TRANS-SYNTHESIS SYSTEM FOR POLYPHONIC MUSICAL RECORDINGS OF BOWED-STRING INSTRUMENTS

Tien-Ming Wang, Wei-Chen Chang, Kuan-Ting Lin,

Wen-Sen Su, and Alvin W.Y. Su

SCREAM Lab,
CSIE., National Cheng-Kung University
Tainan, Taiwan
showmin@csie.ncku.edu.tw

ABSTRACT

A system that tries to analyze polyphonic musical recordings of bowed-string instruments, extract synthesis parameters of individual instrument and then re-synthesize is proposed. In the analysis part, multiple F0s estimation and partials tracking are performed based on modified WGCDV (weighted greatest common divisor and vote) method and high-order HMM. Then, dynamic time warping algorithm is employed to align the above results with the score to improve the accuracy of the extracted parameters. In the re-synthesis part, simple additive synthesis is employed. Here, one can experiment on changing timbres, pitches and so on or adding vibrato or other effects on the same piece of music.

1. INTRODUCTION

In [1], an analysis and trans-synthesis system for solo Erhu music has been reported. In this system, solo Erhu musical recording can be analyzed and re-synthesized using timbres such as violin, trumpet, oboe and so on. To achieve above works, a WGCDV pitch tracking method called [2] was proposed and simple partial tracking was applied based on the pitch tracking results [1]. However, the methods mentioned in [1] and [2] cannot be applied directly to polyphonic recordings and their trans-synthesis system has limited of applications since even violin solo recordings are considered actually polyphonic. For example, a player may bow two strings at the same time while the other strings may still vibrate. Then, there are four sources at the same time.

For polyphonic musical recordings, F0 estimation of multiple sources is necessary. Yeh [3] proposed a scoring function to rate the plausibilities of all possible combinations of F0 hypotheses for the sound sources considered as quasi-harmonic, stationary and non-reverberated signals based on three physical principles: harmonicity, spectral smoothness and synchronous amplitude evolution with respect to a single source. In [4], Wen developed a partial searching algorithm based on 1st-order frequency prediction using a revised dynamic programming method. However, the first difficulty may be the identification of the number of existing sources. Furthermore, the missing partial candidates may procure the incompleteness of the pitch candidate trajectories. In [5], Chang et. al. use high order HMM based on the results produced by the method reported in [3]. In their paper, the scoring machine which involves special processing for close or shared partials, is coupled with a tree searching method for polyphonic transcription task. Excellent results have been reported in [6].

In this paper, a high-order HMM for partial tracking of different F0s modified from the method in [5] is proposed. Hence, one can obtain more information about the partials of individual source in consecutive frames. Though reasonable results can be achieved, errors do occur. Moreover, accurate onset is important musically. The symbolic score-matching algorithm proposed in [7] is employed and the alignment approach [8] of polyphonic musical recording to symbolic score information by using chroma representation is applied to give onset information. The alignment results are also used to correct possible estimation errors produced in the previous stage. Onset timing for each source is further calibrated based on the partial tracking results, too. The proposed analysis system is now tested on polyphonic musical recordings of bowed-string instruments though it is possible to apply it for recordings using other kinds of musical instruments.

Finally, additive synthesis [9] employed in [1] is again used in this paper with acceptable synthesis sound quality. In the system, manual control of each partial is equipped. Other features include changing the timbres and extra sound effects such as vibrato.

The overall trans-synthesis system flow diagram is shown in Figure. 1. In the preprocessing stage, a simple multiple pitch detection is used. Then, the proposed partial tracking method starts by using the previously detected F0s. Score alignment and onset detection are performed with respect to the pitch and partial information. Finally, evaluation of energy of each individual source for each audio frame is done and additive synthesis is applied to re-synthesize the same piece of music using the desired timbre and effects.

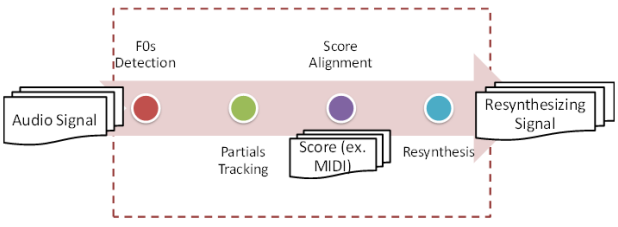


Figure 1: *the overall trans-synthesis system*

The rest of this paper is organized as follows. The proposed partial tracking is presented in section 2. Polyphonic score alignment is described in section 3. The technique to evaluate energy of each individual source is presented in section 4. Experimental results are shown in section 5. Conclusions are given in section 6.

2. MULTIPLE PITCHES AND PARTIALS TRACKING

FFT (Fast Fourier Transform) is applied to the observed signal to obtain the instantaneous spectrum. Then, a simple frame-based multiple-F0 estimation algorithm modified from [2] is applied. Based on sinusoidal model [9], the observed music signal is modeled as a sum of several harmonic partials and noise. In [2], the spectral peak with the highest score produced by WGCDV method is selected as the pitch. In this paper, we simply select the spectral peaks with the scores larger than a certain threshold as the possible pitch candidates. Those candidates which perform poorly in partial tracking stage will be eliminated. The overall detail can be found in [10].

In addition to the pitch information, the partial information is necessary when calibrating the onset timing and calculating the amplitude parameter required in additive synthesis algorithm. Furthermore, it is also useful when analyzing the timbre of a source in the recording.

Given the power spectrum and the estimate of F0s, the search for partials can be associated with F0s under the assumptions weaker than the perfect harmonic model. To associate partial candidates among several consecutive audio frames into corresponding trajectory, the tracking can be considered as a higher-level model based on a probabilistic framework. The note event model used in [11] which describes the temporal evolution of a single note as a sequence of states changing from frame to frame is represented as a high order hidden Markov model (HMM). The states of a node model are usually modeled as attack, decay, sustain, release and silence (ADSR). In our case, a simplified version of the node model which just contains attack state and sustain state, shown in Figure 2(a).

In Figure 2(b), only two possible transitions between the nodes are allowed: "attack to sustain" and "sustain to sustain". The corresponding transition probabilities are assumed to be identical though this may not be very suitable for bowed-string instruments. The connection resulting in the maximal propagation weight is considered the most likely path and is stored as a pointer to the "winning node". After the forward tracking stage, the propagation weights and the related back pointers of all the nodes are recorded. The second part is iterative backward tracking of the candidate trajectories. All the back pointers obtained in the forward tracking result in tree structure within which the paths from the "roots" to the "leaves" are conceivable trajectories. At this stage, it is supposed to iteratively extract the candidate partial trajectories by finding the most likely paths from the leaves to the roots. In each analysis frame, it is reasonable to search in the order of the propagation weights because a large weight is evidence of high probability.

For each partial group, the tracking of partial candidates can thus be understood as decoding process of multiple optimal paths in a three-dimensional trellis structure shown in Figure 3. This is very similar to the work done in [5], except that a partial group of a source is considered rather than just the respective pitch. In Figure 3, each node represents the hidden state of a note, either attack or sustain. The solid lines connect the candidate *C* in frame-t to the possible candidates in frame-(t-1), whereas the dash lines connect the candidate c in the tth frame to the possible candidates in frame-(t-2). Hence, Figure 3 represents a 2nd order HMM. In this paper, higher order is employed because higher partials usually have lower energy than the fundamental and the trajectories may be broken easily if the lower order HMM is used.

Moreover, there are more than one possible best path to be decoded, Since the number of paths is unknown, the traditional Viterbi algorithm is not suitable for this case. Here, the tracking algorithm is divided into two parts. The first one is the evaluation of forward propagation of the connection weights. The probability of each trajectory is associated with the weights propagated from node to node within the related path. The observation probability emitted by the sustain state is defined as a Gaussian distribution and calculated as

$$\psi(\Delta f, \Delta m) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{\left(-\frac{\Delta f^2}{2\sigma_1^2}\right)+\left(-\frac{\Delta m^2}{2\sigma_2^2}\right)\right\}, \qquad (1)$$

where Δf and Δm are the frequency difference and the magnitude difference between two concerned partial candidates, respectively. It is noted that the frequency of a possible partial is first converted to tone scale and its magnitude is converted to dB scale. In this paper, σ_1 is set to 0.25, which corresponds to one quarter tone and σ_2 is set to 1.0, which corresponds to 1dB.

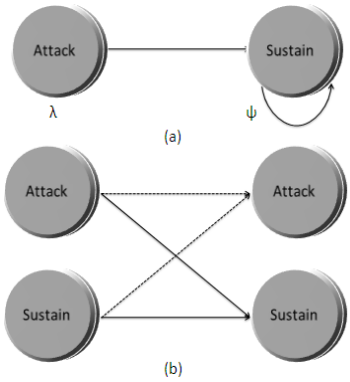


Figure 2: *HMM note model with the attack probability λ and the sustain probability ψ: (a) graphical representation; (b) state transition weight matrix.*

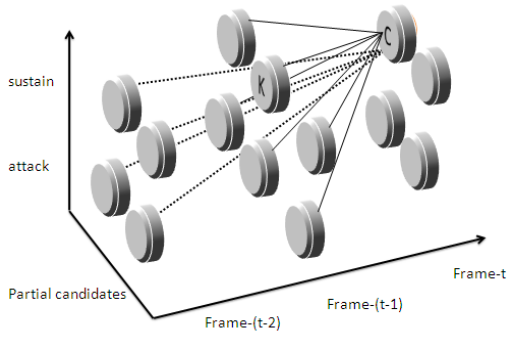


Figure 3: *trellis structure of the 2nd order HMM model.*

3. SCORE ALIGNMENT AND RE-SYNTHESIS

After extracting the trajectories of F0s and partials, errors such as detection miss of octave and/or quint pitches may still occur. Therefore, these errors can be corrected if score information is available. For most pieces of classical or popular music, there are printed scores, audio recordings, and often MIDI files created by manual transcription or score conversion. In this paper, we use MIDI files though other forms are always possible. Techniques used in music information retrieval systems can be applied. For example, the approach to perform some sorts of polyphonic transcription on the music and then use a symbolic score-matching algorithm [7, 12]. Nevertheless, accurate polyphonic transcription is yet to be achieved and errors may make the score-matching difficult in many cases. A representation of both score and audio data is the discrete chromagrams, which are the sequences of chroma vectors [8]. The chroma vector representation is a 12-element vector, where each element stands for the spectral energy corresponding to one pitch class such as C, C#, D, D#, etc. Let N-point FFT be used. $A_i(\bullet)$ is the magnitude spectrum of the *i-th* frame and $W(\bullet)$ is the window function of length $2L+1$. The *k-th* element of the chroma vector of the *i-th* frame is calculated as

$$s_{ik} = \sum_{p=0}^{P} \sum_{j=-L}^{L} A_i(2^p \cdot f_k + j) W(2^p \cdot f_k + j), \qquad (2)$$

where f_k is the index of the central frequency of the *k-th* tone for p, j such that $0 \le 2^p \cdot f_k + j < N/2$. Figure 8(a) and Figure 5(a) show spectrogram and chromagram of the original signal extracted from an acoustic recording of violin solo.

The score alignment procedure is shown in Figure 4. To transform audio data into chroma vectors, FFT is used and each frequency bin is assigned to the pitch class of the nearest step in the chromatic equal-tempered scale. Chroma vectors are not sensitive to spectral shape, yet they are sensitive to prominent pitches and chords. Since we are comparing MIDI data to the correspondent acoustic data, it is good to focus on pitch classes and more-or-less ignores details of timbre and spectral shape. The MIDI data is converted to audio data using a MIDI synthesizer, and then the rendered audio was converted to chromagrams.

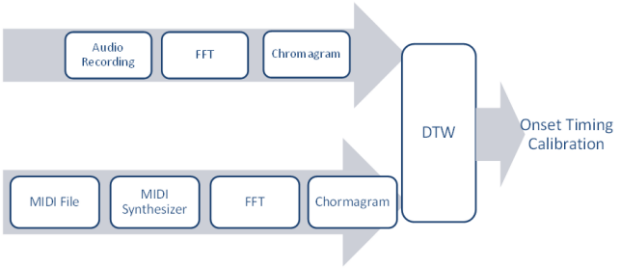


Figure 4: *score alignment procedure*

With two sequences of chroma vectors for both audio recordings and MIDI data, the correspondence between the two sequences is explored. The "agreement between vectors" means the Euclidean distance between two vectors which are normalized in zero mean and one variance. If there is perfect agreement, the distance is zero. Alignment is computed using a dynamic time warping (DTW) algorithm. DTW computes a path in a matrix where the row corresponds to one vector sequence and the column corresponds to the other. The path is a sequence of adjacent cells, and DTW finds the path with the smallest sum of distances. In comparison with traditional DTW, the Euclidean distance in this paper is calculated by using chroma vectors instead of pitch information. The distance measure used in the DTW algorithm is calculated as follows:

$$D(i,j) = \min \begin{pmatrix} D(i, j-1) \\ D(i-1, j-1) \\ D(i-1, j) \end{pmatrix} + dist(i,j), \qquad (3)$$

where

$$dist(i,j) = \sqrt{\sum_{k=1}^{12} (s_{ik} - t_{ik})^2}. \qquad (4)$$

The initial condition is set as: $D(i,1) = \infty, i \ge 2$ and $D(1,j) = dist(i,j)$, for all *i*. In (4), s_{ik} is the *k-th* element of the *i-th* input chroma vector, t_{jk} is the *k-th* element of the *j-th* reference chroma vector, and D(i, j) is the overall DTW distance.

After the DTW score alignment, the partial trajectories that do not coincide with the score are eliminated. Moreover, the onset timing obtained from the score alignment may not be very accurate. It is known that onset timing is very critical for music performance. Usually, the onset timing found in the score alignment part deviates from the correct onset by several audio frames. Therefore, it has to be calibrated by using the results obtained in the partial tracking part. Define the energy of the *i-th* frame as

$$E_i = \sqrt{\sum_{j=0}^{N/2-1} (A_i(j))^2} \qquad (5)$$

The following equation is used to decide whether the *i-th* frame is an onset frame for an individual source.

$$Onset(i) = \begin{cases} 1 & \text{,if } E_i / E_{i-1} > \alpha \text{ and } E_i > T \\ 0 & \text{,otherwise} \end{cases}, \qquad (6)$$

where α is set as 2 in our experiments and energy has to be greater than a certain threshold.

It is noted that the score alignment may fail when the score (the MIDI file) has certain degree of difference from the audio recording. In such cases, manual correction of the score file will be necessary.

After the information about F0, partial tracking and onset is obtained, the energy for each individual source has to be measured. We sum up the spectral energy of the corresponding partial group for each audio frame. Accurate partial tracking produce good results when calculating the individual source energy. In real-world case, human beings are not sensitive to loudness differences among partials [13]. If the energies and frequencies of

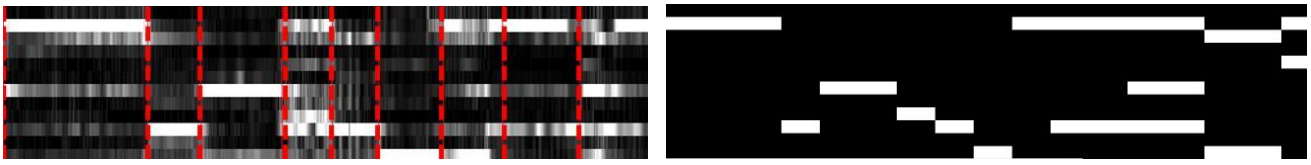
two notes are sufficiently close for a specific instrument, their corresponding partials are also close to each other. In [1], a parametric table is built for a Chinese traditional bowed-string instrument, Erhu, for some F0 frequencies and energies of 20 partials of each F0. The partial energies of non-specified F0s are then calculated using simple linear interpolation method from the existing information of the available F0s. One can re-synthesize the same piece of music with the identical style of playing. It is easy to change the timbre and the expression. Such a convenient tool was developed in [1]. A demo can be found in [14].

4. EXPERIMENTAL RESULTS

In this paper, two test signals and acoustic recording of Bach's BWV1001 violin solo are used. In order to have clear figures, only the results with frequency ranges under 2KHz are shown. The first test signal contains C4, E4 and G4 tones followed by D4, F4 and A4 tones. Figure 6(a) and Figure 6(b) show the spectrogram and its partial tracking result respectively. The second test signal contains C3 and C4 tones followed by F3 and F4 tones. Because of the perfect octave problem, the pitch and partial tracking does not perform well. Figure 7(b) shows the result without score alignment and Figure 7(c) shows the result with score alignment. Both figures show only the zoom-in parts of the whole results with the circle dotted lines representing the fundamentals and the solid lines representing the partials. As shown in Figure 7(a), only one tone is found. After score alignment, two tones separated by a perfect octave are located.

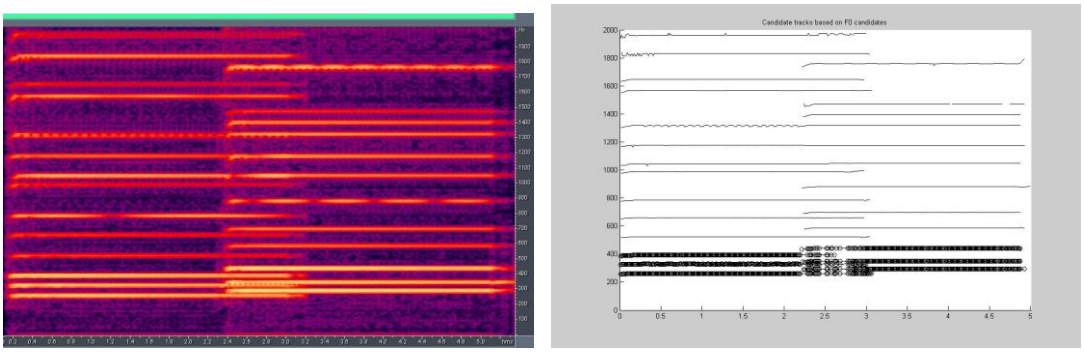
Finally, the partial tracking result for the Bach's BWV1001 is shown in Figure 8(b). Figure 8(c) shows the score alignment result by using Figure 5(b) as the correspondent midi score information. Finally, the re-synthesis sounds by using erhu and trumpet timbres are shown in Figure 8(d) and Figure 8(e).

The re-synthesis results sound similar to the original violin solo in the playing style. However, discontinuities occur occasionally where partial energy is too low. This problem may be reduced once better multiple pitch detection method such as [5] is used. Furthermore, one may notice that the violinist played larger vibrato than the re-synthesis. This is because the proposed tracking method tends to smooth the trajectory in order to avoid some troubles. We will try to solve this problem in the future.



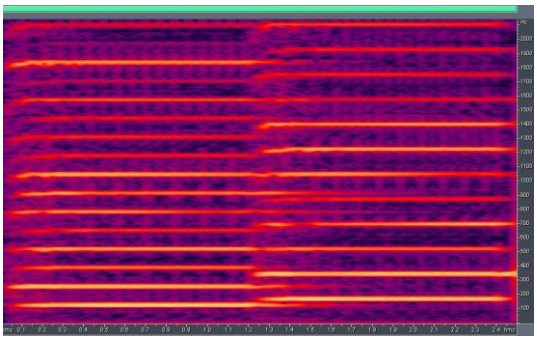
(a) (b)

Figure 5: *chromagrams obtained by using (a) acoustic recording of BWV1001; (b) MIDI file of BWV1001*

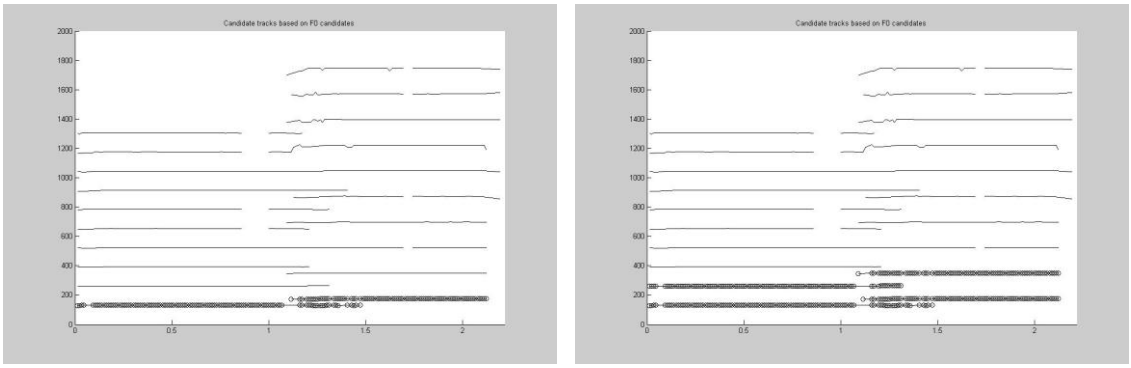


(a) (b)

Figure 6: *The first test signal: (a) spectrogram; (b) partial tracking result*

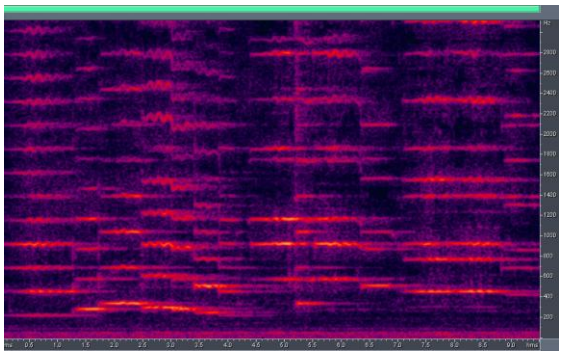


(a)

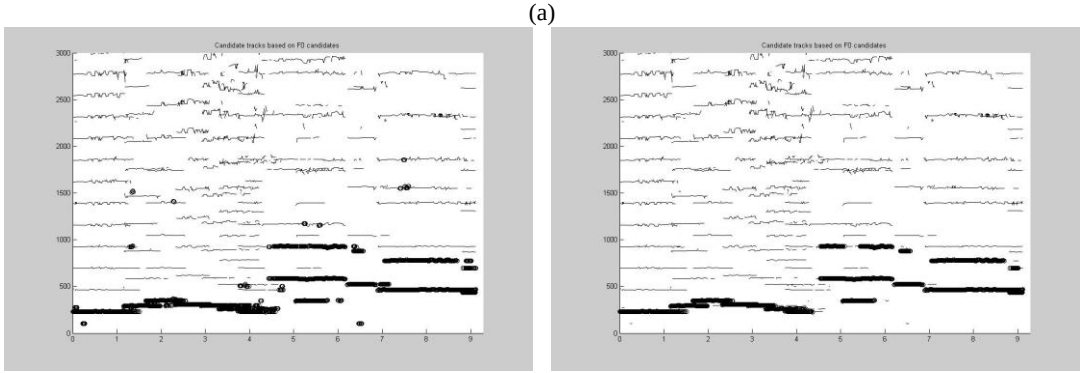


(b) (c)

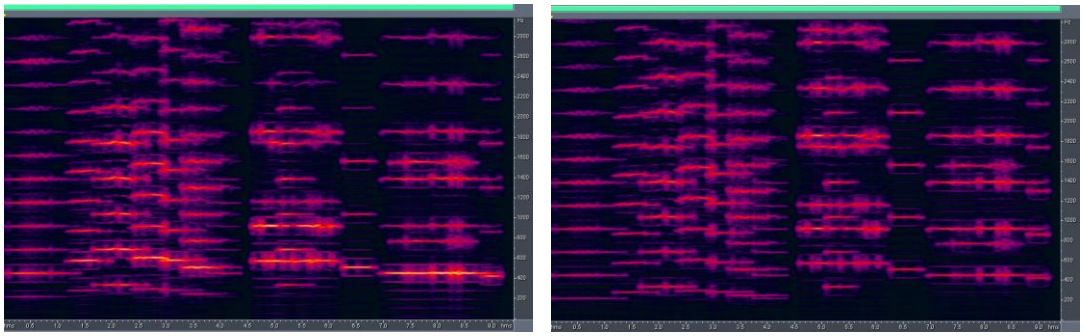
Figure 7: *The second test signal: (a) spectrogram; (b) partial tracking without score alignment; (c) partial tracking with score alignment.*



(a)



(b) (c)



(d) (e)

Figure 8: *Bach BWV1001: (a) spectrogram of acoustic recording; (b) partial tracking without score alignment; (c) partial tracking with score alignment; (d) spectrogram of re-synthesis with erhu timbre; (e) spectrogram of re-synthesis with trumpet timbre.*

5. CONCLUSIONS

A trans-synthesis for polyphonic musical recording of bowed-string instruments has been presented. We employed a simple frame-based multiple F0s estimation algorithm. A high-order HMM model is proposed to provide quality partial tracking. Then, score alignment is performed by using chroma feature transformation and DTW algorithm to give rough onset information of the audio recording based on the corresponding MIDI file. The partial information is refined such that it coincides with the score information. The onset timing is then calibrated. Finally, energy of each individual source is calculated to give the amplitude information of required in additive synthesis. Once the synthesis parameters are available, one can easily re-synthesize the same piece of music. For example, one can change the timbre, the expression, and so on. The results of the analysis part are showed in the paper. The re-synthesis sounds will be presented in the conference.

There are still some weaknesses of the system. First, the pitch estimation part may give incorrect results. This will affect the accuracy the partial tracking part even for such a simple violin solo recordings. For more complicate music such as string quartets, errors happen more frequently. Second, weak partials may be dropped in the current implementation and discontinuities can occur. The HMM of partial tracking can be improved. For example, the transition probabilities of “attack to sustain” and “sustain to sustain” should not be identical in the bowed-string cases. Third, the score alignment usually fails when lots of differences exist between the score file and the audio file. A smarter score alignment algorithm is desired. Finally, the re-synthesis sound quality may be improved. Though additive synthesis sound quite well in the sustain part, more realistic synthesis of the attack part is desired.

Acknowledgment

We would like to thank Dr. Roebel and Dr. Yeh of IRCAM for their valuable comments on this work.

6. REFERENCES

[1] W. C. Chang, Y. S. Siao, and A. W. Y. Su, "Analysis and transynthesis of solo Erhu recordings using additivie/subtractive synthesis," in *120th Audio Engineering Society (AES) Convention* Paris, France, 2006.

[2] Y. S. Siao, W. C. Chang, and A. W. Y. Su, "Pitch detection/tracking strategy for musical recordings of solo bowed-string and wind instruments," *Journal of Information Science and Engineering*, vol. 25, To be published July, 2009.

[3] C. Yeh, A. Robel, and X. Rodet, "Multiple fundamental frequency estimation of polyphonic music signals," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)*. vol. 3 Paris, France, 2005.

[4] X. Wen and M. Sandler, "A partial searching algorithm and its application for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* Tampere Univ. of Technol, Finland, 2005.

[5] W. C. Chang, A. W. Y. Su, C. Yeh, A. Roebel, and X. Rodet, "Multiple-F0 tracking based on a high-order HMM model," in *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08)* Espoo, Finland, 2008.

[6] C. Yeh, A. Roebel, and W. C. Chang, "Multiple F0 estimation for MIREX 2008," *The 4th Music Information Retrieval Evaluation eXchange (MIREX'08).* 2008.

[7] J. Bloch and R. Dannenberg, "Real-time accompaniment of polyphonic keyboard performance," in *International Computer Music Conference*, 1985, pp. 279-290.

[8] N. Hu, R. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* New Paltz, NY, 2003.

[9] X. Serra and J. Smith III, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, pp. 12-24, 1990.

[10] W. S. Su, "Pitch and Partial Tracking of Polyphonic Musical Signals," in *CSIE*. MS Thesis Tainan: NCKU, 2009.

[11] M. Ryynanen and A. Klapuri, "Polyphonic music transcription using note event modeling," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

[12] J. Pickens, J. Bello, G. Monti, M. Sandler, T. Crawford, M. Dovey, and D. Byrd, "Polyphonic score retrieval using polyphonic audio queries: A harmonic modeling approach," *Journal of New Music Research*, vol. 32, pp. 223-236, 2003.

[13] W. Thurlow, "Perception of low auditory pitch : a multicue, mediation theory," *Psychol Rev*, vol. 70, pp. 461-70, 1963.

[14] Y. S. Siao, "Demostration of Erhu trans-synthesis tool." http://www.youtube.com/watch?v=pKF1RWyc8HE