MODEL-MAPPING BASED VOICE CONVERSION SYSTEM

A Novel Approach to Improve Voice Similarity and Naturalness using Model-based Speech Synthesis Techniques

Baojie Li, Dalei Wu and Hui Jiang
Department of Computer Science and Engineering, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada

Keywords: Voice conversion, HMM-based speech synthesis, GMM, Model mapping.

Abstract: In this paper we present a novel voice conversion application in which no any knowledge of source speakers is available, but only sufficient utterances from a target speaker and a number of other speakers are in hand. Our approach consists in two separate stages. At the training stage, we estimate a speaker dependent (SD) Gaussian mixture model (GMM) for the target speaker and additionally, we also estimate a speaker independent (SI) GMM by using the data from a number of speakers other than the source speaker. A mapping correlation between the SD and the SI model is maintained during the training process in terms of each phone label. At the conversion stage, we use the SI GMM to recognize each input frame and find the closest Gaussian mixture for it. Next, according to a mapping list, the counterpart Gaussian of the SD GMM is obtained and then used to generate a parameter vector for each frame vector. Finally all the generated vectors are concatenated to synthesize speech of the target speaker. By using the proposed model-mapping approach, we can not only avoid the over-fitting problem by keeping the number of mixtures of the SI GMM to a fixed value, but also simultaneously improve voice quality in terms of similarity and naturalness by increasing the number of mixtures of the SD GMM. Experiments showed the effectiveness of this method.

1 INTRODUCTION

Voice conversion (VC) is a technique that converts voice of a source speaker to that of a target speaker. Generally speaking, text-dependent and text-independent voice conversion represent two main streams of research directions. In text-dependent voice conversion, target voice can be produced with high-quality of correctness and acceptable smoothness based on the provided transcription for input speech waveform, e.g. (Yoshimura, 2002). By contrast, text-independent systems have no knowledge about the transcription of input waveform, therefore more mismatches between source and target speakers are present and the quality of the generated speech then degrades. For this reason, text-independent voice conversion attracts a wider range of studies. The techniques presented in this paper are also focused on text-independent voice conversion.

In the field of text-independent voice conversion, usually some forms of transforms are estimated from training data of both source and target speakers, such as K-means clustering in VTLN-based voice conversion (Suedermann et al., 2003), codebook based mapping (Arslan et al., 1999) and GMM based clustering (Ye et al., 2006). In some applications, however, no knowledge about source speakers is applicable beforehand. Therefore it is impossible to estimate the transforms between source and target speakers using the conventional techniques. In our previous work, we built a GMM-based VC system using hidden Markov model (HMM) based speech synthesis to address such particular requirements. At the training stage, a SD GMM is trained for the target speaker using his/her pre-recorded training data. In the conversion stage, for each utterance from a source speaker, the best matched Gaussian mixture is chosen from the GMM. Next, the mean vectors of the selected mixtures are concatenated, smoothed and then sent as inputs to the sound synthesizer, which is provided by HTS engine (Tokuda et al, 2000; Yoshimura et al., 2002). By experiments, we found that this approach was quite capable of conducting voice conversion with acceptable quality. However, we also found recognizable discontinuity and flatness in the synthesized voices. Through investigation, we found that the

Li B., Wu D. and Jiang H. (2010).
MODEL-MAPPING BASED VOICE CONVERSION SYSTEM - A Novel Approach to Improve Voice Similarity and Naturalness using Model-based Speech Synthesis Techniques.
In *Proceedings of the Third International Conference on Bio-inspired Systems and Signal Processing*, pages 442-446
DOI: 10.5220/0002747104420446

discontinuity is attributed to frame mismatches. And the flatness was caused by the use of average of training samples. To cope with these two problems, we proposed in this article to renovate our VC system by introducing additional SI models and model-mapping technique. We confirmed by experiments that the new method was quite effective in increasing the quality of the generated speech.

The rest of this paper is organized as follows: we begin with an introduction to our previous GMM-based VC system in Section 2. To deal with the discontinuity and flatness problems, a model-mapping strategy is introduced in Section 3, where its effectiveness is evaluated by experiments. We draw the conclusions in Section 4.

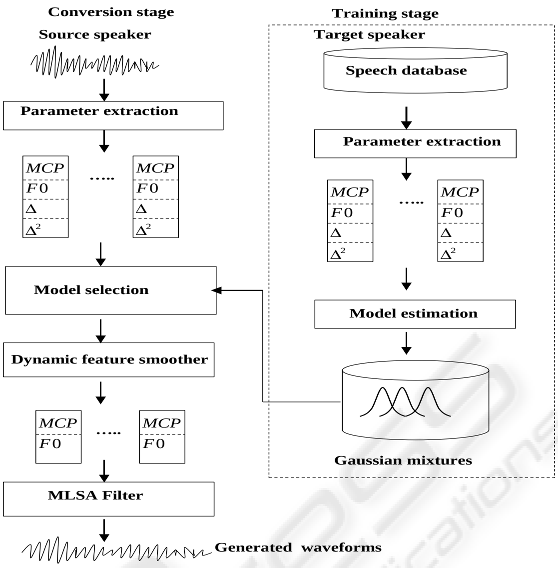

Figure 1: GMM-based voice conversion system.

2 GMM-BASED VOICE CONVERSION SYSTEM

2.1 HMM-based Speech Synthesis Technique

Our voice conversion system is based on HMM-based speech synthesis (HTS) techniques (Tokuda et al., 2000; Yoshimura et al., 2002). The HMM-based speech synthesis system assumes that sufficient training data and their corresponding transcriptions are available. The system models phonetic and prosodic parameters simultaneously. At the training stage, Mel-cepstrum coefficients (MFCC), fundamental frequency (F0) and duration are modeled by multiple-stream HMMs. At the synthesis stage, a given string of words is firstly decomposed into a string of phonemes. The system then searches in the HMM pools the corresponding model for each phoneme. The mean vector of each model is taken as a frame to represent that phoneme. Later, all these frames are then concatenated according to their time order and passed to a smoother, which improves the quality of the synthesized speech by smoothing over the whole sequence of frames using dynamic parameters (e.g., delta and delta delta) of MFCCs and F0. The smoothed frame sequence is then input to the synthesizer, known as MLSA filter(Tokuda, 2000), to produce speech waves.

For a text-to-speech task, HTS has been demonstrated in generating high quality voices especially in terms of continuity and naturalness. However in our current task, neither the transcriptions for the input waveforms nor the knowledge about the source speaker is available. Many conventional transform-based VC approaches are inapplicable in such a situation. Hence we were motivated to combine speech recognition and synthesis techniques to conduct voice conversion.

2.2 Overview of Our VC System

The diagram of our previously proposed VC system is shown in Figure 1. Our VC system consists of two stages. At the first stage, parameter vectors are extracted from training data of the target speaker. Then Gaussian mixture models are estimated. All mixtures have no phoneme labels but sequential numbers. At the conversion stage, a mixture is chosen for each input frame from the GMM set according to distance or probability metrics. Then the mean vectors of the selected mixtures are concatenated to form an output sequence of frames. After smoothed by a smoother using the dynamic features, the sequence of frames is input to the MLSA filter to generate speech wave.

2.3 Feature Extraction

The feature vectors are extracted from speech waveforms using HTK 3.4 (Young et al., 2003) and SPTK (Imar et al., 2008). Each vector includes 25-dimentional MFCCs, 1-dimentional pitch information (F0), along with their dynamic features. The vector is divided into four streams: MFCCs and their dynamics, F0, delta F0 and delta delta F0. All streams take uniform weights.

2.4 Model Estimation

GMM for the target speaker is estimated using the standard EM algorithm. Considering that correlations between F0s and MFCCs introduce inaccuracy into probability calculations when a diagonal covariance matrix is used in recognition process, only the MFCCs are used in mixture clustering process. After the mixtures are obtained, all the training data are aligned into the mixtures. The average of F0 over all frames in a mixture is set as the mean of this mixture. More specifically, assume we have N frames $\{\mathbf{x}_n\}$ and K mixtures. Each mixture has a probability distribution as $\mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)$ with a weight of w_k. Then each frame contributes to this mixture as γ

$$\gamma_{nk} = \frac{w_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(\mathbf{x}_n \mid \mu_j, \Sigma_j)} \tag{1}$$

And the total number of samples accumulated in this mixture is $N_k = \sum_{n=1}^{N} \gamma_{nk}$. Note that only the MFCC part of vector $\mathbf{x}_n$ are used in the above calculation. Then we get the F0 parameters for this mixture as

$$\mu_k^{F0} = \frac{1}{N_k} \sum \gamma_{nk} \mathbf{x}_n^{F0} \tag{2}$$

where μ_k^{F0} is the F0 part of μ_k, and $\mathbf{x}_n^{F0}$ the F0 part of $\mathbf{x}_n$ respectively.

2.5 Model Selection

In conversion stage, each input frame from a source speaker is calculated on the GMMs of the target speaker (excluding the F0 parameters). The mixture with the highest probability is selected as the output.

$$l_t^* = \arg\max_{\mathbf{k}} w_k \mathcal{N}(\mathbf{x} \mid \Sigma_k, \mu_k) \tag{3}$$

where μ_k and Σ_k is the mean and covariance for the k-th Gaussian component, respectively.

2.6 Dynamic Features Smoothing

All the mean vectors of the selected Gaussian components are concatenated to form a sequence. The sequence is smoothed according to its delta and delta delta information in the same way as in HMM-based speech synthesis (Tokuda,2000; Yoshimura,2002). Then a new sequence consisting of only static MFCCs and static F0 features is obtained. We denote the static, delta, delta delta features for time t as $(\mathbf{c}_t, \Delta\mathbf{c}_t, \Delta^2\mathbf{c}_t)$. Then, they are defined as

$$\triangle \mathbf{c}_t = \sum_{-L_-^{(1)}}^{+L_+^{(1)}} w^{(1)(\tau)} \mathbf{c}_{t+\tau} \tag{4}$$

$$\triangle^2 \mathbf{c}_t = \sum_{-L_-^{(2)}}^{+L_+^{(2)}} w^{(2)(\tau)} \mathbf{c}_{t+\tau}. \tag{5}$$

where $w^{(j)}(\tau), j \in \{1,2\}$ are weight coefficients and τ is a length of time. The new static features sequence $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_T]^\top$ can be estimated by solving a system of linear equations as follows:

$$\mathbf{C} = (\mathbf{W}^{\mathbf{T}}\mathbf{U}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{U}^{-1}\mathbf{S}^\top, \tag{6}$$

where

$$\mathbf{S} = [\mu_{l_1^*}^\top, \mu_{l_2^*}^\top, \cdots, \mu_{l_T^*}^\top]^\top \tag{7}$$

$$\mathbf{U}^{-1} = \mathrm{diag}[\Sigma_{l_1^*}^{-1}, \Sigma_{l_2^*}^{-1}, \cdots, \Sigma_{l_T^*}^{-1}] \tag{8}$$

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_T]^\top \tag{9}$$

$$\mathbf{w}_t = [\mathbf{w}_t^{(0)}, \mathbf{w}_t^{(1)}, \mathbf{w}_t^{(2)}] \tag{10}$$

$$\begin{aligned}\mathbf{w}_t^{(n)} = [&\mathbf{0}_{M\times M}, \cdots, \mathbf{0}_{M\times M}, w^{(n)}(-L_-^{(n)})\mathbf{I}_{M\times M}, \\ &\cdots, w^{(n)}\mathbf{I}_{M\times M}, \cdots, w^{(n)}(L_+^{(n)})\mathbf{I}_{M\times M}, \\ &\mathbf{0}_{M\times M}, \cdots, \mathbf{0}_{M\times M}]^\top, n = 0,1,2.\end{aligned} \tag{11}$$

where M is the number of the dimensions of static MFCC features without dynamic features and l_t^* denotes the index of the best-matched Gaussian component, selected as in Eq. (3).

2.7 Experiments

Experiments were conducted on The Continuous Speech Recognition Wall Street Journal Phase I (CSR-WSJ0) Corpus (Linguistic Data Consortium). Speaker 001 (male, 598 utterances) and 002 (female, 600 utterances) were alternatively used as source and target speaker. 1720 Gaussian mixtures were trained for each target speaker. subjectively listening to the converted utterances were conducted. Five listeners were given ten generated utterances for each target speaker. They were asked to give to each utterance a score which is ranged from 1 to 5. The score average of each listener across the ten sentences is listed in Table 1.

Table 1: Subjective evaluation scores for GMM VC.

Listener id	1	2	3	4	5	Ave
naturalness	2.9	3.4	2.6	3.8	3.2	**3.2**
similarity	2.9	3.2	2.3	3.2	2.0	**2.7**

We found that the quality of generated voices was acceptable. Most of the literal contents were identifiable. However there were also noticeable discontinuity and flatness in the voices.

3 IMPROVE VOICE SIMILARITY AND NATURALNESS BY MODEL MAPPING

By investigating the output frames in the previous experiments, We noticed that the mis-selection of models resulted in remarkable discontinuity and the use of mean vectors of Gaussian mixtures made the generated speech sounds flat. More specifically speaking, each mean vector of a Gaussian mixture used for speech synthesis was an average over some training frames. This average effaced the details of the characteristics of the target speaker. we tried to increase the number of Gaussian mixtures to capture more details of the target speaker. However, when a certain number was exceeded, the recognition performance started to decrease thus degrading the quality of the synthesized voice. This was due to that when number of mixtures went up, the trained models perfectly fitted the training data but gave a very poor representation of the test data thus resulting in a bad recognition performance. This behaviour is known as over-fitting. To solve this problem, we introduced an additional model set for recognition, which was a SI model set trained from a number of speakers.

In this method, both the SI models and the target SD models were trained for all phonemes. Each phoneme model consists of a number of Gaussian mixtures. Therefore, between the SI and SD model sets, we could construct a mapping list by the model labels. The modified VC system is depicted as in Figure 2. In conversion stage, we firstly searched over SI GMMs to find out the closest model to the input frame. Then according to the mapping list, its counterpart model in the target speaker model set was selected as the output model. This scheme enabled us to capture more detailed characteristics of the target speaker by increasing the number of mixtures of target speaker models, without losing high recognition performance for the input frames by keeping the number of mixtures of SI models fixed.

3.1 Reduce Mismatch in Recognition

To recognize frames from an unknown speaker, the SI model set performs much better than the target speaker model set. When given an input frame $\mathbf{x}$ from source speaker, it is firstly calculated on the SI model set

$$p(\mathbf{x}) = \sum_{j=1}^{K} w_j \mathcal{N}(\mathbf{x} \mid \mu_j, \Sigma_j). \tag{12}$$

and assigned to a phoneme model which has the highest probability, e.g., model *ah*. Then its counterpart in the target model set can be found according to the mapping list. Next we will try to find a best matched mixture among all the mixtures belonging to the target SD model *ah*. To alleviate model mismatch, we shift the input frame according to

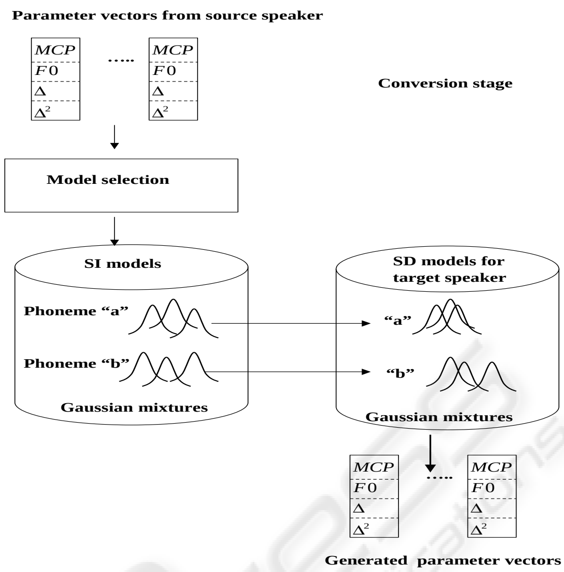

Figure 2: *Modified our voice conversion system.*

$$\mathbf{y} = \mu^{SD} + \sqrt{\frac{\Sigma^{SD}}{\Sigma^{SI}}}(\mathbf{x} - \mu^{SI}) \tag{13}$$

Assume each model in the target model set consists of K mixtures, the nearest mixture to vector $\mathbf{y}$ is selected by calculating the normalized Euclidean distances

$$l^* = \arg\max_{\mathbf{k}} \sum_{m=1}^{M} \frac{(\mathbf{y}_m - \mu_m^k)^2}{\Sigma_m^k} \tag{14}$$

where M is the dimensionality of MFCCs, μ_m^k the m-th dimension of the mean vector of k-th mixture.

Experiments were conducted using the same corpus as in Section 2.7. 7138 utterances from SI-84 speakers were used to train the SI GMMs. Each triphone model consists of 10 mixtures (after mixture sharing, the total number of mixtures was 1720, the same as that in Section 2.7). The same scoring criterion was applied. The results are listed in Table 2.

Table 2: Subjective evaluation scores for modified VC.

Listener id	1	2	3	4	5	Ave
naturalness	3.2	4.1	3.4	4.3	3.7	**3.7**
similarity	3.0	3.1	2.8	3.3	2.5	**2.9**

The SI model did improve the naturalness of the generated speech. The speech sounded more smoothly.

3.2 Improve Similarity by Increasing Number of Mixtures of SD Models

In the following experiments, we keep the number of mixtures of the SI models fixed to maintain a good recognition performance, while increasing the number of mixtures of the target models to capture the detailed characteristics of the target speaker. This is based on the assumption that even a target SD mixture lacks of training data, it is still capable of representing the voice of the target speaker since it came from the real samples of that speaker. Experiments were conducted by fixing the SI models to 1720 mixtures while changing the target models with different number of mixtures. Here we listed the evaluation scores in Table 3 and Table 4 .

Table 3: Naturalness for different number of mixtures.

Listener Id	1	2	3	4	5	Ave
1720 mix	3.2	4.1	3.4	4.3	3.7	**3.7**
3440 mix	4.1	4.3	3.6	3.9	4.1	**4.0**
6880 mix	3.5	4.3	4.0	4.2	3.7	**3.9**
13760 mix	2.8	3.8	3.4	3.7	3.6	**3.5**

Table 4: Similarity for different number of mixtures.

Listener Id	1	2	3	4	5	Ave
1720 mix	3.0	3.1	2.8	3.3	2.5	**2.9**
3440 mix	3.2	3.8	3.2	4.4	3.7	**3.6**
6880 mix	3.8	3.7	4.2	4.0	3.9	**3.9**
13760 mix	3.3	4.1	3.1	4.3	3.3	**3.6**

As shown in the tables, increasing the number of mixtures not only brought improvement in naturalness, but also improved similarity remarkably. As long as the number of mixtures did not go to extreme, the quality of generated voice was improved greatly.

4 CONCLUSIONS

In this paper, we have presented a new approach for voice conversion. By introducing a SI model set into the GMM VC system, naturalness of the converted voice was improved due to the improved recognition performance. Moreover the new system could use more mixtures in target models for parameter generation, therefore more detailed features of input speeches could be captured to improve similarity. In comparison with the original GMM VC system, both naturalness and similarity were improved.

REFERENCES

T. Yoshimura, ”Simultaneous Modelling of Phonetic and Prosodic Parameters and Characteristic conversion for GMM-based Tex-to-speech Systems” ,Ph.D. dissertation, Nagoya Institute of Technology, 2002.

D. Sundermann and H. Ney, ”VTLN-based Voice Conversion”, Proc. of Signal Processing and Information Technology (ISSPIT)”, Dec. 2003, pp. 14-17.

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi and T. Kitamura, ”Speech parameter generation algorithms for GMM-based speech synthesis”, Proc. of ICASSP, pp.1315-1318, June 2000.

L. Arslan and D. Talkin, ”Speaker Transformation Algorithm using Segmental Codebooks (STASC)”, Speech Commun., pp. 211-226, 1999.

H. Ye and S. Young, ”Quality-enhanced Voice Morphing Using Maximum Likelihood Transformation”, IEEE Trans. on Audio, Speech and Language Processing, Vol.14, No.4, pp.1301-1312, 2006.

S. Young et al., ”HTKBook (V3.4)”, Cambridge University Engineering Department, 2006.

S. Imai et al., ”Speech Signal Processing Toolkit Ver.3.2”, http://sp-tk.sourceforge.net, 2008 .

Linguistic Data Consortium, ”The LDC Corpus Catalog”, http://www.ldc.upenn.edu/Catalog/