Voice Conversion by Combining Frequency Warping with Unit Selection

Zhiwei Shuang, Fanping Meng, Yong Qin

IBM China Research Lab
{shuangzw, mengfp, qinyong}@cn.ibm.com

ABSTRACT

In this paper, we propose a novel voice conversion method by combining frequency warping and unit selection to improve the similarity to target speaker. We use frequency warping to get the warped source spectrum, which will be used as estimated target for later unit selection of the target speaker's spectrum. Such estimated target can preserve the natural transition of human's speech. Then, part of the warped source spectrum is replaced by the selected target speaker's real spectrum before reconstructing the converted speech to reduce the difference in detailed spectrum. TC-STAR 2007 voice conversion evaluation results show that the proposed method can achieve about 20% improvement in similarity score compared to only frequency warping.

Index Terms—Voice Conversion, Warping, Selection

1. INTRODUCTION

Voice conversion is to change the characteristics of a source speaker's voice to those of a target speaker. 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. To evaluate the performance of voice conversion systems, there are two criteria for the converted speech: quality and similarity to the target speaker. With state of the art voice conversion technologies, there is always a tradeoff between quality and similarity. Different applications may have different requirements for quality and similarity to the target speaker.

Spectral conversion is the key component in voice conversion system. The two most popular spectral conversion methods are codebook mapping [1][2] and GMM based mapping[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-frequency warping, introduces less quality degradation [7]. Many previous approaches have been proposed on finding good frequency warping functions [8][9][10]. We proposed a new method of generating frequency warping function by mapping formant parameters of the source speaker and the target speaker in 2006 [11]. 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, and can achieve a high quality of the converted speech. However, listeners can still perceive the difference between the converted speech and the target speaker's speech. Based on our observation, part of the reason is the difference in detailed spectrum, which can not be compensated by only frequency warping. For those usage scenarios with high requirements for similarity to the target speaker, only applying frequency warping is not enough.

In this paper, we propose to combine frequency warping and unit selection to improve the similarity to target speaker. We first apply frequency warping to generate warped source spectrum, which is similar to the target speaker's spectrum. Then, the warped source spectrum will be used as estimated target for later unit selection of the target speaker's spectrum. Finally, we replace part of the warped spectrum with the target speaker's real spectrum, and reconstruct the converted speech. TC-STAR 2007 voice conversion evaluation results show that the proposed method can achieve a much better similarity score than only frequency warping and a better quality score than most other systems.

This paper is organized as follows. Section 2 describes our voice conversion system in detail. The TC-STAR voice conversion evaluation data and method are described in Section 3. And Section 4 provides the evaluation results and discussions. We conclude our paper in Section 5.

2. VOICE CONVERSION SYSTEM

The diagram of our voice conversion system is shown as Figure 1. In general, our voice conversion method can be divided into two stages: Training Stage and Conversion Stage. Our speech analysis/reconstruction technique is described in [12]. We decompose speech into complex spectrum envelope and F0 contour. Then we can make both amplitude and phase manipulation, including frequency warping, F0 modification and spectral smoothing etc.

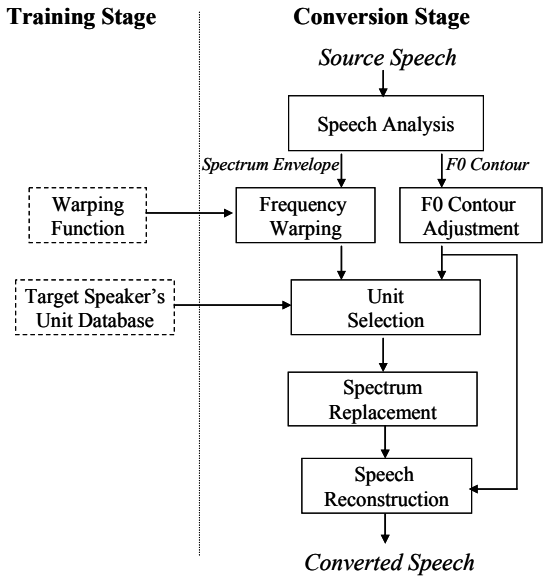

Figure 1: *Diagram of Voice Conversion System.*

2.1. Training stage:

2.1.1. Training of frequency warping function

We train frequency warping function based on mapping formant parameters of selected aligned frames. Manual check can be included in training. The mapping formants are used as 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. This approach needs only a very small amount of training data.

2.1.2. Training of unit database

Our basic candidate unit is frame. Our frame interval is 5ms for woman and 10ms for man. We get amplitude spectrum envelope feature and F0 feature of each unit by our speech analysis technique. Meanwhile, we can get phonetic features for each unit from the alignment information. The phonetic features include the unit's current phoneme information and neighboring phonemes information. Such information is very useful for later unit selection because we find that only acoustic distance is not robust enough. We store the feature vector of each candidate unit in the unit database.

2.1.3. Other trainings

Besides the two trainings above, other trainings can be performed. For example, we can train a linear F0 adjustment function 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 calculated according to the average and variance of $\log f_0$ of the source speaker and the target speaker.

2.2. Conversion stage:

2.2.1. Frequency warping and F0 adjustment

First we use the speech analysis algorithm to decompose the source speech into complex spectrum envelope and F0 contour. Then we use frequency warping to stretch/compress source spectrum along frequency axis to get the warped spectrum, which is similar to the target speaker's spectrum in general. Meanwhile, we use F0 adjustment to transform the average and variance of $\log f_0$.

2.2.2. Unit selection.

This step is similar to unit selection in Text to Speech (TTS) system. However, our target feature vector is generated from the warped spectrum, adjusted F0 contour and alignment information. Compared to the estimated target by models and decision trees in TTS system, our estimated target can preserve the natural transition of human's speech.

Dynamic Programming is used to select the best unit sequences. The target cost is defined as weighted sum of spectral distance, prosodic distance and phonetic distance between the target feature vector and each candidate unit feature vector. The transition cost is defined as the spectral distance between pairs of candidates units. Our spectral distance is calculated by weighted LSF distance (Linear Spectral Frequency). The first 8 LSF parameters' differences are given a larger weight than later ones. Our prosodic distance is calculated by the difference between F0 in log domain. Our phonetic distance is set by whether the phoneme information and the neighboring phonemes information are the same. The weights of spectral distance, prosodic distance and phonetic distance are manually adjusted to make them comparable.

2.2.3. Spectrum replacement

After selecting the proper candidate unit, we replace the warped amplitude spectrum with the amplitude spectrum of the selected candidate unit. However, preferably, we will keep the warped spectrum below a specific frequency (i.e. 1000 Hz) unchanged. It is because our basic unit is frame that the converted speech can have severe discontinuity problem if we replace the whole spectrum. Since the low frequency spectrum is very important for keeping the continuity and not so important for improving the similarity, it is usually better to keep the low frequency amplitude spectrum unchanged. The warped phase spectrum will be kept unchanged for whole frequency range. The partly replaced amplitude spectrum and the warped phase spectrum are combined to the modified complex spectrum.

2.2.4. Speech reconstruction

Finally, we will reconstruct the speech data from the modified complex spectrum and converted F0 contour. Spectral smoothing over time axis can be applied before reconstruction to mitigate the discontinuity problem.

3. TCSTAR VOICE CONVERSION EVALUATION

The TC-STAR project, financed by 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 Text to Speech (TTS) (speech synthesis). The objective of TC-STAR project is ambitious: making a breakthrough in SST that significantly reduces the gap between human and machine translation performance.

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

3.1. Evaluation data

In 2007 TC-STAR voice conversion evaluation, the training data are of 4 speakers noted as F(75), F(76), M(79) and M(80), where F denotes female speaker while M denotes male speaker. 126 sentences are provided as training data for UK English intra-lingual voice conversion. 4 direction conversion are evaluated: F(75)->F(76), F(75)-> M(79), M(80)-> F(76) and M(80)-> F(79).

3.2. Evaluation criteria

3.2.1. TC-STAR similarity evaluation

In this evaluation, the listeners are asked to rate whether 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.

3.2.2. 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.

3.3. Subjective test settings

Subjective tests were carried out via the web. An access to 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. They were 18 to 40 years old native English speakers with no known hearing problem. No one was a speech synthesis expert.

4. RESULTS AND DISCUSSIONS

We submitted two systems, noted as IBM1 and IBM2 for UK English intra-lingual voice conversion. IBM1 is a voice conversion system only by frequency warping, as described in our previous paper [13]. IBM2 is based on combination of frequency warping and unit selection, as described in this paper. In this submission, no spectral smoothing is applied before reconstruction. We used formants in the middle of phoneme "ɜ:"in syllable "Heard" of No.22 training sentence as the mapping formants to generate warping function.

Table 1: *Mapping Formants of 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

Our two systems were evaluated together with another 5 systems. Natural speech of source speaker (SOURCE) and target speaker (TARGET) were also evaluated as reference. TC-STAR Evaluation ranked all systems according to their mean score of quality score and similarity score. Table 2 and Table3 are the evaluation results of UK English intra-lingual voice conversion except other company names are hidden.

Table 2 lists the average voice conversion scores for all 4 direction conversions. As shown in Table 2, IBM1 gets a much higher quality score than all the other systems and also gets the highest mean score of similarity and quality. However, IBM1's similarity score is not very good, which can limit its usage scenarios. Compared to IBM1, IBM2 gets about 20% improvement in similarity score (from 2.32 to 2.76), which is close to the highest similarity score of all systems. Though IBM2's quality is not as good as IBM1's, its quality is still better than most other systems'.

Table 2: *Average voice conversion scores*

System	Similarity Score	Quality Score	Mean Score	Rank
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 3: *Separate similarity scores*

System	F(75) -> F(76)	F(75)-> M(79)	M(80)-> F(76)	M(80)-> F(79)
IBM1	2.10	2.56	1.92	2.71
IBM2	3.20	3.00	2.57	2.25
System3	2.67	2.50	1.60	1.89
System4	1.64	1.50	1.44	2.40
System5	2.00	2.80	2.56	2.40
System6	2.62	3.67	2.33	2.60
System7	2.10	3.67	2.17	3.57
SRC-TGT	1.90	1.00	1.00	1.63
TGT-TGT	4.42	4.21	4.42	4.21

To understand the difference between IBM1 and IBM2 better, we check separate similarity scores for each conversion in Table 3. We find that IBM2 gets much better similarity scores than IBM1 for 3 conversions. In fact, IBM2 gets highest similarity score for conversion F(75)->F(76) and conversion M(80)->F(76) among all systems.

However, to our surprise, IBM2 gets a worse similarity score for conversion M(80) to M(79). When we check the speech data, we find that the target speaker M(79) sounds younger than the source speaker M(80). However, the discontinuity problems in IBM2 make the converted speech sound coarser and feel older than the target speaker. Thus, listeners gave IBM2 a lower similarity score than IBM 1.

In our later experiments, we find spectral smoothing over time axis before reconstruction can be very helpful to mitigate the discontinuity problem and alleviate the coarse feeling in M(80) to M(79) conversion.

5. CONCLUSIONS

In this paper, we propose a novel voice conversion method by combining frequency warping and unit selection to improve the similarity to target speaker. We use frequency warping to get the warped spectrum, which will be used as estimated target spectrum for later unit selection. Then, part of the warped source spectrum is replaced by the selected target speaker's real spectrum before reconstructing the converted speech. This method has two advantages: (1) Compared to the estimated target by models and decision trees, the warped spectrum can keep the natural transition and variation in natural speech. (2). Unit selection and spectrum replacement can reduce the difference in detailed spectrum between speakers, which can not be compensated by frequency warping. TC-STAR 2007 voice conversion evaluation results show that the proposed method can achieve 20% better similarity score than only frequency warping and a better quality score than most other systems.

6. REFERENCE

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice Conversion through Vector Quantization," *Proc. ICASSP*, Seattle, WA, U.S.A., 1998, pp. 655-658.

[2] L.M. Arslan, and D. Talkin, "Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum," *Proc. Eurospeech*, Rhodes, Greece, 1997.

[3] Y. Stylianou, O. Cappe and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Transactions on Speech and Audio Processing*, v. 6, no. 2, March 1998, pp. 131-142.

[4] A. B. Kain, "High Resolution Voice Transformation," *Ph.D. thesis*, Oregon Health and Science University, October 2001.

[5] Z. W. Shuang, 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.

[6] T. Toda, 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.

[7] M. Eichner, M. Wolff, and R. Hoffmann, "Voice Characteristics Conversion for TTS Using Reverse VTLN," *Proc. ICASSP*, Montreal, PQ, Canada, 2004.

[8] H. Valbret, E. Moulines, J.P. Tubach, "Voice transformation using PSOLA technique," *Proc. ICASSP,* San Francisco, 1992

[9] E. Eide, and H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," *Proc. ICASSP*, Atlanta, USA, 1996.

[10] D. Sundermann, and H. Hoge, "TC-Star: Cross-Language Voice Conversion Revisited," *TC-STAR Workshop on Speech to Speech Translation*, Barcellona, June 2006.

[11] Z. W. Shuang, R. Bakis, S. Shechtman and Y. Qin, "Frequency Warping Based on Mapping Formant Parameters", *Proc. ICSLP,* Pittsburgh, U.S.A, 2006.

[12] D. Chazan, , 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," *Proc. ICASSP*, Toulouse, France, 2006.

[13]. Z.W. Shuang, R. Bakis and Y. Qin, "Voice Morphing System Based on Mapping Formant Parameters," *TC-STAR Workshop on Speech to Speech Translation*, Barcellona, June 2006.