Journal of Multidisciplinary Engineering Science Studies (JMESS)
ISSN: 2458-925X
Vol. 8 Issue 7, July - 2022

A Singing Text-to-Speech Conversion

Based on UTAU Software

Hung-Che Shen
Dept. of digital multimedia design
I-Shou University
Kaohsiung, Taiwan
shungch@isu.edu.tw

***Abstract*—This paper describes a singing text-to-speech system that can synthesize a talking voice from an input singing voice and the song lyrics. The system controls four acoustic features that determine the difference between speaking and singing voices: the pitch, phoneme duration, tempo and velocity. By changing these features of a singing voice, the system synthesizes a talking voice while retaining the timbre of the singing voice. Originally, the UTAU software was designed for singing voice synthesis. Controlling the musical note's features to some target values with input lyrics, a singing text-to-speech system is derived. The singing voice becomes talking voice by preserving the same timbre. The system finally generates a talking voice that preserves the timbre of the singing voice but has speech-like features. Currently, only Mandarin text-to-speech is implemented. Experimental results show that this singing text-to-speech system can convert singing voices into speech voices whose timbre is almost the same as the original singing voices and quality is nature.**

Keywords—singing text-to-speech; UTAU; Mandarin;

I. INTRODUCTION

This paper attempts to show that a singing voice synthesis by UTAU software [1] can also be used as a text-to-speech system. The singing text-to-speech is achieved by controlling the acoustic features unique to speech. On the basis of signal generation, singing has a close affinity to speech. Previous study on speech-to-Singing [2,3] focus on converting the speaking voice to the singing voice. The success of this conversion can also suggest that the ability to sing is a good indicator of the ability to imitate speech. Therefore, a singing text-to-speech synthesis, which converts a voice singing any text (e.g., the lyrics of a song) into a speaking voice is possible.

The research of singing text-to-speech synthesis can facilitate the both conversions between singing and speech voices from investigation of their acoustic differences [4,5]. Manipulating singing voices for speech means the singing synthesis and also becomes talking synthesis with the same timbre. As a tool of virtual singer like UTAU, the end users will also find it interesting for rapping purpose using this synthesis technique.

We propose a novel talking voice synthesis system, "Singing Text-to-Speech" that can convert a singing voice to a speaking voice while keeping the voice quality of singer's voice bank. The singing text-to-speech conversion is based on the speech analysis in the spectrogram that allow for visualization of pitch contours. It is helpful for control the singing voice to talking voice. The primary singing text-to-speech conversion requires four acoustic features. They are the pitch (F0 contour), the duration of each phoneme of the lyrics, velocity and tempo, from the input singing voice. The timbre extraction is directly using the UTAU voice bank. The process of singing text-to-speech process is shown in Fig. 1.

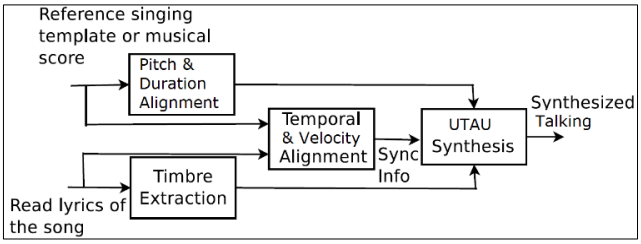


Fig. 1. *The process of singing text-to-speech conversion*

To obtain the target values of these features, singing Text-to-Speech uses an available UTAU software (singing voice synthesis) and supplies the text of the song lyrics to obtain a talking voice. Note that this talking voice obtained by singing is to get the natural quality. Since there's a lot more to both male voice and female voice types for UTAU singing voice bank, the singing text-to-speech applications extend the variety of synthesized voices that can be obtained.

II. SINGING VOICE AND TALKING VOICE

Conventional studies focus on three points: pitch contour, phoneme duration and power, to clarify the differences between singing and speaking voices. These differences are explained in the following. The acoustic difference in speech and singing is explained below using the example of spectrogram analysis shown in Fig. 2.

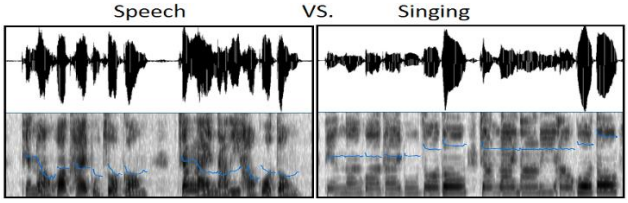


Fig. 2. *The process of singing text-to-speech conversion*

First, confirm that you have the correct template for There are three characteristics of differences in the F0 contours between speech and singing voices [6].

(a) The dynamic range of the speech F0 contours is wider than that of singing voices, while the singing voice has higher pitch than speech.

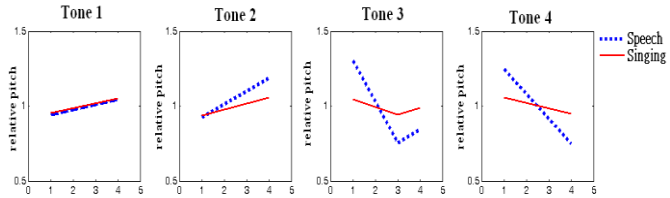
(b) The singing voice has a steady state of an F0 contour corresponds to a Note. The note changes of the F0 contours correspond to melody.

(c) There are many F0 fluctuations that are observed only in singing voices, while the speech has tonal pitch bend.

The thing about talking vs. singing is that singing consists primarily of notes that that work at a consistent tempo, and have mostly straight pitch that can be marked down on a music sheet, and UTAU's interface reflects this, whereas in speech, tempo can be more sporadic, and pitch just starts flying all over the place.

A. Pitch in Tones

In singing, duration of syllables changes according to reference score. A syllable is not stretched equally. To investigate stretch characteristic of consonants, we calculated ratio between the duration of consonants in singing and speech. The result demonstrates that stretch ratio of a consonant is stable and depends on type of the consonant. Mandarin Chinese is a tone language, in which there are four pitched tones. In speech, pitch variety inside a syllable depends on its tone. A composer should be able to find out melody which matches tone of lyric or vice versa. Unfortunately, not all melody matches its corresponding lyric perfectly. Fig. 3 reveals pitch variety ranges in speech and singing. It can be observed that pitch inside a syllable is stable and independent of tone of a syllable. Pitch variety range inside a syllable in singing is about 1.6 semitone, while



that of a tone 4 syllable in Chinese can be 8.4 semitone.

Fig. 3. *Pitch variety ranges in speech and singing*

For the singing voice, a musical note corresponds to a steady state of the pitch contour. A musical score therefore corresponds to the pitch contour that has a step-like shape [7], as shown in Figure 1. For the speaking voice, the pitch contour has a fluid shape that has a low frequency at the beginning and end of each utterance.

B. Phoneme Duration

For the singing voice, the duration of each phoneme changes in accordance with the musical score. For the speaking voice, on the other hand, the duration of each phoneme has a relatively similar length. To be precise, corresponding consonant parts have approximately the same length, the boundary between a consonant and a succeeding vowel also have approximately the same length, and vowel parts have different lengths.

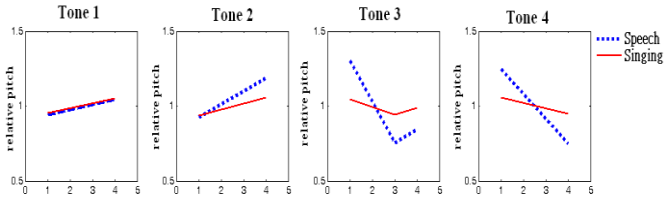
It has been reported that, in singing, note onsets are located at vowel onsets rather than at consonant onsets. The phonemes of the lyrics must be distributed between the notes such that the transitions between notes coincide with the onset of the vowel or set of vowels. In this way, considering one syllable, the consonantal phonemes located before the first vowel will be pronounced within the previous note interval. The result is a redefinition of the borders of the syllables.

C. Accents

Mandarin is a tonal language. The perceptual correlation of phrase accent in Mandarin includes pitch and timing. That is, words are marked with changes in larger fundamental frequency range and longer duration is generally perceived as phrase accent. These two features are the acoustic correlates of pitch and timing respectively. Recently, perceptual experiments and acoustic studies showed that the timing might serve as the primary cue to the prominence and the presence of prominence increases word duration.

D. Tempo

It is well known that each Mandarin character is pronounced as a syllable. For mandarin speaking rate, 120 words per minute are equal to 120 syllables singing in 120 BPM.



E. Velocity

For the singing voice, velocity changes are synchronized with pitch. For the speaking voice, the power always varies continuously.

III. MAKING UTAU TALK

The singing-to-speaking synthesis system has the following input and output:

- Input: Singing voice and lyrics of the song.
- Output: Synthesized speaking voice.

The voice conversion is achieved by changing characteristics of the three different acoustic features,

i.e., phoneme duration, F0 contour, and power, into characteristics of acoustic features generated by TTS. These three features are chosen since they are the main differences between singing and speaking voices, as discussed in Section 2. The system extracts three acoustic features from comparing the singing voice and speech voice using the spectrogram analysis program called PRAAT [8,9]. The following procedures are used to make UTAU talk.

A. Set the Tempo

We can make the tempo whatever you want, but we often find that it's easier to make it faster than the default 120 BPM. Usually, for talking speed, the tempo is set to 180 BPM, as syllables can often comfortably take up an entire eighth note at that speed; of course, you'll probably want to make the BPM higher or lower depending on the sort of talking emotion. For this example, I'm setting the tempo to 175 BPM.

B. Entering Notes and Lyrics

For singing Text-to-Speech, the first thing to do is laying out the number of notes needed for the string of talking words. A music score editor that accepts a MIDI file and lyric input was used as the front-end user interface for this synthesis system. In Fig 4, the Mandarin words have 12 syllables, so 12 notes are placed, and then type in the lyrics. For this particular example, all the notes are quarter notes. We can play it to make sure it's at a speed we want, and adjust the tempo as needed.

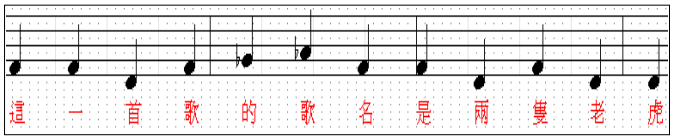


Fig. 4. *The process of singing text-to-speech conversion*

In general, at beginning of a Mandarin phrase, duration of the syllable tends to be larger and as length of the phrase goes to longer one, duration of the syllable at end of phrase turn to be shorter. Therefore, in Fig. 5, the piano view editor of UTAU software can be used to adjust each duration of syllable.

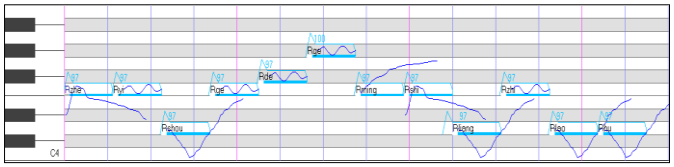


Fig. 5. *The process of singing text-to-speech conversion*

C. Spacing Out Notes

Now we'll adjust the notes in relation to each other to make it sound more natural for speech. Now, unlike singing, speech doesn't fit as comfortably in a consistent rhythm, so some notes are going to working outside of the boundaries of half-notes, quarter notes, eighth notes, etc. Make sure your default quantization and length quantization is at 1/64, otherwise we won't be able to move them around so easily.

D. Pitch Bending

There are four tones in Mandarin Chinese. They differ from each other by the changes of their pitches. As shown in Table I, every syllable in Mandarin can have one of four tones. Every tone can represent different meaning.

TABLE I. TABLE STYLES

Type	Syllable	Tone	Gloss
Tone 1	Ma1	High level	"mother"
Tone 2	Ma2	Rising	"hemp"
Tone 3	Ma3	Low-falling	"horse"
Tone 4	Ma4	Falling	"scold"

Fig. 6. shows that a UTAU has a plugin called pitch bend editor. Control points are used to shape four Mandarin tones in a tune that sound nice and nature.

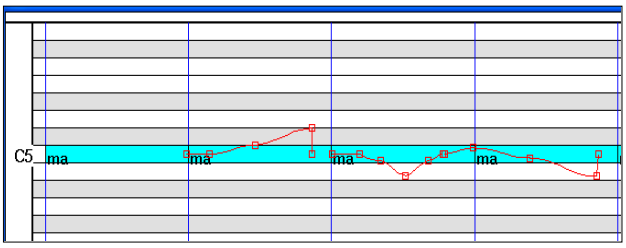


Fig. 6. *Pitch bend editor is used to create Mandarin four tones*

E. Moving Notes Around

Now that it's starting to resemble speech a little bit, the last thing we need to do is move notes around (vertically, that is). Note that on this entire demonstration, I have kept the notes I've made on a single note, that note being C3 on the piano roll, this was intentional, as this way, I can change what parts of the phrase are stressed once I've decided on how the pitch bends are placed. This allows us to change what parts of a phrase are emphasized, and have more stress.

Unfortunately, this is mostly something you need to have intuition on to do properly, so if you don't have that, it'll mostly be guesswork until you get something nice. After having played around with it a little bit, this was what I got.

IV. EXPERIMENTS

We evaluated our system by conducting two psycho acoustic experiments. First, we compared the timbre of synthesized speaking and singing voices, and then evaluated the perceptual similarity in their voice timbre. Second, we evaluated the naturalness of synthesized speaking voices when the mean F0 was varied. In order to evaluate Mandarin MIDI-to-Singing [10], the synthesized songs are available online at http://sing.dmd.isu.edu.tw/en/mandarin.html.

A. Evaluation of Naturalness

In this experiment, we evaluated how well the system retained the voice timbre. To do this, we synthesized speaking voices for a given set of lyrics from different singers, and asked the subjects to "match" the singing voice to the synthesized speaking voice, as depicted in Figure 9. If the timbre unique to each singer was retained, the subjects should be able to match the singing voice to the synthesized speaking voice.

B. Evaluation of Voice Timbre

In this experiment, we evaluated how the perceived naturalness of the synthesized sound changed as the mean pitch contour of the synthesized voice was changed. Each subject will listen to the synthesized talking voice samples.

TABLE II. DESCRIPTION OF THE OPINION SCORE

Description	Score
Highly nature	5
Natural	4
Fair	3
Unnatural	2
Highly unnatural	1

V. CONCLUSION

The proposed singing Text-to-Speech is a novel system to synthesize a talking voice from a singing voice while retaining the timbre of the singing voice. This method has the efficiency of Using UTAU software to create ideal talking voice. The system is based on manipulation of the pitch contour, the phoneme duration, and the power. Experimental results showed that our system is capable of retaining the timbre that is unique to a particular singer while changing aspects other than the timbre to a speaking voice.

In the future, we plan to improve the duration control and to develop a way to gradually change from a singing voice to a speaking voice in order to realize a morphing function between singing and speaking voices. Comparing the quality of the synthesized voice when the target signal is a real talking voice instead of a synthesized text-to-speech is another future work.

References

[1] Singing voice synthesis tool UTAU download site: "http://utau2008.web.fc2.com/"

[2] D. Schwarz, "A system for data-driven concatenative sound synthesis," in Proc. Digital Audio Effects, 2000, pp. 97–102.

[3] T. Saitou, M. Goto, M. Unoki and M. Akagi, "Speech-to-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices," 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2007, pp. 215-218, doi: 10.1109/ASPAA.2007.4393001.

[4] K. Vijayan, H. Li and T. Toda, "Speech-to-Singing Voice Conversion: The Challenges and Strategies for Improving Vocal Conversion Processes," in IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 95-102, Jan. 2019, doi: 10.1109/MSP.2018.2875195.

[5] Lindblom B., Sundberg J. (2007) The Human Voice in Speech and Singing. In: Rossing T. (eds) Springer Handbook of Acoustics. Springer Handbooks. Springer, New York, NY. https://doi.org/10.1007/978-0-387-30425-0_16

[6] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech- to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in Proc. IEEE Workshop on Applica- tions of Signal Processing to Audio and Acoustics, 2007, pp. 215–218.

[7] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrim- ination between singing and speaking voices," in Proc. Eurospeech, 2005, pp. 1141–1144.

[8] Fang Chen, et al. Natural sounding embedded Text-To-Speech systems, proceedings of 5th National Conference on Modern Phonetics, 2001

[9] J. Sundberg, "Articulatory interpretation of the "singing formant"," The Journal of the Acoustical Society of America, vol. 55, pp. 838–844, 1974.

[10] http://sing.dmd.isu.edu.tw/en/mandarin.html.