Proc. 17th National Conference on Communications (NCC 2011), Bangalore, 28-30 January 2011, Paper No. Sp Pr. II, P2

Transformation of Short-Term Spectral Envelope of Speech Signal Using Multivariate Polynomial Modeling

P. K. Lehana
Department of Physics and Electronics
University of Jammu
Jammu 180004, India
Email: pklehana@gmail.com

P. C. Pandey
Department of Electrical Engineering
Indian Institute of Technology Bombay
Powai, Mumbai 00076, India
Email: pcpandey@ee.iitb.ac.in

Abstract— **Speaker transformation uses a mapping between the signal parameters derived from a set of phrases spoken by two speakers to modify the speech signal of one speaker to make it perceptually similar to that of the other speaker. In spectral warping and interpolation technique, the transformation function can be estimated using lesser data, but it needs a different transformation function for each acoustic class. Earlier reported investigations have used class dependent multivariate linear regression or univariate cubic interpolation. In this paper, a technique for modifying spectral characteristics, using a multivariate polynomial modeling for source-target mapping of the spectral parameters, is presented. Harmonic plus Noise Model (HNM) has been used for analysis-synthesis because of ease of the time and frequency scaling.**

Keywords-speaker transformation; speech conversion; spectral chracteristics; harmonic plus noise model.

I. Introduction

Speaker transformation modifies the speech signal of one speaker (source) to make it perceptually similar to that of another speaker (target). It is generally carried out using a speech analysis-synthesis system and a mapping between the signal parameters derived from a set of phrases spoken by source and target speakers. It involves four phases: feature extraction, estimation of source to target mapping, transformation of source parameters, and resynthesis of speech using the transformed parameters. In feature extraction phase, the speech signal is analyzed for extracting the parameters of excitation source and the vocal tract. The parameters related to the vocal tract are considered relatively more important as compared to those related to the excitation in specifying speaker individuality [1]. The pitch contour is also considered to be an important cue for individuality [2].

For estimating the source-target mapping or transformation function from the parameters, the corresponding phonemic segments in the source and target passages are aligned. During resynthesis, the parameters of the source spectrum are modified by this transformation function for obtaining the transformed speech. For transformation, the speech spectrum is commonly represented using parameters such as formant frequencies [3], cepstrum [4], Mel Frequency Cepstrum Coefficients (MFCCs) [5], or Line Spectral Frequencies (LSFs) [6]. As compared to other parameters, the use of MFCCs has been reported to be advantageous as spectral compression due to mel scale reduces the perceptual effect of the errors [7]. The components of MFCCs are almost uncorrelated, making them suitable in the framework of stochastic modeling. Despite over-smooth representation, they do not loose high frequency information [8]. The corresponding components in source and target MFCCs have been reported to be correlated [9], and they are robust with respect to noisy environment [10]. The main problem with analysis-synthesis using MFCCs is that they lose pitch and phase related information. Hence the phases for resynthesis need to be estimated using source phase, predicted waveforms, minimum phase, or phase codebook approaches [11].

The set of source or target parameters for each frame is known as a feature vector. The techniques for estimating the transformation function from the source feature vectors to the corresponding target feature vectors are generally based on Vector Quantization (VQ) [12], Artificial Neural Networks (ANN) [13], Gaussian Mixture Model (GMM) [14], Hidden Markov Model (HMM) [15], Vector Field Smoothening (VFS) [16], [17], Time-Variant Filtering (TVF) [16], Mixtures of Linear Transform (Ms-LT) [18], Multistep-speaker Voice Conversion (MVC) [19], and frequency warping and interpolation [20].

In GMM based systems [14], the acoustic feature spaces of the source and target speakers are modeled by finite number of Gaussian functions, assuming that the speech may be characterized by finite acoustic classes such as vowels, nasals, or fricatives. Each class is represented by average spectral feature vector along with some variability because of pronunciation and co-articulation effects. Stylianou et al. [5] used Harmonic plus Noise Model (HNM) based GMM technique for estimating the transformation function. The cepstral distance was reduced by more than 4 dB between the source and the target frames. The quality of the conversion, assessed through XAB, preference, and opinion tests based on three utterances and 6 listeners was satisfactory although some of the listeners reported a muffling effect when the number of GMM components was small. Bandoin and

Stylianou have reported in [21] that GMM is better than VQ and ANN. The shortcoming of this technique is the over-smoothening of the transformed spectrum because of the weighted sum of the conditional mean vectors [22].

The transformation function may be estimated using text-dependent or text-independent schemes. Jian and Yang [18] used Ms-LT estimated the transformation function using LSFs (order = 16) in a text-independent scheme. The results were slightly inferior for text-interdependent scheme as compared to text dependent and this was attributed to spectral averaging. When a transformation function is to be developed for each combination of speakers from the database, MVC reduces the number of transformation functions by introducing an intermediate speaker and computing the transformation function from each speaker to the intermediate speaker. Masuda and Shozakai [19] used this concept using GMM (mixtures = 64) based transformation function obtained from cepstral coefficients (order = 41) with LPC based analysis-synthesis. The quality of the transformed speech was reported to be unsatisfactory.

One of the simplest techniques for speaker transformation, based on Dynamic Frequency Warping (DFW), used formants, but it required automated estimation of the frequency, bandwidth, and amplitude of the formants. Quality of modified speech was not very high because of limitations of formant based synthesis [3], [23]. Valbret et al. [20] estimated the warping function using DFW between frame-aligned log-magnitude spectra after removing the spectral tilt. The transformation function for each class was estimated by modeling the median of the warping functions by a cubic polynomial. For comparison, Linear Multivariate Regression (LMR) was also used to estimate the transformation function for each class using cepstral coefficients (order = 21). The synthesis was performed using LPC framework and the prosody modified by PSOLA. Training was carried out by CVC logatoms. XAB based evaluation (with CVCs and 3 subjects) showed LMR to be better than DFW. Some audible distortions were reported in the transformed speech.

Iwahashi and Sagisaka [24], [25] investigated speaker interpolation technique, using a transformation of spectral patterns, time-aligned using Dynamic Time Warping (DTW) [5], [26]. Interpolation ratio was determined by minimizing the error between the interpolated and target spectra. For generating the speech of the given target, the spectral vector for each frame of the source speech was compared with the stored spectral vectors to find the nearest one. The set of interpolation ratio for this frame and the given target were used to interpolate the spectral parameters for the target. Evaluation using Japanese utterances showed a need of further refining the technique as the formants of the transformed speech were broadened because of over smoothening.

The various techniques for spectral transformation may be grouped as being based on vector quantization, statistical and ANN based transformation, and frequency warping and interpolation. Vector quantization suffers from the discrete nature of the acoustic space, which hampers the dynamic character of the speech signal. The statistical and ANN based techniques capture the natural transformation function independent of the acoustic unit, but they need a large set of training data and computation. In frequency warping and interpolation, the transformation function can be estimated using lesser data, but a different transformation function is needed for each acoustic class.

Our hypothesis is that a single transformation function between the acoustic spaces of the source and the target may be derived using multivariate polynomial modeling. The objective of this paper is to investigate the modification of spectral characteristics by modeling the source-target relationship for obtaining a single mapping applicable to all acoustic classes. Each parameter for generating the target speech is modeled as a multivariate polynomial function of all the parameters of the source speech, and the set of these polynomial functions is obtained by analyzing a set of time aligned source and target frames. The study has been carried out using univariate linear, multivariate linear, and multivariate quadratic functions. For analysis and synthesis, HNM has been used, as it provides high quality speech output with a reasonable number of parameters, and easily permits pitch and time scaling [27], [28]. As the HNM parameters (harmonic magnitudes and LP coefficients) are not suitable for multivariate polynomial modeling, the harmonic magnitudes in the harmonic band were converted to MFCCs and the LP coefficients in the noise band to LSFs for estimating the transformation function. Application of the multivariate polynomial modeling for transformation of short-term speech spectral envelope along with a linear transformation for time and pitch scaling is used for speaker transformation. The multivariate polynomial modeling is briefly described in Section II. Methodology of the investigations is described in Section III and results are presented in Section IV.

II. Multivariate Polynomial Modeling

If an m-dimensional function g is known at q points, a multivariate polynomial surface f can be constructed such that it approximates the given function within some error at each point [29],

$$g({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m) = f({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m) + \varepsilon_n \qquad (1)$$

where $n = 0, 1, \cdots, q-1$. The multivariate function can be written as

$$f(w_1, w_2, \cdots, w_m) = \sum_{k=0}^{p-1} c_k \phi_k (w_1, w_2, \cdots, w_m) \qquad (2)$$

where p is the number of terms in the polynomial of m variables. By combining (1) and (2), we get a matrix equation

$$\mathbf{b} = \mathbf{Az} + \boldsymbol{\varepsilon} \qquad (3)$$

where vectors $\mathbf{b}$, $\mathbf{z}$, and $\boldsymbol{\varepsilon}$ are given by

$$\mathbf{b}^{\mathrm{T}} = [g_0 \quad g_1 \quad \cdots \quad g_{q-1}]$$

$$\mathbf{z}^{\mathrm{T}} = [c_0 \quad c_1 \quad \cdots \quad c_{p-1}]$$

$$\boldsymbol{\varepsilon}^{T} = [\varepsilon_0 \quad \varepsilon_1 \quad \cdots \quad \varepsilon_{q-1}]$$

and matrix $\mathbf{A}$ is a $q \times p$ matrix, with elements given as

$$a(n,k) = \phi_k ({}^{n}w_1, {}^{n}w_2, \cdots, {}^{n}w_m) \text{, } 0 \le n \le q\text{-}1, \ 0 \le k \le p\text{-}1.$$

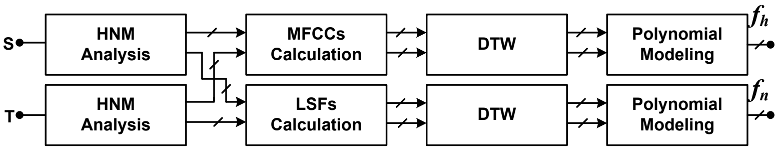

Figure 1. Estimation of the transformation function.

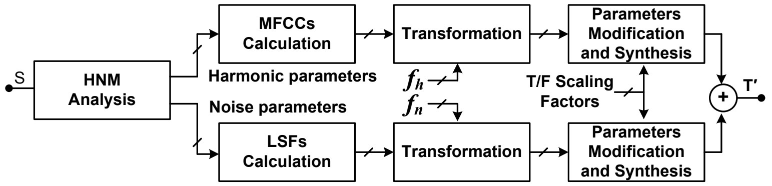

Figure 2. Transformation of the source speech.

If the number of data points is greater than the number of terms in the polynomial ($q \geq p$), then coefficients c_k's can be determined for minimizing the sum of squared errors

$$E=\sum_{n=0}^{q-1}\left[g({}^nw_1,{}^nw_2,...,{}^nw_m)-f({}^nw_1,{}^nw_2,...,{}^nw_m)\right]^2 \qquad (4)$$

and we get the solution

$$\mathbf{z}=(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{b} \qquad (5)$$

where $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ is known as pseudo–inverse of **A** [29].

We model the mapping between the acoustic spaces of the source and the target using multivariate quadratic surface as it provides a smooth transition between two given points in the multidimensional space and hence can be used for interpolation for the values which are not present in the training data. The total number of terms in a quadratic expression with *m* variables is $p=1+2m+{}^mC_2$.

III. Methodology

The methodology of investigations carried out may be divided into five sub-tasks: material recording, estimation of transformation function, transformation of the source speech, and evaluation. These tasks are described as follows.

A. Material

A small story in Hindi consisting of 80 sentences was recorded, in an acoustically treated room, with 16-bit quantization and sampling frequency of 10 kHz. This story was read by four speakers (two males and two females, age: 20-23 years, mother tongue: Hindi). A total of 77 sentences were used for estimating the transformation function and the remaining three were used for testing. In this paper, the two male speakers are referred to as M1, M2 and the female speakers as F1, F2. Speech signals corresponding to all the sentences were manually segmented.

B. Transformation Function Estimation

The scheme for estimating the transformation functions is shown in Fig 1. The speech signals of source S and target T sentences are analyzed using HNM for obtaining the parameters of harmonic band (voicing, pitch, maximum voiced frequency, harmonic magnitudes and phases) and noise band (LP coefficients and energy) [27], [28]. The analysis window length is taken as two pitch period for voiced segments and 10 ms for unvoiced segments. From the harmonic magnitudes a_k's, a continuous spectral magnitude function sampled at $K=256$ points is obtained for minimizing the sum of squared errors [30]. After estimating the energy E_m in each critical band in the spectrum by using a triangular function [31], MFCCs (order=20) are calculated as

$$c_m=A\sum_{m=0}^{M-1}\cos(j\frac{\pi}{M}(m+0.5))\log_{10}E_m \qquad (6)$$

where A is taken as 100 [30]. LP coefficients for the noise band (in both the voiced and the unvoiced frames) are converted to LSFs of order 13. Dynamic time warping [32] is used for further frame-by-frame alignment of the parameters, using MFCCs for the voiced frames and LSFs for the unvoiced frames.

The transformation functions f_h and f_n are estimated using univariate linear, multivariate linear, and multivariate quadratic modeling. In univariate linear modeling, the mapping between the MFCCs of the voiced segments is estimated between the corresponding terms in the feature vectors. This technique is not used for LSFs, as they do not have any correlation between corresponding coefficients. In multivariate modeling, each component in the target feature vector **y** is modeled as a multivariate quadratic function of all the components in the source vector **x,**

$$y_i=f(x_0,x_1,\cdots,x_{M-1}) \qquad (7)$$

where $0\leq i\leq M-1$ and $M=20$ for MFCCs and 13 for LSFs. Coefficients for these functions, for mapping from source to target frame vectors, were obtained using (5). Thus two transformation functions were obtained, one for the harmonic part (f_h) and other for the noise part (f_n).

C. Transformation of the Source Speech

Fig. 2 schematically represents generation of the transformed speech from the source speech. The source speech is analyzed for obtaining HNM parameters. The parameters of the harmonic part are converted to MFCCs and that of the noise part are converted to LSFs. The two transformation functions estimated earlier are used to transform the MFCCs and the LSFs. These parameters are converted back to the HNM parameters (harmonic magnitudes and LP coefficients). The pitch is scaled using mean and standard deviation [33]. For frame *i*, the target pitch P_t^i is obtained from the source pitch P_s^i as

$$P_t^i = \mu_t + (\sigma_t / \sigma_s)(P_s^i - \mu_s) \tag{8}$$

where μ_s and σ_s are the mean and standard deviation of the source pitch and μ_s and σ_s are the mean and standard deviation of the target pitch, obtained from the voiced frames of the source and target sentences. The time-scaling is carried out by a scaling factor obtained as the ratio of the durations of the total voiced frames in the source and target sentences.

Glottal closure instants are marked on the synthesis axis according to the modified pitch contour. The frame parameters at different synthesis time instants are obtained from the source parameters according to time scaling. Harmonic phases are estimated from the harmonic magnitudes by assuming a minimum phase system [34]. For the noise part, the excitation of the source speaker is also modified according to the required pitch and time scaling. The transformed noise part is resynthesized using an all-pole filter. Target speech is obtained by adding the harmonic and noise parts.

D. Evaluation

The level of distortion in the analysis-transformation-synthesis process was assessed by three tests. In the first test, the transformation was carried out for the sentences of the same speakers as both the source and the target. In the second test, only pitch was modified and in the third test, only vocal tract characteristics were modified by using the transformation function for the given source-target pair. In the first two cases, the identity of the speaker was not disturbed. A slight qualitative change was noticed, possibly because of the phase estimation assuming a minimum phase system or parameters modification for pitch scaling. In the third case, change of identity was observed with satisfactory quality.

Evaluation of speaker transformation was carried out for 4 speaker pairs: F1-F2, F1-M1, M1-F2, and M1-M2. Out of the 3 sentences not used in training, two were used for evaluation. For qualitatively studying the transformation, the spectrograms of the source, the target, and the modified speech were visually examined. Further, the closeness of the transformed speech to that of the target was assessed by both objective and subjective evaluations.

Objective evaluation was carried out using Mahalanobis distance [35] in parametric space defined for feature vectors **X** and **Y** as

$$D_{\mathrm{M}}(\mathbf{X},\mathbf{Y}) = \sqrt{(\mathbf{X}-\mathbf{Y})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\mathbf{Y})} \,. \tag{9}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the feature vectors used in training. It was used for estimating the distance of the transformed vectors from the corresponding target vectors, for univariate linear (UL), multivariate linear (ML), and multivariate quadratic (MQ) modeling of all the feature vectors obtained from DTW. In UL transformation, each coefficient of the target feature vector is assumed to be a function of the corresponding coefficient in the source-target feature vectors. On the other hand, ML and MQ transformations assume each coefficient of the target feature vector to be a function of all the coefficients in the source feature vectors. The closeness of the transformed and target speech was subjectively evaluated by XAB test [36], using an automated setup employing randomized presentations and a GUI for controlling the presentation and recording the responses. For each presentation, the subject matched the speech stimulus X with either stimulus A or stimulus B. Stimulus X could be randomly selected as source, target, or modified speech. Either the source or the target sounds were randomly presented as A or B. Subject could listen to the three sounds in any order more than once before finalizing the response. In a test, each stimulus appeared 3 times. This test was conducted with 6 subjects with normal hearing. Tests were conducted for two types of speaker transformations: transformation with spectral and pitch modification, transformation with only pitch modification. To estimate the MOS rating, the subjects were also asked to classify the quality of each X-phrase on the scale 1-5 with 1 as the lowest and 5 the highest.

TABLE I. MAHALANOBIS DISTANCE OF SOURCE-TARGET MFCCS.

Distance	Transformation			
	F1-F2	F1-M1	M1-F2	M1-M
Orig.	0.51	0.65	0.64	0.53
UL	0.68	0.65	0.61	0.64
ML	0.45	0.47	0.44	0.43
MQ	0.38	0.39	0.38	0.33

TABLE II. MOS SCORES (FOR 2 SENTENCES × 3 PRESENTATIONS × 6 LISTENERS, AVERAGED ACROSS THE 4 SPEAKER PAIRS).

Technique	UL	ML	MQ
MOS Score	1.7	2.8	3.1

IV. RESULTS

A visual examination showed the spectrogram of the transformed speech to be very similar to that of the target speech, and the speech obtained using multivariate quadratic function was relatively closer to the target speech as compared to that obtained using the other two functions. The spectrogram of the speech obtained using univariate linear function showed a degradation in the form of randomly distributed white dots.

The parametric distances between the source-target, target-transformed using the three functions (UL, ML, and MQ) are given in Table I. The reduction in target-transformed distance for MQ involving cross-gender transformation is higher as compared to the other two functions.

The results of the XAB test showed that listeners made a small error in identifying the source and target speech: 6 % for the source sentences and 4 % for the target sentences. For transformation using only the pitch modification, the modified speech was labeled as the target in 14 % of the responses. For the transformation involving the pitch and the spectral modification both, the modified speech was labeled as the target in 92 % of the responses. These results show a satisfactory transformation of the speech signal of the source speaker to the target speaker for all the source-target pairs. The results of MOS test are given in Table II. It shows that the quality of the transformed speech using UL is not satisfactory and that the quality is relatively better for MQ among the three techniques used for transformation.

V. Conclusion

We have investigated the use of multivariate polynomial modeling of spectral parameters in HNM based analysis-synthesis for speaker transformation. The technique does not require an extensive training data or labeling of acoustic classes. The investigations showed that univariate modeling did not result in good quality transformation, and multivariate modeling resulted in fair quality speech and satisfactory transformation of speaker identity. Further listening tests involving a larger number of speaker pairs and listeners are needed to refine the technique and assess its performance. Some of the other multivariate regression techniques may also be explored for speaker transformation. A comparison of the performance of speaker transformation using multivariate regression and some of the other speaker transformation techniques also needs to be carried out.

References

[1] D. G. Childers, B. Yegnanarayana, and Wu Ke, “Voice conversion: factors responsible for quality,” in *Proc. ICASSP*, 1985, Tampa, Florida, vol. 1, pp.748-751.

[2] H. Matsumoto, S. Hiki, T. Sone, and T. Nimura, “Multidimensional representation of personal quality of vowels and its acoustical correlates,” *IEEE Trans. Audio and Electro.*, vol. 21, pp. 428-436, 1973.

[3] H. Mizuno and M. Abe, “Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectral tilt,” *Speech Commun.*, vol. 16, no. 2, pp. 153-164, 1995.

[4] C. H. Wu, C. C. Hsia, T. H. Liu, and J. F. Wang, “Voice conversion using duration-embedded Bi-HMMs for expressive speech synthesis,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1109-1116, 2006.

[5] Y. Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech and Audio Process.*, vol. 6, no. 2, pp. 131-142, 1998.

[6] L. M. Arslan, “Speaker transformation algorithm using codebooks (STASC),” *Speech Commun.*, vol. 28, no. 3, pp. 211-226, 1999.

[7] F. Villavicencio, A. Robel, and X. Rodet, “Extending efficient spectral envelope modeling to mel-frequency based representation,” in *Proc. ICASSP*, 2008, Las Vegas, NV, vol. 1, pp. 1625-1628.

[8] D. Sundermann, A. Bonafonte, H. Ney, and H. Hoege, “A study on residual prediction techniques for voice conversion,” in *Proc. ICASSP*, 2005, Philadelphia, PA, vol. 1, pp.512-516.

[9] E. Helander, J. Nurminen and M. Gabbouj, “LSF mapping for voice conversion with very small training sets,” in *Proc. ICASSP*, 2008, Las Vegas, NV, vol. 1, pp. 4669-4672.

[10] Y. Adachi, S. Kawamoto, S. Morishima, and S. Nakamura, “Perceptual similarity measurement of speech by combination of acoustic features,” in *Proc. ICASSP*, 2008, Las Vegas, NV, vol. 1, pp. 4861-4864.

[11] R. J. McAulay and T. F. Quatieri, “Phase modeling and its application to sinusoidal transform coding,” in *Proc. ICASSP*, 1986, Tokyo, Japan, vol. 1, pp. 1713-1716.

[12] K. Shikano, K. Lee, and R. Reddy, “Speaker adaptation through vector quantization,” in *Proc. ICASSP*, 1986, Tokyo, Japan, vol. 1, pp. 2643-2646.

[13] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana “Transformation of formants for speaker transformation using artificial neural networks,” *Speech Commun.*, vol. 16, no. 2, pp. 207-216, 1995.

[14] Y. Stylianou and O. Cappe, “A system for voice conversion based on probabilistic classification and a harmonic plus noise model,” in *Proc. ICASSP*, 1998, Seattle, WA, vol. 1, pp. 281 - 284.

[15] H. Ye and S. Young, “Quality-enhanced voice morphing using maximum likelihood transformation,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp.1301-1312, 2006.

[16] A. Rinscheid, “Voice conversion based on topological feature maps and time variant filtering,” in *Proc. Int. Conf. Spoken Language Process,*, 1996, Philadelphia, PA, vol. 3, pp. 1445-1448.

[17] M. Hashimoto and N. Higuchi, “Training data selection for speaker transformation using speaker selection and vector field smoothening,” in *Proc. Int. Conf. Spoken Language Process,*, 1996, Philadelphia, PA, vol. 3, pp. 1397-1400.

[18] Z. H. Jian and Z. Yang, “Voice conversion without parallel speech corpus based on mixtures of linear transform,” in *Proc. Int. Conf. Wireless Communications, Networking and Mobile Computing*, 2007, Shanghai, China, vol. 1, pp. 2825 – 2828.

[19] T. Masuda and M. Shozakai, “Cost reduction of training mapping function based on multistep voice conversion,” in *Proc. ICASSP*, 2007, Honolulu, HI, vol. 4, pp. 693-.696.

[20] H. Valbret, E. Moulines, and J. P. Tubach, “Voice transformation using PSOLA technique,” *Speech Comm.*, vol. 11, pp. 175-187, 1992.

[21] G. Bandoin and Y. Stylianou, “On the transformation of the speech spectrum for voice conversion,” in *Proc. Int. Conf. Spoken Language Process.*, 1996, Philadelphia, PA, vol. 3, pp. 1405-1408.

[22] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, “Voice conversion with smoothed GMM and MAP adaptation,” in *Proc. EuroSpeech*, 2003, Geneva, Switzerland, vol. 1, pp. 2413-2416.

[23] D. Rentzos, S. Vaseghi, and Q. Yan, “Voice conversion through transformation of spectral and intonation features,” in *Proc. ICASSP*, 2004, Montreal, Canada, vol. 1, pp.21-24.

[24] N. Iwahashi and Y. Sagisaka, “Speech spectrum transformation by speaker interpolation,” in *Proc. ICASSP*, 1994, Adelaide, Australia, vol. 1, pp. 461-464.

[25] N. Iwahashi and Y. Sagisaka, “Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks,” *Speech Commun.*, vol. 16, pp.139-151, 1995.

[26] L. Rabiner, B. H. Juang, Fundaments of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall, 1999.

[27] J. Laroche et. al., “HNS: Speech modification based on a harmonic + noise model," in *Proc. ICASSP*, 1993, Minneapolis, MN, vol. 2, pp. 550–553.

[28] P. K. Lehana and P. C. Pandey, “Harmonic plus noise model based speech synthesis in Hindi and pitch modification", in *Proc. 18th Int. Congress on Acoustics*, 2004, Kyoto, Japan, vol. 1, pp. 3333-3336.

[29] R. L. Branham, Scientific Data Analysis: An Introduction to Overdetermined Systems, London: Springer-Verlag, 1990.

[30] O. Cappe, J. Laroche, and E. Moulines, “Regularized estimation of cepstrum envelope from discrete frequency points”, in *Proc. EuroSpeech*, 1995, Madrid, Spain, pp. 451-454.

[31] J. Picone, “Signal modeling techniques in speech recognition”, *Proc. IEEE*, 1993, vol. 81, no. 9, pp. 1215-1247.

[32] C. S. Myers and L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, vol. 60, no. 7, pp. 1389-1409, 1981.

[33] J. MA and W. Liu, “Voice conversion based on joint pitch and spectral transformation with component group-GMM,” in *Proc. Int. Conf. Natural Language Proces. and Knowledge Eng.*, 2005, Wuhan, China, pp.199- 203.

[34] T. F. Quatieri and A. V. Oppenheim, “Iterative techniques for minimum phase signal reconstruction from phase or magnitude”, *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 6, pp. 1187-1193, 1981.

[35] P. C. Mahalanobis, "On the generalised distance in statistics," *Proc. National Institute of Sciences of India*, vol. 2, no. 1, pp. 49–55, 1936.

[36] J. Kreiman and G. Papcun, “Comparing, discrimination and recognition of unfamiliar voices,” *Speech Commun.*, vol. 10, no. 3, pp. 265-275, 1991.