Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Estimation of Vocal Tract Shape for VCV Syllables for a Speech Training Aid

Milind S. Shah and Prem C. Pandey
Department of Electrical Engineering
IIT Bombay, Mumbai 400 076, India
E-mail: {milind, pcpandey}@ee.iitb.ac.in

***Abstract*–Display of vocal tract shape can be used in speech training aids for the hearing impaired children, as it provides a visual feedback of the articulatory efforts. Estimation of vocal tract shape, based on LPC and other analysis techniques, works satisfactorily for vowels but generally fails during stop closures. Indication of correct place of articulation is very important, particularly for non-labial consonants. In order to study the dynamics of the vocal tract shape estimation during transitions at vowel-consonant boundaries, we have used "areagram", a spectrogram-like two-dimensional (2D) display of estimated vocal tract cross-sectional area as a function of time and position along the tract length. Area estimation is based on reflection coefficients obtained from LPC analysis of speech. Based on estimated area during the transition segments preceding and following the stop closure, bivariate polynomial surfaces are obtained and these are used for estimation of the vocal tract shape during stop closure by performing 2D interpolation. The place of closure for various stop consonants could be estimated satisfactorily from the conic surface approximation.**

***Keywords*–Vocal tract shape estimation, Speech training aids for hearing impaired children, Place of stop closure.**

I. INTRODUCTION

In children with normal hearing, the process of learning to speak is primarily aided by auditory feedback. The absence of this feedback, as in profoundly deaf or prelingually deaf children, severely affects the accuracy of articulation and the nature of stress and intonation patterns [1]-[3]. The tongue shifts little from a neutral position in the mouth during the production of different sounds. The blending of consonant-vowel and consonant-consonant is prolonged. Also, vowels and consonants for which tongue movements are hidden in the mouth are typically indistinct in the speech of the deaf. Thus, despite having proper speech production mechanism, they are generally not able to produce intelligible speech.

It is possible to teach deaf persons to speak by use of appropriate non-auditory feedback, providing supplementary information through tactile or visual feedback [2], [4]. Visual feedback can be provided by display of acoustic and articulatory speech parameters, related to articulatory efforts which are easily controllable by the person undergoing speech training. Speech intensity, pitch, spectrogram, vocal tract shape, lip shape, and consonantal features like voicing, frication, and nasality have been used for providing visual feedback via speech training aids [5]. The earlier electronic speech training systems generally provided visual feedback of one of these aspects. More recent systems based on signal processing techniques provide, selectively but in an integrated way, display of several aspects for speech training [6]-[8].

Speech training systems providing visual feedback of vocal tract shape were found to be useful for improvement in vowel articulation [9]-[11]. Speech training for consonant articulation based on visual feedback of vocal tract shape has not been reported, although it may be more important. Various techniques like measurement of formant frequencies, measurement of acoustic impedance at the lips, LPC analysis, and neural network based system have been reported for the estimation of vocal tract shape [12]-[16]. These work satisfactorily for vowels but fail during most of the consonantal segments [14], [16], [17].

For studying the consistency of shape estimation for sustained vowels and the dynamics during transitions at vowel-consonant boundaries, we have used "areagram", a spectrogram like two-dimensional (2D) display of vocal tract cross-sectional area as a function of time and lip-to-glottis distance, represented by *x*- and *y*-axes respectively. Area value is represented by one of the 128 grey levels, minimum by black and maximum by white. This display is meant only for studying the vocal tract shape estimation, and will not be useful for speech training. For actual speech training, appropriate displays, cartoons, or games based on dynamically varying vocal tract shape will have to be devised and used.

Estimation of vocal tract shape, based on reflection coefficients obtained from LPC analysis of speech [18], showed that estimated shapes for synthesized and natural sustained vowels and semivowels are satisfactory and related to place of articulation [17], [19]. However, during stop closure segments of vowel-consonant-vowel (VCV) syllables, when the signal energy is very low, the estimation gives random area estimates, unrelated to place of closure. With amplitude scaling of synthesized and natural vowels with 16-bit quantization, it was found that satisfactory shape estimation is restricted to a dynamic range of 40 dB. Hence the shape estimation during stop closure fails to provide information on place of closure.

The vocal tract transfer function rapidly changes during the vowel-consonant (VC) and consonant-vowel (CV)

transitions and the dynamic estimation of vocal tract during these segments may be used to get the shape during the closure. In this paper, we present investigations for shape estimation during stop closures by performing 2D interpolation of bivariate polynomial surfaces based on estimated area during the transition segments preceding and following the stop closure.

II. Shape Estimation For VCV Syllables

A. LPC Based Vocal Tract Shape Estimation

Our implementation of vocal tract shape estimation is based on reflection coefficients obtained from LPC analysis [18], using Wakita's speech analysis model [14], and Robinson's algorithm [14] for optimum inverse filtering. Speech signal sampling rate, frame size, and LPC order were selected to obtain satisfactory vocal tract shape for the synthesized and natural vowels. The speech signal was acquired at a sampling rate of 11.025 k Sa/s. Each analysis frame duration was equal to twice the lowest pitch period, and the successive analysis frames were shifted by 5 ms. Optimum order of the linear predictor was found to be 12. The length of the vocal tract was taken as 17 cm.

The consistency of shape estimation for sustained vowels, and dynamics of shape estimation during VC transitions can be studied with the display of estimated area values in the form of areagram. The algorithm was tested with natural and synthesized speech signals for consistency and validity of the estimates [17], [19]. The areagram results are found to be satisfactory for vowels and semivowels. During the vowel segments of VCV syllables, the area estimates are consistent and related to place of articulation. During the stop closure, the segment energy is zero or very low and the area estimates appear to vary randomly across the analysis frames and are not related to place of closure. However, for various VCV syllables, area values during VC and CV transition segments had distinct patterns. Hence we carried out further investigation for predicting tract shape during stop closure from the shapes estimated on either side of the closure.

B. 2D Interpolation

In the articulatory movement for VCV syllables, the dynamic shape of vocal tract during VC and CV transitions is based on vowel preceding and following stop closure, and the actual place of stop closure. We have investigated use of bivariate polynomial surface approximation over the VC and CV transition segments for estimating the area during closure segment. We have used 2nd order (conic) and 3rd order (cubic) bivariate polynomials to model the area values in the areagram. The conic bivariate polynomial approximation is given by

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

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

$$f(x,y) \approx d_0 + d_1 x + d_2 x^2 + d_3 x^3 + d_4 y + d_5 y^2 + d_6 y^3 + d_7 xy + d_8 x^2 y + d_9 xy^2, \quad (2)$$

where d_0-d_9 are the cubic polynomial coefficients. These equations for a set of q points, with $q > 6$ and $q > 10$ for conic and cubic polynomial approximation respectively, result in over-determined system of simultaneous linear equations expressed in matrix notation as

$$\mathbf{Az} = \mathbf{B}, \quad (3)$$

where

$$\mathbf{B}^T = \begin{bmatrix} f(x_0, y_0) & f(x_1, y_1) & \dots & f(x_{q-1}, y_{q-1}) \end{bmatrix} \quad (4)$$

For conic polynomial approximation, $\mathbf{A}$ and $\mathbf{z}^T$ are given by

$$\mathbf{A} = \begin{bmatrix} 1 & x_0 & y_0 & x_0 y_0 & x_0^2 & y_0^2 \\ 1 & x_1 & y_1 & x_1 y_1 & x_1^2 & y_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{q-1} & y_{q-1} & x_{q-1} y_{q-1} & x_{q-1}^2 & y_{q-1}^2 \end{bmatrix} \quad (5)$$

$$\mathbf{z}^T = \begin{bmatrix} c_0 & c_1 & c_2 & c_3 & c_4 & c_5 \end{bmatrix} \quad (6)$$

For cubic polynomial approximation, $\mathbf{A}$ and $\mathbf{z}^T$ are given by

$$\mathbf{A} = \begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & y_0 & y_0^2 & y_0^3 & x_0 y_0 & x_0^2 y_0 & x_0 y_0^2 \\ 1 & x_1 & x_1^2 & x_1^3 & y_1 & y_1^2 & y_1^3 & x_1 y_1 & x_1^2 y_1 & x_1 y_1^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{q-1} & x_{q-1}^2 & x_{q-1}^3 & y_{q-1} & y_{q-1}^2 & y_{q-1}^3 & x_{q-1} y_{q-1} & x_{q-1}^2 y_{q-1} & x_{q-1} y_{q-1}^2 \end{bmatrix} \quad (7)$$

$$\mathbf{z}^T = \begin{bmatrix} d_0 & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 & d_9 \end{bmatrix} \quad (8)$$

The values in matrix $\mathbf{B}$ consist of areagram values in VC and CV transition regions. The transition regions and one possible way of area value selection are shown in Fig. 1. Value m_0 along x-axis corresponds to the starting position of the transition segment along x direction and value n_1 along y-axis corresponds to position along the tract length for which interpolated values for closure duration are to be obtained. Here the points used in the conic and cubic approximation are

$$m_0 \le x \le m_1 \text{ and } m_2 \le x \le m_3, \quad (9)$$

$$n_1 \le y \le n_2. \quad (10)$$

In order to evaluate the coefficients in (1) and (2), we need to have $j = n_2 - n_1 \ge 2$ for conic and 3 for cubic

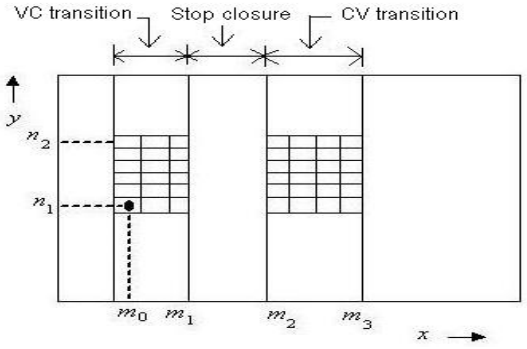


Fig. 1. Transition regions and area value selection.

approximation respectively. As the 12-section area values are plotted along *y*-axis, we get

$$1 \le n_1 \le 12 - j. \tag{11}$$

The polynomial coefficient matrix **z** can be obtained by using least square fit, which minimizes the sum of squares of the deviations of the data from the model. This solution [20] is obtained from (3) by

$$\mathbf{z} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{B}, \tag{12}$$

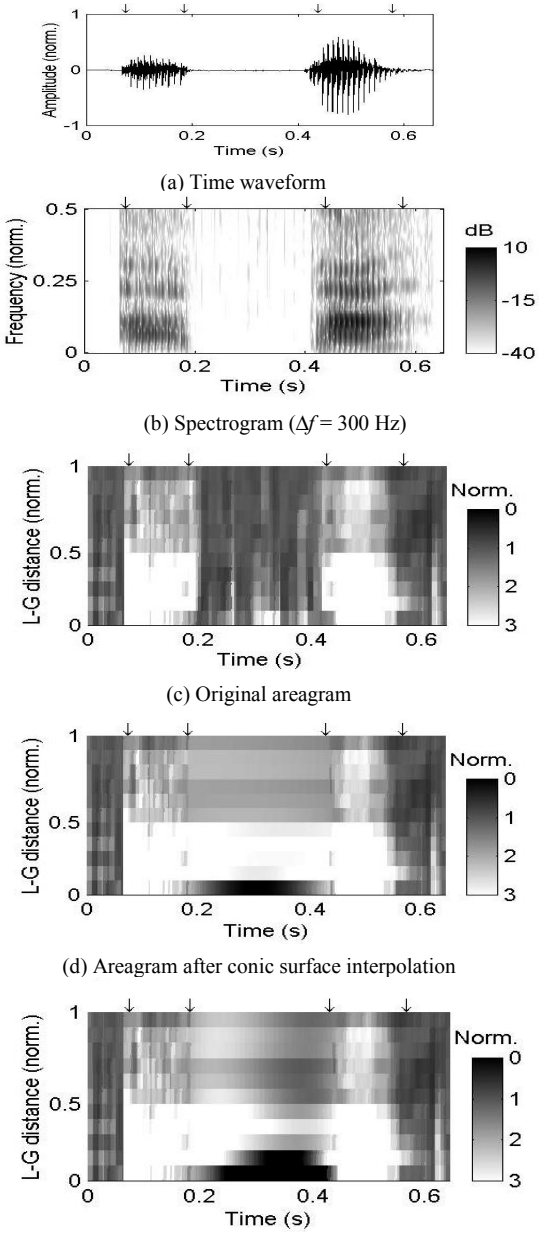
where matrix $(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ is pseudo inverse of **A**.

Detection of silence interval boundary for the stop closure is based on silence interval detection. The end-point detection of VCV syllable is based on short-time average magnitude of the signal [21]. Initial estimate of silence gap boundary is obtained by computing short-time average magnitude of the normalized syllable and comparing it with a particular magnitude threshold.

We have investigated shape estimation by obtaining conic and cubic polynomial surfaces based on estimated area values during both VC and CV transition, and then performing 2D interpolation using (1) or (2) for all the *x*, *y* values during stop closure. The 2D interpolation results obtained are presented in the next section.

III. Results And Discussion

The 2D interpolation results for VCV syllables /*apa*/ and /*aka*/, based on conic and cubic bivariate polynomial surface



(a) Time waveform

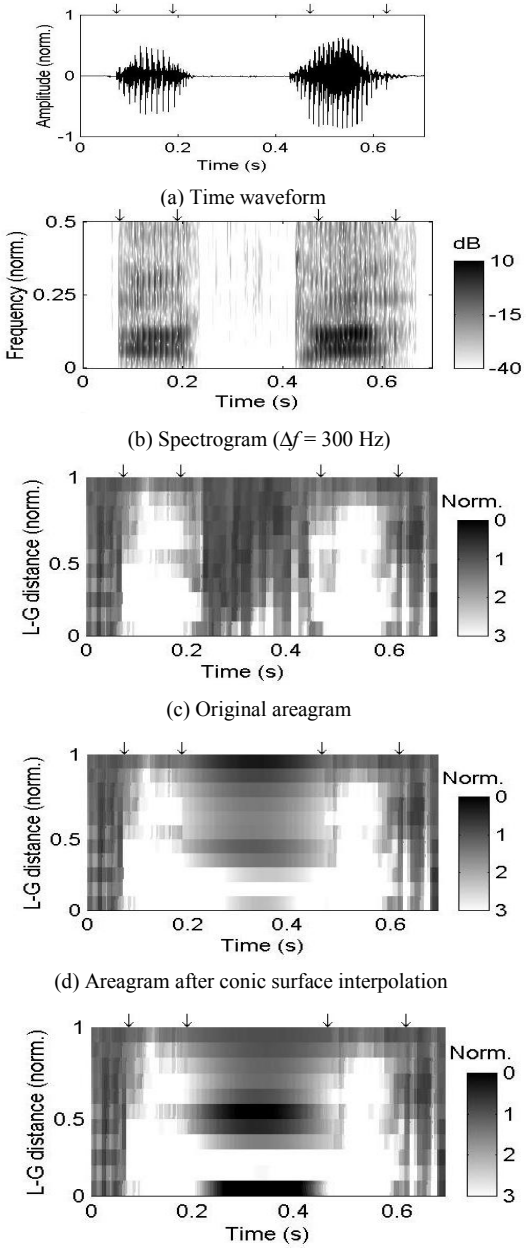
(b) Spectrogram (Δf = 300 Hz)

(c) Original areagram

(d) Areagram after conic surface interpolation

(e) Areagram after cubic surface interpolation

Fig. 2. Interpolation results for /*apa*/.



(a) Time waveform

(b) Spectrogram (Δf = 300 Hz)

(c) Original areagram

(d) Areagram after conic surface interpolation

(e) Areagram after cubic surface interpolation

Fig. 3. Interpolation results for /*aka*/.

approximation of area values, are shown in Figs. 2 and 3 respectively. Parts (a), (b), and (c) show time waveform, its wideband spectrogram, and areagram respectively. Parts (d) and (e) show areagram results obtained after performing 2D interpolation of conic and cubic surfaces. The value of *j* was 5 and 7 for /*apa*/ and /*aka*/ respectively.

We observe that original areagrams show satisfactory results during the vowel segment /*a*/, for which we expect more mouth opening in the front and less opening in the central region due to tongue elevation. The estimated area values appear to vary randomly during the stop closure duration of VCV syllable due to very low signal energy. However, area values during VC and CV transitions have distinct patterns. The results obtained from the silence interval boundary detection algorithm are indicated by downward arrows along the upper side of each of the figures. Outer pair of arrows gives VCV syllable boundary while inner pair of arrows gives stop closure boundary. We observe that both the boundaries are detected satisfactorily. With 2D interpolation based on conic and cubic surfaces, we are able to recover a satisfactory tract shape during closure duration with proper indication of place of constriction for labial /*p*/ and velar /*k*/. As can be seen from the results for /*aka*/, the interpolation results based on conic surface are more realistic than those based on cubic surface.

The analysis of VCV syllables showed that the interpolation technique gave area values, which were correlated with the place of closure for all the unvoiced and voiced stops in VCV syllables, with vowel /*a*/ and consonants /*p*/, /*b*/, /*t*/, /*d*/, /*k*/, and /*g*/.

IV. Conclusions

Our implementation of vocal tract shape estimation algorithm is based on reflection coefficients from LPC analysis of speech signal. In order to study the consistency of the vocal tract shape estimation with sustained vowels, and to study dynamics of shape estimation during the transitions at VC boundaries, we have used areagram, a 2D display of vocal tract area with time and lips-to-glottis distance.

After analyzing VCV syllables, it was observed that the algorithm for shape estimation works satisfactorily for vowel segments but fails during stop closures due to very low signal energy. However, areagram results during transition boundary from VC and CV were distinctly different. Based on estimated area before and after stop closure, bivariate polynomial conic and cubic surfaces are obtained by least square method and are used for shape estimation during stop closure by 2D interpolation. The algorithm for semi-automatic detection of silence interval boundary for the stop closure is found to be working satisfactorily. Conic and cubic surface based interpolation during stop closure could estimate vocal tract shape as well as place of closure in stop consonants. The 2D interpolation results based on conic surfaces are more realistic than the ones based on cubic surfaces.

After the vocal tract areas are estimated with consistency and appropriate dynamic response, these can be used for vocal tract shape display as part of a speech training aid for children with hearing impairment.

References

[1] H. Levitt, "Acoustic analysis of deaf speech using digital processing techniques," *IEEE Trans. Audio Electroacoust*., vol. 20, no. 1, pp. 35-41, 1972.

[2] L. S. Liben, *Deaf Children: Developmental Perspectives*. New York: Academic Press, 1978.

[3] R. S. Nikerson, "Characteristics of the speech of deaf persons," in *Sensory Aids for the Hearing Impaired*, H. Levitt, J. M. Pickett, and R. A. Houde, Eds. New York: IEEE Press, 1980, pp. 540-545.

[4] D. Crystal, *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press, 1997.

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

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

[7] M. L. Hsiao et al., "A computer based software for hearing impaired children's speech training and learning between teacher and parents in Taiwan," in *Proc. 23rd Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society* (Istanbul, Turkey), 2001, pp. 1457-1459.

[8] Dr. Speech Software Group, Software Demo on Dr. Speech 4, and Speech Therapy, http://www. drspeech.com, Tiger DRS, Inc., Seatle, Wa, 2003.

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

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

[11] 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.

[12] 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.

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

[14] 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.

[15] H. Deng, R. K. Ward, M. P. Beddoes, and M. Hodgson, "Estimating vocal-tract area functions from vowel sound signals over closed glottal phases," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2004, pp. 589-592.

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

[17] M. S. Shah and P. C. Pandey, "Areagram display for investigating the estimation of vocal tract shape for a speech training aid," in *Proc. Symposium on Frontiers of Research on Speech and Music* (Kanpur, India), 2003, pp. 121-124.

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

[19] 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.

[20] H. W. Brinkmann and E. A. Klotz, *Linear Algebra and Analytic Geometry*. Massachusetts: Addison-Wesley, 1971.

[21] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297-315, 1975.