HIGH QUALITY VOICE CONVERSION BASED ON GAUSSIAN MIXTURE MODEL WITH DYNAMIC FREQUENCY WARPING

Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science
Nara Institute of Science and Technology, Japan
tomoki-t@is.aist-nara.ac.jp

Abstract

In the voice conversion algorithm based on the Gaussian Mixture Model (GMM), quality of the converted speech is degraded because the converted spectrum is exceedingly smoothed. In this paper, we newly propose the GMM-based algorithm with the Dynamic Frequency Warping (DFW) to avoid the over-smoothing. We also propose that the converted spectrum is calculated by mixing the GMM-based converted spectrum and the DFW-based converted 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 algorithm with the proper weight for mixing spectra.

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.

Since voice conversion is usually performed with an analysis-synthesis method, quality of an analysis-synthesis method is 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 [2][3].

As the voice conversion algorithm that can represent the acoustic space of a speaker continuously, the algorithm based on the Gaussian Mixture Model (GMM) has been also proposed by Stylianou et al. [4][5]. 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.

In the GMM-based voice conversion algorithm applied to STRAIGHT[6], quality of the 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 ("GMM-converted spectrum") and the spectrum of the target speaker ("Target spectrum"). As shown in this figure, the over-smoothing exists on the GMM-based converted spectrum.

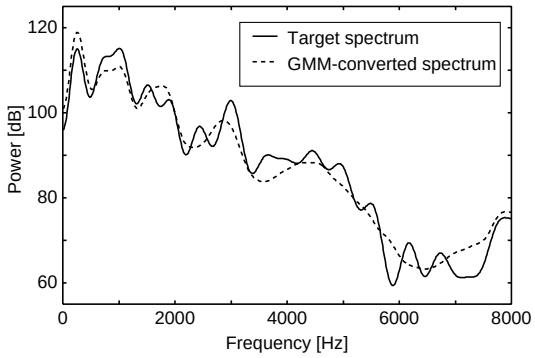

Figure 1: Spectrum converted by the GMM-based voice conversion algorithm and spectrum of the target speaker.

In this paper, we newly propose the GMM-based algorithm with the Dynamic Frequency Warping (DFW) to avoid the over-smoothing. However, conversion-accuracy on speaker individuality with the DFW is a little worse than that of the GMM-based algorithm because the spectral power cannot be converted. So, we also propose that the converted spectrum is calculated by mixing the GMM-based converted spectrum and the DFW-based converted 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[2][3]. 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[2][3].

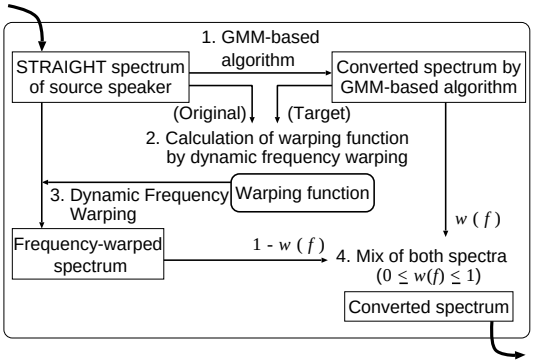

Figure 2: GMM-based voice conversion algorithm with the Dynamic Frequency Warping.

3. GMM-based voice conversion algorithm with Dynamic Frequency Warping

In this paper, we propose the GMM-based algorithm with the Dynamic Frequency Warping (DFW) to avoid the over-smoothing. In this algorithm, the converted spectra are calculated by mixing the GMM-based converted spectra and the DFW-based converted spectra. An overview of the proposed algorithm is shown in Figure 2.

3.1. DFW

In order to avoid the over-smoothing of the converted spectrum, the spectral conversion is performed with the DFW[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 that minimized the normalized mel-spectrum distance between the STRAIGHT logarithmic mel-spectrum of the source speaker and the GMM-based converted logarithmic mel-spectrum. In the GMM-based algorithm, the mel-cepstrum of the smoothed spectrum analyzed by STRAIGHT is used as an acoustic feature. In this paper, the mel-cepstrum order is 40, and the covariance matrix is diagonal.

3.2. Mix of converted spectra

Conversion-accuracy on speaker individuality with the DFW is a little worse than that of the GMM-based algorithm because the spectral power cannot be converted. So, we also propose that the converted spectrum is calculated by mixing the GMM-based converted spectrum and the DFW-based converted spectrum to avoid the deterioration of conversion-accuracy on speaker individuality. In the proposed algorithm, the converted spectrum $S_{\mathrm{c}}(f)$ is written as

$$|S_{\mathrm{c}}(f)| = \exp[w(f)\ln|S_{\mathrm{g}}(f)| + \{1-w(f)\}\ln|S_{\mathrm{d}}(f)|], \quad \text{subject to} \quad 0 \le w(f) \le 1, \tag{1}$$

where $S_{\mathrm{d}}(f)$ and $S_{\mathrm{g}}(f)$ denote the DFW-based converted spectrum and the GMM-based converted spectrum, respectively. Also, $w(f)$ denotes the weight for mixing spectra. As the mixing-weight is closer to 1, the converted spectrum is more close to the GMM-based converted spectrum.

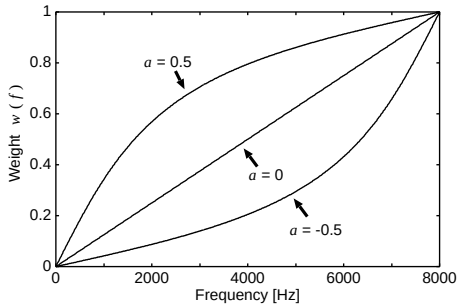

Figure 3: Variations of mixing-weights which correspond to the different parameters a.

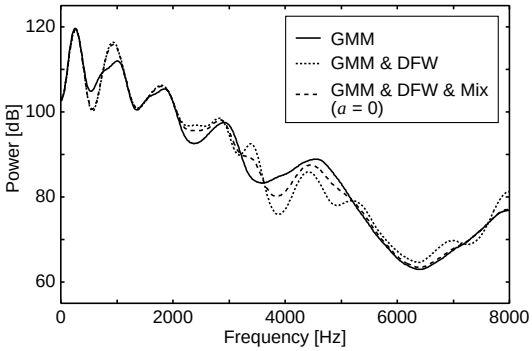

Figure 4: Spectra converted by the GMM-based algorithm, the proposed algorithm without the mix of the converted spectra, and the proposed algorithm with the mix of the converted spectra.

Results of preliminary experiments clarified that quality of the converted speech is degraded considerably when a spectrum is exceedingly smoothed in the low-frequency regions. So, we use the mixing-weight as follows

$$w(f) = \left| \frac{2\pi f}{f_s} + 2\tan^{-1}\left(\frac{a\sin(2\pi f/f_s)}{1 - a\cos(2\pi f/f_s)} \right) \right| \Big/ \pi, \quad \text{subject to} \quad -1 < a < 1, \quad -f_s/2 \le f \le f_s/2, \tag{2}$$

where f_s denotes the sampling frequency, and a denotes the parameter which change the mixing-weight. Figure 3 shows the variations of the mixing-weights which correspond to the different parameters a when the sampling frequency is 16 kHz. Figure 4 shows the example of the GMM-based converted spectrum ("GMM"), the DFW-based converted spectrum ("GMM & DFW"), and the converted spectrum calculated by mixing the converted spectra when the parameter a is set to be 0 ("GMM & DFW & Mix").

4. Experiments using various mixing-weights

Evaluation experiments were performed to investigate effects by the mixing-weight. We performed subjective evaluation experiments on speech quality and speaker individuality. The number of Gaussian mixtures was set to

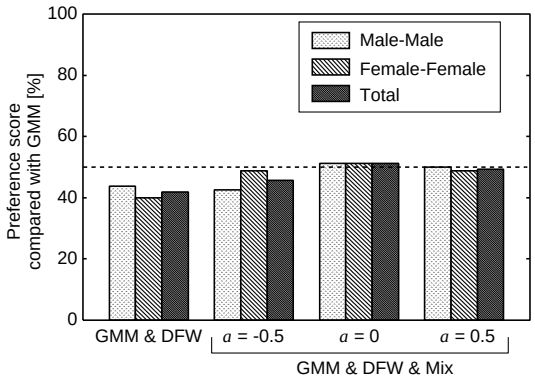

Figure 5: Relation between the conversion-accuracy on speaker individuality and the parameter a of the mixing-weight.

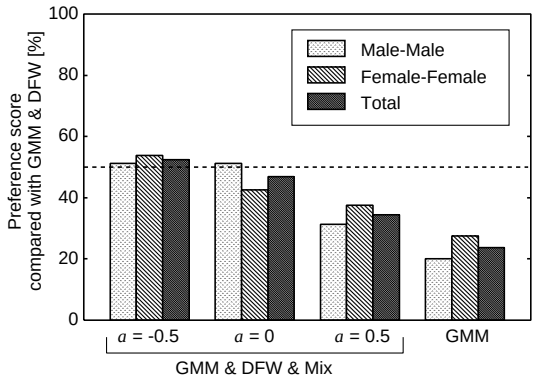

Figure 6: Relation between the converted speech quality and the parameter a of the mixing-weight.

be 64, and the amount of training data was set to be 58 sentences. The total duration of this data is about 4 or 5 minutes. The male-to-male and female-to-female voice conversions were performed and 10 listeners participated in each experiment.

As for the source information, the average of log-scaled F0 of the source speaker was converted to that of the target speaker. The prosodic dynamic characteristics between two speakers were not considered. The converted spectrum of the proposed algorithm was liftered by the 40-th order mel-cepstrum, in the same way as the GMM-based converted spectrum.

4.1. Preference test on speaker individuality

In order to evaluate the relation between the conversion-accuracy on speaker individuality and the parameter a of the mixing-weight, the preference (XAB) test was performed. Two sentences that were not included in the training data were used to evaluate. In the XAB test, X was the synthesized speech by converting of the average log-scaled F0 and replacing the source speaker's spectra with those of the target speaker (this means the perfect spectral conversion). A and B were the converted speech. Listeners were asked to select either A or B as being most similar to X.

The experimental result is shown in Figure 5. The conversion-accuracy on speaker individuality of the proposed algorithm without the mix of the converted spectra ("GMM & DFW") is a little worse than that of the GMM-based algorithm. However, the conversion-accuracy is improved by mixing the converted spectra.

4.2. Preference test on speech quality

In order to evaluate the relation between quality of the converted speech and the parameter a of the mixing-weight, the preference test was performed. Four sentences that were not included in the training data were used to evaluate. Listeners were asked to select either the converted speech as having better speech quality.

The experimental result is shown in Figure 6. The converted speech quality of the proposed algorithm with the mix of the converted spectra is better than that of the GMM-based algorithm. The converted speech quality of the proposed algorithm with the mix of the converted spectra (the parameters a are set to be 0 and -0.5 in the male-to-male and female-to-female voice conversions) is the same as that of the proposed algorithm without the mix of the converted spectra, and the conversion-accuracy on speaker individuality is the same as that of the GMM-based algorithm as shown in Figure 5.

5. Comparison with conventional method

In order to evaluate the performance of the GMM-based algorithm with the DFW, we performed subjective evaluation experiments on speech quality and speaker individuality. The experimental conditions are the same as those of the previous section. The parameters a were set to be 0 and -0.5 in the male-to-male and female-to-female voice conversions.

As for the source information, a log-scaled F0 of the source speaker was converted to that of the target speaker by using GMM-based algorithm. The converted spectrum of the proposed algorithm was not liftered by the 40-th order mel-cepstrum.

5.1. Evaluation experiment on speech quality

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

The experimental result is shown in Figure 7. Error-bars denote standard deviations. The converted speech quality by the proposed algorithm ("GMM & DFW & Mix") is better than that of the GMM-based algorithm ("GMM") because the converted spectrum is not over-smoothed.

5.2. Evaluation experiment on speaker individuality

In order to evaluate the conversion-accuracy on speaker individuality of the proposed algorithm, the preference (ABX) test was performed. In the ABX test, A and B

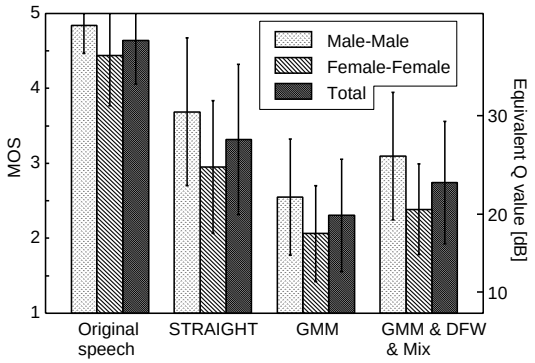

Figure 7: MOS (Mean Opinion Score) for the speech quality.

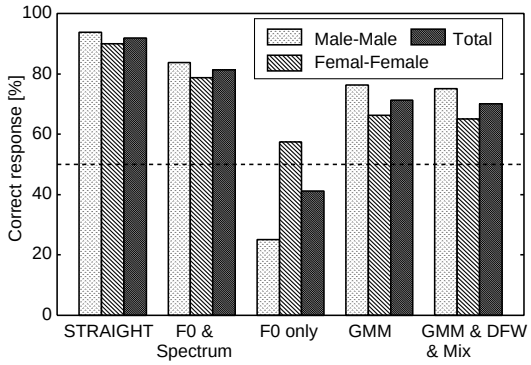

Figure 8: Correct response for speaker individuality.

were the source and the target speaker's speech, and X was one of the converted speech as follows,

- converted speech by the proposed algorithm$\cdots$"GMM & DFW & Mix",
- converted speech by the GMM-based algorithm$\cdots$"GMM",
- synthesized speech by converting of a log-scaled F0$\cdots$"F0 only",
- synthesized speech by converting of a log-scaled F0 and replacing the source speaker's spectra with those of the target speaker$\cdots$"F0 & spectrum",
- target speaker's speech synthesized by STRAIGHT$\cdots$"STRAIGHT".

"F0 & spectrum" was used to evaluate the conversion-accuracy on speaker individuality when conversion of spectra was perfect. "STRAIGHT" was used to evaluate the conversion-accuracy when both the conversion of spectra and the source information were perfect. Listeners were asked to select either A or B as being most similar to X. Two sentences that were not included in the training data were used to evaluate.

The experimental result is shown in Figure 8. The conversion-accuracy on speaker individuality of the proposed algorithm ("GMM & DFW & Mix") is the same as that of the GMM-based algorithm ("GMM"). The conversion-accuracy on speaker individuality of only F0 conversion ("F0 only") is insufficient, and it can be improved by converting spectra.

6. Conclusions

In this paper, we propose the voice conversion algorithm based on the Gaussian Mixture Model (GMM) with the Dynamic Frequency Warping (DFW) of STRAIGHT spectrum. We also propose that the converted spectrum calculated by mixing the GMM-based converted spectrum and the DFW-based converted spectrum. In order to evaluate the proposed algorithm, we performed evaluation experiments on speech quality and speaker individuality, compared with the GMM-based algorithm. The experimental results reveal 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 algorithm with the proper weight for mixing spectra.

7. Acknowledgment

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

[3] 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.

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

[5] 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.

[6] T. Toda, J. Lu, H. Saruwatari, and K. Shikano, "STRAIGHT-based voice conversion algorithm based on Gaussian mixture model," Proc. ICSLP, Beijing, China, pp. 279–282, Oct. 2000.

[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.