A Research of Automatic Composition and Singing Voice Synthesis System for Taiwanese Popular Songs

Chih - Fang Huang
Department of Information Communications, Kainan University, No.1
Kainan Road, Luzhu Shiang,
Taoyuan 33857, Taiwan
jeffh.me83g@gmail.com

Wei - Gang Hong
Master Program of Sound and Music
Innovative Technologies
National Chiao Tung University,
1001 Uni-versity Road, Hsinchu,
Taiwan 300, R.O.C.
weigan99@gmail.com

Min – Hsuan Li
Master Program of Sound and Music
Innovative Technologies
National Chiao Tung University,
1001 Uni-versity Road, Hsinchu,
Taiwan 300, R.O.C
tmac790601@hotmail.com

ABSTRACT

The paper discussed the integration of the automatic composing and singing voice synthesis systems, to let the computer can compose a new song and sing in Taiwanese. First, the automatic composer system analyzes 10 Taiwanese popular songs through a first-order Markov chain and establishes the probability transition matrixes of the pitch and the duration. Second, the singing synthesis is based on STRAIGHT algorithm and 509 Taiwanese basic syllables are analyzed to build a text-to-singing (TTSI) synthesis system. Finally, the MIDI music files which are produced by automatic composer system and lyrics are fed into TTSI synthesis system to synthesis a new song. In order to improve naturalness, the pitch curve adds the vibrato and fine fluctuation.

1. INTRODUCTION

Algorithmic Composition refers to the use of mathematical models or music theory rules to allow the computer to automatically generate music or assist people to compose. Arnold Schoenberg first introduced the idea of Twelve-tone technique [1] in the late 20th century. Then further developed by Anton Webern and other successors, they characterized the music and extract the pitch, duration and volume as parameters to train the model and produce melody. Therefore, we can generate the music in totally random way. But this simple way of automatic composition doesn't fit the rule of music theory. Therefore, the automatic composition should use the result of analysis to produce music. For example, Lejaren Hiller and LeonardIsaacson use music theory rules to enter stylized musical parameters and automatically generate the stylized music through the program, modify and select three steps [2]. Iannis Xenakis use mathematical models and Sieve theory to complete the Formulized Music for automatic composer [3]. Phil Winsor also uses Sieve theory and mathematical algorithms to develop the Composer's Toolbox and Music Sculptor, which are two kinds of automatic composition software.

After the note generation from automatic composition system, the selection of timbre or musical instruments is considered. Sound synthesis technology provides contribution of the note sounds, such as piano, violin, guitar, etc. Synthesis of these instruments has been developed for a long time and is already quite mature. However in the General MIDI, it only offers /a/ and /o/ human voice in 127 kinds of sounds. This simple synthetic sounds unable to synthesize singing voice with semantics but text-to-singing (TTSI) synthesis system can be used to solve this problem. User can enter text and music information to generate synthesized voice by TTSI synthesis system, such as Japan YAMAHA Company launched in 2004 commercial software VOCALOID [4]. In VOCALOID, users not only can write the lyrics to synthesize singing , but also tune the ten acoustic parameters to adjust the naturalness of synthesized singing voice, for example, vibrato, velocity(VEL), clearness(CLE), pitch bend(PIT), etc. However the complexity of the interface causes it only suitable for professional arrangement. And it doesn't provide Taiwanese.

Many scholars developed Mandarin TTSI synthesis system in Taiwan, such as the Professor Jang [5], Professor Gu [6], etc. But few people study in the Taiwanese [7], the main reason is that Taiwanese writing on is not easy and not as a unified national language text representation. But there are more than 1,500 people in Taiwan speaking Taiwanese, accounting for about 70 percent of Taiwan's population [8]. It shows Taiwanese songs of a certain size in Taiwan market, and Taiwanese have a symbolic spiritual heritage of indigenous culture and arts, such as Taiwanese opera and Nanguan. Therefore, the development of Taiwanese TTSI synthesis system is necessary indeed, particularly for recent years, the government in the promotion of local culture positive, and this system may also be able to teach the singing has been applied.

The paper is organized as follows. In Section 2, the background knowledge of the Taiwanese, automatic composing and singing synthesis is described. In Section 3, system and process integration approach are described. In Section 4, experimental evaluations are described. Finally, we summarize this paper in Section 5.

2. RELATED RESEARCH

2.1 Taiwanese

Taiwanese is one of the major common language and also a tonal language like Mandarin and Hakka. Each word consists of only one syllable with the basic tone pronunciation unit, and each syllable is composed by a consonant and vowels. According to the investigation from Ministry of Education, Taiwanese has 17 consonants, 75 vowels and eight kinds of tone and the number of Taiwanese basic syllables are 509 which can be used as synthesis unit.

2.2 Automatic Composition Method

From past studies automatic composition system to learn the rules and produce music can be divided into the implicit rules and explicit rules two methods [9]. The implicit rules need to learn the rules from other music and produced music will have the similar style. On the contrary, explicit rules are based on music theory or experience to set the rules and the user can specify parameters to produce various musical parameters consistent with their music. Theoretically, using explicit rules is based on music theory and not biased. But we are not composers and composers usually create music with experience. In addition, there are no absolutely rules for composition in music theory, especially tonality, form and harmony are difficult to describe with rules. However, implicit rules are based on the extracted musical features to establish the database and rules. It can produce similar style with original music. Hence, analysis of the original score, the more detailed the resulting sample database and composer guidelines will be clearer. Since our main purpose is to obtain the Taiwanese popular songs music, so in this paper we will use the first-order Markov chain algorithm, which is a kind of implicit rules, to derive the implicit rules of composition and the output of the music recorded on MIDI music files.

2.2.1 First-order Markov chain

A Markov chain is a mathematical system that undergoes transitions from one state to another on a state space. It is a random process usually characterized as memoryless: the next state depends only on the current state and not on the sequence of events that preceded it. Markov chains have many applications as statistical modelsof real-world processes. Application of music [10], we assume that each note as state, a state between another states is to change the melody. This relation is defined in the following equations:

$$\begin{aligned} &P(X_n = j_n \mid X_0 = j_0, \ldots, X_{n-2} = j_{n-2}) \\ &= P(X_n = j_n \mid X_{n-1} = j_{n-1}) \end{aligned} \qquad (1)$$

$$\mathbf{P} = \begin{array}{c|ccc} & \mathbf{1} & \mathbf{2}\cdots & \mathbf{M} \\ \hline \mathbf{1} & \mathbf{p}_{11} & \mathbf{p}_{12}\cdots & \mathbf{p}_{1M} \\ \mathbf{2} & \mathbf{p}_{21} & \mathbf{p}_{22}\cdots & \mathbf{p}_{2M} \\ \vdots & & \vdots & \\ \mathbf{M} & \mathbf{p}_{M1} & \mathbf{p}_{M2}\cdots & \mathbf{p}_{MM} \end{array} \qquad (2)$$

where **P** is probability transition matrix. So if we want to randomly generate a particular musical style of music, we can collect and statistic for this style of music. By establishment of a probability transition matrix, which selects the next occurrence of the notes based on this probability, the melody produced can exhibit this particular style of music.

Every note in score contains three musical parameters: pith, duration and volume. Among these parameters, pitch and duration of the most affected auditory perception. Hence we take pitch and duration as a joint random variable in a state, and then create the probability transition matrix. However, this method is the minimum degree of randomness, but the maximum capacity to imitate musical styles.

2.3 Singing Voice Synthesis

The main voice synthesis methods are divided into concatenative synthesis and parametric synthesis [11].

The former method needs to pre-recorded sounds of different pitch and length for the large corpus. And according to the information of score, the corresponding speech waveforms are selected and modulated by some signal process technique such as Pitch Synchronous Overlap and Add (PSOLA). Finally, the modulated waveforms are concatenated to synthesize the singing voice. This method can produce good quality synthetic voice without accurate speech model and has low computation load. But the disadvantage is the need to build a large corpus spends a lot of time and the voice quality is degraded by high modulation.

The latter method is based on source-filter theory which divides the voice into two components: excitation signal and vocal tract response. The advantage of this method is that it can change the voice characteristic by tuning the parameters [12] and do not need to create a large corpus.

The quality of concatenative synthesis is better than parametric synthesis in the past, but in recent years with advances in digital signal processing, there are some good parametric synthesis methods have been proposed [13].

The Speech Transformation and Represen- tation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) algorithm, proposed by Kawahara et al. [14], is one of the parametric synthesis methods. This algorithm is a high-quality analysis and synthesis method, which uses pitch adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region to remove signal periodicity. Due to the purpose of high quality and naturalness, so we chose

STRAIGHT to implement Taiwanese TTSI synthesis system.

3. SYSTEM STRUCTURE

First collected music and voice files are pre-processed and save as .mid and .wav files to the database. The MIDI files are analyzed through the first-order Markov chain and then establish the probability transition matrix. In the meanwhile, the wav files are analyzed by STRAIGHT and the parameters are saved in syllable parameter database. Then, the probability transition matrix produces similar styles of music depending on the initial tone. The corresponding lyrics serve as information of modulation. Finally, the singing voice is produced. The system flowchart is shown in Figure 1.

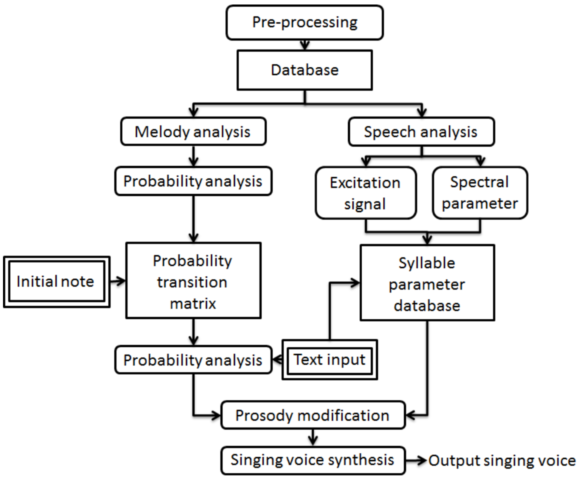


Figure 1. System flowchart

3.1 Pre-processing

We collect ten popular Taiwanese songs and all of songs are pre-processed. The pre-processing procedures for music are list as follows:

1. Collect ten songs and the time signature are 4/4.
2. Remove accompaniment part.
3. Turn the scores into C major and set the tempo to 120.
4. Output the MIDI files.

Next, the 509 syllable units are also pre-processed. The procedures are normalization of the volume and remove all the silence part in the wav files.

3.2 Melody and speech analysis

Ten MIDI files are analyzed by first-order Markov chain and the results are used to construct the probability transition matrix. The melody analysis procedures are list as follows:

1. Using Matlab MIDI toolbox to analyze the MIDI files and extract the parameters. Figure 2 shows an example score and the corresponding result are list in Table 1.

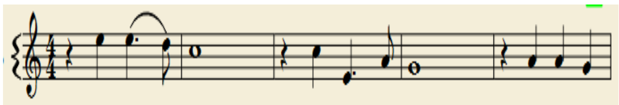


Figure 2. An example score

Note number	Pitch	Volume	Note on	Note off	Duration
1	76	127	0	0.5	0.5
2	76	127	0.5	1.25	0.75
3	74	127	1.25	1.5	0.25
4	72	127	1.5	3.5	2
5	72	127	4	4.5	0.5
6	64	127	4.5	5.25	0.75
7	69	127	5.25	5.5	0.25
8	67	127	5.5	7.5	2
9	69	127	8	8.5	0.5
10	69	127	8.5	9	0.5
11	67	127	9	9.5	0.5

Table 1. MIDI parameter matrix.

2. Find the rest in the score through the note on and note off time and set pitch and duration of the rest. The result shows in Table 2.

Pitch	duration
76	0.5
76	0.75
74	0.25
72	2
0	**0.5**
72	0.5
64	0.75
69	0.25
67	2
0	**0.5**
69	0.5
69	0.5
67	0.5

Table 2. Adding the information of rest to MIDI parameter matrix.

3. Calculate the type of pitch and duration and take each combination as a state in Markov chain. Figure 3 shows an example in this procedure.

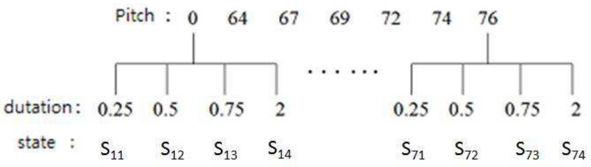

Pitch : 0 64 67 69 72 74 76
dutation: 0.25 0.5 0.75 2
. . . . . .
0.25 0.5 0.75 2
state : S_{11} S_{12} S_{13} S_{14}
S_{71} S_{72} S_{73} S_{74}

Figure 3. An example in procedure 3.

4. Estimate the transition probability and build the probability transition matrix. The result shows in Table3.

To From	S_{12}	S_{23}	S_{32}	S_{34}	S_{41}	S_{42}	S_{52}	S_{61}	S_{72}	S_{73}
S_{12}	0	0	0	0	0	0.5	0.5	0	0	0
S_{23}	0	0	0	0	1	0	0	0	0	0
S_{32}	0	0	0	0	0	0	0	0	1	0
S_{34}	1	0	0	0	0	0	0	0	0	0
S_{41}	0	0	0	1	0	0	0	0	0	0
S_{42}	0	0	0.5	0	0	0.5	0	0	0	0
S_{52}	0.5	0.5	0	0	0	0	0	0	0	0
S_{61}	0	0	0	0	0	0	1	0	0	0
S_{72}	0	0	0	0	0	0	0	0	0	1

S_{73}	0	0	0	0	0	0	0	1	0	0

Table 3. Probability transition matrix.

Finally, 509 syllable units are analyzed by STRAIGHT and the excitation signal and spectral parameter are stored in the syllable parameter database.

3.3 Prosody modification

There are difference in prosody between singing and speaking, for example, the composition ratio of consonants and vowels and the contour of pitch curve. Hence, the following modulation methods are introduced including pitch and duration modulation.

3.3.1 Syllable duration stretching.

Consonant and vowel ratio modulation are defined as follows:

1. Compute the ratio between the note duration and the syllable duration. The equation is defined in (3).

$$\text{Scale} = \text{duration (MIDI)/duration (syllable)} \qquad (3)$$

2. If Scale larger than 1, the duration of vowel is stretched. The equation is defined in (4).

$$\begin{aligned}&\text{duration(syllable)}\\&=\text{duration(consonant)} + \text{Scale} * \text{duration(vowel)}\end{aligned} \qquad (4)$$

3. If Scale less than 1, the duration of syllable is shorten. The equation is defined in (5).

$$\text{duration (syllable)} = \text{duration (syllable)} * \text{Scale} \qquad (5)$$

The result shows in Figure 3. The duration of syllable /tshue/ is 0.8 sec. The upper plot shows the syllable is compressed to 0.5 sec and the bottom plot shows the syllable is stretched to 2 sec.

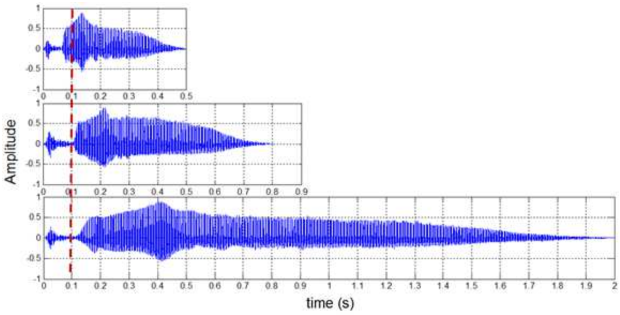


Figure 3. The waveform after duration modulation

3.3.2 Vibrato and Fine fluctuation

Vibrato is a musical effect consisting of a regular, pulsating change of pitch. From the observation of pith contour, the vibrato effect is like a sine wave. The frequency and amplitude of the sine wave are about 5~8 Hz and 3% of fundamental frequency respectively [11]. The vibrato is defined in (6).

$$\Delta F = \frac{3 * F_0}{100} * \sin(2\pi f t) \qquad (6)$$

Fine fluctuation means that the pitch contour has little disturbance when singer maintains a constant pitch [15].The equation is defined in (7),

$$\Delta F = \frac{F_0}{100}(\sin(12.7\pi t) + \sin(7.1\pi t) + \sin(4.7\pi t))/3 \quad (7)$$

where the F_0 in (6) and (7) is fundamental frequency.

3.3.3 Runs and Riffs

Runs and Riffs means that the pitch contour of two or more notes are interpreted by one syllable. The polynomial curve fitting technique is used to estimate the Runs and Riffs effect. The equation is defined in (8),

$$y = \sum_{n=0}^{i} a_n x^n \qquad (8)$$

where n is order and a_n are coefficients. We set n is equal to 21. Figure 4 shows the pitch of Runs and Riffs.

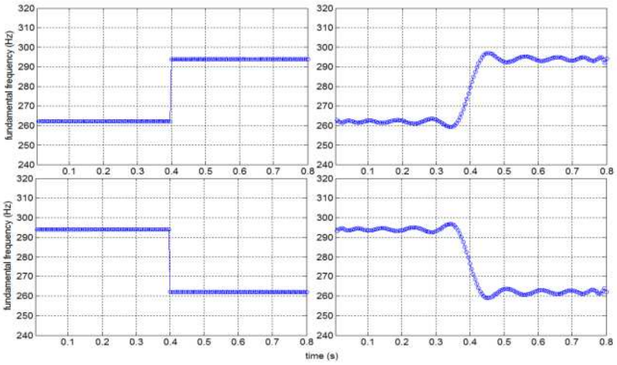


Figure 4. Pitch contour of Runs and Riffs

3.4 Singing voice synthesis

Finally, every modulated syllable units are concatenated into singing voice and saved as a wave file. Figure 4 shows the synthesized singing voice and its pitch contour.

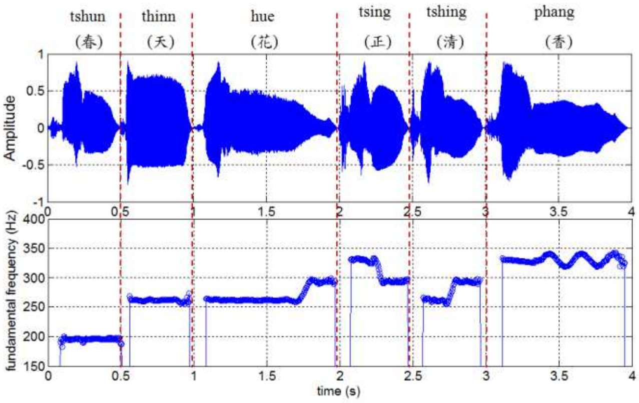
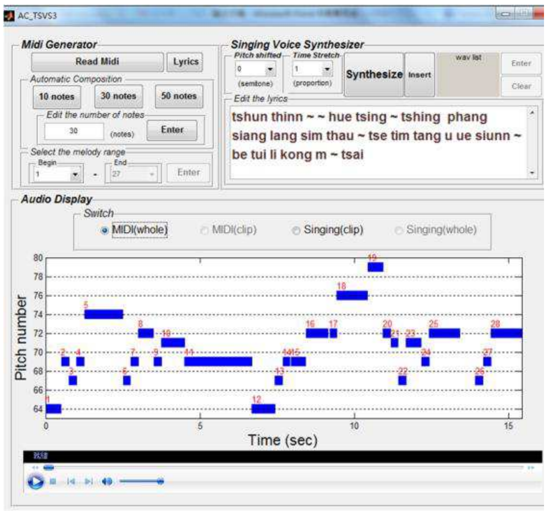


Figure 5. Synthesized singing voice and its pitch contour.

4. RESULT

4.1 Picture based user interface

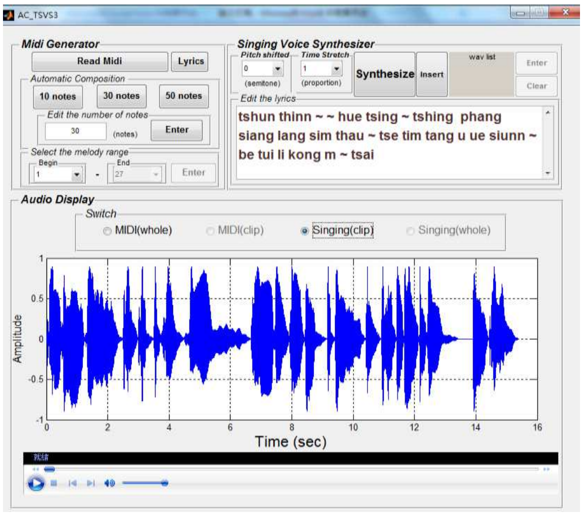
We use an engineering program, Matlab, to show the construction system of combined steps as a structure. Simultaneously we designed a picture based interface to let this approach to be easier to use with the excuse of being user friendly and to produce rhythm and typed lyrics to construct singing voices. As shown in Figure 6, divided by three sections: Midi Generator, Singing Voice Synthesizer, Audio Display. Firstly, Midi Generator can be as the input MIDI files, or to have notes typed in for MIDI files generation.



Pitch shifted: 0 (semitone)
Time Stretch: 1 (proportion)
Edit the number of notes: 30 (notes) Enter
Select the melody range: Begin 1 - End 27 Enter
Edit the lyrics:
tshun thinn ~ ~ hue tsing ~ tshing phang
siang lang sim thau ~ tse tim tang u ue siunn ~
be tui li kong m ~ tsai
Switch: ● MIDI(whole) ○ MIDI(clip) ○ Singing(clip) ○ Singing(whole)

Figure 6. Picture based user interface (1).

This is 30 auto-generated musical notes, which includes 2 rests. To proceed on, we have Audio Display, which will act accordingly to serial number displayed musical note information and transport MIDI file to the media player. The tones of MIDI are piano based. Lastly, input lyrics and Runs and Riffs mark to Singing Voices Synthesizer to synthesize the singing voice. As shown in Figure 7, we use the same steps to combine waveform with wave files to transport to Audio Display.



Pitch shifted: 0 (semitone)
Time Stretch: 1 (proportion)
Edit the number of notes: 30 (notes) Enter
Select the melody range: Begin 1 - End 27 Enter
Edit the lyrics:
tshun thinn ~ ~ hue tsing ~ tshing phang
siang lang sim thau ~ tse tim tang u ue siunn ~
be tui li kong m ~ tsai
Switch: ○ MIDI(whole) ○ MIDI(clip) ● Singing(clip) ○ Singing(whole)

Figure 7. Picture based user interface (2).

4.2 Listening test

The main purpose of survey listening evaluation is to evaluate this system of Taiwanese combined melodically sound quality. There are 3 combined songs, which are respectively quite familiar to the ear popular Taiwanese folk songs: <Romance in Seasons (四季紅)>, <Spring Breeze (望春風)>, <Wife (家後)>, 12 subjects who were tested understood Taiwanese, during the process of listening they will assign to mark the Runs and Riffs lyrics. Then use their own personal sense to make judgments to the combined songs. Points Likert scale given are extensively used as Likert 5-point scale (5 = Very Natural, 4 = Natural, 3 = Normal, 2 = Un-natural, 1 = Very Un-Natural). The results are list in the Table 4.

songs	Naturalness of Runs and Riffs	Naturalness of vibrato	Clearness of pronunciation	Fluency of singing	Acceptance of singing
Romance in Seasons (四季紅)	2.91	2.58	3.08	3.16	2.75
Spring Breeze (望春風)	2.66	2.41	2.91	3.08	2.5
Wife (家後)	2.33	2.16	2.58	2.25	2.08
Average Score	2.63	2.38	2.86	2.83	2.44

Table 4. Results of listening test.

5. CONCLUSIONS

In this study, we use first order Markov chain algorithm to establish Taiwanese popular songs automatic composition system and STRAIGHT algorithm to establish Taiwanese TTSI synthesis system. And we achieve the integration of the two systems for validation. According to this integrated system interface, the user can simply produce music with Taiwanese popular ballad style, and to synthesize singing with any lyrics, so people, who cannot read music, also can compose music.

Acknowledgments

The authors would like to appreciate the support from National Science Council projects of Taiwan: NSC 101-2410-H-155-033-MY3 and NSC 101-2627-E-155-001-MY3.

6. REFERENCES

[1] A. A. SCHOENBERG, Style and Idea: selected writings of Arnold Schoenberg: University of California Pr, 1975.

[2] L. A. H. a. L. M. Isaacson, Experimental Music; Composition with an Electronic Computer by Lejaren A. Hiller, Jr. and Leonard M. Isaacson: New York. McGraw-Hill, 1959.

[3] I. Xenakis, "Free Stochastic Music from the Computer. Programme of Stochastic music in Fortran," Gravesaner Blätter, vol. 26, pp. 54-92, 1965.

[4] H. Kenmochi and H. Ohshita, "VOCALOID-commercial singing synthesizer based on sample concatenation," in INTERSPEECH, 2007, pp. 4009-4010.

[5] C.-Y. Lin, T.-Y. Lin, and J.-S. R. Jang, "A corpus-based singing voice synthesis system for Mandarin Chinese," in Proceedings of the 13th annual ACM international conference on Multimedia, 2005, pp. 359-362.

[6] H.-Y. Gu and Z.-F. Lin, "Mandarin singing voice synthesis using ANN vibrato parameter models," in Machine Learning and Cybernetics, 2008 International Conference on, 2008, pp. 3288-3293.

[7] Sen-Fu Liang, " An F0 Control Model for Synthesis of Taiwanese Children"s Songs, Master’s Thesis, Department of Computer and Communication Engineering National Kaohsiung First University of Science and Technology, 2010.

[8] Yu-Jhe Li, " Using Speech Scoring for the Validation of Taiwanese Speech Corpus, Master’s Thesis, Department of Computer Science, National Tsing Hua University,, 2013.

[9] Chiu,Shih-Chuan, "Computer Music Composition by Discovered Music Patterns," Master’s Thesis, Department of Computer Science, National Chengchi University, 2005.

[10] W. Schulze and B. van der Merwe, "Music generation with Markov models," Multimedia, IEEE, vol. 18, pp. 78-85, 2011.

[11] Shih-Han Chan, "A Singing Voice Synthesis System Based On Pitch Curve Modulation," Master’s Thesis, Department of Computer Science, National Tsing Hua University, 2006.

[12] Yun-Ting Tsai, "HMM-base speech synthesis for Mandarin Chinese," Master’s Thesis, Department of Electrical Engineering, National Tsing Hua University, 2009.

[13] J. P. Cabral, "HMM-based speech synthesis using an acoustic glottal source model," 2011.

[14] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, 1997, pp. 1303-1306.

[15] M. Macon, L. Jensen-Link, E. B. George, J. Oliverio, and M. Clements, "Concatenation-based MIDI-to-singing voice synthesis," in Audio Engineering Society Convention 103, 1997.