Proc. 17th National Conference on Communications (NCC 2011), Bangalore, 28-30 January 2011, Paper No. Sp Pr. II, P4

Improving the Consistency of Vocal Tract Shape Estimation

K. S. Nataraj, Jagbandhu, P. C. Pandey
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai 400076, India
Email: {natarajks, jagbandhu, pcpandey}@ee.iitb.ac.in

M. S. Shah
Department of Electronics & Telecomm. Engineering
Fr. C. Rodrigues Institute of Technology
Vashi, Navi Mumbai 400703, India
Email: milind05in@yahoo.co.in

***Abstract*— Estimation of the vocal tract shape has applications in articulatory synthesis, speech recognition, and speech-training aids. LPC based analysis can be used to obtain the vocal tract shape during speech segments produced with glottal excitation and with fixed as well as transitional tract configurations. During the stop closures of vowel-consonant-vowel (VCV) utterances, the shape can be estimated by bivariate surface modeling of the shapes during transition segments. The shape obtained by LPC analysis of steady vowels shows variability with the position of the analysis frame. Low-pass filtering of the shapes across the frames for improving the consistency cannot be used during transition segments. A windowed energy index is calculated as the ratio of the energy of the windowed signal to the frame energy, and it is shown that the shapes in the frames corresponding to the valleys in this index have a reduced variability. Thus the selection of the frames based on this index can be used for improving the consistency of vocal tract shape estimation for various applications.**

***Keywords*- vocal tract shape; speech training aid; linear predictive coding.**

I. INTRODUCTION

Estimation of the vocal tract shape has applications in articulatory synthesis of speech [1]-[3], speech recognition [4]-[6], and visual speech-training aids [7]-[9]. Hearing-impaired children have great difficulty in acquiring the ability to control position and movements of various articulators due to the lack of auditory feedback. Several speech training aids have been developed to dynamically display important acoustic parameters (such as speech intensity, voicing and pitch, spectral features) [7]. Speech-training aids providing visual feedback of articulatory motion are found to be useful in teaching lingual consonants and vowels [8],[9].

Estimation of the vocal tract shape consists of recovering the sequence of vocal tract cross-sectional area values that produce a given acoustic speech signal. It can be carried out directly from the speech signal using several methods, including linear predictive coding (LPC) [10],[11], formant estimation [12], and articulatory codebook mapping [13]. Despite its limitations, Wakita's LPC based method [11] for direct estimation of the vocal tract shape has been widely used as it is suitable for real-time processing [8]. In this method, the vocal tract is modeled as a lossless acoustic tube with sections of equal length and varying area of cross-section. For plane wave propagation through the acoustic tube, reflections occur at the section interfaces due to different areas on the two sides. The area ratios are calculated from the reflection coefficients obtained from the autocorrelation coefficients of the signal by LPC analysis. The method can be used for estimating fixed as well as transitional vocal tract configurations during speech segments with glottal excitation which can be modeled as produced by an all-pole vocal tract filter. Some of the limitations of the LPC-based estimation of vocal tract shape, as outlined by Wakita [14], are (i) related to the errors in estimation of vocal tract transfer function from the band-limited speech signal, (ii) due to uncertainty in glottal source characteristics, and (iii) due to lack of a method for estimating the scaling factor for converting the area ratios into area values during dynamically varying vocal tract configuration. The method is not applicable during nasalized vowels, nasal stops, and fricatives. It cannot be used for estimating the shape during the stop closures due to very low signal energy. Pandey and Shah [15] proposed a method to estimate the vocal tract shape during the stop closures of vowel-consonant-vowel (VCV) utterances by bivariate surface modeling of the shapes during transition segments. The accuracy of this method depends on the accurate location of the transition segments and accuracy of the estimation of dynamically changing shapes.

LPC based vocal tract shape estimation involves applying Hamming window on the analysis frames of the speech signal, with a short inter-frame interval to track the vocal tract shape changes. The estimated area values are found to vary even for the fixed tract configurations, thus showing an inconsistency. During vowels with fixed tract configurations, low-pass filtering of the estimated area values across the frames can be used for improving the consistency. The variability in the estimation can also be reduced by increasing the length of the analysis frame and by using frame length equal to a multiple of the pitch period. But these methods cannot be used during transitional configurations, e.g. diphthongs, semivowels, and vowel-consonant and consonant-vowel transitions. Hence other means for improving the consistency of the estimated shapes without smearing the transitions are needed.

Rabiner *et al.* [16] observed that LPC prediction error varied substantially with the position of the analysis frames, independent of the analysis method. They proposed two pre-processing methods to reduce the variation in the prediction error: all-pass filtering and pre-emphasis of the speech signal. Although these methods reduced the variability, they increased the prediction error. They also reported that speech synthesized with LPC coefficients obtained from speech frame with maximum prediction error was more nasal-like than the one synthesized from the coefficients obtained from speech frame with the minimum prediction error. It has been reported that the peaks of the prediction error are not always prominent [17]. Hence they cannot be used for selecting frames for improving the estimation of the coefficients. Mizoguchi *et al.* [18] showed that by selecting samples in time domain with prediction error less than a threshold, the variation in the prediction coefficients across the frames was reduced for steady state vowels. Ma *et al.* [19] showed that short time energy based selection of samples was more robust than the LPC prediction error based selection. Mezzalama [20] reported that large errors were introduced in the LPC-based formant estimation as the window position shifted from the glottal pulse and the error could be reduced by repeatedly concatenating the segment in the frame before applying the window. In this method, repeating the frames introduces error in the analysis if the frame length is not equal to the multiple of the pitch period.

An examination of the LPC analysis results for steady-state voiced segments showed that the variations in the vocal tract shape and in the prediction error both were related to the windowed signal energy. Here, we present a method for selecting the frames, based on the windowed energy, for improving the consistency of the estimation without smearing the variations during transitional vocal tract configuration.

II. METHOD

A. Variation in estimated vocal tract shape

For estimation of vocal tract shape using Wakita's method [11], the speech signal was sampled at 10 kHz with 16-bit quantization and a pre-emphasis was applied by taking its first difference. The LPC analysis of order 12 was carried out on analysis frames of duration equal to twice the average pitch period after applying Hamming window. In this analysis, the vocal tract length from glottis to lips was modeled as 12 cylindrical sections of equal length and the area values of the sections were calculated from the reflection coefficients obtained using the autocorrelation coefficients of the windowed frame. Analysis of speech signals of steady-state vowels with an inter-frame interval of 5 ms showed a significant variability in the area values for most of the sections. The variation in the area values may be attributed to the natural variation during phonation or to the errors related to the frame position with respect to the instants of glottal closure.

To eliminate the variations contributed by natural changes in phonation, the analysis was carried out on the vowel sequence / -a-i-u- / synthesized using Klatt synthesizer [21] with a constant pitch frequency of 90 Hz and a constant amplitude. To study the effect of the position of the analysis frame, analysis was carried out for inter-frame interval of one sample. The vocal tract shape as a function of time was displayed using areagram, a two-dimensional display of square root of vocal tract area values plotted as gray levels as a function of frame position along x-axis and glottis-to-lips (G-L) distance along y-axis. The area values of the twelve sections were converted to 40 values using cubic-spline interpolation and square root of these values were used for plotting the areagram [22]. Fig. 1 shows, for the vowel sequence /-a-i-u-/, the speech waveform, the wide-band spectrogram, and the areagram. The estimated area values are in accordance with the respective places of articulation for the three vowels: mid constriction (nearly neutral) for /a/, front constriction for /i/, and back constriction for /u/. However, there is a large variation in the area values as a function of time and the variations are related to the position of the analysis frame.

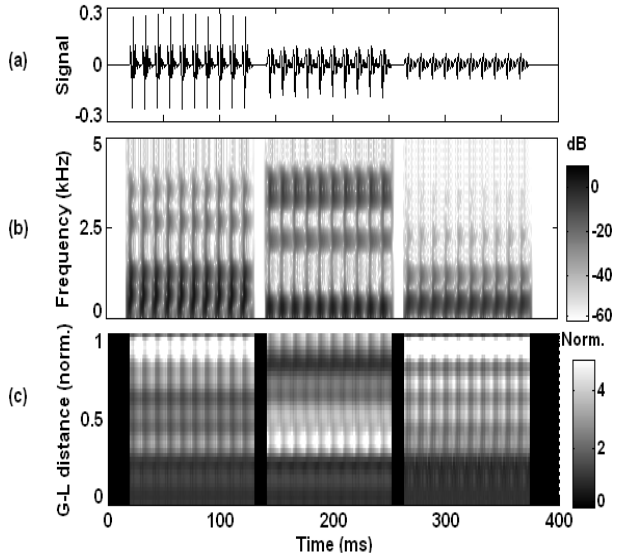

Figure 1. Vocal tract shape estimation of the synthesized vowel sequence /-a-i-u-/: (a) speech waveform, (b) wideband spectrogram, (c) areagram.

B. Windowed energy index for improving the consistency of the estimated area values

A plot of RMS value of the LPC prediction error showed a large variation with the analysis frame position, with the peaks occurring for the analysis frames coinciding with the starting of the glottal pulses. The frame positions corresponding to the minimum in the prediction error are likely to be related to the least estimation error in the vocal tract parameters. However, locating the peaks or the valleys of the LPC prediction error consistently was found to be difficult. In the speech signal, a glottal closure instant (GCI) marks the beginning of the glottal excitation. The GCIs were obtained using Childers and Hu's algorithm [23] and it was found that the variation in the prediction error was related to the GCIs. However, the location of the frame positions for minimum error with respect to the GCIs was found to be different for the three vowels, making it

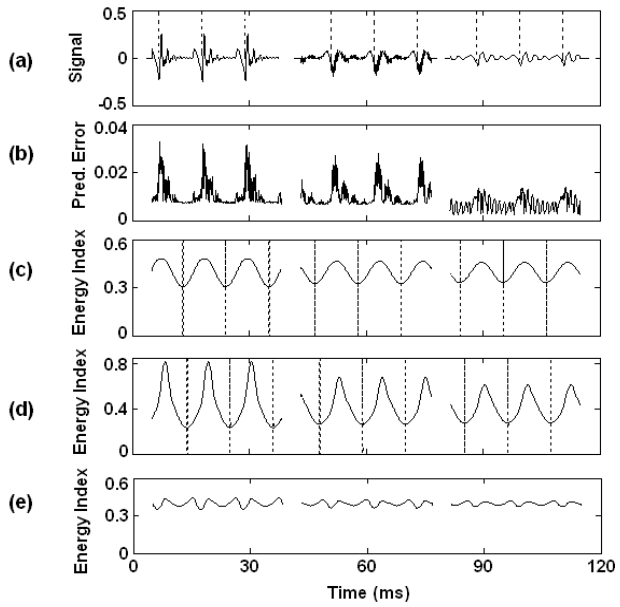

Figure 2. Speech waveform, windowed energy index, and RMS prediction error for three glottal cycles each of synthesized vowels /a/, /i/, /u/: (a) speech waveform with GCI marked by dotted lines, (b) RMS prediction error, (c) windowed energy index for window length of two pitch periods, with the minima marked as dotted lines, (d) windowed energy index with window length decreased by 10 %, (e) windowed energy index with window length increased by 10 %.

difficult to use the GCI locations for automated selection of frame positions for minimizing the estimation error.

The LPC analysis using autocorrelation method involves application of Hamming window on the analysis frame. The prediction error was found to be related to the energy of the windowed frame, with the lowest values of the prediction error coinciding with the minima of the windowed energy for all the three vowels. For automating the selection of the frames with the minimum energy, a function called "windowed energy index" was calculated as the ratio of the energy of the windowed signal to the frame energy, and given as

$$E_w(n) = \sum_{m=0}^{N-1} [s_n(m)w(m)]^2 \Big/ \sum_{m=0}^{N-1} s_n{}^2(m) \qquad (1)$$

where $w(m)$ is the Hamming window function of length N samples and $s_n(m)$ is the speech segment for the frame position n. For steady state segments and frame length equal to an exact multiple of the pitch period, the function is periodic with period equal to the pitch period. It varies smoothly with easily detectable maxima and minima.

Fig. 2 shows the synthesized vowel sequence /-a-i-u-/, the windowed energy index, and the RMS prediction error, plotted as a function of the analysis frame position. The GCI positions are marked on the speech waveform. Natural speech has significant jitter in pitch period. To examine the effect of the window length with respect to the pitch period, windowed energy index function was calculated for the window length equal to two pitch periods, and for window lengths decreased and increased by 10 %, as shown in Fig. 2. The minima of the function, located using valley detection, are also marked. The

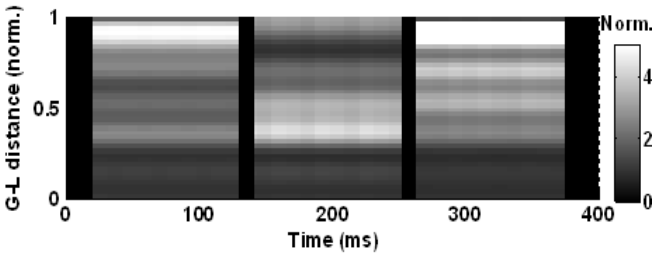

Figure 3. Areagram of the synthesized vowel sequence /-a-i-u-/ as obtained for E_w–minima selected frames

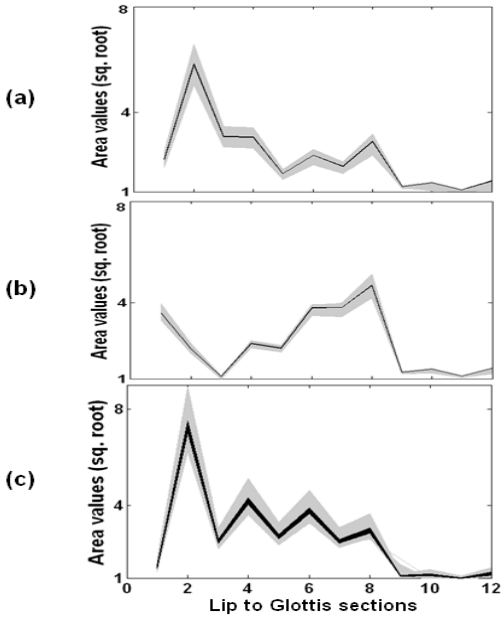

Figure 4. Plot of area values for synthesized vowels: (a) /a/ (b) /i/ (c) /u/ (shaded line: original area values, solid dark lines: area values for E_w–minima selected frames).

function obtained with the window length of two pitch period has distinct minima and these correspond to the low values of the prediction error for all the three vowels. The function with the decreased window length also has distinct minima corresponding to the low values of prediction error, but its shape is different for the three vowels. Increased window length results in a function with relatively indistinct minima. Thus the function can be used for locating the frame positions corresponding to low prediction error by using a window length equal to two pitch periods or slightly shorter. The technique was applied for estimating the vocal tract shape from speech signals of synthesized and natural vowels and vowel-semivowel-vowel utterances.

III. Results And Discussion

In order to reduce the variability in the estimated area values, the values were calculated at the frame positions corresponding to the E_w minima. Fig. 3 shows the resulting areagram for the synthesized vowel sequence /-a-i-u-/. The estimated vocal tract shape for all the three vowels shows much smaller variation as compared to that in the areagram in Fig.1. For a visual comparison of the spread in the area values, the plots of the estimated values for the twelve sections for all

TABLE I. MEAN AND MAX-MIN DEVIATION OF SQUARE-ROOT OF VOCAL TRACT AREA VALUES OF 12 SECTIONS FOR THREE VOWELS.

(A) Inter-frame interval of 1 sample

Vowel	Parameter	Lips-to-glottis section number											
		1	2	3	4	5	6	7	8	9	10	11	12
/a/	mean	2.26	5.94	3.14	3.12	1.70	2.40	1.97	2.90	1.20	1.33	1.07	1.38
	m-m	0.61	1.56	0.79	0.78	0.40	0.60	0.46	0.82	0.12	0.37	0.08	0.41
/i/	mean	3.63	2.20	1.09	2.41	2.23	3.81	3.80	4.68	1.26	1.36	1.10	1.40
	m-m	0.66	0.38	0.14	0.29	0.27	0.45	0.54	0.92	0.09	0.20	0.07	0.25
/u/	mean	1.49	7.33	2.57	4.24	2.76	3.83	2.60	3.04	1.18	1.13	1.02	1.20
	m-m	0.53	2.71	0.87	1.52	0.95	1.40	0.84	1.07	0.74	0.43	0.15	0.44

(B) $\underline{E_w}$ –minima selected frames

Vowel	Parameter	Lips-to-glottis section number											
		1	2	3	4	5	6	7	8	9	10	11	12
/a/	mean	2.22	5.82	3.08	3.06	1.68	2.38	1.97	2.90	1.19	1.33	1.07	1.41
	m-m	0.02	0.05	0.02	0.02	0.01	0.02	0.01	0.02	0.00	0.00	0.00	0.01
/i/	mean	3.59	2.19	1.09	2.39	2.21	3.80	3.81	4.68	1.25	1.38	1.11	1.42
	m-m	0.01	0.00	0.00	0.01	0.01	0.02	0.01	0.01	0.00	0.01	0.00	0.01
/u/	mean	1.45	7.23	2.51	4.17	2.70	3.76	2.51	2.95	1.09	1.12	1.00	1.18
	m-m	0.07	0.35	0.12	0.21	0.13	0.18	0.12	0.15	0.03	0.07	0.02	0.09

[m-m : max-min deviation]

the frame positions are superimposed in Fig. 4. The plots of the area values for the frames selected at the E_w–minima are also superimposed in the same figure. The shaded lines show a large spread in the values, while the dark lines corresponding to the E_w–minima selected frames show a much smaller spread.

The variation in the estimated square root of the area values of each section was quantified by the max-min deviation (the difference between maximum and minimum values). The mean values and the max-min deviations for inter-frame interval of one sample are given in Table I(A). The deviations for the three vowels are significant for all the sections and they are very large for some of the sections. The mean values and the max-min deviations for the E_w–minima selected frames are given in Table I(B). As compared to the corresponding values in Table I(A), there are no significant changes in the mean values. The max-min deviations have significantly decreased for all the sections. They have decreased by more than an order of magnitude, except for a few front sections in case of /u/.

To study the effect of E_w–minima selected frames on the vocal tract shape estimation during transitional tract configurations, the method was used for obtaining areagrams

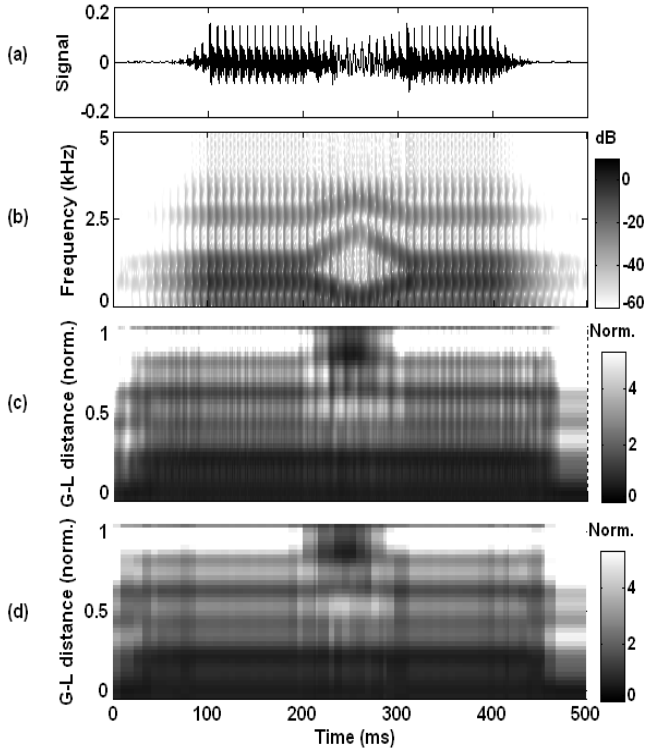

Figure 5. Estimated vocal tract shape variation of synthesized vowel-semivowel-vowel sequence /aja/. (a) speech waveform (b) wideband spectrogram (c) original areagram (d) areagram for E_w–minima selected frames.

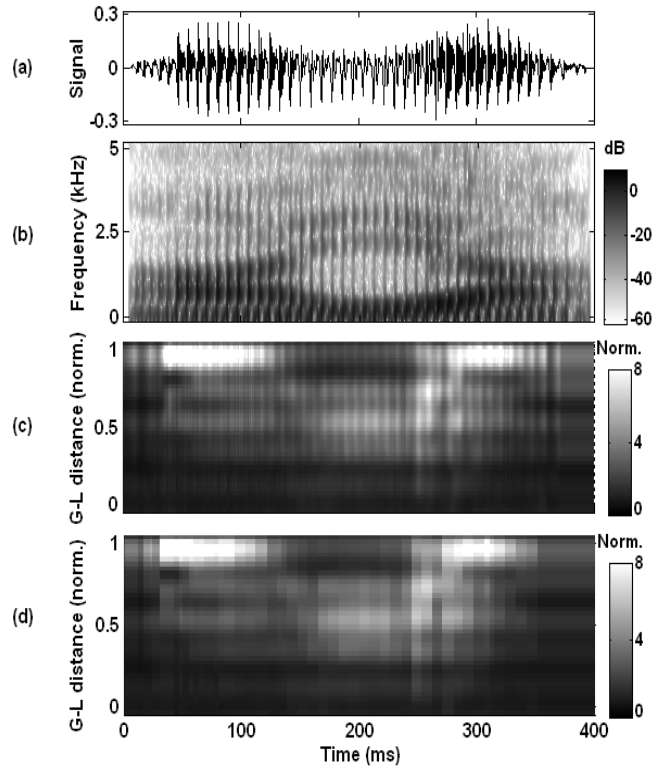

Figure 6. Estimated vocal tract shape variation of natural vowel-semivowel-vowel sequence /aja/. (a) speech waveform (b) wideband spectrogram (c) original areagram (d) areagram for E_w–minima selected frames.

of synthesized vowel-semivowel-vowel utterances. Fig. 5 shows the speech waveform, wideband spectrogram, and areagrams for the utterances /aja/. It is observed that E_w–minima selected frames resulted in areagram with reduced variability during fixed tract configuration without smearing the changes in the area values during the transitional tract configuration.

To examine the effectiveness of the method in estimating the vocal tract shape from natural speech signal, areagrams were obtained for vowel and vowel-semivowel-vowel utterances from several speakers. As an example, Fig. 6 shows the speech waveform, wideband spectrogram, and areagrams for the natural utterance /aja/ from a male speaker. It is observed that E_w–minima selected frames resulted in reduced variability in the area values during the fixed vocal-tract configuration without smearing the changes in the area values during the transitional configuration.

IV. CONCLUSION

The investigations showed that vocal tract shape estimated for steady-state vowel varied with the position of the LPC analysis frame. Vocal tract shape estimated from the analysis frames positioned at the frames with minimum energy resulted in low prediction error and significantly reduced variability in the estimated area values. It is shown that minima of the windowed energy index (ratio of the windowed energy to the frame energy), detected by valley picking, can be used for selecting the frame positions for reducing the variability in the estimated values and improving the consistency of vocal tract shape estimation, without smearing the variations in the shape during transitional tract configuration. The proposed method can be used to estimate the VC and CV transition area values during VCV utterances and these values can be used to more accurately estimate the vocal tract shape and place of articulation during stop closures of VCV utterances. The method may be useful in improving the vocal tract shape estimation for speech training aids and other applications.

ACKNOWLEDGMENT

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

REFERENCES

[1] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737-793, 1987.

[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] H. Kinugasa, H. Kamata, and Y. Ishida, "Spoken word recognition using vocal tract shapes," in *Proc. IEEE Conf. Commun., Comput. Signal Process.*, 1993, vol. 1, pp. 133-136.

[5] 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.

[6] K. Erler and G. Freeman, "Using articulatory feature for speech recognition," in *Proc. IEEE Conf. Commun., Comput. Signal Process.*, 1995, pp. 562–566.

[7] 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.

[8] 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.

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

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

[11] 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.

[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] 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.

[14] H. Wakita, "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 3, pp. 281–285, 1979.

[15] 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.

[16] L.R. Rabiner, B.S. Atal, and M.R. Sambur, "LPC prediction error-analysis of its variation with the position of the analysis frame," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. ASSP-25, no.5, pp. 434-442, 1977.

[17] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, pp. 309–319, 1979.

[18] R. Mizoguchi, M. Yanagida, and O. Kakusho, "Speech analysis by selective linear prediction in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1982, vol. 3, pp. 1573-1576.

[19] C. Ma, Y. Kemp, and L. F. Willems, " Robust signal selection for linear prediction analysis of voiced speech," *Speech Commun.*, vol. 12, no. 2, pp. 69-81, 1993.

[20] M. Mezzalama, "Influence of the position of the analysis frame in LPC pitch synchronous analysis," *Signal Process.*, vol. 1, pp.191-204, 1979.

[21] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, pt. 3, pp. 971–995, 1980.

[22] M. S. Shah and P. C. Pandey, "Estimation of place of articulation in stop consonants for visual feedback," in *Proc. Interspeech*, 2007, pp. 2477–2480.

[23] D. G. Childers and T. H. Hu, "Speech synthesis by glottal excited linear prediction," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2026-2036, 1994.