REDUCED-RANK MODELING OF TIME-VARYING SPECTRAL PATTERNS FOR SUPERVISED SOURCE SEPARATION

Tomonori Fujiwara, Masao Yamagishi, and Isao Yamada

Dept. of Communications and Computer Engineering, Tokyo Institute of Technology, Japan
E-mail: {fujiwara, myamagi, isao}@sp.ce.titech.ac.jp

ABSTRACT

In this paper, we propose a new modeling technique of signals having time-varying spectral patterns for supervised source separation. Typical examples of such signals are instrumental sounds having several segments such as "attack" and "sustain". In the proposed technique, a given signal is modeled as a linear combination of multiple bases which are obtained by using reduced-rank representation of the given signal, where the number of bases is determined automatically. The proposed technique is used to generate the basis matrix in the context of supervised source separation, which improves conventional source separation methods.

***Index Terms*—** low-rank approximation, nonnegative matrix factorization, automatic transcription

1. INTRODUCTION

Source separation is a technique to extract the underlying source signals from a given linear mixture. It is ubiquitous in several applications including array processing, medical image processing, audio signal processing, and so on. A well-known strategy to realize source separation is nonnegative matrix factorization (NMF) [1], [2]. The NMF aims to decompose a given data matrix (generated from the linear mixture) into the product of two nonnegative factor matrices[1] to determine source signals. In practice, since the basis matrix can be constructed by prior information, e.g., instrumental sounds, supervised NMF (SNMF) has been also applied to source separation [3]–[7]. That is, from the data matrix and the basis matrix, the SNMF attempts to estimate the activation matrix. In addition, numerical results in [5] suggest that the SNMF achieves superior performances to the NMF if the basis matrix is constructed suitably, which implies that design of the basis matrix is significant in the SNMF.

In fact, such an issue has been studied, e.g., in the context of sound source separation of polyphonic music [5]–[9]. A common technique to design the basis matrix is realized by extracting a single representative vector from the monophonic magnitude-spectrogram of each instrumental sound, which is embodied essentially by finding the best rank-one approximation of the given spectrogram [5]. Recently, [7]–[9] adopt multiple representative vectors for dealing with time-varying spectral patterns of monophonic spectrogram, due to instrumental sounds of having multiple segments such as "attack" and "sustain" (see Fig. 1), which naturally leads to the selection problem of the number of representative vectors for each instrumental sound.

In this paper, we tackle the selection problem by introducing a novel design of the basis matrix with use of reduced-rank representation of the monophonic spectrogram matrix. Our idea is to reformulate capturing time-varying spectral patterns of the spectrogram matrix into the problem to find a low-rank approximation of the spectrogram matrix (i.e., finding a small number of vectors that is sufficient to accurately describe all the column vectors of the spectrogram matrix). Then, the representative vectors can be obtained from column vectors of the low-rank approximation with discarding negative components.

In addition, after giving examples of iterative algorithms for the SNMF, we apply the proposed basis matrix to automatic transcription of polyphonic music based on the SNMF. A numerical example demonstrates that the proposed technique selects an appropriate number of representative vectors and achieves superior estimation performance to conventional design of the basis matrix, especially in the "attack" segment of each instrumental sound.

(a) (b) (c)

Fig. 1. (a) Time-varying spectral patterns in the constant-Q transform (CQT) data of piano A2. The horizontal axis represents time; (b) Representative vectors learned by the proposed method; (c) The standard basis vector (rank-one approximation).

2. PRELIMINARIES

Let $\mathbb{R}$ and $\mathbb{R}_+$ be the sets of all real numbers and non-negative real numbers, respectively.

Non-negative matrix factorization (NMF) is a problem to decompose a given matrix $\boldsymbol{X} \in \mathbb{R}_+^{M\times D}$ into a basis matrix $\boldsymbol{B} \in \mathbb{R}_+^{M\times K}$ and an activation matrix $\boldsymbol{W} \in \mathbb{R}_+^{K\times D}$ such that

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{W} + \boldsymbol{N}, \tag{1}$$

where $\boldsymbol{N} \in \mathbb{R}^{M\times D}$ is a noise matrix. It is usually assumed that $K < M$. As an application of the NMF, Smaragdis and Brown have proposed automatic transcription with the NMF [1]. In their setting, $\boldsymbol{X}$ is a magnitude-spectrogram matrix of a given time signal $x(t)$ $(t = 0, 1, \ldots, N_x)$:

$$\boldsymbol{X} := \begin{pmatrix} |X(0,0)| & |X(0,1)| & \cdots & |X(0,D-1)| \\ |X(1,0)| & |X(1,1)| & \cdots & |X(1,D-1)| \\ \vdots & \vdots & \vdots & \vdots \\ |X(M-1,0)| & |X(M-1,1)| & \cdots & |X(M-1,D-1)| \end{pmatrix},$$

where $X(f,n)$ is the time-frequency domain representation of $x(t)$ such as the short time Fourier transform (STFT) and constant-Q

This work was supported by JSPS Grants-in-Aid (26730128).

[1]We refer to the two matrices as *basis matrix* and *activation matrix*.

transform (CQT)[2] . Each column vector of the basis matrix $\boldsymbol{B}$ represents an instrumental sound, and the activation matrix $\boldsymbol{W}$ contains temporal information on notes. These two matrices are utilized to transcribe polyphonic music sound.

Recently, supervised NMF (SNMF) has been proposed in consideration that the basis matrix can be learned a priori [3]–[7]. Since the SNMF assumes that the basis matrix $\boldsymbol{B}$ in (1) is known, the SNMF is an estimation problem of the activation matrix $\boldsymbol{W}$ from the given matrix $\boldsymbol{X}$ and $\boldsymbol{B}$. In applications of the SNMF to automatic transcription, each vector[3] of the basis matrix is learned from a monophonic magnitude-spectrogram matrix.[4] Typically, the basis vector is learned by solving a rank-one approximation problem [5]:

$$\underset{\boldsymbol{b}^{\mathrm{mono}}\in\mathbb{R}_+^M,\boldsymbol{w}^{\mathrm{mono}}\in\mathbb{R}_+^{D_1}}{\text{minimize}} \|\boldsymbol{X}^{\mathrm{mono}}-\boldsymbol{b}^{\mathrm{mono}}(\boldsymbol{w}^{\mathrm{mono}})^\top\|_F, \quad (2)$$

where $\boldsymbol{X}^{\mathrm{mono}}\in\mathbb{R}_+^{M\times D_1}$ is a monophonic magnitude-spectrogram, and $\boldsymbol{b}^{\mathrm{mono}}$ is the basis vector. However, the basis vector selected by (2) is not enough to approximate the instrumental sound if $\boldsymbol{X}^{\mathrm{mono}}$ has time-varying spectral pattern, or the rank of $\boldsymbol{X}^{\mathrm{mono}}$ is large.

3. PROPOSED SCHEME

We propose a scheme to generate the basis matrix to describe time-varying spectral patterns efficiently and in detail. In our scheme, each monophonic magnitude-spectrogram matrix is separately utilized to generate basis vectors, so that we focus on the process involving a single monophonic magnitude-spectrogram matrix. Unlike the standard generation process, we utilize multiple basis vectors to describe the monophonic magnitude-spectrogram matrix. To achieve our goal, we extend the typical basis learning problem in (2) to a reduced-rank approximation problem. More precisely, we attempt to find a basis matrix having minimum number r_{**} of column vectors over all candidate basis matrices $\widetilde{\boldsymbol{B}}^{\mathrm{mono}}\in\mathbb{R}_+^{M\times r}$ of enough approximation precision, i.e., $\Upsilon(\widetilde{\boldsymbol{B}}^{\mathrm{mono}})\le\epsilon$ with a predefined precision parameter $\epsilon>0$, where[5]

$$\Upsilon(\widetilde{\boldsymbol{B}}^{\mathrm{mono}}) := \min_{\widetilde{\boldsymbol{W}}^{\mathrm{mono}}\in\mathbb{R}_+^{r\times D_1}} \|\boldsymbol{X}^{\mathrm{mono}}-\widetilde{\boldsymbol{B}}^{\mathrm{mono}}\widetilde{\boldsymbol{W}}^{\mathrm{mono}}\|_F. \quad (3)$$

Considering that (3) is difficult to solve due to nonnegativity constraints, we propose a strategy to generate an approximate solution. The following observation leads to a guideline to generate an approximate solution: the rank-r approximation of $\boldsymbol{X}^{\mathrm{mono}}$ via the singular value decomposition (SVD) tends to have only few negative components. Hence, by ignoring the nonnegativity constraints in (3), we generate the approximate solution through three steps: (i) we determine the number of column vectors, say r_*, via the SVD of $\boldsymbol{X}^{\mathrm{mono}}$ (see Remark 1); (ii) we form a set of vectors of which the span is identical to the column space of the rank-r_* approximation of $\boldsymbol{X}^{\mathrm{mono}}$ as candidate basis vectors (see a thumb rule in Remark 2); (iii) we discard negative components of the candidate basis vectors.

Finally, we collect all the generated basis vectors, for given monophonic spectrogram matrices, into the basis matrix, i.e., $\boldsymbol{B}=[\boldsymbol{B}^{(1)},\dots,\boldsymbol{B}^{(I)}]$, where I is the number of instrumental sounds, and $\boldsymbol{B}^{(i)}\in\mathbb{R}^{M\times n_i}$ consists of basis vectors of ith instrumental sound (n_i corresponds to r_* for ith instrumental sound).

[2]The CQT has a log-frequency resolution, while the STFT does linear-frequency, which facilitates exploiting "constant pattern" structures of harmonic frequency components independent of their fundamental frequency.

[3]We call each column vector of the basis matrix as *basis vector*.

[4]Practically, a monophonic music signal is available (MIDI or [10]).

[5]The Frobenius norm is defined by $\|\boldsymbol{X}\|_F := \sqrt{\sum_{i=1}^M\sum_{j=1}^N x_{i,j}^2}$, where $x_{i,j}$ is the (i,j)-th entry of $\boldsymbol{X}\in\mathbb{R}^{M\times N}$.

Algorithm 1 selection of $\{\tilde{\boldsymbol{b}}_{\ell_1},\tilde{\boldsymbol{b}}_{\ell_2},\dots,\tilde{\boldsymbol{b}}_{\ell_{r_*}}\}$

input: the principal left singular vector $\boldsymbol{u}_1$,
the set $\mathcal{X}_{r_*}^{\mathrm{mono}} := \{(\boldsymbol{X}_{r_*}^{\mathrm{mono}})_{:,1},\dots,(\boldsymbol{X}_{r_*}^{\mathrm{mono}})_{:,D_1}\}$ of column vectors of $\boldsymbol{X}_{r_*}^{\mathrm{mono}}$.

$\tilde{\boldsymbol{b}}_{\ell_1}\leftarrow\boldsymbol{u}_1$
for $i=2$ to r_* **do**
  $\tilde{\boldsymbol{b}}_{\ell_i}\leftarrow\underset{\boldsymbol{x}\in\mathcal{X}_{r_*}^{\mathrm{mono}}}{\arg\min}\,|\frac{\pi}{2}-\theta(P_{\mathfrak{B}_i}(\boldsymbol{x}),\boldsymbol{x})|$
    where $\mathfrak{B}_i=\mathrm{span}\{\tilde{\boldsymbol{b}}_{\ell_1},\dots,\tilde{\boldsymbol{b}}_{\ell_{i-1}}\}$
end for
output: $\{\tilde{\boldsymbol{b}}_{\ell_1},\tilde{\boldsymbol{b}}_{\ell_2},\dots,\tilde{\boldsymbol{b}}_{\ell_{r_*}}\}$

Remark 1: (Detailed description of the first step) In the first step, we determine r_* with the SVD of $\boldsymbol{X}^{\mathrm{mono}}$, say $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. For our observation, a lower bound of the minimum of the criterion Υ,

$$\begin{aligned}\min_{\widetilde{\boldsymbol{B}}_r^{\mathrm{mono}}\in\mathbb{R}_+^{M\times r}}\Upsilon(\widetilde{\boldsymbol{B}}_r^{\mathrm{mono}}) &\ge \min_{\substack{\widetilde{\boldsymbol{B}}^{\mathrm{mono}}\in\mathbb{R}^{M\times r},\\ \widetilde{\boldsymbol{W}}^{\mathrm{mono}}\in\mathbb{R}^{r\times D_1}}}\|\boldsymbol{X}^{\mathrm{mono}}-\widetilde{\boldsymbol{B}}^{\mathrm{mono}}\widetilde{\boldsymbol{W}}^{\mathrm{mono}}\|_F\\ &= \|\boldsymbol{X}^{\mathrm{mono}}-\boldsymbol{X}_r^{\mathrm{mono}}\|_F\\ &= \|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_r\|_F,\end{aligned}$$

is expected to be tight, where the first equation holds provided by [11], and $\boldsymbol{X}_r^{\mathrm{mono}}$ is the best rank-r approximation of $\boldsymbol{X}^{\mathrm{mono}}$ obtained by discarding all singular values except first r values (i.e., $\boldsymbol{X}_r^{\mathrm{mono}}=\boldsymbol{U}\boldsymbol{\Sigma}_r\boldsymbol{V}^\top$, where $\boldsymbol{\Sigma}_r$ is a truncation of $\boldsymbol{\Sigma}$). Hence,

$$r_*=\min\{r\in\{1,2,\dots,M\}\mid\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}_r\|_F\le\epsilon\}$$

is adopted as the number of column vectors.

Remark 2: (An angle-based greedy selection rule of column vectors) As an implementation of the second step, we form $\{\tilde{\boldsymbol{b}}_{\ell_1},\dots,\tilde{\boldsymbol{b}}_{\ell_{r_*}}\}$ by gathering the principal left singular vector $\boldsymbol{u}_1$ and (r_*-1) column vectors of the rank-r_* approximation $\boldsymbol{X}_{r_*}^{\mathrm{mono}}$ in a way to greedily minimize their correlation in the sense of angles[6] (i.e., orthogonal is desired). First, we set[7] $\tilde{\boldsymbol{b}}_{\ell_1}=\boldsymbol{u}_1\in\mathbb{R}_+^M$. Then for any $i=2,3,\dots,r_*$, we select ℓ_i s.t. $\boldsymbol{b}_{\ell_i}$ minimizes the correlation with $\mathfrak{B}_i:=\mathrm{span}\{\tilde{\boldsymbol{b}}_{\ell_1},\tilde{\boldsymbol{b}}_{\ell_2},\cdots,\tilde{\boldsymbol{b}}_{\ell_{i-1}}\}$ (see Algorithm 1).[8] Finally, we normalize $\tilde{\boldsymbol{b}}_{\ell_1},\cdots,\tilde{\boldsymbol{b}}_{\ell_{r_*}}$. Note that, thanks to our observation, the selected column vectors tend to have few negative components.

4. APPLICATION TO SNMF

We apply the proposed scheme to generate the basis matrix in the SNMF problem. Although most iterative algorithms for the SNMF problem are applicable directly with our design of the basis matrix, some of their ideas should be carefully extended. Here we exemplify it in the case of *sound volume smoothness*, considering that sound volume usually changes smoothly: inspired by [6], we introduce iterative algorithms for the SNMF problem, which can exploit the sparseness of the activation matrix and smoothness of the sound volume for improving performance. Note that we follow the idea

[6]Define inner product and induced norm as $\langle\boldsymbol{x},\boldsymbol{y}\rangle:=\boldsymbol{x}^\top\boldsymbol{y}$ and $\|\boldsymbol{x}\|:=\sqrt{\langle\boldsymbol{x},\boldsymbol{x}\rangle}$ ($\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$) respectively. We define an angle $\theta(\boldsymbol{x}_j,\boldsymbol{x}_k)$ between $\boldsymbol{x}_j$ and $\boldsymbol{x}_k$ as $\theta(\boldsymbol{x}_j,\boldsymbol{x}_k):=\arccos\left(\frac{\langle\boldsymbol{x}_j,\boldsymbol{x}_k\rangle}{\|\boldsymbol{x}_j\|\|\boldsymbol{x}_k\|}\right)$.

[7]The nonnegativity of $\boldsymbol{u}_1$ is guaranteed by [12, Theorem 8.3.1].

[8]The metric projection $P_{\mathfrak{B}_i}(\boldsymbol{x})$ of $\mathbb{R}^M$ onto $\mathfrak{B}_i$ is defined by $P_{\mathfrak{B}_i}(\boldsymbol{x}):=\arg\min_{\boldsymbol{y}\in\mathfrak{B}_i}\|\boldsymbol{y}-\boldsymbol{x}\|$.

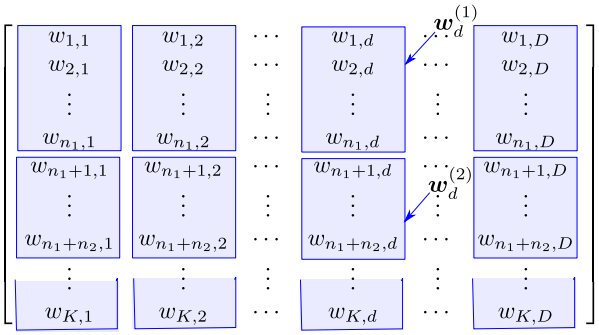

Fig. 2. Partition of $\boldsymbol{W}$ with respect to instrumental sounds. In our setting, multiple basis vectors for each instrumental sound can be arranged to adjoin each other, so that their corresponding activation coefficients are in consecutive rows.

of derivation of iterative algorithms in [6]:[9] we design convex optimization criteria and then, applying a suitable convex optimization technique (in this paper, the alternating direction method of multipliers (ADMM) [15] is adopted), derive iterative algorithms for the SNMF problem.

The convex optimization problem of our interest is as follows:

$$\underset{\boldsymbol{W}\in\mathbb{R}_+^{K\times D}}{\text{minimize}}\ \|\boldsymbol{X}-\boldsymbol{B}\boldsymbol{W}\|_F^2+\lambda_1\|\boldsymbol{W}\|_{\ell_1}+\lambda_2\psi(\boldsymbol{W}), \qquad (4)$$

where $\lambda_1,\lambda_2\in\mathbb{R}_+$ are regularization parameters. The first term maintains data fidelity, and the second term promotes the sparsity with the ℓ_1 norm defined by

$$\|\boldsymbol{W}\|_{\ell_1}:=\sum_{k=1}^{K}\sum_{d=1}^{D}|w_{k,d}|,$$

which is a translation of the fact that few instrumental sounds occur simultaneously (see the use of this prior information in [13], [17]). The third term encourages the smoothness of the volume of each instrument sound. In our setting, since multiple basis vectors are utilized for monophonic spectrogram, sound volume should be related to multiple entries of the activation matrix to determine the volume of each instrumental sound.[10] Here we adopt a simple definition of a sound volume $v_d^{(i)}$ of the ith instrumental sound at time d: define

$$v_d^{(i)}:=\langle\boldsymbol{w}_d^{(i)},\mathbf{1}_{n_i}\rangle, \qquad (5)$$

as the sum of all the activation coefficients corresponding the ith instrumental sound, say $\boldsymbol{w}_d^{(i)}\in\mathbb{R}_+^{n_i}$, where n_i is the number of entries which correspond to the ith instrumental sound at time d, and $\mathbf{1}_{n_i}\in\mathbb{R}^{n_i}$ is a vector whose all entries are 1 (see Fig. 2 for a partition of $\boldsymbol{w}_d^{(i)}$).[11] By using this definition of the sound volume,

[9]We adopt the idea to exploit prior information in a way similar in conventional algorithms: Many algorithms utilize prior information by carefully designing their implicit/explicit optimization criterion. For example, the activation smoothness is encouraged with the Total Variation (TV) $\sum_{k=1}^{K}\sum_{d=1}^{D-1}|w_{k,d+1}-w_{k,d}|$ in [13] and with the so-called Tikhonov-type regularization $\sum_{k=1}^{K}\sum_{d=1}^{D-1}(w_{k,d+1}-w_{k,d})^2$ in [6], [14].

[10]In a typical situation (where a single basis vector is utilized for monophonic spectrogram), sound volume smoothness is translated as the activation smoothness because sound volume relates to activation coefficients directly.

[11]We can also consider the sum of magnitude-spectrogram as the volume of ith instrumental sound at time d, i.e., $v_d^{(i)}:=\langle\boldsymbol{B}^{(i)}\boldsymbol{w}_d^{(i)},\ \mathbf{1}_M\rangle$.

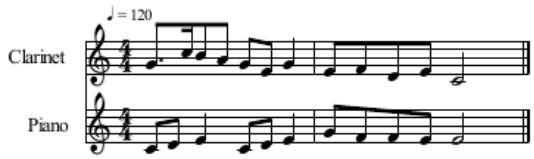

Fig. 3. Score excerpted from [18]: For simplicity, two instruments are utilized in our experiment while four instruments are utilized in [18].

we introduce two variations of the volume-smoothness terms as the sum of time change of volume $v_d^{(i)}$

$$\psi_1(\boldsymbol{W}):=\sum_{i=1}^{I}\alpha_i\sum_{d=1}^{D-1}|v_{d+1}^{(i)}-v_d^{(i)}|, \qquad (6)$$

$$\psi_2(\boldsymbol{W}):=\sum_{i=1}^{I}\alpha_i\sum_{d=1}^{D-1}(v_{d+1}^{(i)}-v_d^{(i)})^2, \qquad (7)$$

where $\alpha_i>0$ is a weight for the ith instrumental sound. Finally, we apply the ADMM to solve (4) (see Appendix for detail).

5. NUMERICAL EXPERIMENTS

We show the efficacy of the proposed design of the basis matrix for the SNMF problem in the context of the automatic transcription of polyphonic music by comparing the proposed basis matrix with the standard one.[12] First, we generate the magnitude-spectrograms of the monophonic signals $\boldsymbol{X}^{\mathrm{mono}}$ and the mixture signal $\boldsymbol{X}$ using MIDI. Monophonic signals of two instruments (Clarinet and Piano) are generated from MIDI signals, where 25 tones per instrument are utilized. The mixture signal is a MIDI signal according to the score in Fig. 3 with additive Gaussian noise of 5dB. All the audio signals are sampled at 44.1k Hz, and their spectrograms are calculated through STFT/CQT[13]. Second, we generate the basis matrix from monophonic magnitude-spectrogram matrices by using one of the two learning techniques, i.e., Algorithm 1 ($\epsilon=\|\boldsymbol{X}^{\mathrm{mono}}\|_F/10$) and the standard rank-one approximation in (2). Third, for solving the SNMF problem with the given spectrogram $\boldsymbol{X}$ and the basis matrix $\boldsymbol{B}$, we apply the ADMM to (4) with the volume smoothness term (6) or (7).[14] In (6) and (7), the uniform weights, $\alpha_i=1$ for all i, are employed, and the parameters of all the iterative algorithms are chosen in such the way that the performance is the best in our experiments. Finally, using the resulting activation matrix $\hat{\boldsymbol{W}}$, we generate an estimated score $T(\boldsymbol{A}\hat{\boldsymbol{W}})$ by thresholding the volume of $\hat{\boldsymbol{W}}$, where T represents the component-wise thresholding operation with the level $\tau=\max(\boldsymbol{X})/5$, $\max(\boldsymbol{X})$ is the largest entry in $\boldsymbol{X}$, and $\boldsymbol{A}\in\mathbb{R}^{I\times K}$ represents the mapping from an activation matrix $\boldsymbol{W}$ to a matrix consisting of their sound volumes, i.e., the (i,d)-th component of $\boldsymbol{A}\boldsymbol{W}$ is $(\boldsymbol{A}\boldsymbol{W})_{i,d}=v_d^{(i)}$.

For performance comparison of basis matrix generation in the second step, we evaluate resulting transcription performances in two criteria, $\|T(\boldsymbol{A}\hat{\boldsymbol{W}})-T(\boldsymbol{A}\boldsymbol{W}^*)\|_F^2$ and MIREX [16], where $T(\boldsymbol{A}\boldsymbol{W}^*)$ is the ground-truth of the score. We also compute them

[12]We already confirm advancement of the proposed scheme compared with [9]. The results and further comparisons will be discussed elsewhere.

[13]The STFT is computed using a Hamming window that is 46.4 ms long with a 23.2 ms overlap. The CQT is performed by CQT toolbox [19] with 24 bins/octave.

[14]Note that the ADMM to (4) with (6) or (7) are supervised versions of [13] and [14] when the basis matrix is learned by the standard way (2).

Table 1. Performance comparisons averaged over 5 trials: the smaller $\|T(\boldsymbol{A}\hat{\boldsymbol{W}}) - T(\boldsymbol{A}\boldsymbol{W}^*)\|_F^2$ and the larger F-measure are preferred. For Standard (2), the ADMM with (6) and (7) are supervised versions of [13] and [14], respectively.

Domain	Basis matrix	Iterative solver for (4)	$\|T(\boldsymbol{A}\hat{\boldsymbol{W}}) - T(\boldsymbol{A}\boldsymbol{W}^*)\|_F^2$ overall	$\|T(\boldsymbol{A}\hat{\boldsymbol{W}}) - T(\boldsymbol{A}\boldsymbol{W}^*)\|_F^2$ attack	F-measure of MIREX [16] overall	F-measure of MIREX [16] attack
STFT	Algorithm 1	ADMM with (6)	**101.0**	**42.2**	**0.8428**	**0.5589**
	Standard (2)	ADMM with (6)	102.4	46.6	0.8425	0.5422
	Algorithm 1	ADMM with (7)	108.2	45.2	0.8305	0.5169
	Standard (2)	ADMM with (7)	110.6	51.0	0.8293	0.4871
CQT	Algorithm 1	ADMM with (6)	**241.4**	69.0	0.9182	0.8269
	Standard (2)	ADMM with (6)	268.8	87.6	0.9080	0.7689
	Algorithm 1	ADMM with (7)	**241.4**	**68.4**	**0.9185**	**0.8285**
	Standard (2)	ADMM with (7)	285.4	98.6	0.9026	0.7373

(a) Ground-truth.

(b) Proposed: Algorithm 1.

(c) Conventional: Standard (2).

Fig. 4. Estimated score around the last two notes of the first measure (i.e., the 1st E3 and its next note of clarinet as well as the 2nd D3 and its next of piano). The results are shown in the case of the ADMM with (6) in the CQT domain. The horizontal axis is time. Upper and lower notes in each activation are clarinet and piano, respectively.

over "attack" columns[15] to evaluate the performance in "attack" segments.

The proposed technique achieves the best performance in all the simulation scenarios as shown in Table 1. Especially, we observe significant performance advancement in "attack" segments. For example, as shown in Fig. 4, the proposed basis matrix realizes an accurate estimation of the "attack" segment of the piano.

6. CONCLUDING REMARKS

This paper has proposed a novel scheme to construct the basis matrix for describing time-varying spectral-patterns. We have designed the basis matrix using a reduced-rank approximation under the nonnegativity constraints. Exploiting knowledge on the reduced-rank representation, our scheme can determine the number of the basis vectors automatically. In addition, as an application of the proposed scheme to automatic transcription, we have discussed the design of iterative algorithms for estimating the activation matrix by exemplifying a mathematical translation of sound volume smoothness. The numerical examples have shown the superior performance of the proposed design of the basis matrix to the standard basis matrix based on the rank-one approximation problem.

[15]As "attack" columns, we extract columns including the start of sound and their three subsequent columns, where the start of sound is known from the ground-truth.

APPENDIX: Iterative solvers for (4)

We adopt the ADMM [15] for solving the problem (4). Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be Euclidean spaces equipped with the standard inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\|$. In general, the ADMM can solve the following convex optimization problem:

$$\underset{\boldsymbol{x}\in\mathcal{X}_1}{\text{minimize}}\, f(\boldsymbol{x}) + g(L(\boldsymbol{x})) \tag{8}$$

where $f\colon \mathcal{X}_1 \to \mathbb{R}\cup\{\infty\}$ and $g\colon \mathcal{X}_2 \to \mathbb{R}\cup\{\infty\}$ are *proper lower semicontinuous convex functions*[16] and $L\colon \mathcal{X}_1 \to \mathcal{X}_2$ is a bounded linear operator. The ADMM iteratively computes

$$\begin{cases} \boldsymbol{x}_{k+1} = \underset{\boldsymbol{x}\in\mathcal{X}_1}{\arg\min}\, f(\boldsymbol{x}) + \dfrac{1}{2\gamma}\|L(\boldsymbol{x}) - \boldsymbol{\eta}_k + \boldsymbol{\xi}_k\|^2 \\ \boldsymbol{\eta}_{k+1} = \mathrm{prox}_{\gamma g}(L(\boldsymbol{x}_{k+1}) + \boldsymbol{\xi}_k) \\ \boldsymbol{\xi}_{k+1} = \boldsymbol{\xi}_k + L(\boldsymbol{x}_{k+1}) - \boldsymbol{\eta}_{k+1}, \end{cases}$$

where $\mathrm{prox}_{\gamma g}\colon \mathcal{X}_2 \to \mathcal{X}_2$ is the *proximity operator*[17] of γg with $\gamma > 0$ (see e.g. [20], [21] for convergence analysis of the ADMM). That is, we can solve (4) by applying the ADMM to the following reformulation of (4) into (8): Let

$$\begin{aligned} f(\boldsymbol{W}) &:= \|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{W}\|_F^2 \\ g(\boldsymbol{Z}) &:= \lambda_1\|\boldsymbol{Z}_1\|_{\ell_1} + \lambda_2\nu(\boldsymbol{Z}_2) + \iota_{\mathbb{R}_+^{K\times D}}(\boldsymbol{Z}_3) \\ L(\boldsymbol{W}) &:= (\boldsymbol{W}, \Phi(\boldsymbol{W}), \boldsymbol{W}), \end{aligned}$$

where $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3) \in \mathbb{R}^{K\times D}\times\mathbb{R}^{I\times(D-1)}\times\mathbb{R}^{K\times D}$, $\iota_{\mathbb{R}_+^{K\times D}}$ is the indicator function[18] of $\mathbb{R}_+^{K\times D}$, $\nu : \mathbb{R}^{I\times(D-1)} \to \mathbb{R}$ is $\nu(\cdot) = \|\cdot\|_{\ell_1}$ if $\psi = \psi_1$; $\nu(\cdot) = \|\cdot\|_F^2$ if $\psi = \psi_2$, and we utilize the fact that the third terms in (4) are the composition

$$\psi_1(\boldsymbol{W}) = \|\Phi(\boldsymbol{W})\|_{\ell_1}, \quad \psi_2(\boldsymbol{W}) = \|\Phi(\boldsymbol{W})\|_F^2,$$

of a function involving a suitable norm and a bounded linear operator $\Phi\colon \mathbb{R}^{K\times D} \to \mathbb{R}^{I\times(D-1)}$ defined by

$$\Phi(\boldsymbol{W}) := \boldsymbol{A}\boldsymbol{W}\boldsymbol{H},$$

where $\boldsymbol{H} = (h_{i,j}) \in \mathbb{R}^{D\times(D-1)}$ computes volume differences, i.e.,

$$h_{i,j} = \begin{cases} -\alpha_i & (i = j) \\ \alpha_i & (i - 1 = j) \\ 0 & \text{otherwise.} \end{cases}$$

[16]A function $f\colon \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ is called proper lower semicontinuous convex if $\mathrm{dom}(f) := \{\boldsymbol{x}\in\mathcal{X} | f(\boldsymbol{x}) < \infty\} \neq \emptyset$, $\mathrm{lev}_{\leq\alpha}(f) := \{\boldsymbol{x}\in\mathcal{X} | f(\boldsymbol{x}) \leq \alpha\}$ is closed for all $\alpha\in\mathbb{R}$, and $f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y}\in\mathcal{X}$ and $\lambda\in(0,1)$, respectively.

[17]The proximity operator of a proper lower semicontinuous convex function $f\colon \mathcal{X} \to \mathbb{R}\cup\{\infty\}$ is given by $\mathrm{prox}_f(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathcal{X}} f(\boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$.

[18]For a given nonempty closed convex set C, the indicator function ι_C is defined by $\iota_C(\boldsymbol{x}) := 0$ if $\boldsymbol{x}\in C$; ∞ otherwise.

7. REFERENCES

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

[2] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing & Management*, vol. 42, no. 2, pp. 373–386, 2006.

[3] Y. Chen, M. Rege, M. Dong, and J. Hua, "Non-negative matrix factorization for semi-supervised data clustering," *Knowledge and Information Systems*, vol. 17, no. 3, pp. 355–379, 2008.

[4] J. Paulus and T. Virtanen, "Drum transcription with nonnegative spectrogram factorisation," in *European Signal Processing Conference (EUSIPCO)*, 2005, pp. 4–8.

[5] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*. Springer, 2013, pp. 341–371.

[6] Y. Morikawa and M. Yukawa, "A sparse optimization approach to supervised NMF based on convex analytic method," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6078–6082.

[7] G. Grindlay and D. P. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1159–1169, 2011.

[8] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 121–124.

[9] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1727–1741, 2013.

[10] U. of Iowa, "University of Iowa electronic music studios." [Online]. Available: http://theremin.music.uiowa.edu/index.html

[11] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[12] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 2012.

[13] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proceedings of International Computer Music Conference (ICMC)*, vol. 3, 2003, pp. 231–234.

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

[15] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.

[16] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking systems." in *The International Society for Music Information Retrieval (ISMIR)*, 2009, pp. 315–320.

[17] S. A. Abdallah and M. D. Plumbley, "An independent component analysis approach to automatic music transcription," *Preprints-Audio Engineering Society*, 2003.

[18] D. Kitamura, H. Saruwatari, Y. Kosuke, K. Shikano, Y. Takahashi, and K. Kondo, "Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 97, no. 5, pp. 1113–1118, 2014.

[19] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conference, Barcelona, Spain*, 2010, pp. 3–64.

[20] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, 1992.

[21] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.