NON-PARALLEL TRAINING FOR VOICE CONVERSION BASED ON FT-GMM

Ling-Hui Chen, Zhen-Hua Ling, Li-Rong Dai

iFLYTEK Speech Lab, University of Science and Technology of China, Hefei, P.R. China

Chenh@mail.ustc.edu.cn, zhling@ustc.edu, lrdai@ustc.edu.cn

ABSTRACT

This paper presents a non-parallel training algorithm for voice conversion based on feature transform Gaussian mixture model (FT-GMM), which is a mixture model of joint density space of source speaker and target speaker with explicit feature transform modeling. In FT-GMM, the correlations between the distributions of two speakers in each component of the mixture model are not directly modeled, but absorbed into these explicit feature transformations. This makes it possible to extend this model to non-parallel training. The joint space model can be easily decomposed into two sub-models for source and target speaker respectively, while taking account of all parameters of original FT-GMM. Non-parallel model training is accomplished by maximizing the likelihood of these two sub-models using the corresponding speakers’ training data. However, we found that the performance of this method degraded severely when spectral distance between source and target is large, e.g. cross-gender conversion. To overcome this problem, we adopt a front-end frequency warping process, and use the warped acoustic features as training data for the source speaker.

The rest of this paper is organized as follows. Section 2 briefly reviews the FT-GMM based voice conversion. Section 3 discusses the extending and model training of FT-GMM to non-parallel case. Experimental results are shown in section 4. Conclusions are presented in 5.

1. INTRODUCTION

Voice conversion is a technique that modifies the input speech of one speaker (source speaker), and makes it sound like that uttered by another speaker (target speaker). This technique has become a hot topic of research because of its widely potential application: a high-performance voice conversion system may be a flexible extension for speech synthesis system, an important part of speech-to-speech translation, etc. It may have many helpful applications in human communication.

Recently, many approaches to voice conversion have been studied, such as codebook quantization [1], GMM [2], joint density GMM (JDGMM) [3, 4] and linear regression [5], etc. GMM based approaches are statistical ones that perform soft clustering on source speaker’s (or joint) acoustic feature space, and established mapping relation via space distributions. It has been widely studied because of its stable performance.

However, most of these methods were based on parallel training data. The requirement of parallel training utterances limited their application in many scenarios. In many online applications, parallel training samples are difficult to obtain, especially in cross-lingual conversion, which is an important component in speech-to-speech translation, parallel data does not even exists. Recently, some researchers have studied on non-parallel training of voice conversion. e.g. Mouchtaris et al. [6] proposed a method that adapts source and target distributions of an existing JDGMM to the new source and target speakers; Erro [7] et al. proposed a frame selection strategy to obtain “parallel” data artificially from non-parallel datasets, etc.

In this paper, we proposed a non-parallel training algorithm for FT-GMM [8], which is a reformed GMM for joint density space with explicit feature transformation in each mixture component. In parallel training case, FT-GMM outperformed conventional JDGMM [8]. In FT-GMM, distribution for target speaker and covariances between source and target speaker are not directly modeled but absorbed into these explicit feature transformations. This is the key factor that makes it possible to easily extend FT-GMM to non-parallel training. The joint space model can be easily decomposed into two sub-models for source and target speaker respectively, while taking account of all parameters of original FT-GMM. Non-parallel model training is accomplished by maximizing the likelihood of these two sub-models using the corresponding speakers’ training data. However, we found that the performance of this method degraded severely when spectral distance between source and target is large, e.g. cross-gender conversion. To overcome this problem, we adopt a front-end frequency warping process, and use the warped acoustic features as training data for the source speaker.

The rest of this paper is organized as follows. Section 2 briefly reviews the FT-GMM based voice conversion. Section 3 discusses the extending and model training of FT-GMM to non-parallel case. Experimental results are shown in section 4. Conclusions are presented in 5.
where $\lambda_G = \left\{ w_m, \mu^{(s)}_m, \Sigma^{(s)}_m \mid m = 1, 2, \ldots M \right\}$ related to the source distribution and $\lambda_T = \left\{ \Gamma_m, \Sigma^{(o)}_m \mid m = 1, 2, \ldots M \right\}$ related to explicit transformations. At conversion phase, $\lambda_G$ is used only for (soft/hard) classifying source features, feature transformation and spectral parameter generation are accomplished by $\lambda_T$.

3. NON-PARALLEL TRAINING FOR FT-GMM

3.1. Extend FT-GMM to non-parallel training

In parallel training, the GMM (JDGMM/FT-GMM) $\lambda^{(s)}_G$ is estimated from joint features $z^{(P)}$ via the ML criterion:

$$\hat{\lambda}^{(s)}_G = \arg\max_{\lambda^{(s)}_G} p(z^{(P)} | \lambda^{(s)}_G)$$

(5)

but in non-parallel cases, this kind of aligned feature pair is not available, all we have are two sub-sequences $x$ and $y$: $z^{(NP)} = \{ x, y \}$. Since we have no information about the relation between $x$ and $y$, instead, all we can assume that they are independent of each other. Under this assumption, we can rewrite equation (5) for non-parallel training of model $\lambda^{(x)}_{NP}$ as:

$$\hat{\lambda}^{(x)}_{NP} = \arg\max_{\lambda^{(x)}_{NP}} p(z^{(NP)} | \lambda^{(x)}_{NP})$$

(6)

$$= \arg\max_{\lambda^{(x)}_{NP}} p(x | \lambda^{(x)}) \cdot p(y | \lambda^{(y)})$$

(7)

where $\lambda^{(x)}$, $\lambda^{(y)}$ represent the sub-models derived from $\lambda^{(x)}_{NP}$ that describe distribution of source speaker and target speaker's acoustic space respectively.

In non-parallel training of FT-GMM, for the convenience of deriving parameter estimation formulae, we define a model set slightly different from that of the parallel training:

$$\lambda^{(x)}_{NP} = \left\{ w_m, \mu^{(y)}_m, \Sigma^{(y)}_m, \Gamma_m, \Sigma^{(o)}_m \mid m = 1, 2, \ldots M \right\}$$

(8)

In this case $\lambda_G = \lambda^{(y)}$. The transformations $\Gamma_m$ are still mapping from source space to target space. As a result, an equivalent JDGMM is different from that of parallel training case:

$$\mu^{(s)}_m = \begin{bmatrix} A_m^{-1} (\mu^{(y)}_m - b_m) \\ \mu^{(y)}_m \end{bmatrix}$$

(9)

$$\Sigma^{(s)}_m = \begin{bmatrix} A_m^{-1} R_m A_m^{-1\top} & A_m^{-1} R_m \\ R_m A_m^{-1\top} & \Sigma^{(y)}_m \end{bmatrix}$$

(10)

where $R_m = \Sigma^{(y)}_m - \Sigma^{(o)}_m$.

As a result, the sub-model for the source speaker’s space can be derived from $\lambda^{(x)}_{NP}$ as:

$$\lambda^{(x)} = \left\{ w_m, A_m^{-1} (\mu^{(y)}_m - b_m), A_m^{-1} R_m A_m^{-1\top} \mid m = 1, 2, \ldots M \right\}$$

(11)

The conversion phase of non-parallel FT-GMM is similar with that of the parallel case. The difference is that the model $\lambda^{(x)}$ is a full covariance model. Since $\lambda^{(x)}$ is only used for determining the sub-optimal Gaussian component sequence, an approximation using diagonal covariances should not bring considerable degradation. So for simplicity, we only used an approximated model with diagonal covariances for $\lambda^{(x)}$ in the conversion phase.

3.2. Model training

In 3.1, we derived sub-models $\lambda^{(x)}$ and $\lambda^{(y)}$ for model training. In this section, we present the parameter estimation algorithm for non-parallel training of FT-GMM. According to (7), a two-step estimation could be employed here:

Step 1: $\hat{\lambda}^{(y)} = \arg\max_{\lambda^{(y)}} p(y | \lambda^{(y)})$

Step 2: $\hat{\lambda}^{(x)} = \arg\max_{\lambda^{(x)}} p(x | \lambda^{(x)})$

$$= \arg\max_{\lambda^{(x)}} p(x | \lambda^{(y)}, \lambda^{(x)})$$

In this case, we still cannot estimate mixture weights of the joint density space, because there isn’t any training sample actually from joint density space. To address this problem, we simply make an assumption that joint space, source space and target space share the same mixture structure. Under this assumption, we simply use mixture weights estimated in Step 1 as that of joint space.

Estimation of $\lambda^{(y)}$ in Step 1 is a typical GMM training problem, which can be easily accomplished using the EM algorithm. So, in the rest of this section, we focus on the estimation of $\lambda^{(x)}$ in Step 2.

The expected log likelihood of outputting feature sequence $x$ from model $\lambda^{(y)}$ can be written as:

$$L(\Gamma_m, \Sigma^{(o)}_m) = K - \frac{1}{2} \sum_{t=1}^{T} \sum_{m=1}^{M} \sum_{i,j} \zeta_m(t) | \log | R_m | - \log | A_m |^2$$

$$+ (\Gamma_m \xi_t - \mu^{(y)}_m)^\top R_m^{-1} (\Gamma_m \xi_t - \mu^{(y)}_m)$$

(12)

where $\zeta_m(t) = P(x_t, m | \lambda^{(y)}) / \sum_{m=1}^{M} P(x_t, m | \lambda^{(y)})$ is the probability of outputting $x_t$ from $m$-th mixture, and $K$ is a constant value that is independent of $\lambda^{(x)}$. The transformations on model parameters in (11) become transformations on features. This makes the estimation problem simpler. Maximizing $L$ is a regularized linear regression problem with the regularization term $\log | A_m |^2$ to help avoiding overfitting. These are the reasons why we adopted model parameter set (8) instead of (3).

Note that this process is very similar with the constrained maximum likelihood linear regression (CMLLR, also named feature-space MLLR) problem, which has been proved to be an effective technique for HMM based adaptation [9].

By setting derivatives of $L(\Gamma_m, \Sigma^{(o)}_m)$ with respect to $\lambda^{(x)}$, we can get its estimation formula:

$$\gamma_{m,i} = (xp_{m,i} + k(i) G^{(s)}_{m,i}) G^{(s)}_{m,i}^{-1} \cdot \rho = \frac{\gamma_{m,i}^2}{p_{m,i}}$$

(13)

$$\Sigma^{(s)}_m = \Sigma^{(y)}_m - \sum_{t=1}^{T} \sum_{m=1}^{M} \zeta_m(t) (\Gamma_m \xi_t - \mu^{(y)}_m)(\Gamma_m \xi_t - \mu^{(y)}_m)^\top \sum_{t=1}^{T} \zeta_m(t)$$

(14)

where $\gamma_{m,i}$ is the $i$-th row of matrix $\Gamma_m$, $\xi_t = [x_t \top, 1]^\top$, $p_{m,i} = [0, c_{m,i,1}, c_{m,i,2}, \ldots, c_{m,i,n}]$, $c_{m,i,j}$ is the $(i, j)$-th element of the adjacent matrix $A^{(s)}_m$ of $A^{(y)}_m$, $n$ is the feature order, $\sigma_{m,i}$ is the $i$-th element of the diagonal matrix $R_m$, and

$$\beta_m = \sum_{t=1}^{T} \zeta_m(t)$$

(15)

$$G^{(s)}_{m,i} = \sum_{t=1}^{T} \zeta_m(t) \xi_t \xi_t^\top$$

(16)
\[ k_m^{(i)} = \mu_m^{(y)} \sum_{t=1}^{T_x} \zeta_m(t) \xi_t^\top \]  

in which \( \mu_m^{(y)} \) is the i-th element of \( \mu_m^{(y)} \).

### 3.3. Frequency warping

In our early experiments, we found that the method introduced in section 3 is effective in intra-gender conversion, but its performance degraded severely when it came to cross-gender cases. We ascribe this degradation to the distinct difference between spectral structures of female and male. Normally, formants of a female are obviously higher than that of a male. In the training phase of \( \Lambda_T \) described in 3.2, \( \zeta_m(t) \) plays an important role; it determines the accuracy of aligning the training sample \( x_t \) to each Gaussian component, as dynamic time warping (DTW) algorithm performs feature alignment in the parallel training case, but this is an alignment between model and data. However, in our training algorithm, we just use \( \Lambda_T^{(y)} \) as the initial model of \( \Lambda_T(z) \), the mismatch between training data and initial model causes significant error in calculating \( \zeta_m(t) \). This error can not be corrected by an iteration process of model training, and this leads to the degradation in performance. In intra-gender cases, the calculation of \( \zeta_m(t) \) is more reliable.

To overcome this problem, we adopt a frequency warping (FW) technique as a front-end process to reduce this mismatch. Since FW technique, also named vocal tract length normalization (VTLN), has an equivalent effect with linear transformation in cepstral space [10], this process can be viewed as a global linear transformation for the source features. An advantage of FW is that it is only a simple modification on the frequency axis and shouldn’t bring notable degradation of speech quality. Although this technique is too simple and cannot capture the entire characteristic of a target speaker, it can successfully change the gender characteristic of acoustic features and turn a cross-gender conversion into an intra-gender one, which is effective with the proposed training algorithm.

In this paper, the bilinear transformation (BLT) is adopted as the FW function. The BLT in the complex z-domain is defined by:

\[
\tilde{z} = \frac{z + \alpha}{1 + \alpha z}, \quad \text{with } z = e^{j\omega}, \tilde{z} = e^{j\tilde{\omega}}, |\alpha| < 1 \quad (18)
\]

where \( \omega \) denotes the original frequency and \( \tilde{\omega} \) denotes the warped frequency, \( \alpha \) is the warping factor. The optimal warping factor is obtained via the ML criterion:

\[
\alpha_{opt} = \arg \max_{\alpha} p(s(\alpha)|\Lambda_T^{(y)}) \quad (19)
\]

where \( \Lambda_T^{(y)} \) is a GMM for target speaker, and \( s(\alpha) \) is the cepstral coefficients sequence extracted from warped spectrum with warping factor \( \alpha \). Unfortunately there seems to be no closed-form solution for this optimization process (19), but a straightforward solution is to traverse some possible candidates of \( \alpha \) to find the optimal factor.

### 4. EXPERIMENTS

#### 4.1. Databases and setups

We used the ARCTIC database [11] to evaluate the performance of proposed method. Two female speakers (slt and clb) and two male speakers (rms and bdl) were selected to perform two cross-gender conversions: a female to male (clb → bdl, **F2M**) and a male to female (rms → slt, **M2F**) conversion. 40 utterances were selected for the model training, and 20 sentences, not including in the training set, were used as a test set for evaluation.

We compared the proposed method with a baseline parallel FT-GMM as introduced in section 2. The same 40 utterances were used for the source speaker in both baseline and proposed method. For the target speaker, parallel utterances were used in the baseline method, but different 40 utterances were used for the proposed method.

24 order warped cepstral coefficients were extracted, by the STRAIGHT [12] analyzer, as spectral parameters. For the target speaker, Mel-cepstral coefficients were used, while for the source speaker, optimal warped coefficients were used. Acoustic features were 48 dimensional, with static features and their delta features. The total duration of training utterances was about 150 seconds for each speaker, not including silence segments.

We trained a 64 mixture FT-GMM for each conversion pair. The linear transformation in each Gaussian component was bi-block diagonal, one block for static feature transformation and another for dynamic features. Both blocks were full matrices.

#### 4.2. Estimate optimal factor for frequency warping

Before model training, we first estimated FW factors for source speakers as described in section 3.3. Since the spectral parameters we adopted were Mel-cepstral coefficients for target speaker, the FW for the source speaker should be performed in the Mel-frequency domain (Mel-FD). For Mel-FD, \( \alpha_m \approx 0.42 \) for warping from the normal frequency domain (FD). BLT warping in Mel-FD with \( \alpha \) equals to a BLT warping in FD with \( \alpha' = \frac{\alpha + \alpha_m}{1 + \alpha_m} \). In these experiments, we perform FW process in FD and search for the optimal \( \alpha' \) instead of \( \alpha \).

We selected 10 utterances from the training set for estimating the warping factor. Non-parallel utterances were used for the source and target speaker. Note that the model trained in this step is different from that in section 3.1. Only static features were used in this process. A GMM with 8 mixtures was trained for the target speaker. We searched \( \alpha_{opt} \) from 0.25 to 0.65, with step size 0.01. The optimal factors for **M2F** and **F2M** were 0.53 and 0.35 in normal frequency domain (corresponding to 0.14 and −0.08 in Mel-FD) respectively.

After obtaining optimal FW factor \( \alpha_{opt} \), the features used for source speaker in the next model training and conversion phase are both extracted from the warped spectrum.

#### 4.3. System evaluation

##### 4.3.1. Objective evaluation

Average cepstral distortion (CD) between converted cepstral coefficients and reference target was calculated for objective evaluation. DTW algorithm was used to perform time alignment between converted and target feature sequences.

<table>
<thead>
<tr>
<th>CD (dB)</th>
<th>M2F</th>
<th>F2M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source (no FW)</td>
<td>7.9311</td>
<td>7.9499</td>
</tr>
<tr>
<td>Source (FW)</td>
<td>7.0261</td>
<td>7.2175</td>
</tr>
<tr>
<td>Proposed (no FW)</td>
<td>6.6335</td>
<td>6.1126</td>
</tr>
<tr>
<td>Proposed (FW)</td>
<td>5.7553</td>
<td>5.8734</td>
</tr>
<tr>
<td>Baseline (no FW)</td>
<td>4.8131</td>
<td>4.9817</td>
</tr>
<tr>
<td>Baseline (FW)</td>
<td>4.7319</td>
<td>4.9325</td>
</tr>
</tbody>
</table>

Table 1. CD(dB) between spectral parameters converted by each systems and referenced target.
decrease is not large enough, FW has successfully changed the gender characteristic in the acoustic features. FW also successfully improved the performance of the proposed method. However, CD of the proposed systems with FW were still much larger than that of the baseline systems. This is because that the baseline method is a “supervised” method, in which transformations are optimized by minimizing average Mahalanobis distance between transformed features and their referenced target, while the proposed method is an “unsupervised” approach without a referenced target. In the baseline method, improvements brought by adopting FW were less obvious than in the proposed method, because DTW algorithm already have a good frame aligning accuracy.

4.3.2. Subjective evaluation

Last, we conducted subjective listening tests to compare the performance of the proposed method with that of the baseline systems. FW is used in both methods in this test. 6 listeners took part in the tests. 20 converted utterances were played to listeners randomly. They were asked to give a 5-scale opinion score (5: very good, 1: very bad) about each utterance they heard.

The method in this work is proposed for spectral transformation. Prosodic conversion was not addressed here. For tests of similarity and naturalness, prosody conversion is performed in a different way. In similarity tests, in order to exclude the negative effect caused by F0 conversion, converted spectral parameters were aligned to the reference targets using DTW algorithm, and speech waveforms were synthesized directly using \( F_0 \)'s of target. In the naturalness tests, speech waveforms were synthesized using \( F_0 \)'s converted linearly in log scale for the purpose of avoiding unnaturalness brought by DTW.

Fig. 1 shows the mean opinion scores on similarity and naturalness of both conversion pairs. As we can see, the proposed method achieved comparable performance with the baseline system. In particular, the baseline method was superior slightly in similarity, which was reflected in CD, while the proposed method outperformed the baseline a little in naturalness. This agrees with the difference between the two methods in model training. As discussed in section 3.2, the proposed method can avoid over-fitting, and over-smoothing in the spectrum is less obvious than that in the baseline method.

5. CONCLUSION

In this paper, we have proposed a non-parallel training method for spectral transformation of voice conversion based on FT-GMM. Although FT-GMM is a model of the joint density space, it can be decomposed into two sub-models for source speaker and target speaker while still including all information of the FT-GMM. In the training phase of FT-GMM, these two sub-models were optimized using corresponding non-parallel training data. However, the proposed didn’t work well on cross-gender cases because of the mismatch between training data and initial model of the source speaker. We overcome this mismatch by adopting a front-end frequency warping process to change the cross-gender conversions into “intra-gender” ones. Subjective test results show that the proposed method achieved comparable performance to FT-GMM with parallel training.

The effect of FW was still somewhat unstable, so a future work will seek a better method to overcome the mismatch problem.

6. REFERENCES