CONVERSION FUNCTION CLUSTERING AND SELECTION FOR EXPRESSIVE VOICE CONVERSION

Chi-Chun Hsia, Chung-Hsien Wu, and Jian-Qi Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, ROC
{shiacj, chwu, glinwu }@csie.ncku.edu.tw

ABSTRACT

In this study, a conversion function clustering and selection approach to conversion-based expressive speech synthesis is proposed. First, a set of small-sized emotional parallel speech databases is designed and collected to train the conversion functions. Gaussian mixture bi-gram model (GMBM) is adopted as the conversion function to model the temporal and spectral evolution of speech. Conversion functions initially constructed from the parallel sub-syllable pairs in the speech database are clustered based on linguistic and spectral information. Subjective and objective evaluations with statistical hypothesis testing were conducted to evaluate the quality of the converted speech. The results show that the proposed method exhibits encouraging potential in conversion-based expressive speech synthesis.

Index Terms— *Speech synthesis, voice conversion, Gaussian mixture bi-gram model, linguistic information, expression.*

1. INTRODUCTION

For high quality expressive speech synthesis, concatenative text-to-speech (TTS) systems have been realized with large-sized expressive speech databases. To overcome the obstructions resulted from the requirement of large-sized speech databases, voice conversion (also called voice morphing) methods have been adopted as a post-processing module of the expressive text-to-speech systems.

In the past decade, stochastic approaches have dominated the development of voice conversion systems. The Gaussian mixture model (GMM) based voice conversion is performed using a frame-by-frame mechanism with the time-independence assumption and disregards spectral envelope evolution. Toda et al. [1] introduced a GMM-based framework considering global variance. In addition, Hidden Markov model (HMM) based methods have recently been proposed [2] [3]. The state transition property in HMM-based methods presents a good approximation of the spectral envelope evolution in the time axis. However, the HMM-based method is too complicated and requires large amount of training data for robust parameter estimation.

Besides stochastic approaches, Duxans et al. [2] considered the phonetic information available in a TTS system for each frame, including phone, vowel/consonant flag, point of articulation, manner and voicing, by adopting a classification and regression tree (CART). However, CART is a sequence of hard decision processes; neither a distance nor a similarity score is output. The framework of CART is designed as a frame-based approach. Each splitting in a node regards only the data in that node. Also, all the linguistic features are treated equal, and do not consider the acoustic similarities between conversion functions.

This work presents a Gaussian mixture bi-gram model (GMBM) [4] based voice conversion model to characterize the temporal and spectral evolution in the conversion process. Figure 1 shows the flowchart of the conversion model construction. The STRAIGHT algorithm, proposed by Kawahara et al. [5], is adopted to estimate the spectrum of source speech. After the alignment using dynamic time warping (DTW) algorithm, the initial models (initial conversion functions) for all source and target paired speech segments of the same sub-syllable are trained by Expectation Maximization (EM) algorithm [6]. The linguistic feature vector corresponding to each initial model is extracted and calculated from the text. The initial models are clustered by the K-means algorithm using spectral and linguistic similarity between conversion functions. Linguistic similarities accounting for the context of different functions are estimated on linguistic feature vectors using cosine measure. Since the conversion functions are derived from the joint distributions on spectral feature space, Kullback-Leibler (KL) divergence [7] and sigmoid function are used to calculate the spectral similarities between conversion functions. For conversion function construction, a small speech database was designed and collected for each emotion to cover all 150 sub-syllables, including 112 context-dependent initial parts and 38 final parts [8] in Mandarin. The synthetic neutral utterances of a TTS system were used as the input speech of the expressive voice conversion model.

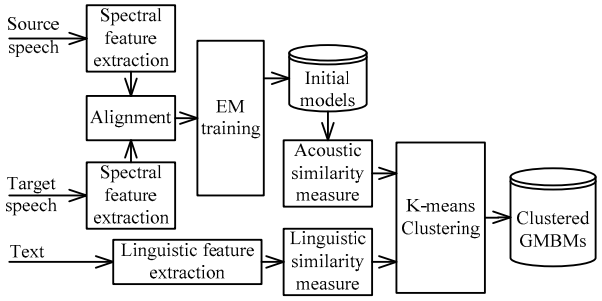

Figure 1: Flowchart of conversion model construction.

2. GAUSSIAN MIXTURE BI-GRAM MODEL

In voice conversion, joint density method has been introduced by modeling the source and target paired spectral feature vectors in a joint GMM distribution [9]. In this study, the Gaussian mixture bi-gram model is adopted to characterize the temporal and spectral evolution in the conversion function. The

probability density function of the joint random variable $\mathbf{z}_t = (\mathbf{y}_t, \mathbf{y}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t-1})$ is modeled by a mixture model as:

$$p(\mathbf{z}_t) = \sum_{m=1}^{M} w_n p(\mathbf{y}_t, \mathbf{y}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t-1}, m) = \sum_{m=1}^{M} w_m N(\mathbf{z}_t, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \quad (1)$$

where $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are the mean vector and covariance matrix for mixture m, respectively. The conversion function is then given by:

$$f(\mathbf{y}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t-1}) = E[\mathbf{y}_t \mid \mathbf{y}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t-1}] \quad (2)$$

$$= E[\mathbf{y}_t] + \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{y}_t,\mathbf{y}_{t-1}} \\ \boldsymbol{\Sigma}_{\mathbf{y}_t,\mathbf{x}_t} \\ \boldsymbol{\Sigma}_{\mathbf{y}_t,\mathbf{x}_{t-1}} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{y}_{t-1},\mathbf{y}_{t-1}} & \boldsymbol{\Sigma}_{\mathbf{y}_{t-1},\mathbf{x}_t} & \boldsymbol{\Sigma}_{\mathbf{y}_{t-1},\mathbf{x}_{t-1}} \\ \boldsymbol{\Sigma}_{\mathbf{x}_t,\mathbf{y}_{t-1}} & \boldsymbol{\Sigma}_{\mathbf{x}_t,\mathbf{x}_t} & \boldsymbol{\Sigma}_{\mathbf{x}_t,\mathbf{x}_{t-1}} \\ \boldsymbol{\Sigma}_{\mathbf{x}_{t-1},\mathbf{y}_{t-1}} & \boldsymbol{\Sigma}_{\mathbf{x}_{t-1},\mathbf{x}_t} & \boldsymbol{\Sigma}_{\mathbf{x}_{t-1},\mathbf{x}_{t-1}} \end{bmatrix}^{-1} \left(\begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{bmatrix} - E\left(\begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{bmatrix} \right) \right)$$

With the assumption that $\mathbf{y}_t$ is independent of $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ is independent of $\mathbf{y}_{t-1}$, the conversion function is further simplified as:

$$f(\mathbf{y}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t-1}) = \quad (3)$$

$$E[\mathbf{y}_t] + \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{y}_t,\mathbf{y}_{t-1}} \\ \boldsymbol{\Sigma}_{\mathbf{y}_t,\mathbf{x}_t} \\ \mathbf{0} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{y}_{t-1},\mathbf{y}_{t-1}} & \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{y}_{t-1},\mathbf{x}_{t-1}} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{x}_t,\mathbf{x}_t} & \boldsymbol{\Sigma}_{\mathbf{x}_t,\mathbf{x}_{t-1}} \\ \boldsymbol{\Sigma}_{\mathbf{x}_{t-1},\mathbf{y}_{t-1}} & \boldsymbol{\Sigma}_{\mathbf{x}_{t-1},\mathbf{x}_t} & \boldsymbol{\Sigma}_{\mathbf{x}_{t-1},\mathbf{x}_{t-1}} \end{bmatrix}^{-1} \left(\begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{bmatrix} - E\left(\begin{bmatrix} \mathbf{y}_{t-1} \\ \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{bmatrix} \right) \right)$$

3. CONVERSION FUNCTION CLUSTERING AND SELECTION

3.1 Conversion Function Clustering

For each sub-syllable, one conversion model containing multiple conversion functions is trained using K-means algorithm by the following steps.

1) For each source-target-paired speech segments of the same sub-syllable label in the parallel speech database, the conversion function f_i, $1 \le i \le I$, is trained using the EM algorithm [6]. The corresponding joint distribution and linguistic feature vector are denoted by g_i and $\mathbf{l}_i$, respectively.

2) Conversion functions F_j, $1 \le j \le J$, are randomly selected as the initial conversion functions for each cluster with the corresponding joint distribution G_j and linguistic feature vector $\mathbf{L}_j$.

3) Calculate the similarity to F_j for each f_i by:

$$Sim(f_i, F_j) = \gamma \cdot S_{spectral}(g_i, G_j) + (1-\gamma) \cdot S_{linguistic}(\mathbf{l}_i, \mathbf{L}_j) \quad (4)$$

where $S_{spectral}(g_i, G_j)$ and $S_{linguistic}(\mathbf{l}_i, \mathbf{L}_j)$ denote the acoustic and linguistic similarities, respectively. γ is a weighting factor. For each f_i the most similar conversion function, F_j, is selected and $b(i)$ is set to j.

4) Re-estimate the conversion function F_j for each cluster by the EM algorithm using the speech data of the conversion function f_i with $b(i) = j$, and re-calculate the corresponding joint distribution G_j and linguistic feature vector $\mathbf{L}_j$.

5) Repeat steps 3) and 4) until there is no change in the assignments of $b(i)$ of two successive iterations.

3.2 Linguistic Similarity

The linguistic feature vector for each conversion function is represented by $\mathbf{l}_i = [l_{i,1}, l_{i,2}, \cdots, l_{i,M}]$, where M denotes the total number of linguistic features. Each element $l_{i,m}$, $1 \le m \le M$, is given in the form similar to the term-frequency-inverse-document-frequency (*tf-idf*) used in the field of information retrieval as:

$$l_{i,m} = \begin{cases} (1 + \log(freq_{i,m})) \times (\log(K/N_m)), & \text{if } freq_{i,m} \ge 1 \\ 0, & \text{if } freq_{i,m} = 0 \end{cases} \quad (5)$$

where $freq_{i,m}$ is the appearance number of the *m-th* linguistic feature in the training data of the conversion function f_i. K is the total number of conversion functions. N_m denotes the number of functions in which the *m-th* linguistic feature appears. If the *m-th* linguistic feature appears in the database, the element $l_{i,m}$ is assigned by the first clause of Eq. (5); otherwise it is set to zero. The formula $\log K/N_m = \log K - \log N_m$ gives a full weight to linguistic features that appear in one conversion function ($\log K - \log N_m = \log K - \log 1 = \log K$). A linguistic feature that appears in all conversion functions has zero weight ($\log K - \log N_m = \log K - \log K = 0$). The linguistic similarity between two conversion functions is estimated by the cosine measure between linguistic feature vectors $\mathbf{l}_i$ and $\mathbf{L}_j$ as:

$$S_{linguistic}(\mathbf{l}_i, \mathbf{L}_j) = \cos(\mathbf{l}_i, \mathbf{L}_j) = (\mathbf{l}_i \cdot \mathbf{L}_j) / (\|\mathbf{l}_i\| \cdot \|\mathbf{L}_j\|) \quad (6)$$

3.3 Spectral Similarity

Since the conversion functions are derived from the joint density of source and target acoustic feature vectors, the spectral similarity between conversion functions f_i and F_j is estimated by the KL divergence on their corresponding joint probability density function g_i and G_j, respectively, and normalized using the sigmoid function as:

$$S_{spectral}(g_i, G_j) = 1 - 1/(1 + \exp(-\alpha \cdot D_{KL}(g_i, G_j))) \quad (7)$$

where α is a slope parameter. $D_{KL}(g_i, G_j)$ denotes the symmetric KL divergence between two distributions g_i and G_j, and is defined as:

$$D_{KL}(g_i, G_j) = (KL(g_i \| G_j) + KL(G_j \| g_i))/2 \quad (8)$$

$KL(g_i \| G_j)$ is the KL divergence between two distributions. As the mixture model is adopted, the KL divergence can be approximated by:

$$KL(g_i \| G_j) \approx \sum_n \alpha_n \min_m KL(g_{i,n} \| G_{j,m}) \quad (9)$$

where $g_i = \sum_n \alpha_n g_{i,n}$ and $G_j = \sum_m \beta_m G_{j,m}$ are two mixture models with mixture weights α_n and β_m, respectively. When the Gaussian distribution is adopted for each component, the KL divergence can be calculated as:

$$KL\left(N(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\,\|\,N(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)\right)$$
$$=\frac{1}{2}\left(\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}+Tr\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-d\right) \quad (10)$$

where $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean vectors $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ with dimension d. The GMM-based method is described below as the baseline.

3.4 Selection Process

The function selection process is shown in Fig. 2. The spectral feature vectors for each new speech segment are extracted by the STRAIGHT algorithm. Each candidate conversion function F_j is described by its linguistic feature vector $\mathbf{L}_j$ and source-target joint distribution G_j trained from the training data. The similarity between the input speech segment $\mathbf{X}$ and the candidate conversion function F_j is measured as a weighted sum of spectral and linguistic similarities as shown in Eq. (4). Since the spectral similarity is measured on joint distributions, for the input speech segment, the candidate converted target feature vectors $\tilde{\mathbf{Y}}_j$ are calculated according to the candidate conversion function F_j. The joint distribution of the input source $\mathbf{X}$ and the converted target feature vectors $\tilde{\mathbf{Y}}_j$ is estimated by the EM algorithm, and is used to calculate the spectral similarity to the joint distribution, G_j which belongs to F_j. The linguistic similarity is estimated between the linguistic feature vector of source $\mathbf{X}$ and the linguistic feature vector $\mathbf{L}_j$ which belongs to F_j. The conversion function with the highest weighted sum of spectral and linguistic similarities is selected as the conversion function for voice conversion.

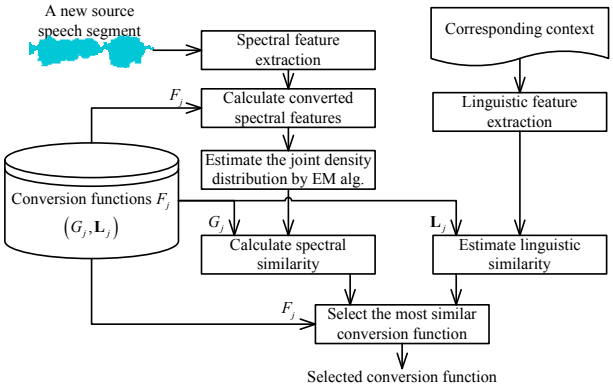

Figure 2: Block diagram for conversion function selection.

Table I: Used linguistic features

Feature level	Features	Number of types
Sub-syllable	Sub-syllable class identity of current sub-syllable (6 types for INITIAL part and 17 types for FINAL part)	23
	Sub-syllable class identity of proceeding sub-syllable	23
	Sub-syllable class identity of succeeding sub-syllable	23
Syllable	Tone type (Tone 1 to Tone 4 and neutral tone)	5
	Position in a word (the first one, the last one, in the middle, or the word has only one syllable)	4
Word	Part-of-Speech	44

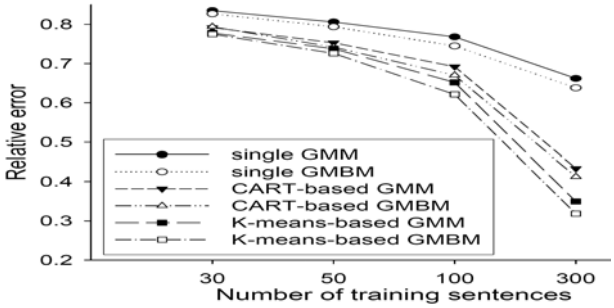

Figure 3: Relative error for the number of training sentences.

4. EXPERIMENTAL RESULTS

This study adopts happiness, sadness and anger as the target emotions. Three phonetically balanced, small-sized speech databases, each for one emotion, were designed and collected to train the voice conversion models. Each database was designed to include all the 150 sub-syllables in Mandarin and resulted in a size of 300 sentences. The speaker was a female radio announcer, and was familiar with our study. All utterances were recorded at a sampling rate of 22.05 kHz and 16 bit resolution. For feature extraction, the mel-frequency cepstral coefficients (MFCCs) were calculated from the smoothed spectrum extracted by the STRAIGHT algorithm. The analysis window was 23 ms with a window shift of 8 ms. The order of cepstral coefficients was set to 45. Table I shows the linguistic features used to calculate the linguistic similarity, which includes the features in sub-syllable, syllable and word levels [10].

4.1 Objective test

Initially, each sentence in the test set was synthesized by the TTS system. The synthesized utterances were further converted into expressive speech using GMBM-based conversion function clustered by the K-means algorithm (so called K-means-based GMBM). The conversion function for GMBM was simplified where all the covariance and cross-covariance matrices were diagonal. The conversion models were built for each of the 150 sub-syllables. The maximum number of mixture components was set to 128 for each mixture model. All the 300 parallel utterances were used as the training and test data for each emotion type. The performance index used for testing is:

$$relative\ error = 1/M\sum_{m=0}^{M-1}\left(D(\tilde{\mathbf{y}}_m,\mathbf{y}_m)/D(\mathbf{x}_m,\mathbf{y}_m)\right) \quad (11)$$

where M is the total frame number of the source speech. $\mathbf{x}_m$, $\mathbf{y}_m$ and $\tilde{\mathbf{y}}_m$ are the *m-th* frames of the source, aligned target and converted speech, respectively. $D(\cdot)$ denotes the log-spectral distortion. The weighting factors for K-means-based GMBM were set to 0.3, 0.4 and 0.3 for happiness, sadness and anger, respectively, in the following experiments. The weighting factors for K-means-based GMM were set to 0.4, 0.3 and 0.3 for happiness, sadness and anger, respectively. Figure 3 shows the average relative error as a function of the number of training sentences. Incorporating temporal information in the conversion process yields lower relative error than the GMM-based method. The proposed k-means-based framework also results in lower distortion than the CART-based methods.

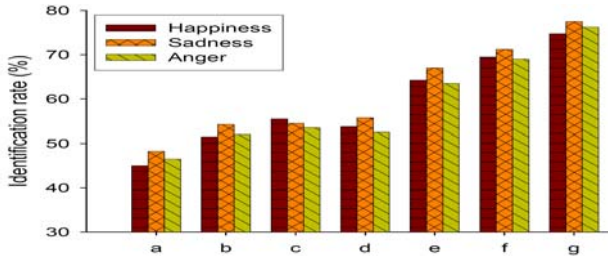

Figure 4: Identification results for different conversion methods.

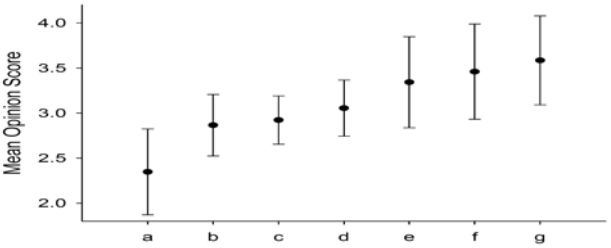

Figure 5: Mean opinion score (MOS) for different methods.

4.2 Subjective test

In order to evaluate the performance of spectral conversion as a post-processing module of the TTS system, a GMM-based prosody conversion model along with a pitch target model [11] were adopted to convert the pitch contour for each syllable. The prosody conversion model is constructed as the regression on the joint GMM distribution of source and target aligned prosody parameters. Each sentence in the test set were synthesized by the TTS system and further converted using the following conversion method for each sub-syllable in each emotion type,

a) Single GMBM spectral conversion function,
b) CART-based GMBM spectral conversion functions,
c) K-means-based GMBM spectral conversion functions,
d) Prosody conversion,
e) Single GMBM spectral conversion function + prosody conversion,
f) CART-based GMBM spectral conversion functions + prosody conversion and
g) K-means-based GMBM spectral conversion functions + prosody conversion.

All the 300 sentence pairs were used to train the spectral conversion models and the GMM-based prosody conversion models for each emotion type. The total number of utterances presented to each listener was 420 (3*7*20). A double-blind experiment was conducted in the subjective study. For each test sentence randomly selected from the test set, 20 converted utterances processed by each conversion method to each emotion type were randomly output to the human subjects. Twenty adult subjects, aged around 22-32, were asked to classify each utterance as one of the three emotion types. The subjects were familiar with our study. Figure 4 shows the identification results, and indicates that K-means-based GMBM method performs better than CART-based method. Although prosody controls most of the emotional cues in speech, spectral conversion is still helpful to emotion expression. The naturalness of the converted utterances was also evaluated, according to a 5-scale scoring method (5 = excellent, 1 = very poor). Figure 5 compares various conversion methods with mean opinion score (MOS) and its standard deviation. The analysis of variance (ANOVA) evaluations were conducted and the results yielded there was significant difference between methods with significance levels of $p < 0.05$.

5. CONCLUSION

A conversion function clustering and selection framework is presented in this work to incorporate linguistic information into the design of spectral conversion process. The K-means algorithm is adopted to cluster the conversion functions in each conversion model. Gaussian mixture bi-gram model is adopted as the conversion model. Results of objective experiments confirm the proposed method outperforms the CART-based method in the reduction of distortion between the converted and target expressive speech. The inclusion of linguistic information improves the modeling of conversion functions. Subjective tests reveal that more accurate spectral conversion would improve the expression of emotional speech.

6. REFERENCES

[1] T. Toda, A. W. Black and K. Tokuda, "Spectral Conversion based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter," in *Proc. of ICASSP'05*, vol. 1, pp. 9-12, Philadelphia, USA, Mar 2005.

[2] H. Duxans, A. Bonafonte, A. Kain and J. van Santen, "Including Dynamic and Phonetic Information in Voice Conversion Systems," in *Proc. of ICSLP 2004*, pp. 5-8, Jeju Island, South Korea, 2004.

[3] 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 Processing*, 14(4):1109-1116, 2006.

[4] W. H. Tsai and W. W. Chang, "Discriminative Training of Gaussian Mixture Bigram Models with Application to Chinese Dialect Identification," *Speech Communication*, 36(3-4): 317-326, 2002.

[5] H. Kawahara, "Speech Representation and Transformation using Adaptive Interpolation of Weighted Spectrum: Vocoder Revisited," in *Proc. of ICASSP 1997*, vol. 2, pp. 1303-1306, Munich, Germany, 1997.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. R. Statist. Soc. B*, vol. 39, pp. 1-38, 1977.

[7] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[8] C. H. Wu and J. H. Chen, "Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis," *Speech Communication*, 35(3-4): 219-237, 2001.

[9] A. Kain and M. W. Macon, "Spectral Voice Conversion for Text-to-Speech Synthesis," in *Proc. of the ICASSP'98*, Seattle, USA, 1998.

[10] S. H. Chen, S. H. Hwang and Y. R. Wang, "An RNN-based Prosodic Information Synthesis for Mandarin Text-to-Speech," *IEEE Trans. Speech and Audio Processing*, 6(3):226-239, 1998.

[11] J. Tao, Y. Kang and A. Li, "Prosody Conversion From Neutral Speech to Emotional Speech," *IEEE Trans. Audio, Speech and Language Processing*, 14(4):1145-1154, 2006.