Vocal Music Cheater Project Final Report

Ruiyang Jin
College of Fine Arts
ruiyangj@andrew.cmu.edu

Zichong Yang
Department of Mechanical Engineering
zichongy@andrew.cmu.edu

Keting Zhao
Electrical and Computer Engineering
ketingz@andrew.cmu.edu

Rohit Manthena
Electrical and Computer Engineering
rmanthen@andrew.cmu.edu

1 Introduction

People love to sing, but only a few of us can sing well. While most of us go off tunes or can't follow the rhythm from time to time, as we are not trained or don't have the gift for singing, almost everyone can speak in their own language naturally and fluently. Therefore, with the help of singing voice synthesis techniques, it's possible to develop a system that allows people to produce singing voice with the right pitch and rhythm by just speaking the lyrics.

This project aims at building a speech-to-sing algorithm. The algorithm, called the Vocal Music Cheater, is designed to generate singing audio from a person's speech of plain lyrics reading. It will take a piece of speech audio and the corresponding score as input, and output a piece of singing voice with the voice of the speech and the pitch and rhythm specified in the score. With the help of Vocal Music Cheater, people will sing without knowing how to sing correctly.

2 Related works

To bolster system fidelity, preprocessing such as denoising can be performed using a method proposed by Wilson et al. (2008) for nonnegative matrix factorization (NMF).

After preprocessing, the very first step is pitch detection. YIN is one of the most popular algorithm for pitch detection proposed by De Cheveigné, Kawahara (2002), and Mauch, Dixon (2014) proposed pYIN based on their work.

To convert speech to singing, the pitch curve and durations of each word needs to be modified according to the score, and phase vocoder, proposed by Flanagan, Golden (1966), is a classical signal processing method which can achieve pitch shift and timescale modification at the same time. Charpentier, Stella (1986) proposed PSOLA algorithm, and Moulines, Charpentier (1990) gave a more comprehensive review of the algorithm and its time-domain and frequency-domain variation (which is TD-PSOLA and FD-PSOLA). Though initially used for improving the quality of diphone-based text-to-speech synthesis, PSOLA algorithm can also be used for pitch correction and duration adjustment, and has been applied to singing voice synthesis, though few of those works have clear metrics and quantified results.

In more recent research, people like Saino et al. (2006) proposed to use HMM-based TTS system for singing voice synthesis, but few of them have explored the possibility of using HMM as a speech-to-sing system. At the same time, some people begin to explore in speech-to-singing problem, but most of their works don't have reliable codes or results, or just use deep neural networks. Saitou et al. (2007) proposed a speech-to-singing system based on the STRAIGHT vocoder, but they only used speaking voices of 2-second length in their experiments and only used subjective metric.

3 Datasets

We used NUS-48E Sung and Spoken Lyrics Corpus dataset proposed by Duan et al. (2013) , which contains some songs and the spoken version of their lyrics, as well as the complete set of transcriptions and duration annotations at the phone-level for all recordings. For one speech-song pair, the system takes the speech audio file as the first input, take the annotation of the speech as the second input to get duration of each word, and take the annotation of the song as the third input to get the score. After that, the system synthesize a singing voice from the input speech, with the pitch and duration in the given score.

Since the unit of the annotations is phoneme, it is too fine-grained and therefore not convenient for determining the boundary of syllables with different pitches, we preprocessed one of the annotation files to use syllable as unit, and added the pitch of each syllable we got from the song file.

4 Proposed methods

The overall structure of our system is shown below in fig 1. To make a more robust system, we first perform denoising with NMF on the input speech, so that if given a noisy speech, the noise won't affect the performance of following steps. Then we extract the pitch curve of the speech and calculate the position and duration of each word. Finally, we modify these pitch and duration of words according to the given score. Since there is currently no standard method with reliable performance, and also we can't promise the success of any method in this new problem, we investigate and compare two different algorithms: phase vocoder and PSOLA.

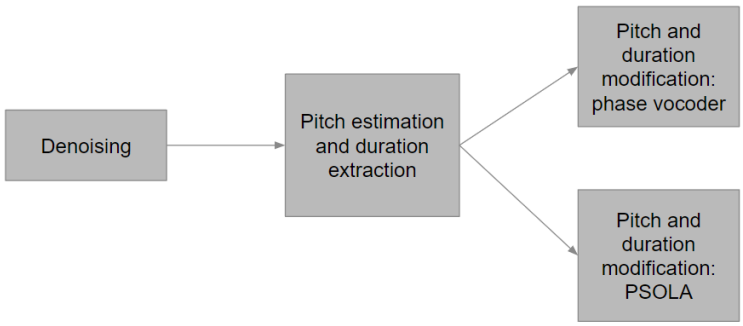

Figure 1: The structure of our speech-to-sing system

4.1 Denoising

We use NMF to denoise the signal before performing pitch detection. The objective function we use here during training is Kullback-Leibler (KL) divergence between clean speech/noise spectrogram and the product of learned basis and weights.

4.2 Pitch detection

We use the YIN algorithm for speech processing. For this, we first compute the autocorrelation of the input signal. To further improve it we calculate the difference and the cumulative mean normalized difference function which uses the concept that if the input signal is periodic then taking the square and averaging over a window will still yield a periodic signal. After this we perform parabolic interpretation and choose a global minima. This global estimate gives us the best pitch estimate within a small window.

4.3 PSOLA

Then we feed the pitch and duration information into the modification algorithm, and one of them is PSOLA. The procedure of the algorithm is shown in Figure 2. This algorithm first put pitch-synchronous marks on the input speech, then window the speech to get short time signals centered

around those marks. The intervals of those short time signals are then changed according to the pitch on the score, while the number of them are adjusted according to the duration. After the adjustment, synthesizing those short time signals by overlap-add will change the speech to a singing audio with the pitch and duration given by the score.

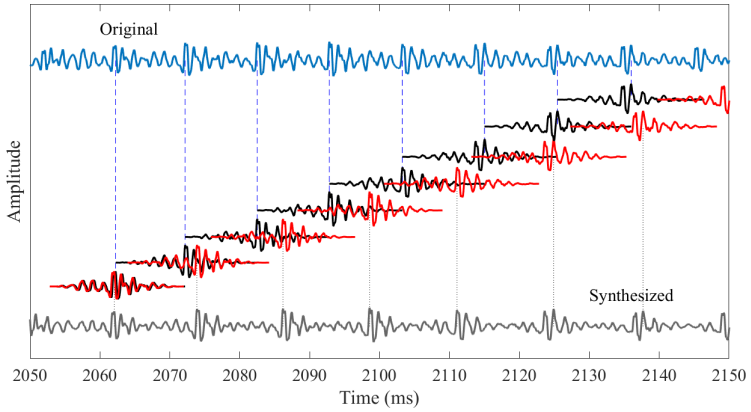

Figure 2: PSOLA algorithm

4.4 Phase vocoder

Another method for time stretching and pitch shifting we experimented is the phase vocoder. The phase vocoder is a variation on the STFT(short-time Fourier transform) which converts a time domain representation of sound into a time-frequency representation, allowing modifications to the amplitudes or phases of specific frequency components of the sound before resynthesis into the time domain by the inverse STFT. Modifying the time position of the STFT frames prior to the resynthesis operation allows for time-scale modification of the original sound file, while the pitch shifting is achieved through resampling of the time-stretched signal.

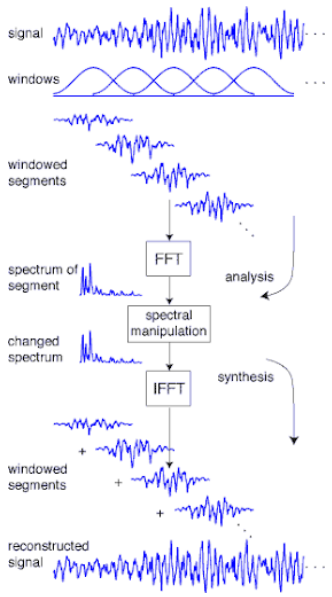

Figure 3: Phase vocoder algorithm

5 Evaluation metrics

If a piece of voice have certain pitch and duration patterns, then it will sound like singing, so the correctness of the pitch contour could me the most important metrics to judge the synthesized singing voice. We use the similarity between the pitch contour of the authentic singing voice and that of the synthesized singing voice as our metrics. However, the common similarity metrics such as L2 distance or cosing similarity doesn't take alignment into consideration, and the score will change a lot if the pitch contour is shifted or not aligned in time. Therefore, we choose to calculate the similarity using dynamic time wraping, and compare the result between PSOLA and phase vocoder.

6 Experiment results and analysis

In the denoising stage, We used the basis matrices of speech and noise that is trained after 250 iteration to get the reconstructed speech spectrogram from the noisy speech spectrogram. The reconstructed speech spectrogram after 500 iteration and the original clean speech spectrogram are shown in figure 4. From this given result, we expected that our system should be able to perform pitch detection given a noisy speech input.

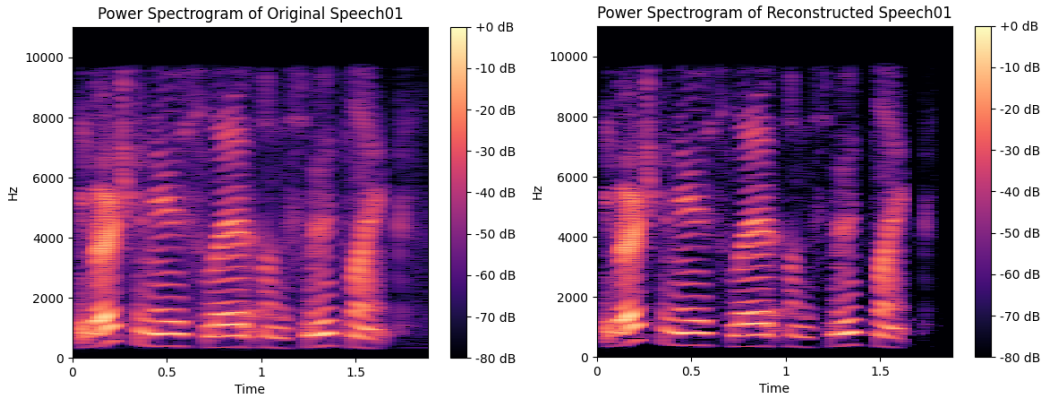

Figure 4: A speech spectrogram (left) and its reconstruction version (right)

Figure 5 shows the result of performing PSOLA with the noisy speech input and with the denoising speech input. Although the contour didn't shows much difference, several unexpected spikes appear in the output when using noised speech. The singing synthesized with noised speech did contain much more background noise than that with the denoise speech.

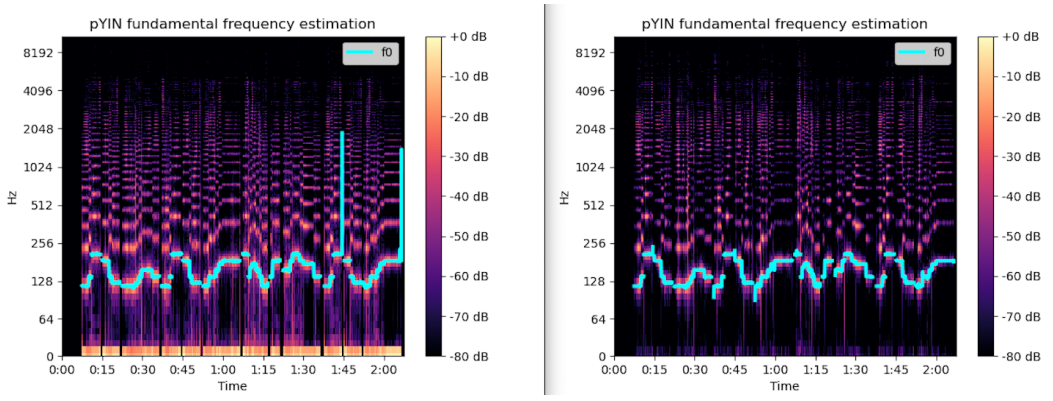

Figure 5: Pitch contour of the noised speech (left) and denoised speech version (right)

Figure 6 shows the pitch estimation curve for one of the audio files in the NUS-48E Sung and Spoken Lyrics Corpus Database. (For consistency, all the results in this section use the first speech-song pair

in the dataset as the input) To get the duration of each word in the speech we computed the difference in times between two consecutive silence's. One limitation with this approach is that if there are no silences in the input file then we cannot extract the duration of words in the speech as we will not know where to split.

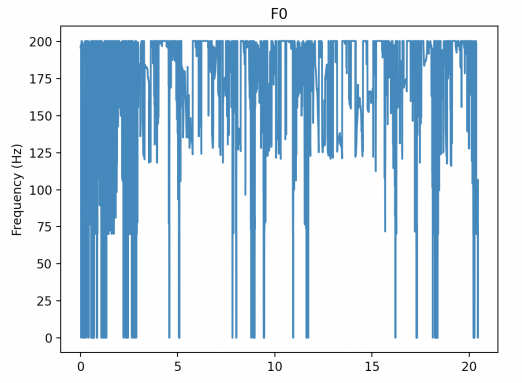

Figure 6: Pitch Estimation Curve

The waveform of one speech segment, the voice synthesized from it by PSOLA and its corresponding song file is shown in figure 7. As shown in figure 8, the pitch contour of the synthesized voice is rather close to that of the authentic singing voice. However, both the waveform and the spectrogram show that apart from the pitch and duration given, PSOLA can only capture basic information like harmonic of the fundamental frequency, and the result still lacks detail and sounds unnatural. Some possible reason could be that PSOLA only utilizes pitch and duration information in the song while ignoring other details, or it just use overlap-add to resynthesize singing voice from short time signals in the speech without further modification, and thus can't fill the gap between the speech and the song.

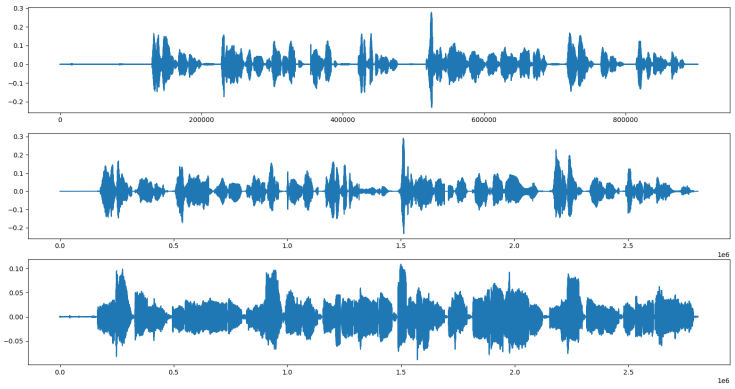

Figure 7: A speech segment (upper), its synthesized version (middle) and the corresponding song file(lower)

For the phase vocoder, we implemented the algorithm with the help of the Rubber Band library as well as its command line tool wrapper pyrubberband. We also use the NUS-48E Sung and Spoken Lyrics Corpus Database. Using the recording of plain reading version of the lyrics, detected pitch of the audio, as well as the score of the song, we are able to reconstruct the singing voice. Since the speaker and the singer are the same person, we also implemented on the speech of reading the same lyrics. The results are shown below in figure 9. The reconstructed pitch of the audio is clearly recognizable, and time stretching successfully recovers the rhythm of the original pitch while matching the duration of the music score. However the generated audio still have the chipmunk effect, indicating the vocal formants are also shifting in the direction of the changing pitch. And the audio would suffer from distortion especially when the time or pitch shifting ratio is very large.

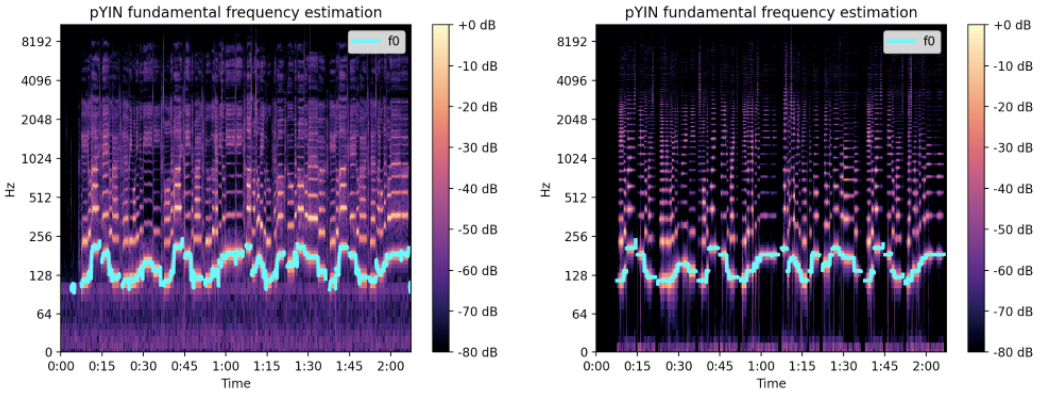

Figure 8: Pitch contour of the authentic song (left) and voice synthesized by PSOLA (right))

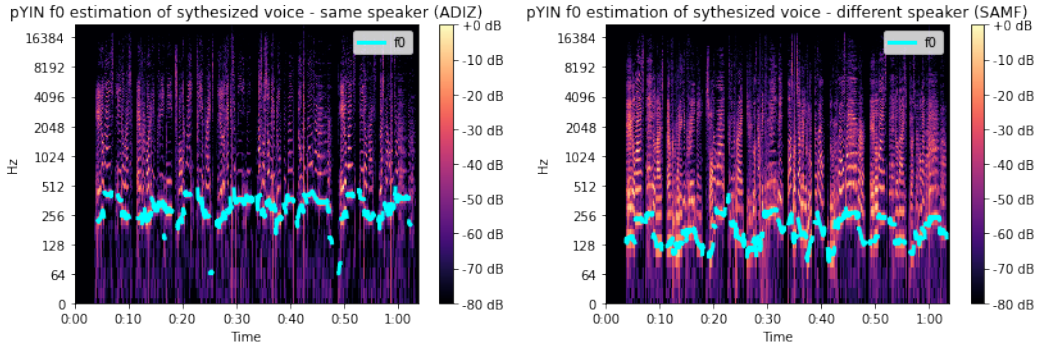

Figure 9: Spectrograms with the pitch contour detected from audio synthesized by phase vocoder, left: using reading voice from the same person as the singer of the reference score, right: generated singing using different speaker to read the lyrics. Pitch of the authentic song can be found in figure 8.

To compare the performance of PSOLA and phase vocoder, we calculated the (unnormalized) similarity score of the two methods, which is shown in table 1. As we can see in the table and in the pitch contour images above, PSOLA can generate singing voice with more stable pitch compared with phase vocoder, and the quality is higher. However, both methods can't produce sufficiently natural singing voice and the voice is a little bit metallic. Also, using speaker different from the reference singing score would not affect the overall performance as the order of magnitude keeps in the same level.

Table 1: Similarity scores calculated by DTW

Algorithm	Speech Speaker	Reference Singer	Compare Object	Similarity score
PSOLA	ADIZ	ADIZ	ADIZ	5679922.97
PSOLA	SAMF	ADIZ	ADIZ	4222097.56
PSOLA	SAMF	ADIZ	SAMF	46852791.86
Phase vocoder	ADIZ	ADIZ	ADIZ	11057549.90
Phase vocoder	SAMF	ADIZ	ADIZ	61154304.42
Phase vocoder	SAMF	ADIZ	SAMF	5672338.69

7 Conclusion and Future work

We proposed to use some existing signal processing and machine learning methods to address a new problem, which is speech-to-singing conversion. We integrate denoising, pitch and duration extraction, and pitch/duration modification algorithms to form a functioning system which convert

speech to singing with given score. Two modification algorithm is compared and PSOLA achieves better result. This project is not good enough for commercial use, but it can serve as a prototype of a basic speech-to-singing system and give ideas to following works.

The methodology of this project still have plenty of room for improvement. One possible future work could be investigating the effect of more machine learning algorithms on shifting the pitch and duration without harming the naturalness of the voice, since PSOLA and phase vocoder are mostly signal processing methods and don't utilize much information in the speech signal. Another possible direction would be finding a way to extract pitch and duration information without annotation, so that the system won't rely on manual annotation and could become automated.

References

Charpentier Francis, Stella M. Diphone synthesis using an overlap-add technique for speech waveforms concatenation // ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing. 11. 1986. 2015–2018.

De Cheveigné Alain, Kawahara Hideki. YIN, a fundamental frequency estimator for speech and music // The Journal of the Acoustical Society of America. 2002. 111, 4. 1917–1930.

Duan Zhiyan, Fang Haotian, Li Bo, Sim Khe Chai, Wang Ye. The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech // 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. 2013. 1–9.

Flanagan James L, Golden Roger M. Phase vocoder // Bell system technical Journal. 1966. 45, 9. 1493–1509.

Mauch Matthias, Dixon Simon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions // 2014 ieee international conference on acoustics, speech and signal processing (icassp). 2014. 659–663.

Moulines Eric, Charpentier Francis. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones // Speech communication. 1990. 9, 5-6. 453–467.

Saino Keijiro, Zen Heiga, Nankaku Yoshihiko, Lee Akinobu, Tokuda Keiichi. An HMM-based singing voice synthesis system // Ninth International Conference on Spoken Language Processing. 2006.

Saitou Takeshi, Goto Masataka, Unoki Masashi, Akagi Masato. 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. 215–218.

Wilson Kevin W., Raj Bhiksha, Smaragdis Paris, Divakaran Ajay. Speech denoising using nonnegative matrix factorization with priors // 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. 2008. 4029–4032.