Speaker Transformation Using Quadratic Surface Interpolation

Parveen K. Lehana and Prem C. Pandey
SPI Lab, Department of Electrical Engineering
Indian Institute of Technology Bombay
Powai, Mumbai 00076, India
{lehana, pcpandey}@ee.iitb.ac.in

***Abstract*-Speaker transformation is a technique that modifies a source speaker's speech to be perceived as if a target speaker has spoken it. Compared to statistical techniques, warping function based transformation techniques require less training data and time. The objective of this paper is to investigate the transformation using quadratic surface interpolation. Source and target utterances were analyzed using harmonic plus noise model (HNM) and harmonic magnitudes were converted to line spectral frequencies (LSFs). Transformation function was found using LSFs of the time aligned source and target frames using dynamic time warping. The transformed LSFs were converted back to harmonic magnitudes for HNM synthesis. This method was able to transform speech with satisfactory quality. Further, the results were better if pitch frequency was included in the frame vectors.**

I. Introduction

Speaker transformation is a technique that modifies a source speaker's speech to be perceived as if a target speaker has spoken it. This is carried out using a speech analysis-synthesis system, in which the parameters of the source speech are modified by a transformation function and resynthesis is carried out using modified parameters. The transformation function is obtained by analyzing the source and target speaker's utterances. Precise estimation of the transformation function is very difficult as there are many features of speech which are difficult to extract automatically, such as meaning of the passage and intention of the speaker [1], [2]. Mostly, the transformation function is derived using dynamics of the spectral envelopes of source and target speakers [3].

Instead of using the whole spectrum, only few formants can also be used for speaker transformation. The problem of using this method is that it requires automated estimation of the frequency, bandwidth, and amplitude of the formants, which can not be accurately estimated. Further, formant based transformation is not suitable for high quality synthesis [4]. Sinusoidal models also have been used for speech modification, but the results are not very encouraging [5]. Many researchers have used codebook mapping for speaker transformation [6]-[8]. In this approach, vector quantization (VQ) is applied to the spectral parameters of both the source and the target speakers. The two resulting VQ codebooks are used to obtain a mapping between source and target parameters. The quality of the converted speech using this method is mostly low as the parameter space of the converted envelope is limited to a discrete set of envelopes. A number of researchers have reported satisfactory quality of the transformed speech using Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), and Artificial Neural Networks (ANN) based transformation systems. The main difficulty with these methods is the dependence of the quality of the transformed speech on training and amount of data [9]-[12].

Iwahashi and Sagisaka [13] investigated speaker interpolation technique. Spectral patterns for each frame of the same utterances spoken by several speakers are stored in the transformation system. The spectral patterns are time-aligned using dynamic time warping (DTW). The values of the interpolation ratios are determined by minimizing the error between the interpolated and target spectra. Set of interpolation ratio is frame and target dependent. For generating the speech of the given target, it is gradually changed from frame to frame. The spectral vector for each frame of the source speech is compared with the stored spectral vectors to find the nearest one. The set of interpolation ratio for this frame and the given target is fetched from the database. The target speech is generated using the spectral parameters estimated by interpolation. Good results using this technique have been reported [14], with a reduction of about 25% in the distance between the speech spectra of the target speaker and the transformed as compared to that for the target speaker and the closest pre-stored speaker.

In dynamic frequency warping (DFW) for speaker transformation [15], spectral envelope and excitation are derived from the log magnitude spectra for source and target speaker. Then a warping function between the spectral envelopes is obtained, one for each pair of source-target spectral vectors within the class. An average warping function is obtained for each class of acoustic units and then it is modeled using a third order polynomial. The target speech is obtained by using an all-pole filter derived from modified envelope and modifying the excitation for adjusting the prosody. They have also used linear multivariate regression (LMR) based transformation between the cepstral coefficients of the corresponding classes in the acoustic spaces of the source and the target. The speech converted by both the methods had audible distortions. Although the number of parameters needed for mapping is lesser in DFW, the quality of the converted speech using LMR was reported to be better [15], [16]. The quality was assessed using ABX test with vowels and CVC.

Most of the techniques for speaker transformation discussed in this section can be grouped into four major categories:

frequency warping, vector quantization, statistical, and artificial intelligence based. Although the statistical and artificial intelligence based techniques try to capture the natural transformation function independent of the acoustic unit, these techniques need a lot of training data and time. Vector quantization is also associated with many problems, such as discrete nature of the acoustic space. It hampers the dynamic character of the speech signal and hence the converted speech loses naturalness. In frequency warping technique, the transformation function can be estimated using lesser data, but a different transformation function is needed for each acoustic class. Estimation of all acoustic classes requires a lot of speech material and computation power.

We have investigated the use of quadratic surface interpolation [17]-[19] for estimating the mapping between the source and the target acoustic spaces, for harmonic plus noise model (HNM) based speaker transformation. HNM is a variant of sinusoidal modeling of speech and divides the spectrum of the speech into two sub bands, one is modeled with harmonics of the fundamental and the other is simulated using random noise. HNM has been chosen as it provides high quality speech output, smaller number of parameters, and easy pitch and time scaling [20], [21]. The other advantage is that it can be used for concatenative synthesis with good quality of output speech.

In general, the system developed can be used for any speech transformation if proper amount of training data are provided for adaptation. Because of the time constraints of alignment of the source and target utterances for training of the model, the investigations have been restricted to vowels. This technique is explained in Section II. Methodology of the investigations is described in Section III. Results and conclusion are presented in Section IV and Section V, respectively.

II. Quadratic Surface Fitting

If a multidimensional function $g(w_1, w_2, \dots, w_m)$ is known only at q points, a quadratic surface $f(w_1, w_2, \dots, w_m)$ can be constructed such that it approximates the given function within some error $\varepsilon(w_1, w_2, \dots, w_m)$ at each point [17]-[19],

$$g({}^n w_1, {}^n w_2, \dots, {}^n w_m) = f({}^n w_1, {}^n w_2, \dots, {}^n w_m) + \varepsilon({}^n w_1, {}^n w_2, \dots, {}^n w_m), \qquad n = 0, 1, \dots, q-1 \tag{1}$$

The multivariate quadratic surface function can be written as

$$f(w_1, w_2, \dots, w_m) = \sum_{k=0}^{p-1} c_k \varphi_k(w_1, w_2, \dots, w_m) \tag{2}$$

where p is the number of terms in the quadratic equation formed by m variables, c_k represents coefficient of quadratic term k, and $\varphi_k(w_1, w_2, \dots, w_m)$ represents the term k itself. For example, this expression for 3 variables becomes

$$f(w_1, w_2, w_3) = c_0 + c_1 w_1 + c_2 w_2 + c_3 w_3 + c_4 w_1^2 + c_5 w_2^2 + c_6 w_3^2 + c_7 w_1 w_2 + c_8 w_2 w_3 + c_9 w_3 w_1 \tag{3}$$

The coefficients c_k are determined for minimizing the sum of squared errors

$$E(c_0, \cdots, c_{p-1}) = \sum_{n=0}^{q-1} \left[g({}^n w_1, {}^n w_2, \dots, {}^n w_m) - f({}^n w_1, {}^n w_2, \dots, {}^n w_m; c_0, \cdots, c_{p-1}) \right]^2 \tag{4}$$

Now (1) and (2) can be combined to form the matrix system of equations

$$\mathbf{B} = \mathbf{AZ} + \varepsilon \tag{5}$$

where the matrices **B, A**, **Z**, and $\boldsymbol{\varepsilon}$ are given by

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

$$A_{n,k} = \varphi_k({}^n w_1, {}^n w_2, \dots, {}^n w_m),$$

$$0 \le n \le q-1, 0 \le k \le p-1$$

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

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

If the number of given data points $q \ge p$, then (3) can be solved for minimizing the error as given in (4), giving the following solution

$$\mathbf{Z}=(\mathbf{A}^\mathrm{T}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}\mathbf{B} \tag{6}$$

where matrix $(\mathbf{A}^\mathrm{T}\mathbf{A})^{-1}\mathbf{A}^\mathrm{T}$ is known as pseudo–inverse of **A** [19].

III. Methodology

A. Analysis-parameter modification-synthesis

Investigations were carried out using recordings of a passage read by five speakers (two males and three females) in the age group of 20-23 years having Hindi as their mother tongue. The recordings were carried out in an acoustically treated room. The total recordings were of about 30-minute duration. The sampling frequency and number of bits used for quantization were 10 k sa/s and 16 bits, respectively. The ten vowels shown in Table 1 were extracted from these recordings taking the context same for all the speakers. The labeled vowels for the speakers were aligned manually in the same sequence for the source and the target and HNM analysis was performed for obtaining parameters such as pitch, voiced/unvoiced decision, maximum voiced frequency, harmonic magnitudes, harmonic phases, and noise parameters (linear predictive coefficients and energy contour) [20], [21].

The harmonic magnitudes were converted to autocorrelation coefficients using Wiener-Khintchine theorem [22]. The autocorrelation coefficients were transformed to line spectral frequencies (LSFs) [23]. The order of the LSFs was fixed as 20. The LSFs are related to formant frequencies and bandwidths, and show good linear interpolation properties [23]. Hence, target vectors can be assumed as linear combinations of source vectors. Further, LSFs can be reliably estimated using a limited dynamic range, and estimation errors have localized effects; a wrongly estimated value of LSF only affects the neighboring spectral components [23].

Before obtaining the transformation function, a number of frames in source and target training data were aligned using dynamic time warping (DTW) [15]. For each aligned frame of source and target speakers, feature vectors consisting of 20 LSFs and pitch frequency were constructed for each frame. Let the source frame vector $\mathbf{X}$ and the target frame vector $\mathbf{Y}$ be

$$\mathbf{X} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{20} \end{bmatrix} \tag{7}$$

$$\mathbf{Y} = \begin{bmatrix} y_0 & y_1 & \cdots & y_{20} \end{bmatrix} \tag{8}$$

Each component in the target feature vector is modeled as a multivariate quadratic function of source components

$$y_i = f_i(x_0, x_1, ..., x_{20}),\ i = 0,1,...,20 \tag{9}$$

Coefficients for these quadratic functions were obtained using (6), providing the mapping from source to target frame vectors.

A few vowels from the speech of the source speaker were taken. These vowels were different from the vowels used for training. These vowels were analyzed using HNM and frame vectors were calculated for each frame. The frame vectors for each frame were transformed using the mapping in (9) with coefficients obtained from the training data. Transformed LSFs were used for obtaining LPC spectrum and sampling of it at modified harmonic frequencies provided the modified harmonic magnitudes. Harmonic phases were estimating from the harmonic magnitudes by assuming minimum phase system [24]. These modified HNM parameters were used for resynthesizing the target speech.

In this paper, we are presenting the investigations regarding transformation of harmonic part of the vowels using HNM based analysis-synthesis. As HNM divides the speech into harmonic and noise parts, both parts should be transformed independently for speech involving phonemes other than vowels. The transformation of harmonic part of all phonemes is similar, but extra steps are needed for transforming noise part. In our present investigations, we are simulating the noise part using only the magnitudes and frequencies of the perceptually important peaks in the spectra. The magnitudes of the frequencies other than these peaks are replaced with zeroes and this spectrum is converted to LSFs before finding the transformation function for the noise part. It is to be noted that transformation functions based on mel frequency cepstrum coefficients (MFCCs) and harmonic magnitudes themselves also need to be investigated.

B. Evaluation

To assess the extent of the closeness of the transformed speech to that of the target, both subjective and objective evaluations were carried out.

Objective evaluation has been done at two levels: for transformed parameters and for the transformed spectra. Mahalanobis distance has been reported to be an efficient measure for multidimensional pattern comparisons [25]-[30] and has been often used for distance in parametric space in speech research [29], [30]. We have used it for estimating the errors in the transformed LSF vectors and the corresponding target LSF vectors. Log spectral distance measure is generally used to estimate the closeness of the spectrum of the modified speech and the spectrum of the target speech [31]-[35]. It is calculated between the spectral values for each frame, and then averaged across frames [31]

$$D = \left[\frac{1}{K} \sum_{k=0}^{K-1} \left[20\log\left|S(k)\right| - 20\log\left|S'(k)\right| \right]^2 \right]^{1/2} \tag{10}$$

where $S(k)$ and $S'(k)$ are the DFT values of the signals for index k with $K = 4096$.

For subjective evaluation of the closeness of the transformed and target speech, generally, ABX test has been often used [4], [6], [36]-[40]. In this test, the subject is asked to match the speech stimuli (X) with either source or target stimuli. The source and target stimuli are represented by A and B. The subjects do not know whether the source, target, or modified stimulus is presented at A, B, or X. For this, an automated test setup employing randomized presentations and a GUI for controlling the presentation and recording the responses was used. In each presentation, sound X could be randomly selected as source, target, or the modified speech. The subject had to select sound A or sound B as the best match to presentation X. Either source or the target sounds were randomly made A or B. Subject could listen to the sounds more than once before finalizing the response and proceeding to the next presentation. In a test, each vowel appeared 5 times. This test was conducted with 10 subjects with normal hearing.

IV. RESULTS

In order to assess the level of distortion in the analysis-transformation-synthesis process, the transformation was carried out for the vowels of the same speaker as both source and target. Informal listening tests have confirmed that the identity of the speaker was not disturbed, except some loss of quality due to phase estimation assuming minimum phase system.

For investigating the speaker transformation abilities of the quadratic surface interpolation method, the transformation function was estimated by using quadratic surface fitting in the parametric space (normalized F_0 and LSF) of the source and target aligned vowels by DTW. Using this function, the vowels not included in the training sets, were transformed and Mahalanobis distances between the source-target (ST), target-synthesized target (TT'), and source-synthesized target (ST') pairs in parametric space were calculated. A plot of the distance for consecutive frames of three cardinal vowels, in Fig.1, shows that the distance between target and the transformed vowel (TT') is less than the original distance between the source and the target. This implies improved transformation from the source to target. It has been observed that the reduction of distance between transformed vowel and the target is maximum for /*a*/ and minimum for /*i*/. Further, this distance is slightly less for the transformation taking pitch as one of the feature components.

Investigations were also carried out using the harmonic magnitude envelopes of the source (S), transformed source

(T'), and the target (T). These envelopes for the three cardinal vowels are shown in Fig. 2. It is clear from this figure that the harmonic magnitudes for the transformed source and the corresponding target are very close to each other. Log spectral distances between the spectra of source and the target (ST) and the target and the converted speech (TT') for various vowels are given In Table 1. It is seen that conversion by including F_0 in the feature vector results in an additional reduction in the distances.

Subjective evaluation showed that the transformed speech was satisfactory in quality and it sounded near to that of the target speech. Analysis of the scores from the XAB listening test showed that more than 90 % responses labeled the modified speech as that of the target.

V. Conclusion

Investigations were carried out to explore the use of quadratic surface interpolation for speaker transformation using HNM based analysis/synthesis. Results from objective and subjective evaluation showed that the method was able to transform vowels with satisfactory quality. Further, the results improved if pitch frequency was included in the feature vectors. We are presently investigating the use of this technique for continuous speech.

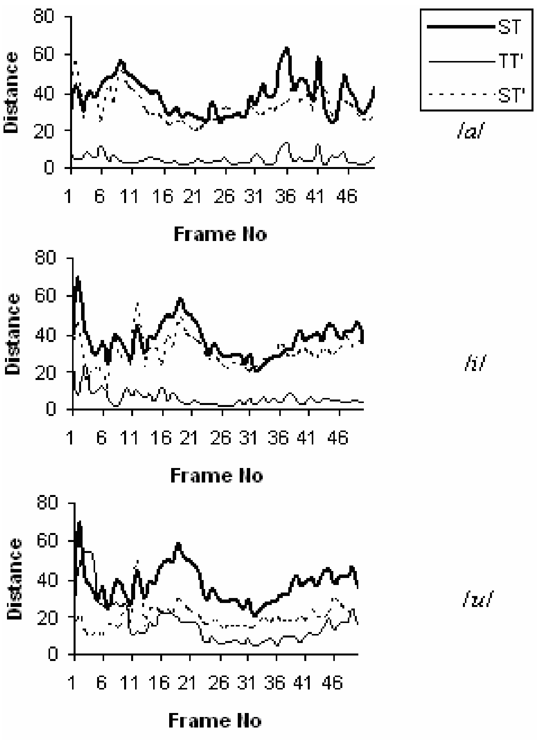

Fig. 1. Mahalanobis distance between the LSFs of source-target (ST), target-modified source (TT'), and source-modified source (ST') cardinal vowel pairs.

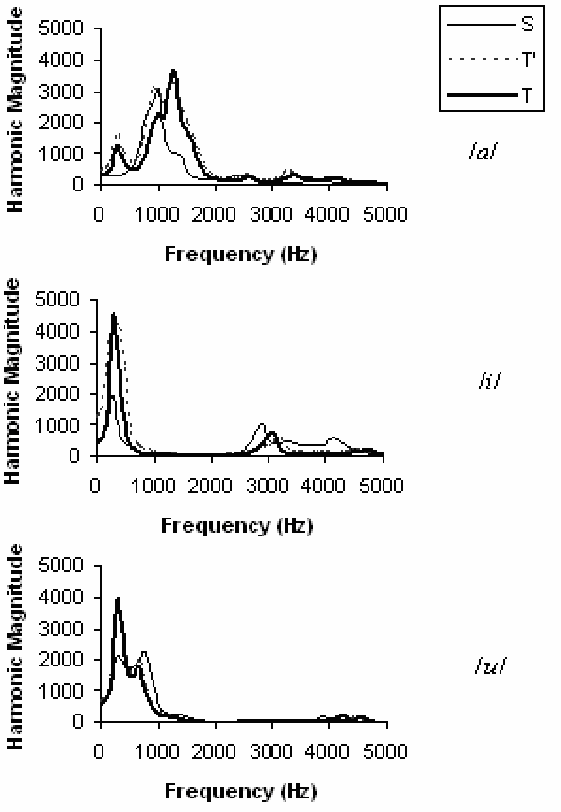

Fig. 2. Harmonic magnitude envelopes for the source (S), modified source (T'), and the target (T) cardinal vowels.

Table 1. Log spectral distances between the vowel spectra

Vowel	Log spectral distance		
	ST	TT'	
		Without F0	With F0
ʌ अ	11.46	5.61	5.32
ɑ आ	9.77	4.52	4.26
ɪ इ	8.25	4.13	3.79
I ई	9.03	4.57	4.19
ɛ ए	8.78	4.35	4.02
æ ऐ	11.93	5.03	4.63
ʊ उ	8.79	4.45	4.15
u ऊ	8.46	4.72	4.16
oʊ ओ	9.21	4.48	4.07
aʊ ऑ	9.99	4.49	4.12

REFERENCES

[1] W. Endres, W. Bambach, and G. Fl¨osser, “Voice spectrograms as a function of age, voice disguise, and voice imitation,” *J. Acoust. Soc. Amer.*, vol. 49, pp. 1842–1848, 1971.

[2] M. R. Sambur, “Selection of acoustic features for speaker identification,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 176–182, 1975.

[3] H. Kuwabara and Y. Sagisaka, “Acoustic characteristics of speaker individuality: Control and conversion,” *Speech Commun.*, vol. 16, pp. 165–173, Feb. 1995.

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

[5] J. Wouters and M. W. Macon, “Spectral modification for concatenative speech synthesis,” in *Proc. ICASSP 2000*, pp. II.941–II.944.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” in *Proc. ICASSP 1988*, New York, NY, pp. 655–658.

[7] M. Abe, S. Nagamuka, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” *J. Acoust. Soc. Japan.*, vol. E-11, pp. 71–77, Mar. 1990.

[8] K. Shikano, K. Lee, and R. Reddy, “Speaker adaptation through vector quantization,” in *Proc. ICASSP 1986*, pp. 2643–2646.

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

[10] L. D. Paarmann and M. D. Guiher, “A nonlinear spectrum compression algorithm for the hearing impaired,” in *Proc. IEEE Fifteenth Annual Bioengineering Conf. 1989*, pp. 21-22, 1989.

[11] L. M. Arslan and D. Talkin, “Speaker transformation using sentence HMM based alignments and detailed prosody modification,” in *Proc. ICASSP 1998*, pp. 289-292.

[12] A. Verma and A. Kumar, “Voice fonts for individuality representation and transformation,” *ACM Trans. Speech, Language Processing*, vol. 2, no. 1, pp. 1-19, 2005.

[13] 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, Feb. 1995.

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

[15] H. Valbret, E. Moulines, and J. P. Tubach, “Voice transformation using PSOLA techniques,” *Speech Commun.*, vol. 11, pp. 175–187, June 1992.

[16] E. Moulines and F. Charpentier, “Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Commun.*, vol. 9, pp. 453–467, 1990.

[17] G. M. Philips, *Interpolation and Approximation by Polynomials.* New York: Springer-Verlag, 2003.

[18] S. A. Dyer and J. S. Dyer, “Cubic-spline interpolation: part 1,” *IEEE Instrum. Meas. Mag.,* vol. 4, no. 1, pp. 44–46, 2001.

[19] R. L. Branham Jr., *Scientific Data Analysis: An Introduction to Overdetermined Systems.* New York: Springer-Verlag, 1990.

[20] J. Laroche, Y. Stylianou, and E. Moulines, “HNS: Speech modification based on a harmonic + noise model,” in *Proc. ICASSP 1993*, vol. 2, pp. 550–553.

[21] P. K. Lehana and P. C. Pandey, “Speech synthesis in Indian languages”, in *Proc. Int. Conf. on Universal Knowledge and Languages 2002* (Goa, India, Nov 2002), paper no. pk1510.

[22] K. M. Aamir, M. A. Maud, A. Zaman, and A. Loan, **“**Recursive computation of Wiener-Khintchine theorem and bispectrum**,”** *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E89-A, no. 1, pp. 321-323, 2006.

[23] K. K. Paliwal, “Interpolation properties of linear prediction parametric representations,” in *Proc. Eurospeech 1995*, pp. 1029–1032.

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

[25] T. Takeshita, S. Nozawa, and F. Kimura, "On the bias of Mahalanobis distance due to limited sample size effect," in *Proc. 2nd IEEE Int.Conf. on Document Analysis and Recognition*, 1993, pp. 171-174.

[26] J. M. Yih, D. B. Wu, and C. C. Chen, "Fuzzy C-mean algorithm based on Mahalanobis distance and new separable criterion," in *Proc. IEEE Int. Conf. on Machine Learning and Cybernetics*, 2007, pp. 1851-1855.

[27] J.C.T.B. Moraes, M. O. Seixas, F. N. Vilani, and E. V. Costa, "A real time QRS complex classification method using Mahalanobis distance," in *Proc. IEEE Int. Conf. on Computers in Cardiology*, 2002, pp. 201-204.

[28] T. Kamei, "Face retrieval by an adaptive Mahalanobis distance using a confidence factor," in *Proc. IEEE Int. Conf. on Image Processing*, 2002, vol. 1, pp. 153-156.

[29] G. Chen, H. G. Zhang, and J. Guo, "Efficient computation of Mahalanobis distance in financial hand-written Chinese character recognition" in *Proc. IEE Int. conf. on Machine Learning and Cybernetics*, 2007,vol. 4,pp. 2198-2201.

[30] J. P. Campbell, “Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, pp. 1437-1462, Sept. 1997.

[31] A. Verma and A. Kumar, “Voice fonts for individuality representation and transformation,” *ACM Trans. Speech, Language Processing*, vol. 2, no. 1, pp. 1-19, 2005.

[32] K. K. Soong and B. H. Juang, "Optimal quantization of LSP parameters," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 1, pp. 15-24, 1993.

[33] T. Ramabadran, A. Smith, and M. Jasiuk, "An iterative interpolative transform method for modeling harmonic magnitudes," in *Proc. IEEE Workshop on Speech Coding*, 2002, pp. 38-40.

[34] J. Samuelsson and J. H. Plasberg, "Multiple description coding based on Gaussian mixture models," *IEEE Signal Processing Letters*, vol. 12, no. 6, pp. 449-452, 2005.

[35] E. R. Duni and B. D. Rao, "A high-rate optimal transform coder with gaussian mixture companders," *IEEE Trans. Audio, Speech and Language Processin*g, vol. 15, no. 3,pp. 770-783, 2007.

[36] Y. Stylianou, O. Cappe, “A system for voice conversion based on probabilistic classification and a harmonic plus noise model,” in *Proc. ICASSP 1998*.

[37] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai"Voice characteristics conversion for HMM-based speech synthesis system," in *Proc.ICASSP-1997*, vol. 3, pp. 1611-1614.

[38] L. Cheng and J. Jang, "New refinement schemes for voice conversion," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, 2003, vol. 2, pp. 725-728.

[39] O. Salor and M. Demirekler, "Spectral modification for context-free voice conversion using MELP speech coding framework," in *Proc. IEEE Int. Sym. on Intelligent Multimedia, Video and Speech Processing*, 2004, pp. 314-317.

[40] K. Furuya, T. Moriyama, and S. Ozawa, "Generation of speaker mixture voice using spectrum morphing," in *Proc. IEE Conf. on Multimedia and Expo,* 2007 pp. 344-347.