VOICE CONVERSION BY PROSODY AND VOCAL TRACT MODIFICATION

K. Sreenivasa Rao

Department of ECE,
Indian Institute of Technology Guwahati,
Guwahati - 781 039, India.
E-mail: ksrao@iitg.ernet.in

B. Yegnanarayana

Department of CS&E,
Indian Institute of Technology Madras,
Chennai - 600 036, India.
E-mail: yegna@cs.iitm.ernet.in

ABSTRACT

In this paper we proposed some flexible methods, which are useful in the process of voice conversion. The proposed methods modify the shape of the vocal tract system and the characteristics of the prosody according to the desired requirement. The shape of the vocal tract system is modified by shifting the major resonant frequencies (formants) of the short term spectrum, and altering their bandwidths accordingly. In the case of prosody modification, the required durational and intonational characteristics are imposed on the given speech signal. In the proposed method, the prosodic characteristics are manipulated using instants of significant excitation. The instants of significant excitation correspond to the instants of glottal closure (epochs) in the case of voiced speech, and to some random excitations like onset of burst in the case of nonvoiced speech. Instants of significant excitation are computed from the Linear Prediction (LP) residual of the speech signals by using the property of average group delay of minimum phase signals. The manipulations of durational characteristics and pitch contour (intonation pattern) are achieved by manipulating the LP residual with the help of the knowledge of the instants of significant excitation. The modified LP residual is used to excite the time varying filter. The filter parameters are updated according to the desired vocal tract characteristics. The proposed methods are evaluated using listening tests.

1. INTRODUCTION

Voice conversion is a method that aims to transform the characteristics of an input (source) speech signal such that the output (transformed) signal is perceived to be produced by another (target) speaker. Its applications include customization of text-to-speech systems (e.g., to speak with a desired voice or to read out email in the sender's voice) as well as entertainment and security applications [1]. In film industry, voice conversion systems can be employed for dubbing and translation to a different language by preserving speaker characteristics. Personification of synthesized speech is another important application as many automated systems use synthesized speech as a computer interaction tool [2].

Speech production is often represented by a source-filter model. Both parts of this model contribute to speaker characteristics. For example, speech rate, duration, pitch and dynamic pitch range are features mainly related to the source, while formant positions and bandwidths are features related to the filter, i.e., vocal tract. A perfect voice conversion should deal with all these features in phase. However, they are often processed separately to make the task easier. Most current voice conversion systems focus on the spectral conversion and often apply simple modification for prosody features, such as shifting the average pitch of a source speaker to a target speaker [3, 4]. In this paper the proposed methods impose the target pitch contour along with the desired durational and intensity (gain) modifications.

There are many ways to implement the transformation function for converting source features to target features, such as mapping codebooks, discrete transformation functions, artificial neural networks, Gaussian mixture models and some combinations of them [5–9]. In codebook mapping, one to one correspondence between the spectral codebook entries of the source speaker and the target speaker is developed by some form of supervised vector quantization method. In general, these methods face several problems such as artifacts introduced at the boundaries between successive speech frames, limitation on robust estimation of parameters (e.g., formant frequency estimation), or distortion introduced during synthesis of target speech. Another issue which has not been explored in detail is the transformation of the glottal excitation characteristics.

In discrete transformation method, codebook mapping is replaced by piecewise linear function [6]. The discrete classification of source and target features results in discontinuity in the reproduced speech. Artificial neural networks are used here for mapping the source and target features using continuous and nonlinear transformation function. GMM and Maximum Aposteriori (MAP) adaptation

9th International Conference on Information Technology (ICIT'06)
0-7695-2635-7/06 $20.00 © 2006 IEEE

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on January 30, 2009 at 02:07 from IEEE Xplore. Restrictions apply.

approach is used for spectral transformation, and Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) framework is used for pitch scale modification [10]. In selective pre-emphasis method, transformation of vocal tract spectrum is carried out using sub-band based framework, where perceptual characteristics of human auditory system is taken into account [2]. This method provides the estimation of detailed spectral envelope and the modification of spectral resolution in different sub-bands.

In harmonic peak-to-valley ratio (HPVR) modification method, voice quality attributes such as breathiness and roughness of a speaker are collectively modeled by harmonic peak-to-valley ratio of the speaker's speech spectrum [11]. The average HPVR is modified through a post-filtering operation, after the conversion of spectral envelope, pitch and other speaker individual features.

In this paper we proposed methods to modify both vocal tract characteristics and prosody (i.e., durational, intonational and gain contours) information according to the target requirements. In Section 2, modification of vocal tract shape is discussed. Section 3 describes the general method for the modification of prosodic characteristics and a specific method to impose the desired (target) pitch contour on a given speech signal. Results of the perceptual evaluation from the listening tests are given in Section 4. Conclusions and future work are provided in Section 5.

2. MODIFICATION OF VOCAL TRACT CHARACTERISTICS

The basic shape of vocal tract can be characterized by the gross envelope of Linear Prediction (LP) spectrum. LP spectrum can be approximated by a set of resonant frequencies (formants) and their associated bandwidths. For each speaker, the shape of the vocal tract will be unique, and correspondingly the set of formants and their bandwidths. Formant frequency values depend on gender and age of the speaker. In this paper we derived a gross relation between formant frequencies and average pitch, with respect to male and female speakers.

Television (TV) broadcast news data is used for analyzing the relationship between pitch (source) and formant frequencies (vocal tract characteristics) [12]. For the analysis, speech data of five female and five male speakers is considered. The steady vowel regions with identical positional and contextual constraints are derived from the database. Formant frequencies are computed by using the group delay function method [13]. All the group delay functions of a particular vowel are plotted on the same figure, one over the other, as shown in Fig. 1. In the figure, the overlapping regions correspond to the formant frequencies of the vowel. The average pitch in the vowel region is also computed. The average pitch and formant frequencies for all vowels and speakers are computed. Table 1 shows the dependence of formant frequencies on pitch. In the analysis the first three formants are considered.

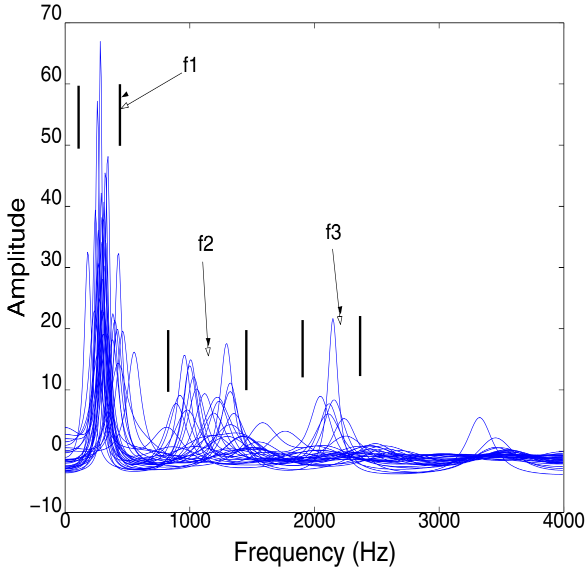

Fig. 1. Group delay functions for the vowel /u/.

The shape of the vocal tract is characterized by major resonances of the vocal tract system. The resonances and their bandwidths are related to the angle and magnitude of the corresponding poles in the z-plane. The formant frequencies can be changed by shifting the poles of a system transfer function in the z-plane. As per the required modification in formant frequencies, the angle and magnitude of the poles are modified. The LPCs are recomputed from these new poles. The basic procedure for the modification of LPCs is given in Table 2.

3. MODIFICATION OF PROSODIC CHARACTERISTICS

Proposed methods modify the prosodic characteristics such as, pitch contour and durational characteristics in residual domain using the knowledge of the instants of significant excitation [14]. The basic reason for choosing the residual domain for modification is that the successive samples in the LP residual are less correlated compared to the samples in the speech signal. Therefore the residual manipulation is likely to introduce less distortion in the speech signal synthesized by using the modified LP residual and LP coefficients. The region around the instant of glottal closure correspond to the significant part of excitation, in the sense that the strength of excitation is maximum in that region of

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on January 30, 2009 at 02:07 from IEEE Xplore. Restrictions apply.

Table 1. Dependence of formant frequencies on pitch.

Vowel	Male (M)				Female (F)				Ratio (F/M)			
	Pitch	Formant frequencies			Pitch	Formant frequencies			Pitch	f1	f2	f3
		f1	f2	f3		f1	f2	f3				
a	115	652	1410	2365	245	845	1630	2732	2.13	1.30	1.16	1.16
e	132	470	1895	2635	263	605	2165	2915	1.99	1.29	1.14	1.11
i	127	378	2086	2757	239	465	2476	3115	1.88	1.23	1.19	1.13
o	122	562	1315	2515	255	678	1438	2835	2.09	1.21	1.09	1.13
u	118	365	1225	2470	228	475	1357	2992	1.93	1.30	1.11	1.21

Table 2. Steps for LPCs modification.

1 Preemphasize the input speech.
2 Compute LPCs with 10^{th} order LP analysis, with a frame size of 20 ms and a frame shift of 5 ms.
3 For each set of LPCs compute the roots in rectangular form $(x_i \pm jy_i)$.
4 Transform the roots to polar form $(re^{\pm j\theta})$ using the relations $r = \sqrt{x^2 + y^2}$ and $\theta = \arctan(y/x)$.
5 As per the required vocal tract shape, modify the magnitude and angle (r and θ) of the poles using the relations $\theta_i^{'} = \alpha_i \theta_i = \frac{f_i^{'}}{f_i}\theta_i$ and $r = e^{-\pi\beta_i T}$. (where θ_i and $\theta_i^{'}$ represents angular components of poles of source and target formant frequencies f_i and $f_i^{'}$, r = magnitude of the poles, β_i and T represents bandwidth of formants and sampling period.)
6 Transform the modified roots into complex conjugate form using the relation $(rcos\theta + jrsin\theta)$.
7 Compute the LPCs from the modified roots.

the pitch period. Therefore, we attempt to retain that region during pitch period modification. Finally, in LP residual manipulation the samples around the instant of glottal closure are preserved, and the samples at the low SNR regions (other than the region around the glottal closure) are allowed for modification. This process offers least audible distortion in the synthesized speech.

There are four main steps involved in the prosody manipulation: (1) Deriving the instants of significant excitation (epochs) from the LP residual signal, (2) deriving a modified (new) epoch sequence according to the desired prosody (pitch and duration), (3) deriving a modified LP residual signal from the modified epoch sequence, and (4) synthesizing speech using the modified LP residual and the LPCs.

Group-delay analysis is used to derive the instants of significant excitation from the LP residual [15, 16]. The analysis involves computation of the average slope of the unwrapped phase spectrum (i.e., average group-delay) for each frame. If $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of the windowed signal *x(n)* and *nx(n)*, respectively, then the group-delay function $\tau(\omega)$ is given by the negative derivative of the phase function $\phi(\omega)$ of $X(\omega)$, and is given by [15, 17]

$$\tau(\omega) = -\phi^{'}(\omega) = \frac{X_R Y_R + X_I Y_I}{X_R^2 + X_I^2},$$

where $X_R + jX_I = X(\omega)$, and $Y_R + jY_I = Y(\omega)$. Any isolated sharp peaks in $\tau(\omega)$ are removed by using a 3-point median filtering. Note that all the Fourier transforms are implemented using the discrete Fourier transform. The average value $\bar{\tau}$ of the smoothed $\tau(\omega)$ is the value of the *phase slope function* for the time instant corresponding to the center of the windowed signal $x(n)$. The phase slope function is computed by shifting the analysis window by one sample at a time. The instants of positive zero-crossings of the phase slope function correspond to the instants of significant excitation. Fig. 2 illustrate the results of extraction of the instants of significant excitation for voiced speech segment.

The time interval between two successive epochs correspond to the pitch period for voiced speech. With each epoch we associate three parameters, namely, time instant, epoch interval and LP residual. We call these as *epoch parameters*. The prosody manipulation involves deriving a new excitation (LP residual) signal by incorporating the desired modification in the duration and pitch period for the utterance. This is done by first creating a new sequence of epochs from the original sequence of epochs. For this purpose all the epochs derived from the original signal are considered, irrespective of whether they correspond to a voiced segment or a nonvoiced segment.

For pitch period modification, the new sequence of epochs are generated using the following steps: (1) The epoch interval plot is generated from the epoch sequence of the original signal. (2) The new epoch interval plot is generated by

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on January 30, 2009 at 02:07 from IEEE Xplore. Restrictions apply.

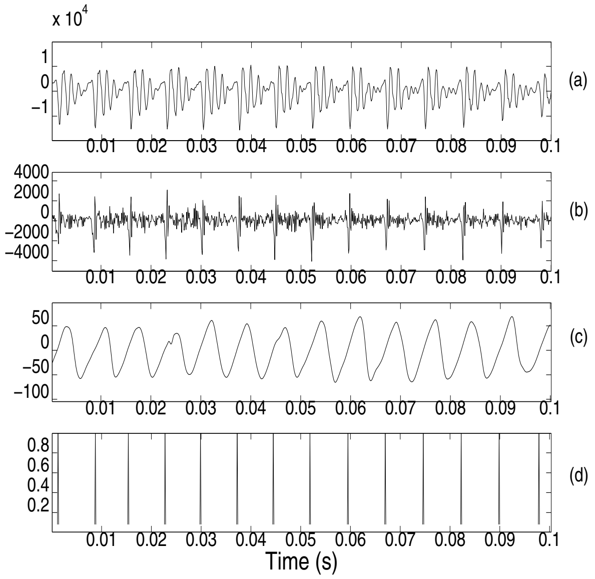

Fig. 2. (a) A segment of voiced speech and its (b) LP residual, (c) phase slope function, and (d) instants of significant excitation.

multiplying the original epoch interval plot by the desired modification factor (i.e., scaling the epoch interval plot). (3) Derive the new epoch sequence from the modified epoch interval plot iteratively. In the case of duration modification, the new epoch interval plot is generated either by stretching or compressing the original epoch interval plot according to the desired modification factor. This is achieved by resampling process.

For each epoch in the new epoch sequence, the nearest epoch in the original epoch sequence is determined, and thus the corresponding epoch parameters are identified. The original LP residual is modified in the epoch intervals of the new epoch sequence, and thus a modified excitation (LP residual) signal is generated. The modified LP residual signal is then used to excite the time varying all-pole filter represented by the LPCs.

So far in the pitch period modification, the pitch contour of the speech utterance is shifted by a constant scale factor. But in voice conversion application, the existing pitch contour needs to be modified as per the desired (target) pitch contour (i.e., imposing the desired pitch contour on a given speech signal). For imposing the desired (target) pitch contour on the existing pitch contour, the durations of voiced and nonvoiced regions in both the contours should be equal. Therefore the voiced and nonvoiced regions of the desired epoch interval plot (desired pitch contour) are resampled according to the requirement. Then, the voiced and nonvoiced regions of the given epoch interval plot are replaced by the resampled voiced and nonvoiced regions of the desired epoch interval plot. Fig. 3 shows the modification of the given epoch interval plot according to the desired epoch interval plot.

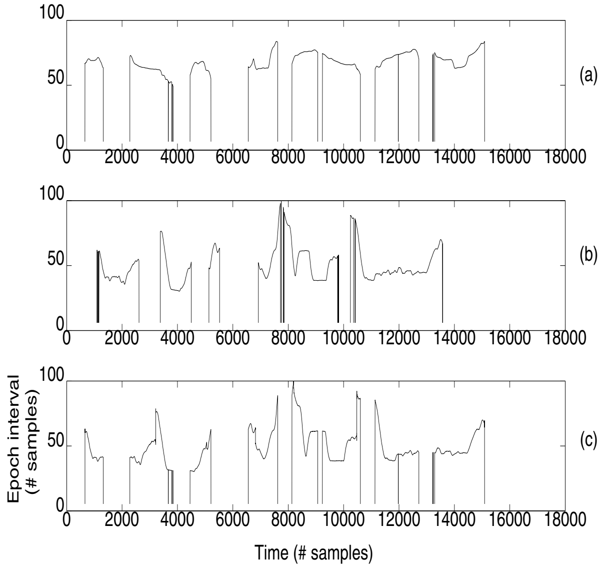

Fig. 3. (a) Epoch interval plot for a given speech utterance, (b) desired epoch interval plot and (c) desired epoch interval plot is superimposed on the given epoch interval plot.

4. RESULTS AND DISCUSSION

In the previous sections we described the methods to modify the vocal tract shape (formant structure) and prosody characteristics such as duration and intonation patterns. In this section we discuss how these methods are used collectively, for converting a male speaker speech to a female speaker speech and vice versa. For performing this study, ten Hindi sentences are recorded by a male and a female speakers. For each utterance, duration, average pitch, pitch contour, average frame energy and energy contour (gain contour) are computed. Using these values, each of the utterance spoken by a male speaker is transformed to it's corresponding female speaker spoken utterance, and vice versa. To perform the above transformation, prosody modification is carried out in LP residual domain, and vocal tract characteristic transformation is carried out by shifting the poles of the transfer function in *z*-plane. The overall transformation procedure is given in Table 3. The transformation of voice characteristics at course level can be done by using the steps mentioned in the Table 3. At finer level, one should incor-

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on January 30, 2009 at 02:07 from IEEE Xplore. Restrictions apply.

porate the target pitch contour and gain contour instead of gross modification factors.

Table 3. Steps for overall transformation.

1 Compute LPCs and LP residual with 10^{th} order LP analysis, with a frame size of 20 ms and shift of 5 ms.
2 Modify the LPCs by shifting the poles in Z-plane as per the desired vocal tract shape.
3 Compute the ratios for the average pitch, duration and average frame energy between source and target speakers.
4 Modify the LP residual as per the required pitch and duration modification factors.
5 Synthesize the speech by exciting the time varying filter representing the modified LPCs with the modified LP residual.
6 Enhance the synthesized speech signal by a required energy transformation factor.

The proposed methods can be evaluated by listening tests. Twenty students are participated in conducting these tests. These tests are used to evaluate the following three aspects: (1) Perceptual distortion, (2) Natural feminine/male characteristics and (3) Target speaker characteristics. The subjects were asked to give their opinion score on a five point scale separately for each of these aspects. The 5-point scale for representing the quality of speech and the distortion level is given in Table 4 [18]. The Mean Opinion Scores (MOS) for the above mentioned three aspects are given in Table 5. The scores in first row correspond to the transformation from male to female, and the second row correspond to female to male transformation. The mean opinion scores indicate that there is no perceivable distortion in the transformed speech. This may be due to modification of prosodic characteristics in LP residual domain. With respect to second aspect (natural gender characteristic), the MOS indicate that the transformation provided the desired gender characteristics in the synthesized speech. In the case of last aspect (target speaker characteristics), the MOS indicate that the performance of the transformation is not up to the mark. This is because the transformation is done at course level. Slight improvement is observed in this aspect, if the transformation is performed at finer level.

Table 4. Ranking used for judging the quality and distortion of the synthesized speech signal.

Rating	Speech quality	Level of distortion
1.	Unsatisfactory	Very annoying and objectionable
2.	Poor	Annoying but not objectionable
3.	Fair	Perceptible and slightly annoying
4.	Good	Just perceptible but not annoying
5.	Excellent	Imperceptible

Table 5. Mean Opinion Scores (MOS) for evaluating the transformation process.

Transformation	Mean Opinion Scores (MOS)		
	Perceptual distortion	Target gender characteristic	Target speaker characteristic
Male to Female	4.56	4.23	2.92
Female to Male	4.71	4.37	3.23

The subjective tests rightly pointed out that the transformation process had not incorporated the target speaker characteristics accurately. This is due to the fact that, in the transformation process, we used the average values for transforming pitch contour, energy contour and vocal tract characteristics. But, this is sufficient for certain applications where it is required to transform male voice to female voice and vice versa. If the application requires specific speaker characteristics, we need to perform the transformation by imposing the exact pitch contour, energy contour, source characteristics and vocal tract characteristics.

5. SUMMARY AND CONCLUSION

In this paper we proposed methods to modify the vocal tract characteristics and prosodic characteristics of a given speech signal. In particular, these methods are useful in voice conversion process. In the proposed method, vocal tract characteristics are modified by shifting the poles of the vocal tract system transfer function in z-plane. Prosodic characteristics (duration and intonation patterns) are modified in residual domain using instants of significant excitation as anchor points. As the samples in the LP residual have low correlation, manipulation of these samples according to desired prosody modification gives minimum distortion. Samples around the instants of significant excitation correspond to high signal-to-noise ratio regions of the speech signal. Therefore in prosody modification, these regions are preserved and only other regions are allowed for manipulation. This process retains the naturalness and gives minimum distortion in the output transformed speech. The basic methods proposed in this paper provides greater flexibility in modification of the parameters by large factors. The proposed methods are evaluated using listening tests. The MOS from the listening tests indicate that the transformation process does not produce any perceptual distortion. It is also evident that the desired gender transformation is achieved successfully, but the specific target speaker characteristics are not

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on January 30, 2009 at 02:07 from IEEE Xplore. Restrictions apply.

completely observed in the transformed speech.

The methods mentioned in this paper provide the basic framework for voice conversion process. In this paper, we have not attempted the modification of glottal pulse characteristics. It is known that the shape of the glottal pulse and its characteristics are unique to a speaker. Therefore by incorporating the glottal pulse characteristics according to the desired speaker, the overall performance of the voice conversion can be improved.

6. REFERENCES

[1] A. Kain and M. W. Macon, “Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 813–816, May 2001.

[2] O. Turk and L. M. Arslan, “Subband based voice conversion,” in *Proc. Int. Conf. Spoken Language Processing*, (Denver-Colorado, USA), 2002.

[3] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 285–288, May 1998.

[4] T. Toda, H. Saruwatari, and K. Shikano, “Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 841–844, May 2001.

[5] M. Abe, S. Nakanura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 655–658, May 1998.

[6] H. Mizuno and M. Abe, “Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt,” *Speech Communication*, vol. 16, pp. 153–164, 1995.

[7] M. Narendranadh, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, “Transformation of formants for voice conversion using artificial neural networks,” *Speech Communication*, vol. 16, pp. 206–216, Feb. 1995.

[8] Y. Stylianou, Y. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 131–142, 1998.

[9] M. Marshimi, T. Toda, K. Shikano, and N. Campbell, “Evaluation of cross-language voice conversion based on GMM and STRAIGHT,” in *Proceedings of EUROSPEECH 2001*, (Aalborg, Denmark), pp. 361–364, Sept. 2001.

[10] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, “Voice conversion with smoothed GMM and MAP adaptation,” in *Proc. Eurospeech*, (Geneva), pp. 2413–2416, 2003.

[11] A. Verma and A. Kumar, “Modification of harmonic peak-to-valley ratio for controlling roughness in voice conversion,” *Proceedings of IEE: Electronic Letters*, vol. 40, Dec. 2004.

[12] *Database for Indian languages*. Speech and Vision lab, Indian Institute of Technology Madras, India, 2001.

[13] H. A. Murthy and B. Yegnanarayana, “Formant extraction from group delay function,” *Speech Communication*, vol. 10, pp. 209–221, Mar. 1991.

[14] K. S. Rao and B. Yegnanarayana, “Prosody modification using instants of significant excitation,” *IEEE Trans. Speech and Audio Processing*, vol. 14, pp. 972–980, May 2006.

[15] P. S. Murthy and B. Yegnanarayana, “Robustness of group-delay-based method for extraction of significant excitation from speech signals,” *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 609–619, Nov. 1999.

[16] R. Smits and B. Yegnanarayana, “Determination of instants of significant excitation in speech using group delay function,” *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 325–333, Sept. 1995.

[17] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*. Upper Saddle River, NJ.: Prentice-Hall, 1999.

[18] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*. New York, USA: Macmilan Publishing Company, 1993.

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY. Downloaded on January 30, 2009 at 02:07 from IEEE Xplore. Restrictions apply.