Voice Conversion with Smoothed GMM and MAP Adaptation

Yining Chen[1], Min Chu[2], Eric Chang[2], Jia Liu[1], and Runsheng Liu[1]*

[1]Department of Electric Engineering, Tsinghua University
Chenyining99@mails.tsinghua.edu.cn
[2]Microsoft Research Asia
5F, Sigma Center, No.49, Zhichun Road, Beijing 100080, P.R.C
{minchu,echang}@microsoft.com

Abstract

In most state-of-the-art voice conversion systems, speech quality of converted utterances is still unsatisfactory. In this paper, STRAIGHT analysis-synthesis framework is used to improve the quality. A smoothed GMM and MAP adaptation is proposed for spectrum conversion to avoid the overly smooth phenomenon in the traditional GMM method. Since frames are processed independently, the GMM based transformation function may generate discontinuous features. Therefore, a time domain low pass filter is applied on the transformation function during the conversion phase. The results of listening evaluations show that the quality of the speech converted by the proposed method is significantly better than that by the traditional GMM method. Meanwhile, speaker identifiability of the converted voice reaches 75%, even when the difference between the source speaker and the target speaker is not very large.

1. Introduction

Voice conversion technology, which converts one's voice to another's, makes it possible to provide various distinctive voices in data-driven text-to-speech systems [1], and provides speaker individuality in ultra low bit-rate communication systems such as the one based on speech recognition and speech synthesis [2].

Speech production is often represented by a source-filter model. Both parts of this model contribute to forming the speaker individuality. For example, speech rate, duration allocation, pitch and dynamic pitch range are features mainly related to the source, while formant positions and bandwidths are features related to the filter, i.e. the vocal tract. A perfect voice conversion should deal with all these features in phase. However, they are often processed separately to make the problem easier. Most current voice conversion systems focus on the spectral conversion and often apply simple adjustment for prosody features, such as shifting the pitch mean from a source speaker to a target speaker [1][3]. The same strategy is adopted in this paper.

The linear predictive framework has been widely adopted in many voice conversion systems. For example, piecewise linear conversion is performed on formant frequencies derived from poles of linear predictive model in [4] and linear spectral frequencies are used [1]. However, since the residual and linear predictive coefficients are not independent, modifying them separately will often degrade the quality of the reproduced speech [5]. Therefore, a high quality analysis-synthesis framework, STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrogram) [6] is utilized in this paper. In the STRAIGHT framework, voice quality stays high not only when the pitch is scaled up to as high as 600% above the original value but also when the STRAIGHT spectrum is coded with Mel-scale DCT (Discrete Cosine Transformation) as illustrated in Section 2.1.

There are many ways to implement the transformation function for converting source features to target ones, such as mapping codebooks [7], discrete transformation functions [4], artificial neural networks [8], Gaussian Mixture Models (GMM) [1] [9] [10] and some combinations of them [3]. In the mapping codebooks method, the spectrum parameters are first vector-quantized with the source codebook, and then are decoded with the mapping codebook[7]. When the mapping codebooks are replaced by other functions like piecewise linear function, the discrete transformation function method is obtained [4]. The discrete classification of source and target features results in discontinuity in the reproduced speech [9]. Artificial neural network is a well-known continuous and nonlinear transformation function, but its performance is often not as good as the GMM method [11]. The GMM method has been shown to be more efficient and robust than the others mentioned above [11]. However, the conventional GMM based conversion tends to generate overly smoothed utterances. A dynamic frequency warping approach has been proposed to reduce the over smoothness in [3]. In this paper, a GMM and MAP (Maximum a Posteriori) adaptation approach is proposed. In addition, the transformation function is smoothed along the time axis to maintain a continuous transformation in consecutive frames.

The new algorithm is described in Section 2. In Section 3, listening evaluations on the speech quality and speaker recognizability of the new method are introduced. Finally, conclusions are provided in Section 4.

2. Smoothed GMM and MAP adaptation algorithm for voice conversion

2.1. Features

STRAIGHT is a high quality analysis-synthesis framework for speech proposed by Kawahara et al. [5]. In the analysis phase, pitch-adaptive spectral analysis and time-frequency domain smoothing are used to obtain the smoothed spectrum, in which the effects of the original pitch are removed. In the synthesis phase, pitch synchronous minimum-phase impulse response overlap-add is used. Thus, in this framework, speech

[*] The work was carried out as a visiting student in Microsoft Research Asia

parameters such as pitch, vocal tract length and speech rate can be easily manipulated without significantly destroying the voice quality. To reduce the dimension of feature vectors, Mel-scale DCT is performed on the STRAIGHT log-spectrum and only the lowest 39 dimensions are kept. According to an informal listening test, the reconstructed utterances from the 39 dimension DCT features sound almost the same as those reproduced directly from the STRAIGHT spectrum.

2.2. Conventional GMM-based voice conversion

There are two mainstream methods of GMM conversion, the mean square error method [9] and Joint Density (JD) method [1]. Their performance is almost equivalent [1] . The JD method is described below as the baseline of our method.

The source speech is represented by an *n*-frame time-series $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, where $\mathbf{x}_i$ is a *d* dimensional feature vector for the *ith* frame, i.e. $\mathbf{x}_i = [x_1, x_2, \cdots, x_d]^T$. The target speech is represented by an *m*-frame time-series $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_m$, where $\mathbf{y}_j = [y_1, y_2, \cdots, y_d]^T$. DTW (Dynamic Time Warping) algorithm is then adopted to align source features to their counterparts in target series to obtain feature pair series $\mathbf{Z} = \mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_Q$ where $\mathbf{z}_k = [\mathbf{x_i}^T, \mathbf{y_j}^T]^T$.

The distribution of $\mathbf{z}$ is modeled by GMM as in Equation (1):

$$p_{GMM}(\mathbf{z}) = \sum_{l=1}^{L} c_l N(\mathbf{z}, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) = p(\mathbf{x}, \mathbf{y}) \qquad (1)$$

where c_l is the prior probabilities of $\mathbf{z}$, given the component *l*, and it satisfies $\sum_{l=1}^{L} c_l = 1$. $N(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the *2d* dimension normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The parameters $(c, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the joint density $p(\mathbf{x}, \mathbf{y})$ can be estimated using the well-known Expectation Maximization algorithm.

The transformation function that converts source feature $\mathbf{x}$ to target feature $\mathbf{y}$ is given by Equation (2)

$$\begin{aligned} F(\mathbf{x}) &= E(\mathbf{y} \mid \mathbf{x}) = \int \mathbf{y} p(\mathbf{y} \mid \mathbf{x}) d\mathbf{y} \\ &= \sum_{l=1}^{L} p_l(\mathbf{x}) \left(\boldsymbol{\mu}_l^{\mathrm{y}} + \boldsymbol{\Sigma}_l^{\mathrm{yx}} \left(\boldsymbol{\Sigma}_l^{\mathrm{xx}} \right)^{-1} (\mathbf{x} - \boldsymbol{\mu}_l^{\mathrm{x}}) \right) \end{aligned} \qquad (2)$$

$$p_l(\mathbf{x}) = \frac{c_l N(\mathbf{x}, \boldsymbol{\mu}_l^{\mathrm{x}}, \boldsymbol{\Sigma}_l)}{\sum_{k=1}^{L} c_k N(\mathbf{z}, \boldsymbol{\mu}_k^{\mathrm{x}}, \boldsymbol{\Sigma}_k)} \qquad (3)$$

where $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{\mathrm{x}} \\ \boldsymbol{\mu}^{\mathrm{y}} \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}^{\mathrm{xx}} & \boldsymbol{\Sigma}^{\mathrm{xy}} \\ \boldsymbol{\Sigma}^{\mathrm{yx}} & \boldsymbol{\Sigma}^{\mathrm{yy}} \end{bmatrix}$, $p_l(\mathbf{x})$ is the probability of $\mathbf{x}$ belonging to the *lth* component.

2.3. GMM and MAP adaptation algorithm

In conversional GMM based algorithm, acoustic features are converted from a source speaker to a target speaker by minimizing the mean squared error. Its main drawback is that the converted features are overly smooth and this makes the reconstructed speech unclear. Some researchers explained the phenomenon as the result of statistical average [3]. This paper would like to argue that since the correlation between features from two speakers is not linear, the correlation term in Equation (2) becomes very small. Consequently, the converted feature is mainly relied on the first item in Equation (2), i.e. most variations in feature $\mathbf{X}$ have been removed in the converted feature. Detailed explanations are given below.

The correlation item $\boldsymbol{\Sigma}_l^{\mathrm{yx}} \left(\boldsymbol{\Sigma}_l^{\mathrm{xx}} \right)^{-1}$ in Equation (2) is calculated from part of feature pairs in the training set and the values of its elements are shown with a grey-scale graph as in Figure 1, in which each small cell represents an element in the matrix $\boldsymbol{\Sigma}_l^{\mathrm{yx}} \left(\boldsymbol{\Sigma}_l^{\mathrm{xx}} \right)^{-1}$. The horizontal axis is for features from source speaker and the vertical axis is for target speaker. The darker a cell is, the larger the element is. The largest element is the one at position (1,1) whose value is 0.82. More than 90% of cells have values smaller than 0.1 and more than 40% of cells have values smaller than 0.01.

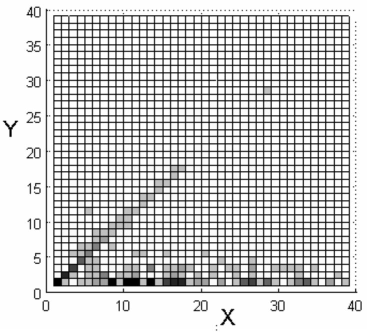

Figure 1. Correlation coefficient between features of two speakers.

The small $\boldsymbol{\Sigma}_l^{\mathrm{yx}} \left(\boldsymbol{\Sigma}_l^{\mathrm{xx}} \right)^{-1}$ reveals that the relationship between the two speakers is difficult to be modeled with simple linear transformations even though both speakers are females. Therefore, when Equation (2) is used to convert voices, the converted features are normally very close to the first item, $\sum_{l=1}^{L} p_l(\mathbf{x}) \boldsymbol{\mu}_l^{\mathrm{y}}$, in Equation (2), i.e., source features that belong to the same mixtures tend to be converted to the same target which is the weighted mean of the target GMM model. This means that the variance of the converted features is much smaller than that of the source features. This makes the reproduced speech sound unclear.

To avoid the over smoothness in converted speech, enough variances have to be maintained in the converted features. Since it is very difficult to map the variances of source features to those of target features, assuming them to be the same is an easy yet reasonable way for solving the problem. With this assumption, the transformation function in (2) is then modified to (4), i.e. the source features are shifted toward the target ones by the weighted distance between the mean vectors of corresponding mixtures in the source GMM and the target GMM respectively.

$$F(\mathbf{x}) = \mathbf{x} + \sum_{l=1}^{L} p_l(\mathbf{x}) \left(\boldsymbol{\mu}_l^{\mathrm{y}} - \boldsymbol{\mu}_l^{\mathrm{x}} \right) \qquad (4)$$

where, $p_l(\mathbf{X})$ is the probability of $\mathbf{X}$ belonging to the *lth* component, given in equation (3). $\boldsymbol{\mu}_l^x$ and $\boldsymbol{\mu}_l^y$ are the means of *lth* component of the source and target GMMs and they are estimated by a GMM and MAP adaptation approach.

When generating new voices for a corpus-driven text-to-speech system, there is normally a large speech corpus from the source speaker. Yet, only a limited number of paired utterances from both the source and target speakers are available. Therefore, the GMM of the source speaker is trained directly from the large speech corpus and the GMM of the target speaker is adapted from the source GMM with the MAP [12] method. Experiments show that only adapting the mean of GMM obtained best performance [12]. In this paper, we use a modified MAP method shown in Equation (5).

$$\boldsymbol{\mu}_l^{\mathbf{y}} = \frac{r}{r+\sum_{q=1}^{Q} p_l(\mathbf{x}_q)}\boldsymbol{\mu}_l^{\mathbf{x}} + \frac{\sum_{q=1}^{Q} p_l(\mathbf{x}_q)\mathbf{y}_q}{r+\sum_{q=1}^{Q} p_l(\mathbf{x}_q)} \qquad (5)$$

where r is a fixed factor [12], Q is the total frame number of adaptation data, and $\mathbf{x}_q$, $\mathbf{y}_q$ are the feature pairs aligned by DTW algorithm. Unlike traditional MAP, $p_l(\mathbf{x}_q)$ is adopted instead of $p_l(\mathbf{y}_q)$ to improve the accuracy of estimation.

2.4. Smoothed GMM and MAP adaptation

Another shortcoming of traditional GMM is that it doesn't take into account the correlation between frames. As we know, the continuity of speech is very important to subjective perception. GMM does not use any information about time correlation so that the conversion function may cause discontinuous. Listening tests show that there are some clicks in the converted speech. Therefore, we use smoothing methods to deal with the problem.

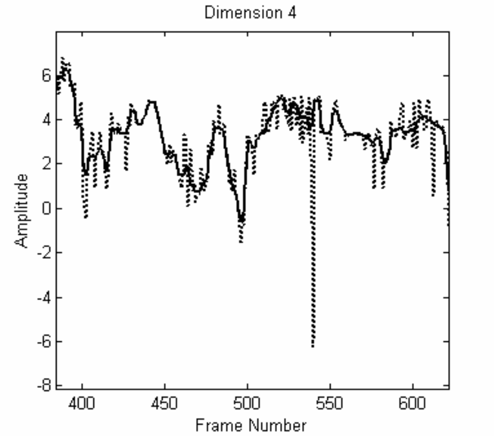

Figure 2. Discontinuity of conversion function.

Although it is hard to use the correlation directly, we know that if a discontinuous signal is added to a continuous signal, a discontinuous signal is obtained. In equation (4), only when $\sum_{l=1}^{L} \frac{c_l N(\mathbf{x}, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}{p_{GMM}(\mathbf{x})}\left(\boldsymbol{\mu}_l^{\mathbf{y}} - \boldsymbol{\mu}_l^{\mathbf{x}}\right)$ is continuous can we assure the continuity in the converted signal. The dotted line in Figure 2 shows the 4th dimension of the feature vectors in a segment of a sentence, from which we can see not only noise but also some unrepresentative points. So a median filter and a low pass filter are employed respectively for wiping off these points and smoothing the signal to get the smoothed offset illustrated by the solid line in Figure 2.

The training and conversion flowcharts of the smoothed GMM and MAP adaptation method are shown in Figure 3 and 4 respectively.

3. Evaluation experiments

Two subjective evaluation experiments are performed to investigate the performance of the proposed method. In the first experiment, voice quality of four types of converted speech is compared. In the second one, speaker identifiability is studied. In both experiments, source GMM is trained from 1000 utterances from the source speaker and the target GMM is adapted from the source one by 300 paired utterances from both speakers. The number of Gaussian mixtures is set to be 256. Both speakers are female, and speaker A has higher pitch than speaker B. Pitch means are adjusted accordingly.

3.1. Experiment on speech quality

Four methods are compared in this experiment. They are the JD GMM method, the GMM+MAP method, the smoothed GMM+MAP method and the Pitch Only (PO) method. The PO method refers to scaling the pitch range of the source speaker to that of the target speaker in the same way as it is done in the GMM+MAP method without any spectrum adjustment. It is provided to investigate how much benefit the spectrum conversion contributes to the voice quality.

Three groups of paired utterances, 10 pairs in each group, were generated for comparing the JD GMM method, GMM+MAP method and the PO method with the smoothed GMM+MAP method respectively. They were played randomly to 10 subjects, who were expected to choose the more natural one from each pair.

3.2. Experiment on speaker identifiability

ABX test was performed in this experiment, in which three utterances were played sequentially to subjects, who were expected to decide which of the previous two utterances (A and B) was closer to the third one (X). In 20 groups of stimulus, A and B were speech from the source and the target speakers that had gone through the STRAIGHT analysis-synthesis framework. X was converted speech with the smoothed GMM+MAP method, either A to B or B to A. In the other 10 groups, A and B were generated with the smoothed GMM+MAP method and the PO method respectively and X were from the target speaker. In both experiments, all test utterances were not included in the training set.

3.3. Results and analysis

The results of the two experiments are given in Figure 5 and Figure 6 respectively. From Figure 5, it is obvious that voice quality of the speech generated by our smoothed GMM+MAP method is significantly better than those generated by JD GMM and slightly worse than that by PO. From Figure 6, the converted speech sounds more like the target speaker (with the probability of 75%). Furthermore, the spectrum conversion carried out in the smoothed GMM+MAP method makes the reconstructed speech closer to the target speaker's individuality than the PO method, although the former causes the quality to drop somewhat. [+]

[+] Samples are available in http://research.microsoft.com/echang/projects/voice_conversion/Eurospeech2003_voice_conversion.htm

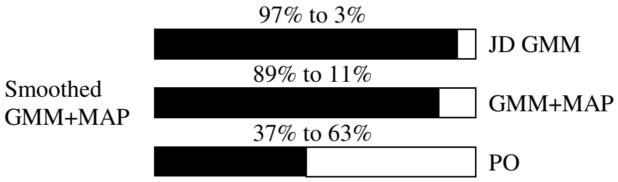

Figure 5. Results for the quality evaluation.

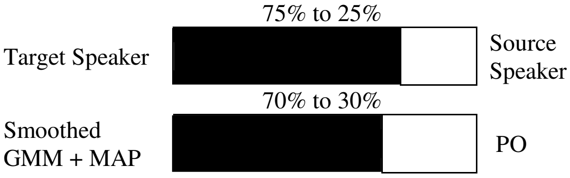

Figure 6. Results for the speaker identification evaluation.

4. Discussion

According to the quality evaluation, the proposed smoothed GMM and MAP adaptation method has been shown to generate speech with higher quality than those reconstructed with the traditional JD GMM method. The subjective speaker identification evaluation confirms that the proposed method produces more similar speech to the target speaker than the solution that only pitch is adjusted. All results show that the smoothed GMM and MAP adaptation method works well even though the difference between the source speaker and the target speaker is not as large as that between a male and a female.

The benefit results from the following factors. First, the Mel-Scale DCT of STRAIGHT spectrum is a reliable feature for reconstructing speech. Second, in the new transformation function, the dynamic changes in source features are copied statistically into the converted features so that the over smoothness problem is overcome. Third, since the source GMM is trained with ‘enough’ data and the target GMM is adapted with paired-vectors, the obtained transformation function is more precise and robust. Fourth, the low pass filter on the transformation function along time axis removes the discontinuities in the converted features. 300 paired utterances are used for training the target GMM, however, the method works well even if less data are available.

The main drawback of the proposed method is that the STRAIGHT analysis is pitch-sensitive. Large errors in pitch estimation will significantly reduce the voice quality of the converted speech.

5. Acknowledgements

The authors would like to thank all the members of the speech group in Microsoft Research Asia for their helps in attending the evaluation tests.

6. References

[1] Kain, A., and Macon, M.W., “Spectral voice conversion for Text-to-Speech synthesis”, in *Proc. of ICASSP,* 1998, pp. 285-299.

[2] Lee, K-S., and Richard V.C, “A very low bit rate speech coder based on a recognition/synthesis paradigm”, *IEEE Trans. Speech and Audio Proc.*, 9(5), 2001, pp. 482-491.

[3] Toda, T., Saruwatari, H., and Shikano, K., “Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum”, in *Proc. of ICASSP,* 2001, pp. 841-944.

[4] Mizuno, H., and Abe, M., “Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt”, *Speech Communication,* vol. 16, 1995, pp. 153-164.

[5] Syrdal, A., Stylianou, Y., Garrison, L., Conkie, A., and Schroeter, J., “TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis”, in *Proc. of ICASSP,* 1998, pp. 273-276.

[6] Kawahara, H., Masuda-katsuse, I., and De Cheveign, A., “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds”, *Speech Communication,* vol. 27, 1999, pp. 187-207.

[7] Abe, M., Nakanura, S., Shikano, K., and Kuwabara, H., “Voice conversion through vector quantization”, in *Proc. of ICASSP,* 1998, pp. 655-658.

[8] Narendranath, M., Murthy, H.A., Rajendran, S., and Yegnanarayana, B., “Transformation of formants for voice conversion using artificial neural networks”, *Speech Communication,* vol. 16, 1995, pp. 207-216.

[9] Stylianou, Y., Cappe, Y., and Moulines, E., “Continuous probabilistic transform for voice conversion”, *IEEE Trans. Speech and Audio Proc.,* Vol. 6, 1998, pp. 131-142.

[10] Marshimo, M., Toda, H., and Shikano, K., Campbell, N., “Evaluation of cross-language voice conversion based on GMM and STRAIGHT”, in *Proc. of Eurospeech,* 2001, pp. 361-364.

[11] Baudoin, G., and Stylianou, Y., “On the transformation of speech spectrum for voice conversion”, in *Proc. of ICSLP,* 1996, pp. 1405-1408.

[12] Reynolds, D., Quatieri, T., and Dunn.R., “Speaker Verification Using Adapted Gaussian Mixture Models”, *Digital Signal Processing,* vol. 10, 2000, pp. 19-41.

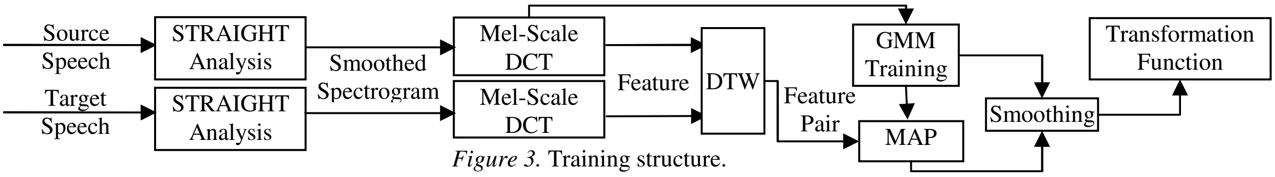

Figure 3. Training structure.

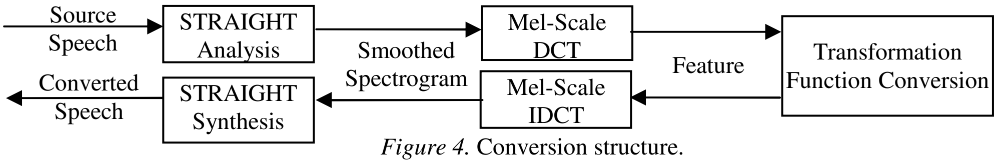

Figure 4. Conversion structure.