MMSP Term Project Report

利用 AHS model 做音源分離

B95901144 電機四 劉俊麟/B97901047 電機二 呂行

***Abstract*-Music source separation is currently a hot topic, its main goal is to separate the single musical instruments or vocal source from a mixed music signal. We take [1] as our main reference in this project. In a limited frequency range, different musical instrument has its unique structure of base frequency and harmonics, called harmonic structure. Taking the average of the harmonic structure of all the frames in a single source sample, we can get the so called Average Harmonic Structure (AHS) of that source. For a mixed music sample with two or more sources, we can compare the harmonic structures in a certain frame with the AHS model we already have to determine which sound belongs to which source. Therefore we can separate the sources in the mixed sample.**

I. Introduction

Most of the music we hear is played by multiple musical instruments or even human voices. Separating the sources in the music can be helpful in many applications. For example, the single source can be very helpful for data mining or music researches. Or we can use this method to automatically generate accompanying music for Karaoke.

Current researches in this field focus on different features of the music signals. In this project, we will use average harmonic model as our basis.

II. Average Harmonic Structure Model

Different musical instruments have different timbre; the physical meaning behind timbre is the relative intensity and time delay between the base frequency and harmonics. This can be used to distinguish different sources in the mixed music.

First look at one frame of the spectrum of the music sample, we can find some significant peaks; they are the base frequency and harmonics. Their relationship is called the harmonic structure. By taking the average of the harmonic structures in all the frames, we can obtain the average harmonic structure (AHS). AHS has some properties which are useful in music source separation:

1. Different musical instruments have different AHS.
2. For a certain instrument, the AHS doesn't change much over different pitches. But this is restricted in a narrow frequency range.

Our music source separation algorithm consists of two steps: the AHS model learning and source separation. In the AHS model learning step, We use music samples of a single source (musical instrument) to let the program extract its AHS model. In the separation step, mixed music samples are taken as input, and we use the AHS models learned in the first step as a reference to separate the mix signal.

III. AHS Model Learning

To simplify the task, we use a single source instead of a mixed sample suggested in the paper to learn the AHS model. The AHS Model learning can also be divided into two steps: peak detection and harmonic structure extraction.

A. Peak detection

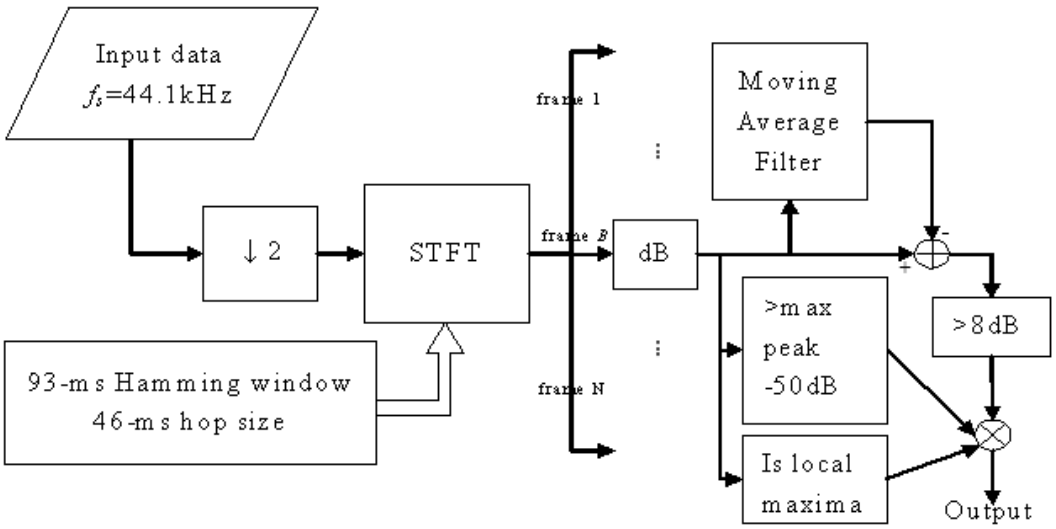

Figure 1: Block diagram of peak detection

The harmonic structure is consisted of relative intensity of base frequency and its harmonics. This appears in the spectrum of each frame as amplitudes (in dB) of corresponding peaks. But we can find that there are a numerous local maxima in the spectrum, most of them are caused by side lobes and noises. We have to set a criterion to filter out those peaks with significance. In this project, we use the following two rules to find significant peaks:

1. The peaks should be higher than a bottom line, which is defined as the global maximum of the spectrum minus 50dB.
2. We use the `smooth()` function (moving average filter) in Matlab to obtain the smoothed envelop of the spectrum. The peaks must be higher than this envelope by 8dB.

By applying these two rules, we can greatly reduce the number of peaks to be considered, and rule out the effect of noises.

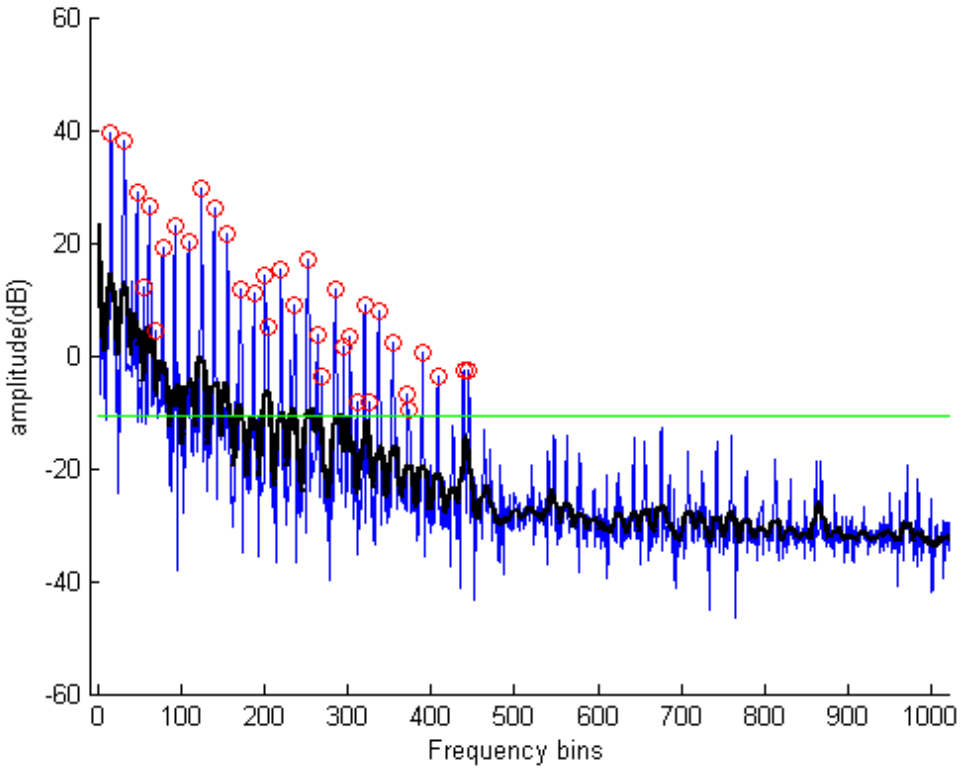

Figure 2: Peak detection: The blue line is the spectrum, the black line is the smoothed envelope, and the horizontal green line is the bottom line. The significant peaks are marked in red.

B. Harmonic Structure Extraction

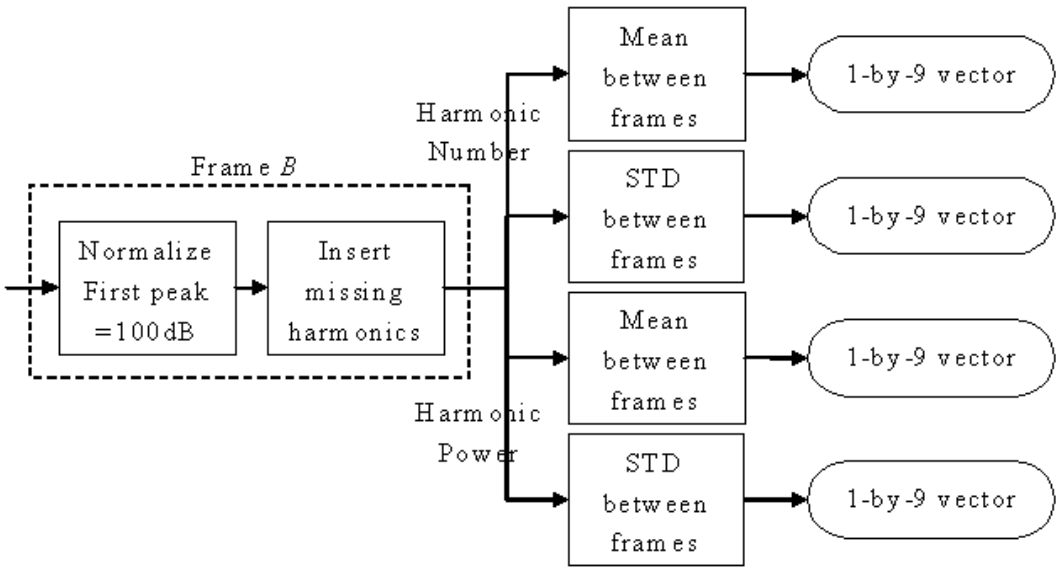

Figure 3: Block diagram of AHS extraction

We then extract the harmonic structures of each frame from the peaks detected in the previous step. After a series of experiments, we found that the peaks at lower frequency range show more consistency than the higher frequency ones over time. So we focus on the first ten harmonics to establish our AHS model. For each frame, we extract the first ten peaks, and we normalize their amplitude to let the first harmonic be 100dB. We also calculate their harmonic number by dividing the frequencies by the frequency of the first harmonic. This is an important feature of the AHS method; we use harmonic numbers instead of frequencies because the frequencies may vary with different pitches.

Ideally, the harmonic numbers should be 1, 2, 3, ..., 9, 10. But in practice, the frequencies may not be perfectly integer multiples. Also some of the harmonics may be too small in amplitude and failed to be recognized in the peak detection step. So we set a tolerance range of 0.5. For example, the i^{th} harmonic should have its harmonic number in the range of (i-0.5)~(i+0.5). Another reason of this loose bound is that there is uncertainty in estimating the fundamental frequency due to discrete Fourier transform. Once the harmonic number *i* grows, it deviates from real harmonic number *i* larger. If no harmonics are detected in this range, we say that the i^{th} harmonic are missing in that frame, and we set its amplitude to -1 and do not take account into this value in averaging.

After the harmonic structures of all the frames are extracted, we take the average of all the frames and calculate the standard deviation of each harmonic between frames. We stored them as two 4-by-9 matrix. One is for amplitude information, in which the first row is the average amplitude of the second to tenth harmonics and the second row is the corresponding standard deviation. And the other one is for harmonic number, with its structure similar to the first one. Note that we drop the first harmonic, since it must be $[100\ 1]^T$, we can simply discard it to save memory space.

In this project, we use two sources: the piano and the clarinet in the MIDI library. We choose these two instruments because they have very different physical structure. The instruments of the same category (strings, woodwinds, etc.) will have similar AHS models, which will complicate our work.

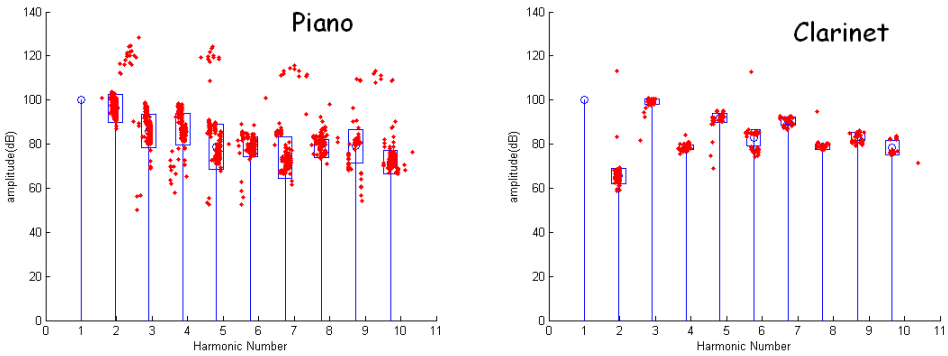

Figure 4: The AHS model of piano and clarinet: The blue stems are the AHS model, the rectangles are the region within one standard deviation, and the red dots are the peaks of each frame.

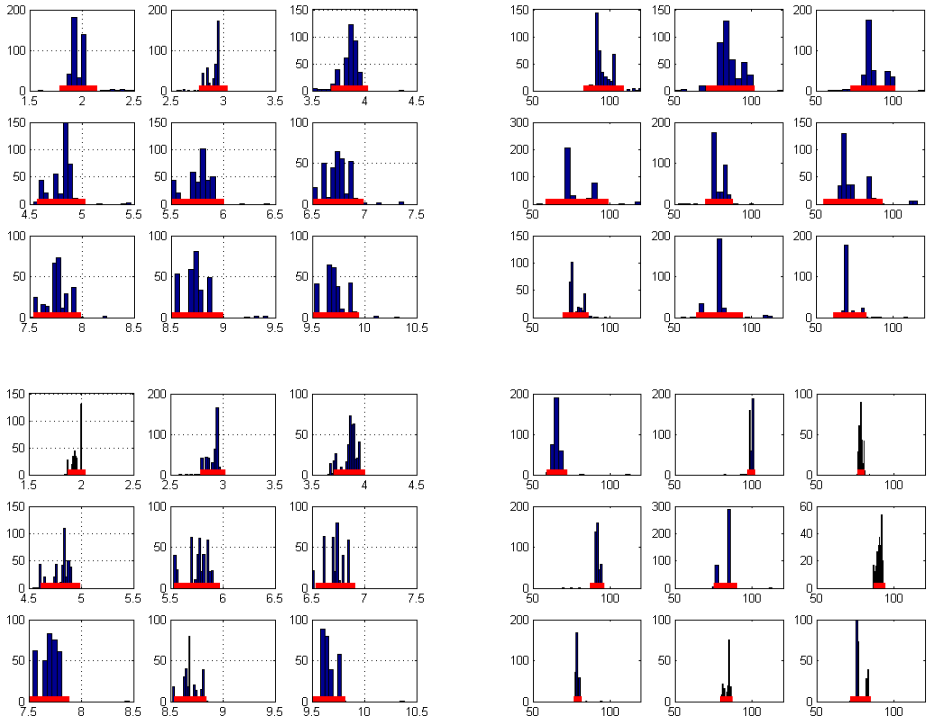

Figure 5: The AHS model distribution of piano (top) and clarinet (bottom): The left figure is the distribution in harmonic numbers. It is close to Gaussian distribution. The right figure shows the distribution in amplitude in dB. For piano, the distribution exhibits two peaks. For clarinet, the distribution is more concentrated.

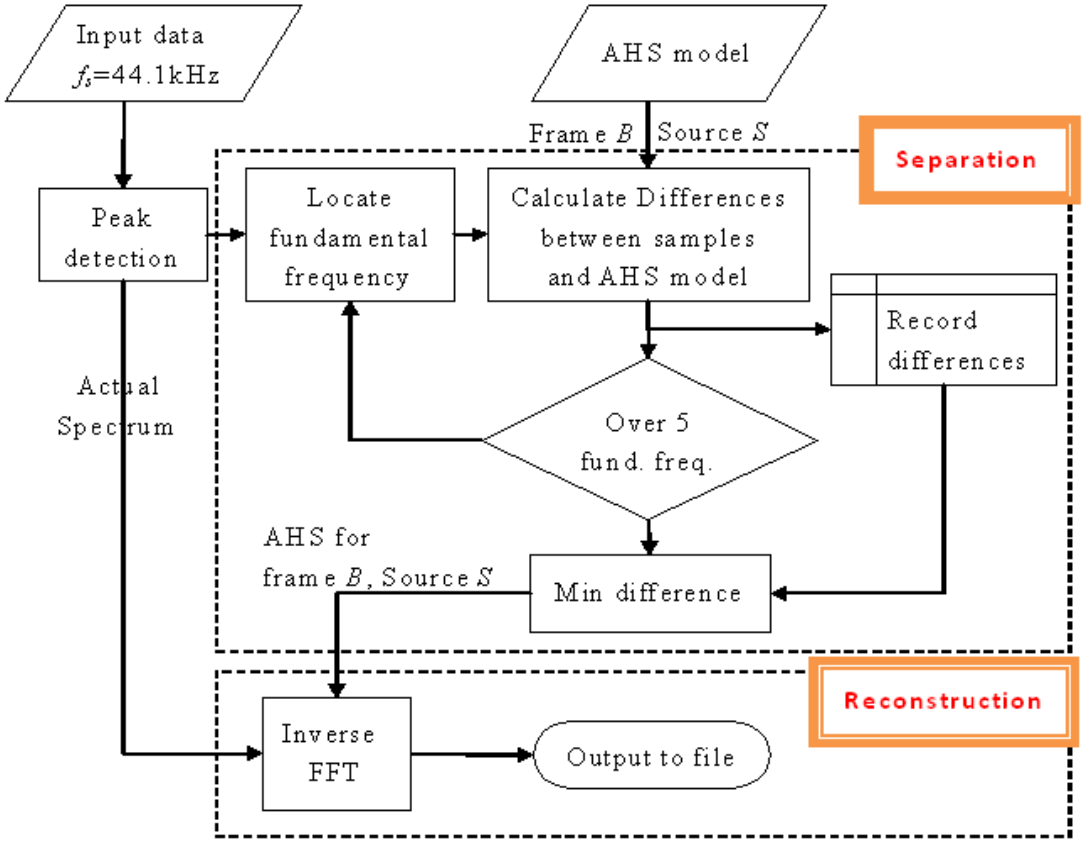

Figure 6: Block diagram of Source Separation

IV. Source Separation based on AHS

After we have the AHS model of the music sources in hand, we can start to separate the mixed sample by matching the harmonics in the sample with the AHS model. In this project, we use mixed music samples played by piano and clarinet (MIDI format). For each frame of the mixed sample, the peaks are detected by the method described in III – A. there may be two sounds playing concurrently, their peaks will interlace. So we make an assumption that the base frequencies will be one of the first five peaks, which is true in most cases.

First, we let the first peak be the base frequency, and calculate the ten harmonics by the relationship described in the piano AHS model. The standard deviation in the AHS model will form a rectangular area on the spectrum plane around the AHS model peak, which will be our search range. In this search range, we try to find the peak of the sample spectrum. Then we calculate the square error of amplitudes between the sample peak and the ideal AHS model peak. If no peak is found in the search range, we find the nearest point from the ideal AHS model peak to the spectrum curve, and store the square error of the two points. Adding up all these errors, we can know if this sound matches the AHS model of piano with the first peak as the base frequency.

We then let the second to fifth peaks be the base frequency one at a time, the one with the lowest difference with the AHS model is considered the best match, and stored as the error of piano for the current frame. All these works are repeated with the clarinet AHS model.

The error and the possibility of a certain source are playing negatively correlated. So we use exp(-(error)) to indicate if the source is playing or not. If this value is close to 1, then that source is probably playing aloud, the pitch is the base frequency of the best match mentioned above.

To reconstruct the two sources, we have to keep a certain source and filter out all the other frequencies. Take piano for example, for each frame, we keep only the peaks that match well with the AHS model, we keep the region within one standard deviation around the peaks (For base frequencies we keep 10% of frequency range before and after them). All the other parts are set to 0. Then we can use the inverse fast Fourier transform to reconstruct the source signal in the time domain.

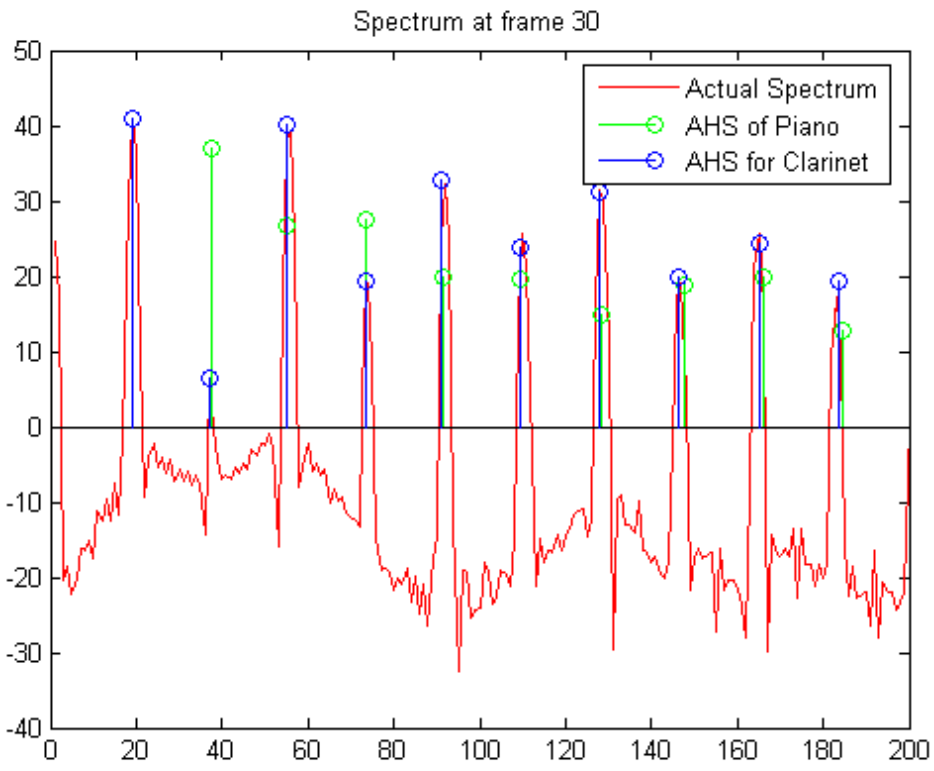

Figure 7: Peak matching illustration: The peaks from the mixed sample (red line) are compared with the AHS of Piano and Clarinet. In this case, the sample fits better with the Clarinet AHS model.

V. Experimental Results

A. Single Source in the AHS Learning Step

If we take the original single source used in the AHS learning step as the input for separation, theoretically the output will be exactly the same as the input. However, the results show that some of the notes failed to be recognized by the AHS model. In Fig.8 the first two notes (C4-D4) is missing, while the other notes are mostly the same as the input. This shows that the AHS of C4 and D4 are different from E4 to B4. Just as we said in II, the AHS model is only suitable for a limited pitch range. In order to get a better result, we may have to build AHS models for different pitch range.

Note that in Fig. 8, the plot have sharp peaks, while in Fig. 9, the plot will keep in high level for a long time. This is because the ADSR envelope [2] of piano and clarinet are different.

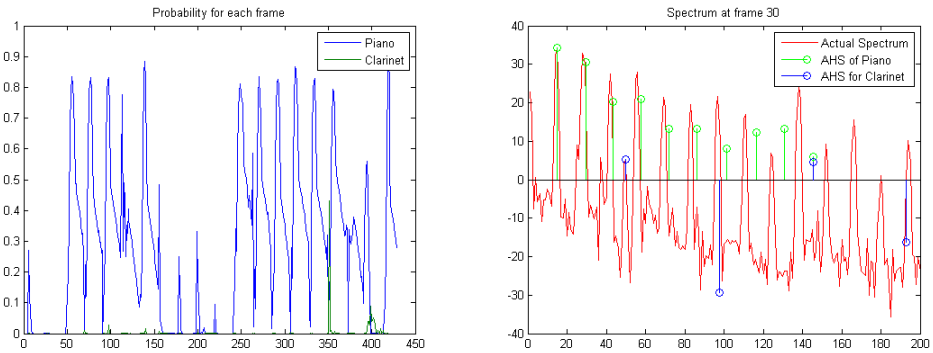

Figure 8: Take the training piano sample as input of the separation program. The left figure is the probability in each frame. The right figure is the spectrum at frame 30. The stem plot shows the matching process.

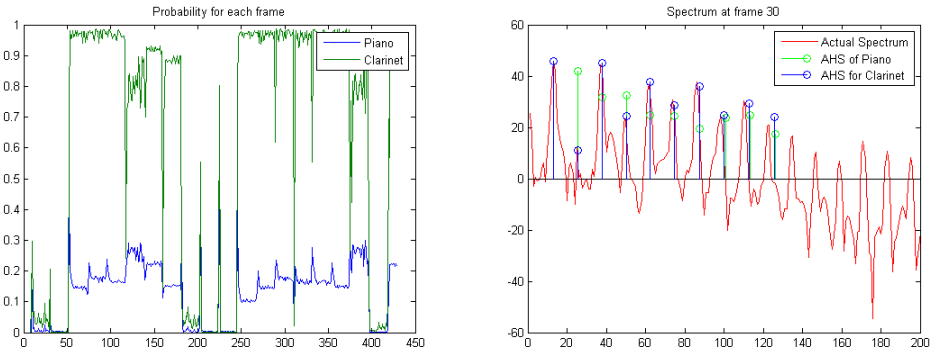

Figure 9: Take the training clarinet sample as input of the separation program. The left figure is the probability in each frame. The right figure is the spectrum at frame 30. The stem plot shows the matching process.

B. Mixed Music Sample

In this experiment, we use two kinds of samples: One is without superimposition while another is with superimposition.

1. No Superimposition

If the two sources don't play at the same time, the result is quite good. The clarinet is perfectly extracted. However, in the piano part, we can still hear the sound of the clarinet. This problem can be solved by another approach. We can generate the clarinet part, and then subtract it from the mixed sample; therefore only piano part will be kept.

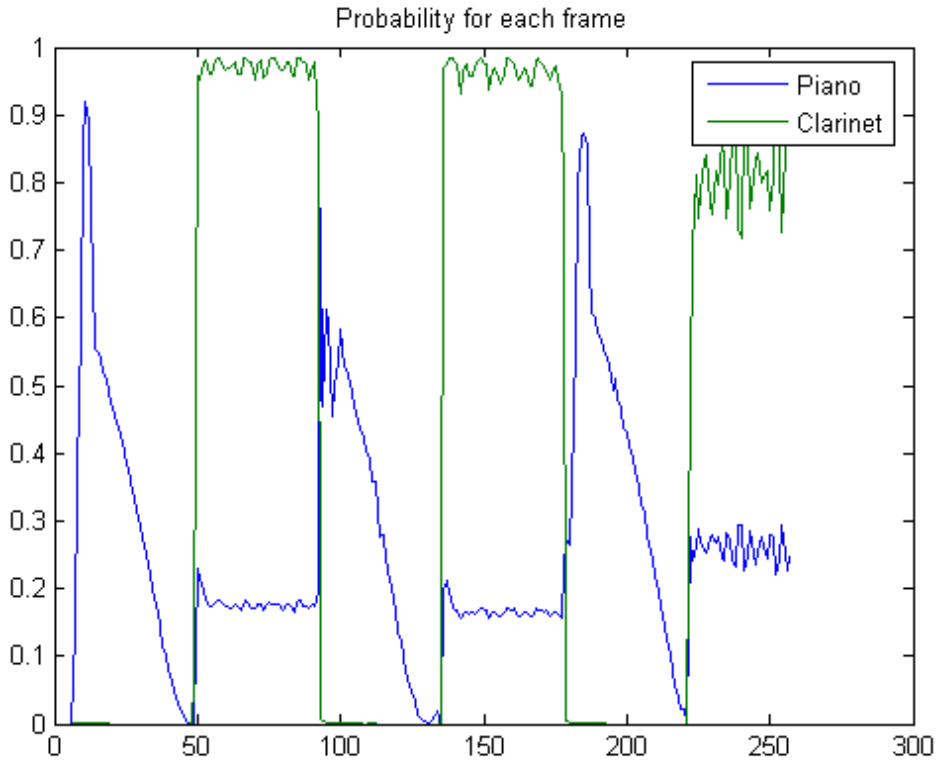

Figure 10: The probability for each frame. The piano and clarinet are alternately played. In the piano slot, the clarinet is rejected well. While in the clarinet part, the interference of piano takes place as a interference.

2. With Superimposition

Now the two sources can play at the same time, the clarinet is still working very well, but in the piano part we can still hear the clarinet voice. This is because the piano sound has a very short attack-decay-sustain envelop, and a rather long release part. In the release part, the voice is much lower than the clarinet sound. The AHS model of piano is not perfectly orthogonal to the clarinet part, so in the release period, some clarinet sound may be recognized as piano sound.

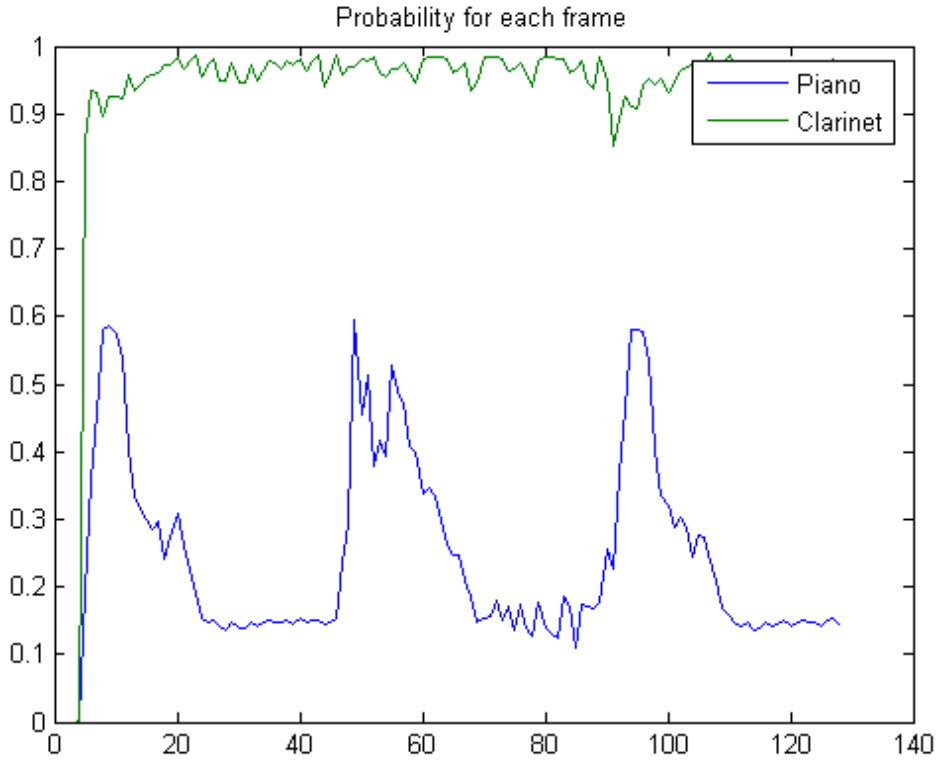

Figure 11: The piano and clarinet are played at the same time. The clarinet occurs at very high probability while the piano is more likely played at the attack region. The interference of piano is quite high.

VI. Conclusion and Discussion

This report adapts the AHS concept in [1] and do source separation. The model-learning process from a reference sample is similar to give lectures to computers about the nature of different instruments. The AHS model is not exact, but we can perform kind of primary separation. The experimental results show the quality.

In the process of doing this project, we know that the probability model is not exact. Take piano for an example, the amplitude exhibits two-peak distribution due to the attack and sustain period. It is not suitable to model it as a Gaussian distribution by estimated mean and variance. But for simplicity, we model it as Gaussian and the performances are not bad. Another observation is that the AHS model is a function of fundamental frequency. Only when it lies in a small pitch, we can approximate it as constant. If we want to improve the probability of correctness, we have to adapt a more complicated model to perform this task.

Finally, thanks to Professor Chen, who gave us inspiration about multimedia. Thanks to the TAs, who discussed the current progress about this topic. Thanks to NTU Philharmonic Club, both of us are in this club. Classical music provided us the

pleasure in life. Thanks to 劉俊麟, who worked mainly on the algorithm and MATLAB code. Thanks to 呂行, who work hard on the report and the presentation. Thanks to everyone who gave us support!

VII. References

[1] Zhiyao Duan; Yungang Zhang; Changshui Zhang; Zhenwei Shi, "Unsupervised Single-Channel Music Source Separation by Average Harmonic Structure Modeling," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.16, no.4, pp.766-778, May 2008

[2]http://en.wikipedia.org/wiki/ADSR_envelope