VOICE CONVERSION: FROM SPOKEN VOWELS TO SINGING VOWELS

Tin Lay Nwe, Minghui Dong, Paul Chan, Xi Wang, Bin Ma, and Haizhou Li

Institute for Infocomm Research (I²R), A*STAR, 1 Fusionopolis Way, Singapore 138632
Emails: tlnma@i2r.a-star.edu.sg; mhdong@i2r.a-star.edu.sg; ychan@i2r.a-star.edu.sg;
wangxi@i2r.a-star.edu.sg; mabin@i2r.a-star.edu.sg; hli@i2r.a-star.edu.sg

ABSTRACT

In this paper, a voice conversion system that converts spoken vowels into singing vowels is proposed. Given the spoken vowels and their musical score, the system generates singing vowels. The system modifies the speech parameters of Fundamental frequency (F0), duration and spectral properties to produce singing voice. F0 contour is obtained using F0 fluctuation information from training singing voice and music score. Duration of each vowel of speech is stretched or shortened according to the length of the corresponding musical note. To transform speech spectrum to singing spectrum the following two approaches are employed. The first method employs spectral mean shifting and variance scaling method. And, the second approach uses weighted linear transformation method to transform speech to singing spectrum. The system is tested on the database including 75 speech and 30 singing voices sung using vowels. The results show that the proposed system is able to convert spoken vowels into singing vowels with a quality very close to the target singing voice.

***Keywords*—** Synthesis, music, speech, singing, voice conversion

1. INTRODUCTION

The synthesis of the singing voice is the artificial production of human-like singing voice. There are many different sets of inputs to the singing synthesis systems. Notational information from a musical score, for example, may be digitally represented in Standard MIDI Format (SMF) or Common Music Notation. Musical score is a printed version of a musical arrangement in notational form which may include lyrics, notes, supplemental text, etc,. Lyrics may be given as ordinary text or phonetic notation and can be either a set of inputs to the singing synthesis system [1], or, in the case of SMF, be incorporated into the digitized notational information. In this case, synthesis system produces singing voice of the given lyrics using the voice of predefined singer [2, 3, 4]. As an alternative, the inputs can be a given performance also represented in some digital format, the performance of a human singer or an instrumental performance. In this case, synthesis system generates the singing voice which mimics the given performance [5].

The speech to singing voice conversion system proposed in this paper is to produce human-like singing voice when the speaking voice of the lyric of the song, the musical score, and synchronization information between each vowel and corresponding musical note are available. The system can also be described as a "voice conversion system" that converts spoken vowels to singing vowels for a particular speaker. This system is useful for developing practical applications in computer-based music production, such as pitch-shifting and automatic tuning systems, where, currently, the pitch of the singing voice may only be manipulated (corrected or intentionally modified) at the expense of its naturalness [6]. The advantage of a voice conversion system is its ability to manipulate singing voices while keeping their naturalness. Additionally, this system is also able to produce pleasant sounding singing voice in the subject's voice if he is not good at singing. Hence, this system is also applicable to karaoke systems.

Although several works have been reported on text-to-singing (lyrics-to-singing) synthesis systems [2, 3, 4], only a few studies have been done on speech-to-singing (reading lyric-to-singing) synthesis [6]. For speech to singing conversion, the system in [6] controls the three acoustic features including the fundamental frequency (F0), phoneme duration, and spectrum. Given the musical score and its tempo, the F0 contour of the singing voice is generated by controlling four types of F0 fluctuations: overshoot, vibrato, preparation and fine fluctuation. The duration of each phoneme in the speaking voice is stretched by using the duration of its musical note. Finally, the spectral envelope of the speaking voice is imposed onto the singing voice by controlling both the singing formant and the amplitude modulation of formants in synchronization with the vibrato.

Several research works have been done to convert a source speaker's speech to sound as if it is spoken by a target speaker [7, 8]. As an example, female's voice is converted to male's voice. In [7], a technique to convert a source speaker's speech to a target speaker's speech is proposed. The technique employs a perceptually weighted linear transformation approach to minimize the distance

978-1-4244-7493-6/10/$26.00 ©2010 IEEE

between spectral envelopes of source and target speakers. In [8], frequency warping approach is employed o achieve an acoustic spectrum of target speaker based on the source speaker's acoustic spectrum. A frequency warping function is generated by mapping formant parameters of the source speaker and the target speaker.

Based on the above research works, the important issue in voice conversion is spectral transformation. In [9], multi-resolution spectral transformation method is proposed for music analysis such as melody extraction or multiple pitch estimation which rely on a spectral representation of the audio signal. In our proposed approach, we use linear resolution to extract spectral information from speech as linear resolution is good enough to represent the speech spectrum. And, the same resolution is kept when speech spectrum is converted to singing spectrum. Regarding speech to singing spectral transformation, applying linear transformation to spectral envelope is the most popular approach on voice conversion. Linear transformation approaches proposed in [10, 11] outperforms other approaches in terms of voice quality.

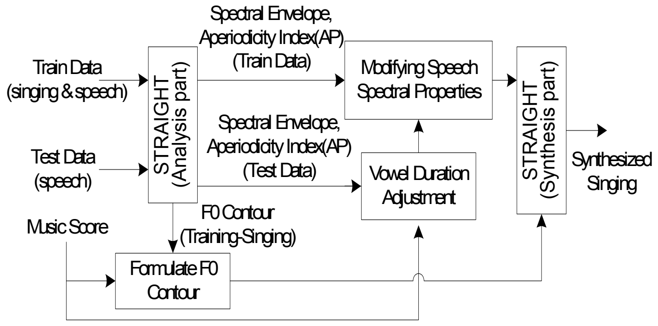


Fig. 1. Speech to singing voice conversion system

In this paper, we propose two new approaches for speech to singing spectral transformation. In these methods, the spectral envelopes of singing are obtained based on the parameters learned from training singing sample. In the first approach, Spectral Mean Shifting and Variance Scaling (SMS-VS) method is employed to adapt the spectral properties of speech to that of singing. In the second approach, Weighted Linear Transformation (WLT) method is adopted to transform the speech to singing spectrum. In addition to spectral transformation, duration of the vowels and Fundamental Frequency (F0) of speech are modified as follows. The duration of the each vowel in speech is lengthened or shortened according to the duration of the musical notes in music score similar to the process as in [6]. Music score in our experiments includes note information for each vowel. Fundamental frequency (F0) of the synthesized singing is obtained using the note information in musical score and information of F0 variations from training singing samples. The proposed system is built on the speech manipulation platform *STRAIGHT* [12] and the system block diagram is shown in Figure 1.

This paper is organized as follows. Section 2 describes the description of the database used in the experiments. In Section 3, formulating the Fundamental frequency, modifications of duration and spectral properties of speech are detailed. In Section 4, experimental results are reported. Finally, Section 5 concludes the paper.

2. SPEECH AND SINGING DATA

A database which includes speech and singing voices were built in the recording studio. Before recording, the musical score was prepared. The vowel 'ah' was used to record for both the reading (where 'ah' was recited to the rhythm of the stanza) and singing while the original lyrics of the song were used as visual cues. Figure 5 shows the vowels and notes to be sung as well as the original lyrics. A total of 5 lines were extracted from the songs listed in Table 1.

Table 1. List of the guided lyrics and songs from which the lyrics were extracted.

No .	Line	Song Title
1	Twinkle twinkle little star	Nursery rhymes – Twinkle Twinkle Little Star
2	I find her standing in front of the church.	Michael Learns to Rock - 25 Minutes
3	Boy I miss your kisses.	Michael Learns to Rock - 25 Minutes
4	Hickory dickory dock	Nursery Rhymes – Hickory Dickory Dock
5	Happy birthday to you	Happy Birthday song

Three subjects participated in the recordings. Each subject was tasked to sing each line twice in 'ah's and similarly speak the rhythm of each line 5 times also in 'ah's. Singing voices were to be used as the reference and in training the system. A total of 75 speech and 30 singing voices were recorded from these subjects. The subjects freely practiced several times before the actual vocal recording. Details of the vowels and their corresponding notes are presented in Figure 5. Each vowel of both the spoken and sung recordings were manually segmented and associated with a musical score file in preparation for experimental analysis. We used the leave-one-out cross validation method to train and test the system.

3. SPEECH PARAMETER MODIFICATION

Although speech and singing are produced from the same vocal organs, there are many differences in their physical characteristics. The parameters such as F0, phone duration and spectral envelopes have the largest differences between two modes of voicing. The following sub-sections study the differences of these parameters in speech and singing as well as modification of these parameters to convert speech to singing voice.

3.1. Fundamental frequency (F0)

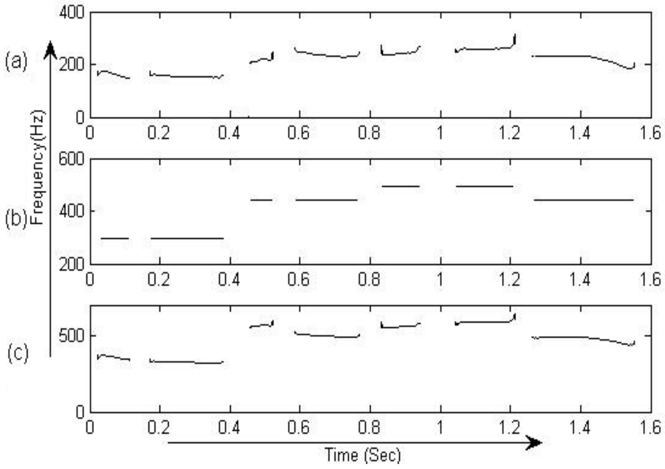


Fig 2. F0 contours of (a) training singing voice, (b) music score file and (c) synthesized singing voice of guided lyric 'twin-kle-twin-kle-lit-tle-star'.

When singing, notes are changed according to the melody without changing personalized singing style such as vocal timbre. Therefore the characteristics of singing voice are more dynamic and complicated than those of a speaking voice. In singing voices, F0 contours correspond to the melody of a song, and include several types of fluctuations such as overshoot, vibrato, and fine fluctuations [13]. To include these fluctuations in F0 contour of singing voice the following processes are carried out. The F0 contour of speaking voice is totally discarded. And, F0 contour of the singing voice is generated from musical score file. To include the effect of fluctuations in the F0 contour of singing voice, the F0 contour of singing voice of training sample is extracted using the *STRAIGHT* [12]. Then, the contour of fluctuations, $F0_{\text{fluct-singing,}}$ is determined from the F0 contour of the singing voice of training sample using equation (1).

$$F0_{\text{fluct-singing}}=F0_{\text{singing}}-E[F0_{\text{singing}}] \qquad (1)$$

where $E[F0_{\text{singing}}]$ is the mean of the F0 of training singing voice. Then, the effect of fluctuation is added to the note of individual vowel in the music score file as follows to obtain the F0 contour of the synthesized singing, $F0_{\text{synthesized-singing}}$.

$$F0_{\text{synthesized-singing}}=F0_{\text{music-score}}+F0_{\text{fluct-singing}} \qquad (2)$$

Figure 2 shows F0 contours of training singing voice sample, music score file and synthesized singing voice.

3.2. Vowel duration

The duration of vowels in speaking voice is different from that of vowels in singing voice. Hence, the duration of the vowels in speaking voice is lengthened or shortened according to the duration of the corresponding musical note. In this process, the duration of the vowel in speaking voice is evenly stretched or shrunk. Figure 3 shows the vowel durations of the synthesized singing voices adjusted to that of training singing voice.

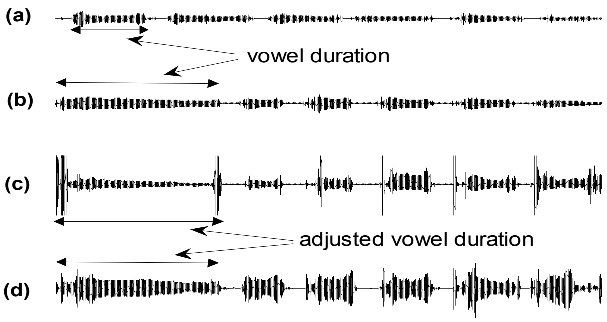


Fig 3. Waveform representations of guided lyric 'Boy I miss your kiss- es' for (a) speaking voice, (b) singing voice (c) synthesized singing voice of SMS-VS method and (d) synthesized singing voice of WLT method. Duration adjustments of the first vowel of speech to the first vowel of singing for the synthesized singing voices are shown.

3.3. Spectral envelope

Spectral envelopes of speaking and singing voices are different even if speaking and singing are produced by a same vocal system (ie, same speaker). The differences are as follows. Singing voice has a remarkable peak called the "singing formant" near 3kHz [14]. And, an aperiodicity index (AP) which can be estimated by using the analysis part of the speech manipulation system *STRAIGHT* [12] has a dip near 3kHz [6]. In addition, the formant amplitude of a singing voice is modulated in synchronization with the frequency modulation of each vibrato in the F0 contour [15]. Hence, speech spectrum is modified to minimize its distance to the spectrum of singing voice. To achieve this, spectral envelope and aperiodicity index (AP) [12] of speaking voice is modified. The following are the details. The spectral envelopes and aperiodicity index (AP) of the speech and singing training samples are extracted using *STRAIGHT* [12]. Then, the spectral envelope and AP of speaking voice are modified to achieve the singing spectral properties for synthesized singing voice. Two approaches are used in modifying them. The first approach employs Spectral Mean Shifting and Variance Scaling (SMS-VS) method. And, the second method uses Weighted Linear Transformation (WLT) approach. Details are explained in the following sections.

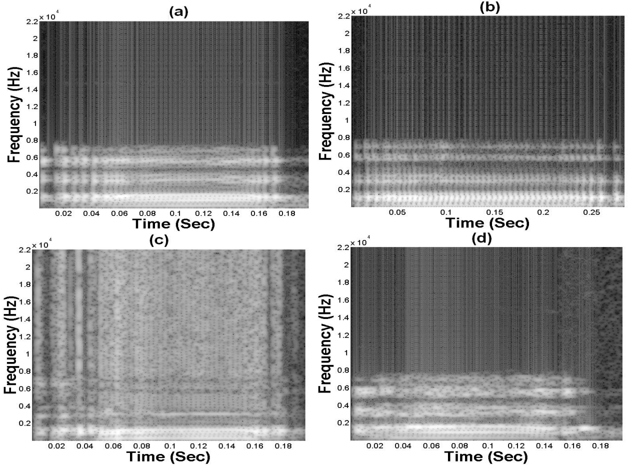


Fig 4. The spectrograms of (a) singing voice (b) speaking voice, and (c) synthesized singing voice with SMS-VS method and (d) synthesized singing voice with WLT method for the third vowel 'ah' of guided lyric 'Boy I miss your kiss- es'

3.3.1. Spectral mean shifting and variance scaling (SMS-VS)

In this process, the spectral envelope of the synthesized singing voice is obtained by spectral mean shifting and variance scaling of the spectral envelope of speaking voice. The transformation of the spectral envelope of speaking voice to that of synthesized singing voice is as follows. First, variance scale, V_{scale}, is determined as follows.

$$V_{scale} = G_{singing\text{-}train}/G_{speak\text{-}test} \qquad (3)$$

where $G_{singing\text{-}train}$ and $G_{speak\text{-}test}$ are the variances of the spectrum of training singing and test speaking voices respectively. Then, the spectral envelope of the synthesized singing voice ($Z_{synthesized\text{-}singing}$) is obtained by employing meaning shifting and variance scaling to the spectrum of test speaking voice, $X_{speaking\text{-}test}$, as follows.

$$Z_{synthesized\text{-}singing}=(X_{speaking\text{-}test}-U_{speaking\text{-}test})V_{scale}+U_{singing\text{-}train} \qquad (4)$$

where, $U_{speaking\text{-}test}$ and $U_{singing\text{-}train}$ are means of the spectrums of the test speaking voice and training singing voice respectively. To obtain Aperiodicity index (AP) for synthesized singing voice, the above processes are applied to the AP of speaking voice.

3.3.2. Weighted linear transformation (WLT)

In this approach, the spectral weights are estimated by dividing the spectrum of the singing voice by that of speaking voice as follows.

$$W=Y_{singing\text{-}train} / X_{speaking\text{-}train} \qquad (5)$$

where, W is spectral weight, $Y_{singing\text{-}train}$ is the spectral envelope of training singing voice sample, and $X_{speaking\text{-}train}$ is the spectral envelope of the training speaking voice sample. Then, the spectral envelope of the speaking voice is transformed to that of the singing voice by multiplying the spectral envelope of testing speaking voice, $X_{speaking\text{-}test,}$ with the spectral weight W as follows.

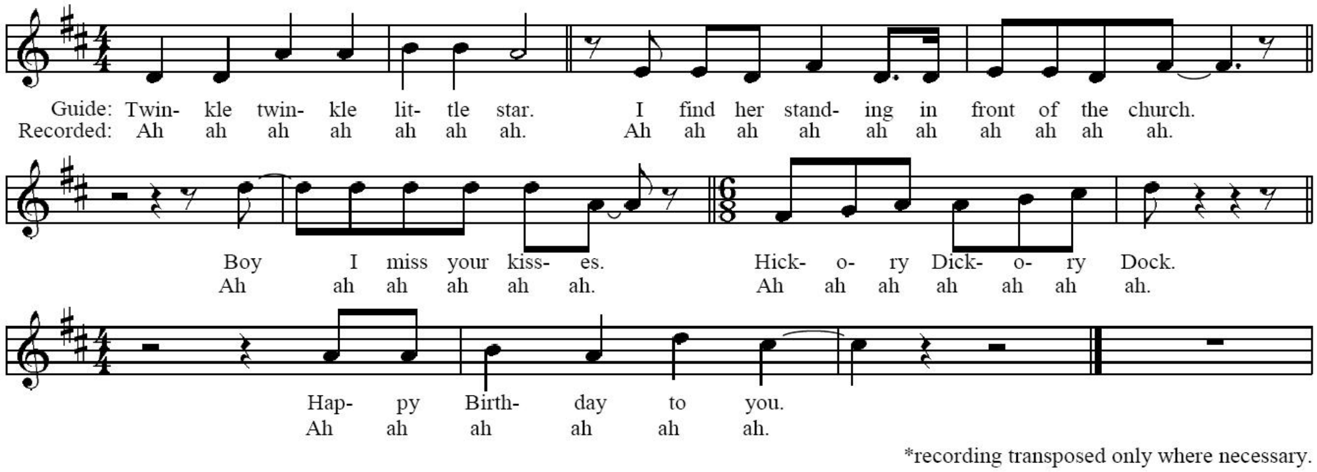
$$Z_{synthesized\text{-}singing} = W\, X_{speaking\text{-}test} \qquad (6)$$

where $Z_{synthesized\text{-}singing}$ is spectral envelope of the synthesized singing voice. To obtain Aperiodicity index (AP) for synthesized singing voice, the above processes are applied to the AP of speaking voice. Figure 4 shows the example spectrograms of target singing vowel, speaking vowel, synthesized singing vowel of SMS-VS method and synthesized singing vowel of WLT method. By multiplying the spectral envelope of speaking voice with the spectral weight W, the spectral envelope of $Z_{synthesized\text{-}singing}$ become near to the spectral envelope of target singing vowel. However, spectral envelope of SMS-VS has spectral distortions especially in the high frequency regions.

4. EXPERIMENTS AND DISCUSSION

The performance of the proposed spoken to singing vowel conversion system is evaluated by conducting speaker dependent singing synthesis experiments. For each speaker, one singing voice of each lyric is used as training data. The leave-one-out cross validation method is used to divide the training and testing data for speaking voices. For example, there are 5 speaking voice samples of each lyric for each speaker. Five rounds of experiments are conducted. In each round, one speaking voice is used as training and four is used as testing. Hence, there are 20 test cases for each lyric for each speaker. A total of 20 X 5 lyrics = 100 test cases for each speaker and hence, 300 test cases for 3 speakers.

Using speech manipulation platform *STRAIGHT* [12], F0, Aperiodicity index (AP) and spectral envelopes are extracted for speech and singing voices. Then, duration, F0, AP and spectral envelopes of the test speech samples are modified as discussed in Section 3. Finally, synthesized singing voices are obtained using modified duration, F0, AP and spectral envelopes information based on STRAIGHT [12]. Processing flow of the system is shown in Figure 1.


Guide: Twin- kle twin- kle lit- tle star. I find her stand- ing in front of the church.
Recorded: Ah ah ah ah ah ah ah. Ah ah ah ah ah ah ah ah ah ah.
Boy I miss your kiss- es. Hick- o- ry Dick- o- ry Dock.
Ah ah ah ah ah ah. Ah ah ah ah ah ah ah.
Hap- py Birth- day to you.
Ah ah ah ah ah ah.

*recording transposed only where necessary.

Fig 5. Musical notes, guided lyrics and vowels used to record speech and singing voices. Each vowel is sung at the corresponding musical note.

The synthesized singing voices are compared between two spectral modification methods: SMS-VS and WLT. The synthesized singing voices from these two methods are compared with reference singing voice to evaluate naturalness and the audio quality. Ten subjects, all graduate students with normal hearing ability, listen the synthesized singing samples using headphones. The naturalness and audio quality of the samples are rated separately using the scales 1 ~ 5. The rating of 1 is the worst and 5 is the best. Table 2 shows the ratings averaged over 100 test cases of each speaker. Examples of synthesized speaking vowels, target singing vowels and synthesized singing vowels generated by SMS-VS and WLT methods can be listened in [16].

Table 2. Average ratings of subjective evaluations for SMS-VS and WLT based singing synthesis methods for 3 different speakers

Speakers	SMS-VS		WLT	
	Naturalness	Quality	Naturalness	Quality
1(Female)	3.1	2.7	3.4	3.7
2(Female)	2.9	2.7	3.5	3.7
3(Male)	3	2.1	3.4	3.5

The results indicate that WLT method achieves the ratings on both naturalness and audio quality close to reference singing. And, WLT method obtains higher ratings for both of the naturalness and audio quality than SMS-VS method. WLT method is able to transform speech spectral properties to singing spectral properties better than SMS-VS method especially for speaker dependent system. The ratings of the naturalness for SMS-VS method are very similar to that of WLT method. However, SMS-VS has lower ratings on audio quality than WLT method. This is because of spectral distortions on the samples of SMS-VS method as shown in Figure 4. In SMS-VS, variance scaling processes results the distortion to the formant amplitude modulation [15] of the synthesized singing.

5. CONCLUSIONS

This paper proposes a voice conversion system that can convert the spoken vowels into the singing vowels for a particular speaker. Speaking and singing voices are recorded by reading and singing the lyrics using vowel 'ah'. The system performs 3 steps: 1) formulating the Fundamental Frequency (F0) of singing voice by integrating the effects such as overshoot, vibrato to the F0 contour of music score file using information from training singing sample 2) adjustment of the vowel duration of speaking vowels using note duration 3) integrating the effects such as singing formant and formant amplitude modulation to the speech spectral properties by using two approaches i) Weighted Linear Transformation (WLT) and ii) Spectral Mean Shifting and Variance Scaling (SMS-VS) processes. The results are compared for two different spectral transformation methods. The subjective evaluation results show that weighted linear transformation method outperforms the spectral mean shifting and variance scaling method. Spectral representations also present that spectrogram of WLT has less distortions compared to that of SMS-VS.

6. REFERENCES

[1] X. Rodet, "Synthesis and Processing of The Singing Voice," *Proceeding of the First IEEE Benelux Workshop on Model-Based Processing and Coding of Audio (MPCA-2002)*, Leuven, Belgium, 2002.

[2] J. Bonada and X. Serra, "Synthesis of The Singing Voice by Performance Sampling and Spectral Models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, March 2007.

[3] YAHAMA Corporation, Vocaloid: "New Singing Synthesis Technology, http://www.vocaloid.com/en/index.html.

[4] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-Based Singing Voice Synthesis System," *Proceedings of INTERSPEECH*, pp. 1141-1144, Pittsburgh. 2006.

[5] T. Nakano, and M. Goto, "Vocalistener: "A Singing-to-Singing Synthesis System Based on Iterative Parameter Estimation," *6th Sound and Music Computing Conference*, pp. 343-348, 2009.

[6] T. Saitou, M. Goto, U. Masashi and A. Masato, "Speech-to-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. pp. 215 – 218, 2007.

[7]H. Ye, and S. Young, "Perceptually Weighted Linear Transformation for Voice Conversion," *Eurospeech*, pp. 2409 – 2412, 2003.

[8] Z-W. Shuang, R. Bakis, Slava Shechtman, D. Chazan, and Y. Qin, "Frequency Warping Based on Mapping Formant Parameters," *Interspeech*, pp. 2290 – 2293, 2006.

[9] P. Cancela, M. Rocamora, and E. Lopez: "An Efficient Multi-Resolution Spectral Transform for Music Analysis," *10th International Society for Music Informational Retrieval Conference (ISMIR)*, pp.309-314, 2009.

[10] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion," *IEEE Trans. On Speech and Audio Processing*, vol. 6, no. 2, pp. 131-142, 1998.

[11] A. Kain, " High Resolution Voice Transformation", *PhD dissertation*, OGI, 2001.

[12] H. Kawahara, I. Masuda-Katsuse, and A**.** Cheveigné, "Restructuring Speech Representations Using A Pitch-Adaptive Time-Frequency Smoothing and An Instantaneous-Frequency-Based F_0 Extraction: Possible Role of A Repetitive Structure in Sounds," *Speech Communication*, vol.27, pp.187–207, 1999.

[13] S. Takeshi, U. Masashi, and A. Masato, "Development of an F0 control Model Based on F0 Dynamic Characteristics for Singing-Voice Synthesis," *Speech Communication,* 46 (2005) 405-417.

[14] J. Sundberg, "Articulatory Interpretation of the Singing Formant," *J. Acoust. Soc. Am*., vol. 55, pp. 838-844, 1974.

[15] P. B. Oncley, "Frequency, Amplitude and Waveform Modulation in the Vocal Vibrato," *J. Acoust. Soc. Am*., vol. 49, Issue 1A, pp. 136, 1971.

[16]http://www1.i2r.a-star.edu.sg/~mhdong/singing/synthesizedsample.htm