Tune Helper: a program performing speech-to-sing and auto-tuning in MATLAB

Chenxing Wu
Northwestern University
chenxingwu2014@u.northwestern.edu

Zhe Chen
Northwestern University
zhechen2014@u.northwestern.edu

Maxin Chen
Northwestern University
maxinchen2014@u.northwestern.edu

ABSTRACT

Tune Helper is a program that performs three different functions: "Speech-to-sing", tuning a human voice in a certain tempo to sing a simple melody; "Auto-tuning", creating perfectly tuned vocals by altering off-key pitches; ”Speech-to Rap”, segmenting speech in beats and matched beats with a background drum set music. It is written and built in MATLAB. We use following audio processing methods to achieve these functions: pitch tracking, to analyze input human voice and melody, obtaining their time, amplitude and frequency information; beat tracking, to detect beat and match voice with input melody; phase vocoder and resampling, to time stretch and compress.

Keywords

Auto-tune, Tune Helper, pitch shifting, beat tracking, onset detection

1. INTRODUCTION

The music processing software Auto-Tune is widely used in music industry. Musicians and singers use Auto-Tune to create cool vocal effects during recording, auto-correcting off-keyed pitches as well. Our project is inspired by music processing Auto-Tune[1]. We want to implement it with the basic function of Auto-Tune in pitch auto-correction on MATLAB by ourselves. Next, we want to create an interesting function, which uses human speech and a simple melody as input, combining human speech with melody via audio processing in MATLAB, and generating an audio which is a tuned human speech whose pitch is exactly matched to that of the input melody. Since this function requires a rhythmed input, like someone read a sentence in certain tempo, we figure out another to implement onset detection by analyzing the amplitude of input speech.

There are two essential challenges in this project. The first one is syllables detection in speaking voice. Pitch tracking system is one of the most widely used methods. Periodicity-time-domain methods which utilize autocorrelation-like operations and frequency-domain methods that rely on Fourier transform-like operations are popular models in speech recognition under pitch-tracking system [2]. Also, beat tracking is a useful and accurate tool in onset detection, as long as there are detectable tempos of the sound [3].

The second challenge is pitch shifting while preserving sound quality. Phase vocoder is the most common method. By matching phase differences between adjacent processing windows after pitch shifting, phase vocoder can generate a high-quality audio signal modification [4,5]. Therefore, we decide to use phase vocoder as our essential method in pitch shifting. Both methods using pitch tracking and beat tracking system of onset detection are applied here.

2. RELATED WORKS

There are several existing software and mobile apps that are related to our project, showed in the following.

2.1 Software: Auto-Tune

Auto-Tune is a software tool to create perfectly tuned vocals. It is based on the phase vocoder principle. Also, it is widely used in music industry.

2.2 Mobile App: Songify by Smule

Songify can automatically turn speech into music, creating the vocal effect shown in the popular video "AutoTune the News" on YouTube.

3. TUNE HELPER

3.1 Structure

Tune Helper program consists of three functions, Auto-tuning, speech-to-sing, and speech-to-rap. With the last two function, it can shift pitches of the speech into a song, and modify rhythm to create a speech-rap sound accompanied with drum set. The structure and core functions of this program are showed in Figure 3.1 as follows.

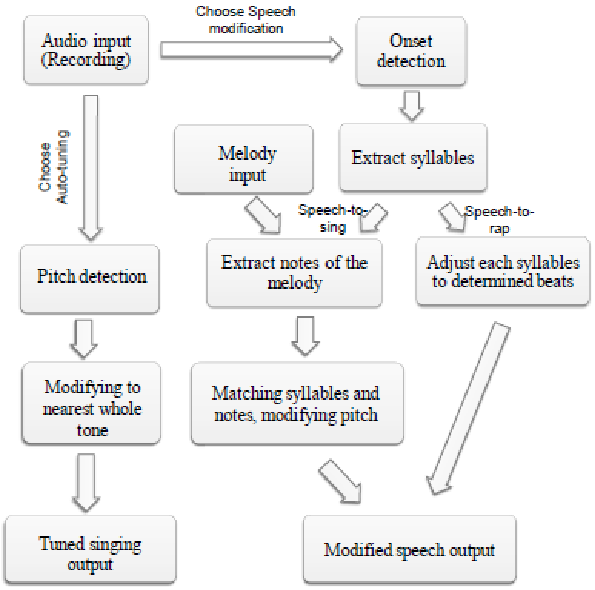
Audio input (Recording)
Choose Speech modification
Onset detection
Choose Auto-tuning
Melody input
Extract syllables
Speech-to-sing
Speech-to-rap
Pitch detection
Extract notes of the melody
Adjust each syllables to determined beats
Modifying to nearest whole tone
Matching syllables and notes, modifying pitch
Tuned singing output
Modified speech output

Figure 3.1 System structure

3.2 User Interface

The user interface of Tune Helper is written and built in MATLAB, as well.

User can record vocals and speech when hitting the "Record" button showing on upper right. For the functions of output audios, it can tune singing, shift speech into a song, or create "rap" by adjusting syllables of input speech to determined beats. User can choose what they want by clicking buttons and import audio files. Output audio file can be listened and shown as waveforms. The small “note” button below the G clef is a “help” button. If user clicks that button, a user guide window will pop out to tell how to use tune helper.

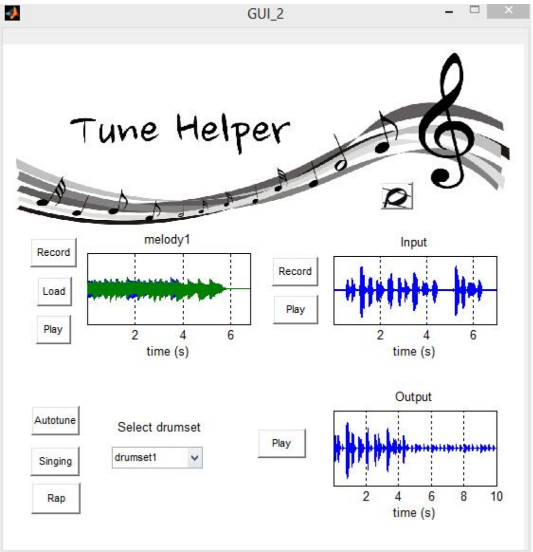

Figure 3.2 User Interface of Tune Helper

3.3 Methods

3.3.1 *Primary methods:*

Pitch tracking: It is used to extract pitches, time, and amplitude information from input audio. [6]

Beat tracking: It is used to detect onsets and match syllables to notes in a melody.

Phase Vocoder and resample: They are used to perform time expansion/ compression and pitch shifting of audio clip.

3.3.2 *Onset detection*

We use two methods in onset detection.

The first one is beat tracking. The program firstly measures tempo of the speech. It then uses the estimated tempo to track beat points in the speech. We use the calculated beat points as onset to segment syllables in the speech.

The second method is localizing amplitude valleys, calculated by the pitch-tracking function, as onsets. In order to avoid including small bumps as onsets, it is set to ignore valleys whose frame numbers are different from neighbor valleys in less than 10.

The first one is more accurate, but requires the speech to have a detectable tempo. It is used in speech-to-sing part, where users are preferred to talk in a tempo. The second one might miss syllables, but doesn’t have requirement of speech input. It is used in speech-to-rap part, as we only care about re-localizing words according to certain beats.

4. RESULT EVALUATION

We use spectrogram to present the time-frequency information for input and output audio, then comparing the difference between input and output. We demoed our program to people, and it works in most cases. They can hear an auto-tuning output, a speech with certain melody or a speech-rap, coming from their own voice. The problem we have is sometime the output is in bad quality, or the syllables in output sound can be barely recognized.

4.1 Speech-to-sing

We use the famous nursery rhyme "Mary had a little lamb" as the input of Speech-to-sing. First, we recorded a speech of the lyrics of“ Mary had a little lamb”, and then created a piano melody on GarageBand. We use these two audio files as system input.

At processing procedure, Tune Helper first uses beat tracking method to analyze both human voice input and piano melody input, extracting the time, frequency and amplitude information. It then uses phase vocoder to shift the pitch of human voice input, according to the relative pitch height information of piano melody input extracted by pitch tracker. After the processing procedure, the result is a singing speech.

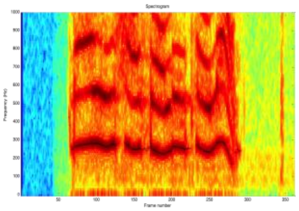

Figure 4.1 Spectrogram of speech

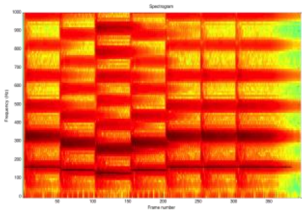

Figure 4.2 Spectrogram of melody

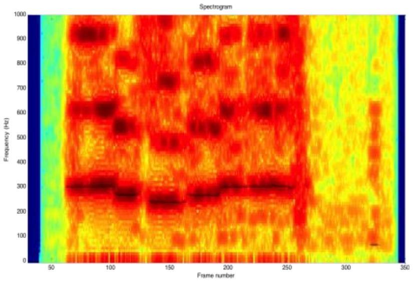

Figure 4.3 Spectrogram of singing-speech

We present the time-frequency spectrograms of human voice input, piano melody input, and output audio, showed below. Comparing these three spectrograms, we can find out that each syllable can be detected and shifted to the notes in melody, showed in Figure 4.3. Using our program, user can import a simple melody, record the speech, and get a singing-speech.

4.2 Auto-tuning

Tune Helper first performs pitch tracking on input vocal, extracting its time, frequency, amplitude and pitch information. It then uses the information to calculate the nearest note for each available frequency point of input vocal. If current point is off-key, auto-tuning function performs a pitch shifting by phase vocoder and resample function, to shift pitch to the nearest music note in C major, which is the seven major notes in one octave. After the processing procedure, the output audio is an auto-tuned vocal. To make output sound similar to electronic sound, we shift each frequency higher by one key.

4.3 Speech-to-rap

A recording of reading a few sentences is imported, processed with "Speech-to-Rap" function, and output a rap sound accompanied with drum set is generated. Speech will split to syllables by the second onset detection method (localize voice valley), applied with pitch-tracking and auto-tuning function to make it sounds better. Then, we split the tuned speech again to voice segment, aligning each of them with beats from drum set we choose. We tried to apply phase vocoder function to adjust each segment to fit the length between two beats. However, the output is distorted, so we choose not to use phase vocoder here. This approach is still on trial, as the amplitude-onset is less accurate, which could not guarantee a consistent segmentation of words.

5. CONCLUSION

In our project, we achieved both auto-tuning and speech-to-sing functions by using pitch tracking, beat tracking and phase vocoder. We use spectrograms to analyze and compare the result with input. We can get an ideal audio output by our program, Tune Helper.

However, the quality of output audio is not very satisfying. We use two different versions of phase vocoder from internet, but neither of them is very reliable. Meanwhile, the output from speech-to-rap function is not very good. Some word segments in rap output sounds blurring and sometimes it has clippings. The algorithms of phase vocoder and speech-to-rap can be modified and improved in the future.

6. FUTURE WORKS

Our program is relatively slow, comparing to commercial software Autotune. For example, it may costs 8 seconds to wait for output from speech-to-sing function. We still need to modify algorithm and codes to shorten the processing time.

Also, we hope to improve speech recognition function to distinguish syllables when the input audio is noisy or the syllables are not really articulated. If this function is achieved, we can tune a talking into melodies, regardless that whether it has a detectable tempo or not; and hence generate a similar effect as seen in Songify. Also, more works need to be done to modify speech-to-rap function, to make the output from rap function sounds more like a real rap.

To improve user's experience of the program, we can add different genre of background music in speech-to-sing/rap function to make sound effects much cooler.

7. ACKNOWLEDGMENTS

We'd like to thank all of the classmates in this course, and Professor Bryan Pardo, for his suggestions and supports for our project.

8. REFERENCES

[1] Tyrangiel, Josh. "Auto-tune: Why pop music sounds perfect." Time Magazine (2009): 1877372-3.

[2] Byung Suk Lee and Daniel P. W. Ellis. "Noise Robust Pitch Tracking by Subband Autocorrelation Classification." in 13th Annual Conference of the International Speech Communication Association, 2012.

[3] B. McFee and D.P.W. Ellis, "Better beat tracking through robust onset aggregation," in International conference on acoustics, speech and signal processing, 2014, ICASSP

[4] Laroche, J., Dolson, M., New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, Oct 1999

[5] Laroche, Jean, and Mark Dolson. "Improved phase vocoder time-scale modification of audio." Speech and Audio Processing, IEEE Transactions on 7.3 (1999): 323-332.

[6] Middleton, Gareth. "Frequency Domain Pitch Correction." Connexions, December 17 (2003).