Regular Paper

Cross-language Voice Conversion Evaluation Using Bilingual Databases

Mikiko Mashimo,[†] Tomoki Toda,[†] Hiromichi Kawanami,[†]
Kiyohiro Shikano[†] and Nick Campbell[†,††]

This paper describes experiments that test an extension of techniques for converting the voice of one speaker to sound like that of another speaker, to include cross-language utterances, such as would be required for spoken language translation or language training applications. In particular, it addresses the issue of evaluation of system performance, and compares objective tests using a perceptually-motivated acoustic measure, with perceptual tests of voice quality and speaker resemblance. The proposed method uses Japanese and English speech databases from 2 female and 2 male bilingual speakers for training in a system based on a Gaussian mixture model (GMM) and a high quality vocoder. Results indicate that training with cross-language models also produces close acoustic matches between source and target speakers' voices. Perceptual tests revealed little significant difference in the performance of mapping functions trained on single-language and cross-language data pairs.

1. Introduction

Speaker 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. This paper describes a method to map from the spectrum of one speaker's voice to that of another so that spectral produced can be changed to sound more like that of the target speaker. It presents results of experiments performed to evaluate the effectiveness of reducing speaker differences in the spectral domain both within a given language and across languages. The method could have applications in automatic speech translation, where the voice of a source-language speaker is used for synthesis in the target language, or in foreign-language training, where it could be used to facilitate automatic evaluation of a student's performance by direct comparison with the speech of the tutor after it has been modified to match the voice of the student. Our goal is to determine the role of spectral information when converting to the target speaker's voice characteristics. Future work include modelling the dynamic speaking style.

The technique that we present here employs signal processing techniques to map between the voices of two people, and the evaluations that were performed test the extent to which language differences affect its performance. Previous work[1] has described how the voice of a reference speaker can be used in concatenative speech synthesis to produce utterances in languages other than that of the original speech database. However, if successful, the present method has the advantage of requiring less data because it makes use of voice-conversion techniques to allow mapping of any target speaker's voice after training with only a small number of source speaker's utterances.

Our proposed method extends the work by Abe, et al.[2),3] in the late 1980's using a codebook mapping method for voice conversion. Their method used a discrete representation of the acoustic features that contribute to speaker individuality, mapping between code vectors by minimizing the acoustic distortion between the pairs. We employ a voice conversion algorithm based on the Gaussian Mixture Model (GMM) proposed by Stylianou, et al.[4] to model the acoustic space of a speaker continuously, and claim that this technique has advantages over the discrete codebook mapping methods. Toda, et al.[5] reported an application of this voice conversion method, in conjunction with a high-quality vocoder STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum)[6),7], 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[8].

[†] Graduate School of Information Science, Nara Institute of Science and Technology
[††] ATR/CREST

Ideally, the quality of the output speech should sound as if the target speaker had spoken the other language and the speaker's 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. In the study presented by Abe, et al., bilingual datasets were not used for evaluating these acoustic distances. We collected Japanese-English bilingual data from the speech of 2 females and 2 males for an investigation into whether differences in the languages have an effect on the quality of the converted voice.

This paper is organized as follows: In Section 2, an overview is given of cross-language voice conversion. In Section 3, details of the proposed method are explained. Experimental techniques are presented in Section 4, results given in Section 5, and discussion of these results in Section 6. Section 7 is a conclusion and outlines for future work.

2. Cross-language Voice Conversion with Bilingual Databases

The backgrounds and the procedures of cross-language voice conversion are discussed in this section. The reasons of using bilingual databases and details of those databases are also introduced.

2.1 Cross-language Voice Conversion

Abe, et al.[3] performed a statistical analysis of spectrum differences between Japanese and English, and they reported some instructive results. Although there were some code vectors that predominate in English or Japanese, there were no perceptual differences between English coded by the English codebook and coded by the Japanese codebook. Another fact was shown that the code vectors' overlap in different language pairs was remarkable than overlap in different speakers. From these points of view, we assume that voice differences between the speakers are more important than the acoustic differences between the languages when performing cross-language voice conversion. In our work, as we used GMM based conversion algorithm, it was possible to represent acoustic spaces more continuously than codebook mapping methods.

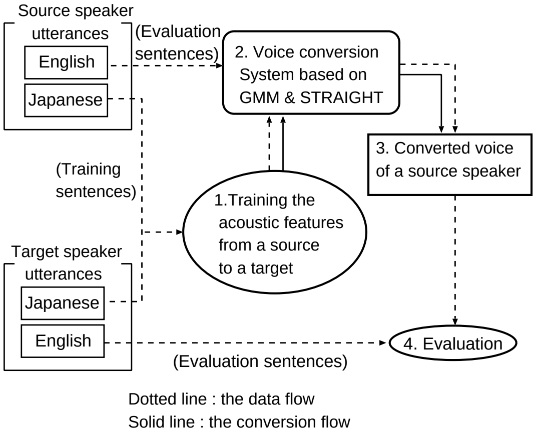
Source speaker utterances
English
Japanese
(Evaluation sentences)
2. Voice conversion System based on GMM & STRAIGHT
3. Converted voice of a source speaker
(Training sentences)
1.Training the acoustic features from a source to a target
Target speaker utterances
Japanese
English
4. Evaluation
(Evaluation sentences)
Dotted line : the data flow
Solid line : the conversion flow

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

To evaluate the conversion success, Abe, et al. used an English speech synthesizer (MITalk) to generate Japanese speech, with conversion of the voice, and measures were made of the acoustic distance between target utterances and voice-modified synthetic speech. Perhaps because of the distortion present in the original synthesised speech, the cross-language voice-conversion was not found to be as effective as single-language conversion. We therefore decided to test the efficiency of our cross-language voice-conversion technique using utterances of builingual speech. By using bilingual speakers, we can be assured of a natural, native speaker utterances both within and across language pairs. The use of bilingual datasets for both source and target speaker's databases allows detailed evaluation for of voice conversion. **Figure 1** shows the flow of processing. Two sets of bilingual data were recorded from native Japanese speakers, and training was performed first using Japanese speech to develop the mapping weights 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.

The procedure is as follows:

(1) Train acoustic mapping functions between speakers (using language A training sentences) to obtain the mapping function,

(2) Apply the mapping function to the source speaker's utterances (using language B evaluation sentences),

(3) Obtain Language B utterances which

Table 1 Mean fundamental frequency and standard deviation of 50 training sentences (Hz).

	English		Japanese	
	mean	std	mean	std
F1	270.0	43.4	248.6	41.2
F2	227.6	41.2	233.9	34.6
M1	114.5	18.0	112.9	15.5
M2	90.8	13.9	89.8	12.3

have the target speaker's spectra and prosodic characteristics,

(4) Evaluate the quality of the voice conversion cross- and single-languages, using both acoustic and perceptual measures.

2.2 Experimental Speech Databases

Utterances of four native Japanese speakers with bilingual language skills (2 females, 2 males) were recorded in a sound-treated room with studio-quality equipment and their speech was digitised using 48 kHz, 16 bit sampling. Since the last target of our voice conversion system is for Japanese to English speech, we selected bilingual speakers who are fluent in these two languages. Speaker F1 had learnt English from an American native speaker from the age of 3-yrs, and speaker F2 has more than 12 years experience of living in an English speaking country. Speaker M1 has a Japanese mother and an English father, and speaker M2 is a specialist in British English phonology.

The speakers each read 60 sentences in each language. After down-sampling to 16 kHz, 50 sentences were used for training data and 10 sentences were used for testing. The mean fundamental-frequency (F_0) and its standard deviation for the 50 training sentences are given in **Table 1**. The male speakers showed little difference in F_0 between Japanese and English. On the contrary, the F_0 for the female speakers differed considerably, even within the readings of the same speaker. Speaker F1 uses a higher pitch for Japanese than for English (21.4 Hz difference on average), while speaker F2 raises her pitch when speaking English (by about 6 Hz).

3. Voice Conversion Method

To carry out the voice conversion, we employed the recently reported method[4),9)] explained below.

In this method, p-dimensional time-alignment of the acoustic features of a source speaker and a target speaker is performed. Dynamic Time Warping (DTW) is used:

source speaker : $\boldsymbol{x} = [x_0,\ x_1,\ \ldots,\ x_{p-1}]^T$,

target speaker : $\boldsymbol{y} = [y_0,\ y_1,\ \ldots,\ y_{p-1}]^T$,

where T denotes transposition.

3.1 Gaussian Mixture Model

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

$$p(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \qquad (1)$$

$$\sum_{i=1}^{m} \alpha_i = 1,\ \ \alpha_i \geq 0, \qquad (2)$$

where $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. α_i denotes a weight of class i, and m denotes the total number of Gaussian mixtures. The merit of employing this model for the voice conversion method is that the acoustic space can be described continuously in GMM, resulting in a reduction of quantisation distortion.

3.2 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:

$$\begin{aligned} F(\boldsymbol{x}) &= E[\boldsymbol{y}|\boldsymbol{x}] \\ &= \sum_{i=1}^{m} h_i(\boldsymbol{x}) E_i[\boldsymbol{y}|\boldsymbol{x}], \end{aligned} \qquad (3)$$

$$E_i[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx} (\boldsymbol{\Sigma}_i^{xx})^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i^x), \quad (4)$$

$$h_i(\boldsymbol{x}) = \frac{\alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{m} \alpha_j N(\boldsymbol{x}; \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})} \qquad (5)$$

where $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$ denote mean vectors of class i for the source and target speakers. $\boldsymbol{\Sigma}_i^{xx}$ is covariance matrix of class i for the source speaker. $\boldsymbol{\Sigma}_i^{yx}$ is the cross-covariance matrix of class i for the source and target speakers. In this paper, these matrices are set to be diagonal. By introducing this mapping function in voice conversion, acoustic features can map continuously using correlation between source and target speakers.

3.3 The Mapping Function Training

To estimate parameters such as $\alpha_i, \boldsymbol{\mu}_i^x, \boldsymbol{\mu}_i^y, \boldsymbol{\Sigma}_i^{xx}, \boldsymbol{\Sigma}_i^{yx}$, the probability distribution of the joint vectors $\boldsymbol{z} = [\boldsymbol{x}^T,\ \boldsymbol{y}^T]^T$ for the source and target speakers is represented by the GMM whose parameters are trained by joint density distribution[9)]. Covariance matrix $\boldsymbol{\Sigma}_i^z$ and mean vector $\boldsymbol{\mu}_i^z$ of class i for joint vectors can be written as

$$\boldsymbol{\Sigma}_i^z = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix}, \quad (6)$$

$$\boldsymbol{\mu}_i^z = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix}. \quad (7)$$

Expectation Maximization (EM) algorithm is used for estimating these parameters.

3.4 Conversion System

The GMM-based voice conversion algorithm has been implemented in STRAIGHT by Toda, et al.[5] 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[6],[7]. 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 keeping with 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. As the waveform power, that of the source speaker was used.

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.

To evaluate the quality of the voice conversion methods, a conversion accuracy evaluation test and two perceptual evaluation tests were carried out. In the following sections, experimental evaluations and the discussions are reported.

4. Voice Conversion Evaluation

On the assumption that a measure of voice-conversion accuracy should be as close to that of human perception as possible, we employed Mel scaling in our cepstral-distance calculations, that is, Mel cepstrum distortion (MelCD).

4.1 Evaluation Measurement

The MelCD we employed for determining acoustic distances between converted and target speech is defined as follows:

$$MelCD = \frac{10}{ln10}\sqrt{2\sum_{i=1}^{40}(mc_i^{(conv)} - mc_i^{(tar)})^2}, \quad (8)$$

where $mc_i^{(conv)}$ and $mc_i^{(tar)}$ denote the long-term averaged 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. For a guide to the minimum expectable distance, we measured differences in the speech of one speaker over different periods of time, as detailed in the next subsection.

4.2 Acoustic Distances of the Same Speaker

It is well known that the voice of a given speaker can change over time. In an attempt to quantify this variability in the short time and determine a baseline minimum distance, we recorded the voice of speaker F1 reading the same set of 10 evaluation sentences 5 times. The first pair of readings differed in time by only a one-minute interval. The last pair differed by an hour, 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 thirty-minute interval,
(5) re-recording after a sixty-minute interval.

Analysis parameters are as given in **Table 2**. MelCD distances calculated between each pair of readings are given in **Table 3**. It can be seen from Table 3 that the smallest value obtained is about 2.2 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. It will also be noticed

Table 2 Analysis parameters.

Analysis window	Compensated gaussian
Sampling frequency	16 kHz
Shift length (training)	5 ms
Number of FFT points	1024
Number of the GMM class	64
Training sentences	50

Table 3 MelCD values between the same speaker, F1 (dB).

	Japanese	English
(1)-(2)	2.32	2.27
(1)-(3)	2.26	2.32
(1)-(4)	3.32	2.44
(1)-(5)	2.36	2.42
mean	2.53	2.36

that a difference of 3.3 dB is found between one pair of readings. The speaker took a nap between these readings and the difference in voice quality upon waking is seen in this measure.

4.3 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 and English data sets of 50 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 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) English sentences voice-mapped using weights trained on Japanese data.

(2) English sentences voice-mapped using weights trained on English data.

(3) Japanese sentences voice-mapped using weights trained on English data.

(4) Japanese sentences voice-mapped using weights trained on Japanese data.

4.4 Conversion Accuracy Evaluation of Cross-language voice conversion

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 2.2 to 2.5 dB. These values almost as low as the distortion measured within-speaker, shown in Table 3. 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. It was also noted that MelCD values were larger for the female results. One cause may be the F_0 differences found even in the same female speaker in Japanese and English.

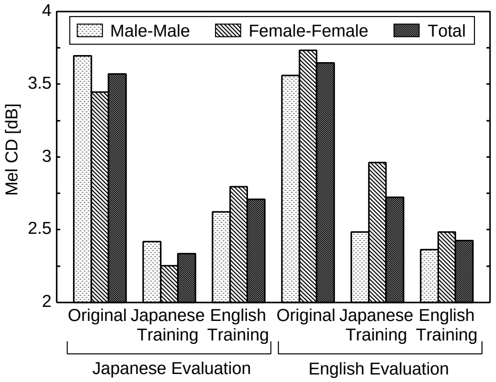

Fig. 2 Results of Mel cepstrum distortion.

5. Subjective Evaluation

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[8] were employed to enables to produce 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. Two kinds of listening tests were performed; one to evaluate the closeness of the mapping from the source speaker's voice to that of the target speaker, and one to evaluate the distortion in voice quality produced by the conversion process. **Table 4** shows the experimental conditions. 3 sentences were selected from the 10 evaluation

Table 4 Conditions of subjective experiments.

Room	Sound treated room
Participants	8 (4 females and 4 males)
Experiment	AB(X)
Presentation style	Random

utterances in both languages. The utterances were produced by both speakers of each sex. This resulted in a set of 24 evaluation sentences. These were randomised and presented to 8 listeners, comprising 4 females and 4 males.

5.1 Conversion Accuracy Evaluation

ABX listening tests were conducted for evaluation of subjective conversion accuracy. Listeners heard speech in turn as follows.

- X: the original unmodified speech of the target speaker
- A or B: the two versions of each voice-converted utterance, using within-language and between-language trained models

Listeners were asked to judge which (A or B) of the converted voices sounded most like the target speaker (X). Responses were marked by circling A or B on the response sheet for each test utterance. Order of presentation (X-A,B/X-B,A) was also randomised.

Figure 3 shows the results. It can be seen from the figure that there is no remarkable difference between the training methodologies. Responses from the listeners were close to 50% in all but one case, indicating that both single-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. We consider 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 discussed in the Section 5.3.

5.2 Sound Quality Evaluation

AB listening tests were also conducted for subjective sound quality evaluation. The same 8 listeners were presented with pairs of voice-converted utterances, produced as above and using the same materials, to further evaluate differences between single-language and cross-language trained models. The listeners were asked to judge which sounded more natural (A or B). In this case, naturalness was defined as meaning less noisy and keeping clarity. The test was designed to evaluate the distortion produced by each of the training procedures. Listeners were required to circle either A or B on the response sheet to indicate the sample with higher naturalness in each pair.

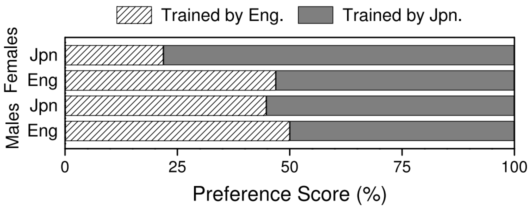

Fig. 3 ABX test results (Conversion accuracy).

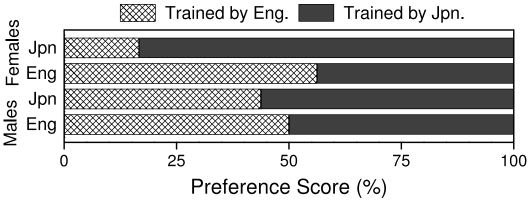

Fig. 4 AB test results (Sound quality).

Figure 4 shows the results. It can be seen that results were also centred well around the 50% mark, indicating that neither method of training the mapping weights produced considerably more distortion than the other, except for the female speaker's Japanese utterances. In the ABX test presented above, it was found that the listeners showed a marked preference for the voice conversion using models trained on Japanese utterances when the test sentences were spoken in Japanese. In the AB listening test presented here, the listeners also judged that the voice quality of the female speaker for Japanese sentences was better when the models used in the voice-conversion process were trained using utterances spoken in the same language. In the following section, we investigated the possible reasons for these results.

5.3 Repeat Evaluation

To determine whether the results of female Japanese sentences were caused by prosody (we saw in the last part of the Section 5.1), the target speaker's utterances were modified to have the source speaker's prosody, as A and B keep source speaker's prosodic information. A repeat conversion accuracy (ABX) and a sound quality (AB) evaluations were performed using these

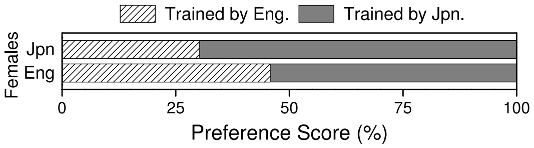

Fig. 5 ABX repeat test results (Conversion accuracy).

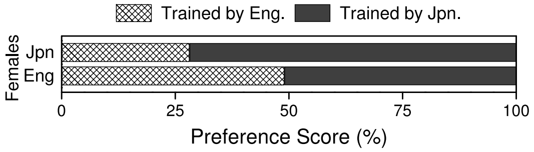

Fig. 6 AB repeat test results (Sound quality).

modified X utterances. When performing ABX repeat tests presenting the target-speech data modified to have source speaker's prosodic feature as a subsequent ABX evaluation, we also presented the same test data at the same time. The presentation order for an AB evaluation of voice naturalness was also randomised.

The results of this further comparison are presented in **Figs. 5** and **6**. Both results showed about 10% improvements in a preference scores for Japanese sentences trained by English. Although it was found to improve the score by modifying prosodic information of the target speaker, the results still kept the same tendency as Figs. 3 and 4. This was not an enough amelioration as we expected as compared with the scores in the English sentences. These results indicated that the other factors had effects when producing these converted voice. Further investigations are left to future work.

6. Conclusion

We confirmed that the effectiveness of cross-language voice conversion extended from single-language voice conversion technique based on GMM and STRAIGHT. Since cross-language voice conversion is normally considered not to be successful because of using mapping function trained by other language, the confirmation is an important step for trying cross-language conversion. To perform and to evaluate the conversion, the bilingual databases were prepared for the source and the target speaker.

Objective measures of performance using a Mel-scaled cepstral distance function on long-term averaged data show that the mapping functions succeed in reducing the distance between the source and target speaker's voice individuality differences. These results all show slightly better MelCD values and perceptual scores in single-language voice conversion, although cross-language voice conversion showed similar trends.

Perceptual tests showed that, in the majority of cases, listeners found little significant difference in the performance of mapping functions trained on single-language and cross-language data pairs. However, for the female speaker's Japanese sentences, the cross-language voice conversion seemed less successful than other 3 types of results. From the repeated ABX and AB test results, we got 10% improvement for reducing prosodic effects.

Future work include inquiring the results of perceptual tests and recording more bilingual data to test for the more detailed experiments.

Acknowledgments The authors thank JST/CREST (Core Research for Evolutional Science and Technology) in Japan for partly supporting this work.

References

1) Campbell, N.: TALKING FOREIGN Concatenative Speech Synthesis and the Language Barrier, *Proc. EUROSPEECH*, Aalborg, Denmark, pp.337–340 (2001).

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

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

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

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

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

7) Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: 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 Communica-*

tion, Vol.27, No.3–4, pp.187–207 (1999).

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

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

(Received November 20, 2001)
(Accepted April 16, 2002)

Mikiko Mashimo recieved the M.S. degree from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST) in 2002. Her current position is a Ph.D. candidate at NAIST. Her research interests include speech analysis, speech synthesis and computer assisted language learning. She is a member of the Acoustical Society of Japan (ASJ).

Tomoki Toda received the B.S. degree from department of electrical engineering, Nagoya University in 1999. He received the M.S. degree in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST) and he is currently a Ph.D. candidate in NAIST. He is an intern researcher at ATR Spoken Language Translation Research Laboratories from 2001. His research topics include speech analysis and speech synthesis. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) and the Acoustical Society of Japan (ASJ).

Hiromichi Kawanami was born in 1969. He received the B.E. degree in electrical engineering in 1994, and received the M.E. and Ph.D. degrees in information and communication engineering in 1997 and 2000, respectively, from the University of Tokyo. During 2000–2001, he joined the Electrotechnical Laboratory (National Institute of Advanced Industrial Science and Technology at present) as a researcher. From 2001 he is an assistant professor of Nara Institute of Science and Technology (NAIST). His research interests are speech analysis and speech synthesis. He is a member of the Institute of Electronics, Information and Communication Engineers and the Acoustical Society of Japan.

Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a professor of Nara Institute of Science and Technology (NAIST), where he is directing speech and acoustics laboratory. His major research areas are speech recognition, multi-modal dialog system, speech enhancement, adaptive microphone array, and acoustic field reproduction. From 1972 to 1993, he had been working at NTT Laboratories, where he had been engaged in speech recognition research. During 1986–1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories. During 1984–1986, he was a visiting scientist in Carnegie Mellon University. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990, Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, IPSJ Yamashita SIG Research Award in 2000, and Paper Award from the Virtual Reality Society of Japan in 2001. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE), Information Processing Society of Japan, the Acoustical Society of Japan (ASJ), Japan VR Society, the Institute of Electrical and Electronics Engineers (IEEE), and International Speech Communication Society.

Nick Campbell has a Ph.D. in Experimental Psychology and is currently Research Director of the JST/CREST Expressive Speech Processing Project. He is a Project Leader in Dept 1 of the ATR Human Information Science Laboratories and a Visiting Professor at NAIST, with a chair in Applied Linguistics. He has been a Guest Researcher at the AT&T Bell Laboratories, and a research Fellow at the IBM UK Scientific Centre. His research areas include speech synthesis, speech databases, and the processing of paralinguistic information from the speech signal. His current interests include speech & emotion, and conversational speech synthesis.