VOICE CONVERSION ALGORITHM BASED ON GAUSSIAN MIXTURE MODEL WITH DYNAMIC FREQUENCY WARPING OF STRAIGHT SPECTRUM

Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan

ABSTRACT

In the voice conversion algorithm based on the Gaussian Mixture Model (GMM) applied to STRAIGHT, quality of converted speech is degraded because the converted spectrum is exceedingly smoothed. In this paper, we propose the GMM-based algorithm with dynamic frequency warping to avoid the over-smoothing. We also propose an addition of the weighted residual spectrum, which is the difference between the GMM-based converted spectrum and the frequency-warped spectrum, to avoid the deterioration of conversion-accuracy on speaker individuality. Results of the evaluation experiments clarify that the converted speech quality is better than that of the GMM-based algorithm, and the conversion-accuracy on speaker individuality is the same as that of the GMM-based algorithm in the proposed method with the properly-weighted residual spectrum.

1. INTRODUCTION

Voice conversion is a technique used to convert one speaker's voice into another speaker's voice [1]. In general, speech databases from many speakers must be required to synthesize speech of various speakers. However, if a high quality voice conversion algorithm is realized, speech of various speakers can be synthesized even with a speech database of a single speaker.

As the voice conversion algorithm which can represent the acoustic space of a speaker continuously, the algorithm based on the Gaussian Mixture Model (GMM) has also been proposed by Stylianou et al. [2][3]. In this GMM-based algorithm, the acoustic space is modeled by the GMM without the use of vector quantization, and acoustic features are converted from a source speaker to a target speaker by the mapping function based on the feature-parameter correlation between two speakers.

Since voice conversion is usually performed with an analysis-synthesis method, quality of an analysis-synthesis method is also important to realize a high quality voice conversion algorithm. As a high quality analysis-synthesis method, STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) has been proposed by Kawahara et al., which is a high quality vocoder type algorithm [4][5].

In the GMM-based voice conversion algorithm applied to STRAIGHT, quality of converted speech is degraded because the converted spectrum is exceedingly smoothed. In this paper, we newly propose the GMM-based algorithm with dynamic frequency warping to avoid the over-smoothing. We also propose an addition of the weighted residual spectrum which is the difference between the GMM-based converted spectrum and the frequency-warped spectrum, to avoid the deterioration of conversion-accuracy on speaker individuality.

2. STRAIGHT

STRAIGHT is a high quality analysis-synthesis method, which uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region in order to remove signal periodicity[4][5]. This method extracts F0 (fundamental frequency) by using TEMPO (Time-domain Excitation extractor using Minimum Perturbation Operator), and designs excitation source based on phase manipulation[4][5].

STRAIGHT can manipulate such speech parameters as pitch, vocal tract length, and speaking rate, while maintaining high reproductive quality.

3. GMM-BASED VOICE CONVERSION ALGORITHM AND ITS SHORTCOMING

3.1. GMM-based Voice Conversion Algorithm

We assume that p-dimensional time-aligned acoustic features $\boldsymbol{x}\{[x_0,\ x_1,\ \ldots,\ x_{p-1}]^{\mathrm{T}}\}$ (source speaker's) and $\boldsymbol{y}\{[y_0,\ y_1,\ \ldots,\ y_{p-1}]^{\mathrm{T}}\}$ (target speaker's) are determined by Dynamic Time Warping (DTW), where T denotes transposition of the vector.

In the GMM algorithm, the probability distribution of acoustic features $\boldsymbol{x}$ can be written as

$$p(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i N(\boldsymbol{x}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}), \quad \sum_{i=1}^{m} \alpha_i = 1, \quad \alpha_i \geq 0, \tag{1}$$

where $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. α_i denotes a weight of class i, and m denotes the total number of the Gaussian mixtures.

The mapping function [2][3] converting acoustic features of the source speaker to those of the target speaker is given by

$$\begin{aligned} F(\boldsymbol{x}) &= E[\boldsymbol{y}|\boldsymbol{x}] \\ &= \sum_{i=1}^{m} h_i(\boldsymbol{x})[\boldsymbol{\mu}_i^{(y)} + \boldsymbol{\Sigma}_i^{(yx)} \left(\boldsymbol{\Sigma}_i^{(xx)}\right)^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i^{(x)})], \end{aligned} \tag{2}$$

$$h_i(\boldsymbol{x}) = \frac{\alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i^{(x)}, \boldsymbol{\Sigma}_i^{(xx)})}{\sum_{j=1}^{m} \alpha_j N(\boldsymbol{x}; \boldsymbol{\mu}_j^{(x)}, \boldsymbol{\Sigma}_j^{(xx)})}, \tag{3}$$

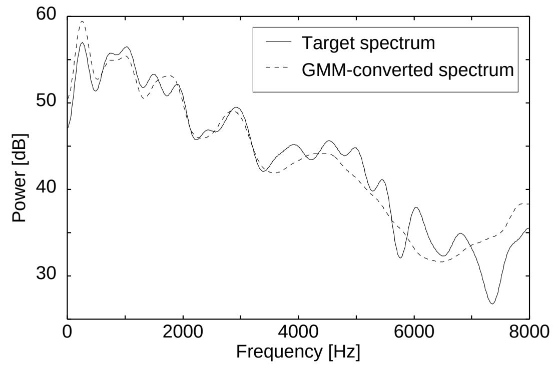

Figure 1: Spectrum converted by the GMM-based voice conversion algorithm ("GMM-converted spectrum") and spectrum of the target speaker ("Target spectrum").

where $\boldsymbol{\mu}_i^{(x)}$ and $\boldsymbol{\mu}_i^{(y)}$ denote the mean vectors of class i for the source and target speakers. $\boldsymbol{\Sigma}_i^{(xx)}$ denotes the covariance matrix of class i for the source speaker. $\boldsymbol{\Sigma}_i^{(yx)}$ denotes the cross-covariance matrix of class i for the source and target speakers. In this paper, we assume that these matrices are diagonal.

In order to estimate parameters (α_i, $\boldsymbol{\mu}_i^{(x)}$, $\boldsymbol{\mu}_i^{(y)}$, $\boldsymbol{\Sigma}_i^{(xx)}$, $\boldsymbol{\Sigma}_i^{(yx)}$), the probability distribution of the joint vectors $\boldsymbol{z} = [\boldsymbol{x}^{\mathrm{T}}, \boldsymbol{y}^{\mathrm{T}}]^{\mathrm{T}}$ for the source and target speakers is represented by the GMM [6]. These parameters are estimated by the EM algorithm.

3.2. Application of GMM-based Algorithm to STRAIGHT

The cepstrum of the smoothed spectrum analyzed by STRAIGHT is used as an acoustic feature. In this paper, the cepstrum order is 40 (the quefrency is 2.5 ms, and the sampling frequency is 16000 Hz). In order to perform voice conversion, the 1 to 40-th order cepstrum coefficients are converted, and the 0-th order cepstrum coefficient, which corresponds the signal power, is kept as the value of the source speaker.

3.3. Shortcoming of GMM-based Algorithm

In the GMM-based algorithm applied to STRAIGHT, quality of converted speech is degraded because the converted spectrum is exceedingly smoothed by the statistical averaging operation. Figure 1 shows the example of the GMM-based converted spectrum and the spectrum of the target speaker. As shown in this figure, the over-smoothing exists on the GMM-based converted spectrum.

4. GMM-BASED VOICE CONVERSION ALGORITHM WITH DYNAMIC FREQUENCY WARPING OF STRAIGHT SPECTRUM

In this paper, we propose the GMM-based algorithm with dynamic frequency warping to avoid the over-smoothing. An overview of the proposed algorithm is shown in Figure 2.

4.1. Dynamic Frequency Warping

In order to avoid the over-smoothing of the converted spectrum, spectral conversion is performed with the dynamic frequency warping [7][8]. In this technique, the correspondence between the original frequency axis and the converted frequency axis is represented by the warping function. This function is calculated as the path which minimized the normalized spectrum distance between the STRAIGHT log spectrum of the source speaker and the GMM-based converted log spectrum.

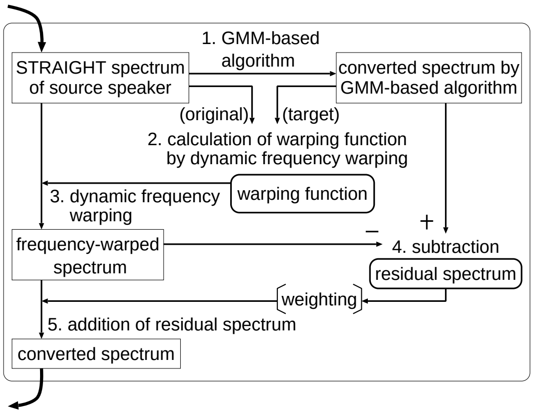
STRAIGHT spectrum of source speaker
1. GMM-based algorithm
converted spectrum by GMM-based algorithm
(original)
(target)
2. calculation of warping function by dynamic frequency warping
3. dynamic frequency warping
warping function
frequency-warped spectrum
−
+
4. subtraction
residual spectrum
weighting
5. addition of residual spectrum
converted spectrum

Figure 2: GMM-based voice conversion algorithm with dynamic frequency warping.

4.2. Conversion of Spectral Power

Conversion-accuracy on speaker individuality with the dynamic frequency warping is worse than that of the GMM-based algorithm because the spectral power cannot be converted. To convert the spectral power, we newly propose the technique to add the weighted residual spectrum which is the difference between the GMM-based converted log spectrum and the dynamic-frequency-warped log spectrum. By using this technique, we can recover the conversion-accuracy on speaker individuality. In the proposed algorithm, the converted spectrum $S_c(f)$ is written as

$$|S_c(f)| = \exp[\ln|S_d(f)| + w(\ln|S_g(f)| - \ln|S_d(f)|)], \quad (4)$$

$$0 \le w \le 1, \quad (5)$$

where $S_d(f)$ and $S_g(f)$ denote the dynamic-frequency-warped spectrum and the GMM-based converted spectrum respectively. Also, w denotes the weight for a residual spectrum. The variations of converted spectra which correspond to the different weights for a residual spectrum are shown in Figure 3.

In this paper, evaluation experiments are performed to investigate effects by the weight for a residual spectrum. In the experiments, we used not only the weights of the constant value but also the frequency-variant weights which change on each frequency as follows

$$w_h(f) = \begin{cases} \dfrac{2}{f_s} f & 0 \le f < \dfrac{f_s}{2} \\ -\dfrac{2}{f_s} f + 2 & \dfrac{f_s}{2} \le f < f_s \end{cases}, \quad (6)$$

$$w_l(f) = \begin{cases} -\dfrac{2}{f_s} f + 1 & 0 \le f < \dfrac{f_s}{2} \\ \dfrac{2}{f_s} f - 1 & \dfrac{f_s}{2} \le f < f_s \end{cases}, \quad (7)$$

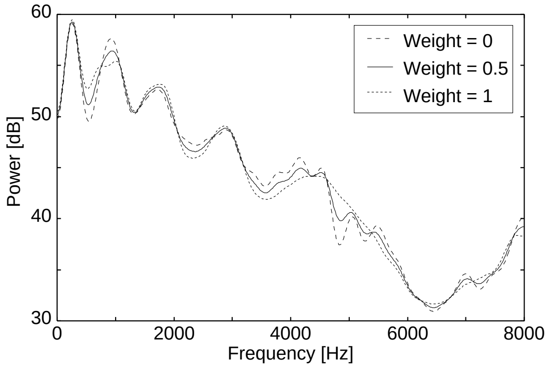

Figure 3: Variations of converted spectra which correspond to the different weights for a residual spectrum.

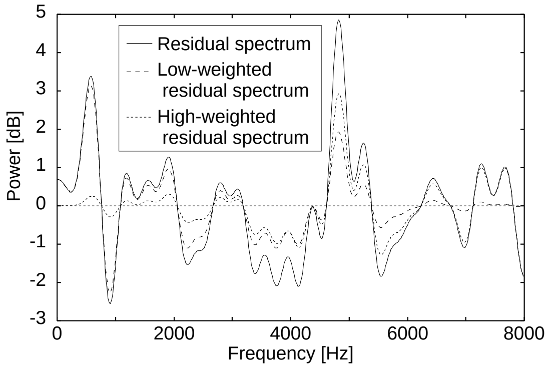

Figure 4: Residual spectra weighted by the frequency-variant weights which increase as the frequency is high ("High-weighted residual spectrum") and the frequency is low ("Low-weighted residual spectrum").

where f_s denotes the sampling frequency. The residual spectra weighted by those frequency-variant weights are shown in Figure 4. For example, if we use the weight $w_h(f)$ which increase as the frequency is high, the converted spectrum is more close to the GMM-based converted spectrum in the high-frequency regions.

5. EVALUATION EXPERIMENTS

In order to evaluate the performance of the GMM-based algorithm with dynamic frequency warping, we performed experiments on speech quality and speaker individuality. We also investigated effects by the weight for a residual spectrum. The number of Gaussian mixtures was set to be 64, and the amount of training data was set to be 58 sentences. The male-to-male and female-to-female voice conversion were performed in each experiment.

As for the source information, the average of log-scaled fundamental frequencies of the source speaker was converted to that of the target speaker. The prosodic dynamic characteristics between two speakers were not considered.

5.1. Evaluation Experiment on Speech Quality

In order to evaluate the quality of the converted speech by the proposed algorithm, the subjective evaluation experiment was performed. Eight listeners participated in the experiment. An opinion score for evaluation was set to be a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Three sentences which were not included in the training data were used to evaluate.

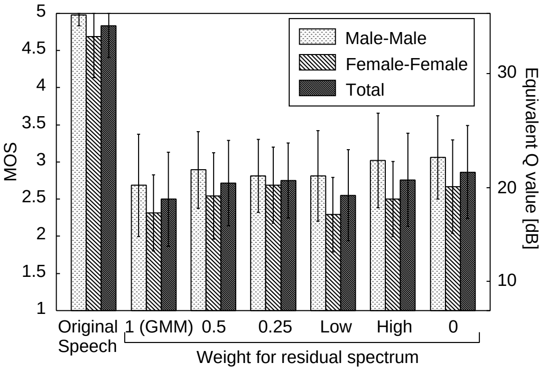

Figure 5: Relation between the weight for a residual spectrum and speech quality.

The experimental result is shown in Figure 5. Errorbars denote standard deviations. The converted speech quality by the proposed algorithm is better than that of the GMM-based algorithm (the weight is 1). About the weight for a residual spectrum, the converted speech quality without a residual spectrum (the weight is 0) is best. Also, the converted speech quality with the weight which increase as the frequency is high ("High") is better than that of the weight which increase as the frequency is low ("Low"). When we use the weight "Low", the converted spectrum is smoothed exceedingly in the low-frequency regions. As this result, it is considered that the speech quality is degraded by the over-smoothing of the converted spectrum in the low-frequency regions.

5.2. Evaluation Experiments on Speaker Individuality

5.2.1. Objective Evaluation Experiment

In order to evaluate the conversion-accuracy on speaker individuality of the proposed algorithm, the objective evaluation experiment was performed by the cepstrum distortion (CD) between the converted speech and the target speech. Ten sentences which were not included in the training data were used to evaluate.

The experimental result is shown in Figure 6. CDs by the proposed algorithm is worse than that of the GMM-based algorithm (the weight is 1). About the weight for a residual spectrum, CDs increase as the weight is more close to 0. When we use the weight which increase as the frequency is high ("High-weighted"), the deterioration of CD is the same as that of using the weight which is 0.5, and the converted speech quality ("High") is better than that of using the weight which is 0.5 as shown in Figure 5.

5.2.2. Subjective Evaluation Experiment

In order to evaluate the conversion-accuracy on speaker individuality of the proposed algorithm, the subjective evaluation experiment (ABX test) was performed. Eight listeners participated in the experiment.

In the ABX test, A and B were the source and the target speaker's speech, and X was either the converted speech as

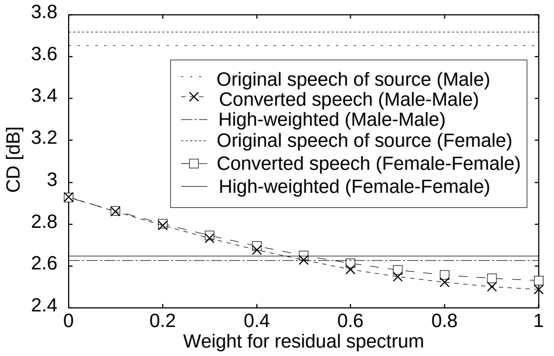

Figure 6: Relation between the weight for a residual spectrum and CD: Cepstrum Distortion.

follow,

- converted speech by the GMM-based algorithm ···“GMM”,
- converted speech by the proposed algorithm without a residual spectrum···“0-weighted”,
- converted speech by the proposed algorithm with the weight which increases as the frequency is high···“High-weighted”,
- synthesized speech by converting of the average log-scaled F0···“F0 only”,
- synthesized speech by converting of the average log-scaled F0 and replacing the source speaker’s spectra with those of the target speaker···“F0 & spectrum”.

“F0 & spectrum” was used to evaluate the conversion-accuracy on speaker individuality when conversion of spectra was perfect. Listeners were asked to select either A or B as being most similar to X. Two sentences which were not included in the training data were used to evaluate.

The experimental result is shown in Figure 7. The conversion-accuracy on speaker individuality of the proposed algorithm without a residual spectrum (“0-weighted”) is worse than that of the GMM-based algorithm (“GMM”). However, we can recover the conversion-accuracy on speaker individuality by using the weight which increases as the frequency is high (“High-weighted”). In order to compare these two algorithms (“GMM” and “High-weighted”), we also performed another subjective experiment (preference test) on speaker individuality. The result clarifies that the conversion-accuracy on speaker individuality of the proposed algorithm with the weight which increases as the frequency is high is the same as that of the GMM-based algorithm.

As shown in Figure 7, the conversion-accuracy on speaker individuality of only F0 conversion (“F0 only”) is insufficient, and it can be improved by converting spectra.

6. CONCLUSION

In this paper, we propose the voice conversion algorithm based on the Gaussian Mixture Model (GMM) with dynamic frequency warping of STRAIGHT spectrum, and evaluate this conversion algorithm. We performed evaluation experiments on speech quality and speaker individuality, compared with the GMM-based algorithm. As the results, it is clarified that the converted speech quality is better than that of the GMM-based algorithm, and the conversion-accuracy on speaker individuality is the same as that of the GMM-based algorithm in the proposed method with the properly-weighted residual spectrum.

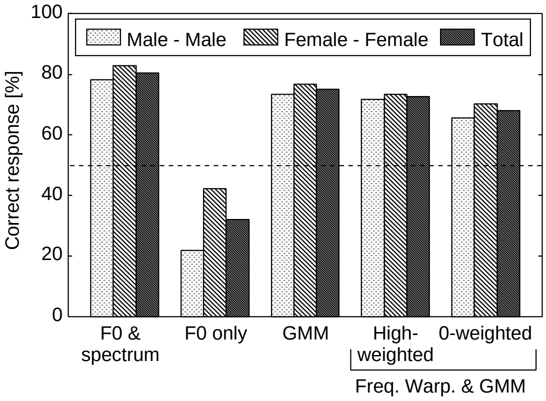

Figure 7: Correct response for speaker individuality.

7. ACKNOWLEDGEMENT

This work was partly supported by CREST (Core Research for Evolutional Science and Technology) in Japan.

8. REFERENCES

[1] H. Kuwabara, and Y. Sagisaka, “Acoustic characteristics of speaker individuality: control and conversion,” Speech Communication, vol. 16, no. 2, pp. 165–173, 1995.

[2] Y. Stylianou, O. Cappé, and E. Moulines, “Statistical methods for voice quality transformation,” Proc. EUROSPEECH, Madrid, Spain, pp. 447–450, Sept. 1995.

[3] Y. Stylianou, and O. Cappé, “A system voice conversion based on probabilistic classification and a harmonic plus noise model,” Proc. ICASSP, Seattle, U.S.A., pp. 281–284, May 1998.

[4] H. Kawahara, “Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited,” Proc. ICASSP, Munich, Germany, pp. 1303–1306, Apr. 1997.

[5] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “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, no. 3–4, pp. 187–207, 1999.

[6] A. Kain, and M.W. Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. ICASSP, Seattle, U.S.A., pp. 285–288, May 1998.

[7] H. Valbret, E. Moulines, and J.P. Tubach, “Voice transformation using PSOLA technique,” Proc. ICASSP, San Francisco, U.S.A., pp. 145–148, Mar. 1992.

[8] N. Maeda, H. Banno, S. Kajita, K. Takeda, and F. Itakura, “Speaker conversion through non-linear frequency warping of STRAIGHT spectrum,” Proc. EUROSPEECH, Budapest, Hungary, pp. 827–830, Sept. 1999.