EVALUATION OF CROSS-LANGUAGE VOICE CONVERSION BASED ON GMM AND STRAIGHT

Mikiko Mashimo[†], *Tomoki Toda*[†], *Kiyohiro Shikano*[†], *and Nick Campbell*[‡]

[†] Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan
[mikiko-m, tomoki-t, shikano]@is.aist-nara.ac.jp
[‡]ATR Information Sciences Division, Kyoto, 619-0288, Japan
nick@slt.atr.co.jp

Abstract

Voice conversion is a technique for producing utterances using any target speakers' voice from a single source speaker's utterance. In this paper, we apply cross-language voice conversion between Japanese and English to a system based on a Gaussian Mixture Model (GMM) method and STRAIGHT, a high quality vocoder. To investigate the effects of this conversion system across different languages, we recorded two sets of bilingual utterances and performed voice conversion experiments using a mapping function which converts parameters of acoustic features for a source speaker to those of a target speaker. The mapping functions were trained using bilingual databases of both Japanese and English speech. In an objective evaluation using Mel cepstrum distortion (Mel CD), it was confirmed that the system can perform cross-language voice conversion with the same performance as that within a single-language.

1. INTRODUCTION

Since voice conversion allows mapping of any target speaker's voice after training using a small number of source speaker utterances (roughly 50-60 sentences), it has a potential for various applications. Our goal is to capture not only the target speaker's voice and speaking style, but also to convert across language pairs which the original speaker may not be capable of. This method would have potential applications in e.g., computer aided language learning system and interpretation systems, although cross-language voice conversion has not yet been well researched.

For cross-language conversion, the quality of converted speech should sound as if the target speaker had spoken the other language, and the speaker individuality should also be preserved across different languages. An attempt was made by Abe et al [1] in the late 1980's between Japanese and English using a codebook mapping method [2], which became the typical voice conversion algorithm. Recently, an algorithm based on the Gaussian Mixture Model (GMM) has been proposed by Stylianou et al [3],[4]. The advantage of this method is that the acoustic space of a speaker is modeled by the GMM, so that acoustic features are converted from a source speaker to a target speaker continuously. The codebook mapping method uses a discrete representation through vector quantization. Voice conversion algorithms based on the GMM method were applied to a high quality vocoder STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) (Kawahara et al [5],[6]), by Toda et al [7],[8]. From their objective and subjective evaluation report, this system succeeded to produce high quality of converted voice within a single-language.

The purpose of the present paper is to apply cross-language voice conversion using the GMM and STRAIGHT-based system and to evaluate the effect of it as a first step towards producing high quality voice conversion between different languages. In the present study, we are working on the assumption that this voice conversion system will be applied to a practical pronunciation learning system. In most language learning systems, there is only a single language dataset for the learner (i.e., that of target speaker). However, we collected a bilingual female speaker's databases of both source and target texts for investigation into whether the differences between the two languages has an effect on the converted voice.

2. VOICE CONVERSION ALGORITHM

In our method, p-dimensional time-aligned acoustic features of a source speaker and a target speaker determined by Dynamic Time Warping (DTW) are assumed as below, where T denotes transposition.

source speaker : $\boldsymbol{x}\{[x_0, x_1, \cdots, x_{p-1}]^T\}$,
target speaker : $\boldsymbol{y}\{[y_0, y_1, \cdots, y_{p-1}]^T\}$,

2.1. GMM-based voice conversion algorithms

In the GMM algorithm, the training data size and the number of trainable parameters are variable [3],[4]. The probability distribution of acoustic features x can be described as

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

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

2.2. Conversion of acoustic features

Conversion of the acoustic features of the source speaker to those of the target speaker is performed by a Mapping Function, defined as follows,

$$\begin{aligned} F(x) &= E[y|x] \\ &= \sum_{i=1}^{m} h_i(x)[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x-\mu_i^x)], \\ h_i(x) &= \frac{\alpha_i N(x;\mu_i^x,\Sigma_i^{xx})}{\sum_{j=1}^{m}\alpha_j N(x;\mu_j^x,\Sigma_j^{xx})}, \end{aligned} \quad (2)$$

where μ_i^x and μ_i^y denote mean vectors of class i for the source and target speakers. Σ_i^{xx} is covariance matrix of class i for the source speaker. Σ_i^{yx} is the cross-covariance matrix of class i for the source and target speakers. These matrices are diagonal.

2.3. Training of The Mapping Function

In order to estimate parameters such as $\alpha_i, \mu_i^x, \mu_i^y, \Sigma_i^{xx}, \Sigma_i^{yx}$, the probability distribution of the joint vectors $z = [x^T, y^T]^T$ for the source and target speakers is represented by the GMM whose parameters are trained by joint density distribution [9]. Covariance matrix Σ_i^z and mean vector μ_i^z of class i for joint vectors can be written as

$$\Sigma_i^z = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}, \qquad \mu_i^z = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}. \quad (3)$$

Expectation maximization (EM) is used for estimating these parameters.

3. ANALYSIS-SYNTHESIS METHOD

In a voice conversion system, not only the voice conversion algorithm but the quality of the analysis-synthesis method determines the quality of the synthesized voice. Therefore, choosing a reliable analysis-synthesis method is of importance. In our work, STRAIGHT was employed as the analysis-synthesis method. STRAIGHT is a very high quality vocoder developed to meet the necessity of a flexible and high quality analysis-synthesis [5],[6]. It consists of pitch adaptive spectrogram smoothing and fundamental frequency extraction (TEMPO), and allows manipulation of speech parameters such as vocal tract length, pitch, and speaking rate.

Table 1: Recording conditions.

Recording place	Sound treated room
Microphone	SONY C355
Recording equipment	DAT SONY DTC-ZA5ES
Sampling frequency	48000 Hz
Number of sentences	60

4. IMPLEMENTATION OF THE CONVERSION ALGORITHMS

The GMM-based voice conversion algorithm has been implemented in STRAIGHT by Toda et al [7], [8]. In their system, acoustic features are described by the cepstrum of the smoothed spectrum analyzed by STRAIGHT. In our work, however, we used Mel cepstrum because of it's closeness to human auditory perception. The prosodic characteristics have not been considered yet but the fundamental frequency (F_0) of the source speaker is adjusted to match the target speaker's F_0 in average of log-scale for the source information. The adjusting function is described as follows,

$$f_0' = \frac{\mu_y}{\mu_x} \times f_0 \quad (4)$$

where f_0 and f_0' denote log scale F_0 of source speaker and converted speech of source speaker, and μ_x and μ_y denote mean log scale F_0 of source speaker and target speaker.

5. EXPERIMENT

5.1. Speech Databases

Bilingual (Japanese and English) speech utterances of two Japanese female speakers were recorded, sampled at 48000 Hz. Each speaker has long experience living in abroad or having learned English from a native speaker since before the age of seven. The speakers read 60 bilingual sentences selected from the ATR phonetically balanced sentences [10]. After down sampling to 16000 Hz, the 50 sentences were used for training data sets, and the remaining 10 were used for evaluation sentences for the converted utterances. Table 1 shows other recording conditions.

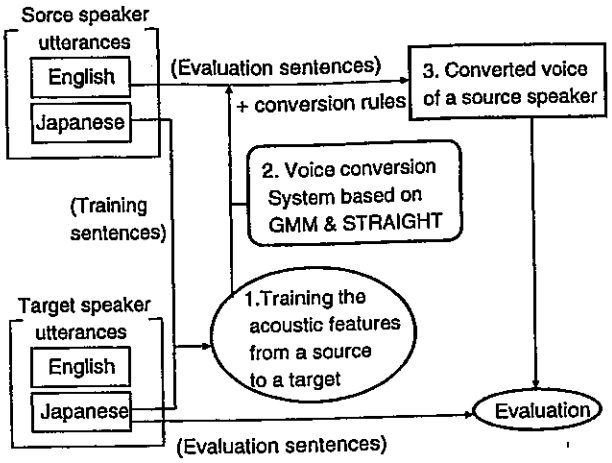

Figure 1: Diagram of cross-language conversion procedure, English converted voice trained by Japanese.

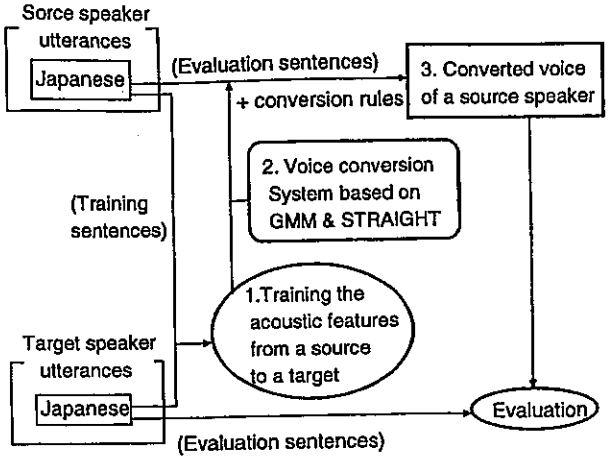

Figure 2: Diagram of single-language conversion procedure, Japanese converted voice trained by Japanese.

5.2. Voice Conversion

In order to investigate the differences between voice conversion across different languages, both Japanese and English trained mapping functions were used for learning the source and target speaker's parameters for conversion of acoustic features. Therefore, the system was tested on 4 types of female to female converted voice, (1) English (Eng) converted voice trained by Japanese (Jpn), (2) Jpn converted voice trained by Jpn, (3) Eng converted voice trained by Jpn, and (4) Jpn converted voice trained by Jpn.The procedure for producing utterances of (1) Eng converted voice trained by Jpn and (2) Jpn converted voice trained by Jpn are depicted in Figure 1 and Figure 2. The procedures of voice conversion trained by English is the same.

According to the previous work of Toda et al [7],[8], the relation between the number of GMM classes and cepstrum distortion (CD) saturates at a certain number of classes, approximately 64, so we used 64 GMM classes. Other analysis parameters were shown in Table 2. Note that the mean F_0 of speakers is different.

Table 2: Analysis parameters.

Analysis Window	Gaussian
sampling frequency	16000 Hz
Shift length	5 ms
Number of FFT points	1024
Number of the GMM class	64
Training sentences	50
Evaluation sentences	10
mean F_0 (source speaker)	Jpn: 270.0 Hz
	Eng: 248.6 Hz
mean F_0 (target speaker)	Jpn: 227.6 Hz
	Eng: 233.9 Hz

6. EVALUATION

To evaluate speaker individuality objectively, a Mel cepstrum distortion (Mel CD) function was calculated between the converted speech and the target speech. Mel CD is calculated as below, where $mc_i^{(conv)}$ and $mc_i^{(tar)}$ denote Mel CD coefficients of converted voice and target voice, respectively.

$$MelCD = 10/ln10\sqrt{2\Sigma_{i=1}^{40}(mc_i^{(conv)} - mc_i^{(tar)})^2}. \quad (5)$$

If the value of Mel CD is smaller, speaker individuality of converted voice is closer to that of target speaker. Figure 3 shows the results of Mel CD whose conversion rules were trained by 50 English and 50 Japanese sentences. Table 6 shows the values numerically.

We can see from the results in the table that characteristics of converted speech were also improved within the cross-language voice conversion. However, since there are normally large spectral differences between Japanese and English speech sounds, the results of the converted speech trained by the same language show closer values (i.e. English converted voice trained by English and Japanese converted voice trained by Japanese is still preferred). In addition, for producing a higher quality converted voice, we must consider the voice quality differences between Japanese and English of the same speaker.

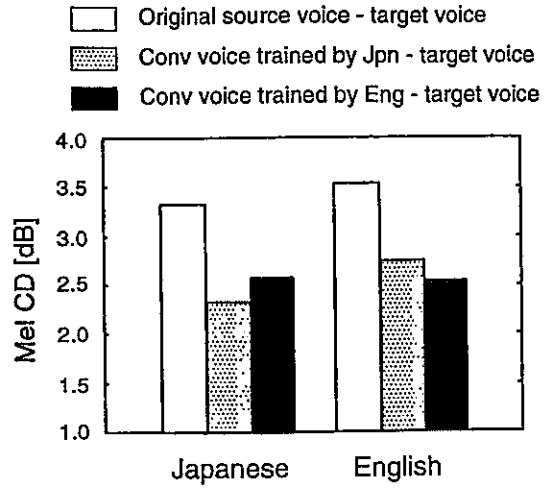

Figure 3: Result of the objective evaluation experiment of speaker individuality.

Table 3: Values of Mel cepstrum distortion.

	Jpn (Hz)	Eng (Hz)
Original Source - Target	3.33	3.53
converted voice - Target (trained by Jpn)	2.32	2.75
converted voice - Target (trained by Eng)	2.59	2.52

7. CONCLUSION

In this paper, we evaluated the effect of applying cross-language voice conversion to a system based on GMM (Gaussian Mixture Model) and STRAIGHT. From the results of the objective evaluation using Mel cepstrum distortion, it was found that the system performs cross-language voice conversion nearly equivalent to that of single-langu- age conversion. This indicates that it has a possibility to be employed to a language learning system. For the next step, the problem of mean fundamental frequency (F_0) differences between Japanese and English utterances of the same speaker and variation of the quality of voice must be considered, as this will cause inconsistency in perception of converted voice sounds. Future work will include developing a method of perceptual evaluation which takes such differences into account.

8. ACKNOWLEDGMENT

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

9. References

[1] M. Abe, K. Shikano and H. Kuwabara, "Statistical analysis of bilingual speaker's speech for cross-language voice conversion," J. Acoust. Soc. Am. 90(1), pp. 76–82, July 1991

[2] 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.

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

[4] Y. Stylianou, 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.

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

[6] H. Kawahara, I. Masuda-Katsuse, 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.

[7] T. Toda, J. Lu, H. Saruwatari, K. Shikano, "STRAIGHT-based voice conversion algorithm based on gaussian mixture model," Proc. ICSLP, PAe(09-10)-K-05, pp. 279-282, Beijing, China, Oct. 2000.

[8] T. Toda, H. Saruwatari, K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," Proc. ICASSP, Salt Lake City, U.S.A., May 2001.

[9] A. Kain, and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP, Seattle, U.S.A., May 1998.

[10] M. Abe, Y. Sagisaka, T. Umeda and H. Kuwabara, "Speech Database Usr's Manual," ATR Technical Report (in Japanese)