-[7]. For improving effectiveness of these systems, it is important to investigate techniques for estimation of vocal tract shape, and particularly the place of constriction, during stop closure.

Various indirect techniques like measurement of acoustic impedance at the lips, measurement of formant frequencies, and LPC based analysis have been reported for estimation of vocal tract shape [8]-[11]. LPC based vocal tract shape estimation, not involving automated tracking of formants, is suitable for real-time processing, and hence despite its several limitations, it can be used for developing speech training aids. Also, LPC coefficients can be transformed into other parameter sets, which can be useful in investigating estimation of vocal tract shape.

From the investigations carried out for LPC based vocal tract shape estimation for various vowel-consonant-vowel (VCV) syllables, it was observed that area values were random and unrelated to place of articulation during stop closure. However, area values during vowel-consonant (VC) and consonant-vowel (CV) transition segments were distinctly different, and could be related to transition in vocal tract shapes, indicating that information for estimating the place of closure may be contained in a few frames preceding and following the stop closure. This paper presents investigations for shape estimation during stop closures by performing 2D interpolation of bivariate polynomial surfaces based on estimated area values and vocal tract transfer function coefficients during transition segments, preceding and following the stop closure, in VCV syllables. The technique implementation permits automated processing of acquired syllables, and can be applied for generating slow motion visual feedback for improving the articulation of stop consonants.

2. LPC based vocal tract shape estimation

Vocal tract shape was estimated from reflection coefficients obtained using LPC analysis of speech signal, using Wakita's speech analysis model and Robinson's algorithm for optimum inverse filtering [10], [12]. From the initial investigations on vowels, it was noted that for proper and consistent shape estimation, LPC based algorithm should operate with analysis window size equal to twice the average pitch period and LPC order 12 with sampling rate around 11 kHz.

In order to study the consistency of the shape estimation with amplitude and pitch variation in vowels, and to study dynamics of shape estimation during transitions, we have used "areagram" display, a spectrogram like 2D display of square-root of cubic-spline interpolated vocal tract area values with time and glottis-to-lips (G-L) distance. In this display, time is plotted along x–axis, y–axis represents distance from glottis-to-lip, and square-root of area value is represented by grey levels. Each new vertical frame corresponds to shifting the analysis window by 5 ms (~ 55 samples, for F_s = 11.025 kHz). Analysis of vowels and consonants [13] showed that LPC based vocal tract shape estimation is proper for vowels and semivowels, and is independent of pitch variation and amplitude variation (over an attenuation range of 0–40 dB).

3. Estimation of place of constriction

Production of VCV syllables involves movement of articulators from the articulatory position of the vowel towards that of the stop closure to that of the vowel. The dynamic variation in the estimated vocal tract area values and various coefficients related to vocal tract transfer function, during VC and CV transition segments, may be used to estimate the shape during closure duration. For this purpose, least-squares bivariate polynomial [14], [15] surfaces representing estimated area values or line spectrum pair (LSP) coefficients [16] over the VC and CV transition segments can be used for estimating the area during closure segment. As the articulatory movements involve low order dynamics, we have investigated the use of second order (conic) and third order (cubic) bivariate polynomials to model the area or coefficient values in the 2D surfaces.

3.1. Least-squares polynomial approximation and 2D interpolation

Polynomial and spline interpolation [14] are useful for curve or surface fitting of a univariate or bivariate data, using the least-squares method [17], [18]. Given a set of q univariate data points g_n, the objective of approximation is to find a function $f(x)$ that matches the data points within a small error r_n,

$$f(x_n) = g_n + r_n, \quad n = 0, 1, \ldots, q\text{-}1 \tag{1}$$

such that the sum of squared errors is minimized. In general

$$f(x) = \sum_{k=0}^{p-1} c_k \Phi_k(x) \tag{2}$$

where c_k represents a set of p parameters (to be determined), and Φ_k represents a set of a priori known functions. In case of univariate polynomial approximation,

$$\Phi_k(x) = x^k . \tag{3}$$

In matrix notation, data-fitting may be expressed as [18]

$$\mathbf{Az} = \mathbf{B} + \mathbf{r} \tag{4}$$

where,

$$\mathbf{A} = \begin{bmatrix} \Phi_0(x_0) & \Phi_1(x_0) & \ldots & \Phi_{p-1}(x_0) \\ \Phi_0(x_1) & \Phi_1(x_1) & \ldots & \Phi_{p-1}(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_0(x_{q-1}) & \Phi_1(x_{q-1}) & \ldots & \Phi_{p-1}(x_{q-1}) \end{bmatrix}$$

$$\mathbf{z}^T = \begin{bmatrix} c_0 & c_1 & \cdots & c_{p-1} \end{bmatrix}$$

$$\mathbf{B}^T = \begin{bmatrix} g_0 & g_1 & \cdots & g_{q-1} \end{bmatrix}$$

$$\mathbf{r}^T = \begin{bmatrix} r_0 & r_1 & \cdots & r_{q-1} \end{bmatrix} .$$

The case $q > p$ corresponds to an overdetermined system of simultaneous linear equations. Usually, p is small compared to q, in order to reduce interpolation errors due to numerical oscillations [18]. The method of least-squares can be applied to determine a set of parameters c_k such that the error function defined as

$$E(c_0, \cdots, c_{p-1}) = \sum_{n=0}^{q-1} \left[g_n - f(x_n; c_0, \cdots, c_{p-1}) \right]^2 \tag{5}$$

is minimized, and this leads to a system of p equations in p unknowns, called as normal equations. A solution to normal equations in matrix notation is given by

$$\mathbf{z} = \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \mathbf{A}^T \mathbf{B} \tag{6}$$

where matrix $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is the pseudo–inverse of $\mathbf{A}$ [17].

For approximating estimated vocal tract area values (or a set of LSP coefficients) as 2D surfaces, least-squares univariate data approximation can be extended to least-squares bivariate data approximation. In this case, the objective of the bivariate approximation is to find a surface $f(x, y)$, that represents estimated values in the least-squares sense. A lower order polynomial surface may model the area values adequately during VC and CV transition regions. Also, if the order of polynomial is large, resulting matrix is typically ill–conditioned [19]. Hence, investigations were restricted to the second order (conic) and third order (cubic) polynomial surfaces. The conic bivariate polynomial surface is given by

$$f(x, y) = c_0 + c_1 x + c_2 y + c_3 xy + c_4 x^2 + c_5 y^2 \tag{7}$$

where $f(x, y)$ approximates g_n, the estimated area (or coefficient) value at analysis frame 'x' (along time axis) and lip-glottis distance (or coefficient number) 'y', and c_0–c_5 are the conic polynomial coefficients. The cubic bivariate polynomial surface is given by

$$\begin{aligned} f(x, y) = {} & d_0 + d_1 x + d_2 x^2 + d_3 x^3 + d_4 y + d_5 y^2 \\ & + d_6 y^3 + d_7 xy + d_8 x^2 y + d_9 xy^2 \end{aligned} \tag{8}$$

where d_0–d_9 are coefficients of the cubic polynomial. Equations (7) and (8) for a set of q points, with $q > 6$ and $q > 10$ for conic and cubic polynomial approximation respectively, result in an overdetermined system of simultaneous linear equations which can be expressed in matrix notation by (4). In this case, $\mathbf{B}$ consists of area (or LSP coefficient) values in vowel–consonant (VC) and consonant–vowel (CV) transition regions. Figure 1 shows the transition regions along with a possible way for selecting values. Value m_0 along x–axis corresponds to the starting position of the transition segment along x–direction, m_1 and m_2 mark the segment over which area values could not be estimated, and m_3 marks the end of the transition. Thus modeled values in surface approximation correspond to

$$m_0 \le x \le m_1 \text{ and } m_2 \le x \le m_3 .$$

Thus, the number of frames to the left and right of stop closure, used for surface modeling are $L_{col} = m_1 - m_0 + 1$ and $R_{col} = m_3 - m_2 + 1$ respectively. In order to evaluate the polynomial coefficients in (7) and (8), we need to have $j = n_2 - n_1 \ge 2$ for conic and 3 for cubic polynomials respectively. As the 12–section values are plotted along y–axis, we get

$$j \le n_2 \le 12 .$$

The polynomial coefficient matrix $\mathbf{z}$ is obtained from (6) for least-squares fit. Two-dimensional interpolation of polynomial surfaces during stop closure duration will result in estimation of area or coefficient values during stop closure duration. For $m_1 < x < m_2$, $y = n_2$, and $\mathbf{z}$ as obtained from (6), the 2D interpolation of area (or coefficient) values during stop closure frames, along the glottis-to-lip position $y = n_2$ can be performed using

$$\hat{\mathbf{B}} = \hat{\mathbf{A}}\,\mathbf{z} . \tag{9}$$

The interpolation process is repeated for n_2 decreasing from 12 to j in steps of 1. For $n_2 = j$, $\hat{\mathbf{B}}$ in (9) is evaluated by varying y over the range $1 \le y \le j-1$, thus giving the interpolated area (or coefficient) values along the first $j-1$ glottis-to-lip positions. This direction of varying n_2 for interpolation gives better estimation at the lip end as compared to that at the glottis end, which is important in this application.

For bivariate polynomial surface generation and its 2D interpolation, the stop closure boundary locations need to be known. These locations were estimated using a two step

process; estimation of the beginning and ending points of VCV syllables, followed by the estimation of stop closure boundary locations within the VCV syllable. Short-time average magnitude with empirically selected threshold values were used for automated location of stop closure boundaries.

3.2. Validation of the proposed technique

To validate the proposed technique, VCV syllables /aja/ and /awa/ were recorded for three male and two female speakers. A central speech segment with increasing duration in each of these speech records (representing different cases) was artificially silenced for assessing proper recovery of vocal tract shape and/or place of articulation based on the proposed technique.

Figure 2 shows analysis results for /aja/ for one of the cases in which a middle segment of 250 ms was artificially silenced and 30 ms of VC and CV transition segments were available for surface modeling. Parts (a), (b), and (c) of Fig. 2 show speech waveform, spectrogram, and areagram respectively while parts (d) and (e) show areagrams obtained after 2D interpolation of conic and cubic surfaces respectively. Part (f) of the figure shows original areagram for /aja/ without any artificially introduced silence gap. Estimated end-points and silence interval boundary locations are indicated by downward arrows along the upper side of each of the figures. Outer pair of arrows shows end-points, while inner pair of arrows shows silence-interval boundary locations. After comparing 2D interpolation results with the known place of articulation for palatal /j/ as shown in part (f) of Fig. 2, it is validated that conic and cubic surface modeling of area values during VC and CV transition segments and its 2D interpolation during silence interval is capable of estimating proper place of articulation. Analysis of /aja/ and /awa/ for three male and two female speakers showed that conic surface based modeling of area values and LSP coefficients was more consistent than cubic surfaces in estimating proper place of articulation.

4. Results and discussion

Vowel-consonant-vowel syllables of the type /aCa/, /iCa/, /aCi/, /iCi/, and /uCu/ with stop consonants /p/, /b/, /t/, /d/, /k/, and /g/ were analyzed for estimation of the place of stop closure. Syllables of the type /aCa/ and /iCa/ were recorded for three male and two female speakers while syllables of the type /aCi/, /iCi/, and /uCu/ were recorded for one male speaker. Estimated place of constriction for the stop consonants was compared against earlier reported articulation places. From the available data based on MRI [20] and X-ray images [21], a typical range for the place of constriction for bilabial, alveolar, and velar stop consonants, on the normalized distance of 0 to 1 (0 corresponds to glottis position and 1 corresponds to lip position), is 1.0, 0.75 to 0.89, and 0.47 to 0.7 respectively. These data were used for validation of the estimated place of stop closures.

Interpolation results for VCV syllable /aka/, based on conic and cubic surface approximation of area values are shown in Fig. 3. Parts (a), (b), and (c) of the Fig. shows speech waveform, wideband spectrogram, and original areagram respectively. Parts (d) and (e) show areagram results obtained after performing 2D interpolation of conic and cubic surfaces respectively. It is observed that the area values are less around the normalized glottis-to-lip distance of 0.6, which corresponds to the place of constriction for velar stops.

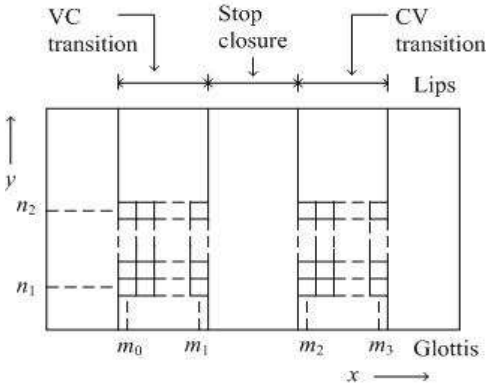

Figure 1: *Selection of area values (or LSP coefficient values) during transition segments for 2D interpolation.*

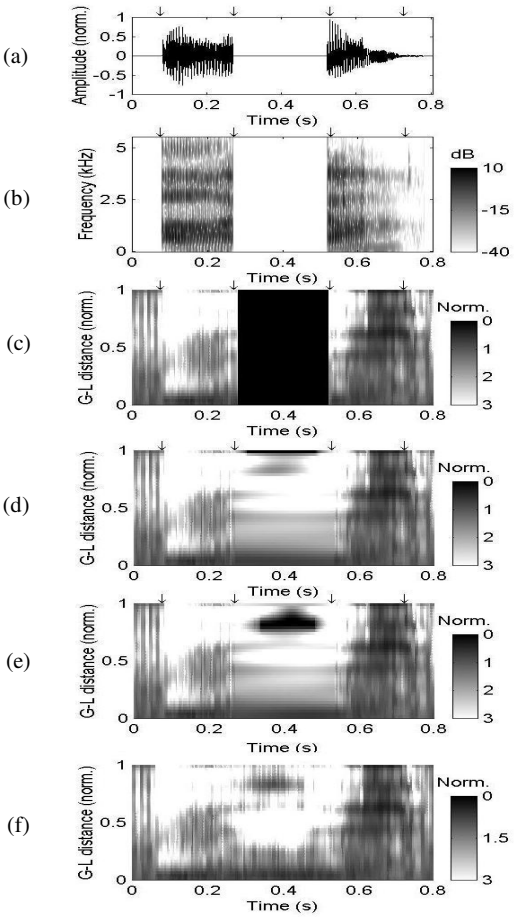

Figure 2: *2D interpolation of area values for /*aja*/ (silence interval = 250 ms): (a) waveform for 0.8 s; (b) spectrogram (Δf = 300 Hz); (c) areagram; (d) and (e) areagrams obtained after interpolation using conic and cubic surfaces respectively (surface generation parameters j = 3, L_{col} = 3, and R_{col} = 3); (f) original areagram for /*aja*/ without any silence gap.*

From the analysis results for VCV syllables of the type /aCa/ across all the speakers, it was observed that estimation of place of constriction for bilabial, alveolar, and velar stop consonants was more consistent for conic surface modeling of

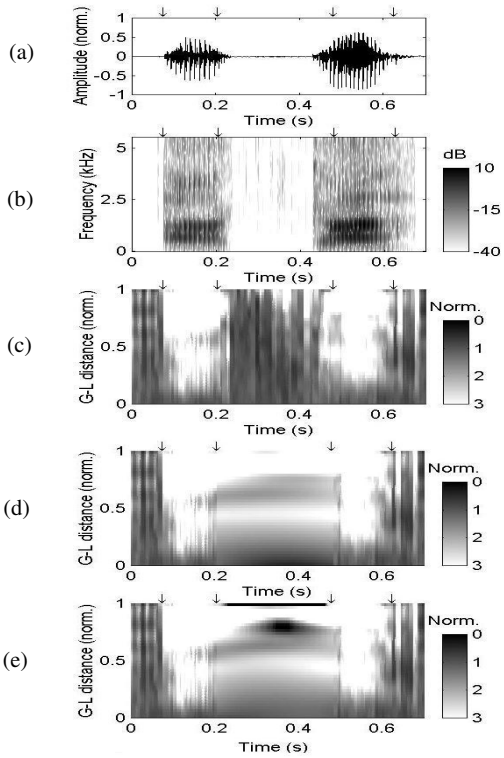

Figure 3: *2D interpolation of area values for /*aka*/: (a) waveform; (b) spectrogram (*Δf *= 300 Hz); (c) original areagram; (d) and (e) areagrams obtained after interpolation using conic and cubic surfaces respectively (surface generation parameters j = 6,* L_{col} *= 4, and* R_{col} *= 4).*

area values and LSP coefficients, which is in conformity with observations during initial validation of the technique. Analysis of /iCa/, /aCi/, and /iCi/ showed that estimation of place of constriction for velar stops is not consistent across the speakers. This indicates that the proposed technique is less effective in estimation of place of constriction when there is transition of place of articulation from front (as for vowel /i/) to back (as for velar /k/ and /g/).

5. Conclusions

It may be concluded that conic surface based modeling of either area values or LSP coefficients during VC and CV transition segments of VCV syllables of the type /aCa/, and its 2D interpolation during stop closure, can consistently estimate the place of closure for unvoiced and voiced bilabial, alveolar, and velar stops. The technique can be used for improving effectiveness of speech-training systems for production of stop consonants by providing visual feedback of place of closure. Use of interpolation may be investigated with various other techniques, for e.g., formant tracking, articulatory analysis by synthesis, etc., for vocal tract shape estimation.

6. References

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

[2] R. S. Nikerson, "Characteristics of the speech of deaf persons," *Volta Rev*., vol. 77, pp. 342–362, 1975; reprinted in [3].

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

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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, 1994.

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

[13] M. S. Shah and P. C. Pandey, "Estimation of vocal tract shape during stop closures," in *Proc. Int. Conf. on Systemics, Cybernetics, and Informatics* (Hyderabad, India), 2004, pp. 304-309.

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

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

[16] D. O'Shaughnessy, *Speech Communications: Human and Machines*. Reading, Massachusetts: Addison-Wesley, 1987.

[17] 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, 2001.

[18] R. L. Branham Jr., Scientific Data Analysis: *An Intro–duction to Overdetermined Systems*. New York: Springer-Verlag, 1990.

[19] S. D. Stearns and R. A. David, *Signal Processing Algorithms.* Englewood Cliffs, NJ: Prentice-Hall, 1992.

[20] 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.

[21] J. L. Flanagan, *Speech Analysis, Synthesis and Per–ception*, 2nd ed. New York: Springer-Verlag, 1975.