EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Musical note analysis of solo violin recordings using recursive regularization

Yi-Ju Lin, Tien-Ming Wang*, Ta-Chun Chen, Yin-Lin Chen, Wei-Chen Chang and Alvin WY Su

Abstract

Composers may not provide instructions for playing their works, especially for instrument solos, and therefore, different musicians may give very different interpretations of the same work. Such differences usually lead to time, amplitude, or frequency variations of musical notes in a phrase in the signal point of view. This paper proposes a frame-based recursive regularization method for time-dependent analysis of each note presenting in solo violin recordings. The system of equations evolves when a new frame is added and an old frame is dropped to track the varying characteristics of violin playing. This method is compared with a time-dependent non-negative matrix factorization method. The complete recordings of both BWV 1005 No. 3 played by Kuijken and 24 Caprices op. 1 no. 24 in A minor played by Paganini are used for the transcription experiment, where the proposed method performs strongly. The analysis results of a short passage extracted from BWV 1005 No. 3 performed by three famous violinists reveal numerous differences in the styles and performances of these violinists.

Introduction

Analyses of performances are mostly subjective in the domain of musicology. Objective analysis has become possible with advances in information technologies and sound/music analysis tools, such as pitch/partial tracking, score alignment/following, melody tracking, and extraction. Non-negative matrix factorization (NMF) [1] is a popular tool for musical signal analysis such as pitch estimation, chord recognition, and automatic transcription [2-4]. In NMF, the matrix of the input magnitude spectrum is decomposed into the product of two matrices. One matrix is formed by a certain number of magnitude spectra and is called the template matrix or dictionary matrix. The other matrix is the intensity information of the notes and is called the intensity matrix or activation matrix. When considering the decomposition of audio recordings, these matrices are, for both NMF and the proposed method, related to several notes, each with a quasi-harmonic spectrum activating during a specific time period. Furthermore, NMF is usually used on the Fourier spectrogram which is easy to apply time-frequency masking but hard to extract time-varying sources. Some additional models are needed to enforce the procedure of decomposition, such as time-dependent parametric and harmonic templates [5] and Markov-chained base [6]. On the other hand, decomposing constant-Q spectrograms is difficult to apply time-frequency masking but shows good potential to deal with spreading of higher harmonic frequencies when the pitch is getting higher, such like scale invariance across linear frequency [7] and shift invariance across log-frequency [8]. State-of-the-art methods in these areas can be found in the annual Music Information Retrieval Evaluation eXchange (MIREX) [9].

Good results can be achieved when the number of notes and/or the spectra information of notes is known *a priori*. In the analysis of polyphonic recordings, to determine the number of notes appearing in a single time frame is firstly discussed. A harmonic structure is generally desirable, and the spectral bases are usually constrained to be harmonic in the applications [10]. A fixed number of templates are usually set in previous works according to the note range of interest. Pitches of a violin can, however, vary continually, and fixed pitch templates are unsuitable in the analysis of bowed string instruments. Two issues are then welcome to be discussed in this work: (a) how to determine the exact number of notes and (b) how to model the time-varying notes with suitable templates.

Methods to estimate the possible number of notes have been discussed in [11,12]. In [13], a dynamic note

*Correspondence: showmin@csie.ncku.edu.tw
SCREAM Laboratory, Department of Computer Science Information Engineering, National Cheng-Kung University, Tainan 701, Taiwan

number detection for NMF is proposed to analyze solo bowed string instrument recordings. Since fixed template is employed in [13], a note with a time-varying spectrum which resulted from performing skills such as vibrato and portamento is encouraged to be obtained by using multiple templates. In [5], the time-dependent parametric and harmonic templates are applied to NMF when the pitch of a note varies. The method provides a parametric representation of the harmonic atoms, which can depend on a fundamental frequency parameter, a chirp parameter, and so on with respect to time. It can therefore represent a time-varying note by using only one single template.

Recursive regularization [14] has been widely applied in the areas of system identification, image restoration, noise reduction, echo cancellation, and blind deconvolution [15,16]. The proposed method decomposes the magnitude spectrogram into the product of a template matrix and an intensity matrix based on the modified version of the previous work in high-resolution image reconstruction [16]. In this work, a new algorithm is developed such that the two matrices are updated whenever a new frame is added and an old frame is dropped. This online scheme is similar to the so-called online dictionary learning but does not keep a global dictionary for identified patterns, i.e., musical notes of the same pitch. To analyze violin solo recordings, we regard each note as one single source in this paper. Some works have been proposed for online dictionary learning using L2 norms [17,18], KL divergence [19], and IS divergence [20]. Here, we considered L2 norms to simplify the derivations of the proposed recursive algorithm. Similar to [16], the new iterative update procedure also eliminates the matrix inversion operation to reduce the computational complexity. Because the convergence of the recursive regularization has been well addressed, those who are interested could find related materials in [16]. The systematic flow proposed in the previous work [13] to find a new note template is modified for the application of this work. The concepts of harmonic and sparseness constraint [21,22] are also adopted. The proposed method is compared with the time-dependent NMF method [5] by using the complete recordings of BWV 1005 No. 3 played by Kuijken [23] and 24 Caprices op. 1 no. 24 in A minor played by Paganini [24]. Finally, the proposed method is tested using Bach solo violin recordings by three violinists, that is, Arthur Grumiaux, Sigiswald Kuijken, and Hilary Hahn [23,25,26]. It is easy to identify the differences in their playing styles when note-by-note spectral and intensity information is available. The insightful discussions will be discussed in the 'Results' section.

The remainder of this paper is organized as follows: The 'Formulation of regularized analysis system' section presents the basic formulation of the decomposition problem using the regularization method. The 'Frame-based recursive regularization analysis' section presents the frame-based recursive regularization analysis method. We then present some experiments and corresponding results in the 'Experiments' and 'Results' sections. Lastly, the 'Conclusions' section offers the conclusion and the discussion of future works.

Formulation of regularized analysis system

Let m be the number of consecutive frames. Each frame has $2n$ samples. Discrete Fourier transform (DFT) is performed for each frame to obtain its magnitude information. We obtain an m-by-n matrix, $\mathbf{V}$, where $\mathbf{V}_{ji}$ represents the magnitude of the ith Fourier coefficient of the jth frame. Let r be the number of pitches present in these frames. We obtain an r-by-n template matrix, $\mathbf{W}$, where $\mathbf{W}_{ki}$ represents the magnitude of the ith Fourier coefficient of the kth template. Finally, we obtain an m-by-r intensity matrix, $\mathbf{H}$, where $\mathbf{H}_{jk}$ represents the intensity of the kth template of the jth frame. Hence, $\mathbf{W}$ and $\mathbf{H}$ are used to construct $\mathbf{V}$, as follows:

$$\mathbf{V} = \mathbf{HW}. \tag{1}$$

A cost function can be set as

$$D = \|\mathbf{V} - \mathbf{HW}\|^2. \tag{2}$$

Subsequently, the template matrix and the intensity matrix can be obtained easily as

$$\mathbf{W} = (\mathbf{H}^\top\mathbf{H})^{-\mathbf{1}}(\mathbf{H}^\top\mathbf{V}) \tag{3}$$

$$\mathbf{H}^\top = (\mathbf{W}\mathbf{W}^\top)^{-\mathbf{1}}(\mathbf{W}\mathbf{V}^\top). \tag{4}$$

The result can be obtained by evaluating (3) and (4) iteratively. Although (1) is similar to NMF in its formulation, (3) and (4) do not enforce the factorization of a non-negative matrix into two non-negative matrices, in comparison to NMF. Since the goal is to get a reasonable distribution of frequency energies, the negative elements of $\mathbf{W}$ and $\mathbf{H}$ can be set to zeros to re-evaluate the equations and obtain a non-negative result in every iteration. Notice that the matrix $\mathbf{V}$ is represented as (1) rather than $\mathbf{V} = \mathbf{WH}$ commonly used in NMF-related literatures to make the derivatives of following formulations more readable without loss of generality.

To improve (3) and (4), a penalty term is added to support a temporal smoothness mechanism in the context of the proposed online scheme. For example, (2) can be modified to

$$J = \|\mathbf{V} - \mathbf{HW}\|^2 + \lambda\|\mathbf{W} - \mathbf{C}_\mathbf{W}\|^2 + \gamma\|\mathbf{H} - \mathbf{C}_\mathbf{H}\|^2, \tag{5}$$

where $\lambda > 0$ and $\gamma > 0$ are carefully chosen to ensure the stability of the solution. Such a process is called regularization [14]. Since both frequency response and intensity of a note evolve slowly in a short time, $\mathbf{C}_\mathbf{W}$ is determined as a template matrix in the previous update iteration,

where λ is a corresponding penalty factor to achieve spectral smoothness. Similarly, $\mathbf{C_H}$ and γ are the terms related to temporal smoothness. The solution becomes

$$\mathbf{W} = (\mathbf{H}^\top\mathbf{H} + \lambda\mathbf{I})^{-1}(\mathbf{H}^\top\mathbf{V} + \lambda\mathbf{C_W}), \tag{6}$$

$$\mathbf{H}^\top = (\mathbf{W}\mathbf{W}^\top + \gamma\mathbf{I})^{-1}(\mathbf{W}\mathbf{V}^\top + \gamma\mathbf{C}_\mathbf{H}^\top). \tag{7}$$

In our experience, the system described by (6) and (7) requires a smaller number of iterations than NMF to converge. Furthermore, the system also requires a smaller number of frames than NMF to obtain reasonably good results. The 'Experiments' section presents the simulation results of the proposed method and a comparison to other methods.

The proposed method is designed by considering the following issues. Firstly, it is crucial to determine the exact number of templates to obtain reasonably good results. Such an issue is widely discussed in [13] for conventional NMF-based methods. Secondly, it is crucial to determine in which manner the penalty term is set in (5). Finally, matrix inversion consumes substantial computing power, compared to the gradient descent algorithms used in NMF. These problems are discussed in the following section.

Frame-based recursive regularization analysis

Refined update rules

This section presents the high computation complexity problem caused by matrix inversions. The template and intensity matrices can be adaptively learned from the current input frame and some previous input frames. For brevity, we present only the derivation to recursively evaluate the time-varying template matrix $\mathbf{W}$. The derivation of the update rule for the intensity matrix $\mathbf{H}$ is omitted. Let the system start with the first m input frames. m must not be large when the signal varies rapidly. Let the r-by-n template matrix for the lth input frame be denoted by $\mathbf{W}(l)$, which is obtained using frame-l, frame-$(l-1)$, ..., and frame-$(l-m+1)$. $\mathbf{V}(l)$ and $\mathbf{H}(l)$ are subsequently defined as

$$\mathbf{V}(l) = [\,\mathbf{v}(l-m+1)\mathbf{v}(l-m+2)\mathbf{v}(l-m+3)\ldots\mathbf{v}(l)]^\top \tag{8}$$

$$\mathbf{H}(l) = [\,\mathbf{h}(l-m+1)\mathbf{h}(l-m+2)\mathbf{h}(l-m+3)\ldots\mathbf{h}(l)]^\top, \tag{9}$$

where

$$\mathbf{v}(q) = [\,v_{q1}v_{q2}v_{q3}\ldots v_{qr}]^\top \tag{10}$$

$$\mathbf{h}(q) = [\,h_{q1}h_{q2}h_{q3}\ldots h_{qr}]^\top, \tag{11}$$

$l-m+1 \le q \le l$. Hence, the cost function in (5) for frame-l is rewritten as

$$J = \|\mathbf{V}(l) - \mathbf{H}(l)\mathbf{W}(l)\|^2 + \lambda\|\mathbf{W}(l) - \mathbf{C_W}(l)\|^2 + \gamma\|\mathbf{H}(l) - \mathbf{C_H}(l)\|^2. \tag{12}$$

Subsequently, the template matrix is obtained as

$$\mathbf{W}(l) = (\mathbf{H}^\top(l)\mathbf{H}(l) + \lambda I)^{-1}(\mathbf{H}^\top(l)\mathbf{V} + \lambda\mathbf{C_W}(l)). \tag{13}$$

Therefore, to obtain the template matrix and the corresponding intensity values for every new input frame, (6) and (7) must be re-evaluated when a new input frame is added. A recursive procedure is proposed to reduce the number of matrix inversions. First, two new matrices are defined as

$$\mathbf{P}(l) = (\mathbf{H}(l)^\top\mathbf{H}(l) + \lambda I)^{-1} \tag{14}$$

$$\mathbf{R}(l) = \mathbf{H}(l)^\top\mathbf{V}(l) + \lambda\mathbf{C_W}(l), \tag{15}$$

where

$$\begin{aligned}\mathbf{P}(l) &= \left(\left[\mathbf{h}^\top(l-m+1)\ldots\mathbf{h}^\top(l)\right]\times\begin{bmatrix}\mathbf{h}^\top(l-m+1)\\ \mathbf{h}^\top(l-m+2)\\ \vdots\\ \mathbf{h}^\top(l)\end{bmatrix} + \lambda I\right)^{-1}\\ &= \left(\sum_{k=l-m+1}^{l}\mathbf{h}^\top(k)\mathbf{h}(k) + \lambda I\right)^{-1}\end{aligned} \tag{16}$$

$$\begin{aligned}\mathbf{R}(l) &= \left[\mathbf{h}^\top(l-m+1)\ldots\mathbf{h}^\top(l)\right]\times\begin{bmatrix}\mathbf{v}(l-m+1)\\ \mathbf{v}(l-m+2)\\ \vdots\\ \mathbf{h}(l)\end{bmatrix} + \lambda\mathbf{C_W}(l)\\ &= \sum_{k=l-m+1}^{l}\mathbf{h}^\top(k)\mathbf{v}(k) + \lambda\mathbf{C_W}(l).\end{aligned} \tag{17}$$

Therefore, $\mathbf{W}(l) = \mathbf{P}(l)\mathbf{R}(l)$ and the template matrix for frame-$(l+1)$ can also be calculated by $\mathbf{W}(l+1) = \mathbf{P}(l+1)\mathbf{R}(l+1)$.

Similar to (16) and (17), we obtain

$$\begin{aligned}\mathbf{P}(l+1) &= \left(\left[\mathbf{h}^\top(l-m+2)\ldots\mathbf{h}^\top(l+1)\right]\times\begin{bmatrix}\mathbf{h}(l-m+2)\\ \mathbf{h}(l-m+3)\\ \vdots\\ \mathbf{h}(l+1)\end{bmatrix} + \lambda I\right)^{-1}\\ &= \left(\sum_{k=l-m+2}^{l+1}\mathbf{h}^\top(k)\mathbf{h}(k) + \lambda I\right)^{-1}\end{aligned} \tag{18}$$

$$\begin{aligned}\mathbf{R}(l+1) = & \left[\mathbf{h}^{\top}(l-m+2)\ldots\mathbf{h}^{\top}(l+1)\right] \\ & \times \begin{bmatrix} \mathbf{v}(l-m+2) \\ \mathbf{v}(l-m+3) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{v}(l+1) \end{bmatrix} + \lambda \mathbf{C}_{\mathbf{W}}(l+1) \\ = & \sum_{k=l-m+2}^{l+1} \mathbf{h}^{\top}(k)\mathbf{v}(k) + \lambda \mathbf{C}_{\mathbf{W}}(l+1).\end{aligned} \tag{19}$$

We define

$$\begin{aligned}\tilde{\mathbf{P}}(l) = & \left(\left[0\ \mathbf{h}^{\top}(l-m+2)\ldots\mathbf{h}^{\top}(l)\right] \right. \\ & \left. \times \begin{bmatrix} 0 \\ \mathbf{h}(l-m+2) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{h}(l) \end{bmatrix} + \lambda I \right)^{-1} \\ = & \left(\sum_{k=l-m+1}^{l} \mathbf{h}^{\top}(k)\mathbf{h}(k) \right. \\ & \left. + \lambda I - \mathbf{h}^{\top}(l-m+1)\mathbf{h}(l-m+1) \right)^{-1}.\end{aligned} \tag{20}$$

By using the Woodbury matrix identity [27],

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{DA}^{-1}\mathbf{B}+\mathbf{C}^{-1})^{-1}\mathbf{DA}^{-1} \tag{21}$$

holds with $\mathbf{A} = \mathbf{P}(l)^{-1} = \sum_{k=l-m+1}^{l} \mathbf{h}^{\top}(k)\mathbf{h}(k) + \lambda I$, $\mathbf{B} = \mathbf{h}^{\top}(l-m+1)$, $\mathbf{C} = -1$, and $\mathbf{D} = \mathbf{h}(l-m+1)$. Therefore, we obtain $\tilde{\mathbf{P}}(l)$ without matrix inversion by

$$\tilde{\mathbf{P}}(l) = \mathbf{P}(l) - \frac{\mathbf{P}(l)\mathbf{h}^{\top}(l-m+1)\mathbf{h}(l-m+1)\mathbf{P}(l)}{\mathbf{h}(l-m+1)\mathbf{P}(l)\mathbf{h}^{\top}(l-m+1)-1}. \tag{22}$$

Next,

$$\begin{aligned}\mathbf{P}(l+1) = & \left(\sum_{k=l-m+2}^{l} \mathbf{h}^{\top}(k)\mathbf{h}(k) \right. \\ & \left. + \lambda I + \mathbf{h}^{\top}(l+1)\mathbf{h}(l+1) \right)^{-1}.\end{aligned} \tag{23}$$

By using (21) with $\mathbf{A} = \tilde{\mathbf{P}}(l)^{-1} = \sum_{k=l-m+2}^{l} \mathbf{h}^{\top}(k)\mathbf{h}(k) + \lambda I$, $\mathbf{B} = \mathbf{h}^{\top}(l+1)$, $\mathbf{C} = +1$, and $\mathbf{D} = \mathbf{h}(l+1)$, we obtain

$$\mathbf{P}(l+1) = \tilde{\mathbf{P}}(l) - \frac{\tilde{\mathbf{P}}(l)\mathbf{h}^{\top}(l+1)\mathbf{h}(l+1)\tilde{\mathbf{P}}(l)}{\mathbf{h}(l+1)\tilde{\mathbf{P}}(l)\mathbf{h}^{\top}(l+1)+1}. \tag{24}$$

A matrix inversion is unnecessary when $\mathbf{R}(l+1)$ is achieved by

$$\tilde{\mathbf{R}}(l) = \mathbf{R}(l) - \mathbf{h}^{\top}(l-m+1)\mathbf{h}(l-m+1) - \lambda \mathbf{C}_{\mathbf{W}}(l) \tag{25}$$

$$\mathbf{R}(l+1) = \tilde{\mathbf{R}}(l) + \mathbf{h}^{\top}(l+1)\mathbf{h}(l+1) + \lambda \mathbf{C}_{\mathbf{W}}(l+1). \tag{26}$$

The oldest frame is removed by using (22) and (25), and the new input frame is added by using (24) and (26). Hence, the template matrix for each new input frame can be computed recursively without matrix inversion by using the results generated by previous input frames. Since both frequency response and intensity of a note evolve slowly in a short time, $\mathbf{C}_{\mathbf{W}}$ and $\mathbf{C}_{\mathbf{H}}$ are determined

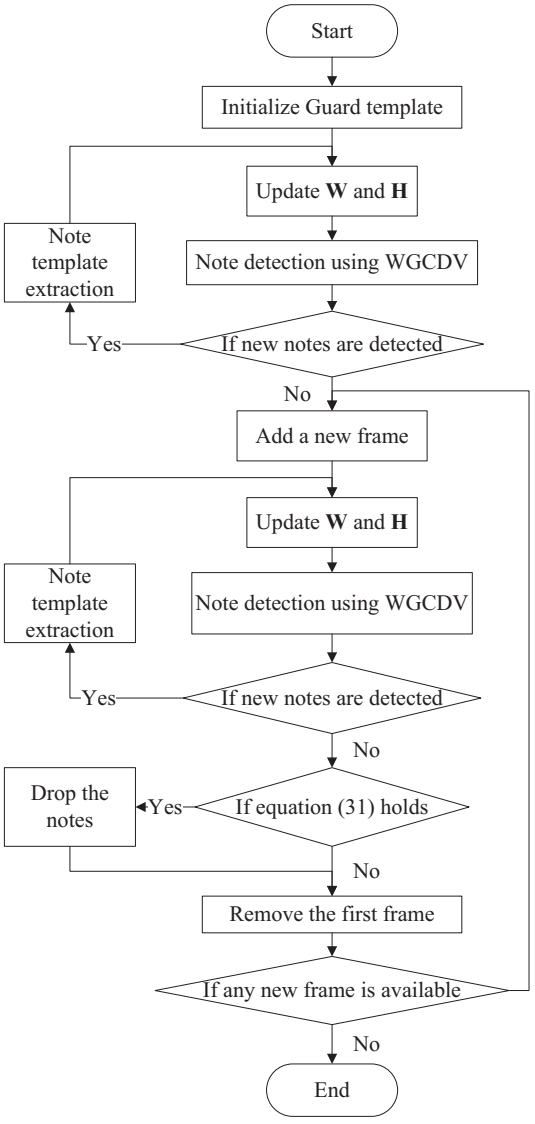

Figure 1 The procedure of the proposed system. The block diagram of the proposed system with the recursive regularization procedure.

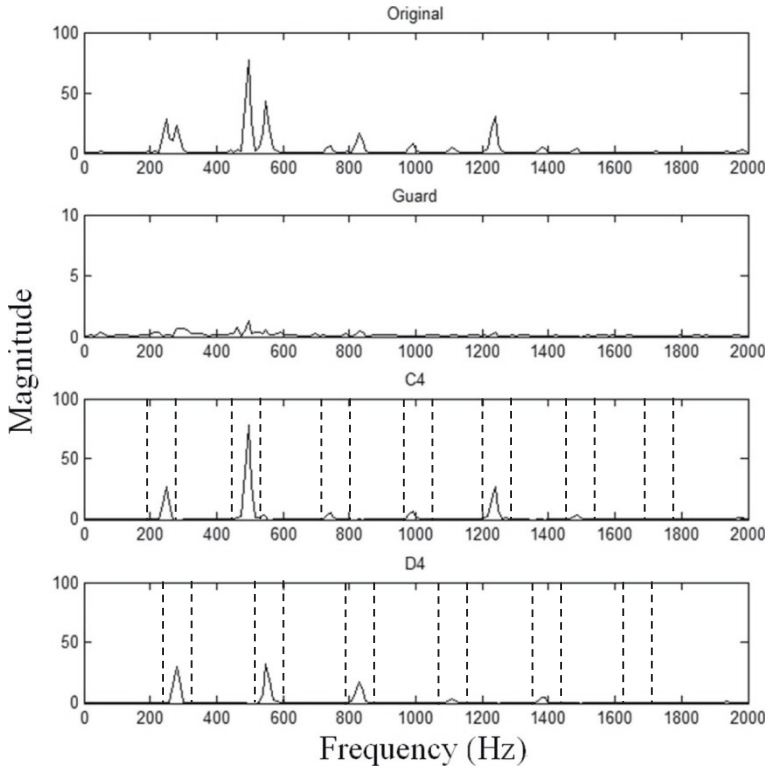

Figure 2 An example of extracted harmonic structures. The original Guard template is separated into C4 and D4 constrained templates and a new Guard template. Sound source: Kuijken's recording on BWV 1005 No. 3 [23].

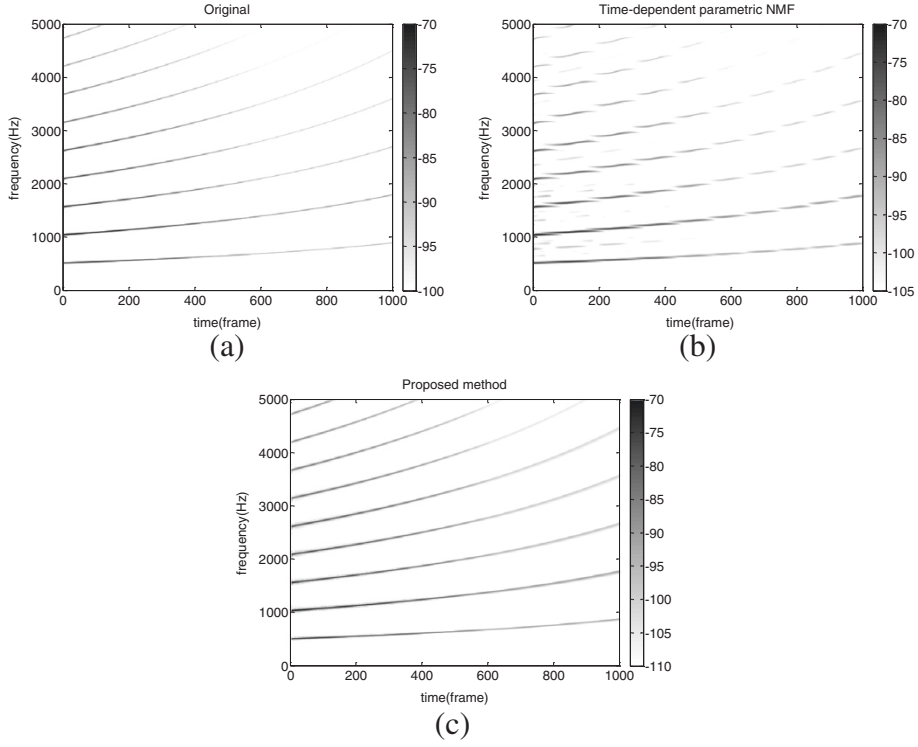

Figure 3 A chirp signal varies from C5 to A5 in 1,000 time frames. (a) The original spectrogram. **(b)** The reconstructed spectrogram of [5]. **(c)** The reconstructed spectrogram of the proposed method.

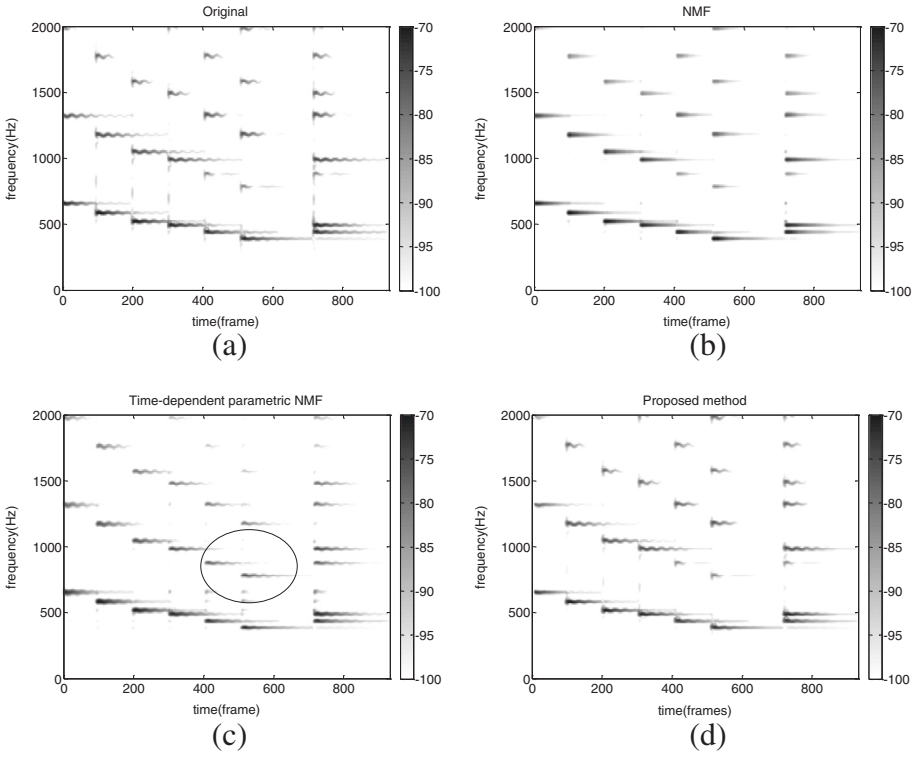

Figure 4 Six vibrating notes presented in the following order: E5, D5, C5, B4, A4, G4, and A4+B4. The spectrograms of **(a)** the original, **(b)** the harmonic-constrained NMF [13], **(c)** time-depend parametric NMF [5], and **(d)** the proposed method of G4, A4, B4, C5, D5, and E5 notes that include vibrato.

by the the results of $\mathbf{W}$ and $\mathbf{H}$ obtained in the previous update iteration, i.e., $\mathbf{C_W}$ is updated by $\mathbf{C_W}(l+1) = \mathbf{W}(l)$.

The intensity matrix is subsequently calculated. Similar to (13), the intensity matrix for the $(l-m+1)$th frame, the $(l-m+2)$th frame, ..., and the lth frame is obtained as

$$\mathbf{H}^\top(l) = \left(\mathbf{W}(l)\mathbf{W}^\top(l) + \gamma I\right)^{-1} (\mathbf{W}(l)\mathbf{V}^\top(l) + \gamma \mathbf{C_H}^\top). \tag{27}$$

Subsequently, the intensity information corresponding to the template matrix of the $(l+1)$th frame is computed as

$$\mathbf{h}^\top(l+1) = \left(\mathbf{W}(l+1)\mathbf{W}^\top(l+1) + \gamma I\right)^{-1} \times \left(\mathbf{W}(l+1)\mathbf{v}^\top(l+1) + \gamma \mathbf{C_H}^\top(l+1)\right). \tag{28}$$

In (28), $\mathbf{C_H}^\top(l+1)$ can be set to $\mathbf{H}^\top(l)$ because it is assumed that the intensity cannot change abruptly. The forgetting factors λ and γ can determine the effects of old frames. They are both set to 100 in this work. The time-varying template matrix and the corresponding intensity matrix can be calculated alternatively when a new input frame is added. Further details and the overall procedure are presented in the following section.

Analysis procedure

As shown in Figure 1, the analysis system started from only one template initialized with random values, which is called Guard template. An initialization loop determines all possible note templates and evaluates the template matrix and the intensity matrix in the same time. After the initialization loop, the main loop takes care of the new frame addition, the new note template detection, the offset note template evaluation, and the old frame removal. For the new note template detection, we used the weighted greatest common divisor and vote (WGCDV) [28] method to detect new tones in Guard template by using a floating point GCD lookup table and a frame-based correction method. WGCDV estimates F0 in three steps: (a) locates the peaks of the frequency response of

Table 1 Objective performance comparison

	SIR	SAR	SDR
NMF with harmonic constraints [13]	26.08	6.19	6.13
Time-dependent parametric NMF [5]	23.06	6.78	6.56
Proposed method	29.08	6.74	6.70

Objective performance comparison of NMF with harmonic constraints [13], time-dependent parametric NMF [5], and the proposed method by using SIR, SAR, and SDR.

Table 2 Transcription results 1

	Accuracy (Acc), %	Precision (*P*), %	Recall (*R*), %	*F*-measure (*F*), %
NMF with harmonic constraints [13]	79.64	86.27	90.97	88.56
Time-dependent parametric NMF [5]	81.94	86.87	93.53	90.08
RR without online scheme [14]	77.45	78.81	97.82	87.29
Proposed method	83.36	97.28	85.35	90.93

Transcription report of the proposed work in the analysis of Kuijken's recording on BWV 1005 No. 3.

the current frame and regards them as possible partials, (b) finds a likely GCD value for each partial pair using a lookup table method, (c) weights the likely GCD values and voting the deterministic GCD according to the spectral energy. Once a harmonic structure is recognized within Guard template at the current frame, these harmonic peaks will be extracted from Guard template and a corresponding new template is added to the original template matrix.

Figure 2 shows an example of extracted harmonic structures of C4 and D4 notes from the original Guard template by using a mask function, *S*. As shown in Figure 2, each mask function represented in a dashed line can be defined by a harmonic set corresponding to the detected note as follows:

$$S_j = \sum_{p=1}^{N} I(pf_j - \epsilon, pf_j + \epsilon), \tag{29}$$

where $I(\alpha, \beta) = 1$ in the interval $[\alpha, \beta]$; otherwise, it is 0. f_j is the fundamental frequency of the jth recognized tone, and p is the partial index. ϵ is set at 3% of the partial frequency, pf_j.

Subsequently, the new template is computed by

$$\mathbf{W}_j^l = \mathbf{S}_j^l \otimes \mathbf{W}_0^l, \tag{30}$$

where $\mathbf{S}_j^l$ is the mask function of frame-l. In (30), $\mathbf{W}_0^l = [\mathbf{W}_{01}^l \mathbf{W}_{02}^l \dots \mathbf{W}_{0n}^l]^T$ represents the original Guard template of frame-l, and $\otimes$ is the element-wise multiplication. The number of templates, r, is equal to $j + 1$. The procedure described in the previous section is performed again for frame-$(l + 1)$ to obtain the new template matrix and intensity matrix.

A re-estimation of the pitch of each note based on the updated template matrix is necessary because all templates, as well as pitches, can vary by frame. Consequently, the mask functions of all templates must be updated. Because each template contains only one harmonic set, $\mathbf{C}_{\mathbf{W}}(l + 1)$ in Equation 13 is computed by $\mathbf{S}(l) \otimes \mathbf{W}(l)$, where $\mathbf{S}(l) = \left[\mathbf{S}_0^l \mathbf{S}_1^l \dots \mathbf{S}_r^l\right]^T$. Based on Equation 5, the iterative update procedure forces $\mathbf{W}(l + 1)$ to retain harmonic structures for all the notes as much as possible, depending on the regularization parameter, λ.

Finally, a musical note is muted eventually after it is played for a while. In this study, a note is removed from the analysis process by removing its template and the corresponding intensity information. A note can be removed if

$$\sum_{q=l}^{l-k} |h_{qi}| \le T, 1 \le i \le r. \tag{31}$$

That is, the ith note is removed after frame-l if Equation 31 holds. T is empirically set to 0.1 in this work. By removing such notes, the computation complexity is also reduced.

Experiments

Database

Two excerpts are generated by a MIDI synthesizer for preliminary tests. Firstly, a synthetic chirp signal is generated, and its pitch varies from C5 to A5 in 1,000 time frames. The second test uses a signal with vibrating notes including six notes in the following order: E5, D5, C5, B4, A4, G4, and A4+B4. A vibrato effect is generated by setting proper MIDI commands. All parameter sets are the same as those in the previous test. Moreover, two recordings, the BWV 1005 No. 3 performed by Kuijken [23] and RWC database

Table 3 Transcription results 2

	Accuracy (Acc), %	Precision (*P*), %	Recall (*R*), %	*F*-measure (*F*), %
NMF with harmonic constraints [13]	56.22	57.09	97.64	71.98
Time-dependent parametric NMF [5]	74.11	75.29	97.94	85.13
RR without online scheme [14]	49.27	50.71	94.93	66.01
Proposed method	76.26	85.99	87.08	86.53

Transcription report of the proposed work in the analysis of RWC database C038.

C038 [24], are used to evaluate the accuracy of all methods, as proposed in [29]. The former contains 587 notes that are manually annotated as the ground truth. The latter contains 1,745 notes whose ground truth is provided from the syncRWC annotations [30].

Parameters

The window size is 4,096 samples, the hop size is 256 samples, and the sampling rate is 44.1 kHz. A Hamming window and 4096-FFT are subsequently applied.

Performance evaluation

An objective measure for evaluating the performance of a source separation method proposed in [31] is adopted for the following discussion. To compare different approaches, the signal-to-distortion ratio (SDR), the signal-to-artifact ratio (SAR), and the signal-to-interference ratio (SIR) are computed with each note as the target. In this work, we considered each note as a separate source. The SIR, SAR, and SDR values of these notes are averaged respectively.

To evaluate the performance of music transcription, we chose note-level metrics with a tolerance of one window before and after the reference onset time. Overall accuracy is defined as

$$\mathrm{Acc} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{FN} + \mathrm{TP}}, \tag{32}$$

where TP (true positives) is the number of correctly transcribed voiced notes (within a quarter tone distance in frequency and 50-ms onset distance in time), FP (false positives) is the number of unvoiced notes transcribed as voiced, and FN (false negatives) is the number of voiced notes transcribed as unvoiced. This measure is defined for note-based onset evaluation in this case and is bounded by 0 and 1, with 1 corresponding to perfect transcription. Another metric, called F-measure, is defined as

$$F = 2\frac{P \cdot R}{P + R}, \tag{33}$$

whereas

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{34}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{35}$$

where P and R represent the precision rate and recall rate, respectively.

Results

The first preliminary test: glissando

The first part of this section presents a comparison of the proposed method by using a glissando signal to the conventional NMF method with harmonic constraints on matrix $\mathbf{W}$ [32] and the time-dependent parametric NMF method [5] that uses the same harmonic structure for different notes of the same instrument. For the time-dependent parametric NMF, 10 templates are used to cover the possible pitch range of the signals. For the proposed method, it started with the initialized Guard template, and the required number of templates is determined adaptively during the analysis process. Figure 3a shows the original spectrogram of the note, and Figure 3b,c shows the reconstructed spectrograms that resulted from the time-dependent NMF method and the proposed method, respectively. In [5], the pitch of a template is constrained to vary within a fixed semitone; therefore, a number of discontinuous effects occurred among the templates, as shown in Figure 3b. Moreover, 10 note templates are required for the time-dependent NMF approach in this case. Because the pitch of a template is allowed to vary freely in the proposed method, only one template is required to represent the chirp signal. This accounts for the superior performance of the proposed method in which the signal varies more than one semitone.

Table 4 Complexity comparison

	Numbers of matrix multiplications	Big-O
Update rules of NMF	$(m-k+1)(2nrk + 2r^2n + 2r^2k + 2rn + 2rk)$	$O(mkrn)$
Proposed update rules	$(m-k)(2r^2n + 9rn + 2r^3 + 15r^2 + 4r)$ $+(2r^2n + 2krn + 4r^3 + (2k+2)r^2)$	$O(mr^2n)$

The complexity of the proposed update rule is compared with the NMF update method.

The second preliminary test: vibrato

In this case, a synthetic signal with six vibrating notes appearing in the following order, E5, D5, C5, B4, A4, G4, and A4+B4, is used. Figure 4 shows the simulation results. Figure 4b shows that the harmonic-constrained NMF is limited in presenting the signal with frequency modulation because of the consistency of the templates. As discussed in [13], separation performance can be improved by increasing the number of templates for each note. The time-dependent parametric NMF performs efficiently in signal that includes vibrato effects; the result is shown in Figure 4c. The only disadvantage is that the amplitudes of the partials are not determined independently

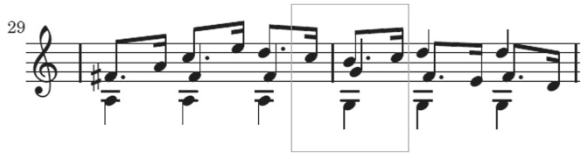

Figure 5 BWV 1005 No. 3. Five notes in the bracket are analyzed.

from time to time, as shown in the circled region of Figure 4c. The amplitude set trends toward the average among all played notes. The spectrogram obtained from the proposed method is closer to the original spectrogram than those obtained from NMF methods. Table 1 shows the separation results of all methods. The time-dependent NMF experiences a number of interference errors from the additional partial energy shown in the circled region of Figure 4c. By considering the SAR, the proposed method preserves most of the vibrato effects in all cases, whereas the harmonic-constrained NMF preserves only steady partial energies. Based on the success of both measures, the proposed method scores the highest SDR as well.

Main experiments

The system is tested on the complete recordings of BWV 1005 No. 3 played by Kuijken [23] and 24 Caprices op. 1 no. 24 in A minor played by Paganini from RWC database [24]. A total of 587 + 1,745 notes are annotated as the ground truth. The analysis is performed blindly; however, pitches outside the possible range are excluded. Additional score information is also excluded from the process.

Table 2 shows the results of analyzing Kuijken's recording. It indicates that 501 notes are correctly detected (TP), 14 detected notes are not in the score (FP), 86 of 587 notes are not detected (FN), 18 notes are missed because their octave notes are too strong, 4 notes are missed because

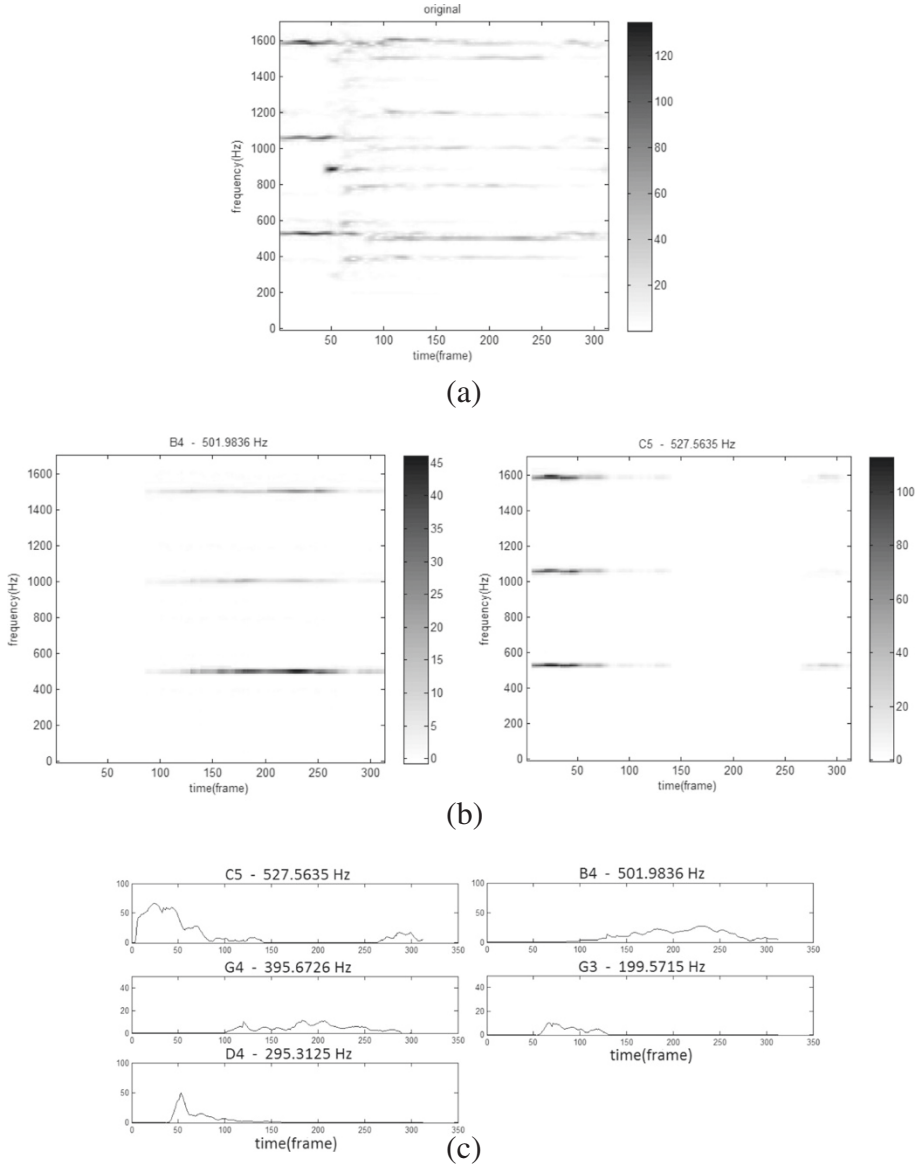

Figure 6 Analysis of Grumiaux's recording. (a) Original spectrogram, **(b)** spectrogram of B4 and C5 notes, and **(c)** intensity matrices.

their perfect fifth notes are too strong, 56 notes remain in Guard templates because the automatic detection method fails to extract them from Guard template, and 8 notes are mixed with other detected notes because their pitches are close to those of certain high intensity notes. Hence, the following results are obtained: precision rate = $501/(501+14) = 97.28\%$, recall rate = $501/(501+86) = 85.35\%$, accuracy = $501/(501+14+86) = 83.36\%$, and F-measure = 90.93%.

The second recording contains more complicated performing styles with larger amount of notes than the first one. False alarms are enormously increased compared with the first recording especially in NMF case, since the energies of overlapping partials of voiced notes and unvoiced notes interfere each other in the intensity matrix $\mathbf{H}$. The performance of the proposed method maintains balance between precision and recall rate. Its number of unvoiced notes transcribed as voiced and voiced notes transcribed as unvoiced is lower than NMF and TD-NMF. That shows our approach is more stable than NMF and TD-NMF.

In Tables 2 and 3, the recursive regularization method without the online scheme means that the sliding frame size is maximized to the whole signal duration. It is

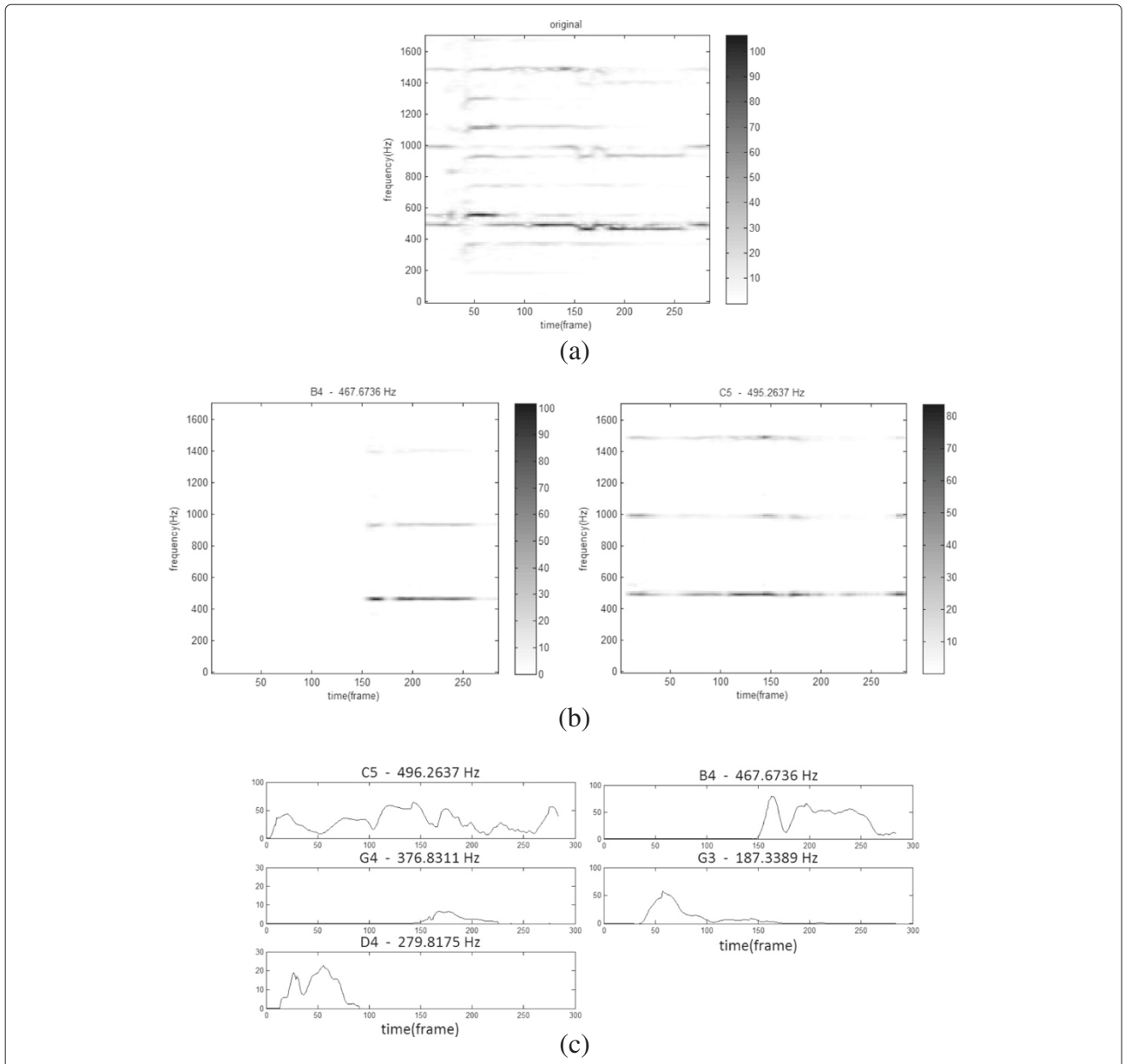

Figure 7 Analysis of Kuijken's recording. (a) Original spectrogram, **(b)** spectrogram of B4 and C5 notes, and **(c)** intensity matrices of all notes.

notable that, according to accuracy and F-measure of the transcription report, the results of this conventional recursive regularization method provides a baseline performance of the proposed algorithm. This work takes advantages of high-speed property of recursive regularization and improves the performance by imposing the proposed online scheme.

Complexity

In addition, the complexity of the proposed update rules is compared with the NMF as shown in Table 4. Let n be the number of frequency bins, m be the number of frames, and r be the number of templates. Suppose m input frames are set to be processed and only k frames are analyzed for each iteration, where $k < m$. For each iteration, the complexities of two update rules are $O(krn)$ and $O(r^2n)$. The complexity of the proposed update rules is similar to traditional NMF update rules in the case of regarding Euclidean distance as the cost function. Under the similar computational complexity, the proposed method is able to handle the time-varying music features and offer better results.

Performance analysis

Here, the proposed method attempts to analyze the performances of three violinists of different generations: Arthur Grumiaux, Sigiswald Kuijken, and Hilary Hahn. The signals are extracted from their CDs [23,25,26]. The following passages excerpted from BWV 1005 No. 3 are used to compare their performances. The musical score is shown in Figure 5. This study analyzes the signals of the short period during five notes in the bracket. The

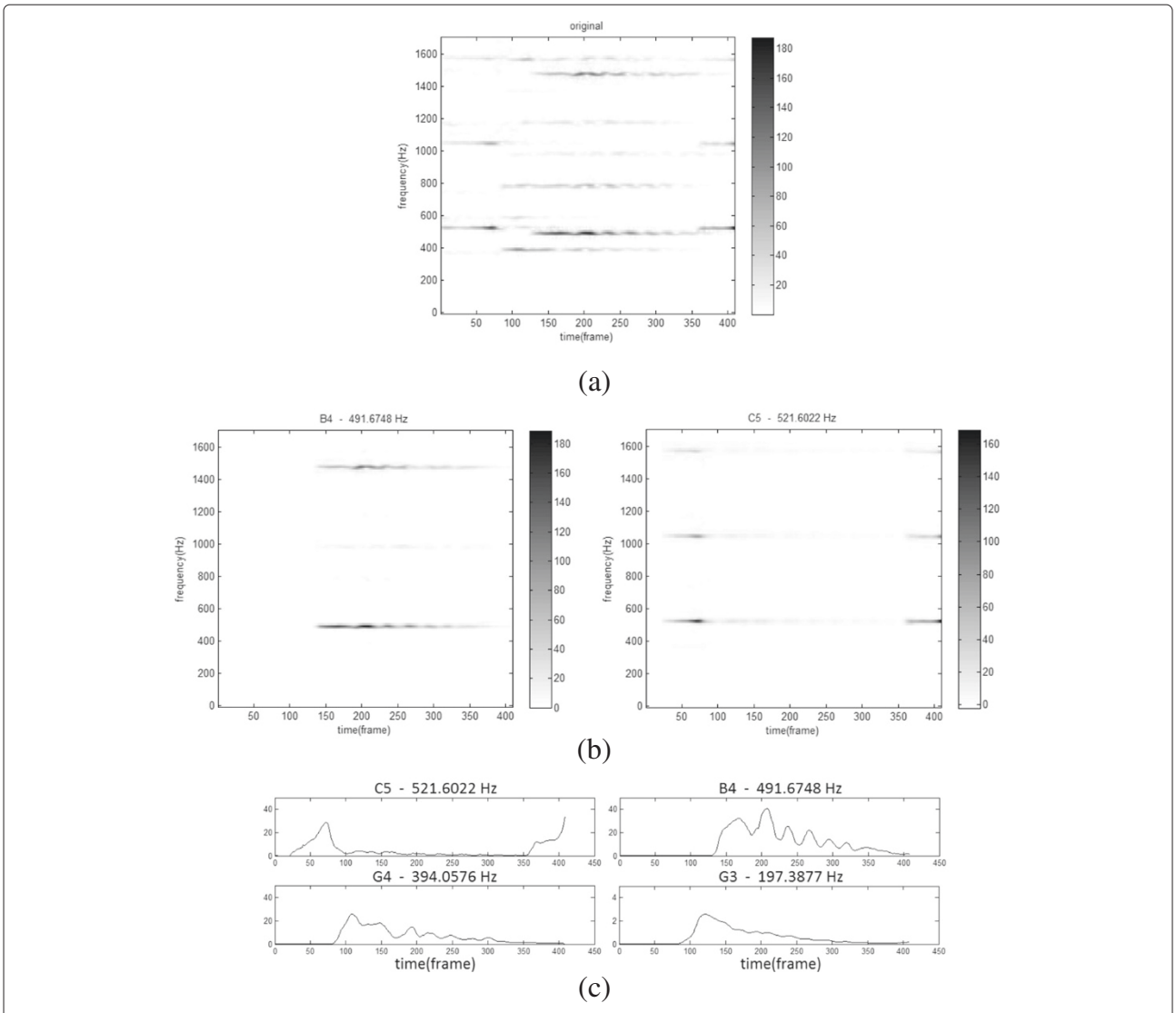

Figure 8 Analysis of Hahn's recording. (a) Original spectrogram, **(b)** spectrogram of B4 and C5 notes, and **(c)** intensity matrices of all notes.

original spectrograms and the analysis results are shown in Figures 6, 7, and 8.

According to the figures, a notable difference is that the pitches of Kuijken's performance are one semitone lower than those of the other violinists' since the instruments are usually used in historically informed performances (HIP), and musicians usually follow the tradition of the period during which the music is composed. Secondly, violinists use vibrato techniques frequently. We can observe the vibrato effects of B4 note in the performance played by Hahn, as shown in Figure 8, compared to those played by Grumiaux and Kuijken, as shown in Figures 6 and 7, respectively. Thirdly, Kuijken's style is distinct. He freely used numerous trills, which were not indicated in the score of Bach's solo violin works. This can be viewed in Figure 7b,c. As shown in Figures 6c and 8c, C5 is off and B4 is on. After B4 continues for a period of time, it is off and C5 is on again. Moreover, C5 is the strongest note in Grumiaux's playing, whereas B4 is the most prominent note in Hahn's playing. Comparatively, Kuijken played equal intensity on these two notes. Hahn's recording sounded brighter than the other two recordings since the spectral energy of the G3 note in Hahn's recording is larger than the others'. This may be caused by the decision of her balance engineer since the lower notes are weak in amplitude in all of her recordings. A notable mistake is also observed, that is, both Kuijken and Grumiaux played an extra D4 note, which is not in the original score. This may be a coincidence, or it is possible that they used a different score edition. Finally, their tempi also differ considerably. Kuijken used 1.68 s (290 frames) to finish the short passage, whereas Grumiaux used 1.83 s, and Hahn used 2.37 s. Hahn's tempo is 40% slower than Kuijken's. As an HIP musician, Kuijken played faster than the other violinists.

Conclusions

A recursive regularization analysis method is proposed to analyze acoustic recordings of solo violin works. Similar to NMF, the proposed method factorizes the matrix formed with the Fourier magnitude coefficients of multiple frames into a template matrix and an intensity matrix. The frame-by-frame-based procedure is designed for time-varying musical signals, such as solo violin recordings. The system of equations is updated by adding a new frame and dropping an old frame to avoid the problems of most NMF methods when the signal varies substantially. The proposed method is compared to the time-dependent NMF method by using two synthesized signals and exhibited superior SDR performances. The objective performance of the proposed method is also verified. For Kuijken's recording of BWV 1005 No. 3, the precision rate is 97.28%, the recall rate is 85.35%, and the F-measure is 90.93%. For a larger recording database from RWC C038, the precision rate is 85.99%, the recall rate is 87.08%, and the F-measure is 86.53%. It shows the stability of our approach. Finally, the proposed method is used to analyze the recordings of J.S. Bach's BWV 1005 No. 3 by three violinists, that is, Arthur Grumiaux, Sigiswald Kuijken, and Hilary Hahn. The results show that the time-varying characteristics of most notes appearing in the recordings can be tracked efficiently. The styles of the three violinists are easily distinguished through the separated results.

We are currently investigating possible approaches to improve the extraction of new notes from Guard template. An octave error may occur in our case because of the overlapping partials of the octave notes. In addition, because of the basis of the least-squares method, the performance of the proposed method with respect to signals of small amplitude, such as higher partials, is not as effective as NMF using other types of cost functions such as KL and IS divergences. The derivation of other cost functions into the proposed method may improve performance. Moreover, a supervised learning procedure can be introduced if note activations are available. Note activations can not only eliminate pitch detection errors but also constrain the intensity matrix for each note. As our approach preserves more musical characteristic details in the note level, nearly perfect decomposition is possible if it incorporates with more constraints, such as timbre, inharmonic bias, and phase. Therefore, many music information retrieval tasks are suitable to take our approach as a preprocessing, for example, player identification or expressive remix.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank the National Science Council, ROC, for its financial support of this work, under contract no. NSC98-2221-E-006-158-MY3.

Received: 17 July 2013 Accepted: 16 May 2014
Published online: 14 June 2014

References

1. DD Lee, HS Seung, Learning the parts of objects by non-negative matrix factorization. Nature. **401**(6755), 788–791 (1999)
2. P Smaragdis, JC 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, 19–22 Oct. 2003), pp. 177–180
3. T Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. Audio Speech Lang. Process. IEEE Trans. on. **15**(3), 1066–1074 (2007)
4. CT Lee, YH Yang, H Chen, Automatic transcription of piano music by sparse representation of magnitude spectra, in *IEEE International Conference on Multimedia and Expo (ICME)* (Taipei, Taiwan, 11–15 July 2011), pp. 1–6
5. R Hennequin, R Badeau, B David, Time-dependent parametric and harmonic templates in non-negative matrix factorization, in *Proc. of the 13th Int. Conference on Digital Audio Effects* (Graz, Austria, 6–10 Sept. 2010)
6. M Nakano, JL Roux, H Kameoka, Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms, in *LVA/ICA'10 Proceedings of the 9th International Conference*

on Latent Variable Analysis and Signal Separation (St. Malo, France, 27–30 Sept. 2010), pp. 149–156
7. R Hennequin, Scale-invariant probabilistic latent component analysis, in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, 16–19 Oct. 2011)
8. P Smaragdis, Relative-pitch tracking of multiple arbitrary sounds. J. Acoust. Soc. Am. **125**, 3406–3413 (2009)
9. MIREX, Music Information Retrieval Evaluation eXchange (MIREX.) http://www.music-ir.org/mirex/wiki/MIREX_HOME. Accessed 19 May 2011
10. N Bertin, R Badeau, E Vincent, Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. Audio Speech Lang. Process. IEEE Trans. on. **18**(3), 538–549 (2010)
11. C Yeh, A Roebel, X Rodet, Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals. Audio Speech Lang. Process. IEEE Trans. on. **18**(6), 1116–1126 (2010)
12. WC Chang, WY Su, C Yeh, A Roebel, 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, 1–4 Sept. 2008)
13. TM Wang, YL Chen, WH Liao, A Su, Analysis and trans-synthesis of acoustic bowed-string instrument recordings–a case study using Bach cello suites, in *International Conference on Digital Audio Effects (DAFX)* (IRCAM, Paris, France, 19–23 Sept. 2011)
14. M Unser, A Aldroubi, M Eden, Recursive regularization filters: design, properties, and applications. IEEE Trans. on Pattern Anal. Mach. Intell. **13**(3), 272–277 (1991)
15. F Nesta, P Svaizer, M Omologo, Convolutive BSS of short mixtures by ICA recursively regularized across frequencies. Audio Speech Lang. Process. IEEE Trans. on. **19**(3), 624–639 (2011)
16. SP Kim, WY Su, Recursive high-resolution reconstruction of blurred multiframe images. Image Process. IEEE Trans. on. **2**(4), 534–539 (1993)
17. J Mairal, F Bach, J Ponce, G Sapiro, Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. **11**, 19–60 (2010)
18. F Wang, C Tan, AC König, P Li, Efficient document clustering via online nonnegative matrix factorizations, in *Proc. SIAM , Hilton Phoenix East/Mesa* (Mesa, USA, 28–30 Apr 2011)
19. Z Duan, G Mysore, P Smaragdis, Online PLCA for real-time semi-supervised source separation. Latent Variable Anal. Signal Sep. **7191**, 34–41 (2012)
20. A Lefevre, F Bach, C Févotte, Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence, in *Proc. WASPAA* (IEEE, New Paltz, NY, 16–19 Oct. 2011), pp. 313–316
21. E Vincent, N Berlin, R Badeau, Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription, in *Proc. of International Conference on Acoustics, Speech and Signal Processing* (IEEE, Las Vegas, Nevada, USA, 2008), pp. 109–112
22. PO Hoyer, Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**, 1457–1469 (2004)
23. S Kuijken, *Bach: Sonatas & Partitas, BWV 1001-1006, CD 1* (Deutsche Harmonia Mundi, 1990)
24. M GOTO, RWC Music Database. http://staff.aist.go.jp/m.goto/RWC-MDB/. Accessed 13–17 Oct. 2002
25. A Grumiaux, *Bach: Sonatas & Partitas (BWV 1001-1006), CD 1* (Philips, 2006)
26. H Hahn, *Hilary Hahn plays Bach* (Sony, 1997)
27. MA Woodbury, Inverting modified matrices. Memorandum Rep. **42**, 106 (1950)
28. YS Siao, WC Chang, WY Su, Pitch detection/tracking strategy for musical recordings of solo bowed-string and wind instruments. J. Inf. Sci. Eng. **25**(4), 1239–1253 (2009)
29. S Dixon, On the computer recognition of solo piano music, in *Proceedings of Australasian Computer Music Conference* (Brisbane, Australia, 17 Jul 2011), pp. 31–37
30. M GOTO, Music synchronization for RWC Music Database (classical music). http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/SyncRWC/. Accessed 24 Sept. 2010
31. E Vincent, R Gribonval, C Févotte, Performance measurement in blind audio source separation. IEEE Trans. on Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
32. S Ewert, M Müller, Using score-informed constraints for NMF-based source separation, in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Kyoto, Japan, 25–30 Mar 2012)

doi:10.1186/s13636-014-0025-6
Cite this article as: Lin *et al.*: **Musical note analysis of solo violin recordings using recursive regularization.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:25.