TELPC BASED RE-SYNTHESIS METHOD FOR ISOLATED NOTES OF POLYPHONIC INSTRUMENTAL MUSIC RECORDINGS

*Ya-Han Kuo, Wei-Chen Chang, Tien-Ming Wang, Alvin W.Y. Su**

SCREAM Lab., Department of CSIE,
National Cheng-Kung University
Tainan, Taiwan
alvinsu@mail.ncku.edu.tw

ABSTRACT

In this paper, we presented a flexible analysis/re-synthesis method for smoothly changing the properties of isolated notes in polyphonic instrumental music recordings. True Envelope Linear Predictive Coding (TELPC) method has been employed as the analysis/synthesis model in order to preserve the original timbre quality as much as possible due to its accurate spectral envelope estimation. We modified the conventional LPC analysis/synthesis processing by using pitch synchronous analysis frames to avoid the severe magnitude modulation problem. Smaller frames can thus be used to capture more local characteristics of the original signals to further improve the sound quality. In this framework, one can manipulate a sequence of isolated notes from two commercially available polyphonic instrumental music recordings and interesting re-synthesized results are achieved.

1. INTRODUCTION

It is sometimes that one wants to manipulate certain notes in a polyphonic music recording. This requires two main operations: extraction of target notes and their analysis/re-synthesis. Note or source extraction/separation for polyphonic signals is one major topic in music information retrieval (MIR) in recent years. Some of the separation techniques consider the entire audio mixture, avoiding any prior source information, such as Independent Subspace Analysis (ISA) [1][2] and Nonnegative Matrix Factorization (NMF) [3]. These methods decompose the amplitude spectrum of the mixture signal into basis spectra vectors in a statistical fashion. These basis vectors are then clustered into disjoint sets corresponding to the different sources. NMF is especially found to be able to efficiently decompose such mixture signals [4], [5]. In [6], a recursive regularization method for time varying analysis of music signals was proposed to track specified notes in polyphonic music recordings. To discuss blind source separation from one channel music signal as a whole is beyond the scope of a brief paper. However, it will become simpler if we have the score information and focus on a single source/note that we are interested in. Most techniques mentioned above can achieve an acceptable separation result in such a case. In this paper, we employed the method in [6] in the stage of note separation preprocessing because of its ability to capture more localized properties of a note. We shall discuss this preprocessing in more details in Section 2.

Then, a flexible re-synthesis method is required for changing smoothly the characteristics of the isolated notes in polyphonic instrumental music recordings. First, it should preserve the timbre of the notes as much as possible. Second, it shouldn't introduce any annoying artifacts when we combine the re-synthesized isolated notes with the residual polyphonic signal. We employed a source-filter based re-synthesizer to reproduce these isolated notes. Compared to FFT-based time-scaling approach, such as Phase Vocoder [7], source-filter based approach encounters little phase alignment problem because the source-filter model is fully parametric and can easily stretch/shorten the note's duration or adjust its pitch through model parameters. In the past, independence between source and filter makes source-filter model not so suitable to deal with instrumental music because musical instruments generally have a strongly coupled relationship between its source (excitation) and filter (instrument body). True envelope (TE) estimate [8] was proposed for accurate linear predictive coding coefficient estimation. As several publications [9]-[12] reported, this method can acquire more accurate spectral envelope and more suitable for music applications than traditional linear predictive estimate [13]. In this paper, TE estimate and linear predictive method are combined to make an appropriate music synthesis model. In order to have flexible pitch and time controls in the re-synthesis stage, an all-pole filter is generated based on the TE estimation of the note. Then, sound is synthesized by feeding the filter with impulses. Furthermore, pitch and time scaling becomes easy by manipulating the input impulses and the filter coefficients.

Conventionally, TE estimate is applied on short-term Fourier Transform (STFT) of the segmented signal to estimate the spectral envelope of the harmonic components. However, the acquired spectral envelope introduces noticeable magnitude modulation among analysis frames in practice. This causes artifacts in re-synthesized sounds if the above time domain source filter operation is applied. One of the reasons may be the magnitude modulation of the interested frequency component may not fall in the center of analysis frequency bin of STFT when we used a fixed analysis frame size. More discussions on TELPC based analysis and synthesis will be made in the later section. To overcome this issue, we employed a pitch period tracker to divide the original signal into analysis frames of appropriate sizes to minimize such magnitude modulation of harmonic components.

Moreover, it is now possible to use smaller frames to calculate the coefficients of the all-pole filters. Therefore, the filters can represent more local characteristics of the note because the frame size is purposely made equal to the sum of several neighbor periods. Transition from a synthesized frame to the next one is smoothened by interpolation of the all pole filters. More details will be covered in the later section. In this framework, one can manipulate a sequence of interested notes from polyphonic instrumental music recordings to result in a nearly artifact-free resynthesized music. Two commercially available recordings are used to demonstrate our results. One is “It’s easy to remember” played by John Coltrane [14], and the other is Ravel’s violin and cello duet [15].

The next section presents the note separation preprocessing employed in this work. Section 3 describes the TELPC based source filter model and the detailed analysis and re-synthesis process. Section 4 shows the re-synthesis results. We demonstrate flexible time-frequency scaling in two commercial polyphonic instrumental music recordings. Finally we present the conclusions and future perspectives.

2. NOTE EXTRACTION

The proposed system consists of a note extraction preprocessing and a note re-synthesizer. As shown in Fig. 1, the target polyphonic instrumental music recording is first decomposed into the target isolated notes and the residual signal by our recursive time-varying analysis. These notes are then feed into the proposed note re-synthesizer. Here the information of manipulation is provided by users. According to the user-defined time-frequency scaling commands, the isolated note can be resynthesized through the re-synthesizer and its characteristics can also be preserved. It is possible to change its characteristics by manipulating extra parameters. At the end, one can also combine the re-synthesized notes and the separated residual signal back into a new polyphonic instrumental music if desired. We will go to re-synthesis details in the later section.

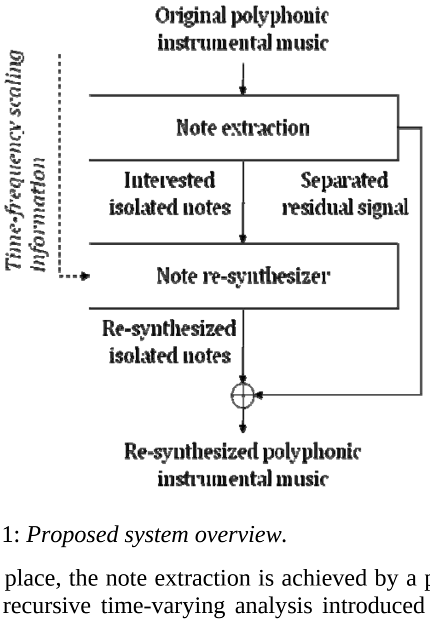
Original polyphonic instrumental music
Time-frequency scaling information
Note extraction
Interested isolated notes
Separated residual signal
Note re-synthesizer
Re-synthesized isolated notes
Re-synthesized polyphonic instrumental music

Figure 1: *Proposed system overview.*

In the first place, the note extraction is achieved by a previously developed recursive time-varying analysis introduced by Wang et al. [6]. This recursive time-varying analysis is based on the same assumption of NMF but decomposes target spectrogram in a regressive approach. Compared to the common practice of NMF, the analysis method extracts the notes by tracking the local characteristics of the target notes with much less number of adjacent frames. The local characteristic contains information such as pitch, volume and timbre. Since score information is introduced to mark those notes we are interested, there are only a few basis spectra vectors, those for the target notes and one for the residual signal. Its computation is also much less than a full note separation task. More detailed discussion can be found in [6].

3. NOTE RE-SYNTHESIZER

The note re-synthesizer contains a pitch period tracker and a source-filter model for analysis/synthesis of the target isolated notes. After the notes are extracted from the polyphonic mixture by the note extraction preprocessing described in the previous section, they are consequentially analyzed by a pitch period tracker and segmented into analysis frames of different appropriate sizes as shown in Fig. 2. After frame segmentation, we applied TE estimate[8] on short-term Fourier Transform(STFT) of the segmented frames to estimate the spectral envelopes of the harmonic components. The estimated spectral envelopes are then transformed into coefficients of LPC all pole filters for synthesis.

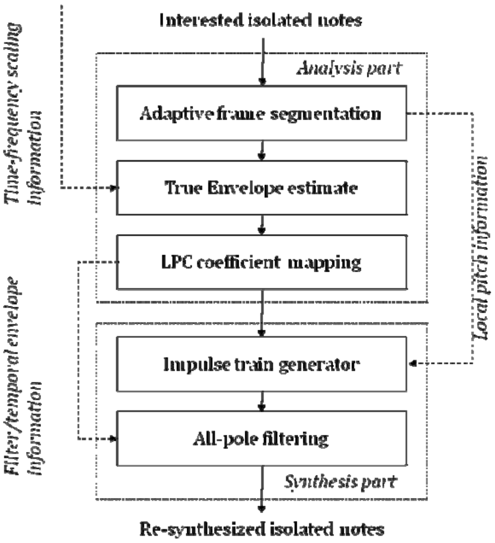
Interested isolated notes
Analysis part
Time-frequency scaling information
Adaptive frame segmentation
True Envelope estimate
LPC coefficient mapping
Local pitch information
Filter/temporal envelope information
Impulse train generator
All-pole filtering
Synthesis part
Re-synthesized isolated notes

Figure 2: *Flowchart of the proposed note re-synthesizer.*

3.1. Pitch Synchronous Segmentation and Synthesis

While TELPC was used for analysis/synthesis of the musical notes using a fixed frame size, we encountered a severe magnitude modulation problem. In many circumstances, low level noise can be easily perceived. To make TELPC a quality music synthesizer, we must deal with such problems. To reveal this problem, we tried to analyze a test 441Hz harmonic waveform consisting of several sinusoids and to re-synthesize using TELPC in different synchronous manners. One could observe that the greater magnitude modulation occurred in the waveform after TELPC synthesis using a fixed frame size, or so called time synchronous method, as shown in Fig. 3(b). On the other hand, the synthesized result using the pitch-period-based frame sizes, or formally named pitch synchronous method, showed less magnitude modulation in Fig. 3(c). Artifacts were observed even though the frame size of pitch synchronous method (=2000) is close to that of time synchronous method (=2048) in this experiment. If vibrato tones are used in the experiment, the problem can be more serious. Such pitch synchronous approaches had been applied to LPC-based methods to increase speech synthesis quality for decades [16]-[18]. A method named autocorrelation with maximum alignment described in [19] was employed to implement our pitch period tracker. This method can provide better energy compactness and less spectral leakage in spectral analysis. We will stop the discussion here because full exploration of such problems and the solutions will be out of the scope of this paper.

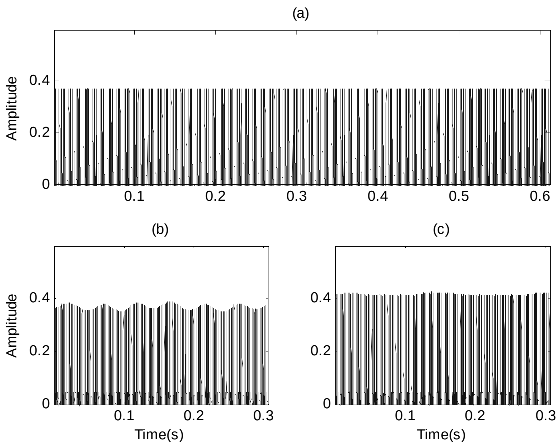

Figure 3: *Magnitude modulation after TELPC synthesis. (a) original test waveform, (b) time synchronous analysis/synthesis, (c) pitch synchronous analysis/synthesis.*

3.2. Analysis Part

In many cases, the pitch periods of a note may vary dramatically. Therefore, it is necessary to detect such variation accurately in order not to generate problems in TELPC. This is done by the pitch tracker in the previous subsection. When the isolated notes in which we are interested are ready for analysis, adaptive frame segmentation is performed based on the result of the pitch tracker. The frame size is kept small in order that the TELPC filter model parameters are obtained based on the local characteristics as much as possible. It is important to notice that the size of any analysis frame must be the collection of several local pitch periods. Moreover, all segmented signals should be basically inphase such that it becomes easier to perform time domain scaling operation.

The segmented frame is then zero-padded to the nearest 2-radiux size for computation efficiency. In our experiments, most cases adopted the size of 2048. The zero-padded signal is transformed by FFT and its spectral envelope is estimated by the TE estimate. The true envelope estimation is based on cepstral smoothing of the amplitude spectrum. The algorithm iteratively updates the smoothing input spectrum with the maximum of the original spectrum and the current cepstral representation. An all-pole filter is then derived from this envelope.

After analysis stage, the local pitch periods and filter coefficients are paired in a frame-based manner and ready for time-frequency scaling. Following steps of re-synthesis procedure according to user-defined time-frequency scaling commands described in next section.

3.3. Synthesis Part

The impulse train generator basically creates a continuous pitch-synchronous excitation based on the previously analyzed pitch information. The period of the impulse train is able to be adjusted dynamically if user wants to perform the pitch-shift operation. This excitation is fed to the all-pole filters which shape its spectral envelope. If user wants to perform the time-scaling operation, the impulse train and corresponding filter coefficients can be interpolated to stretch or shorten the re-synthesized signal length, especially in the stationary period of the signal. To guarantee the stability of filter interpolation, the interpolation of filter coefficients is performed by pole migration in z-plane. For example, if we have divided the target signal into five analysis segments, five sets of pitch periods and synthesis filter coefficients are consequently estimated. We first find the corresponding pole sets in the five frames. The trajectory formed by one pole sets is illustrated in z-plane as shown in Fig. 4.

In Fig. 4, the time index in roman indicates the analysis time and the time index in italic indicates the re-synthesis time. In this case, we tried to stretch the middle of the target signal and kept the whole signal duration unchanged. Suppose the original time indices of three middle analysis poles are 0:14, 0:18, and 0:21 respectively. They are mapped to 0:12, 0:17, and 0:24 after re-synthesis. The number and duration of synthesis frames is basically the same as those of analysis frames for keeping balance between computation complex and re-synthesis quality. Therefore, the new poles for re-synthesis are then interpolated linearly for the synthesis time indices 0:14, 0:18, and 0:21. After the pole sets are interpolated according to the above pole migration steps, the interpolated filter coefficients can be derived consequently. Furthermore, the filter coefficients can be changed in a pitch-period-wise way to improve the sound quality. The last step in the synthesis part is to add a time-varying temporal gain to the output of the all-pole filter. It is noted that the interval of every two adjacent impulses has to be adjusted accordingly if one wants to preserve the sound property before time scaling operation.

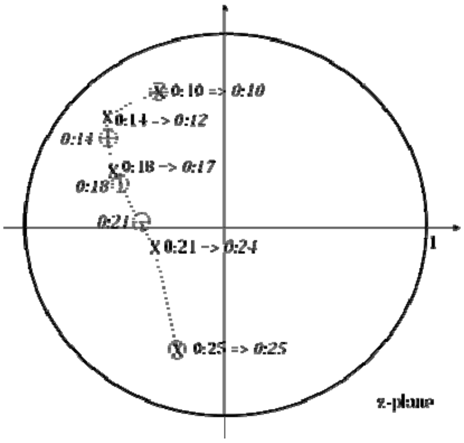

Figure 4: *Filter coefficient interpolated by pole migration diagram. The trajectory of a set of five poles is plotted to illustrate pole migration in both analysis and synthesis aspects. There are five analysis segments whose original time indices are 0:10, 0:14, 0:18, 0:21 and 0:25. Their time indices re-mapped to 0.10, 0.12, 0.17, 0.24 and 0.25 respectively. The synthesis poles are linearly interpolated at the original time indices alignment. (×: Analysis pole. ⊕: Synthesis pole. Time index in roman: analysis time. Time index in italic: synthesis time).*

4. RE-SYNTHESIS RESULTS

We used two commercially available recordings to demonstrate our results. One is John Coltrane's performance of "It's easy to remember" which appeared in the album 'Ballads' [14]. John Coltrane played the saxophone part accompanied by bass, piano and drums. The other is Ravel's violin and cello duet [15]. The first isolated note in "It's easy to remember" begins with a slight upward glide followed by a long sustain as shown in Fig. 5(a) and Fig. 6(a). We tried to enhance the musical tension in this beginning. In Fig. 5(b), we re-synthesized a deeper and longer glide started with a little bit lower pitch than the original note. The second case is to let the note have a dramatic beginning by replacing the glide with a vibrato effect whose pitch varied between 460Hz and 477 Hz as shown in Fig. 6(b).

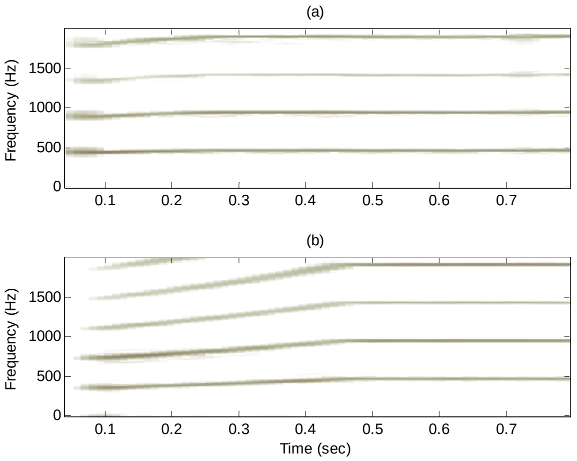

Figure 5: *First isolated note in recording, "It's easy to remember" (gliding effect). (a) Original interested clip. (b) Re-synthesized result with a deeper glide.*

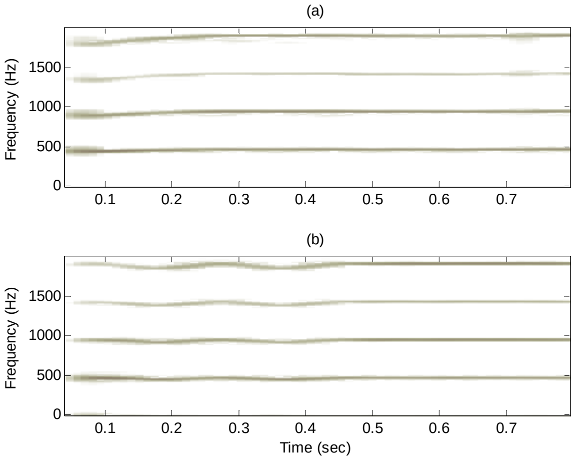

Figure 6: *First isolated note in recording, "It's easy to remember" (vibrato effect). (a) Original interested clip. (b) Re-synthesized result with a vibrato.*

In Fig. 7, the spectrogram of the original performance from 0 second to 3 second is shown. We re-synthesize this clip by changing the first note into four short notes covered from the original pitch contour, keeping the second note unchanged, mitigating the beginning glide of the third note, increasing the vibrato rate and depth of the fourth note. The result is shown in Fig. 8. One can compare the differences in the rectangular boxes in the two figures. It sounded like John Coltrane playing the same piece using different kinds of styles. It is difficult to remove the effect of cymbals because of the wide band characteristics when Elvin Jones, the drummer, came in. One can hear this in the provided sound samples at our website [20]. This may be due to the limitation of the note extraction part.

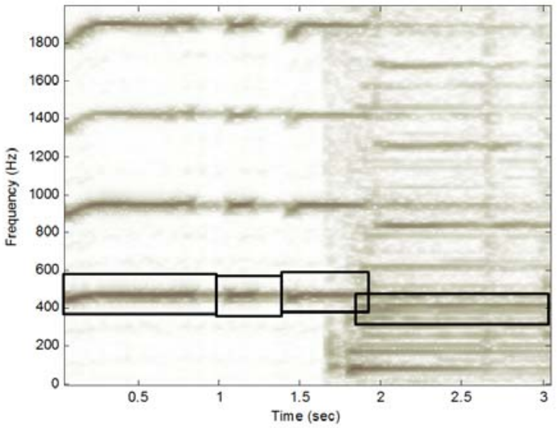

Figure 7: *John Coltrane's "It's easy to remember" (original). Rectangular box indicates interested saxophone notes.*

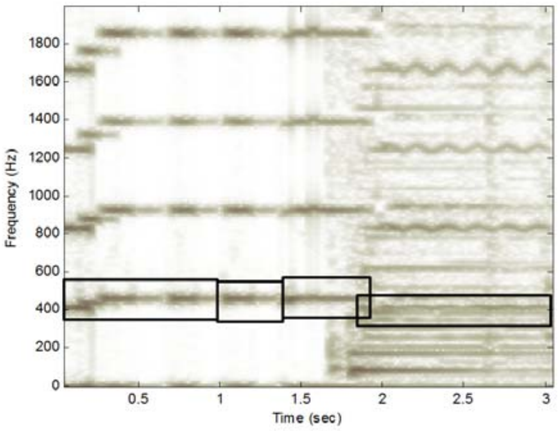

Figure 8: *John Coltrane's "It's easy to remember" (re-synthesized). Rectangular box indicates re-synthesized saxophone note. The first note is reassembled by four short notes. The second note is unchanged. The glide part of the third note is mitigated and the fourth note is changed into a vibrato style.*

The last demonstration is Ravel's violin and cello duet. We tried to change the vibrato behavior of the second violin note into an open-string style. Comparing with the original clip as shown in Fig. 9, the interested violin notes, indicated by rectangular, were re-synthesized properly in Fig. 10. Finally, all demonstrated music pieces can be heard at our website [20].

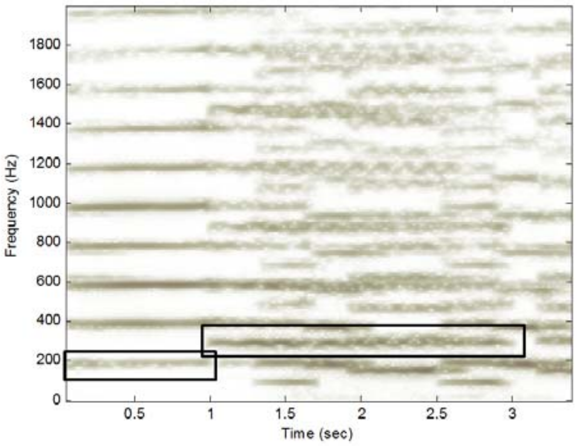

Figure 9: *Ravel's violin and cello duet (original). Rectangular indicates interested violin notes: the first one is an open-string tone and the second one is a vibrato tone.*

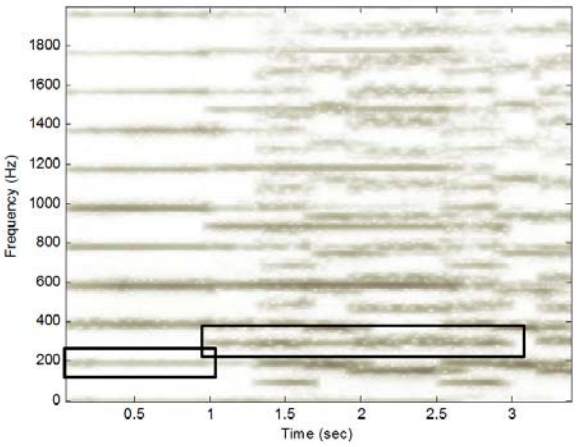

Figure 10: *Ravel's violin and cello duet (re-synthesized). Rectangular indicates re-synthesized violin notes: the first one keeps unchanged and the second one is changed to an open-string tone.*

5. CONCLUSION AND FUTURE WORK

A TELPC based re-synthesis method for isolated notes of polyphonic instrumental music recordings was proposed in this paper. Our proposed system described a framework from interested note extraction to isolated note analysis and time-frequency scaling re-synthesis. The note extraction was achieved by a score-informed recursive regularization method for time varying analysis of polyphonic music signals. TELPC analysis/synthesis is processed pitch synchronously to avoid audible artifacts. Smooth changes of the properties of isolated notes are done by the pole migration method that allows every pitch period to have its own filter coefficient set if necessary. In this framework, one can manipulate a sequence of interested notes extracted from polyphonic instrumental music recordings to result in interesting re-synthesized music such like our demonstrations.

Although the re-synthesized sounds showed encouraged results, there are still problems needed to be solved. Our proposed system is based on harmonic instrumental music; therefore, it is not able to take good care of non-harmonic notes appeared in the accompaniment, such like drums or cymbals. Besides, the note extraction is not so powerful to get clear notes for all kind of music sources. We will continue to improve it in the future.

6. ACKNOWLEDGEMENT

The authors would like to thank the National Science Council, ROC, for its financial support of this work, under Contract No.NSC 100-2221-E-006-247-MY3.

7. REFERENCES

[1] M. A. Casey and A. Westner, “Separation of mixed audio sources by independent subspace analysis,” in *Proc. ICMC*, pp. 154-161, 2000.

[2] C. Uhle, C. Dittmar, and T. Sporer, “Extraction of drum tracks from polyphonic music using independent subspace analysis,” in *Proc. ICA*, pp. 843-848, 2003.

[3] D. D. Lee and H. S. Seung, “Learning the parts of objects by nonnegative matrix factorization,” *Nature*, vol. 401, pp. 788-791, 1999.

[4] B. Wang and M. D. Plumbley, “Musical audio stream separation by non-negative matrix factorization,” in *Proc. DMRN Summer Conference*, Glasgow, 2005.

[5] S. A. Abdallah and M. D. Plumbley, “Unsupervised analysis of polyphonic music using sparse coding,” *IEEE Trans. Neural Networks*, vol.17, no. 1, pp. 179-196, 2006.

[6] T.-M. Wang, T.-C. Chen, Y.-L. Chen, Alvin W.Y. Su, “Time-dependent recursive regularization for sound source separation,” in *Proc. of the 3rd International Conference on Audio, Language and Image Processing* (*ICALIP2012*), Shanghai, China, Jul. 16-18, 2012.

[7] M. Dolson, “The phase vocoder: A tutorial,” *Computer Music J.*, vol. 10, no. 4, pp. 14-27, 1986.

[8] A. Roebel and X. Rodet, “Efficient spectral envelope estimationand its application to pitch shifting and envelope preservation,” in *Proc. DAFx*, 2005.

[9] F. Villavicencio, A. Röbel, and X.Rodet, “Improving LPC spectral envelope extraction of voiced speech by true-envelope estimation,” in *Proc. of the ICASSP'06*, France, 2006.

[10] V. Villavicencio, A. Röbel, and X. Rodet, "Applying improved spectral modeling for high quality voice conversion," *ICASSP2009*, pp. 4285-4288, 2009.

[11] J.J. Burred, A. Roebel, and T. Sikora, “Dynamic spectral envelope modeling for timbre analysis of musical in strumentsounds,” *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 3, March 2010.

[12] M. Caerano and X. Rodet, “A source-filter model for musical instrument sound transformation,” *ICASSP2012*, pp. 137-140, 2012.

[13] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[14] John Coltrane, McCoy, Tyner, Jimmy Garrison and Elvin Jones, *Ballads*, Compact disc. 1962.

[15] Kodaly, Ravel, Maggio Ormezowski, Bianchi, *Works for Violin and Cello*, Compact disc. 1995.

[16] M. V. Mathews, “Pitch synchronous analysis of voiced sounds,” *Journal of the Acoustical Society of America*, vol. 33, no. 2, 1961.

[17] Y. Medan and Eyal Yair, “Pitch synchronous spectral analysis scheme for voiced speech”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 9, September 1989.

[18] H. Yang, S.N. Koh, and P. Sivaprakasapillai, “Pitch synchronous multi-band (PSMB) speech coding”. in *Proceedings ICASSP-95*, 1995.

[19] H. Ding, I. Y. Soon, and C. K Yeo, “A DCT-based speech enhancement system with pitch synchronous analysis,” *IEEE Trans. On Audio, Speech, and Lang. Proc.*, vol. 19, no. 8, pp. 2614-2623, Nov. 2011.

[20] Ya-Han Kou, Re-synthesizing Isolated Notes from Polyphonic Instrumental Music Recordings. URL available at: http://screamlab-ncku-2008.blogspot.tw/2013/04/music-files-of-telpc-based-time.html [accessed 14 April 2013]