Supervised Nonnegative Matrix Factorization Using Active-Period-Aware Structured ℓ_1-Norm for Music Transcription

Yu Morikawa* Masahiro Yukawa† Hisakazu Kikuchi*
* Department of Electrical and Electronic Engineering, Niigata University, JAPAN
† Department of Electronics and Electrical Engineering, Keio University, JAPAN

***Abstract*—An active-period-aware supervised nonnegative matrix factorization (NMF) approach for music transcription is proposed. Supervised NMF relies on a set of known spectrograms associated with all musical instruments that may possibly be involved with given music data; this is supported by the availability of large database of a variety of musical instruments. It is free from the source-number determination problem and this is a significant advantage over the unsupervised NMF approaches. The proposed approach is composed of three steps. Step 1: Apply the existing supervised NMF algorithm. Step 2: Estimate the 'active' periods (during which musical sounds are present) based on the outcomes of Step 1. Step 3: Optimize a refined cost function reflecting the estimate of active periods. The awareness of active periods leads to avoidance of the so-called octave-errors which is a central issue of the existing supervised NMF method. Simulation results show the efficacy of the proposed approach.**[1]

I. Introduction

We address the supervised nonnegative matrix factorization (NMF) problem for music transcription in which a given nonnegative matrix $\boldsymbol{Y}$ wants to be factorized into two nonnegative matrices $\boldsymbol{W}\boldsymbol{H}$ with $\boldsymbol{W}$ known a priori. Here, $\boldsymbol{W}$ a basis matrix (a set of basis vectors) and $\boldsymbol{H}$ is an activation matrix (a set of activation vectors). The supervised NMF method proposed in [1] has no need to determine the *source number;* this is a significant advantage over the unsupervised NMF approaches [2–7]. The method in [1] is based on convex optimization and it involves two regularizers: (i) the sum of the ℓ_2 norms of the row vectors of $\boldsymbol{H}$ (which promotes 'row sparsity' of $\boldsymbol{H}$ and which is referred to as a group ℓ_1 norm) and (ii) the ℓ_1 norm of $\mathrm{vec}(\boldsymbol{H})$ (which promotes sparsity of $\boldsymbol{H}$). The former regularizer aims to pick up only those notes which are present in the music, while the latter one is based on the fact that each note is present only in a part of the time frame (for which NMF is performed). Typically, the spectrograms for two pitches (of the same musical instrument) that differ by an octave, e.g., piano-C3 and piano-C4, are fairly close to each other. Hence, when such octave-different notes are present in the time frame, the spectrogram of one of such two notes tends to be confused with the other one (see Section II-B for an illustration of such a phenomenon). This is due to the use of the group ℓ_1 norm where the group is given by the set of row vectors of $\boldsymbol{H}$. The phenomenon, referred to as the *octave-error issue*, causes significant performance degradation, and hence it has been a central issue to be addressed.

In this paper, we propose an efficient supervised NMF scheme that involves a certain structured ℓ_1-norm regularizer to pick up the notes correctly. The proposed scheme first performs the existing supervised NMF method [1]. Based on its outcome, it then estimates *active periods* which are defined for each note as time frames during which the note is present. The structured ℓ_1 norm is then defined as the sum of ℓ_2 norms of the subvectors (not the whole row vectors) corresponding to the active periods. (The structured ℓ_1 norm has been studied in detail in [8]; it is similar to group ℓ_1 norms, but the groups may overlap although their union should cover all the elements.) The proposed scheme finally optimizes a refined cost function including the structured ℓ_1 norm regularizer. By doing so, the activeness of a part of some row does not affect the other parts of the row, which leads to avoidance of the octave errors. The simulation results show that the proposed scheme outperforms the existing supervised NMF methods.

II. Background and Motivation

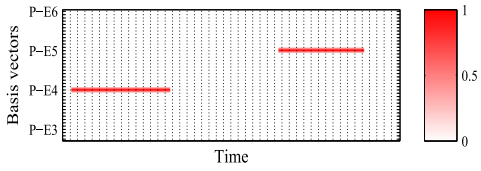
A. Background of supervised NMF

We briefly describe the supervised NMF approach proposed in [1], which has no need to determine the *source number* prior to decomposition. We assume that the basis matrix $\boldsymbol{W}$ is given; the given basis matrix may contain such basis vectors that are 'irrelevant' to $\boldsymbol{Y}$ as well as 'relevant' ones. To select 'relevant' basis vectors, the supervised NMF has been formulated as the following sparse optimization problem:

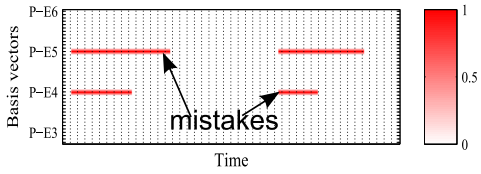
$$(\mathrm{P}_1)\ \min_{\boldsymbol{H}\in\mathcal{H}} J_1(\boldsymbol{H}) = \|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2 + \underbrace{i_C(\boldsymbol{H})}_{\text{(a)}} + \underbrace{\lambda_1 \sum_{l=1}^{L} \|\hat{\boldsymbol{h}}_l\|_2}_{\text{(b)}}$$
$$+ \underbrace{\lambda_2 \sum_{l=1}^{L}\sum_{n=1}^{N-1} (h_{l,n+1}-h_{l,n})^2 + \lambda_3 \sum_{l=1}^{L}\sum_{n=1}^{N} |h_{l,n}|}_{\text{(c)}},$$

where $\boldsymbol{Y} \in \mathbb{R}_{\geq 0}^{M\times N}$, which denotes the set of all nonnegative valued matrices of size $M \times N$, $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{M\times L}$, $\boldsymbol{H}$ =

[1]This work was supported by the Support Center for Advanced Telecommunications Technology Research (SCAT). This work was done when Yu Morikawa was with Department of Electrical and Electronic Engineering, Niigata University, Japan. He is currently working for Sanei Hytechs, Japan.



(a) Ground-truth reference



(b) The solution of (P_1)

Fig. 1. Single sounds of piano (pitches E4 and E5).

$\left[\hat{\boldsymbol{h}}_1\ \hat{\boldsymbol{h}}_2\ \cdots\ \hat{\boldsymbol{h}}_L\right]^{\mathsf{T}} \in \mathcal{H} := \mathbb{R}^{L\times N}$, $h_{l,n}$ denotes the (l,n)-entry of $\boldsymbol{H}$, $C := \mathbb{R}^{L\times N}_{\geq 0} \subset \mathcal{H}$, $\|\cdot\|_2$ denotes the ℓ_2 norm, and $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm. Here, $(\cdot)^{\mathsf{T}}$ stands for *transpose*, $\mathcal{H}$ is the Hilbert space of the activation matrix $\boldsymbol{H}$ to be optimized, and $\hat{\boldsymbol{h}}_l$s are referred to as activation vectors. The indicator funciton $i_C(\boldsymbol{H}) := 0$ if $\boldsymbol{H} \in C$, $i_C(\boldsymbol{H}) := \infty$ otherwise, enforces the solution of (P_1) to be nonnegative. The other penalty terms in (P_1) are used for (b) basis-vector selection and (c) correct activation-vector estimation, respectively. (The first term of (c) enhances temporal continuity.) Typically, $M \gg L$ and hence (P_1) has a unique solution because of the strict convexity of $\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2$.

B. Motivation for this study

Our preliminary experiments revealed that the method in [1] tends to fail when $\boldsymbol{Y}$ contains multiple notes whose frequency spectra are close to each other. Let us show a simple example. Fig. 1(a) describes the case that the matrix $\boldsymbol{Y}$ has two notes, P-E4 (Piano-E4) and P-E5, whose pitches differ by an octave. The result is depicted in Fig. 1(b). The following observations indicate the reasons for the mistakes illustrated in the figure.

1) Those frequency spectra whose pitches differ by an octave tend to be close to each other. Such frequency spectra are intrinsically difficult to distinguish.
2) The formulation in (P_1) uses a group ℓ_1 norm associated with the row-vector groups as shown in Fig. 2, seeking for row-sparse solutions. This means that those basis vectors which are regarded 'relevant' tend to be regarded active by mistake even in inactive periods.
3) There is a model mismatch in general; i.e., there is a gap between the frequency spectra of the model ($\boldsymbol{W}$) and the true audio signals.

From the above observations, one can see that the frequency spectra of the true audio signals can be expressed — at the smallest cost, i.e., with the minimal $J_1(\boldsymbol{H})$, — as a linear combination of P-E4 and P-E5 (rather than solely by one of them).

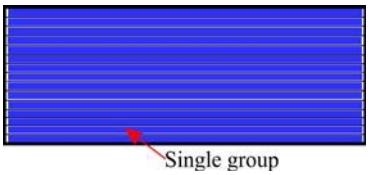


Fig. 2. The group for the structured ℓ_1 norm.

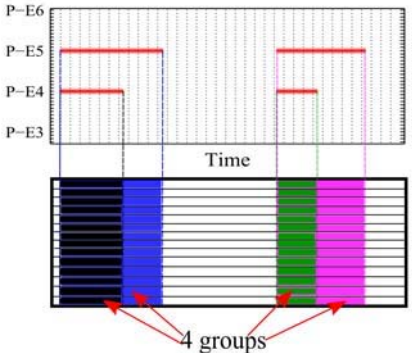


Fig. 3. The groups for the structured ℓ_1 norm constructed by using the solution of (P_1). Each group is the set of row vectors of each submatrix of $\boldsymbol{H}$ corresponding to each active period.

III. Proposed Approach with a Structured ℓ_1 Norm

In Fig. 1 (b), a resulting $\boldsymbol{H}$ for the problem (P_1) would have zero row vectors (corresponding to the 'irrelevant' basis vectors) and nonzero row vectors (corresponding to the 'relevant' basis vectors). It should be emphasized that the selected basis vectors are maybe incorrect in situations that the input matrix $\boldsymbol{Y}$ has two notes whose pitches differ by an octave. However, each estimated active period (duration time) is correct to some extent. Active period is formally defined as follows.

Definition 1 (Active period): For the solution $\boldsymbol{H}$ of (P_1), we define the set of active periods as follows:

$$\mathcal{A} = \{[\![n_1, n_2]\!] := \{n_1, n_1+1, n_1+2, \cdots, n_2\} \mid n_1, n_2 \in [\![1, N]\!],\ \exists l \in [\![1, L]\!] \text{ s.t. } (C1) \text{ and } (C2) \text{ hold}\}. \tag{1}$$

C1. $h_{l,n_1} \neq 0, h_{l,n_1+1} \neq 0, \cdots, h_{l,n_2} \neq 0$.
C2. $h_{l,n_1-1} = h_{l,n_2+1} = 0$ if $n_1 \geq 2$ and/or $n_2 \leq N-1$.
We construct the groups for the structured ℓ_1 norm based on the following rules.

- The overlapping periods are counted individually. In the case of Fig. 3, for instance, the number of active periods is four.
- The group construction does not take into account what pitch is active.

A. Problem Formulation

Define

$$S := \{(i,j) \mid i \in [\![1, L]\!], j \in [\![1, N]\!]\}, \tag{2}$$

and denote by 2^S its power set (the collection of all the subset of S). Let $\mathcal{G} \subset 2^S$ satisfy $\bigcup_{G\in\mathcal{G}} G = S$; i.e., the sets in

TABLE I
SUMMARY OF THE PROPOSED SCHEME.

Step 1. Solve (P_1) by the method in [1].
Step 2. Detect the active periods based on the result of Step 1 and construct the groups based on the active periods (See (1)).
Step 3. Solve (P_2) based on (10a) and (10b).

$\mathcal{G}$ may overlap but their union covers all the elements in S (cf. [8]). We formulate the supervised NMF problem using the overlapping groups as the following sparse optimization problem:

$$(\mathrm{P}_2)\ \min_{\boldsymbol{H}\in\mathcal{H}} J_2(\boldsymbol{H}) = \|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2 + i_C(\boldsymbol{H}) + \Omega(\boldsymbol{H}),$$

where

$$\Omega(\boldsymbol{H}) = \sum_{G\in\mathcal{G}} \left\| \boldsymbol{R}^G \circ \boldsymbol{H} \right\|_{\mathrm{F}}. \tag{3}$$

Here, $\boldsymbol{R}^G$ is an $L\times N$ matrix such that its (s,t)-entry $r^G_{s,t} > 0$ if $(s,t)\in G$ and $r^G_{s,t}=0$ otherwise, and $\circ$ denotes the Hadamard product. A design example of $\mathcal{G}$ is presented below.

Example 1 (Group design with active-period estimate): Assume that an estimate of active periods of the activation matrix $\boldsymbol{H}$ is available by solving the problem (P_1) [1]. We consider (P_2) with

$$\Omega(\boldsymbol{H}) = \sum_{q=1}^{Q}\lambda_q \sum_{l=1}^{L} \|\hat{\boldsymbol{h}}_{q,l}\|_2 + \tilde{\lambda}\sum_{l=1}^{L}\sum_{n=1}^{N}|h_{l,n}|, \tag{4}$$

where $\hat{\boldsymbol{h}}_{q,l}$ denotes the l-th row vector of the submatrix $\boldsymbol{H}_q \in \mathbb{R}^{L\times(n_{q,2}-n_{q,1}+1)}$ corresponding to the q-th active period $[\![n_{q,1}, n_{q,2}]\!]$, and Q is the cardinality of the active set $\mathcal{A}$. Each $\sum_{l=1}^{L}\|\hat{\boldsymbol{h}}_{q,l}\|_2$ is the structured ℓ_1 norm penalty of the submatrix of $\boldsymbol{H}$ for selecting the basis vectors 'relevant' to $\boldsymbol{Y}$ for the period.

B. Proposed Scheme

The proposed approach has three steps as summarized in Table 1. First, the solution of (P_1) is computed to estimate the set $\mathcal{A}$ of active periods (Step 1). Then, the groups for the structured ℓ_1 norm are constructed by using the estimated $\mathcal{A}$ (Step 2). Finally, the solution of (P_2) is computed (Step 3) as described below.

The cost function in (P_2) with the penalty $\Omega(\boldsymbol{H})$ given in Example 1 can be written in the following form:

$$J_3(\boldsymbol{H}) = \underbrace{\varphi(\boldsymbol{H})}_{\text{smooth}} + \underbrace{\sum_{j=1}^{Q+2}\psi_j(\boldsymbol{H})}_{\text{nonsmooth}}, \tag{5}$$

where

$$\varphi(\boldsymbol{H}) := \|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2, \tag{6}$$
$$\psi_1(\boldsymbol{H}) := i_C(\boldsymbol{H}), \tag{7}$$
$$\psi_{j+1}(\boldsymbol{H}) := \lambda_j \sum_{l=1}^{L}\|\hat{\boldsymbol{h}}_{j,l}\|, \quad j=1,2,\cdots,Q, \tag{8}$$
$$\psi_{Q+2}(\boldsymbol{H}) := \tilde{\lambda}\sum_{l=1}^{L}\sum_{n=1}^{N}|h_{l,n}|. \tag{9}$$

Here, φ is a differentiable convex function with the Lipschitz-continuous gradient (i.e., φ is *smooth*) while ψ_j, $j=1,2,\cdots,Q+2$ are nonsmooth but *proximable* convex functions. (See [9, 10] for details about convex analysis in Hilbert spaces.)

The problem (P_2) with the penalty $\Omega(\boldsymbol{H})$ given in Example 1 can iteratively be solved by generating the sequence of the auxiliary variables $(\boldsymbol{Z}_j^{(k)})_{k\in\mathbb{N}}\subset\mathcal{H}$, $j=1,2,\cdots,Q+2$, and $(\boldsymbol{H}^{(k)})_{k\in\mathbb{N}}\subset\mathcal{H}$, with initial estimates $\boldsymbol{Z}_j^{(0)}$, $j=1,2,\cdots,Q+2$, and $\boldsymbol{H}^{(0)} := \sum_{j=1}^{Q+2}\omega_j\boldsymbol{Z}_j^{(0)}$ as follows [11]:

$$\begin{aligned}\boldsymbol{Z}_j^{(k)} := \boldsymbol{Z}_j^{(k-1)} + \alpha\Big(&\mathrm{prox}_{\frac{\gamma}{\omega_j}\psi_j}(2\boldsymbol{H}^{(k-1)} - \boldsymbol{Z}_j^{(k-1)} \\ &-\gamma\nabla\varphi(\boldsymbol{H}^{(k-1)})) - \boldsymbol{H}^{(k-1)}\Big), \\ &j=1,2,\cdots,Q+2,\end{aligned} \tag{10a}$$

$$\boldsymbol{H}^{(k)} := \sum_{j=1}^{Q+2}\omega_j\boldsymbol{Z}_j^{(k)}, \tag{10b}$$

where $\omega_j\in(0,1)$ s.t. $\sum_{j=1}^{Q+2}\omega_j=1$, $j=1,2,\cdots,Q+2$, $\alpha\in\left(0,\min\left\{\frac{3}{2},\frac{\eta\gamma+2}{2\eta\gamma}\right\}\right)$, $\gamma\in\left(0,\frac{2}{\eta}\right)$, $\eta=2\sigma_{\max}(\bar{\boldsymbol{W}}^{\mathsf{T}}\bar{\boldsymbol{W}})$, $\bar{\boldsymbol{W}}=\mathrm{diag}(\boldsymbol{W}\ \boldsymbol{W}\ \cdots\ \boldsymbol{W})\in\mathbb{R}^{NM\times NL}$, and $\sigma_{\max}(\bar{\boldsymbol{W}}^{\mathsf{T}}\bar{\boldsymbol{W}})$ is the maximum modulus of the eigenvalues of $\bar{\boldsymbol{W}}^{\mathsf{T}}\bar{\boldsymbol{W}}$. The gradient $\nabla\varphi$ and the proximity operators $\mathrm{prox}_{\frac{\gamma}{\omega_j}\psi_j}, j=1,2,\cdots,Q+2$, can be computed as follows:

$$\nabla\varphi(\boldsymbol{H}) = 2\boldsymbol{W}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{H} - 2\boldsymbol{W}^{\mathsf{T}}\boldsymbol{Y}, \tag{11}$$

$$\begin{aligned}\mathrm{prox}_{\frac{\gamma}{\omega_1}\psi_1}(\boldsymbol{H}) &= P_C(\boldsymbol{H}) \\ &= \begin{bmatrix}\max\{h_{1,1},0\}\cdots\max\{h_{1,N},0\}\\ \vdots\qquad\ddots\qquad\vdots\\ \max\{h_{L,1},0\}\cdots\max\{h_{L,N},0\}\end{bmatrix},\end{aligned} \tag{12}$$

$$\begin{aligned}&\mathrm{prox}_{\frac{\gamma}{\omega_{j+1}}\psi_{j+1}}(\boldsymbol{H}) = \\ &\sum_{l=1}^{L}\boldsymbol{e}_l\left[\hat{\boldsymbol{h}}_{j,l,\mathrm{L}}^{\mathsf{T}}, \max\left\{1-\frac{\lambda_j\gamma}{\omega_j\|\hat{\boldsymbol{h}}_{j,l}\|_2},0\right\}\hat{\boldsymbol{h}}_{j,l}^{\mathsf{T}}, \hat{\boldsymbol{h}}_{j,l,\mathrm{R}}^{\mathsf{T}}\right] \\ &\qquad j=1,2,\cdots,Q,\end{aligned} \tag{13}$$

$$\begin{aligned}&\mathrm{prox}_{\frac{\gamma}{\omega_{Q+2}}\psi_{Q+2}}(\boldsymbol{H}) = \\ &\sum_{l,n=1}^{L,N}\mathrm{sgn}(h_{l,n})\max\left\{|h_{l,n}|-\frac{\tilde{\lambda}\gamma}{\omega_{Q+2}},0\right\}\boldsymbol{E}_{l,n},\end{aligned} \tag{14}$$

where $\{\boldsymbol{e}_l\}_{l=1}^N$ denotes the standard basis for $\mathbb{R}^L$, $\hat{\boldsymbol{h}}_{j,l,\mathrm{L}}^{\mathsf{T}} \in \mathbb{R}^{1\times(n_{j,1}-1)}$ and $\hat{\boldsymbol{h}}_{j,l,\mathrm{R}}^{\mathsf{T}} \in \mathbb{R}^{1\times(N-n_{j,2})}$ are the lth rows of the submatrices $\boldsymbol{H}_{j,\mathrm{L}} \in \mathbb{R}^{L\times(n_{j,1}-1)}$ and $\boldsymbol{H}_{j,\mathrm{R}} \in \mathbb{R}^{L\times(N-n_{j,2})}$ of $\boldsymbol{H} =: [\boldsymbol{H}_{j,\mathrm{L}}\ \boldsymbol{H}_j\ \boldsymbol{H}_{j,\mathrm{R}}]$, respectively, and $\boldsymbol{E}_{l,n}$ is the $L\times N$ matrix having one at the (l,n)-entry and zeros elsewhere. The definitions of the proximity operator and the projection are given in the appendix.

IV. Simulation Results

We show the efficacy of the proposed scheme for music transcription. As the input audio signal, we use simple single sounds of piano. The basis matrix $\boldsymbol{W}$ is composed of amplitude spectra which are respectively obtained by the short-time Fourier transform (STFT) of piano sounds of 88 pitches, which have different timbre from those of the input signal, violin sounds of 46 pitches, and flute sounds of 37 pitches [12].

We compare the performance of the proposed scheme with the method proposed in [1], the Beta Nonnegative Decomposition (BND) method [13], and the unsupervised Euclidean-NMF (EUNMF) method [4]. All the audio signals for both the input audio signal and the audio signals to learn $\boldsymbol{W}$ are sampled at 16 kHz. STFT is computed using a Hamming window that is 64 ms long with a 32 ms overlap. The parameter of unsupervised EUNMF (source number) is set to the number $L = 3$ of actual sources that are present in the music under consideration. The initial estimates for unsupervised EUNMF are selected randomly, and the algorithm is run for 300 iterations. The parameter of BND is set manually to $\beta = 0.95$ to attain reasonable performance. The initial estimate for BND is set to a matrix with all entries equal to one, and the algorithm is run for 300 iterations at each period. The parameters of (P_1) are set manually to $\lambda_1 = 900, \lambda_2 = 0, \lambda_3 = 5$, respectively, to attain reasonable performance. The initial estimates $\boldsymbol{Z}_j^{(0)}, j = 1,2,3$, are set to random matrices, and the algorithm is run for 300 iterations. The parameters of (P_2) are set manually to $\lambda_1 = 100, \lambda_2 = 500, \lambda_3 = 600, \lambda_4 = 100, \lambda_5 = 500, \lambda_6 = 500, \lambda_7 = 400, \lambda_8 = 400, \tilde{\lambda} = 5$, respectively, to attain reasonable performance. (Note that the resulting $\boldsymbol{H}$ for (P_1) indicates $Q = 8$ as seen in Fig. 4(d).) The initial estimates $\boldsymbol{Z}_j^{(0)}, j = 1,2,\cdots,10$, are set to random matrices, and the algorithm is run for 300 iterations. As post-processing, for all algorithms, each entry of $\boldsymbol{H}$ is divided by the maximum value of $\boldsymbol{H}$. We consider each entry of $\boldsymbol{H}$ to be active (or inactive) if it is greater (or smaller) than the threshold 0.05.

Fig. 4(a) illustrates the ground-truth reference which consists of piano sounds E4, E5, and G5 over 5 seconds, and Figs. 4(b), 4(c), 4(d) and 4(e) depict the post-processed activation matrix $\boldsymbol{H}$ obtained by the unsupervised EUNMF, BND, the method in [1], and the proposed method, respectively. Table II summarizes the results in the standard evaluation metrics from the MIREX [14]. Note that the groud-truth reference for the unsupervised EUNMF is *not* the matrix illustrated in Fig. 4(a) itself, but its submatrix. It is seen that the proposed scheme attains higher score in F-measure $\mathcal{F}$ and lower score in total error $\mathcal{E}_{\mathrm{tot}}$ (measuring different types of errors), meaning that it outperforms the unsupervised EUNMF [4], BND [13], and the method in [1] in both metrics. This is due to the use of the structured ℓ_1 norm reflecting the active periods (which is unexploited in [13] and [1]).

TABLE II
Transcription evaluation from the MIREX [14]. The metrics $\mathcal{F}$ and $\mathcal{E}_{\mathrm{tot}}$ stand for F-measure and total error, respectively.

Algorithm	$\mathcal{F}$	$\mathcal{E}_{\mathrm{tot}}$
unsupervised EUNMF [4]	83.3	34.3
BND [13]	77.8	45.1
supervised NMF [1]	69.2	58.8
Proposed	**89.6**	**17.7**

V. Conclusion

This paper presented a systematic approach to supervised NMF for music transcriptions. The supervised NMF problem was formulated as a sparse optimization problem under a structured ℓ_1-norm regularization reflecting the active periods. The simulation results showed that the proposed approach effectively prevents the octave errors and attains excellent performance.

APPENDIX

Definition 2 ([9, 10]): Let $(\mathcal{H}, \|\cdot\|_{\mathrm{F}})$ be a real Hilbert space.

(a) Given any proper lower-semicontinuous convex function[2] $\psi : \mathcal{H} \to \mathbb{R}$, the proximity operator of ψ of index $\gamma > 0$ for any $\boldsymbol{X} \in \mathcal{H}$ is defined as

$$\mathrm{prox}_{\gamma\psi}(\boldsymbol{X}) := \underset{\boldsymbol{Y}\in\mathcal{H}}{\mathrm{argmin}} \left(\psi(\boldsymbol{Y}) + \frac{1}{2\gamma} \|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}^2 \right).$$

Here, the minimizer of the function $f_1(\boldsymbol{Y}) := \psi(\boldsymbol{Y}) + \frac{1}{2\gamma}\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}^2$ uniquely exists because of its coercivity and strict convexity.

(b) Given any nonempty closed convex set $K \subset \mathcal{H}$, the metric projection of any $\boldsymbol{X} \in \mathcal{H}$ onto the set K is defined as

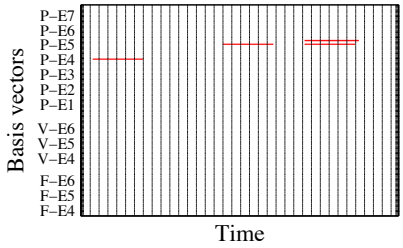
$$P_K(\boldsymbol{X}) := \underset{\boldsymbol{Y}\in K}{\mathrm{argmin}} \|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}.$$

Here, by the convex projection theorem, the minimizer of the function $f_2(\boldsymbol{Y}) := \|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}$ over K uniquely exists due to the closedness and convexity of $K \neq \emptyset$.

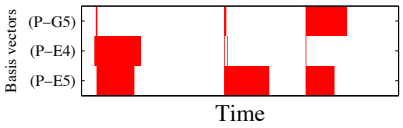
References

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

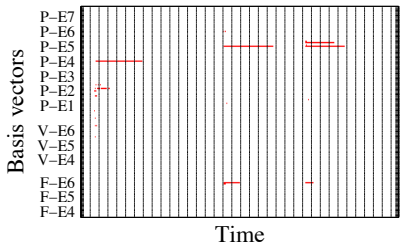
[2] A function $f : \mathcal{H} \to (-\infty, \infty]$ is convex if $f(t\boldsymbol{X}_1 + (1-t)\boldsymbol{X}_2) \leq tf(\boldsymbol{X}_1) + (1-t)f(\boldsymbol{X}_2)$ for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathcal{H}$ and $t \in [0,1]$. It is proper if $\{\boldsymbol{X} \mid f(\boldsymbol{X}) < \infty\} \neq \phi$. It is lower semicontinuous if the set $\{\boldsymbol{X} \mid f(\boldsymbol{X}) \leq a\}$ is closed for any $a \in \mathbb{R}$.



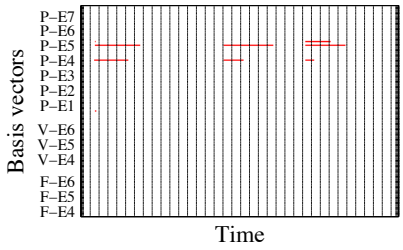
(a) Ground-truth reference



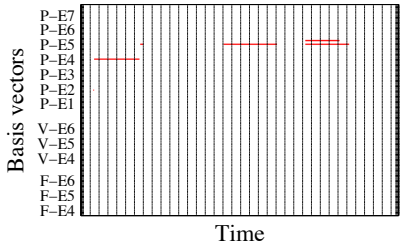
(b) The resulting $\boldsymbol{H}$ for unsupervised EUNMF



(c) The resulting $\boldsymbol{H}$ for BND [13]



(d) The resulting $\boldsymbol{H}$ for (P_1) [1]



(e) The resulting $\boldsymbol{H}$ for (P_2) (proposed)

Fig. 4. Simulation results for single sounds of piano (pitches E4 and E5).

[2] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," in *Environmetrics*, 1994.

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

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*, 2000, pp. 556–562.

[5] D. Guillamet and J. Vitriá, "Analyzing non-negative matrix factorization for image classification," in *Proc. IEEE International Conference on Pattern Recognition*, 2002, pp. 116–119.

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

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

[8] R. Jenatton, J-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, pp. 2777–2824, 2011.

[9] H. H. Bauschke and P. L. Combettes, *Convex Analysis And Monotone Operator Theory in Hilbert Spaces*, Springer, New York: NY, 1st edition, 2011.

[10] I. Yamada, M. Yukawa, and M. Yamagishi, "Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 345–390. Springer, 2011.

[11] H. Raguet, J. Fadili, and G. Peyré, "A generalized forward-backward splitting," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, 2013.

[12] M. Goto, "Development of the RWC music database," in *Proc. ICA*, 2004, pp. 553–556.

[13] A. Dessein, A. Cont, and G. Lemaitr, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*, pp. 341–371. Springer, 2012.

[14] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-F0 estimation and tracking systems," in *Proc. ISMIR*, 2009, pp. 315–320.