Spectral Transformation of Singing Vowels by Dynamic Frequency Warping

Minghui Dong, Paul Chan, Ling Cen, S. W. Lee
Institute for Infocomm Research (I2R), A*Star, 1 Fusionopolis Way, Singapore 138632
{mhdong, yzchan, lcen, swylee}@i2r.a-star.edu.sg Tel: +65-64082757

***Abstract*— There has been an ongoing demand for singing voice transformation. Most existing singing voice synthesis methods are focused on implementing the correct fundamental frequency (F0) trajectory. However, to produce high quality singing voice, the spectrum of the singing voice should also be carefully considered. In this research we look at the spectral differences of the singing voice at different fundamental frequency levels. To investigate the transformation of spectrum in singing transposition, we propose to use the Dynamic Frequency Warping (DFW) approach to learn the mapping of the spectrum curves of vowels at different music note level. From our experiment, we found that a natural sounding singing vowel can be generated by warping its spectrum according to a target vowel.**

I. INTRODUCTION

Singing voice synthesis and transformation is an important topic in a variety of entertainment applications. It has drawn a lot of research interests [1-7]. There is a popular demand for singing processing, such as the conversion of speech into a singing voice, changing the melody of a singing voice.

When synthesizing singing voice from pre-recorded speech or singing signal, the melody of the song is realized by changing the original pitch of the voiced phone fragments. The melody line may be initially humanized by modifying F0-contour to portray vibrato, overshoot, and similar artifacts [3]. However, the F0-only modification cannot achieve a natural sounding singing in some cases. Some degree of unnaturalness is still perceived especially where there is a great change in pitch, e.g., when transposition factor is more than one octave. This is attributed to the differences in spectra between natural utterances of different pitch. Because of this, spectral characteristics need to be taken into consideration in the synthesis of natural singing.

With the importance of the spectrum in the singing voice, there have been several works on the modification of the spectrum of the singing voice. One significant finding of previous research is an additional formant in the singing voice [8]. In speech to singing conversion scenarios, this may be simulated by enhancing the frequencies of the corresponding band [3]. Other spectral transformation methods include spectral mean shift and variance scaling, and weighted linear transformation [4]. However, their improvements are very limited.

In this research, we will investigate whether it is possible to transform the original singing voice at a certain F0 level to a natural singing at a new level. This is done by using frequency warping method. Frequency warping method has been used before in speech recognition to normalize speech features [9-12] and in voice conversion to transform a speaker's identity [13]. Little research has been carried out on the spectral change of the singing voice by frequency warping. We will use the frequency warping approach to transpose a voiced phone from one frequency to another. The warping rule will be learned from data by Dynamic Frequency Mapping.

II. PROBLEM IN SINGING CONVERSION

Let's first look at the spectral differences for the same sound at different F0 levels. In our research, Tandem-STRAIGHT [14] is used to examine the spectrum. Due to the complexity of the singing voice, we will focus our experiments on the simple vowel 'a'.

STRAIGHT decomposes speech into three parameters: an interference-free spectrogram, an aperiodicity map, and a fundamental frequency (F0) trajectory. By changing the three parameters, and reconstructing the signal with the changed parameters, we are able to change the voice signal. For example, if we want to transpose the singing voice from one pitch to another, we might change only the F0 trajectory and re-synthesize the signal. In this work, we will try to modify the spectrogram of the voice in addition to F0 trajectory.

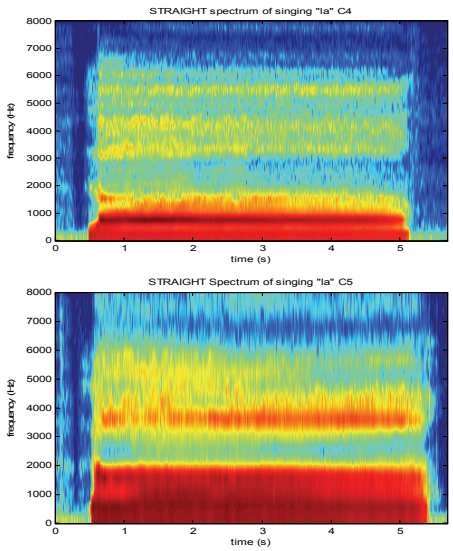

Fig. 1 Comparison of spectrograms of the syllable 'la' sung at C4 and C5 respectively.

In STRAIGHT spectrum, harmonic peaks have been removed, which is different from spectrum of Short Time Fourier Transform (STFT). Hence, the STRAIGHT spectrum actually represents the spectrum envelope of STFT spectrum. In this paper, spectrum shall mean STRAIGHT spectrum.

To understand the differences of STRAIGHT spectrum at different F0 level, we look at the syllable 'la', which is sung by a female singer at frequencies C4 (F0 = 261.63 Hz) and C5 (F0 = 523.25 Hz). The spectra of the two sounds are shown in Fig. 1. Though the general conclusion on spectrum envelope is that it reflects the identity of a phoneme and is independent of pitch [15], from the figure we notice that there is a significant difference between the spectra of the vowel 'a' in the two cases. For example, the spectral width of the dominant band of 'la' sung at C5 is clearly larger than that at C4. Hence, it is important to pay attention to spectral considerations if we want to generate 'la' sung at C5 by transforming that sung at C4.

In the work, we use spectral warping to do spectrum transformation. It changes the spectrum of a signal by using a certain mapping of power of a frequency in a source signal to that of another frequency in a target signal. Frequency warping methods have been previously used in speech recognition to normalize speech features, and used in voice conversion to change the identity of a voice and make it sound like a different speaker.

In our research, we attempt to find out if frequency warping works well in transforming spectrum when attempting to preserve the naturalness of a voice while transposing it from one pitch to another.

III. Spectral Modification Method

We will describe the spectral modification method in the section.

A. Overview

Given two singing signals from the same singer, the source signal A sung at frequency level p1 and target signal B sung at frequency level p2, we wish to convert signal A to make it sound like signal B. The conversion is done with the following steps:

1. Time domain alignment of the two signals: As the two signals may not be in the same length, and unsynchronized, in order to make the conversion, we first need to match the frames in A and B. This is done with a Dynamic Time Warping (DTW) process [16].
2. Time domain conversion: Once we have got the alignment information by DTW, we can construct a new signal A1 from signal A (by sampling or repeating the frames in A) to make its length the same as signal B's.
3. Frequency domain alignment: From each frame pairs from A1 and B, we map the spectra of the two corresponding frames with Dynamic Frequency Warping (DFW) method. Thus a frequency warping function is generated for each frame pair.
4. Frequency domain conversion: Now we construct a signal A2 by warping the frequency of each frame in A1 in the spectral parameter with the frequency warping function for this frame.

B. Time Domain Alignment

Time domain alignment is done using the DTW method, in which MFCC feature is used for alignment.

To reduce the mismatch between the two signals, we have normalized the feature with the following way:

$$F_n = (F_r - \mu)/\sigma, \quad (1)$$

where F_n and F_r are normalized and original feature vectors, μ and σ are mean and standard deviation of feature vectors respectively.

Alignment with DTW is carried out on a similarity matrix to achieve the best possible mapping between two sequences. In this alignment, similarity s between two frames a and b from two signals is calculated with cosine distance as follows:

$$s = (F_a \cdot F_b)/(\|F_a\|\|F_b\|) \quad (2)$$

where F_a and F_b are features from frame a and b.

C. Frequency Domain Alignment

Frequency domain alignment is performed with DFW [10, 12] based on the STRAIGHT spectra of the two frames that are to be aligned. DFW is analogous to DTW in the frequency domain. For the case of single vowel, the frames in the stable part of the vowel can be considered have the same spectrum.

In our work, the sample-rate of the singing signal is 16 kHz, and the default dimension of the STRAIGHT spectrum is 513 corresponding to frequencies 0 to 8 kHz. The power is first converted into dB scale. Then normalization of the spectrum is performed according to the following equation:

$$a_i = a_i / (\max_{j=1}^{n}(a_j) - \min_{j=1}^{n}(a_j)) \quad (3)$$

where a_i is the i-th element of the spectrum curve, n is the number of dimensions.

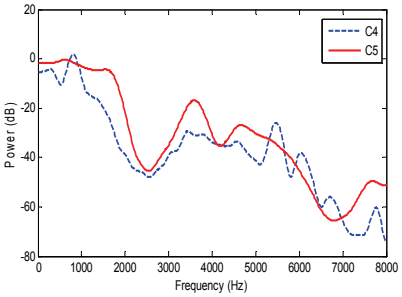

Fig. 2 Comparison of spectra of singing vowel "a" at frequency level C4 and C5.

The spectrum curve is divided into frames with a window size of 20 and a window shift of 5. In this alignment, similarity between two windows is calculated with Euclidean distance.

Fig. 2 shows the spectra of two stable frames from the vowel 'a' at C4 and C5. If the two spectra are used to generate a sound at C5, there is a notable difference. The signal

generated using C5 pitch trajectory and C4 spectrogram contains a hoarse voice. The purpose of the DFW is to find the mapping between the spectra of C4 and C5, so that we are able to minimize the imperfectness when generating C5 signal by modifying C4 spectrogram.

Fig. 3 shows a sample frequency warping function that is obtained by DFW. In the figure, the reference line represents the mapping function for an unwarped spectrum. From the figure, we can see that there is a compression at lower frequency band, and there is an elongation at upper band. Although there is only a small warp in the spectrum for this voice signal, we are able to notice a significant improvement of the synthesized voice.

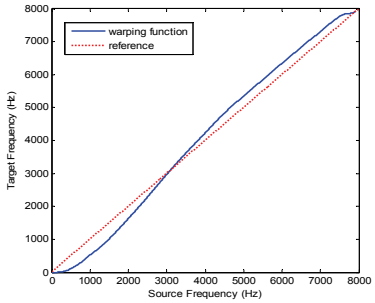

Fig. 3 Frequency warping function obtained by DFW

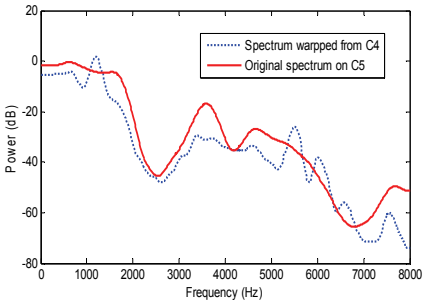

Fig. 4 Comparison of warped spectrum generated from the vowel "a-C4" with the target spectrum of the vowel "a-C5".

D. *Conversion of Singing Signal by Frequency Warping*

When the DFW mapping function is available, we can generate spectrum parameters by warping the frequency of the source sound. Tandem-STRAIGHT needs three elements to synthesize the new sound. For the source sound, we update the spectrum and the F0 curve, but keep the aperiodicity map unchanged when generating the new voice.

Fig. 4 shows the spectra of the stable part of vowel 'a'. From the figure, we can see that the generated spectrum is closer to the target one, compared with Fig. 2.

IV. EXPERIMENT

Spectral modification of singing voice is a complex problem. To simplify the problem, we start with the vowel 'a'.

The data in the research is recorded in a sound proof studio. A female singer participated in the recording. She was asked to sustain the syllable 'la', starting from the lowest note she could comfortably sing to the highest one. A keyboard was used to provide a reference sound of each note before she sang it. The lowest note we recorded is E2 and the highest is C6 for this singer.

E. *Transposition Test*

In our research, we tried to convert the voice from one frequency to another. As we know, when the pitch of voice is raised 2 times or more, there will be some notable unnaturalness in the generated voice. Therefore, in our experiment, we would test whether we can improve the quality of generated voice by spectral modification when the pitch is one octave higher, i.e. the fundamental frequency is doubled.

TABLE I
TRANSPOSITION OF SINGING VOWELS TO HIGHER FREQUENCY
(TRANSFORMATION IS DONE FROM SOURCE SOUND TO TARGET SOUND)

Source voice		Target voice	
Note level	Frequency	Note level	Frequency
A3	220.00	A4	440.00
B3	246.94	B4	493.88
C4	261.63	C5	523.25
D4	293.66	D5	587.33
E4	329.63	E5	659.26
F4	349.23	F5	698.46
G4	392.00	G5	783.99
A4	440.00	A5	880.00
B4	493.88	B5	987.77

We chose 9 recorded singing of 'la' as source voices, of which each recording is sung at a different frequency as shown in Table I. Each source recording is also paired with its one octave higher recording as target voice.

Using the target recordings as references for spectral transformation, 9 pairs of voice files are generated with the following two methods:

- Method A: Among the three STRAIGHT parameters calculated from the source sound, only the pitch trajectory is modified. The spectrum and aperiodicity map are kept the same. This is the baseline method.
- Method B: Both the pitch trajectory and the spectrum are modified. The frequency warping method is used to transform the spectrum. The aperiodicity map is kept the same. This is the proposed frequency warping method.

TABLE II
PAIR-WISE PREFERENCE TEST RESULT

Preference	Score
A	18.9%
No preference	35.5%
B	46.7%

A pair-wise preference test (PPT) was conducted to evaluate the performance of the proposed spectral transformation method B against the baseline method A. Listeners were asked to select which stimulus from the two methods achieves better quality. Each of the listeners was asked to select their preferences on the two sounds and was presented with 3 choices, 'A', 'B' or 'no preference'. 10 listeners participated in the listening test, and 90 responses were collected from the test. The listening test result is as shown in Table II. From the table, we can see that there is a

preference rate of 46.7% for the proposed method as opposed to 18.9% for the baseline method. This verifies that the proposed method helps to improve the quality of the singing vowel ‘a’ in the one octave transposition test.

We also examined the frequency warping function of each transformation in the test as shown in Fig 5. From the figure, we can see that most of the warping functions present only a slight warp in the spectrum. For lower frequency notes, the warping is not significant, while for higher frequency notes, the warping is more obvious. In general, the warping functions cannot be represented in a straightforward manner by a simple curve.

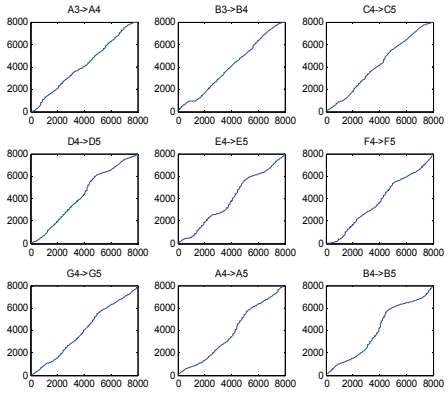

Fig. 5 Frequency warping curves for transposition test
(x-axis represents the source frequency, while y-axis represents the target frequency)

F. Discussion

This work is an attempt to transform the spectrum of singing voice by learning from source and target voices. Though this work is an experiment on a vowel of a particular singer only, from this work, we had a number of observations of the problem, some of which are described below.

Clear warping of the spectrum is not always found in all transpositions. In such case, the warping method does not help much in improving the voice quality. For these cases, we will need to investigate what other method can be applied to improve the spectrum of the signal.

In the frequency warping process, we used Euclidean distance to measure the spectral differences. However, the acoustic difference may not fully reflect the perceptual difference accurately. Therefore, the warping function can be further improved by using better distance measure.

Due to frame-to-frame differences, there are variations between adjacent frames. This results in large variations between warping functions between frames as well. To make the warping function stable across frames, warping functions were smoothed across frames. The way to smooth the warping functions can be further improved.

This work focused on transformation of vowel ‘a’ only. The results obtained from the experiment need to be evaluated and verified on other phones as well. We also wish to test the method on other singers’ voice.

V. CONCLUSION

Spectral envelope transformation is a necessary process during the transposition of singing voices. This research investigated the possibility of using frequency warping method to transform the spectrum of singing vowels by the machine learning method. We used the dynamic frequency warping method to establish the mapping of the source singing to the target singing. The experiment shows that there is a possibility of generating better singing voices with the proposed spectral transformation method.

REFERENCES

[1] Saino, K. Zen, H. , Nankaku, Y., Lee, A. and Tokuda, K., “HMM-based singing voice synthesis system,” in Proc. Interspeech, pp. 1141-1144, Sep. 2006.

[2] Kenmochi, H. and Ohshita, H. “VOCALOID – Commercial singing synthesizer based on sample concatenation,” in Proc. Interspeech, Aug. 2007.

[3] Saitou, T., Goto, M., Unoki, M., and Akagi, M., “Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices,” in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 215-218, Oct. 2007.

[4] Nwe, T. L., Dong, M., Chan, P. Wang, X. , Ma, B. and Li, H., “Voice conversion: From spoken vowels to singing vowels,” in Proc. ICME AdMIRe Workshop, pp. 1421-1426, Jul. 2010.

[5] Bonada, J. et al., “Synthesis of the Singing Voice by Performance Sampling and Spectral Models,” IEEE Signal Processing Magazine, Vol. 24, Iss.2, pp. 67-79, 2007.

[6] Akagi, M., “Rule-based voice conversion derived from expressive speech perception model: How do computers sing a song joyfully?” in Proc. ISCSLP, Tutorial 01, Nov. 2010.

[7] Saitou, T. et al., “Analysis of acoustic features affecting “singingness” and its application to singing voice synthesis from speaking voice,” Proc. ICSLP 2004, Vol. III, pp. 1929-1932, 2004.

[8] Sundberg, J. “Articulatory Interpretation of the ‘Singing Formant’,” J. Acoust. Soc. Am., Vol. 55, pp. 838-844, 1974.

[9] Lee, L. and Rose, R., “A frequency warping approach to speaker normalization”. IEEE Trans. Speech Audio Process. 6 (1), 49–59. 1998.

[10] Umesh, S., Cohen, L. and Marinovic, N., “Frequency-Warping in Speech,” 414 – 417, Proc. ICSLP 1996.

[11] Zhan, P. Westphal, M., “Speaker normalization based on frequency warping,” ICASSP 1997.

[12] Neuburg, E.P, "Frequency warping by dynamic programming”, ICASSP 1998.

[13] Toda, T, Saruwatari, H, Shikano, K, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of straight spectrum", ICASSP 2011.

[14] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H., “TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation,” in Proc. ICASSP, pp. 3933-3936, Mar. 2008.

[15] Schwarz, D. “Spectral Envelopes in Sound Analysis and Synthesis”, Diplomarbeit Nr. 1622, Universit¨at Stuttgart, Fakult¨at Informatik, Stuttgart, Germany, 1998.

[16] Sakoe, H. and Chiba, S., “Dynamic programming algorithm optimization for spoken word recognition”, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, no. 1, pp. 43- 49, 1978.