Frequency Warping Based on Mapping Formant Parameters

Zhi-Wei Shuang[1], *Raimo Bakis*[2], *Slava Shechtman*[3], Dan Chazan[3], *Yong Qin*[1],

[1] IBM China Research Lab,
[2] IBM T.J. Watson Research Center,
[3] IBM Haifa Research Lab.
shuangzw@cn.ibm.com

Abstract

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 performed to ensure that the 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 morphing system has been built based on this approach. And experimental results show its effectiveness.

Index Terms: voice morphing, frequency warping, formant

1. Introduction

Frequency warping is a hot research topic which attempts to compensate for the differences between the acoustic spectra of different speakers. Given a spectral cross section of one sound, it creates a new spectral cross section by applying a frequency warping function. Many previous works have been proposed on finding good frequency warping functions. L.F.Uebeland [1] suggested a Maximum Likelihood Linear Regression method. However, a large training dataset is requested which limits its usage scenarios. Matthias [2] proposed to select the frequency warping function from some pre-defined one-parameter family of functions, but the effectiveness is not satisfying. David [3] adopted dynamic programming to train linear or piecewise linear warping functions, which minimizes the distance between the converted source spectrum and the target one. However, this method can be greatly degraded by noise in the input spectra.

A complete different approach was proposed by Eide and Gish [4], 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 [5]), "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.

Here we propose a novel method to generate a frequency warping function based on mapping formant parameters of selected aligned frames. Alignment and selection process are added to ensure the selected mapping formants can represent speakers' difference well. This approach needs only a very small amount of training data, which can greatly facilitate its application.

This paper will be arranged this way: Section 2 will introduce our frequency warping method based on mapping formants. Section 3 will introduce our voice morphing system. Evaluations and discussions will be made in Section 4.

2. Frequency warping by mapping formants

The schematic diagram of our method is shown in Figure 1. Different from the previous works, we use alignment and selection process to ensure the selected mapping formants can represent speakers' difference well.

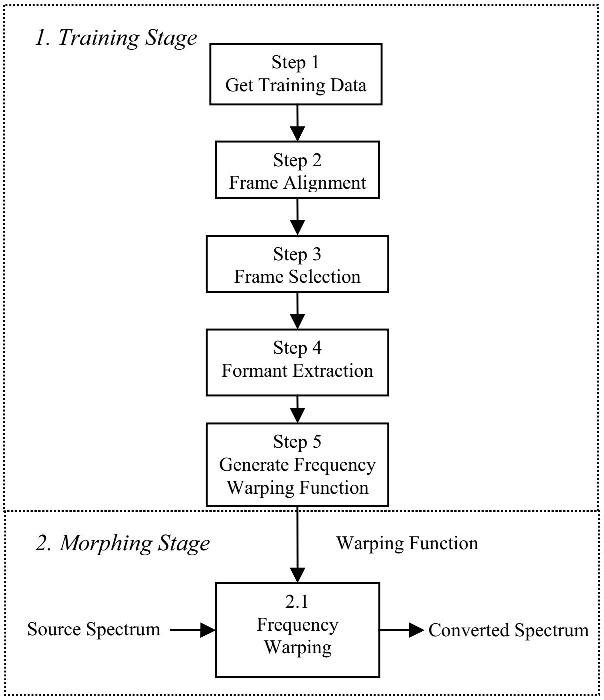

Figure 1: *Diagram of frequency warping.*

2.1. Training stage

The training stage includes 5 steps. The sequence of steps can be modified to achieve similar results.

2.1.1. Get training data

In this step, the training speech of the source speaker and the target speaker are prepared. The training speech can be prepared using various known methods, such as by recording, by extracting from audios, videos or other multimedia resources, etc. There may contain noise and music in the training speech. We do not need to acquire a large training dataset. In fact, one sentence is sufficient. And in practice, often just one phoneme uttered by the target speaker will do, provided that a frame matching with the speech of the source speaker can be extracted there from it.

Depending on different embodiments, the training speech of the source speaker and the target speaker can be required to be either of the same contents, or of different contents.

2.1.2. Frame alignment

If the training speech data is of the same contents, we can do alignment directly. The alignment can be performed by using the known Dynamic Time Warping (DTW) algorithm.

If the training speech data is of different contents, we need to select pairs of occurrences from the training speech of the source speaker and the target speaker first. Each pair of occurrences should belong to the same or similar phonemes with the same or similar contexts in the training speech of the source speaker and the target speaker. The context as used herein includes but is not limited to: neighboring phonemes, position in the word, position in the phrase, position in the sentence, etc. For each pair of aligned occurrences, the middle frame of the source speaker's occurrence is aligned with the middle frame of the corresponding target speaker's occurrence.

2.1.3. Frame selection

We can obtain many aligned frames after the alignment. And then the best aligned frames are selected.

A first selection method is based on the phoneme to which the source frame belongs. Some phonemes, such as "e", are preferred, because the formant parameters of these phonemes are of less variance than those of others. Thus these phonemes can better represent the speaker's characteristics.

A second selection method is based on the context of the frame. Some contexts are preferred, because the formants of the phonemes therein are less affected by the neighboring phonemes. For example, in an embodiment of the present invention, the phonemes with "plosives", "fricatives" or "silence" as their neighboring phonemes are selected.

A third selection method is based on the position of the frame in the phoneme. The aligned frames both located in the middle part of one phoneme are preferred. The frame in the middle is deemed as to be of less variance, because it is less easily affected by the transition from the neighboring phonemes' formants.

A fourth selection method is based on the difference in spectrum between that frame and other frames belonging to the same phoneme with the same or similar context. Smaller difference is preferred to reduce the disturbance of speakers' spontaneous variation.

The above four selection methods can also be performed in combination. 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.

2.1.4. Formant extraction

In this step, we will extract the formant parameters 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.

2.1.5. 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 $[F_{1s}, F_{2s}, F_{3s}, ..., F_{ns}]$, while the Formant parameters of the target speaker are noted as $[F_{1t}, F_{2t}, F_{3t}, ..., F_{nt}]$. 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.

Suppose that the speech of both speakers have the same maximum frequency, which is noted as $F_{\max}$. To facilitate the interpolation outside the minimum formant and the maximum formant frequencies, we add [0,0] $[F_{\max}, F_{\max}]$ as end points. Note that other end points can also be used. For example, we get one pair of mapping first four formant frequencies: [690, 2290,3080,4450]of the source speaker and[560,2180,2750,4040] of the target speaker. And if the maximum frequency is 8000 for both speakers, then the warping function from the target frequency axis to the source frequency axis is shown as Figure 2.

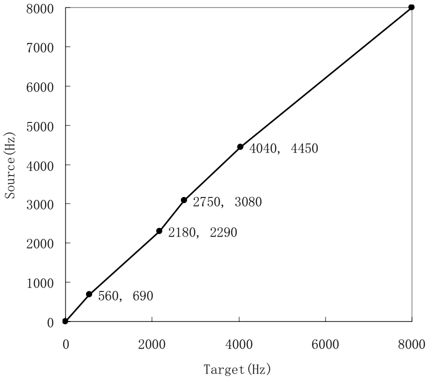

Figure 2: *Frequency warping function generation*

2.2. Morphing stage

In the morphing stage, 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 calculated as:

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

Two strategies can be used in the morphing stage.

2.2.1. *Multiple frequency warping functions strategy*

Using this strategy, different frequency warping functions are applied for their corresponding segments respectively. Alignment or classification information is required to choose the warping function. One possible implementation is: 1. Generate different frequency warping functions for different phonemes. 2. Apply the frequency warping functions according to which phoneme that frame belongs to. This strategy may improve the similarity of converted speech to the target speaker. However, spectral smoothing, or smoothing of the warping function, will be required to avoid discontinuities at phoneme boundaries.

2.2.2. *Single frequency warping function strategy*

Using this strategy, 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.

3. Voice morphing system

We use the analysis/reconstruction technique, proposed in [6], 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 morphing 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.

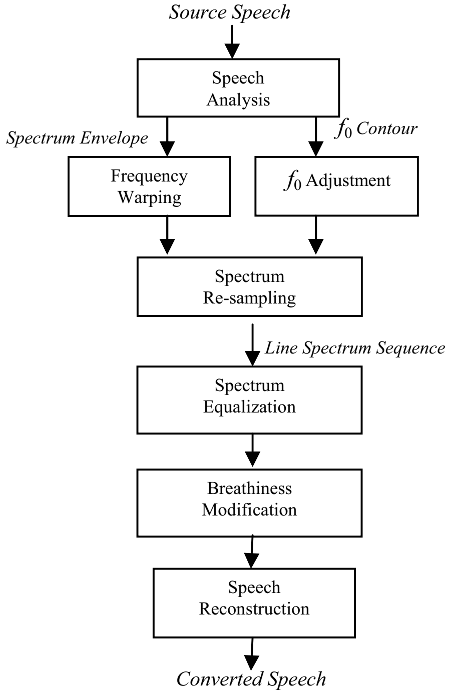

Figure 3: *Diagram of Voice Morphing System*

4. Evaluations and discussions

We run a set of evaluations to evaluate the performance of our voice morphing system.

4.1. Evaluation data

The evaluation included one target voice, LAH, and two source voices, BAS, informally deemed "similar" to LAH, and HOK, deemed "dissimilar" to LAH; all are female native speakers of U. S. English. For each voice, 20 sentences of same contents are used as training data. All speech is at 22.05 kHz sampling

We use the first four formants in the middle position of one occurrence of phoneme "e" in syllable "better" as the mapping formants to generate frequency warping functions. Besides the context, a preliminary estimate of the warping function based on average spectrum is also used as reference in the frame selection step. Single frequency warping function strategy is used.

Table 1: *Mapping Formants Frequencies*

Speaker	F1	F2	F3	F4
LAH	560	2180	2750	4040
BAS	610	2150	2870	3960
HOK	690	2290	3080	4450

As shown in Table 1, we can see that the mapping formants of BAS and LAH are quite similar. It indicates that 1) BAS and LAH are similar to each other, and 2) the generated BAS to LAH frequency warping function will only make small modifications along the frequency axis.

4.2. Preference evaluation of similarity

Here, subjects listen to a triplet of different sentences from the source speaker, the morphed source speaker, and the target speaker. Then they make a preference judgment on the morphed sentence, A: more similar to source, B: more similar to target or X: no preference. Seven listeners, who are researchers at IBM's China Research Lab, evaluated 20 HOK-LAH triplets and 20 BAS-LAH triplets. Table 2 shows the results. Preference evaluation shows the morphed speech is more similar to the target speaker in general.

Table2. *Preference evaluation results of similarity*

Source/ Target	A: More similar to source	B: More similar to target	X: No preference
HOK/LAH	0	76.4%	23.6%
BAS/LAH	23.6%	64.3%	12.1%

4.3. Opinion evaluation of similarity

Here, subjects listen to the following sentences in random sequence: 10 HOK sentences (HOK), 20 Morphed HOK sentences (HOK_M), 10 BAS sentences (BAS), 20 Morphed BAS sentences (BAS_M), and 10 LAH sentences (LAH). Then for each sentence, they give a 1-5 similarity score to LAH by comparing with LAH's 10 seconds' recording where score 5 means the sentence is felt spoken by the LAH, and 4, 3, 2 and 1 means the sentence is felt very similar to, fairly similar to, not similar to and quite different from LAH's voice respectively. Seven listeners, who work for IBM China Research Lab, are asked to make the evaluations. Average scores of each group are given in Figure 4.

Figure 4 shows that: BAS is already similar to LAH, and voice morphing only gives a small improvement on the similarity; HOK is quite different from LAH, and voice morphing greatly improves the similarity to LAH. The results match the indications of mapping formants.

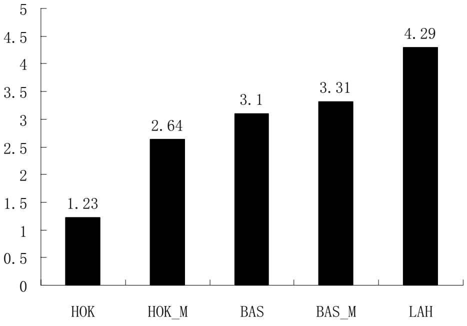

Figure 4: *Opinion evaluation results of similarity*

4.4. MOS quality evaluation

Here, subjects listen to 45 sentence recordings in random sequence: five HOK sentences, ten morphed HOK sentences (H>L), five BAS sentences, ten morphed BAS sentences (B>L), five LAH sentences, and ten LAH sentences reconstructed from an AMR 8.8-Kbps code stream [7]. All speech here is at 22.05 kHz sampling except LAH_AMR files are at 16 kHz. We include LAH_AMR to give a frame of reference since MOS scores are quite subjective. Eighteen listeners, who work for IBM in the United States, are asked to give a 1-5 score for each sentence, where 5 is excellent, 4 is good, 3 is fair, 2 is poor and 1 is unacceptable. Figure 4 shows results, which indicate that the quality of morphed speech is comparable to AMR 8.8-Kbps reconstructed speech.

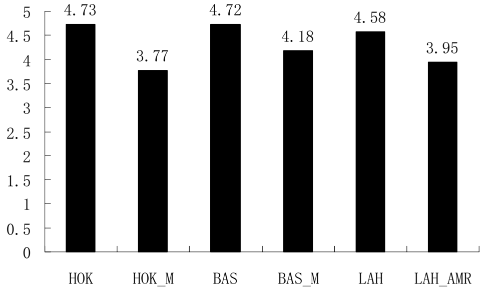

Figure 5: *Quality evaluation results (MOS).*

5. Conclusion

This paper proposes a novel approach of generating a frequency warping function by mapping formants of the source speaker and the target speaker. The advantages of this approach are that 1) it requires very small amount of training data, and 2) it preserves high quality of the converted speech while successfully converting a speaker's identity. Evaluations of a practical voice morphing system based on this approach show its effectiveness.

6. References

[1] L.F.Uebeland P.C.Woodland, "An Investigation into Vocal Tract Length Normalization", in EUROSPEECH’ 99, Budapest, Hungary,1999, pp. 2527-2530.

[2] Matthias Eichner, Matthias Wolff and Rüdiger Hoffmann, "Voice Characteristics Conversion for TTS using Reverse VTLN", in ICASSP 2004, pp. I- 17-20 vol.1.

[3] David Sundermann and Hermann Ney, "VTLN-BASED Voice Conversion", in ICSLP 2004, Jeju, Korea, 2004.

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

[5] Puming Zhang and Alex Waibel, “Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition”, Carnegie Mellon University, Language Institute Technical Report: CMU-LTI-97-150.

[6] 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, Toulouse, France, 2006.

[7] Bessette, B. *et al.*, "The Adaptive Multirate Wideband Speech Codec (AMR-WB)", IEEE Transactions on Speech and Audio Processing, v.10, no. 8, Nov. 2002, pp. 620-636.

[8] E. Eide, A. Aaron *et al*, "Text-to-Speech: Bridging the Flexibility Gap Between Humans and Machines", in SpeechTek West 2006, San Francisco, USA, 2006.