<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>An innovative approach to transcription of polyphonic signals (Accepted version)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Lee, Edwin Wei Thai; Foo, Say Wei</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>2001</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10220/4613">http://hdl.handle.net/10220/4613</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>© IEEE. <a href="http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=5224">http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=5224</a></td>
</tr>
</tbody>
</table>
AN INNOVATIVE APPROACH TO TRANSCRIPTION OF POLYPHONIC SIGNALS

Edwin Lee, Say Wei Foo

Department of Electrical and Computer Engineering
National University Of Singapore
10 Kent Ridge Crescent
Singapore 117576

ABSTRACT

In this paper, a new method is proposed to separate and identify polyphonic signals generated by an acoustic piano. The method consists of a top down extraction architecture using a harmonics-tracking algorithm followed by a bottom up reconstruction architecture based on a set of the expected profiles of single notes. In addition to the architectures mentioned above, a heuristic function is also included to further re-enforce the robustness and the ability to recognize multiple notes and to pull apart octave unison. The algorithm was tested with chords that consist of multiple notes. Some of the notes are one octave apart. It is shown that a recognition accuracy of 96% can be achieved using the proposed method.

1. Introduction

Some of the applications of musical notes recognition include providing for enthusiast pianists who would like to play a particular song for which the score for reference is not available. Other applications include helping pianists spot mistakes in a song and helping composers with the composition of their works.

There were previous attempts to build computer systems that are able to process and recognize musical notes, extract features from acoustic signals and use them in practical applications. One such algorithm that is very robust in the detection of monophonic signal is the use of the Correlogram [1]. There have also been attempts to transcribe polyphonic music. Some of the published methods include the blackboard architecture [2], the use of Neural Networks [3] and heuristic signal processing approach [4]. However, there are not many methods that are able to transcribe polyphonic music reliably. The difficulty lies in the detection of multiple notes played at the same time, especially when the harmonics of the notes are superimposed.

As in the requirements for MPEG 7 coding scheme, it is of much desire to gather descriptions of polyphonic signals. The algorithm presented in this paper aims to separate and identify multiple notes that are played at the same time with special attention to octaves unison.

The method presented in the paper arises from the analogy of the situation in a piano lesson. During a piano lesson where the student learns to play by hearing, a set of notes will be played by the piano instructor. The student will then mentally try to break down the notes based on what he has heard. This accounts for a top down extraction architecture using a harmonics-tracking algorithm. The student then tries his hands on the piano to reaffirm what he perceives the notes to be. This accounts for a bottom up reconstruction architecture based on a set of the expected profiles of single notes.

2. The Equal Tempered Scale

The relationship between a tone and the adjacent one is governed by Equation 2.1 for a semitone resolution.

\[ f_{n+1} = 2^{(1/12)} f_n \]  

(2.1)  

\( f_n \) and \( f_{n+1} \) are the fundamental frequencies of the \( n^{th} \) and the \( (n+1)^{th} \) note respectively. The factor 2\(^{(1/12)}\) is the interval of a semitone separation. This spacing corresponds to a 0.059 increment or a 6% separation between adjacent notes.

For a quartertone resolution, the equation becomes:

\[ f_{n+1} = 2^{(1/24)} f_n \]  

(2.2)  

The fundamental frequency correspond to F#2 is 92.5 Hz and the frequencies of others notes are obtained from this fundamental frequency based on Equation 2.1 for a semitone resolution and Equation 2.2 for a quartertone resolution. The musical signals used to test the algorithm are recorded in a wave format sampled at 22050 kHz using an eight-bit Pulse Code Modulation method.

3. THE CONSTANT Q TRANSFORM

For the analysis of musical signals, a constant Q transform [5] gives a better representation of the spectral data as compared to the fast Fourier Transform. The constant Q transform may be efficiently determined using Equation 3.2 and Equation 3.3.
\[ X^{cq}[k_{cq}] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] K[k,k_{cq}] \]  

\( X^{cq}[k_{cq}] \) is the \( k_{cq} \)th component of the Q transform and \( X[k] \) is the \( k^{th} \) spectral component of the \( n^{th} \) sample of the digitized temporal function \( x[n] \).

\[ K[k,k_{cq}] = \sum_{n=0}^{N-1} w[n] \left( -\frac{N(k_{cq})}{2}, k_{cq} \right) e^{jnk_{cq}(n-N/2)} e^{-j2\pi kn/N} \]  

\( K[k, k_{cq}] \) is the spectral kernels of the transformation and \( w[n, k_{cq}] \) is a Hamming window that is symmetric about the center of the interval. \( N \) is equal to the data frame size used. The index \( k \) corresponds to the indexing of the data vector, while \( k_{cq} \) indexes the frequency components in the transform. The derivation of the above equations can be found in Reference [6].

In the Constant Q transform, the window size varies according to Equation 3.4:

\[ N[k] = \frac{SR \cdot Q}{f_s} \]  

\( N[k] \) is the size of the variable window. \( SR \) represents the sampling rate and \( Q \) is a constant of value 17 for semitone resolution. The maximum window size can be obtained from Equation 3.5:

\[ N_{\text{max}} = \frac{SR \cdot Q}{f_{\text{min}}} \]  

The number of points to transform, or what is termed as the frame size, limits the window size. The maximum window size must be fixed in such a way that it must not exceed the frame size. The minimum frequency of analysis, \( f_{\text{min}} \), determines the maximum window size. A semitone resolution is chosen as this increases the range of notes which can be analyzed. The minimum frequency is taken to be F#2 which is 92.5 Hz, when a maximum window size of 4052 samples is selected. Using a radix 2 FFT, a frame size of 4096 is chosen and this corresponds to a time interval of 0.18576 sec, which is a reasonable time interval for most instances of notes.

When a semitone resolution is used, it becomes a problem when resolving notes that are one semitone apart. To resolve notes that are a semitone apart, a variable resolution of about 3% of the frequency is needed. This corresponds to a quartertone resolution. However, in most common piano pieces, chords that consist of notes that are one semitone apart are seldom played. As such, a compromise of a semitone resolution is used.

For the purpose of detection and separation of octaves unison, quartertone resolution is used. For octaves unison, the fundamental frequency of the higher note is twice that of the fundamental frequency of the lower note. This is illustrated in Equation 3.6:

\[ h_{11} = 2 \cdot f_{10} = f_{20} \]  

\( h_{11} \) is the frequency of the first harmonic of Note 1. \( f_{10} \) is the fundamental frequency of Note 1 and \( f_{20} \) is the fundamental frequency of Note 2 which is one octave higher than Note 1.

For the purpose of detecting octave, the minimum frequency \( f_{\text{min}} \) of interest from Equation 3.5 is doubled, thus the value of \( Q \) can also be doubled without affecting the maximum window size. The resolution is increased as a result to give a better recognition performance.

4. Characteristics Of Musical Signal

A study of musical signals in the frequency domain enables the characteristics of the fundamental and its harmonics to be analyzed. Vital information can be obtained from the spectral envelope of the signal. Musical signals are highly periodic and regular if equal amount of pressure is applied to the keys of the piano.

To study the pattern of the spectral envelope, a single note, C4, is played a few times at a time interval of about four seconds. The pressure applied to the keys is kept as consistency as possible. The spectral is generated and plotted on a constant Q transformed domain. A quartertone resolution is used. Figure 4.1, 4.2 and 4.3 shows the spectral of the notes played at different instances. All of them represent the same notes but they were played at different instances.

The corresponding \( k_{cq} \) values that represent the fundamental are 10 and 11. For quartertone resolution, two \( k_{cq} \) values are used to denote one tone. Each \( k_{cq} \) value corresponds to a particular frequency. A \( k_{cq} \) value of zero corresponds to a frequency of 195.997 Hz, while a \( k_{cq} \) value of 106 correspond to 4186 Hz. The corresponding frequency is not linearly proportion to the \( k_{cq} \) values but is calculated based on Equation 2.2.

![Figure 4.1: Spectral of a C4 signal](image-url)
Comparing Figures 4.1, 4.2 and 4.3, the relative amplitudes of the spectral are close with respect to the different spectra of the same note taken at different time instances.

However, the spectral envelope may differ slightly even for the same note played at different time instances. Such cases arise mostly when the pianist varies the way he or she depresses the key. An example is given in Figures 4.4 and 4.5. In this case the pressure applied to the keys is different. The first onset, which is shown in Figure 4.4, corresponds to a key that is depressed with high pressure. The second onset, which is shown in Figure 4.5, corresponds to a key that is depressed with low pressure but with sustain. The difference in the pattern of the spectral at different instances is clearly demonstrated in these two figures.

From Figure 4.4, it can be seen that the fundamental has lower amplitude than its first harmonic. However in Figure 4.5, the first harmonic has lower amplitude than the fundamental. Also the higher harmonics of the note in Figure 4.5 are more prominent.

The pattern of the spectral envelope varies at different instances when different pressures are applied. The relative amplitudes or the pattern of the spectral depend heavily on the pressure applied to the piano keys. Given such a case, the expected ratios of the spectral are taken into considerations. It is found that the expected ratios give better results than using the absolute value of the spectral profiles for notes that are played with sustain. This is so as when sustain is added when notes are played, the higher harmonics tends to be more prominent.

Four ratios are defined and calculated based on Equation 4.1.

\[ r_{ij} = \frac{f_i}{f_j} \quad (4.1) \]

\( f_i \) and \( f_j \) are amplitudes of the \( i \)th and the \( j \)th harmonic respectively. The fundamental is denoted by \( f_0 \). \( r_{ij} \) is the ratio of the amplitude of the \( i \)th harmonic to the amplitude of the \( j \)th harmonic. \( i = 0,1,2,3 \) and \( j = i+1 \).

To obtain a statistical average of the ratio, Equation 4.2 is used.

\[ \text{Ave}[r_{\text{ave}ij}] = \frac{1}{N} \sum_{k=0}^{N-1} r_{kij} \quad (4.2) \]

\( N \) is the number of samples of identical notes played at different instances.

An average of a large number of samples of the same note is taken for every individual notes. This average is stored in a database and is used for reference when the expected profile of a signal note is needed to determine if that note does indeed exist. This set of ratios of the expected spectral profiles forms the basis for a bottom up reconstruction architecture.

### 5. Top Down Extraction Architecture

The top down extraction architecture consists of a harmonics-tracking algorithm. The purpose of this part of the algorithm is to track the frequencies of the fundamental and its harmonics. The knowledge of the positions of the harmonics relative to the fundamental in a Q-transformed domain is exploited. The first peak of the spectrum is taken as the fundamental of the first note of the chord and its first four harmonics are tracked using Equation 5.1 for semitone resolution. Position of the \( N \)th harmonic in the Q domain is

\[ K_{cq} = 12 \log \left( \frac{(N + 1) f_i}{f_{\text{Lowest}}} \right) \quad (5.1) \]

\( f_i \) is the fundamental frequency of the note and \( f_{\text{Lowest}} \) is the lowest possible frequency of analysis based on Equation 3.5.
The ratios of the harmonics and its adjacent ones are taken as expressed in Equation 4.1. The note is provisionally accepted as a possible note based on the relative position of the harmonics and other possible notes. The next spectral component is then tracked. If the spectral component appears as an octave of the provisional notes, higher harmonics are then examined to justify whether that spectral corresponds to a note. If indeed an octave is played, the fundamental of that octave played must have substantial amplitude and at least one of its harmonics must be prominent.

The frame is then advanced by 512 samples and its harmonics are again tracked and stored. An advancement of 512 samples corresponds to a time interval of 0.02 seconds. The reason of taking the same onset for a specific number of frames at an interval of 0.02 seconds is to obtain a statistical average of the presence of the notes.

This process is iterated for a number of frames. The number of frames to be taken depends on the time interval between the current onset and the next one. After a specific set of frames has been taken, a set of provisional notes in the chord would have been successfully identified. Further processing is required to determine if those notes truly exist.

Before the next onset is analyzed, a frame of samples just before the next onset is taken into consideration to prevent the interference from the decaying signals of the current onset to the signals from the next onset.

### 6. Bottom Up Reconstruction Architecture

The bottom up reconstruction architecture is based on a set of the expected profiles of single notes. The expected profiles of all the individual notes were obtained based on Equation 4.2 and stored in a database prior to running the algorithm on the test signal.

From the set of provisional notes obtained from the top down extraction architecture, the notes are then reconstructed based on the ratios of the expected spectral profiles stored in the database. The amplitude of the lowest note obtained from the top down extraction architecture is used as the basis for the first note to be reconstructed. The amplitude of the first harmonic is obtained from the ratios in the database and the subsequence amplitudes of the harmonics are obtained in a similar manner.

Since the constant Q transform is a linear transformation, the summation of the amplitudes of the expected spectral profiles is compared with the spectral profile of the combined set of notes played at the same time from the test signal. During the top down extraction process, if an additional false note is detected, the reconstructed spectral profile from the bottom up reconstruction process will consist of additional harmonics. This indicates the presence of a false note and the false note should be discarded.

This false note is sieved out based on the relative positions of its fundamental and its harmonics on a constant Q frequency domain. The remaining spectral should be a closer match to the test spectral.

Although the constant Q transform is a linear transformation, the signals in the time domain may not be of the same phase. In the worst-case situation, two different sets of notes with overlapping harmonics may be anti-phase with each other in the time domain. This results in the subtraction instead of addition of the harmonics. The bottom up reconstruction architecture takes that into account. When reconstructing the overall profile of the chord, the harmonics of the spectral profiles of the individual notes are either added or subtracted to obtain the various spectral profiles of the combined notes. The optimum profile is then chosen from all the possible sets of profile reconstructed.

Given the set of possible notes from the top down extraction architecture, the spectral are reconstructed. This algorithm reconstructs the optimal spectral profile, which has the closest resemblance to the test signals. That set of notes that gives the closest resemblance to the spectral of the test signal will be taken as the actual notes that are played.

The advantage of having the bottom up reconstruction architecture is the ability to determine the relative strength of every individual note played. In constructing the optimum match for spectral profile of the test sample, the spectral profile of the individual notes are assigned a certain energy level based on the amplitude of the fundamental. By knowing the energy level of the individual notes, the relative strength of all the notes played can be estimated.

### 7. Heuristic Function

The next level of processing includes a heuristic function, which calculates the probability of an octave being played given the set of detected notes. It looks at the higher harmonics when an octave is deemed to have exists and calculates the probability of that occurrence based on those harmonics. For the purpose of better recognition for octave unison, a quartetone resolution is used.

The purpose of the heuristic function is to increase the robustness of the algorithm to octave unison. All notes that were detected previously are analyzed again for the possibility of the presence of an octave.

It uses a quartetone resolution to examine the higher harmonics of the octave. A quartetone resolution is used without the compromise of a larger window size as the minimum frequency of interest is doubled. This is so as an octave, which occurs at the first harmonic, has a fundamental frequency that is twice that of the lower note.
As the minimum frequency of interest is doubled, the value of $Q$ from Equation 3.5 can also be doubled without increasing the maximum window size for the worst-case scenario.

The number of notes, the notes detected and the energy level of combined set of notes are also taken into account in the calculation of this heuristic function. When any of these features meets a predefined criterion, the probability is set to increase by 0.1.

If the total sum of the probability of the heuristic function is above a certain value, taken to be 0.8 in this case, the octave note is accepted as a true note, else it is discarded.

8. Results And Discussions

The algorithm was tested with notes from an acoustic piano. Before any piece of music is played, a progression of notes is generated from the piano. This is for the purpose of constructing the expected spectral profile of every single note. Many samples of each individual note are taken to give a good statistical average of the ratios of the spectral profile.

After which, a set of 24 chords of multiple notes were put to the test. These sets of multiple notes consist of 21 chords of 4 notes unison and 3 chords of 3 notes unison. Out of these 24 chords, 15 chords consist of one octave unison. Out of a total 93 notes, 90 notes were detected correctly. There were only three missing notes. This corresponds to a test result of 96.7% accuracy. Two sets of chords that consist of six notes were then put to the test. Out of these two sets of six notes unison there were only two missing notes, one from each set.

9. Conclusion

The method proposed consists of a top down extraction architecture using a harmonics-tracking algorithm followed by a bottom up reconstruction architecture based on a set of the expected profiles of single notes in additional to a heuristic function.

The algorithm is tested using a set of 24 sets of chords that consists of 3 to 4 notes. An accuracy of 97.6 % was achieved. The notes that the algorithm is able to detect, range from F#2 to C7. Within this range, there are altogether 55 notes. This algorithm is able to separate chords that consist of up to 4 notes with much reliability.

The proposed method is able to detect most of the notes even when chords that consist of 6 notes are played.

REFERENCES


