Maximum Likelihood Voice Conversion Based on GMM with STRAIGHT Mixed Excitation

Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science,
Nara Institute of Science and Technology, Japan
{yamato-o, tomoki, sawatari, shikano}@is.naist.jp

Abstract

The performance of voice conversion has been considerably improved through statistical modeling of spectral sequences. However, the converted speech still contains traces of artificial sounds. To alleviate this, it is necessary to statistically model a source sequence as well as a spectral sequence. In this paper, we introduce STRAIGHT mixed excitation to a framework of the voice conversion based on a Gaussian Mixture Model (GMM) on joint probability density of source and target features. We convert both spectral and source feature sequences based on Maximum Likelihood Estimation (MLE). Objective and subjective evaluation results demonstrate that the proposed source conversion produces strong improvements in both the converted speech quality and the conversion accuracy for speaker individuality.

Index Terms: Speech synthesis, Voice conversion, Gaussian mixture model, STRAIGHT, Mixed excitation

1. Introduction

Voice conversion techniques can convert the speech of a certain speaker to that of an another speaker. This technique can modify speech features based on conversion rules extracted from a small amount of training data. One typical application of voice conversion is speaker conversion [1], and this application can be extended to cross-language speaker conversion [2][3]. Cross-language speaker conversion is a technique that makes it possible for us to speak any language with own voice.

Although progress in research on statistical modeling of spectral sequences has improved the performance of voice conversion techniques, artificial sound is still evident in the converted speech. To alleviate the artificial sound, it is necessary to statistically model a source sequence as well as a spectral sequence.

Several researchers have proposed the source conversion methods such as the residual codebook [4], residual selection [5][6], and phase prediction [5]. The residual codebook uses speech coders with a speaker-dependent excitation codebook. Residual selection is a refinement of the residual codebook. This method selects appropriate residuals from a database extracted from the target speaker's training data. For phase prediction, the required phases are obtained from the predicted waveform shapes of converted spectra.

In our research, the STRAIGHT mixed excitation is used as our source model. STRAIGHT [7] is a high-quality vocoder. Advantages of STRAIGHT mixed excitation are that (1) the extracted features are statistically modeled in the same manner as that for spectral modeling, and (2) robust feature extraction is possible without pitch marks because of not using phase information. This source model is also used in the Nitech HTS system [10]. In this paper, we introduce STRAIGHT mixed excitation to maximum likelihood voice conversion based on a Gaussian Mixture Model (GMM) [8]. We convert both spectral and source feature sequences based on Maximum Likelihood Estimation (MLE). The proposed conversion's effectiveness is demonstrated through objective and subjective evaluations.

The paper is organized as follows. In Section 2, we describe STRAIGHT mixed excitation. In Section 3, the MLE-based spectral conversion is briefly explained, and in Section 4, we evaluate the experimental results. Finally, we summarize this paper in Section 5.

2. STRAIGHT Mixed Excitation

The mixed excitation using STRAIGHT [7] is defined as the frequency-dependent weighted sum of white noise and a pulse train with phase manipulation. The weight is determined based on an aperiodic component in each frequency bin [9]. Figure 1 shows a process for designing the STRAIGHT mixed excitation.

2.1. Aperiodic Component Analysis [9]

Figure 2 depicts aperiodic component extraction from a liftered power spectrum remaining the periodicity. The aperiodic component is calculated as a subtraction of an upper spectral envelope from a lower spectral envelope, where the upper one shows periodic components and the lower one represents noise components. Because the subtracted value should be less than 0 dB, the range of the aperiodic component is between 0 and 1. In the figure, aperiodicity is large when the lower envelope is close to the upper one. Figure 3 shows a normalized frequency distribution of aperiodic components in each frequency band. There is a noticeable tendency that periodicity is dominant in the lower frequency bands

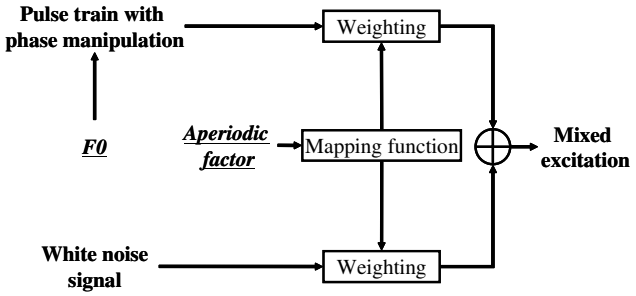

Figure 1: Design process for STRAIGHT mixed excitation.

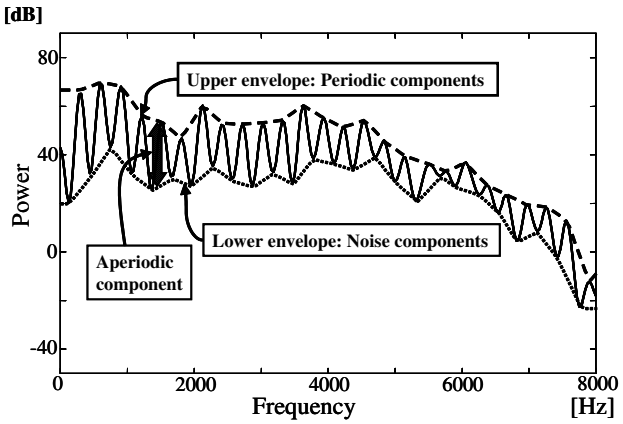

Figure 2: Aperiodic component extraction from liftered power spectrum keeping periodicity.

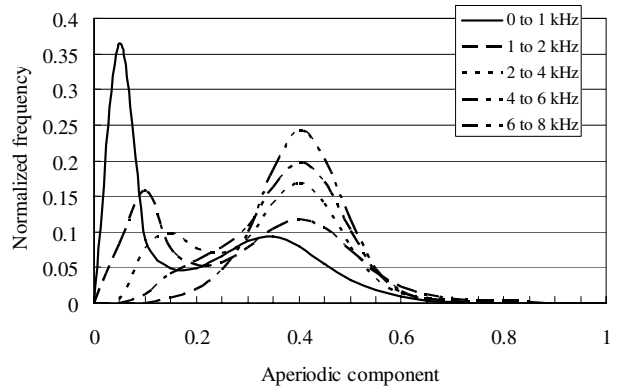

Figure 3: Normalized frequency distribution of aperiodic component on each frequency band.

and that aperiodicity is dominant in the higher ones.

2.2. Design of Excitation

The aperiodic component at each frequency bin is converted to the weight for a noise signal used in the mixed excitation as follows :

$$s\left(a_f\right) = \frac{1}{1+\exp\{-\alpha\left(a_f-0.25\right)\}}, \quad (1)$$

$$W\left(a_f\right) = \frac{s\left(a_f\right)-s\left(0\right)}{s\left(1\right)-s\left(0\right)}, \quad (2)$$

where a_f denotes the aperiodic component at each frequency bin and $W\left(a_f\right)$ is a mapping function. This mapping function varies according to the mapping parameter α as shown in Figure 4. As the mapping component α is larger, the aperiodic component is mapped onto the larger weight.

The mixed excitation is defined as follows:

$$S\left(f\right) = \sqrt{1-W\left(a_f\right)^2}\widetilde{P}\left(f\right)+W\left(a_f\right)N\left(f\right), \quad (3)$$

where $\widetilde{P}\left(f\right)$ denotes a pulse train with phase manipulation [7], and $N\left(f\right)$ denotes a white noise signal.

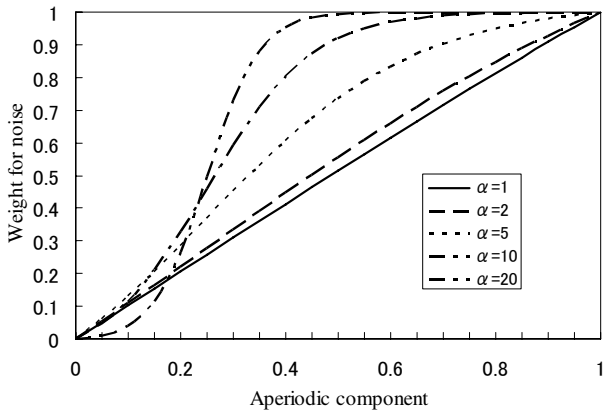

Figure 4: Mapping function from an aperiodic component into weight for noise when varying the mapping parameter α

3. Voice Conversion with STRAIGHT Mixed Excitation

3.1. MLE-based Conversion with GMM [8]

We use 2D-dimensional acoustic features, $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ (source speaker's) and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ (target speaker's), consisting of D-dimensional static and dynamic features, where $\top$ denotes transposition of the vector. By using time-aligned source and target features determined by Dynamic Time Warping, we train a GMM to model the joint probability density $p\left(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\Theta}\right)$, where $\boldsymbol{\Theta}$ denotes model parameters.

When converting the source static and dynamic feature vectors $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ to the target static feature vectors $\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top\right]^\top$, the following function is maximized with respect to $\boldsymbol{y}$,

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} \log\left\{p\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Theta}\right)^\omega \cdot p\left(\boldsymbol{\nu}(\boldsymbol{y})|\boldsymbol{\theta}_\nu\right)\right\} \quad (4)$$
$$\text{Subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y},$$

where $p\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\Theta}\right)$ denotes the likelihood of conditional probability density functions (pdfs) on the target static and dynamic feature vectors, and $p\left(\boldsymbol{\nu}(\boldsymbol{y})|\boldsymbol{\theta}_\nu\right)$ represents the likelihood of a pdf on the global variance (GV) of the target static feature vectors.

3.2. Applying STRAIGHT Mixed Excitation

Figure 5 shows the process of the proposed voice conversion. Our proposed method employs two GMMs. One is used for the spectral conversion and the other is for the aperiodic conversion. Both conversions are performed with MLE. We consider global variance (GV) only in the spectral conversion because GV does not cause any large difference to the converted speech in the aperiodic conversion. We synthesize the mixed excitation from the converted aperiodic components. Finally, we synthesize the convert speech by filtering the excitation with the converted spectra.

4. Experimental Evaluation

We used the speech data of two male speakers and two female speakers from ATR's phonetically balanced sentence database [11]. We considered 50 sentences for training data, and

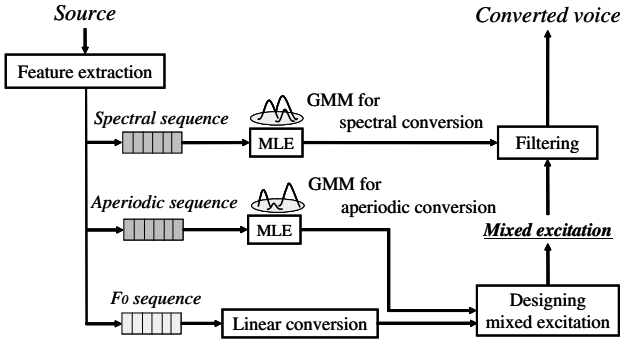

Figure 5: Process of the proposed voice conversion.

another 50 sentences for the evaluation. The total number of combinations of source and target speakers was 12.

For the spectral feature, we take the first through the 24th mel-cepstral coefficients from the STRAIGHT smoothed spectrum. For the aperiodic feature, we used average dB values of the aperiodic components on five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 kHz).

In each feature conversion, we used full covariance matrices, and set the number of mixtures for the spectral conversion to 32 based on our preliminary experiment.

4.1. Optimization of Mapping Parameter

To optimize the mapping parameter α for each speaker, we evaluated the aperiodic component distortion between natural speech and analysis-synthesized speech. Figure 6 shows the aperiodic component distortion as a function of the mapping parameter α. It is apparent that 8 is the optimal value for every speaker, thus we designed STRAIGHT mixed excitation using this value in the following.

To demonstrate the effectiveness of the mixed excitation in the analysis-synthesis, we evaluated the speech quality of natural speech, analysis-synthesized speech without mixed excitation, and analysis-synthesized speech with mixed excitation. Figure 7 shows the result of a preference test. The number of listeners in the case was five. The figure shows that the speech quality of analysis-synthesized speech using mixed excitation is higher than that without mixed excitation.

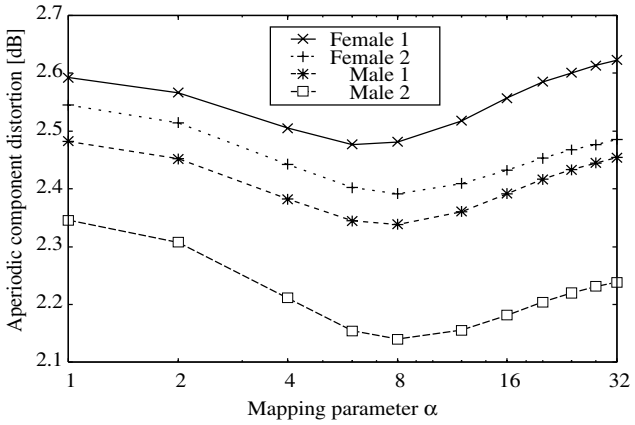

Figure 6: Aperiodic component distortion as a function of the mapping parameter α.

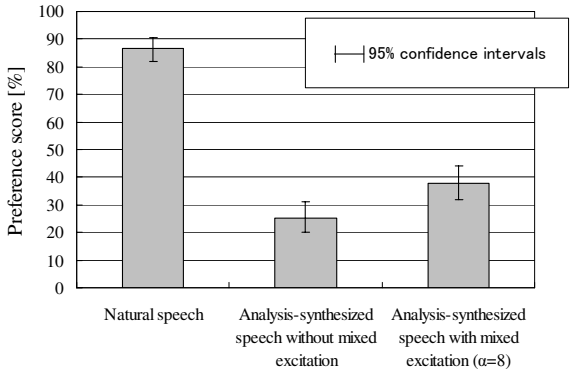

Figure 7: Result of preference test on speech quality.

4.2. Objective Evaluation of Conversion Quality

We evaluated the distortion between the target aperiodic component and the converted one. Figure 8 shows the aperiodic component distortion as a function of the number of mixtures. The aperiodic conversion causes a reduction of the aperiodic distortion. Therefore, the conversion causes the source signal of which characteristics are much more similar to those of the target speaker compared with those of the source speaker. The optimum number of mixtures is 32. However, it is shown that the conversion performance is not very sensitive to the number of mixtures.

4.3. Subjective Evaluation of Speech Quality and Speaker Individuality

We subjectively evaluated the converted speech quality and the conversion accuracy for the speaker individuality. In this evaluation, we employed the following converted voices:

- Converted voice without the mixed excitation
- Converted voice with the mixed excitation based on source speaker's aperiodic component
- Converted voice with the mixed excitation based on the converted aperiodic component

In the preference test for speech quality, we randomly presented a pair of voices from three kinds of voice to eight listeners.

In the XAB test on speaker individuality, we presented the target speaker's voice and after that a pair of converted voices randomly. Then we asked listeners which converted voice is similar to the target speaker's. The number of listeners was six.

4.3.1. Speech Quality

Figure 9 shows the result of the preference test. The STRAIGHT mixed excitation greatly improved speech quality when using mixed excitation. Moreover, the results reveal that the aperiodic conversion slightly improves the converted speech quality.

4.3.2. Speaker Individuality

Figure 10 shows the result of the XAB test. We can see that the conversion accuracy for speaker individuality was also improved by using STRAIGHT mixed excitation. In addition, we can slightly improve it further by converting the aperiodic components.

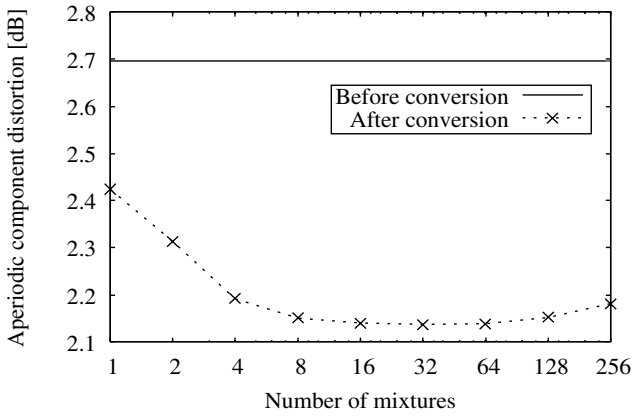

Figure 8: The aperiodic component distortion as a function of the number of mixtures.

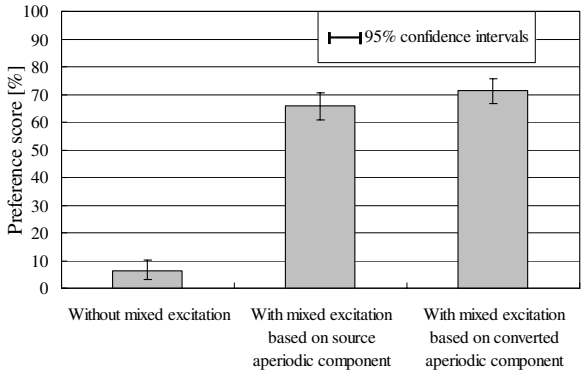

Figure 9: Result of preference test on speech quality.

5. Conclusions

In this paper, we introduced STRAIGHT mixed excitation to Maximum Likelihood Estimation (MLE)-based voice conversion with a Gaussian Mixture model (GMM) in order to improve the converted speech quality and the conversion accuracy for speaker individuality. We statistically converted a source feature sequence of the STRAIGHT mixed excitation as well as a spectral sequence. In addition, we subjectively evaluated the proposed conversion method, finding that proposed method improved both converted speech quality and conversion accuracy for speaker individuality.

6. Acknowledgements

This research was supported in part by MEXT's (the Japanese Ministry of Education, Culture, Sports, Science and Technology) e-Society project.

7. References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, Vol. 11, No. 2, pp. 71–76, 1990.

[2] M. Abe, K. Shikano, and H. Kuwabara, "Statistical analysis of bilingual speaker's speech for cross-language voice conversion," *J. Acoust. Soc. Am*, Vol. 90, No. 1, pp. 76–82, 1991.

[3] M. Mashimo, T. Toda, H. Kawanami. K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," *IPSJ Journal*, Vol. 43, No. 7, pp. 2177–2185, 2002.

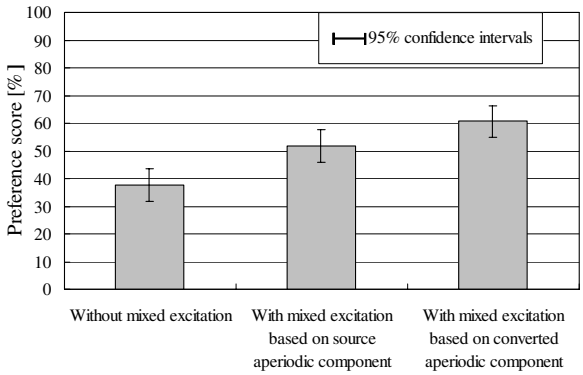

Figure 10: Result of preference test on conversion accuracy for speaker individuality.

[4] A. Kain and M.W. Macon. "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," *Proc. ICASSP*, pp. 813–816, May 2001.

[5] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A Study on residual prediction techniques for voice conversion," *ICASSP*, pp. 13–16, Philadelphia, USA, 2005.

[6] H. Ye and S.J. Young, "High quality voice morphing," *ICASSP* , pp 9–12, Montreal, Canada, 2004.

[7] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and instantaneous- frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication.* Vol. 27, No. 3-4, pp. 187–207, 1999.

[8] T. Toda, A.W. Black and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," *Proc. ICASSP*, pp. 9–12, March 2005.

[9] H. Kawahara, Jo Estill and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *MAVEBA 2001*, Sept.13-15, Firentze Italy, 2001.

[10] H. Zen and T. Toda, An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005, *INTERSPEECH 2005 - EUROSPEECH*, pp. 93–96, Lisbon, Portugal, September 2005

[11] M. Abe et al, "ATR technical report," TR-I-0166, 1990