Reach a Song by Humming the Tune

Tsun-Han Huang
102061221
Department of Electrical Engineering
National Tsing Hua University

Cheng-Ju Lee
102061139
Department of Electrical Engineering
National Tsing Hua University

Abstract

In this project, we create two programs, Query by Singing & Humming (QBSH) and Evaluation of Singing Voice Accuracy (ESVA), which extract the pitch of the singing person. QBSH utilizes Cepstrum to track the pitch of singing voices. It uses the characteristics of the log function to transform the non-linear multiplication into linear addition. ESVA uses the Harmonic Product Spectrum. It accumulates the compact audio data in the frequency domain to magnify the fundamental frequency and thereby extract the singing pitch. We also align the mean of the two person's different frequencies to make the recognition accuracy higher.

1. Introduction

Sometimes, people may remember a song by its melody only, forgetting the exact lyrics. In this situation, reaching a song by humming the tune might be helpful. Therefore, this report focuses on the study of how to recognize a song by humming the tune. In our system, not only can the computer recognize a song, but it can also evaluate one's singing skill.

2. Method

2.1 Pitch Tracking by Cepstrum

Cepstrum can be used to extract the pitch characteristic of one piece of music. The word "cepstrum" comes from the word "spectrum", reversing "SPEC" into "CEPS" to become the new word. Here is the definition of cepstrum:

cepstrum = | IFFT (Log (|FFT(Frame)|)) |

The physical meaning of cepstrum can be illustrated by Figure 2.

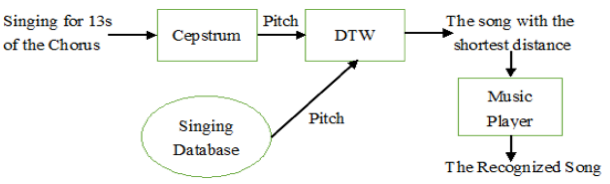

Figure 1. The flow chart of QBSH.

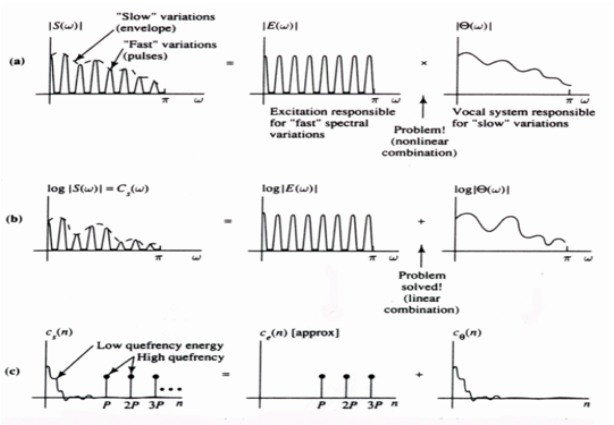

Figure 2. Cepstrum pitch extraction method illustration.

Below are the steps to get the cepstrum :
1. Divide a piece of singing tune into several frames.
2. Perform Fast Fourier Transform on one frame to get the spectrum.
3. Log Conversion on the magnitude spectrum.
4. Perform Inverse Fast Fourier Transform on the log spectrum to get the cepstrum.
5. Use High Pass Filter to get the high quefrency pitch features.

A spectrum can show the frequency components of one frame, but what is more useful for pitch tracking is to get the "pulse-like" components of the cepstrum. Log conversion on the spectrum can change multiplying sign into adding sign, turning the non-linear combination into linear combination, so that the fast variations of the spectrum can be separated from the slow variating envelope. Therefore, performing IFFT back into "quefrency" domain can turn fast variations into pulse-like components that contain the pitch features, and change the slow envelope into low quefrency ones, which is not useful for pitch tracking and can be eliminated by high pass filter. Take the maximum quefrency components of every frame, and the pitch notes of the piece of singing tune can be obtained.

2.2 Query by Singing and Humming

The database is made up of 15 Songs (13-second chorus) sung by Cheng-Ju Lee.

Steps: (Figure 1.)
1. Record one's 13-second singing tune
2. Pitch Tracking by Cepstrum
3. Dynamic time wrapping with the songs in the database (also take the Pitch values).
4. Find the song with the shortest distance.
5. Play the recognized song.

2.3 Harmonic Product Spectrum

In the project of evaluating a person's singing skills, another method, harmonic product spectrum (HPS), is used for pitch tracking in the song. The HPS method is shown in the figure below.

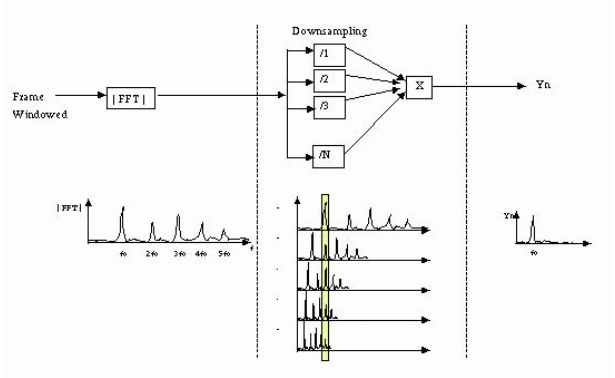

Figure 3. The overall concept of Harmonic Product Spectrum.

After transforming the input audio data from time domain to frequency domain, downsampling is performed. There are many downsampling frequencies applied. They are 1/2, 1/3, 1/4,… of the original one, respectively. When sampling frequency is 1/2 of the original, only half of the data is remained. When sampling frequency is 1/3 of the original, only 1/3 of the data is remained.

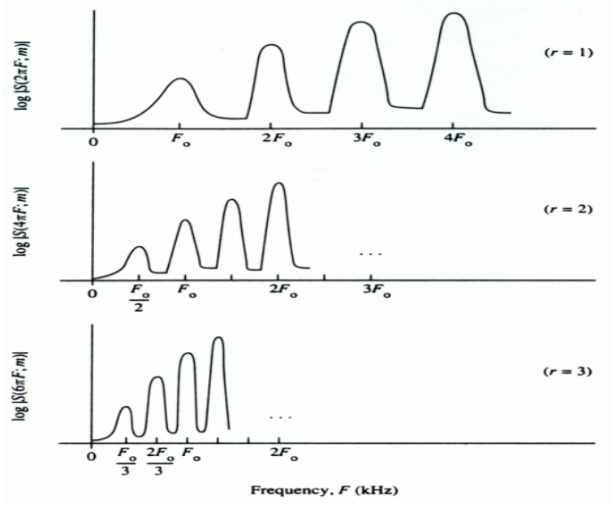

Figure 4. The compact audio data in the frequency dimension.

Until the downsampling process is done, those compact downsampling data is accumulated. Like the figure shown above. Due to the reason that every compact signal will contain a peak point around the fundamental frequency, the result of the accumulation will make this peak point become prominent, which helps us recognize the fundamental frequency more easily. Finally, the pitch of a person's voice can be successfully acquired.

2.4 Dynamic Time Warping

Dynamic Time Warping (DTW) is an algorithm used to measure the similarity between two temporal sequences which may vary in speed. Dynamic Programming is used to perform DTW. The sequences are warped non-linearly in the time domain and a warping path and a distance are produced. It finds an optimal match between the recorded audio signal with those in the database.

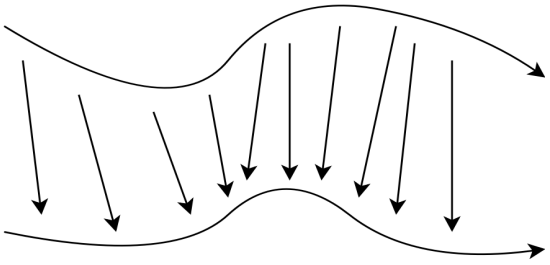

Figure 5. The process of Dynamic Time Warping.

A song recognition can be performed by using the DTW algorithm.

2.5 Evaluation of Singing Voice Accuracy

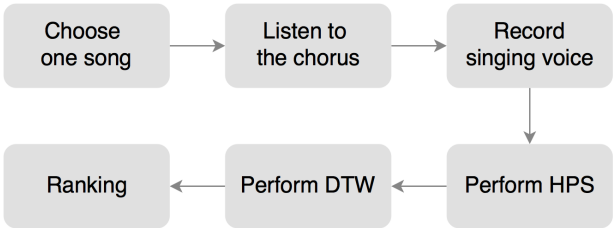

Figure 6. The flow chart of ESVA.

20 songs are collected as the database. They are all of 13 seconds long and mostly contain only pure voice. A user has to pick one from these 20 songs and sing direct to the microphone to record his own pure singing voice. Then, HPS is performed on the selected song and the audio of the user to track the pitch of them. Next, DTW is used to match the user's singing voice with the selected song. Finally, after acquiring the distance between the two audio data, a review and a rank is produced.

3. Result

3.1 Query by Singing and Humming

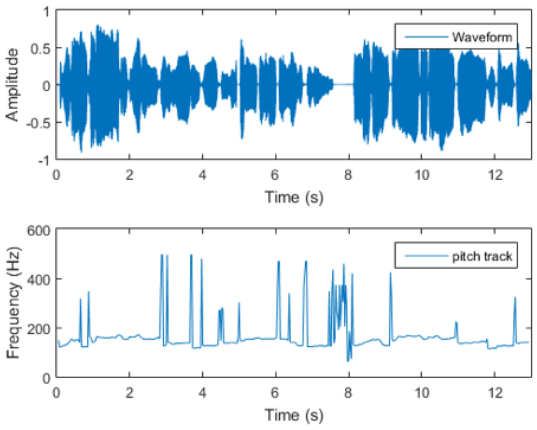

Figure 7. Tsun-Han Huang's Bad Romance Chorus (by Lady Gaga)

Because the database is made up of Cheng-Ju Lee's voice, so the test voice is sung by Tsun-Han Huang. The upper figure is the original time domain data. The Lower one is the pitch tracking data by the cepstrum.

Match the pitch tracking data with those in the database by DTW and get the following recognition results :

```
Command Window

    Academic License

>> demo1
start recording
stop recording
analyzing...
I think the song you sang is :
Lady Gaga-Bad Romance
Good Recognition Result !
fx >>
```

Figure 8. The recognition result

The computer correctly recognize that what Tsun-Han Huang sang is Lady Gaga's Bad Romance, and he DTW Distance is 7749.6. The system will show "Good Recognition Result!" when the DTW distance is smaller than 10000, which represents that the test singing tune is closer to Cheng-Ju Lee's sing voice in the database.

3.2 Evaluation of Singing Voice Accuracy

The standard for evolution is determined by the average distance resulting from many people's singing tests. Most people with nice singing skills receive a distance of around 50000. Therefore, an A-rank is set to those with result distance of 47500~50000, and the rest is shown in the Table below.

Distance	Rank	Distance	Rank
< 47500	A+	60000 ~ 70000	C+
47500 ~ 50000	A	70000 ~ 80000	C
50000 ~ 52500	A-	80000 ~ 90000	C-
52500 ~ 55000	B+	90000 ~ 100000	D
55000 ~ 57500	B	> 100000	F
57500 ~ 60000	B-		

Table 1. The ranking standard of ESVA.

Figures below are the result pitch data of part of the song, "Counting Stars", by OneRepublic and a user's singing voice

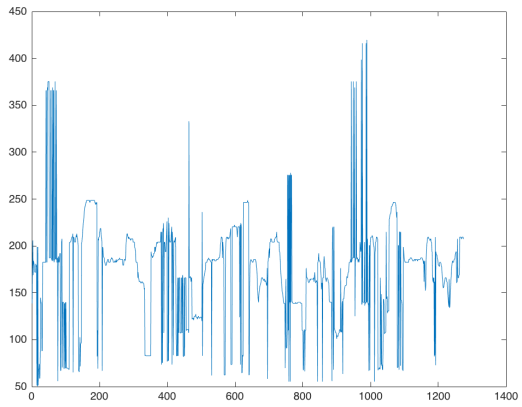

Figure 9. The pitch data of part of the song "Counting Stars".

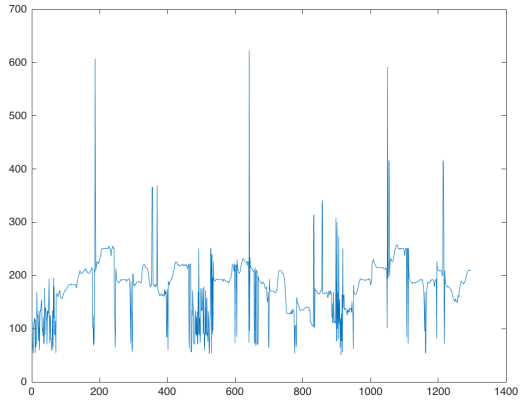

Figure 10. The pitch data of a user's singing voice.

The pitch variation of the singing voice can be observed from the figures. Those pitch data are used to perform DTW and then for evaluation.

```
Command Window

Pick one song and enter the number of its order: 6
Do you want to hear the song(yes:1/no:0)? 1
Are you sure(yes:1/no:0)? 1
Press enter to start recording
start recording
stop recording
Excellent!!! Rank:A+
```

Figure 11. The result of ESVA.

4. Discussion

The average singing pitches may vary from person to person. For example, the average frequency of female singing voice is higher than male one.

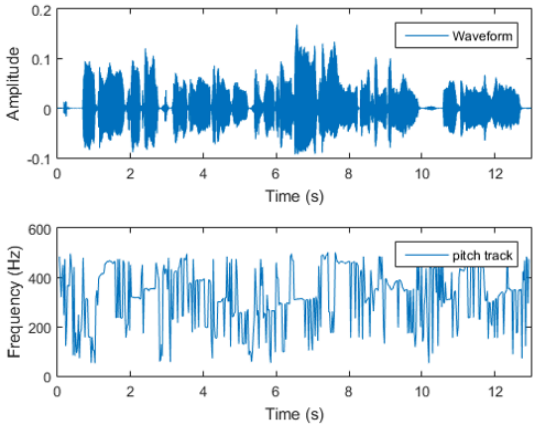

Figure 12. A female tester's singing voice.

The frequency can be up to 400Hz, and the pitch tracking data has more noises.

In the Query by Singing and Humming experiment, because the database is made up of male voice, if tested by female singing voice, the computer may recognize it as the song with the highest pitch, such as Adele's Rolling in the deep. Therefore, shift of frequency is necessary. We adopt the following method to shift the frequency:

D1 = D1 + (mean(D2) - mean(D1));
D1 : The pitches of the test voice
D2 : The pitches from one of the songs in the database

After using the above formula, the mean frequency of D1 and D2 will be equal, so that the distance of D1 and D2 can be reduced and less influenced by the difference of average singing frequency. Therefore, the computer can focus on the pitch change of the song and be more likely to reach the right song.

After using the modification, the query system can recognize female's singing with higher accuracy in some female singers' songs and some male singer's song with higher average pitches.

5. Conclusion

By using Cepstrum and Harmonic Product Spectrum with Dynamic Time Wrapping Method, the pitch can be extracted from a piece of singing or humming, so that the computer can recognize a song just by the pitch without the original lyrics, which is helpful for the query of a song. Also, shifting the mean frequency of the test voice with respect to the mean frequency of the songs in the database can increase the recognition accuracy. With the methods mentioned in this report, the computer is able to reach a song by humming the tune and evaluate people's singing performance.

6. Reference

7-6 Frequency-domain: Cepstrum.
https://mirlab.org/jang/books/audioSignalProcessing/ptFreqDomainCepstrum.asp?title=7-6%20Frequency-domain%3A%20Cepstrum

Project: Pitch Detection.
http://note.sonots.com/SciSoftware/Pitch.html

7-5 Frequency-domain: HPS
https://mirlab.org/jang/books/audioSignalProcessing/ptFreqDomainHps.asp?title=7-5%20Frequency-domain%3A%20HPS