ESTIMATION OF VOCAL TRACT SHAPE DURING STOP CLOSURES

Milind S. Shah
Department of Electrical Engineering
IIT Bombay
Powai, Mumbai 400 076, India
(+91-22) 2576 4417

milind@ee.iitb.ac.in

Prem C. Pandey
Department of Electrical Engineering
IIT Bombay
Powai, Mumbai 400 076
(+91-22) 2576 7445

pcpandey@ee.iitb.ac.in

ABSTRACT

The display of intensity, pitch, and vocal tract shape is considered to be helpful in speech training of the hearing impaired. Vocal tract shape estimation, based on LPC and other analysis techniques, works satisfactorily for vowels but generally fails during stop consonants due to very low signal energy. Our estimation of vocal tract shape is based on reflection coefficients obtained from LPC analysis of speech. In order to study the consistency and dynamics of the vocal tract shape estimation with amplitude variation of vowels, and to study the transitions at vowel-consonant boundaries, we have used "areagram", a spectrogram-like two-dimensional display of estimated vocal tract cross-sectional area as a function of time and position along the tract length. Display of areagram for vowel-consonant-vowel (V-C-V) syllables is satisfactory for vowel part, but during stop closure, area values become random. Based on estimated vocal tract cross sectional area before and after stop closure, conic surfaces are obtained and used for estimation of vocal tract shape during stop closure by a two-dimensional interpolation.

Keywords

Vocal tract shape estimation, speech training aids for hearing impaired children, place of stop closure.

1. INTRODUCTION

In children with normal hearing, the process of learning to speak is aided by auditory feedback. The hearing impaired children lack this feedback and therefore experience difficulty in acquiring normal speech characteristics. Thus, in spite of proper speech production mechanism, they may not be able to produce intelligible speech. It is possible to teach such persons to speak by use of appropriate feedback. Visual feedback can be provided by display of certain speech parameters, which are easily controllable by the person undergoing speech training.

These parameters should provide necessary cues for uttering a specific speech segment. Vocal tract shape, lip shape, pitch, and energy variations can be used for this, since these are directly linked with speech production.

Efforts have been made to develop speech training aids involving spectrographic displays, display of formant tracks, voicing and pitch variations, palatograph, nasal vibration etc. [3, 7, 13, 14, 28, 32, 33]. In order to provide information on articulatory efforts, speech training aids have been reported for displaying vocal tract shape [8, 10, 15, 18, 19, 25]. In the literature, various techniques like measurement of formant frequencies, measurement of acoustic impedance at the lips, LPC analysis, neural network based system, and magnetic resonance imaging have been reported for the estimation of vocal tract shape [2, 8, 9, 11, 12, 15, 16, 17, 18, 19, 22, 23, 25, 26, 27, 29, 30, 31].

Watanabe et al [32] developed a speech visualization system which extracts the consonantal features using neural network and creating visual images by adding all the consonantal patterns whose brightness is controlled by the strength of the extracted phonemic features. In another development, a speech training aid, which integrates acoustic and several type of instrumentally measured articulatory data like palatograph, nasal vibration, airflow, and presence/absence of voicing has been reported in [7]. PC based speech training systems developed by Tiger DRS, Inc. [3] and Video Voice [28] use different games and cartoons based on pitch, loudness, formants etc.

Mahdi [15] has reported a system for visualization of the vocal tract shapes during vowel articulation. The vocal tract cross sectional areas and first three formant frequencies are estimated by LPC analysis and estimated areas are mapped to corresponding mid-saggital distances for displaying 2-D vocal tract lateral graphics. To compensate for possible errors in the estimated area function due to variations in vocal tract length, the first two section distances are determined by the three formants. The system reported by Park et al [19] displays intensity, fundamental frequency, and nasality along with vocal tract shape. Fundamental frequency and nasality are detected using separate vibration sensors. Vocal tract area function from lips to glottis is found using Wakita's method [29], supported by lip-to-lip distance (amount of lip opening) found from first three formant frequencies as given by [9]. Successful training of deaf children for five Korean vowels was reported. Language Vision Inc. [10] has developed a PC based training system which can estimate the vocal tract shape for vowels in real time and provide speech

articulation training by dynamically displaying estimated vocal tract shape by moving mid-saggital head.

Thus the vocal tract shape estimation techniques are based on measurement of formant frequencies, LPC analysis, neural networks etc. These are reported to work satisfactorily for vowels but have a limitation in vocal tract shape estimation for consonants. Our investigation of vocal tract shape estimation is based on reflection coefficients obtained from LPC analysis of speech. In order to study the consistency of the shape estimation with amplitude variation in vowels, and to study dynamics of shape estimation during transitions at V-C boundaries, we have used "areagram", a two-dimensional display of vocal tract cross-sectional area as a function of time and position along the tract length. Display of areagram for vowels and semivowels is satisfactory but for V-C-V syllables, area values become random during stop closure, due to very low signal energy. It is to be noted that the indication of place of closure is critical for the success of the system as a speech training aid. In this paper we propose a new method for shape estimation during stop closure by obtaining conic surfaces based on estimated area values before and after stop closure, and then performing interpolation to get the area values during stop closure.

2. INVESTIGATION OF VOCAL TRACT SHAPE ESTIMATION

Our implementation of vocal tract shape estimation is based on reflection coefficients obtained from LPC analysis of speech signal [21]. Wakita's speech analysis model [29] is used for the estimation of vocal tract area, and Robinson's algorithm [29] for optimum inverse filtering is implemented. Order of linear predictor is chosen to be 12. The 12-section area function is interpolated to 176 points by Beizer form algorithm [5] to obtain a more realistic display. The length of the vocal tract is assumed to be 17 cm. The pitch is estimated by short time autocorrelation method [4], and energy as zeroth autocorrelation coefficient.

Difficulties have been earlier experienced in vocal tract shape estimation algorithm implementation using fixed-point real-time processing due to recursive errors in computation and dynamic range limitation [8]. For the investigations reported here, the algorithm is implemented with floating-point arithmetic.

In order to study the consistency and dynamics of the vocal tract shape estimation during vowels, and to study the transitions at vowel-consonant boundaries, we have used "areagram", a spectrogram [21] like two dimensional display of vocal tract area with time and lip to glottis distance. Time is plotted along x axis; y axis represents distance from lip to glottis. Each new vertical frame corresponds to shifting the analysis window by 128 samples ~ 11.6 ms., and each horizontal frame corresponds to normalized distance of 1/170 ~ 0.0057. Minimal opening is represented by black and maximal opening is represented by white. It is to be noted that the areagram is meant only for studying the vocal tract shape estimation, and will not be useful for speech training. For actual speech training, appropriate displays, cartoon, or games based on dynamically varying vocal tract shape, will have to be devised and tested.

The speech signal is recorded with PC sound card at a sampling rate of 11.025 k Sa/s. Each analysis frame consists of 256 samples, and successive frames are positioned with 50% overlap.

For processing and display, package "VTAG-1" is developed in Matlab for displaying speech waveforms (recorded signal and selected segment), pitch and energy contours, spectrogram, and areagram. Recorded speech signal, selected speech segment, and pitch/energy contours are displayed in the left half of the screen while spectrogram and areagram for selected segment is displayed in the right half. The lower and upper range of various scales can be modified with graphical user interface menu in order to select the most appropriate dynamic range.

The package was tested with natural and synthesized speech signals for consistency and validity of the estimates. The areagram results were found to be satisfactory for vowels and semivowels. The areagram display for V-C-V syllables is proper for vowel segments but is random during stop closure [24].

Figs. 1 and 2 show the time waveform, spectrogram and areagram for the sequence of synthesized and natural vowels /a/, /i/, /u/, respectively. We observe that the shape of the vocal tract and place of tongue elevation for the three vowels is indicated in the areagram.
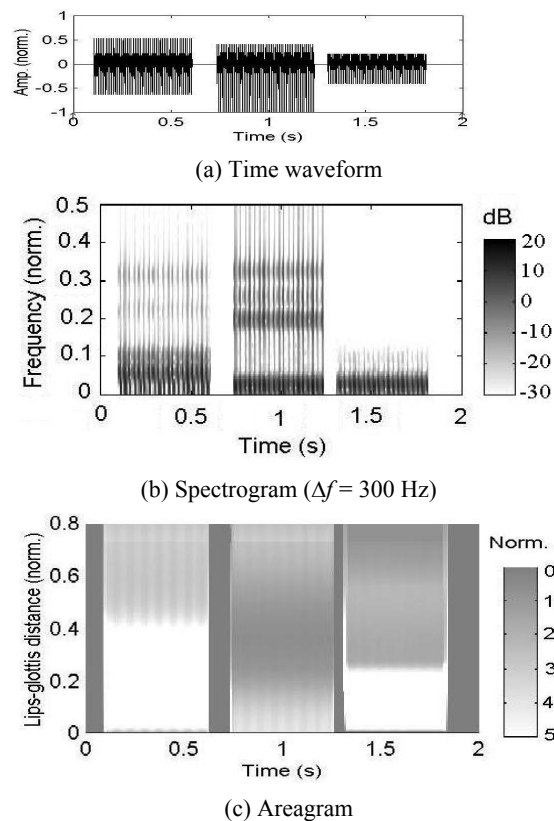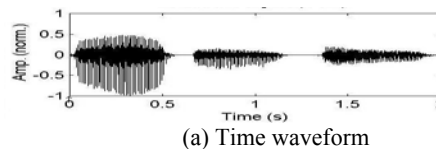

(a) Time waveform

(b) Spectrogram (Δf = 300 Hz)

(c) Areagram

Fig. 1 Spectrogram and areagram for synthesized vowels /a/, /i/, /u/ (Fo = 100 Hz).

(a) Time waveform

(b) Spectrogram (Δf = 300 Hz)
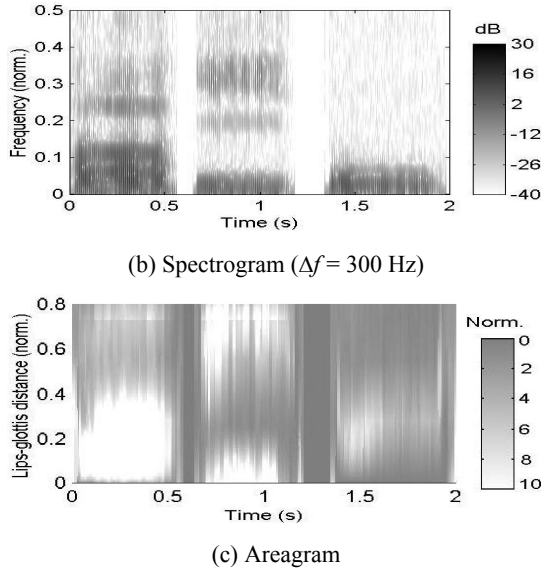

(c) Areagram

Fig. 2 Spectrogram and areagram for natural vowels /a/, /i/, /u/, speaker: male.

The algorithm for vocal tract shape estimation fails when the signal intensity is very low. Experimentation was carried out on synthesized and natural sustained vowels for obtaining minimum signal energy below which the estimation of vocal tract shape becomes inconsistent. For this study, vowels were recorded as well as synthesized, using 16-bit quantization and subsequently scaled to floating point in the ±1 range. All of these segments had rms value of approximately 0.2. The waveforms were scaled, and the lowest rms value that gave acceptable vocal tract shape estimation was noted. The speech signal amplitude scaling was carried out, as if they were scaled by factor "k" before quantization. The speech record of 2 s duration was scaled in time steps of 0.2 s with k = 1, 1/2, 1/10, 1/20, …, 1/20000. Fig. 3 shows areagram result for synthesized vowel /a/. We can observe that vocal tract shape estimation is proper up to 1.4 s duration corresponding to rms value of 0.101×10^{-3}. In general it was observed that the vocal tract shape estimation works satisfactorily for various sustained vowels over an attenuation range of 0–40 dB.

After studying areagram of amplitude scaled vowels, we carried out investigations for semivowels, which are weaker phonemes than vowels. Various vowel–semivowel–vowel syllables were recorded and analyzed for areagram results. LPC analysis was used for obtaining spectral envelope (by removing harmonic structure due to glottal excitation), and the smoothened spectra were plotted as spectrogram in order to study transitions of vocal tract resonances. The result for /aja/ is shown in Fig. 4. We can observe that LPC smoothened spectrogram shows proper transition of vocal tract resonances and areagram result is satisfactory as the place of constriction (in this case palatal /j/) is properly reflected in the areagram.

(a) Scale factors (dB)

(b) Amplitude scaled speech waveform

(c) Areagram

Fig. 3 Areagram analysis results for synthesized vowel /a/ with amplitude scaled over 0-86 dB range.

(a) Time waveform

(b) Spectrogram (Δf = 300 Hz)

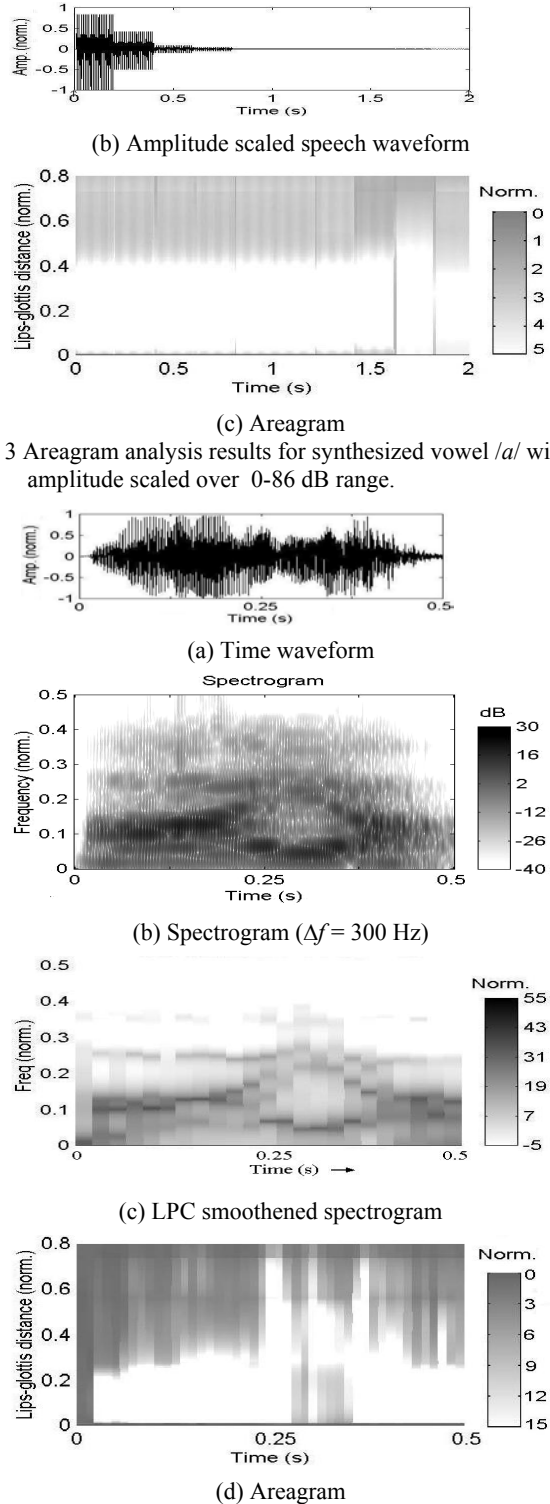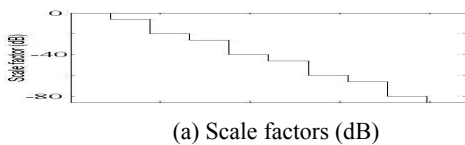(c) LPC smoothened spectrogram

(d) Areagram

Fig. 4 Areagram analysis results for /aja/

Next, investigation was carried out for vocal tract shape estimation for stop consonants, which involve very low signal energy during closure duration. Fig. 5 shows areagram for V-C-V syllables /aba/ and /aga/. It is observed that areagram is satisfactory during vowel segment. However, during stop closure, area values are random. For various V-C-V syllables, areagram

results during transition boundary from vowel to stop closure and stop closure to vowel were distinctly different. Hence we carried out further investigations for predicting tract shape during stop closure from the shapes estimated on either side of the closure.
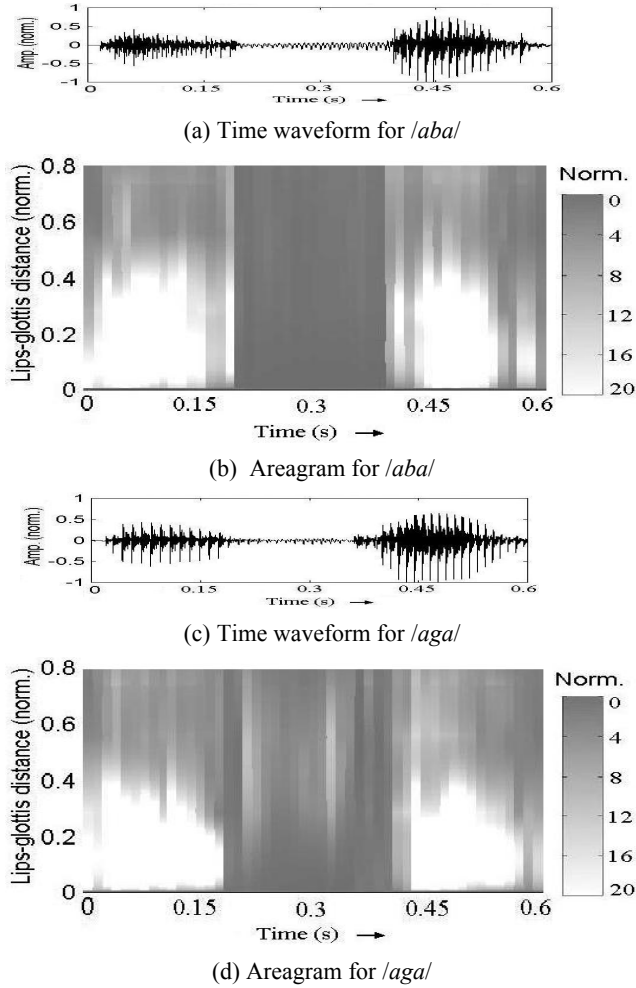

(a) Time waveform for /aba/

(b) Areagram for /aba/

(c) Time waveform for /aga/

(d) Areagram for /aga/

Fig. 5 Areagram analysis results for V-C-V- syllables /aba/ and /aga/, speaker: male.

3. ESTIMATION DURING STOP CLOSURES

It is observed that the area information during the stop closure can not be obtained from the speech waveform by LPC based analysis. However, the area estimates during the transition segments may provide information about the place of closure. In the raw form, this can not be used for speech training, because a hearing impaired child's visual process can not get the place of closure from the transition in the vocal tract shape. It needs to be estimated and displayed. We have investigated use of polynomial two-dimensional surface fitting over the transition segment for estimating the area during the closure segment. We have used a conic surface approximation in order to avoid the problem of unwanted oscillatory behavior arising with higher degree polynomials [6]. Thus, we approximate the areagram with the conic equation

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

where $f(x, y)$ = estimated area at analysis frame 'x' (along time axis) and lip-glottis distance 'y'. This equation for set of q points, with $q > 6$, results in over determined system of simultaneous linear equations expressed in matrix notation as

$$\mathbf{Az = B} \qquad (2)$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & x_0 & y_0 & x_0 y_0 & x_0^2 & y_0^2 \\ 1 & x_1 & y_1 & x_1 y_1 & x_1^2 & y_1^2 \\ M & M & M & M & M & M \\ 1 & x_{q-1} & y_{q-1} & x_{q-1} y_{q-1} & x_{q-1}^2 & y_{q-1}^2 \end{bmatrix}$$

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

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

The values in matrix $\mathbf{B}$ consist of areagram values in vowel to stop closure transition region (VC transition) and closure to vowel transition region (CV transition). The transition regions and one possible way of area value selection is shown in Fig. 6. Value of x equal to m_0 corresponds to lower limit of start transition along x direction and value of y equal to n_1 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 approximation are

$$m_0 \leq x \leq m_1 \quad \text{and} \quad m_2 \leq x \leq m_3$$

$$n_1 \leq y \leq n_2$$

In order to evaluate the coefficients in (1), we need to have $j = n_2 - n_1 \geq 2$. For plotting areagram, the 12-section area values are interpolated to 176 points, and hence
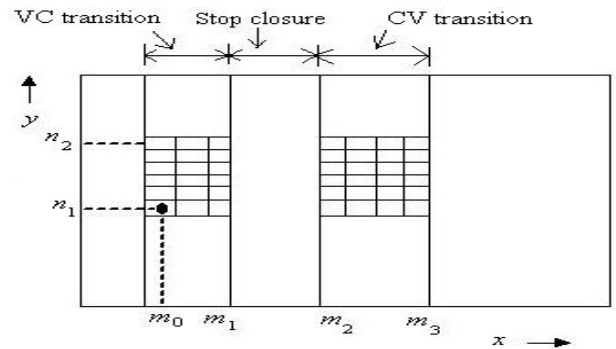
$$0 \leq n_1 \leq 175 - j$$

Fig. 6 Transition regions and area value selection.

The unknown coefficient matrix $\mathbf{z}$ can be obtained by using least square fit, which minimizes the sum of squares of the deviations of the data from the model. This least square solution [1, 20] can be obtained from equation (2) by performing pseudo inverse operation on matrix $\mathbf{A}$

$$\mathbf{z = (A^T A)^{-1} A^T B} \qquad (3)$$

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

In the articulatory movement for V-C-V syllables, the dynamic shape of vocal tract during transition from V-C and C-V is based

on vowels preceding and following stop closure, and the actual place of stop closure. We have carried out investigations for shape estimation during stop closure by obtaining conic surfaces based on estimated area values during V-C transition and C-V transition, and then performing forward and backward extrapolation during stop closure respectively, using (1). Similarly, we have investigated shape estimation by obtaining conic surfaces based on estimated area values during both V-C and C-V transition, and then performing interpolation using (1) for all the *x*, *y* values during stop closure.

4. RESULTS

Fig. 7 shows areagram results obtained after performing least square fitting of conic surface and extrapolation during stop closure for V-C-V syllable /*aga*/. Part (a) of the Fig.7 shows result when only area values from VC transition region were used for surface fitting operation while for part (b) only CV transition values of areagram were used. These results are for *j* =5.
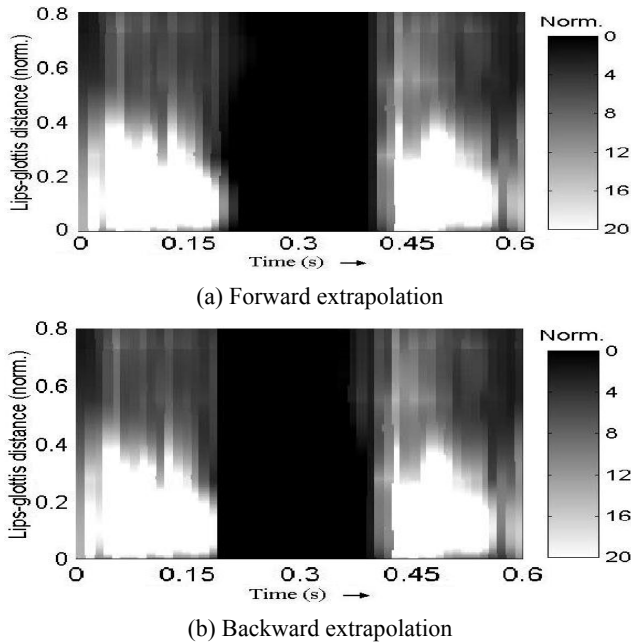

(a) Forward extrapolation

(b) Backward extrapolation

Fig. 7 Areagram results for /*aga*/ after extrapolation (*j* = 5).

We can observe that forward and backward extrapolation of estimated surfaces for shape estimation during stop closure results in having monotonically decreasing surface slope of very large value, and these have to be limited to saturation values.

Fig. 8 shows areagram results after performing conic surface fitting using both VC and CV transition region area values, and then interpolating during stop closure for V-C-V syllables /*aba*/ and /*aga*/. We can see that the process of interpolation has helped in estimating vocal tract shape and place of closure for consonant /*b*/ and /*g*/. Both the results were obtained for *j* = 20.

(a) Areagram for /*aba*/
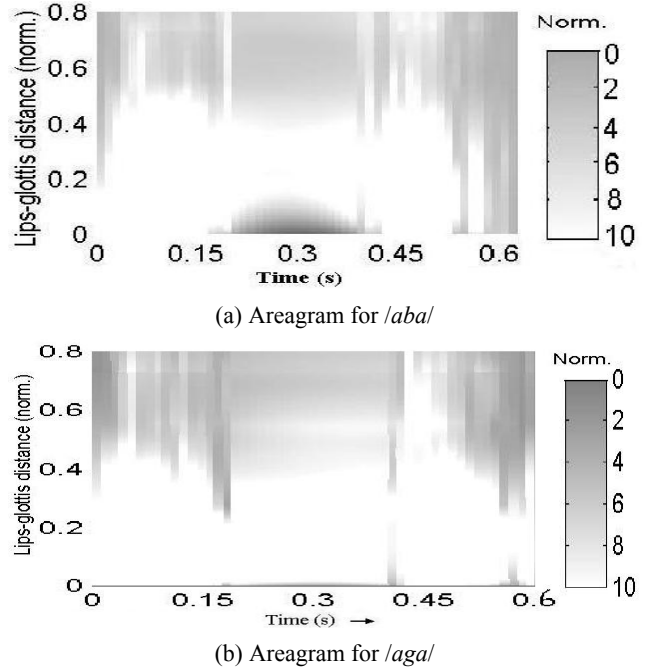

(b) Areagram for /*aga*/

Fig. 8 Areagram results after interpolation.

5. CONCLUSIONS

We carried out various investigations for vocal tract shape estimation based on reflection coefficients obtained from LPC analysis of speech signal. In order to study the consistency and dynamics of the vocal tract shape estimation with amplitude variation of vowels, and to study the transitions at vowel-consonant boundaries, we have used "areagram", a two-dimensional display of vocal tract area with time and lips-to-glottis distance. The areagram results are satisfactory for vowels, and semivowels, but during stop closure when the signal energy is very low, they fail to estimate the vocal tract shape. Investigation with speech amplitude scaling of synthesized and natural vowels with 16-bit equivalent quantization and rms value of 0.2 revealed that our vocal tract shape estimation algorithm works satisfactorily for the vowels over an attenuation range of 0-40 dB.

From the areagram results for various V-C-V syllables, it was observed that areagram during vowel to stop closure and stop closure to vowel transition regions are distinctly different. Based on estimated area before and after stop closure, conic surfaces are obtained by least square method and are used for estimation of vocal tract shape during stop closure by interpolation. Surface fitting using only either V-C transition or C-V transition region and then its extrapolation for stop closure duration did not help in shape estimation during stop closure. But, surface fitting based on area values from both V-C and C-V transition region and then interpolation during stop closure could estimate vocal tract shape as well as place of closure in stop consonants.

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.

6. REFERENCES

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

[2] DeLucia, J., and Kochman, F. A new non iterative algorithm for computing acoustically constrained vocal tract area functions. IEEE Trans. Speech Audio Processing, 8, 2 (2000), 177-183.

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

[4] Dubnowski, J. J., Schafer, R. W., and Rabiner, L. R. Real-time digital hardware pitch detector. IEEE Trans. Acoust., Speech, Signal Processing, 24, 1 (1976), 2-8.

[5] Foley, J. D., and Vandam, A. Fundamentals of Interactive Computer Graphics. Addison- Wesley , New York, 1983.

[6] Hocking, R. P. Methods and Applications of Linear Models: Regression and the Analysis of Variances, 2nd ed. Wiley, Hoboken, New Jersy, 2003.

[7] Javkin, H., Barroso, N. A., Das, A., Zerkle, D., Yamda, Y., Murata, N., Levitt, H., and Youndelman, K. A motivation-sustaining articulatory/acoustic speech training system for profoundly deaf children. In Proc. ICASSP '93, (1993), 145-148.

[8] Kshirsagar, S. A. A Speech Training Aid for Hearing Impaired. M.Tech. Dissertation (Supervisor: P. C. Pandey), Dept. of Electrical Engg., IIT Bombay, 1998.

[9] Ladefoged, P., Harshman, R., Goldstein, L., and Rice, L. Generating vocal tract shapes from formant frequencies. J. Acoust. Soc. Am., 64, 4 (1978), 1027-1035.

[10] Language Vision, Inc., http://www.langvision.com, Idaho Falls, ID, 2003.

[11] Laprie, Y. and Mathieu, B. A variational approach for estimating vocal tract shapes from the speech signal. In Proc. ICASSP '98, (1998), 929-932.

[12] Levinson, S. E., and Schmidt, C. E. Adaptive computation of articulatory parameters from the speech signal. J. Acoust. Soc. Am., 74, 4 (1983), 1145-1154.

[13] Levitt, H. Speech processing aids for the deaf: an overview. IEEE Trans. Audio Electroacoust., 21, 3 (1973), 269-273.

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

[15] Mahdi, A. E. Visualization of the vocal tract based on estimation of vocal area functions and formant frequencies. In Proc. Eurospeech, (Geneva , 2003), 2381-2384.

[16] Milenkovic, P. Vocal tract area functions from two-point acoustic measurements with formant frequency constraints. IEEE Trans. Acoust., Speech, Signal Processing, 32, 6 (1984), 1122-1135.

[17] Paige, A. Computation of vocal tract area functions. IEEE Trans. Audio Electroacoust., 18, 1 (1970), 7-18.

[18] Pardo, J. M. Vocal tract shape analysis for children. In Proc. ICASSP '82, (1982), 763-766.

[19] Park, S. H., Kim, D. J., Lee, J. H., and Yoon, T. S. Integrated speech training systems for hearing impaired. IEEE Trans. Rehab. Engg., 2, 4 (1994), 189-196.

[20] Pratt, V. Direct least-square fitting of algebraic surfaces. Computer Graphics, 21, 4 (1987), 145-152.

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

[22] Schroeder, M. R. Determination of the geometry of the human vocal tract by acoustic measurement. J. Acoust. Soc. Am., 41, 4 (1967), 1002-1010.

[23] Schroeter, J., and Sondhi, M. M. Techniques for estimating vocal-tract shapes from the speech signal. IEEE Trans. Speech Audio Processing, 2, 1 part II (1994), 133-150.

[24] Shah, M. S., and Pandey, P. C. 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), 121-124.

[25] Shigenaya, M., and Kubo, H. Speech training systems for handicapped children using vocal tract lateral shapes. In Proc. ICASSP '86, (1986), 637-640.

[26] Sondhi, M. M. Estimation of vocal-tract areas: the need for acoustical measurement. IEEE Trans. Acoust., Speech, Signal Processing, 27, 3 (1979), 268-273.

[27] Story, B. H., and Titze, I. R. Vocal tract area functions from magnetic resonance imaging. J. Acoust. Soc. Am., 100, 1 (1996), 537-554.

[28] Video Voice Speech Training System, http://www. videovoice.com, Micro Video Corp., Ann Arbor, Michigan, 2003.

[29] Wakita, H. Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveform. IEEE Trans. Audio Electroacoust., 21, 5 (1973), 417-427.

[30] Wakita, H. Estimation of vocal tract shapes from acoustical analysis of the speech wave: the state of the art. IEEE Trans. Acoust., Speech, Signal Processing, 27, 3 (1979), 281-285.

[31] Wakita, H., and Gray, A. H. Numerical determination of the lip impedance and vocal tract area functions. IEEE Trans. Acoust., Speech, Signal Processing, 23, (1975), 574-580.

[32] Watanabe, A., Tomishige, S., and Nakatake, M. Speech visualization by integrating features for the hearing impaired. IEEE Trans. Speech Audio Processing, 8, 4 (2000), 454-466.

[33] Zahorian, S. A., and Venkat, S. Vowel articulation training aid for the deaf. In Proc. of ICASSP '90, (1990), 1121-1124.