Duration-Embedded Bi-HMM for Expressive Voice Conversion

Chi-Chun Hsia, Chung-Hsien Wu, Te-Hsien Liu

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

Abstract

This paper presents a duration-embedded Bi-HMM framework for expressive voice conversion. First, Ward's minimum variance clustering method is used to cluster all the conversion units (sub-syllables) in order to reduce the number of conversion models as well as the size of the required training database. The duration-embedded Bi-HMM trained with the EM algorithm is built for each sub-syllable class to convert the neutral speech into emotional speech considering the duration information. Finally, the prosodic cues are included in the modification of the spectrum-converted speech. The STRAIGHT algorithm is adopted for high-quality speech analysis and synthesis. Target emotions including happiness, sadness and anger are used. Objective and perceptual evaluations were conducted to compare the performance of the proposed approach with previous methods. The results show that the proposed method exhibits encouraging potential in expressive voice conversion.

1. Introduction

Speech is probably the only medium conveying linguistic, emotional, and other para-/extra-linguistic information simultaneously. Recently, many concatenative TTS (text-to-speech) systems have been developed for high quality speech synthesis. However, these systems suffer from synthesizing an utterance of different speaking styles, speakers, or emotions, without a large set of corresponding speech databases. The goal of this paper is to build a voice conversion system as a post-processing module for the TTS system, instead of storing several large speech databases with different expressions.

In the past years, many methods have been proposed to convert a source speech to the target one. VQ (vector quantization) is the earliest approach to spectral conversion [1], which uses a codebook mapping method between source and target feature vectors. In the past decade, stochastic approaches dominated the development of voice conversion systems. These approaches used conversion functions trained with MMSE (minimum mean square error) criterion under some specific probability distributions. In these approaches, GMM (Gaussian Mixture Model) is most widely used [2-3]. In recent years, a joint HMM-based method was proposed, which considered the dynamic characteristics in the acoustic model [4]. Besides, many model refinements including DFW (dynamic frequency warping) [5] and MAP (maximum *a posteriori*) adaptation [6] were also introduced to voice conversion.

Previous GMM-based framework is a frame-by-frame procedure. The time-independence assumption was adopted and the dynamic characteristic of speech sound was ignored. In the HMM-based methods, the state transition property gives a well approximation of the spectrum envelope evolution in time axis. A joint HMM was trained using source and target speech simultaneously. It can model the probability distribution of any feature vector, according to its actual state. Besides, it can model the dynamics of sequences of vectors with transition probabilities between states. However, modeling all the source and target speech in a joint HMM may introduce more confusion not only in the mixture density but also the transition probabilities. Moreover, it is difficult to model the state duration in a joint HMM for source and target speech separately.

In this paper, a duration-embedded Bi-HMM framework is proposed for expressive voice conversion. As shown in Fig. 1, the STRAIGHT algorithm (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum), proposed by Kawahara et al., [8][9] is used to estimate the spectrum of the speech signal. According to the linguistic information, the conversion models are selected according to the corresponding sub-syllable class of the source speech segment. The sub-syllable classes are trained by the Ward's minimum variance clustering method [7]. The target spectrum is predicted by the duration-embedded Bi-HMM using the Viterbi algorithm. The prosody of converted speech is modified according to the dynamic features modeled by template tree. After prosodic transformation, the target speech waveform is re-synthesized by the STRAIGHT algorithm.

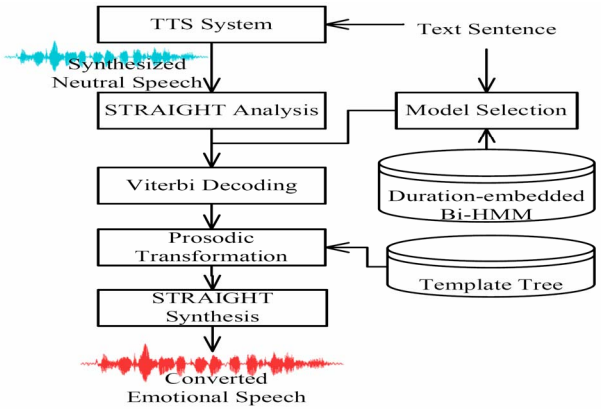

Figure 1: System architecture of proposed method.

Since voice conversion is performed with an analysis-synthesis method, the quality of the synthesized speech is important. For this purpose, the STRAIGHT algorithm is adopted in the voice conversion system. This method extracts F0 by using TEMPO (Time-domain Excitation extractor using Minimum Perturbation Operator), and designs excitation source based on phase manipulation. The STRAIGHT algorithm allows flexible manipulation of speech parameters such as vocal tract length, pitch and

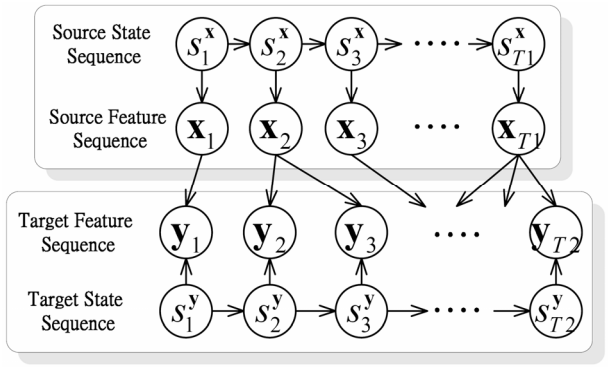

Figure 2: Bayesian network of the Bi-HMM

speaking rate while maintaining high quality of the synthesized speech.

In the following sections, the probabilistic framework of the proposed duration-embedded Bi-HMM model is detailed in Section 2. It also introduces the model clustering and the modification of prosodic features. Section 3 gives the experimental results, and finally, Section 4 gives some concluding remarks.

2. Proposed Voice Conversion Framework

In this study, a Bi-HMM framework is introduced to model the source and target speech signals in a separate but simultaneous way for voice conversion. Figure 2 gives the Bayesian network of the Bi-HMM for voice conversion. The round nodes are used to denote random variables, and the arrows connecting two nodes represent no conditional assumption between them.

The upper part of figure 2 shows the conditional dependence relationship of the source observation sequence $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_{T1}\}$ with length $T1$ and the state sequence $\mathrm{S}_\mathbf{X}=\{s_{\mathbf{x}_1},s_{\mathbf{x}_2},\cdots,s_{\mathbf{x}_{T1}}\}$ given source HMM. The current observation $\mathbf{x}_t$ is conditionally independent of the previous observations given the current state $s_t^{\mathbf{x}}$. Also the state $s_{t+1}^{\mathbf{x}}$ is conditionally independent of the past given the immediately preceding state $s_t^{\mathbf{x}}$. The lower part is the target observation sequence, $\mathbf{Y}=\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_{T2}\}$ and the state sequence $S_\mathbf{Y}=\{s_{\mathbf{y}_1},s_{\mathbf{y}_2},\cdots,s_{\mathbf{y}_{T2}}\}$. The dependence relationship between $\mathbf{x}$ and $\mathbf{y}$ is modeled as the conditional probability $P(\mathbf{y}\,|\,\mathbf{x})$.

2.1. Bi-HMM voice conversion model

As shown in Fig. 3, a Bi-HMM is used to model the dynamics of source and target speech, separately but simultaneously. In the training stage, the source and target feature vector sequences $\mathbf{X}$ and $\mathbf{Y}$ are aligned. The EM algorithm is adopted to estimate the HMM parameter set Λ. The continuous conversion functions resulting from the conditional probability $P(\mathbf{y}\,|\,\mathbf{x})$ and Normal assumption are estimated using the MMSE (minimum mean square error) criterion.

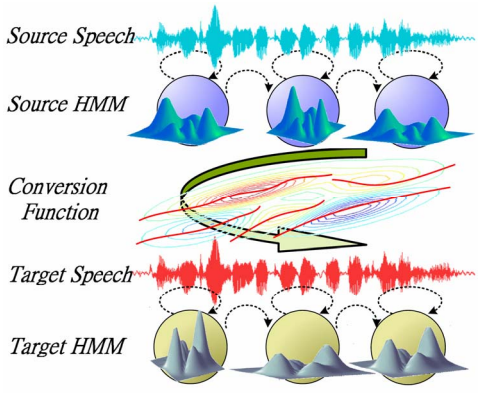

Figure 3: Diagram of the duration-embedded Bi-HMM

For a given state $s_t^{\mathbf{x}}=i$ and an input source feature vector $\mathbf{x}_t$, the converted target vector $\tilde{\mathbf{y}}_t$ is predicted using the following conversion function.

$$\begin{aligned}\tilde{\mathbf{y}}_t &= F\left(\mathbf{x}_t, s_{\mathbf{X}_t}=i\right)\\ &= \sum_{j=1}^{J} p_i\left(j\,|\,\mathbf{x}_t\right)\left[\boldsymbol{\mu}_{i,j}^{\mathbf{Y}}+\boldsymbol{\Sigma}_{i,j}^{\mathbf{YX}}\left(\boldsymbol{\Sigma}_{i,j}^{\mathbf{XX}}\right)^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}_{i,j}^{\mathbf{X}}\right)\right]\end{aligned} \tag{1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and the covariance matrix, respectively. $p_i\left(j\,|\,\mathbf{x}_t\right)$ is the conditional probability that $\mathbf{x}_t$ belongs to the mixture j in state i. Given the Bi-HMM Λ corresponding to the source feature vector sequence $\mathbf{X}$, the state sequences $\mathrm{S}_\mathbf{X}$ and $S_\mathbf{Y}$ with maximum joint probability $p\left(\mathbf{Y},\mathrm{S}_\mathbf{Y},\mathbf{X},\mathrm{S}_\mathbf{X}\,|\,\Lambda\right)$ is obtained as follows.

$$\begin{aligned}\left(\mathrm{S}_\mathbf{Y}^*,\mathrm{S}_\mathbf{X}^*\right) &= \arg\max_{\mathrm{S}_\mathbf{Y},\mathrm{S}_\mathbf{X}} p\left(\mathbf{Y},\mathrm{S}_\mathbf{Y},\mathbf{X},\mathrm{S}_\mathbf{X}\,|\,\Lambda\right)\\ &= \arg\max_{\mathrm{S}_\mathbf{Y},\mathrm{S}_\mathbf{X}} p\left(\mathbf{Y},\mathrm{S}_\mathbf{Y}\,|\,\mathbf{X},\mathrm{S}_\mathbf{X},\Lambda\right)p\left(\mathbf{X},\mathrm{S}_\mathbf{X}\,|\,\Lambda\right)\\ &= \arg\max_{\mathrm{S}_\mathbf{Y},\mathrm{S}_\mathbf{X}} p\left(\mathbf{Y}\,|\,\mathbf{X},\mathrm{S}_\mathbf{X},\mathrm{S}_\mathbf{Y},\Lambda\right)p\left(\mathrm{S}_\mathbf{Y}\,|\,\Lambda\right)p\left(\mathbf{X}\,|\,\mathrm{S}_\mathbf{X},\Lambda\right)p\left(\mathrm{S}_\mathbf{X}\,|\,\Lambda\right)\\ &\approx \arg\max_{\mathrm{S}_{\tilde{\mathbf{Y}}},\mathrm{S}_\mathbf{X}} p\left(\tilde{\mathbf{Y}}\,|\,\mathrm{S}_{\tilde{\mathbf{Y}}},\Lambda\right)p\left(\mathrm{S}_{\tilde{\mathbf{Y}}}\,|\,\Lambda\right)p\left(\mathbf{X}\,|\,\mathrm{S}_\mathbf{X},\Lambda\right)p\left(\mathrm{S}_\mathbf{X}\,|\,\Lambda\right)\end{aligned} \tag{2}$$

where $p\left(\mathrm{S}_\mathbf{X}\,|\,\Lambda\right)$ and $p\left(\mathrm{S}_\mathbf{Y}\,|\,\Lambda\right)$ represent the state transition probabilities for source and target feature vector sequences $\mathbf{X}$ and $\mathbf{Y}$. $\left(\mathrm{S}_\mathbf{X}^*,\mathrm{S}_\mathbf{Y}^*\right)$ is the best state sequence pair, which with maximum joint probability, obtained from the Bi-HMM Λ using the Viterbi algorithm. For the obtained best state sequence pair $\left(\mathrm{S}_\mathbf{X}^*,\mathrm{S}_\mathbf{Y}^*\right)$, the target vector sequence with maximum probability in Bi-HMM is obtained from the following conversion function:

$$\hat{\mathbf{Y}}=F\left(\mathbf{X},S_\mathbf{X}^*\right)=\left\{F\left(\mathbf{x}_1,s_{\mathbf{x}_1}^*\right),F\left(\mathbf{x}_2,s_{\mathbf{x}_2}^*\right),\cdots,F\left(\mathbf{x}_{T1},s_{\mathbf{x}_{T1}}^*\right)\right\} \tag{3}$$

2.2. Embedded duration model

In standard HMM, the state duration probability decrease exponentially with time. When the speech signal stays in state i, $i=1,\cdots,I$ with self-transition probability a_{ii}, for τ

frames, the implicit duration probability density is a geometric distribution $d_i(\tau) = a_{ii}^{\tau-1}(1-a_{ii})$. However, this exponential state duration density is inappropriate for most speech signals. In this study, a Gamma duration model is adopted as follows [10]:

$$d_i(\tau \mid \eta_i, \nu_i) = \frac{\eta_i^{\nu_i}}{\Gamma(\nu_i)} \tau^{\nu_i - 1} \exp(-\eta_i \tau) \tag{4}$$

where $\Gamma(\cdot)$ is gamma function; $\eta_i > 0$ and $\nu_i > 0$ are parameters of the gamma distribution. The Q-function corresponding to gamma duration parameters for state m becomes

$$\begin{aligned} &Q_\gamma(\hat{\eta}_i, \hat{\nu}_i \mid \eta_i, \nu_i) \\ &= \sum_{t=1}^{T} \xi_{t \in t_s}(i) \cdot \left[-\log\Gamma(\hat{\nu}_i) + \hat{\nu}_i \log\hat{\eta}_i + (\hat{\nu}_i - 1)\log\tau_t - \hat{\eta}_i \tau_t \right] \end{aligned} \tag{5}$$

where t_s represents the starting frame of a state, and $\xi_{t \in t_s}(i)$ is defined as:

$$\xi_{t \in t_s}(i) = \delta(s_t - i)\delta(t - t_s) = P(s_{t \in t_s} = i \mid \mathbf{X}, \Lambda) \tag{6}$$

where $\delta(\cdot)$ is the Kronecker delta function. Unfortunately, no closed-form solution to new estimate was derived. Because gamma distribution $d_i(\tau \mid \eta_i, \nu_i)$ has mean ν_i/η_i and variance ν_i/η_i^2, the parameters are computed empirically from the sample mean and variance.

2.3. Model Clustering

In Mandarin speech, there are 150 sub-syllables (Initial or Final of a syllable) which are basic units for speech recognition. The proposed conversion model is therefore constructed on the sub-syllable level. However, it is not necessary for each emotional speech database to contain all the 150 sub-syllables. In order to reduce the size of the training database, all the sub-syllables are clustered to reduce the model number using a large neutral speech database.

For sub-syllable clustering, the Ward's minimum variance method [7] is adopted. This algorithm is a bottom-up approach and two classes with minimum distance will be merged in each iteration. In Ward's method, the distance between two classes is estimated as:

$$D(A,B) = n_A \cdot \left\| \bar{\mathbf{x}}_A - \bar{\bar{\mathbf{x}}} \right\|^2 + n_B \cdot \left\| \bar{\mathbf{x}}_B - \bar{\bar{\mathbf{x}}} \right\|^2 \tag{7}$$

where $\bar{\mathbf{x}}_A$ and $\bar{\mathbf{x}}_B$ are the mean vectors of classes A and B, respectively. $\bar{\bar{\mathbf{x}}}$ is the mean vector of the merged class of A and B, and is calculated as:

$$\bar{\bar{\mathbf{x}}} = \frac{n_A}{n_A + n_B} \bar{\mathbf{x}}_A + \frac{n_B}{n_A + n_B} \bar{\mathbf{x}}_B \tag{8}$$

2.4. Prosody Selection from the Template Tree

In [3], the authors show that prosodic features provide important cues to emotional speech expression. Although it is difficult to model the prosody without a very large speech database, a syllable-balanced database meets the minimum requirement for performance evaluation. In this study, the word prosody template tree proposed in [11] is adopted to model the prosodic features of different emotion expression. Only monosyllable, 2-syllable and 3-syllable words are considered. Each word prosody template contains the syllable duration, average energy and pitch contour of the word. The pitch contour in the word prosody template records the prosodic features for the syllables in the word. For each word in a sentence/phrase, word length is first determined and used to traverse the template tree. A sentence intonation module and a template selection module are adopted to select the target prosody templates [11]. Finally, The STRAIGHT algorithm is adopted for prosodic modification.

3. Experimental Results

This study adopts happiness, sadness and anger as the target emotions. These emotions are primary emotions and have been used in many emotional speech synthesis/recognition studies. To evaluate the performance of the proposed method, a small balanced set consisting of 300 text sentences is selected as the script of the speech database. Sentence selection is designed to include all the 150 sub-syllables in Mandarin. The average appearance time of each sub-syllable is about 31. The number of sub-syllable classes trained with the Ward's algorithm is 61. In order to remove the emotional cues from text in the following subjective test, the neutral textual sentences are used.

A female speaker was asked to read the text with four kinds of emotions (neutral, happy, sad and angry) separately. She was also asked to use consistent, not exaggerative emotional presentation for the same emotion. The speech waves were sampled at 22.05 Hz and quantized in 16 bits per sample. The recording environment is a silent room.

In the following experiments, 15 randomly selected speech utterances for each emotion were used as the testing data, and the others were used to train the conversion functions and the prosody model.

4.1. Objective test

Objective evaluation was conducted to confirm the distortion between different spectral conversion methods. We employed mel-scale cepstrum coefficients (MFCCs) to calculate the spectral distortion. Distortion between the converted (from neutral) and the target (happy, sad and angry) speech is calculated using the following equation:

$$Avg\_Dis = \frac{20}{\ln 10} \left(\frac{1}{T} \sum_{t=1}^{T} \left\| 2(\mathbf{y}_t - \hat{\mathbf{y}}_t) \right\| \right) \tag{7}$$

where $\mathbf{y}_t$ and $\hat{\mathbf{y}}_t$ are the target vector and the converted feature vectors, represented in MFCCs, respectively. Table 1 gives the distortions of different methods. The number of GMM mixture is 32.

Table 1: Results of objective test

Methods	Distortion (dB)		
	Happy	Sad	Angry
VQ	7.65	7.31	7.76
GMM	6.34	6.13	6.57
HMM	5.42	5.24	5.37
Bi-HMM	***5.06***	***4.89***	***5.11***

Table 2: MOS of different conversion methods

Methods	Mean Opinion Score		
	Happy	Sad	Angry
VQ	2.7	2.1	2.4
GMM	3.5	3.8	3.4
HMM	4.0	4.1	3.7
Bi-HMM	***4.1***	***4.3***	***4.0***

4.2. Subjective test

Two evaluations were conducted in the subjective test. One is for naturalness, and the other one is for emotion expression. Ten adult listeners participated in this experiment. They were asked to identify the emotion of each converted speech and gave a mean opinion score (ranges from 1 to 5) to each converted speech. It was conducted as a forced-choice (neutral, happy, sad or angry) test. 15 speech sentences converted from the source neutral speech to each target emotional speech (happy, sad or angry) were used in this experiment. These 45 converted speech sentences were presented to the listeners randomly.

Table 2 shows the MOS of each method. As a result, the proposed method outperforms the conventional approaches. Figure 4 give the result of emotion identification from listening test. Although, prosody controls mainly the emotion expression, spectrum conversion still plays an important role for emotion perception.

4. Conclusion

This work has presented a Bi-HMM framework to expressive voice conversion. In this framework, two HMMs are used separately but simultaneously to model the source and target speech. State duration model is embedded to the stochastic voice conversion model, Bi-HMM. Prosodic features modeled by template tree are included in the post processing. The result of objective experiments confirms the reduction of distortion between source and target expressive speech. Subjective tests were conducted and the result shows that although prosody represents most of the expression cues, spectrum conversion is also important for emotion expression.

5. Acknowledgements

The authors would like to thank Dr. Kawahara for the support of STRAIGHT analysis/synthesis system.

6. Reference

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, ``Voice conversion through vector quantization", *in Proc. of ICASSP*, pp. 655-658, 1988

Figure 4: Listening test results of emotion identification for different approaches. a) VQ, b) GMM, c) HMM, d) Bi-HMM, e) Duration-embedded Bi-HMM, f) Duration-embedded Bi-HMM with prosody modification

[2] Y. Stylianou, O. Cappé and E. Moulines, ``Continuous Probabilistic Transform for Voice Conversion", *IEEE Trans. Speech and Audio Processing*, 6(2):131-142, 1998

[3] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari and K. Shikano, ``GMM-based Voice Conversion Applied to Emotional Speech Synthesis", *in Proc. of EuroSpeech*, pp. 2401-2404, 2003

[4] H. Duxans, A. Bonafonte, A. Kain and J. van Santen, ``Including Dynamic and Phonetic Information in Voice Conversion Systems", *in Proc. of ICSLP*, pp. vol. 1, pp. I-5-8, 2004

[5] T. Toda, H. Saruwatari and K. Shikano, ``High Quality Voice Conversion Based on Gaussian Mixture Model with Dynamic Frequency Warping", *in Proc. of EuroSpeech*, pp. 349-352, 2001

[6] Y. Chen, M. Chu, E. Chang, J. Liu and R. Liu, ``Voice Conversion with Smoothed GMM and MAP Adaptation", *in Proc. of EuroSpeech*, pp. 2413-2416, 2003

[7] Jr. J. H. Ward, ``Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association*, 58:236-244, 1963

[8] H. Kawahara, ``Speech Representation and Transformation using Adaptive Interpolation of Weighted Spectrum: Vocoder Revisited", in Proc. of ICASSP, pp. 1303-1306, 1997

[9] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigné, ``Restructuring Speech Representations using a Pitch Adaptive Time-Frequency-based F0 Extraction: Possible Role of a Repetitive Structure in Sounds", *Speech Communication*, 27(3-4):187-207, 1999

[10] S. E. Levinson, ``Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition", Comput. Speech Lang., 1:29-45, 1986

[11] 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