EVALUATION OF CROSS-LANGUAGE VOICE CONVERSION USING BILINGUAL AND NON-BILINGUAL DATABASES

Mikiko Mashimo[†], *Tomoki Toda*[†], *Hiromichi Kawanami*[†], *Hideki Kashioka*[†‡], *Kiyohiro Shikano*[†] *and Nick Campbell*[†§]

[†] Nara Institute of Science and Technology, 8916-5, Nara, 630-0101 Japan
[‡]ATR Spoken Language Translation Research Laboratories
[§]ATR Human Information Sciences, Kyoto, 619-0288, Japan
mikiko-m@is.aist-nara.ac.jp

ABSTRACT

Cross-language voice conversion is useful for many applications, and we are trying to apply the technique to a language training system for reducing voice individuality differences. In this paper, we describe experiments that test effectiveness of an extension of single-language voice conversion, to include cross-language utterances. The performance was investigated by objective and perceptual evaluation using bilingual-speakers data for training. Then, the correlations between a computed distance measure and a human perceptual pronunciation evaluation score were compared before and after applying conversion. From these results, it was found that the cross-language voice conversion reduces speakers' voice differences between the pairs, and the phoneme based measures show somewhat clearer correspondences to the human perceptual score in vowels' test after applying voice conversion.

1. INTRODUCTION

Voice individuality plays an important role in human speech, and accordingly, much work has been performed in the field of speech synthesis to model the characteristics of individual speakers. One such technique is voice conversion, a method for converting one voice to another. If this method is extended across languages, it could have applications in automatic speech translation, or in foreign-language training.

In a computer language training system, voice individuality differences between tutor and student prevent direct comparison of the pairs. Our goal is to facilitate automatic evaluation of a student's performance on a computer language training system using a voice conversion technique, by direct comparison with the speech of the tutor after it has been modified to match the voice of the student. Therefore, we need to evaluate the effectiveness of cross-language voice conversion.

Cross-language voice conversion was developed by Abe et al. [1] [2] in the late 1980's using a codebook mapping technique. Their method used a discrete representation of the acoustic features that contribute to speaker individuality, mapping between code vectors by minimising the acoustic distortion between the pairs. We employ an algorithm based on the Gaussian Mixture Model (GMM) proposed by Stylianou et al. [3] to model the acoustic space of a speaker continuously, and this technique has advantages over the discrete codebook mapping methods. Toda et al. [4] reported an application of this voice conversion method, in conjunction with a high-quality vocoder STRAIGHT (Speech Transformation and Representation using Ad- aptive Interpolation of weiGHTed spectrum) [5] [6], which was shown to produce better speech quality than the codebook-based methods. Further improvements were gained from inclusion of Dynamic Frequency Warping (DFW) and spectral conversion combining GMM-based and DFW-based algorithms [7].

In this paper, we apply the voice conversion method across languages, and report the effectiveness through objective and perceptual experiments. Also, the possibility of using cross-language voice conversion for a language training is discussed.

2. CROSS-LANGUAGE VOICE CONVERSION

Abe et al.[2] performed a statistical analysis of spectrum differences between Japanese and English. From their findings, we assume that voice differences between the speakers are more important than the acoustic differences between the languages when performing cross-language voice conversion.

Ideally, the speaker's voice individuality should be preserved across the different languages. A measure of acoustic distances between the converted voice and the target voice is therefore essential for the evaluation of success in cross-language voice conversion. To take these points into account, we test the efficiency of the cross-language voice conversion technique using utterances of natural speech. By the use of bilingual datasets for both source and target speaker's data, we can be assured of natural, native speaker utterances both within and across language pairs, and these datasets allow detailed evaluation for voice conversion.

Figure 1 shows the flow of processing, an example of English converted after training on Japanese. Two sets of bilingual data were recorded from native Japanese speakers, and training was performed first using Japanese speech to develop the mapping-weight vectors. These weights were then used to perform voice-conversion of English utterances from the same speakers. The process was then repeated using English utterance pairs for training and Japanese utterance pairs for testing. Tests were also performed using monolingual speech pairs to produce baseline performance ratings. To evaluate the quality of the voice conversion methods, a conversion accuracy evaluation test and perceptual evaluation test were carried out. In the following sections, experimental details and results are reported.

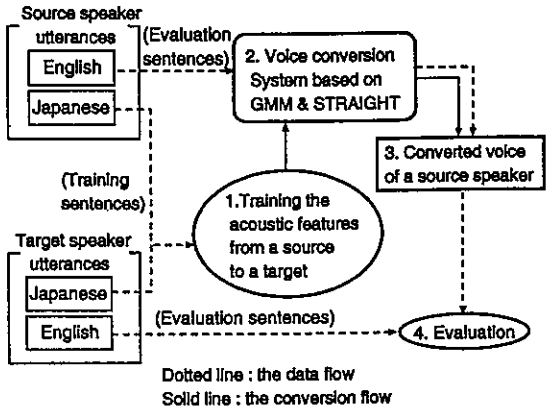

Fig. 1. Diagram of cross-language conversion procedure, English converted after training on Japanese.

3. EXPERIMENTAL SPEECH DATABASES

For the tests of cross-language voice conversion, utterances of four native Japanese speakers with bilingual language skills (2 females who have an American English accent, 2 males who have a British English accent) were recorded in a soundproof room with studio-quality equipment and their speech was digitised using 48 kHz, 16 bit sampling. Since the final target of the voice conversion system is for Japanese to English speech, we selected bilingual speakers who are fluent in these two languages. Speaker Fe-A had learnt English from an American native speaker from the age of 3-yrs, and speaker Fe-B has more than 12 years experience of living in an English speaking country. Speaker Ma-A has a Japanese mother and an English father, and speaker Ma-B is a specialist in British English phonology. Additionally, to investigate the possibility of using cross-language voice conversion in a language training system, we recorded one native Japanese female who has not particular English skills as the students' utterances. These speakers read 60 sentences in each language. Those sentences were selected from the ATR phonetically balanced corpus for Japanese and the TIMIT databases for English, respectively.

4. VOICE CONVERSION SYSTEM

The GMM-based voice conversion algorithm has been implemented in STRAIGHT by Toda et al [4]. It was confirmed that the system could reliably convert synthetic voices in a single-language test. STRAIGHT is a high quality vocoder developed to meet the need for a flexible and high quality analysis-synthesis method [5][6]. It consists of pitch adaptive spectrogram smoothing and fundamental frequency extraction using TEMPO (Time-domain Excitation extractor using Minimum Perturbation Operator), and allows manipulation of speech parameters such as vocal tract length, pitch, and speaking rate maintaining a high quality voice sounds.

As our acoustic feature, we employed the Mel cepstrum of the smoothed spectrum analyzed by STRAIGHT. The cepstral order was set to be 40, and the 1 to 40th order cepstrum coefficients were used for voice conversion. These 40 cepstra were used to map between source and target speaker's frames by DP matching based on cepstral distances. The number of the GMM classes were 64. The waveform power of the source speaker was not manipulated.

Table 1. MelCD values within the same speaker, Fe-A [dB]

	Japanese	English
(1)-(2)	4.64	4.54
(1)-(3)	4.52	4.64
(1)-(4)	4.72	4.84
mean	4.63	4.67

We have not yet considered full conversion of all prosodic characteristics but for these experiments, the fundamental frequency (F_0) was also included as a factor in the GMM conversion method. The feature is ln F_0 and the number of the GMM classes were 2.

5. OBJECTIVE EVALUATION ON VOICE DIFFERENCES

5.1. Evaluation Measurement

We employed Mel cepstrum distortion (MelCD) for determining acoustic distances between converted and target speech defined as follows:

$$MelCD = \frac{20}{ln10}\sqrt{2\sum_{i=1}^{40}(mc_i^{(conv)} - mc_i^{(tar)})^2}, \qquad (1)$$

where $mc_i^{(conv)}$ and $mc_i^{(tar)}$ denote the MelCD coefficients of the converted voice and the target voice, respectively. In this study, these coefficients are calculated from STRAIGHT spectra. As the MelCD measure decreases we can infer that the mapping between source and target voice qualities is improving. The distance between unmodified source and target speakers' voices is taken as a baseline maximal distance.

5.2. Within Speaker Acoustic Distances

For a guide to the minimum expectable distance, we measured differences in the speech of one speaker over different periods of time. In order to find out how much the voice of a given speaker can change with different readings over time. We quantified this variability in the short time to determine a baseline minimum distance, the voice of speaker Fe-A reading the same set of 10 evaluation sentences 4 times was recorded, as shown below:

(1) recording the first set of 10 sentences
(2) re-recording after a one-minute interval
(3) re-recording after a ten-minute interval
(4) re-recording after a sixty-minute interval

MelCD distances calculated between each pair of readings are given in Table 1. It can be seen that the smallest value obtained is about 4.5 dB. This represents the differences in the long-term averaged cepstrum between readings of identical sentences from the same speaker, and can be taken as a minimum baseline for measuring the performance of the voice conversion metric.

5.3. Making Conversion Sets

To investigate the differences between voice conversion across different language pairs, we compared the performance of mapping functions trained on both Japanese (J) and English (E) data sets of 50 training sentences each. Comparisons were made for both male and female speakers, after conversion by mapping the voice of speaker 1 to that of speaker 2 for each of the 10 evaluation

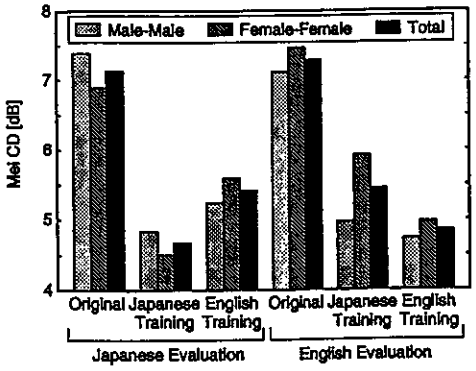

Fig. 2. Results of Mel cepstrum distortion

sentences. The Mel scaled cepstrum distance functions were calculated using the long-term averaged cepstrum of the combined evaluation sentence readings after voice-conversion. The following combinations were tested: (1) E sentences voice-mapped using weights trained on J data. (2) E sentences voice-mapped using weights trained on E data. (3) J sentences voice-mapped using weights trained on E data. (4) J sentences voice-mapped using weights trained on J data.

5.4. Results

Figure 2 shows the values obtained for each of the above four conditions as well as the distances obtained by comparing the voices of the original speakers before the voice-conversion technique was applied. The black bars show the averaged results for both male and female speakers combined. No attempt was made to map between the voice of a male speaker and that of a female speaker (or vice versa) in these experiments.

It can be seen that the acoustic distances between the voice-converted source-speaker's speech and that of the conversion-target speaker decrease in both cross- and single-language voice conversion. In single-language cases, the mean values are around 4.5 to 5.0 dB. These values almost as low as the distortion measured within-speaker, shown in Table 1. This indicates that individual voice differences are significantly reduced. The MelCD for the cross-language conditions are not as close as those of the single-language cases. However, to conclude that voice difference reduction is successful, when compared with the original distances and even when the mapping function training and target languages are different. The differences between same-language and cross-language training results can be taken to show the effects of mapping across languages.

6. PERCEPTUAL EVALUATION ON VOICE DIFFERENCES

6.1. Making Conversion Sets

To confirm the reliability of the objective experiments, we performed a perceptual evaluation of the performance of the cross-language voice-conversion algorithm. For a method of making converted speeches, the algorithm of mixing the GMM-based converted spectrum and the Dynamic Frequency Warping (DFW)-based converted spectrum[7] were employed to enables to produce

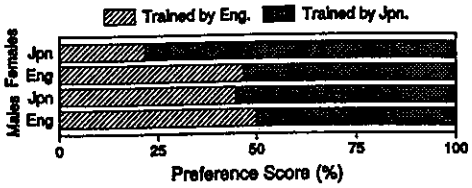

Fig. 3. ABX test results

better quality sounds which need for the listening tests, since the GMM-based converted spectra were over-smoothed than the original target speaker's spectra. Here, the converted speeches keep source speaker's prosodic features except F_0 at first.

6.2. Experimental details

ABX listening tests were conducted for evaluating the closeness of the mapping from the source speaker's voice to that of the target speaker. Here, X is the original unmodified speech of the target speaker, and A or B are the two versions of each voice-converted utterance, using within-language and between-language trained models. 3 sentences were selected from the 10 evaluation utterances in both languages. The utterances were produced by both speakers of each sex. This resulted in a set of 24 evaluation sentences. These sentences were randomised and presented in turn to 8 listeners in a soundproof room, comprising 4 females and 4 males. Order of presentation (X-A,B/X-B,A) was also randomised.

6.3. Results

Figure 3 shows the results. It can be seen from the figure that there is remarkably little difference between the training methodologies. Responses from the listeners were close to 50% in all but one case, indicating that both same-language and cross-language training performed equally well. However, in the case of the female speaker's voice conversion, there was a marked preference for sentences of Japanese with models trained on Japanese utterances. It can be considered that one reason for this difference came from prosodic differences in the utterances of the two female speakers, whose F_0 level and overall speaking rate (and perhaps segmental timing relationships) were indeed different. Since A and B have the source speaker's prosody and X has the target's prosody, participants may have been affected by the prosodic differences. This issue is left for future work.

7. INVESTIGATION ON A POSSIBILITY FOR LANGUAGE TRAINING SYSTEM

7.1. Experimental Details

For the first step towards a language learning system, we investigated the potential of using voice conversion for reducing voice differences when performing pronunciation evaluation. The correlations between a computed distance measure and a human perceptual pronunciation evaluation score were compared per vowel before and after applying voice conversion. As reported the section above, voice differences were reduced even in different languages. Therefore, the direct comparison of the utterances between Teacher (T) and Student (St) after applying conversion must

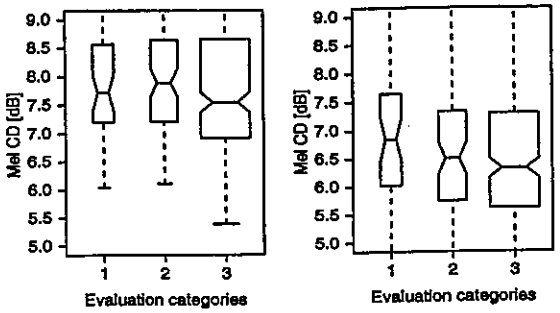

Fig. 4. Relations between MelCD and perceptual score, left: without voice conversion, right: with voice conversion

be possible. As a distance measure for evaluating the students pronunciation of each phoneme, the MelCD was employed.

Each evaluated English phoneme was labeled and extracted from utterances of the evaluation sentences. The bilingual speaker Fe-A was selected as a T. For the training data, 50 Japanese sentences were selected, and for the evaluation data, 34 English sentences were selected from the recorded utterances. Both phoneme labels datasets of T and St were created by hand labeling from 34 English evaluation sentences. The total phoneme number of those sentences were 950, including 346 vowels. To investigate the agreement between the proposed pronunciation distance measurement and human perception, a perceptual pronunciation score was assigned to each St phoneme. Each phoneme was classified into one of these 3 levels as follows, score1 : Mispronunciation including difficult phoneme for Japanese, score2 : Not a mistake, but not the same as the teacher's pronunciation and score3 : The same or close to the teacher's pronunciation. The number of the vowels in each category were as follows, score1: 45, score2: 57, score3: 244.

7.2. Results

We assume that when applying cross-language voice conversion, the relations of MelCD values will be clearer such as score1 > score2 > score3 as those values must directly show spectral differences caused by mispronunciation after reducing voice differences of two speakers.

Relations between MelCD and perceptual score without voice conversion and with voice conversion are shown as boxplots in Figure 4. From these results, it can be seen that there are no clear correlations between MelCD and perceptual score before applying voice conversion. However, after the conversion, the expected correlation appears especially between score1 and score3. To estimate if the difference in the score1 and score3 is significant, a t-test was carried out. The results are shown in Table 2. The t values before and after applying voice conversion are $t(59.9) = 1.06$ and $t(64.4) = 2.13$, respectively. While there is no significant difference before applying voice conversion, the difference is significant at $p < 0.05$ after the conversion. This indicates that the proposed method has a potential for using in a language training system to reduce voice differences. It remains as future work to elaborate the relationship between the other pairs of scores.

Table 2. The results of the t-test

Voice conversion	t value	Significance
Not applied	1.064	Not signif.
Applied	2.133	$p < 0.05$

8. CONCLUSIONS

We have presented the effectiveness of voice conversion across different language pairs using bilingual-language speaker's speech. Results showed that the conversion between cross-language was also effective to reduce voice differences, though it does not reach the quality of single-language voice conversion. Next, the potential of applying voice conversion for reducing voice differences and detecting pronunciation error was tested using vowels included in Student's English evaluation sentences. Although the possibility was found for applying cross-language voice conversion to a language training system, further study is needed for precise confirmation. More investigations using much data are now on going.

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

9. 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] Y. Stylianou and O. Cappé: A system for voice conversion based on probabilistic classification and a harmonic plus noise model, *Proc. ICASSP*, Seattle, U.S.A., pp. 281–284 (1998).

[4] T. Toda, J. Lu, H. Saruwatari and K. Shikano: Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum, *Proc. ICSLP*, Beijing, China, pp. 279–282 (2000).

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

[6] 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).

[7] T. Toda, H. Saruwatari and K. Shikano: High Quality Voice Conversion Based on Gaussian Mixture Model with Dynamic Frequency Warping, *Proc. EUROSPEECH*, Aalborg, Denmark, pp. 349–352 (2001).

[8] A. Kain and M. W. Macon: Spectral voice conversion for text-to-speech synthesis, *Proc. ICASSP*, Seattle, U.S.A., pp. 285–288 (1998).