Resurrecting past singers: Non-Parallel Singing-Voice Conversion

Fernando Villavicencio, Hideki Kenmochi

Corporate Research and Development Center, YAMAHA Corporation, Japan
{villavicencio,kenmochi}@beat.yamaha.co.jp

We present in this work a strategy to perform timbre conversion from unpaired source and target data and its application to the singing-voice synthesizer VOCALOID to produce sung utterances with a past singer voice. The conversion framework using unpaired data is based on a phoneme-constrained modeling of the timbre space and the assumption of a linear relation between the source and target features. The proposed non-parallel framework resulted in a performance close to the one following the traditional approach based on GMM and paired data. The application to convert an original singer database using sung performances of a past singer observed a successful perception of the past singer’s timbre on the singing-voice utterances performed by VOCALOID.

Index Terms: Speech synthesis, speech analysis, linear prediction, pattern recognition

1. Introduction

The emergence of Voice Conversion in the speech community, representing the ability to modify the voice timbre of a person in order to give the perception of another one, allows us to ponder a number of applications in which this technology can be applied. Besides the significant amount of work already done on spoken speech, the application to singing-voice represents a clear opportunity to expand the possibilities of music-related technology. In particular, the appearance of high-quality singing-voice synthesizers, such as VOCALOID [1] may lead us to make possible some interesting applications.

The application of Voice Conversion to perform high-quality singing-voice conversion through VOCALOID was presented by the authors in [2]. This commercial singing-voice synthesizer, widely popular in Japan, is based in the concatenation of diphone units taken from a library consisting of selected recordings of a singer. In that work, GMM-based conversion ([3],[4]) of accurate spectral envelope estimates was applied to perform voice-timbre conversion in order to achieve a multi-speaker capacity from a single database.

GMM-based Voice Conversion is based on a statistical mapping of the timbre space. Typically, source-target paired data is required to perform the mapping of the timbre features. More precisely, the data corresponds to recordings where both source and target utterances follow the same phonetic content. Clearly, by following this approach, the choice of the target speaker (or singer) is limited to those for which such parallel corpus is available.

We present in this work a strategy to achieve non-parallel conversion by exclusively using unpaired data. Our interest is to achieve singing-voice conversion from a VOCALOID DB to match the voice of a singer from whom only some conventional sung material is available (in particular, past singers). Our proposition is based on two main modifications of the conventional GMM technique. Firstly, the modeling of the features corresponding to a phoneme is restricted to a gaussian component, resulting in what we will call as phoneme-constrained multi-gaussian modeling. Then, we consider an approximation based on a proposition found in [5] consisting in a linear relation between the timbre features of two speakers to estimate the joint statistics required to derive the mapping function. The resulting strategy was objectively evaluated by processing paired data as unpaired. The results showed a conversion performance close to the parallel case. Finally, after converting an original singer DB using some recordings from a past singer it was successfully perceived the target timbre on the synthesized singing-voice issued from VOCALOID.

This article is structured as follows. The phoneme-constrained modeling is presented and compared to the conventional GMM approach in section 2. Then, the derivation of the conversion model from unpaired data is described in section 3. The results of an experimental evaluation of the whole proposed methodology are presented in section 4. Finally, the work ends at section 5 with our conclusions and future work.

2. Phoneme-based timbre modeling

2.1. Unsupervised GMM-based conversion

Conventional GMM training in Voice Conversion is achieved in an unsupervised way by maximization of the likelihood of the probabilistic model related to observed paired data of the source and target speakers [3], [4]. The resulting statistical model is assumed to cover the acoustic space of the speakers timbre. Moreover, the components of the resulting mixture are expected to model phonetically meaningful data clusters due to the assumption that, in general, a spectral envelope pattern is attributed to each phoneme.

In general, the size of the GMM (number of gaussian components) showing a robust modelization of the features space depends on the amount of data used to fit the model as well as on the configuration of the covariance matrices. Following the typical values reported on the bibliography a stabilization of the learning can be achieved using a GMM with 8 components and full-matrices or with a significantly increased number of components ([32 – 128]) if using diagonal ones. A clustering of the timbre space by 8 gaussians appears to lead to a large averaging of the phonetic regions regarding the number of phonemes of a language (around 35 for Japanese). Note also that the class membership (conditional probability) resulting of such mixture configuration on continous speech is commonly found to be highly competitive (one single gaussian keeps fully-activated along a stationary region). This may be seen as a full-modeling of different phonemes by the same GMM component.

On the other hand, by using a number of components similar to the number of observed phonemes the behavior of the mixture is less competitive but rather unstable. An unstable switching of the components on a continuous signal may pro-
the amount of data required to fit the Gaussian model is significant and smaller energy values (mainly consonants). Moreover, statistics may not be robust if only a small number of data points correspond to the corresponding data of a phoneme, the resulting Gaussian components based on these associations might be studied in the future.

Voiced plosives (\( \text{g,g',d,d',b,b'} \)), voiced plosives (\( \text{g,g',d,d',b,b'} \)), nasals (\( \text{J,m,N,m',N'} \)), and nasals (\( \text{J,m,N,m',N'} \)) found on phonemes belonging to the same phonetic type as a Gaussian component.

The computation of the Gaussian components is expected to result in a more judicious clustering of the phoneme. This \( \mu_{q}^{y} \) be restricted according to the data corresponding to the same phoneme,\( \mu_{q}^{y} \) is expected to result in a more judicious clustering of the timbre space than the unsupervised way, allowing an individual source-to-target mapping for each phoneme. Fig. 2 shows the matching matrix representing the resulting average membership when using the proposed phoneme-constrained computation of the GMM. The phonetic list correspond to our phonetic segmentation of the Japanese language. Clearly, the cluster-to-phoneme correspondence is significantly increased although it is not perfect.

Note some shared associations are found on phonemes belonging to the same phonetic type as nasals (\( \text{J,m,N,m',N'} \)), voiced plosives (\( \text{g,g',d,d',b,b'} \)), and unvoiced plosives (\( \text{k,k',t,t',p,p'} \)). A reduction in the number of gaussian components based on these associations might be studied in the future.

Note that when restricting the computation of the gaussian elements to the corresponding data of a phoneme, the resulting statistics may not be robust if only a small number of data points is available. However, these cases correspond commonly to phonemes less perceptually important, having the shortest durations and smaller energy values (mainly consonants). Moreover, fixing the covariance matrices of the gaussians to be diagonal the amount of data required to fit the gaussian model is significantly reduced.

2.2. Phoneme-constrained Multi-Gaussian Model

Accordingly, it appears reasonable to directly assign the data of a single phoneme to an individual gaussian component by performing a phonetic segmentation analysis [6] on the signals. Then, the computation of each gaussian component can be restricted according to the data corresponding to the same phoneme. This supervised computation of the gaussian components is expected to result in a more judicious clustering of the timbre space than the unsupervised way, allowing an individual source-to-target mapping for each phoneme. Fig. 2 shows the matching matrix representing the resulting average membership when using the proposed phoneme-constrained computation of the GMM. The phonetic list correspond to our phonetic segmentation of the Japanese language. Clearly, the cluster-to-phoneme correspondence is significantly increased although it is not perfect. Note some shared associations are found on phonemes belonging to the same phonetic type as nasals (\( \text{J,m,N,m',N'} \)), voiced plosives (\( \text{g,g',d,d',b,b'} \)), and unvoiced plosives (\( \text{k,k',t,t',p,p'} \)). A reduction in the number of gaussian components based on these associations might be studied in the future.

Note that when restricting the computation of the gaussian elements to the corresponding data of a phoneme, the resulting statistics may not be robust if only a small number of data points is available. However, these cases correspond commonly to phonemes less perceptually important, having the shortest durations and smaller energy values (mainly consonants). Moreover, fixing the covariance matrices of the gaussians to be diagonal the amount of data required to fit the gaussian model is significantly reduced.

\[
\hat{y} = \sum_{q=1}^{Q} p(q|x) \left[ \mu_{q}^{y} + \sum_{x}^{p} (x - \mu_{q}^{x}) \right] \tag{1}
\]

\[
p(q|x) = \frac{N(x; \mu_{q}^{x}, \Sigma_{q}^{x})}{\sum_{q=1}^{Q} N(x; \mu_{q}^{x}, \Sigma_{q}^{x})} \tag{2}
\]

Despite the average competitive behavior of the mixture, an unstable switching of components is still found in terms of the conditional probability at some regions of the signal. This is mainly due to a significant variability of the envelope features in some phonemes (mainly non-vowel cases) depending the phonetic context. Accordingly, since a phonetic segmentation is available, we replace the conditional probability \( p(q|x) \) by a binary flag based on the phonetic label of each frame (1 for the current phoneme, 0 otherwise) on the well-known expression of eq. 1 [4] to perform features conversion. The new conversion function is therefore expressed as shown in eq. 3. Then, since the membership of the components is not controlled anymore in a probabilistic way and the behavior of the mixture is forced to be fully-competitive (not as a real mixture), we refer to our strategy as phoneme-constrained Multi-Gaussian Modeling (MGM) rather than GMM.

\[
\hat{y} = \mu_{q(x)}^{y} + \sum_{x}^{p} (x - \mu_{q(x)}^{x}) \tag{3}
\]

Note that the full-competitive behavior of the proposed conversion function may lead to abrupt modifications of the spectral envelope at the phoneme boundaries over continuous speech. This may lead to perceived artifacts on the converted signal. However, we solve this by smoothing the components activation. We interpolated the components activation (and de-activation) over a duration of 10 frames centered at each phoneme border. This strategy is well adapted to our interests since the singing-voice material to convert corresponds to controlled recordings of phonetic sequences restricted to a small pitch range, stable phonemes transitions, and duration [1].

2.3. Performance comparison

We were interested to evaluate the performance of the proposed MGM and to compare it with the typical approach consisting of a GMM trained by Expectation-Maximization (EM). We considered two GMM models with 8 and 38 components with a full and diagonal configuration respectively. This choice is based on the interest in comparing our proposition with the typical configuration reported in Voice Conversion (full matrix, 8 components) as well as another using the same number of components resulted of the phoneme-constrained MGM in Japanese [38].

The envelope model corresponds to the one proposed by the authors in [2], consisting in an autoregressive model estimated on a mel-scaled interpolation of the harmonic peaks (mel-AR). This model shows improved envelope estimation accuracy compared to the traditional technique based on Linear Prediction (LPC). The envelope order was set to 50. Also, following the same work of [2], we restricted our experimental framework to the the use of singing-voice corresponding to recordings of similar pitch range (\( C_{3}, 130 Hz \)) for both source (female) and target (male) singers. The error measure corresponds to the spectral distortion (MSE) measured between the resulting converted spectra and the real target spectra (interpolation of the harmonic peaks) considering a mel-based scaling of the frequency axis.
The results of the conversion performance evaluation are shown in Table 1. The performance of the proposed phoneme-constrained modeling was found to be just slightly lower than the one based on EM (0.4 dB approx). These results were obtained using a number of 50,000 data vectors for the training. However, a generalization of the performance was already observed using only a bit more than 10,000 vectors and kept stable for increasing amounts. However, the MGM showed a smaller gap between the training and testing performances, denoting a favorable effect of the phoneme-constrained clustering in terms of a possible faster generalization of the source-target mapping. We also show in Fig. 3 the resulting conversion error measured at perceptual bands. Although the performance was found to be lower in some frequency regions, the performance of the MGM approach follows, in general, the one of the EM-based models. Finally, there was no perceived difference of the conversion effect on the converted utterances issued following both approaches. Moreover, the individual phoneme modeling of MGM reduced some average smoothing on the converted voice in terms of an increased perceived clearness at the phoneme transitions.

3. Non-parallel Timbre Conversion

3.1. Parallel training

In Voice Conversion, the training of the GMM is done using parallel corpus of the source and target speakers. Following the original GMM approach of [3], this condition allows a one-to-one correspondence between the components of the estimated source and target GMMs. Moreover, the joint source-target statistical modeling proposed in [4] allows a straightforward computation of the joint statistics term $\Sigma_{yx}$ required in the conversion function of eq. 1.

However, as it was already outlined at the introductory section, this condition restricts the target voice to those for which such specific training corpus is available, reducing significantly the possibilities of application of this technology whether it is applied to spoken or singing voice. Clearly, a main interest of this technology might be found on the reproduction of the voice of a person that is not available to record specific data. Thus, a wide universe of voices of interest, such as past celebrities or past singers, can hardly be considered using a parallel-training framework. Accordingly, we present in the next sections a methodology to perform timbre conversion using unpaired data.
3.2. The cross-covariance issue

There are two main restrictions for deriving eq. 1 by using unpaired data. Firstly, the classes (GMM components) of the source and target feature spaces are assumed to be in correspondence, in other words, to observe the same phonetic clustering of the timbre space. However, fitting individual statistical models this condition cannot be warranted, specially if using unpaired data. Secondly, the cross covariance term can only be estimated from paired data ($\Sigma_{yx} = E(Y|X)$).

In [5], it is assumed that the features of two speakers are related by a probabilistic linear transformation. The assumption is applied between the features of an unpaired case and the corresponding source and target of a paired one. Accordingly, an unpaired conversion function is obtained by adaptation of the one trained on paired data. The resulting non-parallel conversion performance will depend on the one obtained by simply applying the original parallel-trained model on the unpaired data.

The phoneme-constrained modeling presented in the previous section solves the class correspondence issue and it limits $\Sigma_{yx}$ to depend exclusively on the data of the corresponding phonetic class. Then, considering the linear relation directly between the source and target features this term can be approximated for each phonetic unit using the corresponding data. This term, commonly called transformation matrix after normalization by the source variance, represents a weighting of the deviation of the source input with respect to the mean value on the conversion function (eq. 1). The result of this weighting, added to the corresponding mean target value, represents the predicted target feature. After an exhaustive observation using the EM-based approach, it was found that the values kept by this matrix are typically small (closer to 0 than 1), producing a low variance of the predicted parameters. This results, accordingly, in poor dynamics on the converted features.

Therefore, we state that the actual contribution of this term on the value of the predicted features following the EM-based approach is poor. On the other hand, its effect is limited to impact the variance of the predicted parameters. In particular, for the case of a LSF parameterization (autoregressive-based envelope modeling), a manipulation of the variance of the parameters trajectories is translated to a smoothness control of the spectral formants. This represents a way to reduce the muflling effect commonly perceived on the converted signals [7]. Thus, we claim that the impact of using an approximation of $\Sigma_{yx}$ may not be critical on the conversion performance (mainly given by the target means). Moreover, the ability to control the variance of the converted features appears to be an efficient way to improve the quality of the converted utterances.

3.3. Non-parallel MGM-based features conversion

The assumption considered in [5] consists in relating the spectral envelope features of two speakers through a probabilistic linear transformation. By using phoneme-constrained modeling, we simplify that assumption considering it exclusively inside the source and target features belonging to the same phonetic class

\[ y = A_q(x) + b_q(x) \]  

where $D$ is the dimensionality of $x$ and $b_q$ is a vector of the same dimensionality as $x$. Then, considering this relation in the computation of $\Sigma_{yx}$ for each phonetic-component of the MGM we obtain

\[ \Sigma_{yx} = E(Y, X) = E[(y - \mu^y)(x - \mu^x)] \]  

(5)

(5)

\[ \Sigma_{yx} = E\{[(Ax + b) - (A\mu^x + b)](x - \mu^x)\} \]  

(6)

\[ = E[(Ax - \mu^x)^2] = A^2 \Sigma_{xx} \]  

(7)

Where the factor $A$ can be obtained in a similar way resolving for $\Sigma_{yy}$

\[ \Sigma_{yy} = E\{[(Ax + b) - (A\mu^x + b)]^2\} \]  

(8)

\[ = E[A^2(x - \mu^x)^2] = A^2 \Sigma_{xx} \]  

(9)

\[ A = \sqrt{\frac{\Sigma_{yy}}{\Sigma_{xx}}} \]  

(10)

Note that the approximation of the joint statistics does not depend on the bias vector $b$ and that the covariance matrices $\Sigma_{xx}$ and $\Sigma_{yy}$ are obtained separately from unpaired data when fitting individually the source and target MGMs.

The relation $y = Ax + b$, although it is done inside a phonetic class imposes a strong assumption on the features. Considering that the LSF parameters are mutually uncorrelated [8] and reinforcing this statement by using a diagonal matrix configuration, the resulting covariance region for the unidimensional case is clearly restricted to a narrow line. However, as the norm of $A$ decreases, the form of the covariance region becomes progressively an ellipse, finishing as a circle for the full-unrelated case ($A = 0$). Since it is rather observed an elliptoidal form on the most correlated dimensions on real paired-data and the orientation of the ellipse is exclusively defined by $\Sigma_{xx}$ and $\Sigma_{yy}$, the proposed $\Sigma_{yx}$ appears to be just un upper bound for $\Sigma_{yx}$ related to them. Accordingly, we apply a weighting factor $\alpha$ ($0 < \alpha < 1$) to $\Sigma_{yx}$ on the conversion function in order to impose a more realistic form on the approximated cross-covariance region. Then, based on eq. 3, the final expression for the conversion features will be as follows

\[ \tilde{y} = \mu^y_{q(x)} + \alpha \sqrt{\frac{\Sigma_{yy}}{\Sigma_{xx}}} (x - \mu^x_{q(x)}) \]

In Fig. 4 we compare one LSF dimension of one phoneme from real paired-data and the one generated by the resulting gaussian if trained by EM. Then, in Fig. 5, we compare the same real data and the one generated following the proposed approximation for several values of $\alpha$. Clearly, the approximated region following strictly the relation $y = Ax + b$ ($\alpha = 1$) is not representative of the data. However, it can be seen that for an $\alpha$ value around 0.75 the resulting region is similar to that obtained from EM. On the other hand, $\alpha$ values within the range $[0.5 - 0.7]$ offered the best perceptual results. In general, due to the effect already described as a variance regulator of the converted features it appears reasonable to set it manually. Nevertheless, for clarity, an objective evaluation will be presented in the results section. We note that for dimensions with a low correlation, the use of an approximation denoting a correlation higher than the one observed on the real data was found to be rather beneficial since it reduced the oversmoothing on the corresponding predictions, increasing the naturalness of the converted signals.

Finally, considering the pseudo-correspondance of the different LSF dimensions to a frequency region (especially true for
We aimed to evaluate objectively the performance of the proposed non-parallel strategy and to compare it with the traditional parallel one. Note however that paired data is required to compute the conversion error and that, therefore, an objective evaluation using real unpaired data is not straightforward. Accordingly, we used the same data sets described in the previous section in an "unpaired" way to evaluate the non-parallel approach and make feasible a straight comparison with the parallel one.

We centered our interest on comparing both strategies using a similar clustering capacity. Accordingly, we kept the same configuration of 38 components and diagonal covariance matrices for both GMM and MGM cases. Note that a consideration of diagonal matrices simplifies the computation of $A$ in eq. 10.

Finally, we note the interest in evaluating the performance in terms of the parameter $\alpha$ although it should be rather set manually following a subjective evaluation.

The results are shown in Fig. 6. We plot the resulting average spectral conversion error as a function of $\alpha$ of both training and test sets for the proposed conversion and we compare it with the one obtained with the test set of the traditional GMM one. The results show that the non-parallel performance is comparable to the parallel one. The resulting conversion error for the non-parallel approach was found just slightly higher (0.2 dB) at the best $\alpha$ interval. A similar conversion effect was also perceived on the converted speech following informal subjective tests. We note that the resulting best $\alpha$ range ($\alpha = [0.2-0.4]$) produces values for $\Sigma_X$ similar to those obtained through paired data. However, as we already outlined, the best quality on the converted signals is perceived when the variance of the converted features is slightly higher than the one providing the best objective results ($\alpha = [0.5-0.7]$).

We finally applied the non-parallel framework to real unpaired data. As the title says, the main objective of this work was to "resurrect" the voice of past singers through VOCALOID. Accordingly, following the same experimental framework just described, we used singing-voice material of the famous past Japanese actor and singer as target singer and a male VOCALOID database as source. The singing voice material corresponds to conventional recordings of three songs. Following, we built an "target-like" converted database by applying the resulting conversion model to all the recordings of the source singer.

We note that, for this case, the $\alpha$ value providing the best perceptual results was found lower than the previous case ($\alpha = 0.3$). This can be explained by the fact that the nature of the target data differs significantly from that of the source one, resulting in a reduction of the correlation between the source and target features if compared to the material used for the objective evaluation. This translates in a reduction of the ellipsoidal characteristic of the covariance region. Then, as shown in Fig. 5, a
lower $\alpha$ should be applied for a decreasing correlation.

The results obtained by producing musical sung utterances with VOCALOID using the converted database were as expected. In general, the timbre of the celebrity was successfully perceived on the synthesized singing-voice. The proposed methodology applied to the VOCALOID system allowed us to make a past singer “sing” again. Note however that the conversion of some other aspects of the voice quality (beyond the spectral envelope) strongly identifying some voices should be studied in the future in order to achieve a full perception of the target singer.

5. Conclusions
We presented a methodology to perform non-parallel timbre conversion and its application on the Singing-Voice synthesizer VOCALOID. The proposition is based in a phonetic-based modeling of the timbre-space of the singers and a linear assumption between the source and target features in order to derive a mapping function between timbre features from unpaired data. The success of the proposed methodology was verified through an objective evaluation using originally paired data and confirmed subjectively using real unpaired data from a past singer. Additional evaluations considering more singers will be conducted by the authors to exhaustively validate the performance and resulting converted Singing-Voice quality of the proposed non-parallel singing-voice conversion approach.

6. References