Online Model Adaptation for Voice Conversion using Model-based Speech Synthesis Techniques

Dalei Wu¹, Baojie Li³, Hui Jiang¹, Qian-Jie Fu²

¹Department of Computer Science and Engineering, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, CANADA
²House Ear Institute, 2100 West Third Street, Los Angeles, CA 90057, USA
{daleiwu, lbj, hj}@cse.yorku.ca, qfu@hei.org

Abstract

In this paper, we present a novel voice conversion method using model-based speech synthesis that can be used for some applications where prior knowledge or training data is not available from the source speaker. In the proposed method, training data from a target speaker is used to build a GMM-based speech model and voice conversion is then performed for each utterance from the source speaker according to the pre-trained target speaker model. To reduce the mismatch between source and target speakers, online model adaptation is proposed to improve model selection accuracy, based on maximum likelihood linear regression (MLLR). Objective and subjective evaluations suggest that the proposed methods are quite effective in generating acceptable voice quality for voice conversion even without training data from source speakers.

Index Terms: voice conversion, HMM-based speech synthesis, GMM, model adaptation

1. Introduction

Voice conversion (VC) is a technique in which a source speaker’s voice characteristics are modified to those of a target speaker. Conventional VC methods generally rely on several types of talker transformation, e.g., vocal tract length normalization (VTLN) [6], continuous probabilistic transform [7], and multiple linear transformations [10]. Normally, training data is collected from both target and source speakers to jointly train VC transformations and models, e.g., k-means clustering [6], codebook based mapping [1] and Gaussian mixture model (GMM) modeling [10, 4]. Some training methods even require that the target and source speakers produce the same sentences (i.e., parallel data), allowing the transformation or mapping functions to be estimated by aligning the parallel data.

However, in many VC applications, it may be impractical or difficult to collect training data from both target and source speakers. In the present work, we study voice conversion for some application scenarios where we can only collect training data from target speakers but we cannot access any training data or any prior knowledge for source speakers. There are many interesting applications under this new VC scenario, especially for the internet gaming and toy industries. A typical VC application in this case is that we need to build a VC system which can convert voices from any user (without any training or enrollment) to make them sound like one particular target speaker saying the same content. In such an application, extensive training data can be collected for the target speaker but there is no prior information regarding source speakers and therefore no way to train the VC transformation using source speaker data. Each utterance from the source speaker must be converted independently towards the fixed target speaker. This type of voice conversion is quite challenging and many popular VC techniques may not be applicable as they rely on joint training of probability density using training data from both source and target speakers.

To address the particular requirements for this typical VC scenario, we may argue that the techniques of HMM-based speech synthesis (HMM-SS) may be used for this purpose, due to its general and strong modelling capacity, independent of speech contents [8, 12]. By experiments, we firstly found that this approach is quite effective to address the VC application without prior knowledge. Secondly we also found that frame mismatch between source and target speakers is very critical to affect the VC performance for the proposed approach. For this reason, a technique to reduce mismatch between source and target speakers and therefore improve Gaussian selection accuracy may be helpful to further enhance system performance. To this end, we propose to use the technique of maximum likelihood linear regression (MLLR) for online model adaptation to adapt the pre-trained target GMM model based on each utterance from source speaker. Although MLLR adaptation is well-known to be effective to address the task of speaker independent speech recognition, it is still unclear if this technique is also effective to the HMM-based VC algorithm. Based on our both objective and subjective evaluations, we confirmed that the MLLR adaptation technique is quite effective to enhance the voice quality for the proposed HMM-based approach.

The rest of this paper is organized as follows: In section 2, we describe the overall voice conversion system using the HMM-based speech synthesis technique. In section 3, we introduce the techniques for online model adaptation based on MLLR. In section 4, we present objective and subjective evaluations of the proposed VC system. Finally, we present our conclusions in section 5.
2. Voice conversion system using HMM-based speech synthesis

2.1. System overview

The proposed voice conversion system is composed of six major modules, shown in Fig. 1: feature extractor, model estimator, GMM selector, concatenator, smoother and MLSA filter. As described earlier, training data from a target speaker is used to build a GMM model. All source speaker utterances are then processed frame by frame. For each frame, we select the best matched Gaussian component from the GMM model. Then, we concatenate the selected Gaussian parameters and generate the speech waveform using standard model-based speech synthesis techniques [8, 12].

2.2. Feature Extractor

In the first module, feature vectors are extracted to capture the characteristics of speech signals. In our system, mel-scaled cepstral coefficients (mel-ceps) along with dynamic delta and acceleration features, are extracted for every speech frame. Pitch information (F0) as well as its delta and delta-delta are also computed for voiced speech segments and included as part of the feature vectors.

2.3. GMM Estimator

Speech samples are collected for each target speaker and features are extracted as described in section 2.2. The extracted feature vectors are used to build a GMM model to represent the target speaker. The GMM model is trained based on maximum likelihood estimation (MLE) criterion with a special two-stream topology: 1) extracted mel-ccep feature vectors and 2) F0 pitch information. During the GMM estimation process, the first stream is built using the standard EM algorithm, where only the mel-ccep part of the feature vector is used for clustering speech frames and estimating Gaussian components and their weights. After the first stream is built, the second stream is calculated as the average F0 pitch vector for each Gaussian component in GMM. Because only voiced speech segments have valid pitch information and unvoiced segments always have zero pitch, voiced and unvoiced segments are first separated according to the pitch information (F0) extracted in the first module. Given extracted feature vectors from all voiced speech segments, denoted as \( x_t, c_t, \Delta c_t, \Delta^2 c_t \) for all \( t \), where \( x_t \) denotes mel-ccep feature and its dynamics, \( \mu_t \) normalized F0, \( \Delta c_t \) delta F0 and \( \Delta^2 c_t \) delta-delta F0, for each Gaussian mixture component in GMM, to say \( k = 1, \cdots, K \), a 3-element average pitch vector is computed as follows:

\[
\begin{align*}
\mu_t = & \sum_{k} c_t \cdot Pr(k|x_t) \\
\Delta c_t = & \sum_{k} \Delta c_t \cdot Pr(k|x_t) \\
\Delta^2 c_t = & \sum_{k} \Delta^2 c_t \cdot Pr(k|x_t)
\end{align*}
\]

where \( Pr(k|x_t) \) denotes the posterior probability of a voiced speech frame belonging to \( k \)-th Gaussian component, which is calculated according to its mel-ccep feature \( x_t \).

Note that the above GMM model (including the pitch stream) is estimated for each target speaker beforehand.

2.4. GMM Selector

During voice conversion processing, for any input speech frame from any source speaker, the target GMM model is used to select the best matched Gaussian component based on its mel-ccep feature \( y_t \):

\[
l_t^* = \arg \max_{l} w_k \cdot N(y_t | \mu_k, \Sigma_k).
\]

where the Gaussian distribution \( N(y_t | \mu_k, \Sigma_k) \) is with the mean \( \mu_k \) and covariance matrix \( \Sigma_k \).

2.5. Concatenator

For each source speaker utterance, all selected Gaussian mean vectors \( \mu_{ij} \), covariance matrices \( \Sigma_{ij} \) and their corresponding average pitch vectors \( \{ \mu^*, \Delta c^*, \Delta^2 c^* \} \) are concatenated in the same order as the original source utterance to construct a new sequence of feature vectors as:

\[
\gamma = \{ m_t | \forall t \}, \quad m_t = (\mu_{1t}, \Sigma_{1t}^{-1}, c_{t1}, \Delta c_{t1}, \Delta^2 c_{t1}).
\]

2.6. Smoother

Next, the concatenated feature sequence is smoothed according to its delta and delta-delta information in the same way as in HMM-based speech synthesis [8, 12]. A new sequence of static mel-ccep features \( \{ c_t \} \) and static pitch features \( \{ f_t \} \) is generated accordingly. This smoothing method was proved more effective than the simple shift of the average pitch F0 of the source speaker towards the target speaker in HMM-based speech synthesis [8], so that we decided to use this technique in our VC method.

In feature extractor, dynamic features, including deltas and accelerations, are computed using fixed weighting coefficients, \( w^{(j)}(\tau), j \in \{1, 2\} \), as follows:

\[
\Delta c_t = \sum_{\tau=1}^{L(1)} w^{(1)(\tau)} c_{t+\tau} \\
\Delta^2 c_t = \sum_{\tau=1}^{L(2)} w^{(2)(\tau)} c_{t+\tau}.
\]

Based on the above concatenated feature sequence, new static mel-ccep features \( C = [c_1, c_2, \cdots, c_T]^T \) can be estimated by solving a system of linear equations as follows [8, 12]:

\[
C = (W^T U^{-1} W)^{-1} W^T U^{-1} S^T,
\]

Figure 1: The overall voice conversion system using model-based speech synthesis techniques.
where
\[
S = [\mu_{l_1}^T, \mu_{l_2}^T, \ldots, \mu_{l_T}^T]^T \quad (8)
\]
\[
U^{-1} = \text{diag}[\Sigma_{l_1}^{-1}, \Sigma_{l_2}^{-1}, \ldots, \Sigma_{l_T}^{-1}] \quad (9)
\]
\[
W = [w_1, w_2, \ldots, w_T]^T \quad (10)
\]
\[
w_t = [w_t^{(0)}, w_t^{(1)}, w_t^{(2)}] \quad (11)
\]
\[
w_t^{(n)} = [0_{M \times M}, \ldots, 0_{M \times M}, w_t^{(n)}(\mathbf{L}_t^{(n)})_{M \times M},
\ldots, w_t^{(n)}(\mathbf{L}_t^{(n)})_{M \times M}, w_t^{(n)}(\mathbf{I}_M)_{M \times M}, \ldots, \mathbf{0}_{M \times M}]^T, n = 0, 1, 2.
\]

where \(M\) is the number of the dimensions of static mel-cep features without dynamic features and \(l_t^*\) denotes the index of the best-matched Gaussian component of \(y_t\), selected as in Eq. (4).

The second stream of pitch vectors, we repeat the same procedure to calculate a new smoothed pitch \(\mathbf{F}_0\) sequence, i.e., \(\{f_t\}\).

2.7. MLSA filter

Finally, the smoothed static mel-cep features \(\{c_t\}\) and pitch features \(\{f_t\}\) are converted into waveform signals using a mel log spectrum approximation (MLSA) filter, which has been widely used in HMM-based speech synthesis [12].

3. Online Model Adaptation

In the above framework of voice conversion using the HMM-based speech synthesis technique, the Gaussian selection will significantly impact the voice quality of the transformed speech signals. When the input source speech is well-matched to the target-speaker trained GMM, Gaussian selection accuracy will generally produce acceptable converted voice quality. However, too many errors in Gaussian selection will result in very poor converted voice quality, e.g., pitch discontinuity, hoarse vowel sounds and other audible distortions. In the proposed VC system, we use an online model adaptation method based on maximum likelihood linear regression (MLLR) to improve Gaussian selection and reduce mismatches between the target and source speakers. MLLR-based model adaptation has been widely used for speech recognition, HMM-based speech synthesis and etc., however, it is still unclear if the MLLR is also applicable to our particular scenario. It is also one of the purposes of this article to explore this point. More specifically, in our VC application, given a source speaker utterance, we estimate a single linear transformation to adapt the pre-trained GMM \(G\) into a new GMM \(G'\), which is expected to better match a particular source speaker. The adapted GMM \(G'\) is used for Gaussian selection in the GMM selector, and the corresponding Gaussian components in the original GMM \(G\) are used for smoothing and voice generation. This is quite feasible, as we maintain a clear one-to-one index mapping between Gaussian components of \(G\) and those of \(G'\) during model adaptation.

The goal of the MLLR model adaptation is to seek a linear transform \(W\) based on one utterance \(X\) from the source speaker, thereby projecting each Gaussian mixture in GMM \(G\) into a new position under an MLE framework. In this work, we adapt only mean vectors of GMM, i.e., mean vector of \(i\)-th Gaussian \(G_i\), \(\mu_i\), is mapped into \(\mu_i'\) using a linear regression as:
\[
\mu_i' = W^{(0)} \mu_i + b = W \xi, \quad \text{where } \xi = [\mu_i; 1] \text{ is an extended mean vector for Gaussian mixture } G_i.
\]

The linear regression \(W\) is estimated online from a given utterance \(X = \{x_1, \ldots, x_T\}\) based on maximum likelihood criterion as follows:
\[
W = \arg \max_{\mathbf{W}} \prod_{t=1}^{T} \Pr(X_t \mid G') = \arg \max_{\mathbf{W}} \sum_{t=1}^{T} \log \Pr(X_t \mid G').
\]

4. Evaluations

Performance of the proposed VC system is evaluated in terms of the quality of the converted voice as well as the similarity between the target and converted speech. Sentences from the database [2] produced by one male and one female talker were used for evaluation. The sampling rate for all input speech waveforms was 16 kHz. The male speaker was the target speaker and the female speaker was the source speaker. Note that the spectral characteristics of the source female speaker were clearly different from those of the male target speaker, making the transformation more difficult than VC within the same gender.

As in [8, 12], for each speech frame, we use the HTS analysis tool to extract 75-d mel-cep, consisting of 25-d static, delta and acceleration features. A GMM with 1024 Gaussian components was estimated for the male target speaker from 680 utterances. Each utterance from the female source speaker was transformed towards the target male speaker.

4.1. Objective measure based on average frame log-probability

We first examine the effectiveness of the online model adaptation. Note that the online model adaptation is based on MLE. Thus, given a test data \(X\), the probability \(\log p(X \mid G')\) based on the adapted GMM \(G'\) will always be larger than the probability \(\log p(X \mid G)\) based on the original model \(G\) prior to adaptation. We use this measure to objectively evaluate the online model adaptation. Ten sentences are randomly selected to calculate average frame log-probability (AFLP) for each utterance. As shown in Fig. 2, the AFLPs with online adaptation are significantly larger than those without adaptation. This indicates that the online adaptation significantly reduces the mismatch between source and target speaker and that the adapted GMM \(G'\) better models the source speaker.

4.2. Subjective evaluation

Next, VC quality is subjectively evaluated in terms of naturalness and in terms of similarity between the converted and target speech.

4.2.1. Voice quality

Similar to previous studies [10, 7], a preference test regarding speech quality is conducted. Forty utterances are randomly selected from the source female speaker to transform towards the male target speaker. Six normal-hearing (NH) listeners participate in the experiment. During testing, a pair of transformed utterances with or without online model adaptation is presented in random order. Listeners are asked to specify which sample
sounds more natural. The preference scores are listed in Table 1.

Table 1: Preference score of the ABX test

<table>
<thead>
<tr>
<th>Listener id</th>
<th>no adaptation</th>
<th>with adaptation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABX</td>
<td>35%</td>
<td>65%</td>
</tr>
</tbody>
</table>

Results indicate that listeners prefer the transformation with online adaptation more than without adaptation. The results suggest that the online model adaptation can reduce speaker mismatch and significantly enhance VC voice quality.

4.2.2. Naturalness and Similarity

VC quality is also evaluated in terms of naturalness and similarity to the un-processed target speech, which is quite similar to the standard ABX test widely applied in the VC evaluation. A numerical scale from 1 to 5 is used, similar to evaluations of speech synthesis systems. Based on the previous results in section 4.2.1, only the quality of the converted voices with online model adaptation is evaluated. Five NH listeners with previous experience in voice quality evaluation for speech synthesis and voice conversion systems have participated in the experiment. Forty utterances from the female source speaker are randomly selected for conversion. For each converted utterance, listeners are asked to give a score between 1 and 5 to indicate the naturalness of the conversion. Similarly, subjects are asked to give a score between 1 and 5 regarding the similarity between the converted speech and the male target speech. Mean scores (across all forty utterances) are shown in Table 2. The average score (across subjects) for naturalness and similarity is 2.59 and 2.88, respectively, within the range of typical scores for HMM-based speech synthesis systems (2.0 ~ 3.0). This suggests that the proposed VC system can generate acceptable voice quality without using training data from a source speaker.

Table 2: Subjective evaluation of naturalness and similarity.

<table>
<thead>
<tr>
<th>Listener id</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>naturalness</td>
<td>3.09</td>
<td>2.86</td>
<td>2.59</td>
<td>2.45</td>
<td>1.95</td>
<td>2.59</td>
</tr>
<tr>
<td>similarity</td>
<td>3.16</td>
<td>3.06</td>
<td>2.76</td>
<td>3.50</td>
<td>1.88</td>
<td>2.88</td>
</tr>
</tbody>
</table>

Figure 2: Comparison of average frame log-probability of ten utterances randomly selected from the source speaker before and after model adaptation.

5. Conclusion

In this paper, we have presented a new approach for voice conversion when training data from source speaker is unavailable. The proposed VC system employs state-of-the-art HMM-based speech synthesis techniques. Moreover, MLLR-based online model adaptation is used to improve model selection accuracy to further enhance voice quality. The objective evaluation and subjective evaluations demonstrate that the proposed VC techniques can generate acceptable voice quality without using any training data from source speakers.

6. Acknowledgements

We sincerely thank Zhen-Hua Lin and Yu Hu from iFlyTek Co. Ltd. (www.iflytek.com/english) for arranging the subjective evaluation experiment in section 4.2.2 and John J. Galvin III from House Ear Institute for editorial assistance.

7. References
