The Seventh Western Pacific Regional Acoustics Conference
WESTPRAC VII
Kumamoto, Japan, 3-5 October 2000

VOICE CONVERSION ALGORITHM BASED ON GAUSSIAN MIXTURE MODEL APPLIED TO STRAIGHT

Tomoki TODA, Jinlin LU, Satoshi NAKAMURA, Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 JAPAN
Fax: +81-743-72-5289
E-mail: {tomoki-t,lu,nakamura,shikano}@is.aist-nara.ac.jp

ABSTRACT

Voice conversion is a technique used to convert one speaker's voice into another speaker's voice. As a typical voice conversion algorithm, the codebook mapping method has been studied by Abe et al. The main shortcoming of this method is the fact that the acoustic space of a speaker is limited to a discrete representation. To represent the acoustic space continuously, the algorithm based on the Gaussian mixture model (GMM) has also been proposed by Stylianou et al. In this paper, we apply this GMM-based voice conversion algorithm to STRAIGHT proposed by Kawahara et al., which is recognized as a high quality vocoder. In order to evaluate this voice conversion algorithm, we performed subjective and objective experiments on speaker individuality and speech quality, comparing with the method based on the codebook mapping. As results, the performance of the GMM-based voice conversion algorithm is better than that of the codebook mapping method. Effects by the amount of training data for the voice conversion algorithms were also investigated, as well as the number of the Gaussian mixtures. These evaluation results clarify that the GMM-based voice conversion algorithm is successfully applied to STRAIGHT.

KEYWORDS: voice conversion, codebook mapping, Gaussian mixture model, STRAIGHT

INTRODUCTION

As a typical voice conversion algorithm, the codebook mapping method has been studied by Abe et al. [1]. The main shortcoming of this method is the fact that the acoustic space of a speaker is limited to a discrete representation because of vector quantization usage. To represent the acoustic space continuously, the algorithm based on the Gaussian

Mixture Model (GMM) has also been proposed by Stylianou et al. [2]. In this GMM-based algorithm, the acoustic space of a speaker is modeled by the GMM, and acoustic features are converted from a source speaker to a target speaker by mapping function based on the Gaussian mixture.

Voice conversion is usually performed with an analysis-synthesis method, where quality of the synthesized speech is also important to realize a high quality voice conversion algorithm. STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) proposed by Kawahara et al. is an analysis-synthesis method and can synthesize high quality speech [3].

In this paper, we apply the GMM-based voice conversion algorithm to STRAIGHT, and evaluate this voice conversion algorithm, comparing with the method based on the codebook mapping.

VOICE CONVERSION ALGORITHM BASED ON GMM

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

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

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

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

Since the acoustic space of a speaker is modeled by the GMM without the use of vector quantization, the GMM-based algorithm distortion for the represented acoustic space is less than that of the codebook mapping method.

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

$$F(\boldsymbol{x}) = \sum_{i=1}^{m} h_i(\boldsymbol{x})[\boldsymbol{\mu}_i^{\boldsymbol{y}} + \Sigma_i^{\boldsymbol{yx}}(\Sigma_i^{\boldsymbol{xx}})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i^{\boldsymbol{x}})], \quad h_i(\boldsymbol{x}) = \frac{\alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i^{\boldsymbol{x}}, \Sigma_i^{\boldsymbol{xx}})}{\sum_{j=1}^{m} \alpha_j N(\boldsymbol{x}; \boldsymbol{\mu}_j^{\boldsymbol{x}}, \Sigma_j^{\boldsymbol{xx}})}, \tag{2}$$

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

In order to estimate parameters $(\alpha_i, \boldsymbol{\mu}_i^{\boldsymbol{x}}, \boldsymbol{\mu}_i^{\boldsymbol{y}}, \Sigma_i^{\boldsymbol{xx}}, \Sigma_i^{\boldsymbol{yx}})$, the probability distribution of the joint vectors $\boldsymbol{z} = [\boldsymbol{x}^T,\ \boldsymbol{y}^T]^T$ of the source and target speakers is represent by the GMM [4]. Covariance matrix $\Sigma_i^{\boldsymbol{z}}$ and mean vector $\boldsymbol{\mu}_i^{\boldsymbol{z}}$ of class i of joint vectors are estimated by the EM algorithm and can be written as

$$\Sigma_i^{\boldsymbol{z}} = \begin{bmatrix} \Sigma_i^{\boldsymbol{xx}} & \Sigma_i^{\boldsymbol{xy}} \\ \Sigma_i^{\boldsymbol{yx}} & \Sigma_i^{\boldsymbol{yy}} \end{bmatrix}, \qquad \boldsymbol{\mu}_i^{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{\mu}_i^{\boldsymbol{x}} \\ \boldsymbol{\mu}_i^{\boldsymbol{y}} \end{bmatrix}. \tag{3}$$

Since acoustic features are converted by this mapping function that utilizes feature parameter correlation between two speakers, the converted speech is represented more continuously than that of the codebook mapping method.

APPLICATION OF THE VOICE CONVERSION ALGORITHMS TO STRAIGHT

The cepstrum of the smoothed spectrum analyzed by STRAIGHT is used as acoustic features. In this paper, the cesptrum order is 40 (quefrency is 2.5 ms, 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, power, is kept as the value of the source speaker. As for the source information, the average of log-scaled pitch frequencies of the source speaker is converted to that of the target speaker. The prosodic dynamic characteristics of both speakers are not considered.

EVALUATION

In order to evaluate the performance of the GMM-based voice conversion algorithm that was applied to STRAIGHT, we performed experiments on speaker individuality and speech quality, comparing with the method based on the codebook mapping. The male-to-male and female-to-female voice conversion was performed in each experiment.

Objective Evaluation Experiments on Speaker Individuality. In order to evaluate converted speaker individuality of the GMM-based voice conversion algorithm, objective evaluation experiments were performed by the cepstrum distortion (CD) between the converted speech and the target speech. Ten sentences were used to evaluate, which were not included in the training data.

First, in order to investigate the relation between the number of classes and CDs, CDs for the converted speech by both voice conversion algorithms were calculated. Fifty-eight sentences were used as the training data. The experimental result is shown in Fig. 1. CDs decrease at both voice conversion algorithms as the number of classes increases. The CD performance of the GMM-based voice conversion algorithm is superior to that of the codebook mapping method.

Next, in order to investigate the relation between the amount of training data and CDs, CDs for the converted speech by the GMM-based voice conversion algorithm (16, 64 classes) and the codebook mapping method (16, 64, 256, 1024 classes) were calculated. The experimental result for the female-to-female voice conversion is shown in Fig. 2. CDs increase when the amount of training data is insufficient, because training of parameters of the mapping function is not enough. The result for the male-to-male voice conversion is similar to that of the female-to-female voice conversion.

Subjective Evaluation Experiments on Speech Quality. In order to evaluate quality of the GMM-based converted speech, subjective evaluation experiments were performed. Eight listeners participated in the experiments. 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 were used to evaluate, which were not included in the training data.

First, in order to investigate the relation between the number of classes and speech quality, the converted speech by the GMM-based voice conversion algorithm (16, 64 classes) and the codebook mapping method (16, 64, 256, 1024 classes) was used. Fifty-eight sentences were used as the training data. The experimental result is shown in Fig. 3. Speech quality becomes better at both voice conversion algorithms as the number of classes increases. The performance of the GMM-based voice conversion algorithm is superior to that of the codebook mapping method.

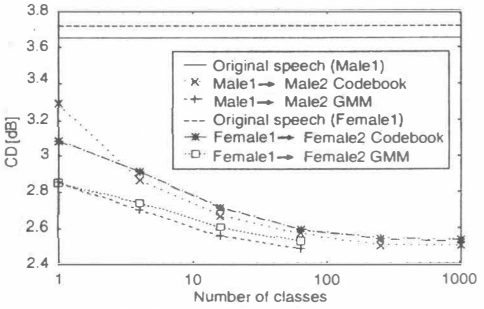

Fig. 1: Relation between the number of the classes and CD.

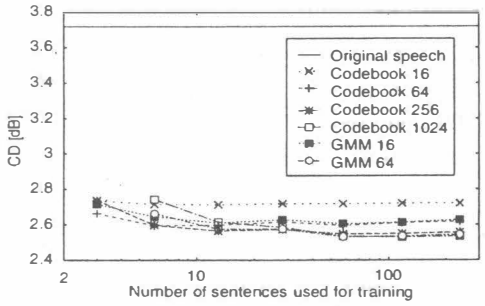

Fig. 2: Relation between the amount of training data and CD.

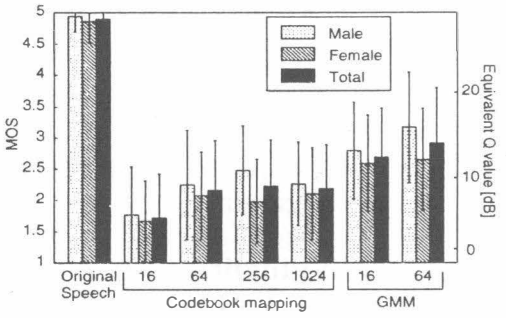

Fig. 3: Relation between speech quality and the number of the classes.

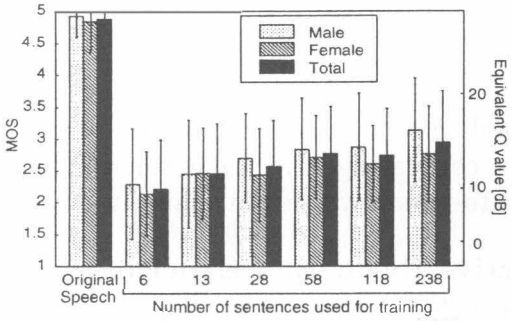

Fig. 4: Relation between speech quality and the amount of training data on GMM (64 classes).

Next, in order to investigate the relation between the amount of training data and speech quality, the converted speech by the GMM-based voice conversion algorithm (64 classes) was used. The experimental result is shown in Fig. 4. Speech quality becomes better as the amount of training data is large. When the amount of training data is insufficient, speech quality is also low, because training of parameters of the mapping function is not enough.

CONCLUSIONS

We apply the voice conversion algorithm based on the Gaussian Mixture Model (GMM) to STRAIGHT, and evaluate this voice conversion algorithm. We performed evaluation experiments on speaker individuality and speech quality, comparing with the method based on the codebook mapping. As a result, the performance of the GMM-based voice conversion algorithm is better than that of the codebook mapping method. Effects by the amount of training data for the voice conversion algorithms were also investigated, as well as the number of the Gaussian mixtures. These evaluation results clarify that the performance becomes better as the number of mixtures increases and the amount of training data is large.

REFERENCES

1. M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice conversion through vector quantization," *Proc. ICASSP*, pp. 655–658, 1988.
2. Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," *Proc. EUROSPEECH*, pp. 447–450, 1995.
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*, **27**, pp. 187–207, 1999.
4. A. Kain, M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, pp. 285–288, 1998.