TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 13/22 pp510-514
Volume 13, Number 4, August 2008

IBM Voice Conversion Systems for 2007 TC-STAR Evaluation

SHUANG Zhiwei (双志伟)**, Raimo Bakis†, QIN Yong (秦 勇)

IBM China Research Lab, Beijing 100084, China;
† IBM Watson Research Center, Yorktown Height, NY 10598, USA

Abstract: This paper proposes a novel voice conversion method by frequency warping. The frequency warping function is generated based on mapping formants of the source speaker and the target speaker. In addition to frequency warping, fundamental frequency adjustment, spectral envelope equalization, breathiness addition, and duration modification are also used to improve the similarity to the target speaker. The proposed voice conversion method needs only a very small amount of training data for generating the warping function, thereby greatly facilitating its application. Systems based on the proposed method were used for the 2007 TC-STAR intra-lingual voice conversion evaluation for English and Spanish and a cross-lingual voice conversion evaluation for Spanish. The evaluation results show that the proposed method can achieve a much better quality of converted speech than other methods as well as a good balance between quality and similarity. The IBM1 system was ranked No. 1 for English evaluation and No. 2 for Spanish evaluation. Evaluation results also show that the proposed method is a convenient and competitive method for cross-lingual voice conversion tasks.

Key words: voice conversion; frequency warping; mapping formants

Introduction

Voice conversion involves conversion of a source speaker's voice to make it sound like a target speaker's voice. There are many applications for voice conversion. An important application is in the construction of customized text-to-speech (TTS) system for different companies, in which systems using each company's favorite voice can be created quickly and inexpensively by modifying the original speaker speech corpus. Voice conversion can also generate voices of special characters in the movie industry or for preserving the speaker's identity in speech-to-speech translation. There are always two criteria for evaluating the performance of voice conversion: the speech quality and similarity to the target speaker. Even with state of the art voice conversion technology, there is always a tradeoff between quality and similarity. A mean combined quality score and similarity score is used by recent worldwide evaluations, where a high mean score means a higher overall system performance.

Spectral conversion is the key component in a voice conversion system. The most popular two spectral conversion methods are based on codebook mapping[1,2] and the Gaussian mixture model[3,4]. However, though both methods have been improved recently, the quality degradation introduced is still severe[5,6]. In comparison, another spectral conversion method, namely frequency warping, introduces less quality degradation[7-9] and is chosen for our spectral conversion. We propose a novel method of generating a frequency warping function by mapping formant parameters of the source speaker and the target speaker[10]. Alignment and selection processes are added to ensure that the

Received: 2007-09-10

** To whom correspondence should be addressed.
E-mail: shuangzw@cn.ibm.com; Tel: 86-10-58748097

selected mapping formants can represent different speakers well. This approach requires only a very small amount of training data for generating the warping function, thereby greatly facilitating its application.

1 Voice Conversion System Overview

The diagram of our voice conversion system is shown in Fig. 1.

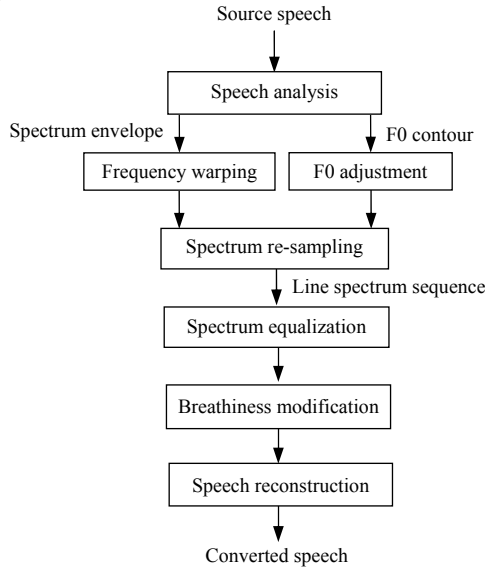

Fig. 1 Diagram of voice conversion system

We use the analysis/reconstruction technique proposed by Chazan et al.[11] to get an enhanced complex envelope model and pitch contour. The technique allows both amplitude and phase manipulations, such as F0 modification, frequency warping, and spectral smoothing.

Here, we use frequency warping to stretch/compress the spectrum envelope along the frequency axis. Meanwhile, we use F0 adjustment to transform the average and variance values of log F0 of the source speaker to those of the target speaker. The warped spectrum envelope is then re-sampled in multiplies of converted F0 to yield a new complex line spectrum. After this, we apply a filter to the spectrum to compensate for the different energy distribution along the frequency axis. We can also add breathiness by adding random numbers to the phase of the complex line spectrum if needed. Finally, the converted speech can be reconstructed from the line spectrum sequence. Duration modification can be achieved by repeating or deleting some frames of line spectrum during the reconstruction.

2 Frequency Warping Method

We propose here a novel method to generate a frequency warping function based on mapping formant parameters of selected aligned frames. Alignment and selection processes are added to ensure that the selected mapping formants can represent differences between speakers well. First, we align the source speech data and target speech data to ensure that they are in similar context. Next, we make a selection from the aligned frames. The selection can be made by combining experimental rules and statistical analysis. For the experimental rules, some phonemes such as “e” in English, and some contexts such as “plosive” as neighboring phonemes are preferred because they are found experimentally to be more stable. For the statistical analysis, segments with stable formants are preferred. The predefined maximal variation ranges for the first three formant frequencies are set to 30 Hz, 50 Hz, and 100 Hz, and the minimum length of the stable segment is set to 40 ms.

The formants of the selected aligned frames are then used as the key positions to generate the frequency warping function. For example, we get one pair from mapping the first four formant frequencies [690, 2290, 3080, 4450] of the source speaker and [560, 2180, 2750, 4040] of the target speaker. The warping function from the target frequency axis to the source frequency axis for a maximum frequency for both speaker of 8000 Hz is shown in Fig. 2.

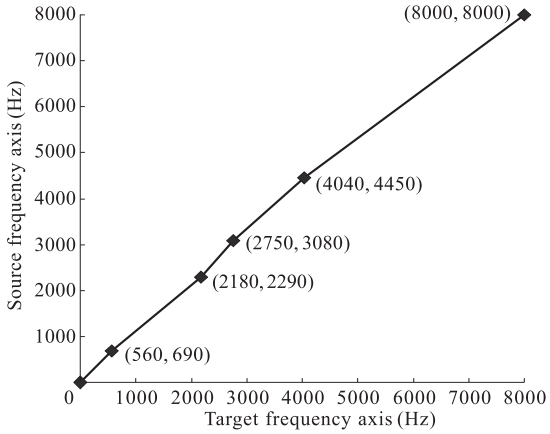

Fig. 2 Frequency warping function generation

Finally, we use the generated frequency warping function to perform frequency warping of the spectrum. Suppose that one frame of the source speaker spectrum

is $S(w)$ and that the frequency warping function from the target frequency axis to the source frequency axis is $F(w)$. The converted spectrum $\mathrm{Conv}(w)$ is then given by

$$\mathrm{Conv}(w) = S(F(w)) \quad (1)$$

The same frequency warping function is applied for all frames. This strategy can avoid discontinuity problems in applying different warping functions to different frames. This strategy also does not require the alignment information of input speech data, which makes it applicable for various usage scenarios.

3 TC-STAR Voice Conversion Evaluation

The TC-STAR project, financed by the European Commission within the Sixth Program, is envisaged as a long-term effort to advance research in all core technologies for speech-to-speech translation (SST). SST technology is a combination of automatic speech recognition (ASR), spoken language translation (SLT), and TTS (speech synthesis). The objectives of TC-STAR project are ambitious: namely to make a breakthrough in SST that significantly reduces the gap between human and machine translation performance.

Since the TC-STAR project aims at translating speech from one language to another, it is important to assess how "close" the translated voice is to the original voice. Voice conversion evaluation is therefore one of a series of evaluations hosted by the TC-STAR project.

3.1 Evaluation data

In 2007, the TC-STAR project hosted a voice conversion evaluation for English and Spanish. The training data are taken from 4 speakers noted as 75(F), 76(F), 79(M), and 80(M), where (F) denotes a female speaker and (M) denotes a male speaker. In total 126 sentences are provided as training data for English, while 154 sentences are provided for Spanish. However, for the selection of mapping formants and calculation of the F0 coefficients, only the first 30 training sentences of each speaker are used.

3.1.1 Selected mapping formants

We select formants in the middle of the phoneme "e:" in the syllable "Heard" of the No. 22 training sentence as mapping formants (Table 1). The first 4 formants are denoted as F1, F2, F3, and F4.

Table 1 Mapping formants of 4 speakers

Speaker	F1	F2	F3	F4
75(F)	717	1762	3031	4162
76(F)	727	1617	2970	4073
79(M)	585	1617	2533	3651
80(M)	593	1464	2530	3767

3.1.2 Other conversion settings

A fundamental-frequency adjustment and a spectral-envelope equalization are applied for each conversion. We add breathiness in the conversion from 75(F) to 76(F). Duration modification is applied for some conversions. Table 2 gives the duration modification ratios for each conversion.

Table 2 Duration modification ratios

Source/target	Duration modification ratio
75(F)/76(F)	1.05
75(F)/79(M)	1.00
80(M)/76(F)	0.90
80(M)/79(M)	1.00

3.2 Evaluation criteria

3.2.1 TC-STAR similarity evaluation (VC1)

In this evaluation, listeners are asked to rate if a given voice pair comes or not from the same person according to following scale: (5) definitely identical, (4) probably identical, (3) not sure, (2) probably different, and (1) definitely different. The arithmetic mean of all subjects' individual scores of all subjects is used as the evaluation result.

3.2.2 TC-STAR quality evaluation (VC2)

In this evaluation, listeners are asked to assess certain sentences according to the following scale: (1) bad, (2) poor, (3) fair, (4) good, and (5) excellent. The mean opinion score (MOS) is the arithmetic mean of all the scores from each individual.

3.3 Subjective test settings

Subjective tests were carried out via the web. Access to a high-speed/ADSL internet connection and good listening material were required. A total number of 20 judges were recruited and paid to perform the subjective tests. The judges were aged 18 to 40 years old and were native English speakers with no known hearing problems. None were speech synthesis experts.

4 Results and Discussions

We submitted two systems for evaluation, noted in the following as IBM1 and IBM2. Both are based on our proposed frequency warping method. However, IBM2 combines other conversion method and utilizes all training data to improve similarity to target speaker.

4.1 English evaluation result

The IBM1 and IBM2 systems were evaluated together with another 5 systems submitted by Nokia, Siemans, and UPC. The natural speech of a source speaker and a target speaker is also evaluated as a reference.

The global voice conversion scores (Table 3) show that the IBM1 system achieves a much higher quality score than all the other systems and also yields the highest mean score of similarity and quality. Compared to IBM1, the IBM2 system results in a better similarity score. However, IBM2 also produces more quality degradation.

To better understand the differences between the IBM1 and IBM2 systems, the similarity scores for each conversion are examined separately (Table 4). The IBM2 system gives much better similarity scores than IBM1 for 3 conversions. In fact, IBM2 gets the highest similarity score for the conversion 75(F) to 76(F) and for the conversion 80(M) to 76(F). However, IBM2 gives a worse similarity score for the conversion 80(M) to 79(M). The target speaker 79(M) is in fact younger than the source speaker 80(M). However, the more severe degradation in IBM2 makes the converted speech sound coarser and more elderly than the target speaker. IBM2 thus has great potential to get high mean scores if the quality can be further improved.

Table 3 Global voice conversion scores for English

System	Global VC score for English			Rank
	Similarity	Quality	Mean	
IBM1	2.32	3.63	2.98	1
IBM2	2.76	2.71	2.73	2
System3	2.17	1.45	1.81	7
System4	1.75	3.11	2.43	5
System5	2.44	2.63	2.54	4
System6	2.81	2.00	2.40	6
System7	2.88	2.50	2.69	3

Table 4 Separate similarity scores for English

Conversion system	Conversion 75(F) to 76(F)		Conversion 75(F) to 79(M)		Conversion 80(M) to 76(F)		Conversion 80(M) to 79(M)	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
IBM1	2.10	5	2.56	6	1.92	6	2.71	3
IBM2	3.20	2	3.00	4	2.57	2	2.25	7
System3	2.67	3	2.50	7	1.60	7	1.89	8
System4	1.64	9	1.50	8	1.44	8	2.40	5
System5	2.00	7	2.80	5	2.56	3	2.40	5
System6	2.62	4	3.67	2	2.33	4	2.60	4
System7	2.10	5	3.67	2	2.17	5	3.57	2
SRC-TGT	1.90	8	1.00	9	1.00	9	1.63	9
TGT-TGT	4.42	1	4.21	1	4.42	1	4.21	1

Note: SRC is the source and TGT is the target.

4.2 Spanish evaluation result

Two types of voice conversion tasks are used in the Spanish evaluation: (1) intra-lingual voice conversion (IVC) and (2) cross-lingual voice conversion (CVC). In the CVC case, the language of the training data for speaker B (the target) is different from the language of speaker A (the source). In this paper the CVC task was only done in the Spanish-to-English direction.

Because the authors have no knowledge of Spanish and to save time, we directly use the mapping formants extracted from the English data to generate a frequency warping function for both IVC and CVC tasks for Spanish. However, we use the F0 mean and variance extracted from the Spanish data for the intra-lingual voice conversion evaluation. Values extracted from English data are used for the cross-lingual voice conversion evaluation.

For the intra-lingual voice conversion evaluation,

our systems, noted as IBM1 and IBM2, are evaluated together with another 4 systems submitted by Siemans and UPC. The results are listed in Table 5.

Table 5 Global IVC scores for Spanish

System	Global IVC score			Rank
	Similarity	Quality	Mean	
IBM1	2.20	3.48	2.84	2
IBM2	2.35	2.92	2.64	4
System3	1.38	3.30	2.34	5
System4	1.98	2.35	2.16	6
System5	2.75	2.85	2.80	3
System6	3.70	2.55	3.13	1

For the cross-lingual voice conversion evaluation, we only submit one system IBM1 because the IBM2 system requires Spanish training data. Our system, IBM1, is evaluated together with another 3 systems submitted by Siemans and UPC. The results are shown in Table 6.

Table 6 Global CVC scores for Spanish

System	Global CVC score			Rank
	Similarity	Quality	Mean	
IBM1	1.78	3.52	2.65	2
System2	1.38	3.23	2.30	3
System3	1.93	2.02	1.97	4
System4	2.63	2.80	2.71	1

In both evaluations, IBM1 achieves the highest quality score. IBM1 and IBM2 also get a moderate similarity score using the mapping formants extracted from English training data. This result is acceptable considering that we do not adjust the formants for the different language.

5 Conclusions

We have described in this paper a novel voice conversion method by frequency warping. The frequency warping function is generated based on mapping formants of the source speaker and the target speaker. Our method has two advantages: (1) it requires a very small amount of training data, and (2) it can preserve high quality in the converted speech while improving similarity to a target speaker. Our proposed method was evaluated at the 2007 TC-STAR voice conversion evaluation. The evaluation results show that our method can achieve much better converted speech quality than other methods and a good balance between quality and similarity. Our method is both competitive and convenient for cross-lingual voice conversion tasks.

References

[1] Abe M, Nakamura S, Shikano K, Kuwabara H. Voice conversion through vector quantization. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Seattle, WA, USA, 1998: 655-658.

[2] Arslan L M, Talkin D. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In: Proceedings of European Conference on Speech Communication and Technology. Rhodes, Greece, 1997.

[3] Stylianou Y, Cappe O, Moulines E. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 1998, **6**(2): 131-142.

[4] Stylianou Y. High resolution voice transformation [Dissertation]. Portland, USA: Oregon Health and Science University, 2001.

[5] Shuang Zhiwei, Wang Zixiang, Ling Zhenhua, Wang Renhua. A novel voice conversion system based on codebook mapping with phoneme-tied weighting. In: Proceedings of International Conference on Spoken Language Processing. Jeju, Korea, 2004.

[6] Toda T, Black A W, Tokuda K. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Philadelphia, PA, USA, 2005, **1**: 9-12.

[7] Eichner M, Wolff M, Hoffmann R. Voice characteristics conversion for TTS using reverse VTLN. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Montreal, PQ, Canada, 2004.

[8] Eide E, Gish H. A parametric approach to vocal tract length normalization. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Atlanta, USA, 1996.

[9] Zhang P M, Waibel A. Vocal tract length normalization for large vocabulary continuous speech recognition. Carnegie Mellon University, Language Institute Technical Report: CMU-LTI-97-150, 1997.

[10] Shuang Z, Bakis R, Shechtman S, Qin Y. Frequency warping based on mapping formant parameters. In: Proceedings of International Conference on Spoken Language Processing. Pittsburgh, USA, 2006.

[11] Chazan D, Hoory R, Sagi A, Shechtman S, Sorin A, Shuang Z, Bakis R. High quality sinusoidal modeling of wideband speech for the purposes of speech synthesis and modification. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Toulouse, France, 2006.