Voice Conversion Using Generative Trained Deep Neural Networks with Multiple Frame Spectral Envelopes

Ling-Hui Chen\textsuperscript{1,2}, Zhen-Hua Ling\textsuperscript{1}, Li-Rong Dai\textsuperscript{1}

\textsuperscript{1}University of Science and Technology of China, Hefei, P.R. China
\textsuperscript{2}iFLYTEK Research, USTC iFLYTEK Co., Ltd., Hefei, P.R. China

\{chenlh,zhling,lrdai\}@ustc.edu.cn

Abstract

This paper presents a deep neural network (DNN) based spectral envelope conversion method. A global DNN is employed to model the complex non-linear mapping relationship between the spectral envelopes of source and target speakers. The proposed DNN is generatively trained layer-by-layer by cascade of two restricted Boltzmann machines (RBMs) and a bidirectional associative memory (BAM), which are considered as generative models estimated using the contrastive divergence algorithm. Further, multiple spectral envelopes are adopted instead of dynamic features for better modeling using the DNN. The superiority of the proposed method is validated by the subjective experimental results.

Index Terms: Deep neural network, spectral envelope, voice conversion

1. Introduction

The conventional joint density Gaussian mixture model (JDGMM) based voice conversion methods \cite{1} suffer a severe quality degradation in the converted speech. There are three reasons that could cause this degradation. The first one is the over-smoothing effect in generated speech parameters caused by the statistical averaging of the model \cite{2}. The second one is the use of low-dimensional spectral features, e.g. mel-cepstra. Lots of spectral details are lost during extracting them from raw spectral envelopes. The last one is the derived piece-wise linear mapping relationship which may be insufficient to describe the complex non-linear relationship between two speakers.

A spectral trajectory conversion approach was proposed to integrate dynamic features \cite{3} and the global variance (GV) into the maximum output probability parameter generation (MOPPG) criterion \cite{4} in order to compensate the over-smooth effect in the parameter generation. Recently, there have been some approaches that adopt neural networks to model the non-linear mapping relationship between the spectral features of source and target speakers \cite{5, 6, 7}. Previously, we have proposed to use the mixture of restricted Boltzmann machines (MoRBM) \cite{8} and mixture of Gaussian bidirectional associative memories (MoGBAM) \cite{9} for spectral envelope conversion. Although these two models showed their advantages in modeling and converting high-dimensional spectral envelopes, they have their limitations. The restricted Boltzmann machine (RBM) has a strong ability in modeling the joint distribution of spectral envelopes, but its not straightforward to derive the conversion function from the MoRBM. On the other hand, although the Gaussian bidirectional associative memory (GBAM) can model the inter-dimensional correlations, it is theoretically equivalent to a single Gaussian distribution. Therefore, the modeling ability of a MoGBAM is limited and the derived mapping function is still a piece-wise linear transformation.

In this paper, we propose to use a deep neural network (DNN) to construct a global non-linear mapping relationship between the spectral envelopes of source and target speakers. The proposed four-layer DNN is trained layer-by-layer by a cascade of a Bernoulli BAM (BBAM) and two RBMs. The RBMs are employed to model the distributions of spectral envelopes of the source and target speakers respectively. The BBAM is employed to model the joint distribution of hidden variables extracted from the two RBMs. Our proposed method is different from the conventional feedforward neural network based methods \cite{5, 6}, which are trained using the back propagation (BP) algorithm with the minimum mean square error (MMSE) criterion, in three aspects: 1) the model is proposed to convert the spectral envelopes directly, 2) the intermediate network that connects the two RBMs is a BBAM, which is also a generative model, 3) no further fine-tuning is performed for jointly optimizing the parameters in all layers. The third point is essential in the proposed method, because the MMSE criterion in the conventional methods isn’t consistent with the human auditory perceptions. The proposed generatively trained DNN (GTDNN) takes both the advantages of the strong modeling and generating abilities of RBMs and the superiority of BBAMs in deriving the conditional distributions. Further, the effect of the proposed method using multiple consecutive frames of spectral envelopes as features is studied.

This paper is organized as follows: In section 2, we briefly review the spectral envelope conversion methods using MoRBM and MoGBAM. In section 3, we explain the details of our proposed method. The experimental conditions and results are shown in section 4. The conclusions are given in section 5.

2. Voice Conversion Using MoRBMs and MoGBAMs

2.1. RBMs and BAMS

The RBM \cite{10} and BAM \cite{11} are two kinds of stochastic neural networks. They both have bipartite structures. The difference between an RBM and a BAM is that there are a visible and a hidden layer in an RBM while there is no hidden layer in a BAM. Both these two-layered stochastic neural networks can be
considered as probability distribution models, their model parameters can be estimated under the maximum likelihood (ML) criterion by the contrastive divergence (CD) algorithm [12].

2.2. Spectral conversion using MoRBM and MoGBAM

In the trajectory modeling and conversion method [4], the spectral features consist of static, velocity and acceleration components. E.g. for the source speaker, the t-th frame of feature in the feature sequence \( x = [x_1^t, \ldots, x_T^t] \) is composed by \( x_i^t = [x_i^v(s), x_i^a(s), x_i^a(a)^T]^T \), where \( x_i^v(s) \) is the static feature at t-th frame,

\[
x_i^v(s) = 0.5x_i^v_{i-1} - 0.5x_i^v_i,
\]

\[
x_i^a(s) = x_i^a_{i-1} - 2x_i^v_i + x_i^v_{i-1}
\]

are the corresponding velocity and acceleration components respectively. \( x \) can be represented by a linear transformation of the static feature sequence \( x^v(s) = [x_1^v(s), \ldots, x_T^v(s)]^T \) with a fixed transform matrix \( M \). In a similar way, the spectral sequence of target speaker \( y = [y_1, \ldots, y_T] \) can be obtained from the corresponding static sequence \( y^v(s) = [y_1^v(s), \ldots, y_T^v(s)]^T \).

In the modeling of spectral envelopes using the MoRBM and MoGBAM [8,9], RBMs and GBAMs are adopted to describe the distributions of joint spectral envelopes \( z = [x^v(s), y^v(s)]^T \) instead of single Gaussian distributions in each mixture component of a GMM. Note that a GBAM is equivalent to a single Gaussian distribution with a full structured covariance matrix. Then at the conversion stage, given a sequence of input spectral envelopes \( x = [x_1, \ldots, x_T]^T \), a sequence of conditional distributions of output spectral envelopes \( P(y|x_t) \) can be derived from the MoRBM and MoGBAM. Similar with that in the JDGMM based spectral conversion method [4], a sub-optimal mixture sequence is adopted for computational efficiency. Further, in order to use the MOPPG algorithm [13] to generate static spectral envelopes, the conditional distribution derived from an RBM is approximated by a single Gaussian distribution [14], which is

\[
P(y_t|x_t) \simeq \mathcal{N}(y_t; \mu_t, \Sigma_t).
\]

where \( \mathcal{N} \) denotes the Gaussian distribution, \( \mu_t \) is the mode of the conditional distribution and \( \Sigma_t \) is the diagonal covariance matrix of the training data in the corresponding mixture component. As for the GBAM, the derived conditional distribution is already a single Gaussian distribution with diagonal covariance matrix. Therefore, a closed-form solution can be derived to generate a sequence of converted static spectral envelopes \( y^{(s)} \), which is

\[
y^{(s)} = (M^T U^{-1} M)^{-1} M^T U^{-1} E,
\]

where \( E \) and \( U \) are the concatenations of all mean vectors and covariance matrices of all the derived conditional distributions.

3. Voice Conversion Using DNNs

Both the MoRBM and MoGBAM have their limitations in spectral envelope conversion. Although RBMs have strong abilities in modeling the distributions of spectral envelopes, it is not straightforward to derive the conditional distributions from the joint distributions at the conversion stage. Some rough approximations were made at the conversion stage of MoRBM based method [8]. The performance of this method could be degraded by those approximations. On the other hand, GBAMs have no hidden units and therefore it is straightforward to derive the conditional distributions from GBAMs. Although a GBAM can effectively model the inter-dimensional correlations, it is theoretically equivalent to a Gaussian distribution whose modeling ability is limited for complex distributions. Besides, the mapping relationship constructed by MoGBAM is still a piece-wise linear transformation.

In this section, in order to take the advantages of both models and avoid their disadvantages, we proposed to use a DNN to combine BAMS and RBMs. As illustrated in the right part of Figure 1, the proposed model is a four-layer feedforward DNN, including an input layer, an output layer and two hidden layers. The input and output layers denote the stochastic variables corresponding to the spectral envelope of source and target speakers respectively.

3.1. Generative training of DNN

Different from the conventional feedforward neural networks for regression tasks, which are usually trained using the BP algorithm under the MMSE criterion, the proposed DNN is trained layer-by-layer by generative learning. The training procedure is illustrated in Figure 1, where each two directly connected layers is considered as an undirected stochastic neural network and trained under ML criterion. In other words, two RBMs are adopted to describe the distributions of spectral envelopes of the source and target speakers respectively. Then, a BBAM is employed to model the joint distribution of the hidden variables extracted from two RBMs. The extracted hidden variables can be considered as the high-order binary representations of the raw spectral envelopes. Therefore, the parameter set of the proposed DNN is given by

\[
\psi = \{W_s, W_g, W_\theta, b_s, b_g, b_\theta\},
\]

where \( \theta \) and \( \theta_s \) are copied from the parameters of the RBM trained using the training spectral envelopes of source speaker, \( \theta_g \) is copied from those of the BBAM trained using the extracted paired data. Note that the bias terms \( b_s \) and \( b_g \) are fixed to those learned in \( \theta_s \) and \( \theta_g \) and not updated in the training of the BBAM. The learning algorithm for the BBAM is similar to that of the GBAM as introduced in [9]. A subtle difference is that the conditional distributions for
sampling data on the Gibbs chain when performing CD algorithm are Bernoulli distributions. It is worthwhile to note that the proposed DNN can be easily extended to a deeper network by replacing the RBMs with deeper stochastic neural networks, such as deep belief networks (DBNs) [15, 16] or deep Boltzmann machines (DBMs) [17].

3.2. Spectral envelope conversion using DNN

Since each layer, except the input layer, in the DNN only depends on its previous layer. Therefore, the conversion can be performed layer-by-layer at the conversion time. E.g. for an input frame $x_t$, the conditional distribution of the output $y_t$ can be derived from the DNN layer-by-layer. It can be approximated by

$$P(y_t|x_t) \approx P(y_t|h_{y,t}, \theta_y),$$

where

$$h_{y,t} \sim P(h_{y,t}|h_{x,t}, \theta_y),$$

$$h_{x,t} \sim P(h_{x,t}|x_t, \theta_x),$$

are the samples for hidden variables drawn from the corresponding conditional distributions. In this paper, the binary data for $h_{x,t}$ is sampled as follows

$${\tilde{h}_{x,t,i}} = \begin{cases} 1, & P(h_{x,t,i} = |1| x_t, \theta_x) \geq 0.5 \text{,} \\ 0, & \text{otherwise} \end{cases},$$

where $\tilde{h}_{x,t,i}$ is the $i$-th element in $h_{x,t}$, $\tilde{h}_{y,t}$ is sampled in the same manner. The conditional distribution (6) can be derived from the RBM $\theta_y$, it is a single Gaussian distribution

$$P(y_t|\tilde{h}_{y,t}, \theta_y) = \mathcal{N}(y_t; {\Sigma_y}^{-1} (W_y^T \tilde{h}_{y,t} + b_y), {\Sigma_y}),$$

where the covariance $\Sigma_y$ is diagonal because the units in the output layer are independent on each other once the variables in layer $h_y$ are given. The conditional distributions given different input spectral envelopes share the same diagonal covariance matrix. Therefore, the parameter generation algorithm in the conventional methods (4) can be applied to generate the converted spectral envelopes.

3.3. Spectral conversion using multiple frames

In conventional trajectory conversion methods, the spectral envelope features usually consist of static and dynamic (velocity and acceleration) components. From Fig. 2 we see the correlations between static and dynamic components are much weaker than that of triple frame. Empirically, neural networks are good at modeling spectral features with strong inter-dimensional correlations [16]. In acoustic modeling of speech recognition, it is observed that the performance of DNN using raw filter-bank features is better than that of using the weak inter-dimensional correlated acoustic features, e.g. mel frequency cepstral coefficients (MFCCs) [18, 19].

In this paper, we investigate the effect of using multiple frame spectral envelopes instead of those with static and dynamic components. In order to keep the same information in input features, three consecutive frames of spectral envelopes are adopted as the input of the DNN, e.g. the feature at $t$-th frame of the spectral sequence is $x_t = [x_{t-1}^T, x_{t-1}^T, x_{t+1}^T]^T$. This triple-framed feature sequence can also be obtained from the static sequence $x$ by a fixed linear transformation. The feature extraction and parameter generation can be implemented simply by modifying the coefficients in the linear transform matrix. Therefore, the training and conversion method using conventional features can be exactly followed.

4. Experiments

4.1. Experimental conditions

A Chinese speech database containing a male and a female speaker was used in our experiments to construct a female to male conversion. 100 parallel sentences were adopted for each of the two speakers. 80 sentences were selected randomly as the training set and the remaining 20 sentences were used as the test set. The spectral envelopes were directly used as the spectral features in the proposed method and some of the other comparing methods. The spectral envelopes were calculated by the STRAIGHT vocoder [20]. The FFT length of the STRAIGHT analysis was set to 1024, which leads to a 513 dimensional spectral envelope vector for each frame. The frame shift for calculating spectral envelopes was set to 5ms. In the JDGMM based baseline system in our experiments, 40 order mel-cepstra (not including the frame-power dimension) were extracted from the spectral envelopes and used as spectral features. Paired training samples were constructed by performing time alignment between the parallel mel-cepstral sequences of source and target speakers using the dynamic time warping (DTW) algorithm. Then, joint spectral envelopes were constructed using the DTW paths derived from their corresponding mel-cepstra.

In the spectral envelope modeling, the spectral amplitudes of all frames were normalized to the same power and the logarithmic spectral amplitudes were adopted as the spectral features for modeling. The CD learning with 1-step Gibbs sampling (CD-1) was employed to train RBMs and BAMs [12, 21]. This paper focuses on the spectral conversion, the fundamental frequencies ($F_0$) were converted simply using a linear transformation in the log-scale of $F_0$s to equalize the mean and variance of converted and target log $F_0$s [4].

4.2. System construction

We constructed several systems for evaluating the performance of the proposed method (GTDNN), including the conventional JDGMM-based method considering GV in parameter generation (GMM-GV), spectral envelope conversion using two-layer stochastic neural networks (MoRBM and MoGBAM) as introduced in section 2. A DNN initialized by a GTDNN and fine-tuned by MMSE criterion (FTDN) was also constructed for

\[ \text{http://staff.ustc.edu.cn/~chenlh/IS14_DNNVC/demo.html} \]
Table 1: LSDs between the spectral envelopes of natural speech and those converted by the GTDNN, FTDNN and JDGMM.

<table>
<thead>
<tr>
<th>method</th>
<th>LSD (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JDGMM</td>
<td>4.61</td>
</tr>
<tr>
<td>GTDNN</td>
<td>5.38</td>
</tr>
<tr>
<td>FTDNN</td>
<td>4.55</td>
</tr>
</tbody>
</table>

Table 2: The results of preference tests (%) between GTDNN and FTDNN. N/P is short for “no preference”.

<table>
<thead>
<tr>
<th></th>
<th>GTDNN</th>
<th>FTDNN</th>
<th>N/P</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim.</td>
<td>53.57</td>
<td>32.14</td>
<td>14.29</td>
<td>0.006</td>
</tr>
<tr>
<td>nat.</td>
<td>77.14</td>
<td>14.28</td>
<td>8.58</td>
<td>0.000</td>
</tr>
</tbody>
</table>

spectral envelope conversion. The “dropout” regularization [22] was employed in training FTDNN.

A JDGMM with 128 components was trained for GMM-GV system. Then the JDGMM was used to partition the joint spectral space into sub-spaces for modeling and conversion using MoRBM and MoGBAM. 300 hidden units were adopted for RBMs in MoRBM. The number of units in each of the two hidden layers of GTDNN and FTDNN were set to 2048.

4.3. Comparison between GTDNN and FTDNN

The first experiment was conducted to compare the GTDNN with FTDNN. Dynamic features were used in this experiment. The average log-spectral distortions (LSDs) between the converted and target reference spectral envelopes in the test set are shown in Table 1. For a reference, the average LSD of the conventional JDGMM based method when using mel-cepstral was also calculated. We see that the average LSD of the GTDNN is much larger than that of the JDGMM system. After the fine-tuning, the average LSD of the FTDNN system decreases to that slightly smaller than the JDGMM based method. The complex non-linear mapping function described by the DNN can reasonably generate spectral envelopes with lower average LSDs.

Next, we conducted a preference evaluation to compare the subjective performances between GTDNN and FTDNN systems. Seven Chinese-native listeners were involved in these tests. The preference test results are shown in Table 2. We see that GTDNN method significantly outperformed the FTDNN in both similarity and naturalness. These results are the opposite of objective LSD results. The subjective results are not consistent with the objective LSD results because the LSD measurement is not consistently related with human auditive perceptions. The LSD is guaranteed by the generative training criterion of GTDNN. Therefore, the LSDs were not evaluated in next experiments.

4.4. Comparing with conventional methods

In the second experiment, we conducted several preference listening tests to compare the proposed method with some conventional methods, including GMM-GV, MoRBM and MoGBAM. The listening conditions were the same with those in the first experiment. The preference test results are shown in Table 3. Comparing MoRBM with GMM-GV, the subjects preferred the speech generated by GMM-GV although MoRBM performs conversion on spectral envelopes. This can be attributed to the limitations of MoRBM based method as mentioned in section 3. Both MoGBAM and GTDNN systems outperformed the GMM-GV system. Due to the highly non-linear mapping relationship described by a GTDNN, the GTDNN system can generate better voice than the MoGBAM system, in which the mapping relationship is a piece-wise linear transformation. The p-values given by the t-test show the significance in difference between the compared systems.

4.5. Comparison between using dynamic feature and triple frames

In the last experiment, we conducted a preference test between conversion by GTDNN using dynamic feature (GTDNN-df) and triple frames (GTDNN-tf). The listening test conditions were the same with those in the previous tests. We can see from the results shown in Table 4 that although triple-frame features contains the same information with the conventional features with static and dynamic features, using them as features can improve the performance of GTDNN, especially in naturalness (p < 0.05). This can be attributes to the superiority of DNNs/RBMs in capturing the highly inter-dimensional correlations in triple frame spectral envelope features.

5. Conclusions

A DNN based spectral conversion method was proposed for direct spectral envelope conversion in this paper. The DNN is generatively trained by two RBMs and a BBAM. The RBMs are used to extract high-order binary representations from raw spectral envelopes of the source and target speakers. The BBAM is employed to model the joint distribution of the extracted high-order representations. The experimental results show that the proposed method significantly outperformed the conventional methods. However, in this paper, only a DNN with two hidden layers was studied. This is not really a “deep” model. The deeper model can better describe the non-linear mapping relationship between two speakers. Therefore, the future work is studying spectral envelope conversion using generatively trained DNN with deeper structure.
6. References


