ANALYSIS AND TRANS-SYNTHESIS OF ACOUSTIC BOWED-STRING INSTRUMENT RECORDINGS – A CASE STUDY USING BACH CELLO SUITES

Yin-Lin Chen, Tien-Ming Wang, Wei-Hsiang Liao, Alvin W.Y. Su
SCREAM Lab., Department of CSIE,
National Cheng-Kung University
Tainan, Taiwan
p76981057@mail.ncku.edu.tw

ABSTRACT

In this paper, analysis and trans-synthesis of acoustic bowed string instrument recordings with new non-negative matrix factorization (NMF) procedure are presented. This work shows that it may require more than one template to represent a note according to time-varying behavior of timbre, especially played by bowed string instruments. The proposed method improves original NMF without the knowledge of tone models and the number of required templates in advance. Resultant NMF information is then converted into the synthesis parameters of the sinusoidal synthesis. Bach cello suites recorded by Fournier and Starker are used in the experiments. Analysis and trans-synthesis examples of the recordings are also provided.

Index Terms—trans-synthesis, non-negative matrix factorization, bowed string instrument

1. INTRODUCTION

In recent years, trans-synthesis is an interesting topic in musical processing [1-2]. Depending on either time-domain or frequency-domain properties of audio signal processing, the authors attempt to overcome the problems when the real audio recordings are analyzed, transformed, and re-synthesized. In particular, non-negative matrix factorization (NMF) is recently well-known to factorize spectrum into basis spectra and temporal activation in music signal analysis [3]. It is widely used for music transcription [4-5], pitch detection and onset detection. To improve original NMF, temporal smoothness [6], sparseness [7], and harmonicity/inharmonicity [8] have been considered as primary constraints. For accurate piano music transcription proposed in [9], trained note templates are obtained in advance.

[5] indicates that one needs to use enough number of required templates to give good results. In [8], 88 templates are used due to the pitch rage of piano, and this takes lots of computation and memory space. In real case, however, the spectral behavior of one note played by an instrument is always time-varying, especially in the case of bowed string instruments. That means it is not reasonable to achieve the NMF task by using only one template per note. Fig. 1 shows the spectra of two different frame of the note B3 played by pianoUPM project [10]. It is interesting to note that spectral contours of two frames (20^{th} frame and 300^{th} frame in this case) are apparently in different shapes.

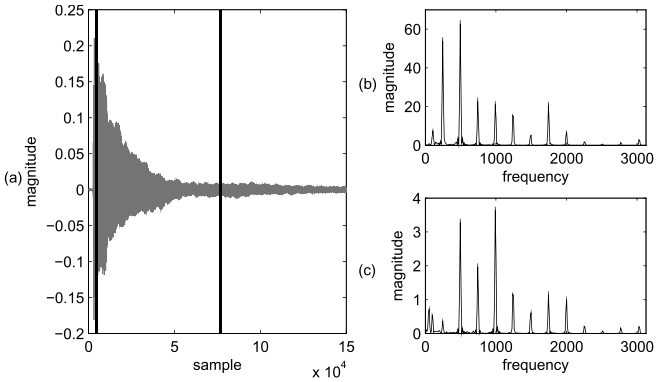

Figure 1: *(a) The waveform of the note B3. Two bold lines indicated 20th and 300th frame. (b) The spectrum of 20th frame. (c) The spectrum of 300th frame.*

Moreover, keyboard music is usually employed in the evaluations in previous works. How NMF performs for other types of instruments is seldom addressed. In this paper, an iterative procedure for deciding the appropriate number of templates is proposed for NMF. No trained tone model is required in advance. The proposed method is applied to acoustic cello recordings.

In order to reproduce the results of the proposed NMF procedure, appropriate synthesis technique is preferably applied. Spectral modeling synthesis (SMS) is proposed to divide music signal into deterministic part and stochastic part. This model was then extended by including transient modeling [11], called sinusoids plus transient and noise. In this paper, deterministic part of conventional SMS is applied to re-synthesize polyphonic musical signals analyzed by NMF. Sinusoidal synthesis [12] is employed here. Sound transformation examples are accomplished and provided in [13] as well.

This paper is organized as follows. In section 2, NMF is briefly reviewed and its modification is proposed. In section 3, experimental results of NMF analysis and music re-synthesis are given. Conclusion is given in section 4.

2. NMF-BASED MUSIC SIGNAL ANALYSIS

2.1. Brief review of NMF

In [3], given an $m \times n$ nonnegative matrix Y, NMF is used to factorize Y into an $m \times r$ nonnegative matrix W and an $r \times n$ nonnegative matrix X such that:

$$Y \approx \tilde{Y} = WX \tag{1}$$

A cost function such as KL divergence shown in equation 2 is designed as a measurement to evaluate how well the multiplication of W and X can approximate Y.

$$D_{KL}\left(Y \mid WX\right) = \left\| Y \otimes \log\left(\frac{Y}{WX}\right) - Y + WX \right\|_F \tag{2}$$

$\|.\|_F$ is the Frobenius norm. $\otimes$ is element-wise multiplication. By iteratively updating W and X, and the cost is minimized. For example, the update rules for (2) are shown as follows.

$$X_{rn} \leftarrow X_{rn} \frac{\sum_m W_{mr} Y_{mn} / (W \cdot X)_{mn}}{\sum_v W_{vr}} \tag{3}$$

$$W_{mr} \leftarrow W_{mr} \frac{\sum_n X_{rn} Y_{mn} / (W \cdot X)_{mn}}{\sum_k X_{rk}} \tag{4}$$

In music analysis, Y is used to represent signal spectrogram. For example, the column vector Y_j of Y is the spectrum of the j^{th} time frame. Frame size and total number of time frames are m and n, respectively. Hence, the column vector W_i of W represents the template of the i^{th} note contained in the signal, and the element X_{ij} of X indicates the intensity of the i^{th} note which appears in the j^{th} time frame.

For NMF, it is important to decide the required number of note templates, r, in advance, in order to give good factorization of Y. Two methods are usually used to decide r. One uses the number of different notes appearing in the signal [5]. The other sets r as 88 if piano music is analyzed [8]. However, the number of different notes is usually unknown and computation complexity is huge if $r = 88$.

Moreover, it is questionable if a template is enough for a piano tone. A sound clip containing 2 different notes obtained from [10] is analyzed. r is set as 2 and 88, respectively. NMF in [5] is used. The results are shown in Fig. 2. The left-hand-side figures represent X, and the right-hand-side represent the corresponding W. In Fig. 2(a), the 1st template contains both notes, and the 2nd template contains one of the notes. A note appears in both templates at the same time. It also shows that 2 templates are not enough to factorize the signal. In Fig. 2(b), only 4 templates give large enough intensity. The 1st note appears in the 33th and 45th templates, but with different spectrum envelopes. Similarly, the 2nd note appears in the 35th and 47th templates. It seems 88 templates are enough, but the computation time is huge.

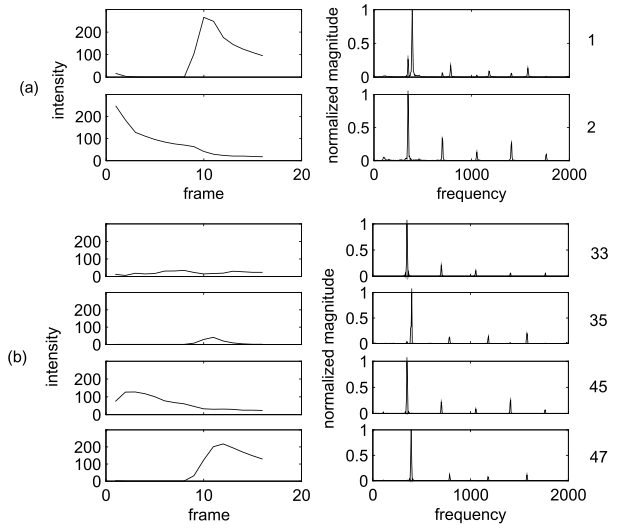

Figure 2: *NMF for piano sound clip: (a) r = 2 (b) r = 88.*

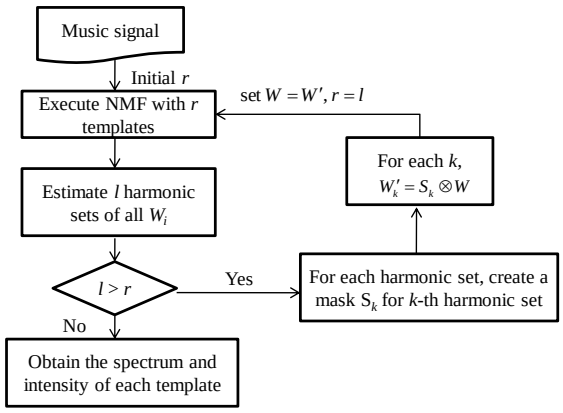

Figure 3: *Flow chart of the proposed NMF procedure.*

2.2. New NMF Procedure for Harmonic Music Signals

The proposed NMF depicted in Fig. 3 is proposed to solve the problem stated above without the knowledge of exact number of required templates. One first starts with a small r. The audio recording is then analyzed by means of NMF described in [5]. Two preliminary non-negative matrices, which stand for intensities and spectra of all templates, come out as result. Next, the spectrum of each template is evaluated to find if it contains more than one harmonic set by using the method in [14]. If so, a mask function is used to extract the k^{th} harmonic set. It can be represented as

$$\begin{aligned} W_k' &= S_k \otimes W \,, \\ S_k &= \sum\nolimits_p (2\sigma)\, f\left(p \cdot \mu_k - \sigma, p \cdot \mu_k + \sigma; x\right), \end{aligned} \tag{5}$$

where $f(x)$ is uniform distribution for the interval $[p \cdot \mu_k - \sigma, p \cdot \mu_k + \sigma]$, μ_k is the fundamental frequency corresponding to W_k' and p is partial index. σ is set as 3% of the fundamental frequency.

l represents the number of harmonic sets extracted from r templates. If $l \le r$, the loop will be stopped and the eventual matrices are obtained. Otherwise, we set r as l, use the temporal matrices as initial conditions and NMF process is then executed again.

After the intensity and spectral information of all note templates are obtained, it can be found that FFT spectrum and $\tilde{Y}$ in (1) are very close to each other (the result is shown in Section 3). Therefore, such NMF information can be used as parameters of the synthesis method stated below.

3. EXPERIMENTAL RESULTS

3.1. Results of the proposed NMF procedures.

Two music passages of Bach's cello suites No.1 (BWV1007) recorded by Starker [15] and Fournier [16] are analyzed. They are both polyphonic. There are 4 different pitches in the first 16 notes, shown in Fig.4. Cost function in (2) is implemented. Frame size is 8192 and hop size is 2048. Initially, the number of template, r, is set as 4. Templates are initialized with random numbers. 100 iterations are performed for NMF update rules. The outer loop in Fig. 3 runs only twice to reach the final results in both cases.

After obtaining the NMF result, W and X are used to obtain synthesis parameters, described in Section 3.2 due to the similarity between Y and $\tilde{Y}$. The 95th frame of Y and $\tilde{Y}$ are shown in Fig. 5. It shows that the envelopes of two representations are close.

Figure 4: *Score of first 16 notes.*

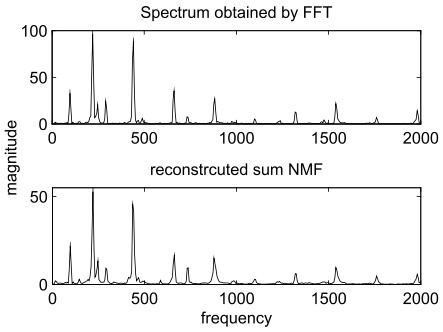

Figure 5: *FFT spectrum and $\tilde{Y}$ of the 95th frame obtained from the Fournier recording.*

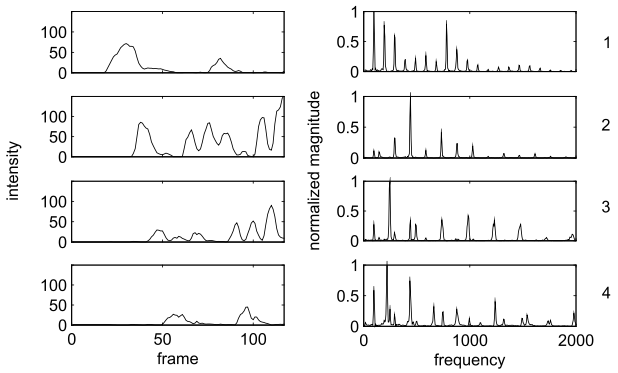

Figure 6: *NMF result when the 1st outer loop factorization is finished (r=4). [Fournier]*

Fig.6 shows the results of conventional NMF procedure of the Fournier recording with 4 initial templates ($r = 4$). In this case, more than one harmonic set exists in 3rd and 4th template such that these templates cannot completely represent all notes of the recording. One of the intuitive ways is to extend template dimension and makes these additional harmonic sets be located in new templates.

The final results of the proposed NMF procedure of the Starker and Fournier recordings are shown in Fig. 7 and 8 respectively. In the figures, each column vector W_i is normalized such that the intensity information is solely represented by the corresponding row vector X_i of X as

$$W_i = \frac{W_i}{\max(W_i)}, \; X_i = X_i \cdot \max(W_i). \tag{6}$$

The note information is stated as follows. In Fig. 7, pitches of the first three templates are G2, the 4th to the 6th ones are D3, the 7th and the 8th ones are B3, the 9th one is A3, and the last template with no obvious f0 is regarded as noise. In Fig. 8, pitches of the first three templates are G2, the 4th and the 5th ones are D3, the 6th and the 7th ones are B3, the 8th one is A3, and the last template with no obvious f0 is also regarded as noise.

By taking the first three templates of Fig. 7 as examples, both of their spectral envelopes and the corresponding intensity functions are quite different. The 1st template can be regarded as the attack template of G2 notes because it has the largest intensity of three and its onset points appear early. Moreover, the duration of the 2nd template is longer and the onsets of the 3rd template appear between those of the 1st and the 2nd ones. It is reasonable to regard them as a sustain template and a decay template respectively. It is interesting to note that the number of templates is solely related to the waveform behavior of the note. It usually depends on the physical architecture of the instrument and the gestures of the musician while he/she playing that note. Three templates are needed to represent G2 notes in this case. The required numbers of templates of D3 notes in the two recordings are different (2 and 3 for the Fournier and Starker recording, respectively).

Our experiments described above show that to use one template for a note may not be enough to sufficiently model a note, especially when the information is to be used in the re-synthesis process. It is interesting to see that when r is set as 10 without using the proposed procedure, one can't successfully factorize the signal from the Starker recording. Due to the harmonic mask $S_{i,k}$, the proposed NMF procedure outperforms in this case. Comparing to Fig. 7, the 6th and the 9th templates contain 2 notes in Fig. 9. This shows that the proposed method is advantageous even when the number of templates is enough.

3.2. Resynthesis

Eventually, W and X are converted to parameters of sinusoids. The number of partials is set as 50. The original and the synthetic signals are compared in Fig.10. The two spectral envelopes are close. Sound transformation such as timbre modification can be easily accomplished by replacing NMF templates of the corresponding notes. Sound examples are provided in [13].

4. CONCLUSION AND DISCUSSION

Analysis and trans-synthesis of acoustic cello recordings made by Fournier and Starker with modified NMF procedure is presented. It is not required to have pre-trained tone model and to know the necessary number of templates in advance for the modified NMF to give good results. It is also found that more than one template can be used to preferably represent a note according to its different sounding states. Spectrum and intensity information of NMF is then converted into the synthesis parameters of the sinusoids. Trans-synthesis sound examples of Bach cello suites can be heard in [13].

Comparing to other state-of-art methods, this paper puts emphasis on the applications of sound reproduction rather than music transcription. That means spectral behavior of the timbre is more significant than the statistical results like F-measure or mean overlap ratio. Without applying any temporal or spectral constraints on NMF update rules, the proposed method models each note by extending template dimension such that these templates can be in charge of different states of one note. According to the aspect, the adequate number of templates will be unpredictable if one music note is played by various kinds of instruments with different gestural representations. It therefore takes amount of computation by means of the iteratively procedure of harmonic verification and multiple NMF updates.

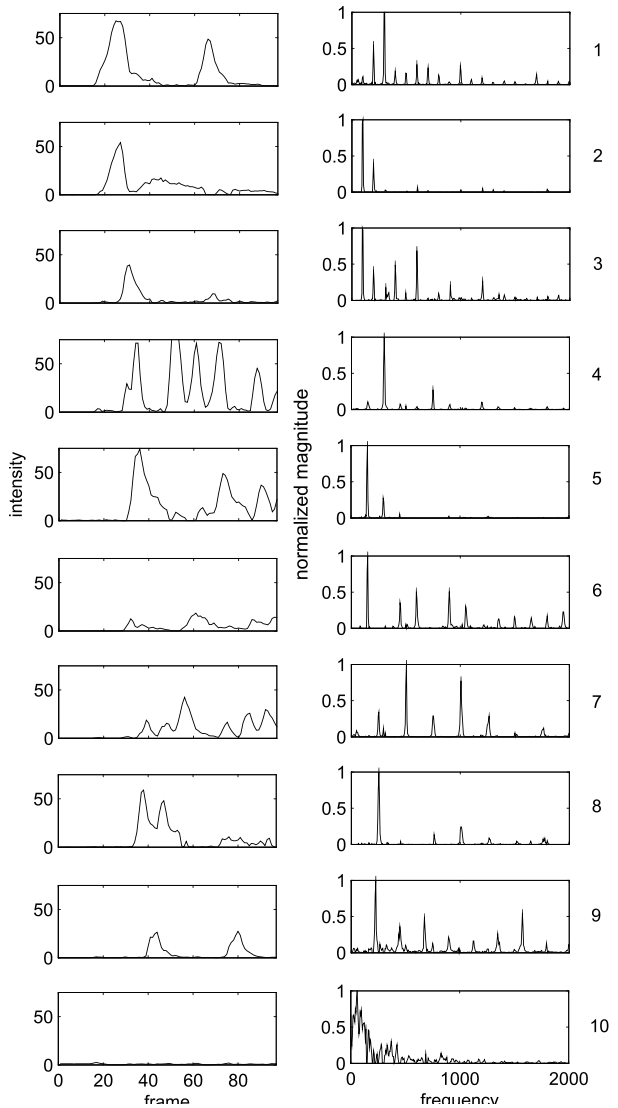

Figure 7: *the result of the proposed NMF procedure. [Starker]*

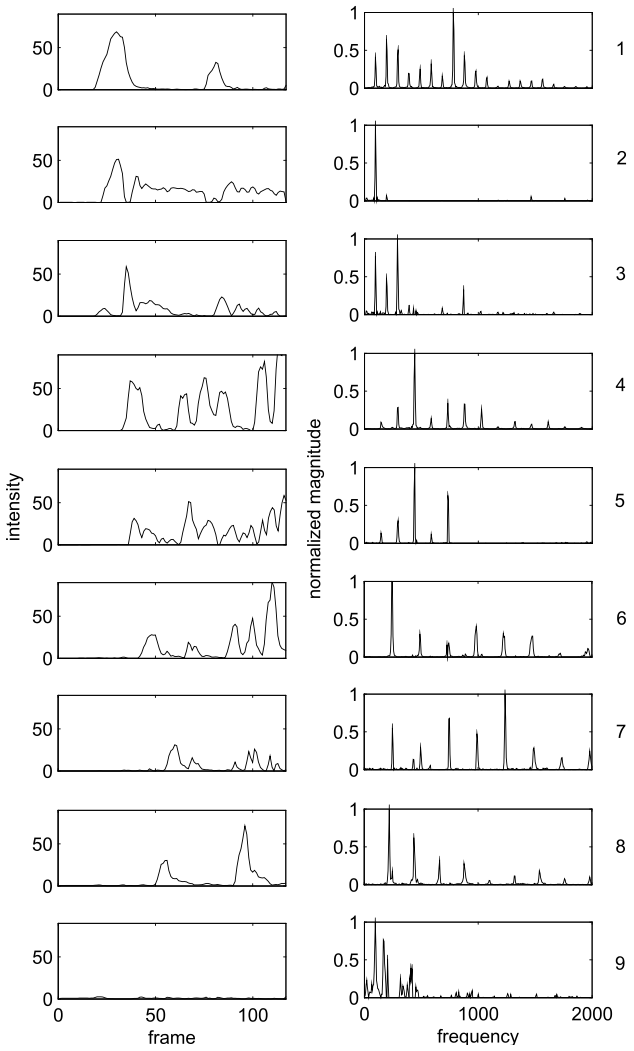

Figure 8: *the result of the proposed NMF procedure. [Fournier]*

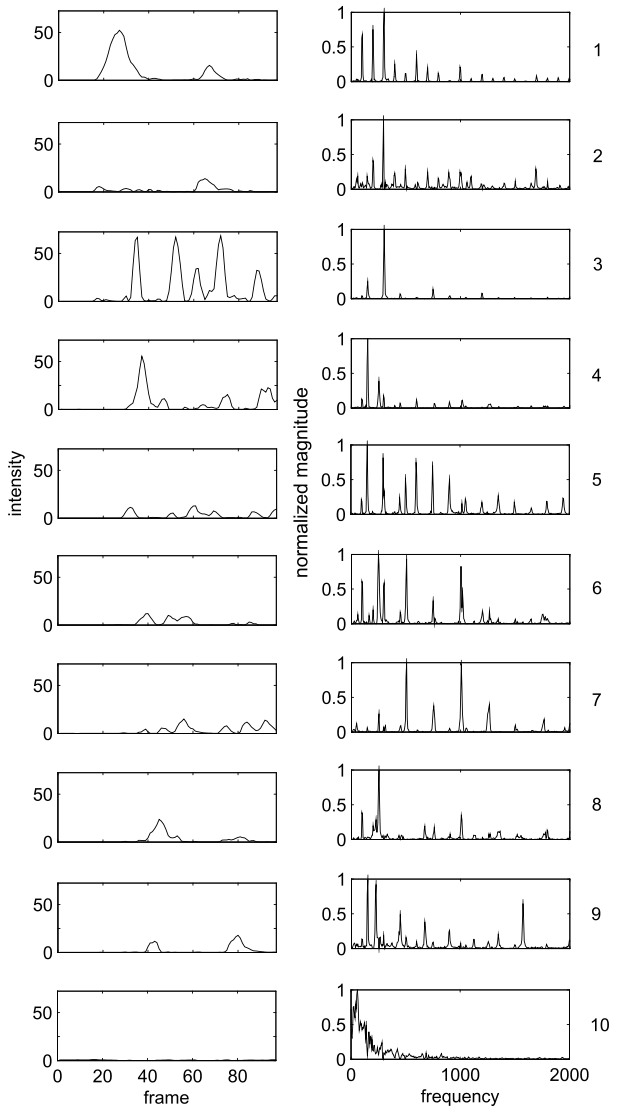

Figure 9: *the result of convention NMF procedure (r=10). [Starker]*

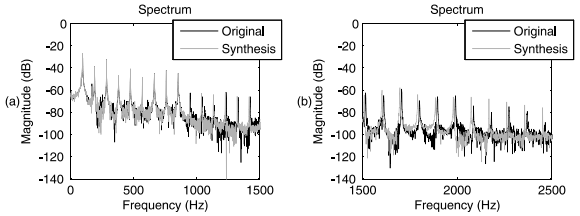

Figure 10: *FFTs of (a) lower partials and (b) higher partials of 32th original and synthetic frame of the Fournier recording.*

The modeling of time-varying templates is one of possible improvements. The refinement of temporal prior may suit for modeling the sustain and decay parts of the note. The attacks or other transient parts are however disfavored. Therefore, the templates are considerable to 'morph' as time goes by when the notes are activated. A time-varying multiplicative gradient approach with adaptive templates may be investigated on in the future.

5. REFERENCES

[1] W.-C. Chang, Y.-S. Siao, and 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] T.-M. Wang, W.-C. Chang, K.-T. Lin *et al.*, "Trans-Synthesis System for Polyphonic Musical Recordings of Bowed-String Instruments," in *Proc. of the 12th Int. Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, 2009.

[3] D. Lee, and H. Seung, "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, pp. 556-562, 2001.

[4] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 538-549, 2010.

[5] P. Smaragdis, and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2003, pp. 177-180.

[6] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066-1074, 2007.

[7] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, 2004.

[8] E. Vincent, N. Berlin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA, 2008, pp. 109-112.

[9] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *proc. 11th International Society for Music Information Retrieval Conference*, Utrecht, Netherlands, 2010.

[10] L. Ortiz-Berenguer, E. Blanco-Martin, A. Alvarez-Fernandez *et al.*, "A Piano Sound Database for Testing Automatic Transcription Methods," in *125th AES Convention*, San Francisco, 2008.

[11] T. Verma, and T. Meng, "An Analysis/Synthesis Tool for Transient Signals that Allows a Flexible Sines+ Transients+ Noise Model for Audio," in *proc. IEEE Conf. on Acoustics, Speech and Signal Processing*, 1998.

[12] R. J. McAulay, and T. F. Quatieri, "Speech Analysis Synthesis Based on a Sinusoidal Representation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 34, no. 4, pp. 744-754, Aug, 1986.

[13] T.-M. Wang. "Experimental Results," http://www.scream.csie.ncku.edu.tw/index.php/research/asp/dafx2011exp.

[14] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 804-816, 2003.

[15] J. Starker, "Bach: Six Suites; Sonatas in G & D," Mercury Records, CD 1, Track 1, 1991.

[16] P. Fournier, "Bach: 6 Suiten für Violoncello solo," Archiv Records, CD 1, Track 1, 1997.