

Voice Conversion Based On Mapping Formants

Zhiwei Shuang, Raimo Bakis*, Yong Qin

IBM China Research Lab,
IBM US Yorktown Research Lab*
shuangzw@cn.ibm.com

Abstract

We propose 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. Alignment and selection process are performed to ensure the mapping formants can represent speakers' difference well. In addition to frequency warping, we also use fundamental frequency adjustment, spectral envelope equalization, breathiness adding and duration modification to improve the similarity to target speaker. Our voice conversion method needs only a very small amount of training data for generating the warping function, which can greatly facilitate its application. We attended TC_STAR intra-lingual voice conversion evaluation for UK English and CN Mandarin. Evaluation results show that our method can achieve much better quality of converted speech than other methods.

1. Introduction

Voice conversion is to convert source speaker's voice to let it sound like target speaker's voice. There are many applications for voice conversion. An important application is to build customized text-to-speech system for different companies, in which a TTS system with one company's favorite voice can be created quickly and inexpensively by modifying origin speaker's speech corpus. Voice conversion can also be used for generating special characters' voice for movie making or keeping speaker's identity in speech to speech translation. In most of the applications, there are two practical requirements: 1). The quality of converted speech should be good enough. Companies are most likely to reject a low-quality customized TTS system even if it provides their favorite voice. 2). The training dataset of the target speaker should be small. Few users are patient to recite hundreds of sentences before they can use the speech translation system. These two requirements are important criteria for us to choose voice conversion methods.

Spectral attributes are one of the most important features to represent speakers' identity. The most popular two spectral conversion methods for are codebook mapping (Abe 1998; Arslan 1997) and GMM (Stylianou 1998; Kain 2001). However, though both methods have been improved recently, the quality degradation introduced is still severe (Shuang 2004; Toda 2005). In comparison, another spectral conversion method-frequency warping, introduces less quality degradation (Eichner 2004), and so is chosen for our spectral conversion.

Many previous works have been proposed on finding good frequency warping functions. One approach was proposed by Eide and Gish (Eide 1996), in which the warping function is based on the median of the third formant for each speaker. Some researchers extended this approach by generating warping functions based on the formants belong to the same phoneme. The underlying

assumption of such an approach is that formants parameters are related to vocal tract length. However, as commented by G.Fant (as quoted by Puming Zhang(Zhang 1997)), "formant frequency and its relationship with VTL are highly dependent on the context, and could vary largely with different context for the same speaker". The mix of formant frequencies of different contexts may not reflect the differences of vocal tract between different speakers. Thus even if a large amount of data is given to get a reasonable average, the mixture can blemish the elaborate differences between speakers.

We propose a novel method of generating a frequency warping function by mapping formant parameters of the source speaker and the target speaker. Alignment and selection process are added to ensure the selected mapping formants can represent speakers' difference well. This approach requires only a very small amount of training data for generating the warping function, which can greatly facilitate its application. It can also achieve high quality of the converted speech while successfully converting a speaker's identity. A practical voice conversion system has been built based on this approach. And experimental results show its effectiveness.

This paper is organized as follows. Section 2 gives an overview of the voice conversion system. Section 3 details the frequency warping method based on mapping formants. In Section 4, we briefly introduce several other conversion methods used in our system. The TC-STAR intra-lingual voice conversion evaluation data and method are described in Section 5. And Section 6 provides the evaluation results with discussions. We conclude our paper in Section 7.

2. Voice Conversion System Overview

We use the analysis/reconstruction technique, proposed by D. Chazan (Chazan 2006), to get an enhanced complex envelope model and pitch contour. The technique is based on efficient line spectrum extraction and frequency

dithering noise insertion during the synthesis. Frame alignment procedures during analysis and synthesis are provided to allow both amplitude and phase manipulation during speech manipulations, e.g. pitch modification, spectral smoothing, vocal tract conversion etc.

Then we use frequency warping to stretch/compress spectrum along frequency axis. Meanwhile, we use f_0 adjustment to transform the average and variance of $\log f_0$ of the source speaker to those of the target speaker. Then, we re-sample the warped spectrum envelope in multiples of converted f_0 to get new complex line spectrum. After this, we apply a filter on the spectrum to compensate for the different energy distribution along the frequency axis. We can apply breathiness adding by adding random number to the phase of the complex line spectrum if needed. Finally, we reconstruct the converted speech from the line spectrum sequence. Duration modification can be achieved by repeating or deleting some frames of line spectrum in reconstruction. Details will be introduced in Section 3 and 4.

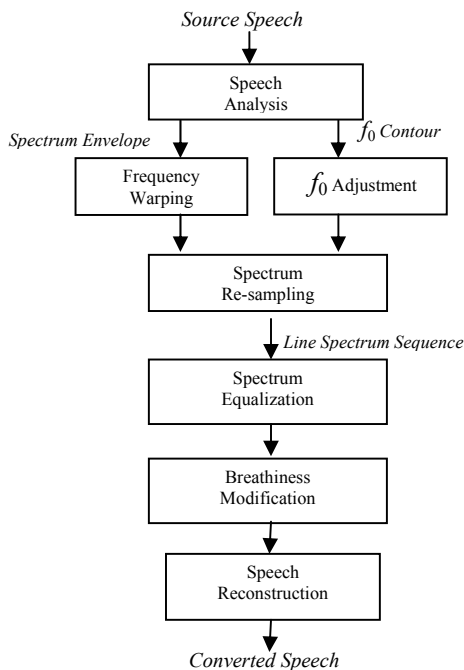


Figure 1: Diagram of Voice Conversion System

3. Frequency warping by mapping formants

3.1. Get Mapping Formants

This process comprises: preparing the training speeches of a source speaker and target speaker; performing frame alignment on the training speech of the source speaker and target speaker; selecting aligned frames from the frame-aligned training speech of the source speaker and target speaker; extracting formant parameters of source speech and target speech.

We manage to align the source speech data and target speech data first to ensure they are in similar context. However, not all aligned frames' formants parameters will be used for generating the frequency warping function. For example, the formants in transition part can not represent the speaker's vocal tract. We need to select the well aligned frames to get mapping formant parameters.

The step of selecting aligned frames comprises one or more of the following: selecting from the phonemes with the formant parameters of less variance, selecting from the phonemes with "plosives", "fricatives" or "silence" as their neighboring phonemes, selecting from the middle portion of the phonemes, and selecting the aligned frames with the minimal acoustic dissimilarity. After the selection step, either one pair of matching frames or multiple pairs of matching frames can be selected. Moreover, the multiple pairs of the matching frames can either belong to different phonemes or belong to the same phoneme.

Then we will extract the formant parameters of the selected frames of source speech and target speech. Many tools can be used to automatically extract formant tracks from speech, such as PRAAT. We suggest manually checking the automatically extracted formants of these occurrences to avoid error.

3.2. Generate Frequency Warping Function

In this step, we will use the formants of the selected aligned frames as the key positions to generate the frequency warping function.

To facilitate illustration, the formant frequencies of the source speaker are noted as $[F_{1s}, F_{2s}, F_{3s}, \dots, F_{ns}]$, while the Formant parameters of the target speaker are noted as $[F_{1t}, F_{2t}, F_{3t}, \dots, F_{nt}]$. The mapping formants $[F_{it}, F_{is}]$ will be the key positions to define a piecewise linear frequency warping function from the target frequency axis to the source frequency axis. Linear interpolation is proposed to generate the part between two adjacent key positions while other interpolation schemes may also be used.

3.3. Frequency Warping

In this step, we will use the defined frequency warping function or functions to perform frequency warping of spectrum. Suppose one frame of the source speaker's spectrum is $S(w)$, and the frequency warping function from the target frequency axis to the source frequency axis is $F(w)$, then the converted spectrum $Conv(w)$ is:

$$Conv(w) = S(F(w)) \quad (1)$$

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

4. Additional Voice Conversion Methods

4.1. Fundamental-frequency Adjustment

Adjustment of f_0 contours consists of a linear transform applied to $\log f_0$. Thus, if f_{0s} is the source f_0 and f_{0t} is the target f_0 , then $\log f_{0t} = a + b \log f_{0s}$, where a and b are chosen to transform the average and variance of $\log f_0$ of the source speaker to those of the target speaker.

4.2. Spectral-envelope Equalization

Spectral-envelope equalization is implemented as a filter on the spectrum to compensate for the different energy distribution along the frequency axis. We calculate the difference curve between average power spectra of the source and target speakers after frequency warping. Then we smooth the difference curve to get a smoother spectral filter to serve as the spectral envelope equalization filter.

4.3. Breathiness Adding

When the target speaker's voice was breathier, on average, than the source, we added random values to the phases of f_0 harmonics above 2500 Hz.

4.4. Duration Modification

Duration modification is implemented by an overall lengthening or shortening to compensate for the difference between speakers in utterance speed. In general, elder people tend to speak slower than younger person. So duration modification can usually be helpful for conversion between speakers of different ages.

5. Evaluation Data and Method

5.1. Evaluation Data

We attend the TC-STAR intra-lingual voice conversion evaluation for UK English and CN Mandarin. The training data are of 3 mandarin speakers noted as 01(F), 02(M), 03(F) and 4 UK speakers noted as 75 (F), 76 (F), 79(M), 80(M), where (F) denotes female speaker while (M) denotes male speaker. We select mapping formants and calculate f_0 coefficients from the first 30 training sentences of each speaker.

5.1.1. Selected Mapping Formants

For UK speakers, we use formants in the middle of phoneme “ɜ:” in syllable “Heard” of No.22 training sentence as mapping formants.

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

Table 1. Mapping Formants of UK English Speakers

For mandarin speakers, we use formants in the middle of phoneme “i:” in syllable “Ji1” of No.28 training sentence as mapping formants.

Speaker	F1	F2	F3	F4
01(F)	301	2746	3583	4287
02(M)	300	2137	2861	3664
03(F)	303	2744	3452	4345

Table 2. Mapping Formants of Chinese Speakers

We find the mapping formants of 01(F) and 03(F) are quite similar. It indicates that 1) 01(F) and 03(F) are similar to each other, and 2) the generated 01(F) to 03(F) frequency warping function will only make small modifications along the frequency axis.

5.1.2. Other Conversion Settings

Fundamental-frequency adjustment and spectral-envelope equalization are applied for each conversion. We apply breathiness adding for conversion from 75(F) to 76(F).

Duration modification is applied for some conversion. Below is the duration modification ratio for each conversion.

Source/Target	Duration Modification Ratio
75(F)/76(F)	1.05
75(F)/79(M)	1.0
80(M)/76(M)	0.9
80(M)/79(F)	1.0
01(F)/02(M)	0.95
01(F)/03(F)	1.0
02(M)/03(F)	1.05

Table 3. Duration Modification Ratio

5.2. Evaluation Method

5.2.1. TC-STAR Quality Evaluation

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

5.2.2. TC-STAR Similarity Evaluation

In this evaluation, the listeners are asked to rate if a given voice pair come or not from the same person according to following scale: (5) Definitely identical, (4) Probably identical, (3) Not sure, (2) Probably different, (1) Definitely different. Arithmetic mean of all subjects' individual score is used as the evaluation result.

6. Results and Discussions

6.1. UK English Evaluation Result

Our system, noted as IBMc, is evaluated together with another 5 systems noted as UPC1, UPC2, NOKa, UPC3 and

SIE1. Natural speech of source speaker (SOURCE) and target speaker (TARGET) are also evaluated as reference. These systems are sorted in ascending order of their quality scores, as shown in Figure 2. And their similarity evaluation results are shown in Figure 3.

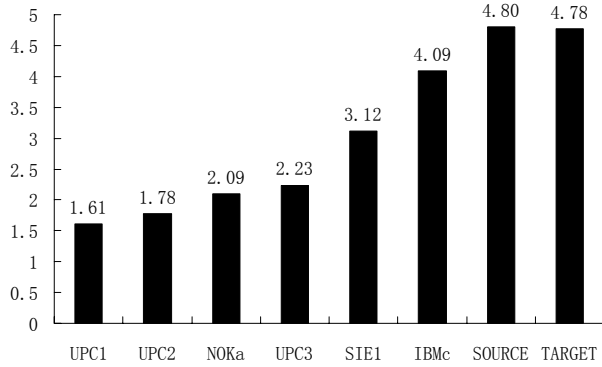


Figure 2. Quality Evaluation of UK English

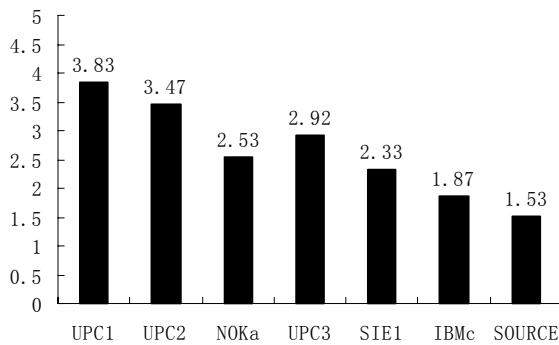


Figure 3. Similarity Evaluation of UK English

Evaluation results show that our system (IBMc) gets a much higher quality score than all the other systems. However, our system gets lower similarity score than the other systems. An interesting thing is that, for most of these systems, the higher is its quality score, the lower is its similarity score. And the only exception occurs for the system NOKa. Though this may due to the characteristics of different methods, we argue that worse quality makes it more difficult for the listeners to decide whether the speech comes from the same speaker or not. For example, listeners may choose “(3)NOT sure” when the quality is “(1) bad” just because he can hardly discern any speaker’s characteristics with such bad quality.

6.2. CN Mandarin Evaluation Result

Our system, noted as IBMc, is evaluated with another system noted as CAS. The quality evaluation result is shown as Figure 4 while the similarity evaluation result is shown as Figure 5.

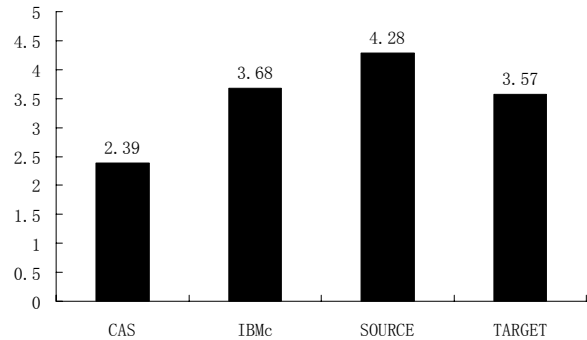


Figure 4: Quality Evaluation of CN Mandarin

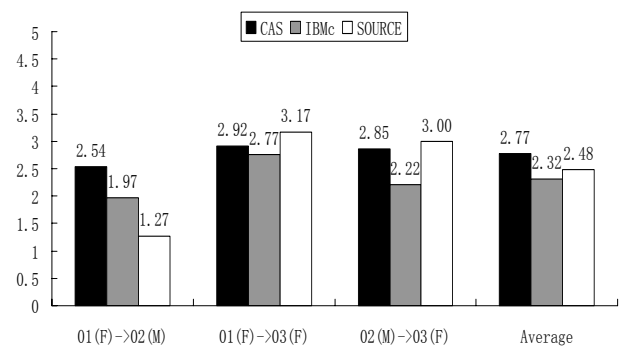


Figure 5: Similarity Evaluation of CN Mandarin

The quality evaluation results show that our system (IBMc) gets a much higher quality score than CAS, surprisingly even higher than the score of natural speech of target speaker. The similarity evaluation results include separate scores for three conversions: from 01(F) to 02(M), from 01(F) to 03(F), from 02(M) to 03(F) and the average score. For conversion from 01(F) to 02(M), both CAS and IBMc achieve a much higher score than source speech.

For conversion from 01(F) to 03(F) and from 02(M) to 03(F), both systems get a lower score than the original source speech. As to conversion from 01(F) to 03(F), it can be explained that 01(F) and 03(F) are already very similar to each other, so voice conversion does not improve the similarity. This explanation matches the indication of similar mapping formants of 01(F) and 03(F).

It is surprising that a high similarity score is obtained for comparing original 02(M) speech and 03(F) speech because they are from two speakers of different genders and sound quite different. After checking with the raw evaluation data, we found that the standard deviation of all scores for comparing original 02(M) and 03(F) is 1.46385. Since its average score is 3.0, it suggests that different people have much different opinion when comparing these two voices. One reason may be that the evaluation criteria is not clear enough, thus different people have different explanation for the criteria. We strongly suggest that TC_STAR refine the similarity evaluation criteria to make the evaluation result more reliable.

7. Conclusions

We propose 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. The advantages of this method are that 1) it requires very small amount of training data, and 2) it preserves high quality of the converted speech while improving the similarity to the target speaker. Evaluations of a practical voice conversion system based on this approach show its effectiveness.

8. Reference

- Abe, M., S. Nakamura, K. Shikano, and H. Kuwabara, "Voice Conversion through Vector Quantization," Proc. ICASSP, Seattle, WA, U.S.A., 1998, pp. 655-658.
- Arslan, L. M. and D. Talkin, "Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum," Proc. Eurospeech, Rhodes, Greece, 1997.
- Stylianou, Y. et al., "Continuous Probabilistic Transform for Voice Conversion," IEEE Transactions on Speech and Audio Processing, v. 6, no. 2, March 1998, pp. 131-142.
- Kain, A. B., "High Resolution Voice Transformation," Ph.D. thesis, Oregon Health and Science University, October 2001.
- Shuang, Z. W., Z. X. Wang, Z. H. Ling, and R. H. Wang, "A Novel Voice Conversion System Based on Codebook Mapping with Phoneme-Tied Weighting," Proc. ICSLP, Jeju, Korea, 2004.
- Toda, T., A. W. Black, and K. Tokuda, "Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter," Proc. ICASSP, Philadelphia, PA, U.S.A., 2005, v. 1, pp. 9-12.
- Eichner, M., M. Wolff, and R. Hoffmann, "Voice Characteristics Conversion for TTS Using Reverse VTLN," Proc. ICASSP, Montreal, PQ, Canada, 2004.
- Eide, E. and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization", in ICASSP 1996, Atlanta, USA, 1996.
- Zhang, P.M. and A. Waibel, "Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition", Carnegie Mellon University, Language Institute Technical Report: CMU-LTI-97-150, 1997.
- Chazan, D., R. Hoory, A. Sagi, S. Shechtman, A. Sorin, Z.W. Shuang, and R. Bakis, "High Quality Sinusoidal Modeling of Wideband Speech for the Purposes of Speech Synthesis and Modification," in ICASSP 2006.
- Eide E., A. Aaron *et al*, "Text-to-Speech: Bridging the Flexibility Gap Between Humans and Machines", in SpeechTek West 2006, San Francisco, USA, 2006.
- Shuang, Z. W., R. Bakis, S. Shechtman, Y. Qin, "Frequency Warping Based on Mapping Formant Parameters", submitted to ICSLP 2006.