Int. Symposium on Frontiers of Research on Speech and Music 2008 (FRSM 2008), Kolkata, India, pp. 161-166.

ESTIMATION OF PLACE OF CONSTRICTION DURING STOP CLOSURES BY BIVARIATE SURFACE MODELING

M. S. Shah and P. C. Pandey

SPI Lab, Department of Electrical Engineering
Indian Institute of Technology Bombay
Powai, Mumbai 400 076, India
E-mail:{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 low signal energy and unavailability of spectral information. A technique has been investigated for the estimation of place of constriction during stop closures by performing bivariate polynomial surface modeling on vocal tract area values during transition segments preceding and following the closure and their interpolation during closure segment. The place of constriction estimated with our technique compared well with the actual constriction locations observed from the articulatory data in /∧Ca/ utterances. The technique can be used for improving effectiveness of speech-training systems for the production of stop consonants.

Keywords–Estimation of place of constriction, Speech training aids, Vocal tract shape estimation.

1. Introduction

Children having prelinguistic profound hearing impairment have great difficulty in acquiring speech because of lack of auditory cues. It is possible to teach them to speak intelligibly by use of appropriate non-auditory feedback by providing supplementary information through visual feedback of acoustic and articulatory speech parameters [1]–[3]. The acoustic parameters include: speech intensity, fundamental frequency, spectral features, etc., while the articulatory parameters include: voicing, nasality, lip movement, tongue movement, etc.

Electromyography data of the hearing impaired and normal hearing persons are almost similar in case of lip movement, but it differs 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 the ease with which hearing impaired persons learn to produce specific sounds. Visual feedback of the 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]–[8].

Estimation of the vocal tract area function from speech signal, an inverse problem, can be carried out by one of the several techniques: LPC analysis [9], use of formants and factor analysis [10], use of formants and perturbation theory [11], mapping via articulatory codebook [12], etc. Other indirect methods for estimation of the vocal tract shape are based on measurement of input impedance (either in frequency domain [13], [14] or in time domain [15]) at the lips using an impedance tube. Most of these techniques are reported to work satisfactorily for vowels. However, shape estimation fails if spectral information is not available, for example during oral stop closure duration [9], [12], [16]. Hence, it is important to investigate a technique for vocal tract shape estimation for consonant articulation which can be used for improving effectiveness of the speech-training systems for the hearing impaired persons.

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 speech spectral zeros, it is used for developing speech training aids. From the investigations carried out for LPC based vocal tract shape estimation for various vowel-stop consonant-vowel (VCV) utterances, it was observed that area values were random and unrelated to place of constriction during stop closure. However, variation in the area values during vowel-consonant (VC) and consonant-vowel (CV) transition

segments present before and after stop closure appear to be different for different places of closure. Hence, the information for estimating the place of closure may be contained in the short transition segments preceding and following the stop closure. This paper presents investigations for vocal tract shape estimation during stop closures by performing 2D interpolation of bivariate polynomial surfaces based on estimated area values during the transition segments in VCV utterances. The technique implementation, reported earlier in [17] and [18], permits automated processing of the acquired utterances. Utterances of the type /aCa/ were analyzed and the estimated place of constrictions was compared with the typical ranges for the place of constrictions for bilabial, alveolar, and velar stop consonants. For direct validation of the technique, it is applied to acoustic signals that have been simultaneously acquired with articulatory data [19] so that estimated location of constrictions can directly be compared to the actual constriction locations.

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 [9], [20]. The vocal tract was assumed to be 17 cm long from the glottis to the lips, a typical value for an adult male. For avoiding fixed-point arithmetic related recursive errors and dynamic range limitation, the processing was carried out with floating-point arithmetic. The vocal tract shapes were obtained with analysis window size equal to twice the average pitch period and LPC order 12 on speech signal sampled with F_s of 11.025 kHz.

In order to study the consistency of shape estimation with amplitude and pitch variation in vowels, and to study the dynamics of shape estimation during transitions, we have used "areagram", a spectrogram [20] like 2D 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. In this display, each new vertical frame corresponds to shifting the analysis window by 5 ms (~ 55 samples, for F_s = 11.025 kHz). Areagram 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.

The consistency and validity of the estimated shapes was checked by analyzing synthesized and natural vowels and VCV utterances and the results have been reported in [21]. LPC based vocal tract shape estimation for vowels were found to be independent of pitch variation and amplitude variation (over an attenuation range of 0–40 dB). For verifying shape tracking ability of LPC based estimation, vowel-semivowel-vowel utterances were analyzed. Areagram results showed proper transition in vocal tract shapes and correct estimation of place of articulation for semivowels. Next, VCV utterances involving stop consonants were analyzed. Estimated place of articulation was proper for vowel segments, but area estimates were random and unrelated to place of constriction during stop closure due to zero or low signal energy and unavailability of relevant spectral information. However, analyses of a large number of VCV utterances showed that variation in area values during transition segments before and after stop closure were distinctly different for different places of stop closures. It indicated that variation in estimated area values during transition segments may contain information related to place of articulation for stop consonants.

3. Shape Estimation during Stop Closure

In a VCV utterance, the dynamic movement of articulators before and after the stop closure (represented here as VC and CV transition segments) is related to the articulatory positions for the vowel preceding and following the stop closure, and the actual place of the stop closure. The dynamic variation in the estimated vocal tract area values during VC and CV transition segments were used to estimate the shape during closure duration. Polynomial and spline interpolation [22] are useful for curve or surface fitting of a univariate or bivariate data, using the least-squares method. We have investigated bivariate polynomial [22], [23] based surface modeling of estimated area values during the transition segments for estimating the place of articulation during closure segment. Movement of articulators is characterized by formant transitions and second degree differential equations are reported to provide good fits to the formant transitions [24]. Hence, investigations were carried out with the second and third degree bivariate polynomials to model the area values.

The estimated vocal tract area values, $g(x, y)$, during VC and CV transition segments were

modeled by the bivariate polynomial surface, $f(x, y)$, within a small error $r(x, y)$. Here 'x' represents analysis frame number along time axis and 'y' represents one of the integer values from 1 to 12 (as 12 area values are estimated per frame) along glottis-to-lips distance. Figure 1 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 can not be estimated, and x_d marks the end of the transition. Thus estimated area values which are used for surface approximation correspond to $x_a \le x \le x_b$, $x_c \le x \le x_d$, and $y_a \le y \le y_b$. The number of frames to the left and right of stop closure used for surface modeling are $L_{col} = x_b - x_a + 1$ and $R_{col} = x_d - x_c + 1$ respectively.

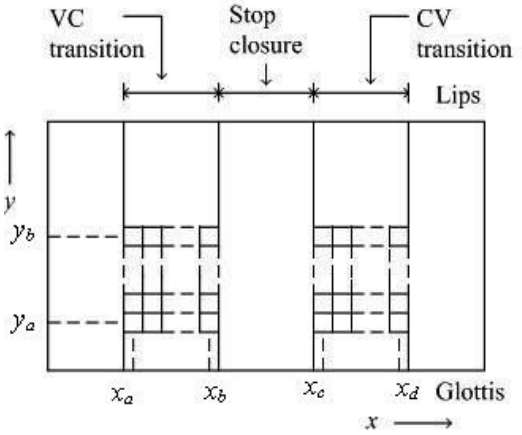

Fig. 1. Selection of area values during transition segments for 2D interpolation.

The second degree bivariate polynomial surface is given by

$$f(x,y) = c_0 + c_1x + c_2y + c_3xy + c_4x^2 + c_5y^2 \quad (1)$$

where $f(x, y)$ approximates $g(x, y)$ in the least-squares sense and c_0–c_5 are the polynomial coefficients. The third degree bivariate polynomial surface is given by

$$f(x,y) = d_0 + d_1x + d_2x^2 + d_3x^3 + d_4y + d_5y^2 + d_6y^3 + d_7xy + d_8x^2y + d_9xy^2 \quad (2)$$

where d_0–d_9 are the polynomial coefficients. In order to evaluate the polynomial coefficients in (1) and (2), we need to have $j = y_b - y_a \ge 2$ for second degree and 3 for third degree polynomials respectively. As the 12–section values are plotted along the y–axis, we get $j \le y_b \le 12$. Equations (1) and (2) for a set of q points, with $q > 6$ and $q > 10$ for the second and third degree polynomial approximations respectively, result in an overdetermined system of simultaneous linear equations. The matrix formulation of linear equations followed by its pseudo–inverse operation, as reported earlier in [17] and [18], were used for obtaining polynomial coefficients in (1) and (2). Two-dimensional interpolation of polynomial surfaces during stop closure duration resulted in estimation of area values during stop closure [17], [18]. For bivariate polynomial surface generation, the stop closure boundary locations need to be known. These locations were estimated using short-time average magnitudes and empirically selected threshold values. The fricative burst release before the transition was excluded for bivariate surface modeling.

4. Results and Discussion

The technique for estimation of the place of constriction during stop closures was applied to the speech utterances of the type /aCa/ with stop consonants /p/, /b/, /t/, /d/, /k/, and /g/. The results obtained, reported in [17] and [18], showed that the place of constriction could be consistently estimated for both voiced and unvoiced stops. Estimated place of constriction was compared with the typical range for the place of constriction available from MRI data and X-ray images. It was observed that estimation of place of constriction for bilabial, alveolar, and velar stop consonants was more consistent with the second degree surface modeling of area values than with the third degree surfaces [17], [18].

For a direct validation, the technique was applied to acoustic signals that have been simultaneously acquired with articulatory data, available from the University of Wisconsin [19] so that estimated location of constrictions can directly be compared to the actual constriction locations. The database incorporated point-parameterized representation of lingual, labial, and mandibular movements recorded using X-ray microbeam system (XRMB), in synchrony with the resulting acoustic wave. Utterances of the type /∧Ca/ involving voiced stop consonants /b/, /d/, and /g/ were analyzed for the estimation of the place of stop closure. The estimated place of constriction for the

stop consonants was compared against the actual place of constriction recorded with the XRMB system.

Interpolation results based on second degree polynomials were more consistent in estimating the place of constriction than the third degree polynomials. Interpolation result for the utterance /∧da/ spoken by an adult male speaker, based on second degree polynomial surface modeling of area values is shown in Fig. 2. Parts (a), (b), and (c) show spectrogram, areagram, and waterfall diagram for /∧da/. The figure caption shows the number of frames (L_{col} and R_{col}) and the number of rows (j) required for the surface modeling and interpolation for the proper estimation of place of articulation. It is observed from the result in part (b) that the estimated location of constriction during stop closure is around the normalized distance of 0.8. The estimated location of constriction obtained from our algorithm was compared with the actual constriction location recorded with X-ray microbeam system. Part (d) of Fig. 2 shows x-y articulatory plot for the maximum vocal tract constriction during stop closure for the utterance /∧da/. Based on x-position of the pellet markers and with the assumption that the length of the vocal tract from glottis-to-lips is equal to 17 cm, the actual location of maximum vocal tract constriction was calculated. It was normalized on the scale of 0 (glottis location) to 1 (lips location) for its comparison with the estimated place of constriction. The actual place of maximum constrictions for /∧da/ was found to be 0.8. Thus, the estimated place of constriction matched exactly with the actual place. The estimated places of maximum constrictions for /∧ba/, /∧da/, and /∧ga/ for all the speakers matched well with the actual places of constrictions.

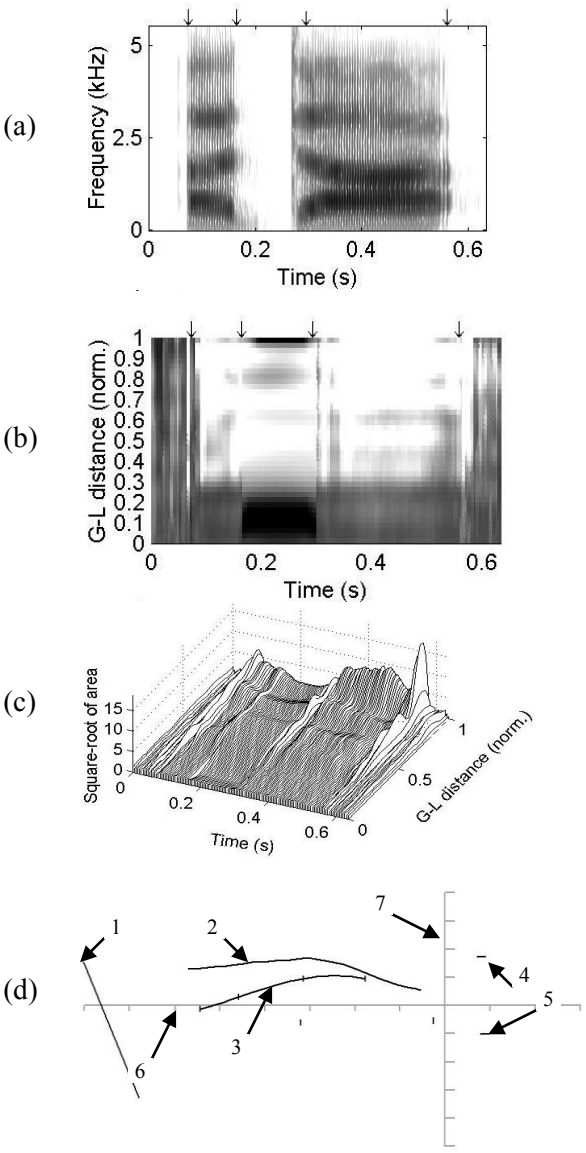

Fig. 2. Second degree polynomial surface interpolation results for the utterance /∧da/: (a) wideband spectrogram; (b) and (c) areagram and corresponding waterfall diagram (surface generation parameters j = 5, L_{col} = 5, and R_{col} = 5); (d) x-y articulatory plot from X-ray microbeam data for maximum vocal tract constriction during stop closure for the utterance /∧da/. Labels in part (d) correspond to: 1-Back pharyngeal wall; 2-Palate; 3-Tongue and four pellet markers; 4 & 5-Upper and lower lip pellet markers respectively; 6 & 7-Pellet position along x- and y-axis respectively in mm.

5. Conclusion

It may be concluded that least-squares surface modeling, with the second degree bivariate polynomials, of area values during transition segments before and after stop closure for utterances of the type /aCa/ and /∧Ca/ and its 2D interpolation during stop closure can estimate the place of constriction of the original stop consonants. The technique can be used for improving the effectiveness of speech-training systems for the production of stop consonants by providing visual feedback of the vocal tract shape, and specifically the place of maximum constriction during stop closures.

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: *Sensory Aids for the Hearing Impaired*, H. Levitt, J. M. Pickett, and R. A. Houde, (Eds.). New York: IEEE Press, 1980, pp. 540–545.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] 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 Engg. Med. Biol. Soc.*, 2005, pp. 6642–6645.

[17] M. S. Shah and P. C. Pandey, "Estimation of place of articulation in stop consonants for visual feedback," in *Proc. Int. Conf. Interspeech 2007-Eurospeech* (Antwerp, Belgium), 2007, pp. 2477–2480.

[18] 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. Congress on Acoust.* (Madrid, Spain), 2007, pp. CAS-03-011-1–CAS-03-011-5.

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

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

[21] 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.

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

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

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