Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 63-80

HMM-based Mandarin Singing Voice Synthesis Using Tailored Synthesis Units and Question Sets

Ju-Yun Cheng*, Yi-Chin Huang*, and Chung-Hsien Wu*

Abstract

Fluency and continuity properties are essential in synthesizing a high quality singing voice. In order to synthesize a smooth and continuous singing voice, the Hidden Markov Model-based synthesis approach is employed in this study to construct a Mandarin singing voice synthesis system. The system is designed to generate Mandarin songs with arbitrary lyrics and melody in a certain pitch range. In this study, a singing voice database is designed and collected, considering the phonetic converge of Mandarin singing voices. Synthesis units and a question set are defined carefully and tailored the meet the minimum requirement for Mandarin singing voice synthesis. In addition, pitch-shift pseudo data extension and vibrato creation are applied to obtain more natural synthesized singing voices.

The evaluation results show that the system, based on tailored synthesis units and the question set, can improve the quality and intelligibility of the synthesized singing voice. Using pitch-shift pseudo data and vibrato creation can further improve the quality and naturalness of the synthesized singing voices.

Keywords: Mandarin Singing Voice Synthesis, Hidden Markov Models, Vibrato

1. Introduction

In recent years, Mandarin text-to-speech synthesis systems have been proposed and have achieved satisfactory performance (Ling, 2012; Wu, 2007). These systems are able to synthesize fluent and natural speech, even with personal characteristics (Huang, 2013). Recently, singing voice synthesis has been one of the emerging and popular research topics. Such systems enable computers to sing any song.

There are two main methods in the research on corpus-based singing voice synthesis. The first one is the sample-based approach. The principle of this method is to use a large database

* Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan
E-mail: { carrie771221; ychin.huang; chunghsienwu}@gmail.com

of recordings of singing voices that are further segmented into units. In the synthesis phase, based on a given score with the lyrics, the system then searches and selects appropriate sub-word units for concatenation. VOCALOID (Kenmochi, 2007) is such a singing voice synthesizer that enables the user to input lyrics and the corresponding melody. Given the score information, the system selects the necessary samples from the Singer Library and concatenates them to produce the synthesized singing voice. Finally, the system performs pitch conversion and timbre manipulation to generate smoothed concatenated samples. The software was originally only available in English and Japanese, but VOCALOID 3 has added support for Spanish, Chinese, and Korean. A Mandarin singing voice system using a unit selection method was proposed in (Zhou, 2008). Singing units in this method are chosen from a singing voice corpus with the lyrics of the song and the musical score information embedded in a MIDI file. To improve the synthesis quality, synthesis unit selection and the prosody and amplitude modification are applied. This system uses a Hanning window to smooth instances where speech segments were concatenated. Although the unit selection method is able to synthesize high quality speech at the waveform level, the concatenation-based methods suffer from the discontinuity problem at the boundaries between concatenated units. As different samples that make up the singing voice are recorded in different pitches and phonemes, discontinuity might exist in the resulting singing voice.

The other method is the statistical approaches, where hidden Markov models (HMMs) (Oura, 2010; Saino, 2006) are the most widely used. Acoustic parameters are extracted from a singing voice database and modeled by the context-dependent HMMs. The acoustic parameters are generated by the concatenated HMM sequence. Finally, vocoded waveforms of the singing voice are generated from the inverse filter of the acoustic parameters. Sinsy (Oura, 2010) is a free-online HMM-based singing voice synthesis system that provides Japanese and English singing voices. Users can obtain synthesized singing voices by uploading musical scores. Synthesizing singing voices based on HMMs sound blurred due to the limitation of the current vocoding technique. Nevertheless, it can generate a smooth and stable singing voice, and its voice characteristics can be modified easily by transforming the parameters appropriately.

In addition to the concatenation-based method and statistical method, there are also some other methods proposed to generate a Mandarin singing voice, *e.g.*, Harmonic plus Noise Model (HNM) (Gu, 2008), which adopted HNM parameters of a source syllable to synthesize singing syllables of diverse pitches and durations. This method can generate singing voices with good quality. Nevertheless, the discontinuity problem occurring at the concatenation points is still a major problem. Speech-to-singing method (Saitou, 2007) is another approach. Instead of synthesizing from a singing database, the speech-to-singing method converts speech into a singing voice by a parameter control model. Similarly, text-to-singing (lyrics-to-singing)

synthesis (Li, 2011) is used to generate synthesized speech of input lyrics by a TTS system followed by a melody control model that converts speech signals into singing voices by modifying the acoustic parameters. These two methods are based mainly on conversion rules that could be patchy.

Research on speech and singing synthesis has been closely linked, but there are important differences between the two methods with respect to the generated voices. The major parts of singing voices are voiced segments, whereas speech consists of a relatively large percentage of unvoiced sounds (Kim, 2003). Besides, fluency and continuity in singing voices are very important properties. In order to synthesize a smooth and continuous singing voice, an HMM-based synthesis approach is adopted in this study to build our singing voice synthesis system. To the best of our knowledge, the currently available HMM-based singing voice synthesis systems have not been applied to the Mandarin singing voice. By carefully defining and tailoring the synthesis units and the question set, a Mandarin singing voice synthesis system based on HMM-based framework has been constructed successfully in this study.

The rest of the paper is organized as follows. The proposed HMM-based singing voice synthesis system is introduced in Section 2. Section 3 consists of subjective and objective evaluations of the proposed system, compared to the original HMM-based singing voice synthesis system. Concluding remarks and future work are given in Section 4.

2. Proposed Mandarin Singing Voice Synthesis System

In recent years, the number of studies on HMM-based speech synthesis has grown. Some research has made progress on prosody improvement (Hsia, 2010; Huang, 2012) to obtain more natural speech. Recently, an HMM-based method has been applied to singing voice synthesis (Saino, 2006). There are more combinations of contextual factors in singing voice synthesis than that in speech synthesis. Applying a unit selection method to singing voice synthesis is quite difficult, because it needs a huge number of singing voices. On the contrary, an HMM-based system can be constructed using a relatively small amount of training data. As a result, the HMM-based approach is easier for constructing a singing voice synthesizer.

The system proposed in this study is based on the HMM-based approach that was developed by the HTS working group (Zen, 2007). The proposed structure of the singing synthesis system based on HMM is shown in Figure 1.

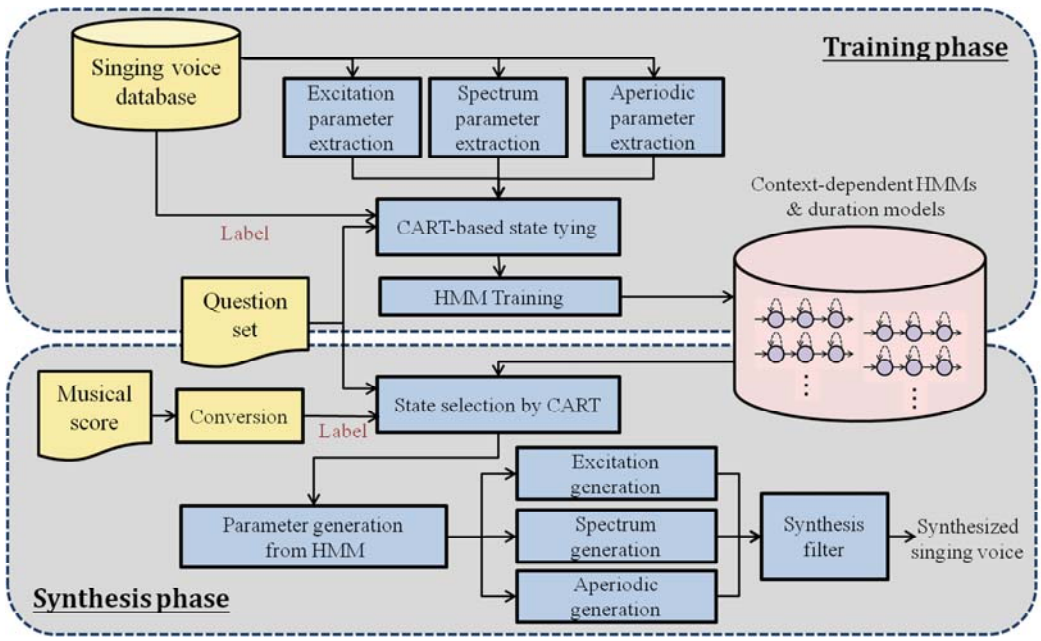

Figure 1. Structure of the HMM-based Singing Voice Synthesis System

In the training phase of the proposed system, excitation, spectral, and aperiodic parameters are extracted from a singing voice database. Lyrics and notes of the songs in the singing corpus are considered as contextual information for generating context-dependent label sequences. Then, the sequences are split and clustered with context-dependent question sets and the context-dependent HMM models are trained based on the clustered phone segments. In the synthesis phase, a musical score and the lyrics to be synthesized also are converted into a context-dependent label sequence. Based on the label sequence, a sequence of parameters, consisting of excitation, spectral, and aperiodic parameters, corresponding to the given song is obtained from the concatenated context-dependent HMMs. Finally, the obtained parameter sequences are synthesized to generate the singing voice.

2.1 Model Definition

Singing is the act of producing musical sounds with one's voice, and one main difference between a singing voice and speech is the use of the tonality and rhythm of a song. Therefore, the contextual factors should consist of not only linguistic information but also note information. In addition, the cue information obtains the actual timing of each phone in the singing data. The details of the model definition are described in the following section.

2.1.1 Linguistic Information

In the HMM-based Mandarin speech synthesis, "segmental Tonal Phone Model, STPM" (Huang, 2004) is often adopted to define the HMM-based phone models. Only a relatively small number of phone models are defined to characterize all Mandarin tonal syllables. Furthermore, in order to represent the five lexical tones for Mandarin syllables, each Mandarin syllable is defined to consist of three parts, based on phonology (Lin, 1992), as: **C+V1+V2**. In the phonological structure, **C** denotes the first extended initial phone and the following units (**V1** and **V2**) are tonal final phones. Tonal final phone conveys tonal information using the extended tone notations, such as H (high), M (middle), and L (low), *i.e.*, Tone 1: H+H, Tone 2: L+H, Tone 3: L+L, Tone 4: H+L, and Tone 5: M+M).

Although STPM can describe all pitch patterns in Mandarin speech, pitch patterns in singing voices are quite different from read speech. Figure 2 shows the pitch contours (blue lines) of the read speech and singing voice of the same sentence produced by the same person. As the figure shows, the pitch contour of the read sentence is controlled by the tone of each syllable. In contrast, the pitch contour of a singing sentence is relatively flat and corresponds to the musical notes of the corresponding syllables. The musical note is more of a requirement than the tones of the syllables for the pitch contour in singing voice. Therefore, the definition of each syllable for a singing voice is redefined as **C+V**, where **C** is still the extended initial sub-syllable and **V** is the final sub-syllable without tonal information.

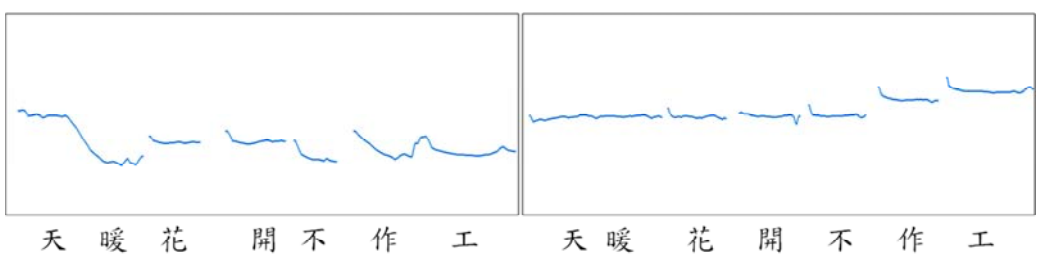

Figure 2. An example of the read speech and singing voice for the sentence "天暖花開不作工," which is uttered and sung by the same person.

Rhythm is one major difference between read speech and a singing voice. Vowels usually convey the rhythm of a singing voice since the vocal tract remains open while uttering a vowel, allowing the resonance frequencies of the vocal tract to remain stable. Because of these characteristics, vowels are probably one of the most important factors to represent a good singing voice. A Mandarin syllable consists of two parts: *initial* and *final*. The *initial* part is optional and is composed of consonants. The *final* part, namely vowels, includes *medial* and *rime*. The *medial* is located between the *initial* and the *rime*. The *medial* phonologically is connected with the *rime* rather than the *initial*. So, in the definition of singing sub-syllable

models, each *medial* is combined with a *final*. Furthermore, the combination of *medial* and *rime* is collectively known as a *final*, and some examples are listed in Table 1. For the singing model definition, the phonetic annotation is based on the Hanyu Pinyin. Note that the tone information (Arabic numbers) of the original tonal syllable is ignored for the *initial* or *final* models in the singing sub-syllable definition. Besides, we define the *final* models with *medial* as separate models to ensure that each vowel can have a specific model representing this property.

Table 1. Examples of finals with medial

	Tonal Syllable	C	V
ㄉㄧㄠˋ	diau4	d	iau
ㄌㄨㄛˇ	luo3	l	uo
ㄒㄩㄝˊ	shiueh2	sh	iueh

The syllable with only an *initial* is generally followed by an empty *rime* “ㄭ” . The empty *rime* does not have word phonetic annotation. In order to represent this property, we define a phoneme “zr” as the empty *rime* of the retroflex, which is connected only to the retroflex class of *initial* phonemes. Correspondingly, the phoneme “sr” is the empty *rime* of the alveolar, which is connected only to the alveolar class of *initial* phonemes.

In general, a long duration note is sung differently from short duration note. For shorter notes, temporal variation is relatively small and stable. Nevertheless, temporal variation of a longer note is much larger and unstable. Lengthening a syllable with a short duration note cannot precisely represent the expression of syllable with long duration note. So, when the word corresponds to a half note or above, the *finals* followed by an “L” are defined to denote the long duration model.

According to the above rules, 95 Mandarin signing sub-syllables are obtained according to the definition for singing voice. There are 21 *initial* sub-syllables, 18 *final* sub-syllables (2 of *finals* are empty *final* phonemes), 20 *medials* combined with *final* sub-syllables, and 36 long duration sub-syllables. In addition to the 95 signing sub-syllables, the silence and pause models are further included. Silence is an unvoiced segment in the beginning and the end of a song. Pause is an unvoiced segment in the middle of a song.

2.1.2 Note Information

In addition to lyrical information, note information is one of the vital factors for singing voice synthesis. Contextual factors of note information consist of three categories to fully describe singing characteristics, including pitch and duration of the note and the song structure. Note pitch refers to the melody of a song and determines if the song sounds great or not. In this

category, absolute pitch, relative pitch, pitch difference between previous and current notes, and pitch difference between current and next notes are included. Duration is the length of a note and is one of the bases of rhythm. In this category, the length of the note can be expressed by three kinds of standards. Song structure means which overall musical form or structure the song adopted and the order of the musical score. Different note positions in the measure or phrase may have different expressions due to breathing. In this category, beat, tempo, key of the song, and position of each note are included.

2.1.3 Cue Information

Cue information considered in the contextual factors consists of the timing and the length of a sub-syllable. We manually segment all of the songs at the sub-syllable level. The timing information of a sub-syllable measured based on a time interval of 0.1 seconds will be converted into the absolute length of the note. The position of note identity in the measure or phrase is also converted according to the cue information.

2.2 Question Set for Decision Trees

Based on unit definition and contextual factors, we define five categories for the questions in the question set. The five categories of the question set are sub-syllable, syllable, phrase, song, and note. The details of the question set are described as follows.

(1) Sub-syllable: (current sub-syllable, preceding one and two sub-syllables, and succeeding one and two sub-syllables) Initial/final, final with medial, long model, articulation category of the initial, and pronunciation category of the final

(2) Syllable: The number of sub-syllables in a syllable and the position of the syllable in the note

(3) Phrase: The number of sub-syllables/syllables in a phrase

(4) Song: Average number of sub-syllables/syllables in each measure of the song and the number of phrases in this song

(5) Note: The absolute/relative pitch of the note; the key, beat, and tempo of the note; the length of the note by syllable/0.1 second/thirty-second note; the position of the current note in the current measure by syllable/0.1 second/ thirty-second note; and the position of the current note in the current phrase syllable/0.1 second/thirty-second note

2.3 Baseline Model

There are 5364 different questions defined in the question set. The HMMs for the baseline Mandarin singing voice synthesis system were trained based on the entire question set, and the resulting clustered HMMs are shown in Table 2 and Table 3. As shown in these tables, the

number of leaf nodes in the tree clustered using fundamental frequency (F0) is 3951. The number of each state for the clustered F0 models is shown in Table 2. The most frequently used questions for every clustered tree of each state were sub-syllable types, position of note in measure or phrase, and phrase level.

The number of leaf nodes in the trees for mel-cepstral coefficients (mgc) is 2844. The number of the leaf nodes in each state is shown in Table 3. The most frequently used questions are the same as the results for F0.

Table 2. Number of leaf nodes in each state in F0 tree

State	State 2	State 3	State 4	State 5	State 6
Number of nodes	1146	509	366	626	1304

Table 3. Number of leaf nodes in each state in mel-cepstral coefficient tree

State	State 2	State 3	State 4	State 5	State 6
Number of nodes	244	849	938	604	209

2.4 System Refinement

The baseline of singing voice system can synthesize arbitrary songs, but it still has a lot of room to improve. The approaches we implemented to refine our system include question set modification, singing voice database extension using pitch-shift pseudo data, and vibrato creation.

2.4.1 Pitch-Shift Pseudo Data

Pitch is highly related to the notes and the sounds we hear when someone is singing. The quality of the song strongly depends on the accurate pitch of all notes produced by the singer. The quality of the HMM-based synthesized singing voices depends strongly on the training data, owing to its statistical nature. Therefore, the singing database should cover the pitch range of the notes in the song. Using the pitch-shift pseudo data, it is helpful to cover the missing pitch of sub-syllables and increase the size of the training data. We examine whether all Mandarin sub-syllables we defined cover the whole pitch range (C4~B4) or not. Since shifting too much frequency of a note will change the timbre, the missing pitches of sub-syllables could be obtained using the nearby notes from other songs.

2.4.2 Question Set Modification

The parameters generated from the clustered HMMs are highly correlated to the speech quality of the synthesized singing voice. A large number of contextual factors are not suitable when the size of the training data is not large enough to be clustered by various contextual

factors, and this may cause a data sparseness problem. The selection of the question set is crucial for generating proper models. In the baseline system, the most frequently used questions in the trees for F0 and mel-cepstral coefficients are sub-syllables types, position of note, and phrase level. Nevertheless, our singing database is not large enough to obtain every contextual factor. Thus, the question set should be tailored to remove some unsuitable questions. The removed questions consist of three types, including duplicate questions, indirect questions, and relative questions.

Duplicate questions refer to times when the note length can be represented by two types of units, 0.1 second and thirty-second note. Although 0.1 second is an absolute length and thirty-second note is the relative pitch of the recorded waveform, both units describe the same information. So, we delete the note length question with 0.1 second. Indirect question means the questions at the level of phrase and song, which are called paralinguistic information. These questions do not directly represent the information of one note, because they are mainly about how many sub-syllables and syllables there are in phrases and the average numbers of sub-syllables and syllables in each measure of the songs. The essential information of a note is its pitch and length, so the questions about position of note are also indirect questions. The paralinguistic information, however, could be useful when the size of corpus is large. Every song has different keys, so the standard of the relative pitch also is different. Two notes with the same relative pitch may have different absolute pitch values. Therefore, we delete the question sets related to relative pitch.

Furthermore, we modify the absolute pitch questions by keeping the questions with absolute answers and remove the questions with comparative answers. Thus, we can ensure that the leaf node that is divided by the absolute pitch questions can be clustered with the same absolute pitch.

2.4.3 Vibrato Creation

Vocal vibrato is a natural oscillation of musical pitch, and singers employ vibrato as an expressive and musically useful aspect of the performance. Adding vibrato can make the synthesized singing voice more natural and expressive. The frequency and the amplitude can be considered since they are the two fundamental parameters affecting the characteristic sound of a vibrato effect. The method to create vibrato is to vary the time delay periodically (Zölzer, 2002), and it uses the principle of Doppler Effect. Our system implemented the vibrato effect by a delay line and a low frequency oscillator (LFO) to vary the delay.

3. Evaluations

3.1 Singing Voice Database

For the construction of the signing voice database, the musical scores from nursery rhymes and children's songs are considered as the candidates. The major selection criterion for choosing the songs is the phonetic coverage for synthesizing universal Mandarin singing voices. The lyrics of the selected songs should cover all of the sub-syllables in Mandarin. A total of 74 songs were selected. Some of the selected songs have two or more versions with the same melody but different lyrics. Considering the variation of pitch and timbre, a female singer who has been participating in a singing contest and is a member of the a cappella team was invited as the signer to provide a stable and natural-sounding signing voice. The singer used the built-in microphone of a MAC notebook for recording. The songs were recorded using Audacity. The environment where the signer recorded was quiet. Noises, including the metronome, were not allowed. Besides, each song has two versions in order to increase the quantity of the database. The singing data with low signal-to-noise ratio or energy exceeding a limit were not included. The amplitude of all singing data was normalized. The overview of this database is summarized in Table 4. To improve the quality of the database, sub-syllable boundaries and musical scores were manually corrected.

Table 4. Details of NCKU singing voice database

Songs	Nursery rhymes (children's songs) Total 148 songs
Singer	One female
Pitch range	C4~B4
Version	2
Total time	About 102 minutes
Sample rate	48 kHz
Resolution	16 bits
Channels	Mono

3.2 Experimental Conditions

Singing voice signals were sampled at a rate of 48 kHz and windowed by a 25ms Blackman window with a 5ms shift. Then, mel-cepstral coefficients were obtained from the STRAIGHT algorithm (Kawahara, 2006). The feature vectors consisted of spectrum, excitation, and aperiodic factors. The spectrum parameter vectors consisted of 49-order STRAIGHT

mel-cepstral coefficients, including the zero-th coefficient, their delta, and delta-delta coefficients. The excitation parameter vectors consisted of log F0, its delta, and delta-delta.

A seven-state (including the beginning and ending null states), left-to-right Hidden Semi-Markov Models (HSMM) (Zen, 2007) was employed, in which the spectral part of the state was modeled by a single diagonal Gaussian output distribution. The excitation stream was modeled with multi-space probability distributions HSMM (MSD-HSMM), each of which consisted of a Gaussian distribution for "voiced" frames and a discrete distribution for "unvoiced" frames.

The term *Riffs and runs* implies a syllable with multiple notes. In other words, it is a quick articulation of a series of pitches sustained on a single vowel sound. In the proposed method, the generation of riffs and runs repeats the last final in previous words to mimic the singing skill.

Furthermore, in the middle of a song, vibrato is combined with the amplitude in 4E-4 millisecond, frequency in 6 Hz, and start timing in 25% of the sub-syllable. At the end of a song, vibrato is combined with the amplitude in 8E-4 millisecond, frequency in 5 Hz, and the start timing is at the position of 50% of the sub-syllable.

3.3 Evaluation Results

To evaluate the constructed Mandarin singing voice synthesis system, we conducted a subjective listening test. Ten songs not included in the training data were divided into two parts. Therefore, we obtained 20 parts for testing. The testing waveforms generated by different systems were presented to the subjects in a random order. 12 native Mandarin speaking subjects were asked to participate in the evaluation test. Mean Opinion Score and Preference test were used as evaluation measures for the subjective test.

In order to evaluate the effectiveness of the refinements we proposed, four different settings of the synthesis models were used. These models were evaluated on the effect of the refinements, *i.e*. question set modification and inclusion of pitch-shift pseudo data. The settings and the descriptions are described in Table 5.

Table 5. Four different settings of models and their descriptions

Model	Description
Baseline	All question set
QM	Question set modification
PS	Pitch shift pseudo data
QM+PS	Question set modification and pitch shift pseudo data

3.3.1 Pitch Contour Comparison

Figure 3 shows the Mandarin singing voice synthesis system can generate the F0 patterns similar to the actual F0 patterns of the musical score. Figure 4 shows the Mandarin singing voice synthesis system can generate the pitch contour of the synthesized singing voice with almost the same as the pitch contour of the original singing voice. Nevertheless, some of the singing phenomena, such as overshoot and preparation were smoothed after HMM training.

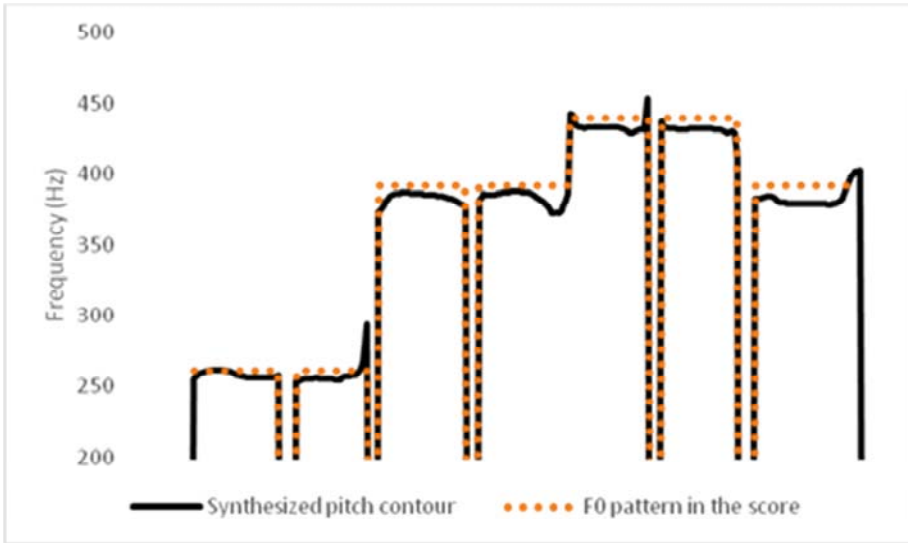

Figure 3. Comparison with generated F0 patterns and F0 patterns in the score

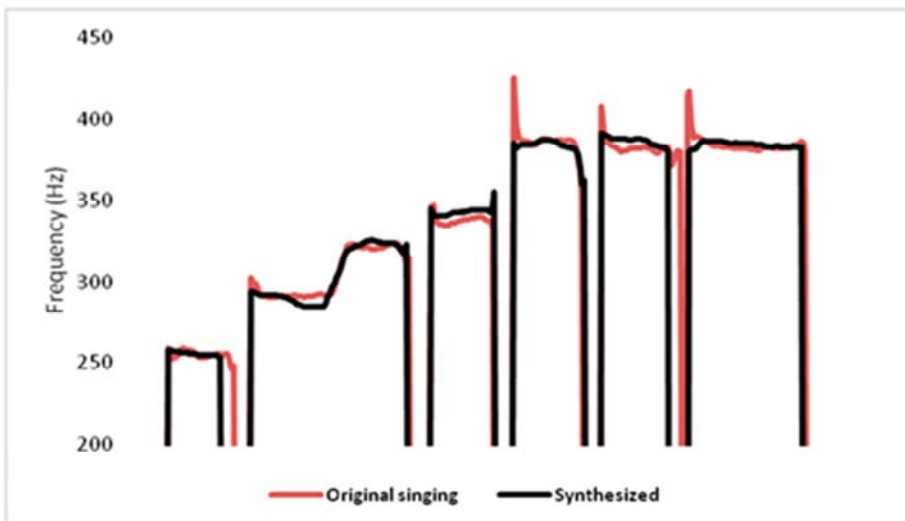

Figure 4. Comparison between the original singing and the synthesized singing pitch contours

3.3.2 Preference Test

We evaluated the nature of the synthesized singing voice with a long duration model. Figure 5 shows the system with the long duration model has 62% preference, which is higher than 38% for the system without the long duration model. This shows that the long duration model can improve the nature of phones with long duration. Therefore, all of the evaluated systems use the long duration model in the following tests.

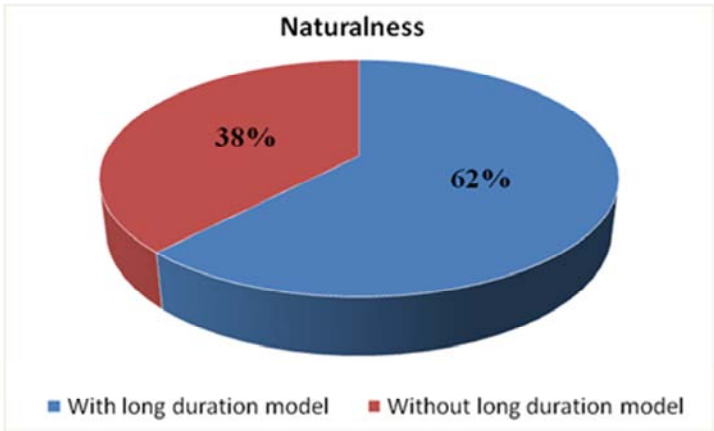

Figure 5. Result of preference test with long duration model

In addition, we evaluated the nature of the synthesized singing voice with vibrato. The preference result is shown in Figure 6. The subjects only slightly preferred the synthesized singing voice with vibrato over that without vibrato. The main reasons are that two combinations of parameter settings are insufficient and that different pitches and situations must correspond to different combinations of vibrato parameters. Moreover, vibrato is not essential in children' songs. Subjects preferred simple over skillful singing styles in these kinds of songs.

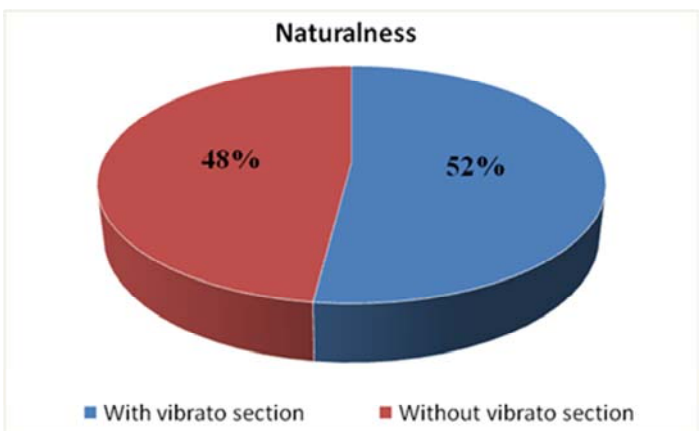

Figure 6. Result of preference test with/without vibrato

3.3.3 Mean Opinion Scores (MOS)

The MOS of quality in four evaluation settings is shown in Figure 7, and the MOS of intelligibility is shown in Figure 8.

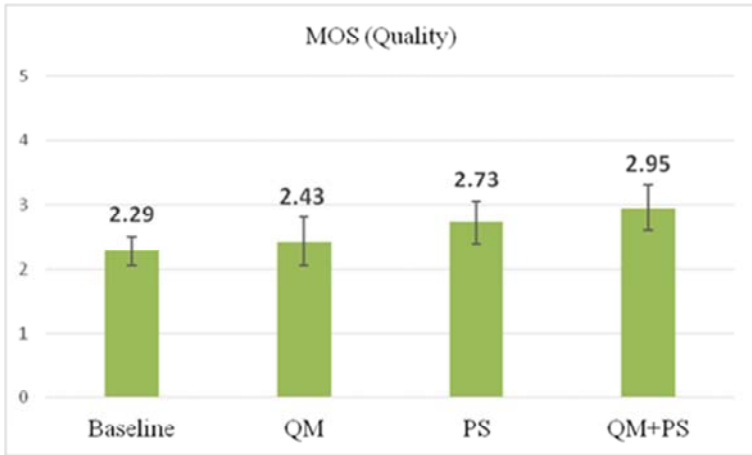

Figure 7. MOS of the synthesized singing voice in quality

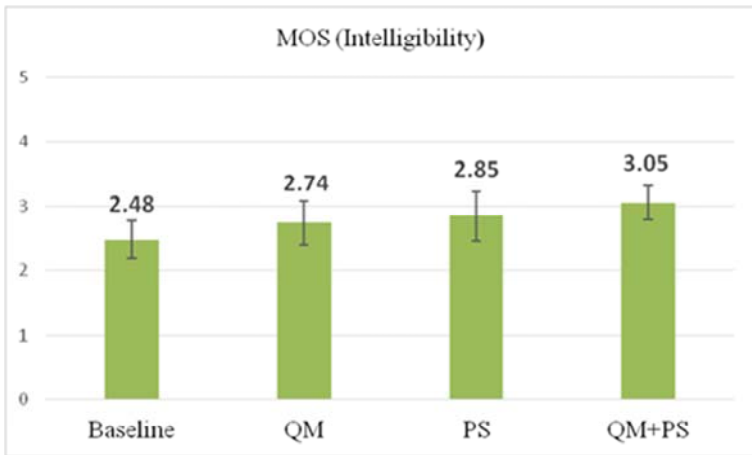

Figure 8. MOS of the synthesized singing voice in intelligibility

The results show that the baseline singing voice system has the lowest MOS, because the training data is insufficient for clustering using a large number of questions and because some sub-syllables are not covered in some of the pitch frequencies. After question modification, MOS is 2.43 in quality and 2.74 in intelligibility, which are higher than those for the baseline system. The PS model has MOS of 2.73 in quality and 2.85 in intelligibility, which are higher than those for the baseline system. This shows that adding pitch-shift pseudo data is one of the useful refinements. Finally, the MOS of QM+PS model is 2.95 in quality and 3.05 in intelligibility. These scores are higher than those for the PS model with the modified question

set and the QM model with pitch-shift pseudo data. According to the results, we can conclude that, although all question sets take all of the contextual factors into account, some contextual information might not be included in the corpus, which may cause bad clustering results. By tailoring the question set appropriately, the system can improve the quality and intelligibility of the synthesized singing voice. In addition, adding pitch-shift pseudo data also can improve the quality of the synthesized singing voice.

4. Conclusions and Future Work

In this paper, a corpus-based Mandarin singing voice synthesis system based on hidden Markov models (HMMs) was implemented. We defined the Mandarin phone models and the question set for model clustering. Linguistic information and musical information both are modeled in the context-dependent HMM. Furthermore, three methods were employed to refine the constructed system, *i.e*. question set modification, pitch-shift pseudo data, and vibrato creation. Experimental results show that the proposed system could synthesize a satisfactory singing voice. The performance of the corpus-based synthesis system is highly dependent on the training corpus, and the quality of the corpus can directly affect the synthesized voice quality. The environment for data recording should be professional and silent, such as in an anechoic chamber or using sound-absorbing equipment. Furthermore, the training corpus should be as large as possible to cover all contextual factors. Although our singing database was designed with high phonetic coverage and enhanced by adding pseudo data for better pitch coverage, there are some factors that were not covered, such as the coverage of duration and higher level information.

Besides, a more accurate model is essential for synthesizing a better singing voice. Model clustering should be categorized and labeled with priority, since some factors are more important than others for singing characteristics. The process of clustering decision trees should be guided based on the priority of clustering questions to obtain a more accurate model.

The singer's timbre and pronunciation are also important factors that affect synthesized singing voice quality. The nasal tone of a singer's voice might cause acoustic information disappearance when uttering syllables with higher pitches. Unclear utterances also cause the synthesized singing voice to become unintelligible. For further improvement, these problems should be carefully considered in order to generate better synthesized singing voices.

References

Gu, H.-Y., & Liau, H.-L. (2008). Mandarin Singing Voice Synthesis Using an HNM Based Scheme. *International Congress on Image and Signal Processing (CISP)*, 347-351.

Hsia, C.-C., Wu, C.-H., & Wu, J.-Y. (2010). Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 1994-2003.

Huang, Y.-C., Wu, C.-H., & Chao, Y.-T. (2013). Personalized Spectral and Prosody Conversion using Frame-Based Codeword Distribution and Adaptive CRF. *IEEE Trans. Audio, Speech, and Language Processing*, *21*(1), 51-62.

Huang, Y.-C., Wu, C.-H., & Weng, S.-T. (2012). Hierarchical prosodic pattern selection based on Fujisaki model for natural mandarin speech synthesis. *2012 8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 79-83.

Huang, C., Shi, Y., Zhou, J., Chu, M., Wang, T., & Chang, E. (2004). Segmental tonal modeling for phone set design in Mandarin LVCSR. *Proceedings of ICASSP 04*, 901-904.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, *27*(6), 349-353.

Kenmochi, H., & Ohshita, H. (2007). VOCALOID-Commercial singing synthesizer based on sample concatenation. *INTERSPEECH 2007*, 4009-4010.

Kim, Y. E. (2003). *Singing Voice Analysis/Synthesis*. Ph.D. dissertation, Massachusetts Institute of Technology.

Li, J., Yang, H., Zhang, W., & Cai, L. (2011). A Lyrics to Singing Voice Synthesis System with Variable Timbre. *Applied Informatics and Communication Communications in Computer and Information Science*, 225, 186-193.

Lin, T., & Wang, L.-J. (1992). *Phonetic Tutorials*. Beijing University Press, 103-121.

Ling, Z.-H., Xia, X.-J., Song, Y., Yang, C.-Y., Chen, L.-H., & Dai, L.-R. (2012). *The USTC System for Blizzard Challenge 2012*. Blizzard Challenge Workshop.

Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y., & Tokuda, K. (2010). Recent Development of the HMM-bases Singing Voice Synthesis System-Sinsy. *The 7th ISCA Speech Synthesis Workshop*, 211-216.

Saino, K., Zen, H., Nankaku, Y., Lee, A., & Tokuda, K. (2006). An HMM-based Singing Voice Synthesis System. *International Conference on Spoken Language Processing (ICSLP)*, 1141-1144.

Saitou, T., Goto, M., Unoki, M., & Akagi, M. (2007). Speech-to-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices. *Applications of Signal Processing to Audio and Acoustics Workshop*, 215-218.

Wu, C.-H., Hsia, C.-C., Chen, J.-F., & Wang, J.-F. (2007). Variable-length unit selection in TTS using structural syntactic cost. *IEEE Trans. Audio, Speech, Lang. Process.*, *15*(4), 1227-1235.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., & Tokuda, K. (2007). The HMM-based Speech Synthesis System (HTS) Version 2.0. *The 6th ISCA Workshop on Speech Synthesis*, 294-299.

Zen, H., Tokuda, K. T., Masuko, T., Kobayasih, T., & Kitamura, T. (2007). A Hidden Semi-Markov Model-Based Speech Synthesis System. *IEICE Trans. Inf. & Sys*., *90*(5), 825-834.

Zhou, S.-S., Chen, Q.-C., Wang, D.-D., & Yang, X.-H. (2008). A Corpus-Based Concatenative Mandarin Singing voice Synthesis System. *2008 International Conference on Machine Learning and Cybernetics*, 2695-2699.

Zölzer, U. (2002). *DAFX- Digital Audio Effects*. John Wiley & Sons, Chapter 3, 68-69.