Proc. 13th Annual Conf. Int. Speech Communication Association (Interspeech 2012), Portland, Oregon, 9 - 13 September 2012, Paper No. 497

Detection of Transition Segments in VCV Utterances for Estimation of the Place of Closure of Oral Stops for Speech Training

Jagbandhu, K. S. Nataraj, P. C. Pandey

EE Dept., Indian Institute of Technology Bombay, Mumbai 400076, Maharashtra, India

{jagbandhu, natarajks, pcpandey} @ ee.iitb.ac.in

Abstract

Speech-training aids providing a visual feedback of articulatory efforts can be used for improving articulation by the hearing-impaired children. LPC-based estimation of vocal tract shape works satisfactorily for vowels but fails during stop closure. The vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterances can be estimated by bivariate surface modeling of the vocal tract area function during the transition segments preceding and following the stop closure. The accuracy of the estimated shape during the closure depends on the detection of the transitions. A technique for detecting the VC and CV transitions in VCV utterances based on a measure of the rate of change of vocal tract area function is presented. The automatically marked start and end points of transitions showed a good match with the manually marked ones and resulted in a consistent estimation of the place of closure of velar, alveolar, and bilabial stops.

Index Terms: speech training aid, vocal tract shape, transition segment detection

1. Introduction

The acoustic-to-articulatory mapping involves estimating the sequence of vocal tract shapes that produce a given speech signal [1]. It has potential applications in speech synthesis [2],[3], speech recognition [4], and speech training aids [5]-[7]. The shape of the vocal tract can be specified by its cross-sectional area values as a function of position along the tract length [1]. Several methods can be used for directly estimating the vocal tract shape from the speech signal: linear predictive coding (LPC) based analysis [8],[9], formants and factor analysis [10], articulatory codebook mapping [1], etc. Speech-training aids providing a visual feedback of articulatory efforts have been found to be useful in improving vowel articulation by the hearing-impaired children [6],[7]. Most of these systems use Wakita's LPC-based method [9] for direct estimation of vocal tract shape as it is suitable for real-time processing [6]. It works satisfactorily for vowels, but fails during the closure segments of oral stops due to low energy and lack of spectral information [9],[11]. Hence, a technique for estimating the place of articulation of stop consonants is needed for improving the effectiveness of speech-training aids.

Pandey and Shah [11] reported a method to interpolate the vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterances by using a bivariate surface model fitted on the vocal tract shapes during the transition segments preceding and following the stop closure. Use of least-squares based bivariate quadratic approximation resulted in satisfactory estimation of place of closure for different unvoiced and voiced oral stops. The estimated place of closure for /ΛCa/ utterances, involving stop consonants /b/, /d/ and /g/ from 20 male and 20 female speakers, showed a good match (with a correlation coefficient of 0.94) with those obtained from direct X-ray imaging from XRMB database [12]. A detailed investigation showed that the accuracy of the estimation was highest if the segments used for the modeling corresponded closely to the VC transition segment preceding the closure and the CV transition segment following the closure [13]. Inclusion of a part of the closure or the release burst in the segments used in the modeling can introduce errors, while that of a part of the vowel on either side can decrease the sensitivity of the fitted model in interpolating the place of closure. Hence a method for automatically detecting the VC and CV transitions is needed for improving the accuracy of the estimated place of closure and improving the feedback given by the visual speech training aids.

Several methods for detecting the acoustic landmarks associated with stop consonants have been reported [14], but they do not mark the start and end points of the transitions. As accurate formant tracking [15] is still a challenging task, it cannot be used to accurately detect the transition segments. Onset of voicing after the stop closure is generally detected as the onset of periodicity in the acoustic waveform. Several methods to detect the voicing onset for calculating the voice onset time have been reported [16]. But they are not useful in detecting the CV transition, because presence of aspiration decreases the accuracy of voice onset detection [17] and association of this landmark with the start and end of the CV transition varies with voicing and place of the stop consonant.

Plots of vocal tract area function of VCV utterances with different oral stops show a distinct pattern during the transition segments as compared to the steady state segments. A method based on a measure of the rate of change in the area values is presented for marking the VC and CV transitions for estimating the place of closure of oral stops in VCV utterances.

2. Estimation of place of closure

Vocal tract shapes are estimated using Wakita's method [9]. The speech signal is sampled at 10 kHz with 16-bit quantization. A first difference of the signal is taken for providing an approximate 6-dB/octave pre-emphasis. Hamming window is applied on analysis frames with duration equal to twice the average pitch period, and the successive analysis frames are shifted by 5 ms. A 12th order LPC is used to compute the reflection coefficients from the autocorrelation coefficients of the windowed frame. The vocal tract from glottis to lips is modeled as 12 cylindrical sections of equal length and the vocal tract area function is obtained as ratios of the areas on both sides of the section interfaces and these ratios are converted into areas by assuming a constant normalized area of unity at the glottis end. The amount of opening in a section of the vocal tract is obtained as the square

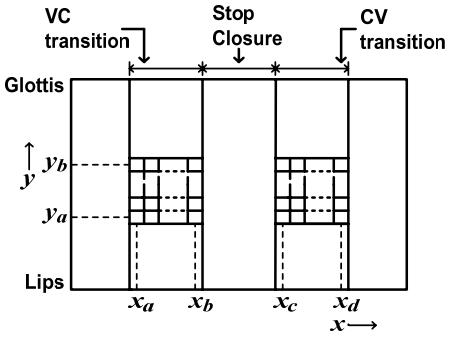

Figure 1: *Selection of transition area values for 2D modeling and interpolation [11].*

root of the area value for use in subsequent analysis and display, although it is referred to as area.

LPC-based vocal tract shape estimation fails during stop closure due to unavailability of relevant spectral information. The vocal tract shape during the stop closures of VCV utterances can be estimated by bivariate polynomial modeling of the shapes during transition segments preceding and following the stop closure as they tend to show different two-dimensional patterns for different places of closure [11]. In this method, the estimated vocal tract area values, given as $g(x, y)$ at analysis frame x (along the time axis) and the section number y from the lip end, during the VC and CV transitions, are approximated by a bivariate quadratic function. The function coefficients are obtained for minimizing the sum of the squared approximation error. The section numbers $y_a \le y \le y_b$ for the frames $x_a \le x \le x_b$ and $x_c \le x \le x_d$ corresponding to VC and CV transitions, respectively, are used for the modeling as shown in Figure 1. The functions with the estimated coefficients are used for interpolating $g(x, y)$ during the closure segment ($x_b \le x \le x_c$). In [11], the transition start and end points were estimated using a two-step process. The start and end points of the utterance were estimated using the short-time average magnitude of the signal [18]. The start and end of the closure were marked using an empirically selected threshold of 0.2 times the RMS value of the VCV utterance, and the end point was delayed to exclude the closure burst. The VC transition start x_a and the CV transition end x_d were marked using empirically selected durations. The transition durations vary across speakers, and it is difficult to empirically set the values for utterances from speakers with speech impairment.

3. Transition segment detection

Vocal tract area values estimated using LPC analysis with fixed-length window show a significant variability with the position of the analysis frame. It has been shown that the area values obtained at the minima of the windowed energy index (the ratio of the energy of the windowed signal to the energy of the signal within the frame) can be used for improving the consistency of vocal tract shape estimation [19]. This method was used to interpolate the values for frames at 5 ms intervals. Plots of area values for VCV utterances from a number of speakers were examined for a relationship between estimated area values and transition segments. These plots indicated that a measure of the rate of change of the vocal tract shape may be

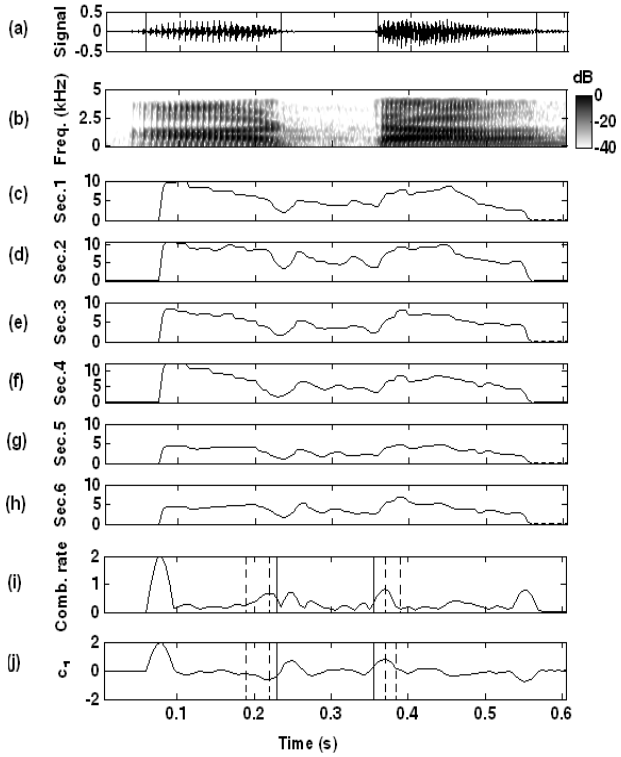

Figure 2: *Rate of change functions: (a) waveform (0.6 s) of /ata/ from a male speaker; (b) wideband spectrogram; (c)-(h) areas of first 6 sections, starting from the lips, (i) RMS of rate of change of areas, (j)* c_1 *from bivariate linear approximation.*

useful in marking the transition segments during VCV utterances. As an example, Figure 2 shows results for /ata/. The areas for first 6 sections are relatively steady as the vocal tract shape does not change significantly during the vowel. The estimated values show transition during the VC and CV transitions, and the transitions extend over approximately the same interval as the formant transitions.

3.1 Computation of rate of change

A rate of change measure of the two-dimensional vocal tract area function may be computed by combining the rate of change of individual sections. As the first difference was found to be noisy, slope estimated from a moving 7-point linear regression was used as a rate of change measure. The root-mean-square of slopes of the first 6 sections was used as a combined rate of change. This function, as shown in Figure 2(i), has distinct valleys at the VC transition start and the CV transition end and peaks at the VC transition end and the CV transition start. However, an examination of the plots of this measure for different utterances from several speakers showed that the valleys and peaks were not consistently associated with the transition points, and hence this measure is not suitable for automated marking of the transitions.

Time-slope of the moving bivariate polynomial fitted on the area values $g(x, y)$ was used as another measure of rate of change. We used linear and quadratic approximations. The bivariate linear approximation is given by

$$g(x, y) = c_0 + c_1 x + c_2 y + \varepsilon(x, y) \qquad (1)$$

where ε is the approximation error. Modeling is carried out for first six sections, i.e. $1 \le y \le 6$ and for moving 7-frame segment centered at the current frame x_n, i.e. $x_n - 3 \le x \le x_n + 3$. The coefficients c_0 - c_2 are obtained for least squared error. In matrix notation, the bivariate linear polynomial can be expressed as,

$$\mathbf{Az} = \mathbf{b} + \mathbf{r} \tag{2}$$

where $\mathbf{r}$ represents the approximation error, and $\mathbf{A}$, $\mathbf{z}$, and $\mathbf{b}$ matrices are given by

$$\mathbf{A}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ x_n-3 & x_n-3 & \cdots & x_n-3 & x_n-2 & \cdots & x_n+3 \\ 1 & 2 & \cdots & 6 & 1 & \cdots & 6 \end{bmatrix}$$

$$\mathbf{z}^T = \begin{bmatrix} c_0 & c_1 & c_2 \end{bmatrix}$$

$$\mathbf{b}^T = \begin{bmatrix} g(x_n-3,1) & g(x_n-3,2) & \cdots & g(x_n+3,6) \end{bmatrix}$$

The polynomial coefficient vector $\mathbf{z}$ for least squared error is,

$$\mathbf{z} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{3}$$

The coefficient c_1 is taken as the combined rate of change of area values along time axis x.

Figure 2(j) gives a plot of c_1. It has a small value during the steady state vowel segments. At the VC transition start, a constriction begins and c_1 becomes negative, reaching a negative peak almost near the closure. After the closure release, the opening starts increasing and it is indicated by a positive value of c_1. Its value reaches a peak at the CV transition start and then falls to almost zero near its end. This pattern was observed across the utterances examined from several speakers, indicating the suitability of c_1 for marking the VC and CV transitions. It may be noted that the peaks and valleys in c_1 may occur at the start and end of the utterance and also during the closure, but being outside the search region they do not affect the transition detection. A similar investigation with bivariate quadratic approximation showed that the change in vocal tract shape during the transitions was represented by two coefficients, but neither captured the change as effectively and consistently as c_1 from bivariate linear approximation.

3.2. Detection of transition segment boundaries

An example of markings for locating the boundaries of the transition segments is shown in Figure 2. The start and end of the stop closure are estimated using short time average magnitude with a threshold value of 0.2 times the RMS value of the signal [11]. These points marked as solid vertical lines on the waveform in Figure 2(a), are used as first estimates of the VC transition end and the CV transition start, respectively. Moving left from the closure, a negative peak c_{1n} is searched to mark a refined estimate of the VC transition end. The values in the frames further left are compared with a threshold to mark the VC transition start. To make the detection adaptive to different rates of transition in the utterances, the threshold is kept as $0.2c_{1n}$ for the first 5 frames and then changed by $0.1c_{1n}$ for each frame. Similarly, a positive peak c_{1p} is searched on the right side of the closure to mark the CV transition start. The positive peak tends to enter into the CV transition during aspirated stops, leading to an error in marking the CV transition start. In such stops, the time delay between the short time average magnitude based estimate and the positive peak was observed to be consistently more than 15 ms. Hence if the time delay between the initial estimate and the positive peak is more than 15 ms, a point corresponding to $0.5c_{1p}$ is located before the positive peak and is marked as the CV transition start. The values in the frames further right are compared with a threshold to mark the CV transition end. The threshold is kept as $0.2c_{1p}$ for the first 5 frames and then changed by $0.1c_{1p}$ for each frame. The detected transition start and end points are marked by dotted lines in Figure 2(j).

4. Results

Utterances of the type /aCa/ with stop consonants /p,b,t,d,k,g/ from two male and two female speakers were analyzed for estimation of place of maximum constriction during the stop closures. The start and end points of VC and CV transitions were marked by the method presented in the previous section. The accuracy of the automated markings was evaluated by comparing them with the manual markings obtained by visual inspection of the wideband spectrograms using PRAAT. The errors in the marking of the four transition points were calculated. There was one large error, of 30 ms in marking VC transition start for /ada/ from one speaker. As the transition in this utterance was much longer than the error, the error did not affect the bivariate modeling. All the other errors were within ±15 ms. The range, mean and standard deviation of the errors are given in Table 1.The standard deviations are less than 10 ms, indicating a match with the manually marked locations.

The marked transition segments were used for estimating the vocal tract shape during the stop closure using the bivariate quadratic modeling and interpolation. The results for VCV utterances with stops /p, t, k/ are shown in Figure 3. The vocal tract area values are displayed as areagram, a two-dimensional plot of the square root of area values as gray levels as a function of time along x-axis and normalized distance from glottis towards lips (G-L) along y-axis. The four arrows indicate the detected transitions. The largest constrictions during the closure in the areagrams are at the normalized distance of 1, 0.8, and 0.6 for /p/, /t/, and /k/, respectively, matching with the values of 1, 0.75 - 0.89, and 0.47 - 0.70, for the bilabial, alveolar, and velar oral stops, respectively, as estimated from MRI [20] and X-ray images [21].

Table 1: *Errors in transition segment estimation*

Transition point	Error (ms)			
	Range		Mean	Std. dev.
VC transition start	-10	30	8.1	8.3
VC transition end	-15	15	1.2	7.4
CV transition start	-5	10	2.7	3.9
CV transition end	-15	10	-1.0	7.7

5. Conclusion

A technique using the slope of the bivariate linear approximation fitted on the vocal tract area function for automatically marking the VC and CV transitions during VCV utterances has been presented. The automatically marked points have a

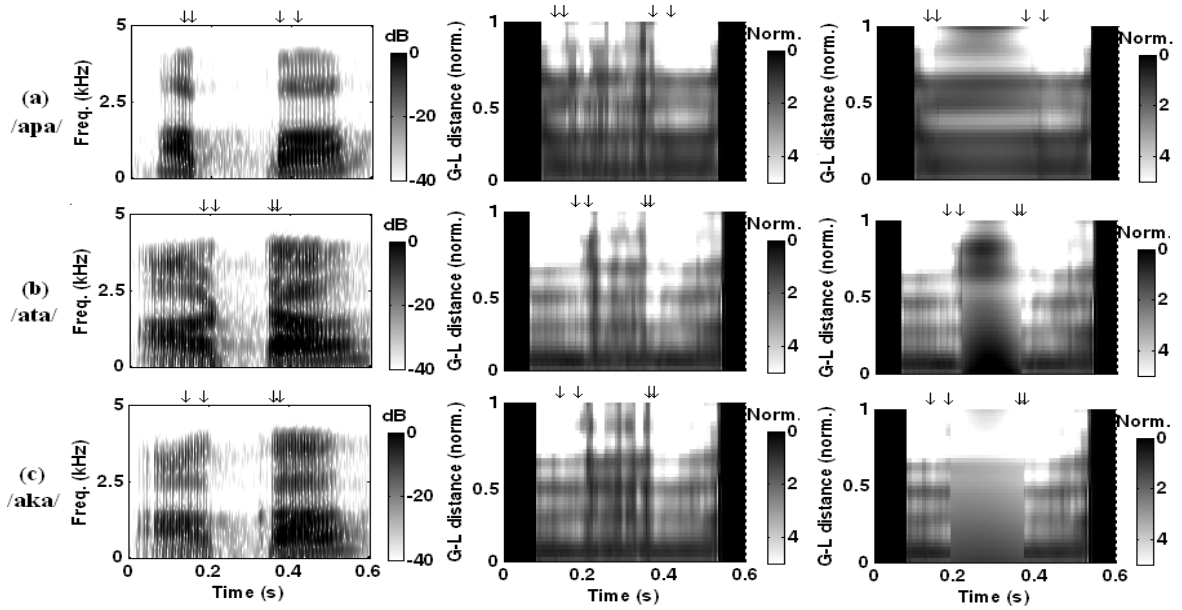

Figure 3: *Interpolation results for the VCV utterance with unvoiced bilabial, alveolar and velar stops: (a)/apa/; (b) /ata/; and (c) /aka/. Left side: wideband spectrogram, middle: original areagram, right side: areagram obtained using bivariate quadratic surface modeling of the automatically detected transition segments.*

good match with the manually marked ones, and resulted in satisfactory estimates of place of closure of oral stops. The technique needs to be further evaluated on VCV utterances from speakers with different languages and accents, and particularly on utterances from children needing speech training. In addition to its potential for improving the visual speech training aids, the technique may be useful in other non-real time speech processing applications requiring an accurate marking of the transitions.

6. Acknowledgements

The research work is supported by the project "National Program on Perception Engineering", sponsored by the Department of Information Technology, MCIT, Government of India.

7. 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, 1994.

[2] C. H. Coker, "A model of articulatory dynamics and control," Proc. IEEE, vol. 64, no. 4, pp. 452-460, 1976.

[3] P. Mermelstein, "Articulatory model for the study of speech production," J. Acoust. Soc. Am., vol. 53, no. 4, pp. 1070-1082, 1973.

[4] M. Naito, L. Deng, and Y. Sagisaka, "Speaker clustering for speech recognition using vocal-tract parameters," Speech Commun., vol. 36, no. 3, pp. 305-315, 2002.

[5] H. Levitt, J. M. Pickett, and R. A. Houde, (Eds.), "Speech training aids," part VII in Sensory Aids for the Hearing Impaired. New York: IEEE Press, 1980, pp. 349-419.

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

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

[8] B. S. Atal, "Determination of the vocal tract shape directly from the speech wave," J. Acoust. Soc. Am., vol. 47(A), p. 64, 1970.

[9] 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, 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] P. C. Pandey and M. S. Shah, "Estimation of place of articulation during stop closures of vowel-consonant-vowel utterances," IEEE Trans. Audio, Speech, Language Process., vol. 17, no. 2, pp. 277-286, 2009.

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

[13] M. S. Shah, "Estimation of place of articulation during stop closures of vowel-consonant-vowel syllables," Ph.D. thesis, Dept. Elect. Eng., Indian Inst. of Technology Bombay, India, 2008.

[14] S. A. Liu, "Landmark detection for distinctive feature based speech recognition," J. Acoust. Soc. Am., vol. 100, pp. 3417-3430, 1996.

[15] C. Glaser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and Bayesian estimation for robust formant tracking," IEEE Trans. on Audio, Speech, and Language Processing, vol. 18, no. 2, pp. 224–236, 2010.

[16] J. H. Hansen, S. S. Gray, and W. Kim, "Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification," Speech Commun., vol. 52, pp. 777–789, 2010.

[17] A. L. Francis, V. Ciocca, and J. M. C. Yu, "Accuracy and variability of acoustic measures of voicing onset," J. Acoust. Soc. Am., vol. 113, no. 2, pp.1025–1032, 2003.

[18] 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.

[19] K. S. Nataraj, Jagbandhu, P. C. Pandey, and M. S. Shah, "Improving the consistency of vocal tract shape estimation," in Proc. National Conf. Commun. (NCC), 2011, Bangalore, India, paper SpPrII.4.

[20] 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.

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