Probabilistic spectral envelope modeling of musical instruments within the non-negative matrix factorization framework for mixed music analysis

Toru Nakashika\textsuperscript{1,\ast}, Tetsuya Takiguchi\textsuperscript{2,\dagger} and Yasuo Ariki\textsuperscript{2,\ddagger}

\textsuperscript{1}Graduate School of System Informatics, Kobe University, 1–1 Rokkodai, Nada-ku, Kobe, 657–8501 Japan
\textsuperscript{2}Organization of Advanced Science and Technology, Kobe University, 1–1 Rokkodai, Nada-ku, Kobe, 657–8501 Japan

(Received 26 April 2013, Accepted for publication 7 February 2014)

Abstract: Non-negative matrix factorization (NMF) has been one of the most useful techniques for musical signal analysis in recent years. In particular, supervised NMF, in which a large number of instrumental samples are used for the analysis, is garnering much attention with respect to analytical accuracy and speed. The accuracy, however, deteriorates if the system does not have enough samples. Therefore, in principle, such methods require as many samples as possible in order for the analysis to be accurate. In this paper, we propose an analysis method that 1) does not require the collection of a large number of training samples, and 2) combines the NMF and probabilistic approaches. In this approach, it is assumed that each instrumental category has a model-invariant feature, called a probabilistic spectral envelope (PSE). As an extension of a spectral envelope, this feature represents the probabilities of spectral envelopes belonging to the instrumental category in a two-dimensional (frequency-amplitude) space. The analysis of an input musical signal is carried out using a supervised NMF framework, where the basis matrix contains the optimum spectra that have been generated from pretrained PSEs.

Keywords: Multipitch analysis, Music transcription, Non-negative matrix factorization, Probabilistic spectral envelope, Gaussian process

PACS number: 43.75.–z, 43.66.Jh, 43.75.Xz, 43.75.Yy, 43.75.Zz, 43.66.Hg
[doi:10.1250/ast.35.181]

1. INTRODUCTION

Mixed music analysis\textsuperscript{\ast}, such as music transcription, where the instrumental labels of each musical note are estimated along with the pitches from a single-channel polyphonic music signal produced by multiple instruments, has been recognized as one of the most challenging tasks in musical signal processing. Unlike monophonic music, various instruments are played polyphonically at the same time, which makes the observed spectrum complicated and hard to analyze. Nevertheless, mixed music analysis has a broader range of applications than monophonic music analysis, such as for music retrieval and music encoding.

Monophonic music can be analyzed with relatively high accuracy [1–3] since the spectrum is not mixed at all. However, with polyphonic music, the analysis becomes much harder, even if only one instrument is being analyzed [4–8]. This is mainly because an acoustic signal contains fundamental frequencies and harmonics, but it is unknown which peak corresponds to the fundamental frequency or the harmonics. In addition, the difficulty of analyzing polyphonic music (mixed music) increases because unknown instrumental labels need to be estimated.

To tackle mixed music analysis, several approaches have been proposed to date including methods based on factorial hidden Markov models (HMMs) [9], independent component analysis (ICA)-based methods [10,11], harmonic-temporal-timbral clustering (HTTC) [12], and instrstroms [13]. In particular, the methods based on NMF have attracted considerable attention recently as a way of analyzing signals more effectively and more easily. In
many of the NMF-based approaches, an observed spectrogram matrix can be represented as the product of two matrices: a basis matrix $W$ and an activity matrix $H$ (collectively referred to as “NMF matrices”). The columns of a basis matrix roughly indicate the spectrum of each source signal associated with the instrument and the pitch, and the activity matrix shows the temporal information of each source.

The methods of analysis based on NMF are broadly divided into two categories: unsupervised [14–16] and supervised [17–19]. Because the former approach decomposes the spectrogram without the assumption of the spectral structures of audio sources, we often obtain incorrect matrices. Thus, it is hard to analyze music signals correctly using an unsupervised approach.

On the other hand, supervised NMF uses the spectral templates of each musical source, which are learned beforehand. Compared with an unsupervised approach, this technique tends to produce better results in terms of analysis speed and accuracy. However, if unlearned sources are included in an input signal, the accuracy deteriorates. Even if sources with the same instrumental category are contained in a basis matrix, if the types (or models) are different, the input signal cannot be analyzed accurately because the spectra are different. For instance, a “Piano” category has several models, such as “Grand Piano,” “Upright Piano,” and “Electric Piano.” The supervised NMF decomposes a matrix relying completely on the same spectra in the basis matrix. In order to improve the decomposition accuracy, one solution is to prepare a basis matrix that contains all the possible kinds of spectral templates (not only various categories but also various models in each category). However, this is extremely difficult to build into a real system since the system does not always include exactly the same sources as those in the test data.

To solve this problem, we propose a novel method of mixed music analysis using a model-invariant feature—the probabilistic spectral envelope (PSE). In this paper, we particularly focus on a mixed music analysis method rather than music transcription, which means that we do not estimate the tempo, the note values, the dynamics markings, the key, and so forth, but we do estimate the pitch, the note-on time, and the instrumental labels. In this approach, all the spectral envelopes of various models belonging to the category are categorized in a corresponding PSE. This feature implies a probabilistic template of the spectral envelopes that belongs to an instrumental category, in a similar manner to the spectral template model [20]. Once the PSE is estimated, any spectrum of the category can be obtained in a random-sampling manner, by varying the shape of the spectral envelopes depending on the degree of variance. Therefore, even if unknown models are heard in an input musical signal, the sound can be analyzed using the preferred basis matrix of the generated spectra. In our approach, the PSE is simply estimated using a Gaussian process (GP)-based method called SPGP+HS [21] (heteroscedastic sparse pseudo-input Gaussian process). While Fujihara and Goto employed a spectral template model for vowels in a singing voice to estimate F0 and phonemes directly [20], our approach uses the PSE to categorize instrument sounds belonging to the same category to make it robust to an open test condition in NMF-based mixed music analysis.

Another important aspect of the proposed system is that the PSE of a certain category is distinguishable from the PSEs of other categories. Therefore, we can analyze a music signal that contains sounds from multiple instruments.

This paper is organized as follows. In Sect. 2, we introduce our proposed instrumental feature, the PSE. Then, we present a concrete way of estimating the PSE and analyzing a signal using the PSE in Sect. 3. We show our setup and experimental results in Sect. 4, and Sect. 5 is our conclusion.

2. PROBABILISTIC SPECTRAL ENVELOPE

The idea of the spectral envelope has been developed in speech analysis. Under this framework, any spectrum of the target signal (Fig. 1(c)) can be modeled as the product of the spectral envelope (Fig. 1(a)) using a harmonics filter (comb filter) (Fig. 1(b)). Note that a spectral envelope indicates the pitch-invariant features of a particular sound source, and a harmonics filter determines the pitch.

When we model the spectra of musical instrument sounds, the conventional (static) spectral envelope is not sufficient because in general the envelope varies with the pitch. For example, the spectral envelope of the C1 sound of the piano differs slightly from that of the C6 sound. Our proposed feature, the PSE, attempts to capture such variations in the instrumental category. Since we estimate the variations associated with all the pitches, this feature becomes independent of pitch.

We also consider that the PSE is independent of instrumental models belonging to the same category. An instrument’s spectrum differs slightly from that of other instruments due to various factors including the type of
instrument, the manufacturer, the materials used in its manufacture, the temperature, humidity, and playing style. However, the way the spectrum fluctuates is not completely random as it depends on the instrument’s category. The PSE, which is categorized by a mean envelope and a variance envelope, attempts to capture such fluctuations within the category. Some examples of PSEs are shown in Fig. 2. These examples were actually estimated using SPGP+HS [21], which is described in Sect. 3.1. There are four PSEs in Fig. 2: two different models for the ‘Piano’ category, and two different models for the ‘Strings’ category. The vertical axis indicates the linear amplitude (Fig. 4). Introducing the symbol \( E \) for the PSE, \( E \) is defined as

\[
E(\mu, \sigma^2) = \{e_f \mid e_f \sim \mathcal{N}(\mu_f, \sigma_f^2); f = 1, 2, \cdots, F\},
\]

where \( F \) is the number of frequency bins, \( \mu = \{\mu_f\} \) is the mean spectral envelope, and \( \sigma^2 = \{\sigma_f^2\} \) is the variance of the spectral envelope. \( e_f \) denotes a variable of the envelope at frequency \( f \) and has a normal distribution \( \mathcal{N}(\mu_f, \sigma_f^2) \).
In the training stage, the PSE analysis stage (to analyze a test signal using the PSEs).

![Flowchart of proposed method involving modeling of probabilistic spectral envelopes and analysis of mixed music signals using the envelopes.](image)

Since one PSE is estimated for each instrumental category, it involves spectral variances related to time. Spectra of the piano, for example, vary in time and are generally modeled separately at three different time points: at the attack time, at the sustain time, and at the release time. In our approach, we attempt to model all the time-varying spectra using the PSE.

As described above, the estimated PSEs are similar among various models of a particular instrumental category. Therefore, even if the PSE is estimated using only one model (e.g., “Grand Piano”), the spectra of other models (e.g., “Bright Piano”) that are similar to the training spectra may be generated from the PSE. It is thus expected that the PSE approach can analyze an input that includes an unknown source signal, in contrast to the conventional supervised NMF.

In most cases, if the instrumental categories are different, the PSEs also have different parameters (mean and variance envelopes) and thus give differently distributed spectra. This difference in the spectral distribution makes it possible to distinguish instrumental categories in a probabilistic manner. It is also possible that different PSEs generate almost the same spectrum, depending on the value of the variance. Fortunately this case rarely occurs since the probability of generating a spectrum is low on one side.

### 3. METHODOLOGY

In this section, we explain a practical method of mixed music analysis using a PSE, as shown in the flowchart in Fig. 5. Our approach can be broadly divided into two stages: the “training stage” (to estimate PSEs) and the “analysis stage” (to analyze a test signal using the PSEs). In the training stage, the PSE $E$ for each instrumental category is trained using prepared acoustic signals, where each musical note is played separately. Then, unsupervised NMF is applied to the magnitude spectrograms of the training signals. Since the training signals are very pure (the signals do not include other instrumental sources and each note is played at a different time) and the number of bases (sources) is known, the spectrogram can be correctly decomposed using NMF, obtaining the ideal basis matrix in which the column vectors contain the spectrum of each musical note. In the next step, all the spectral peaks (a pair of each harmonic frequency and its amplitude) are extracted from the obtained basis matrix. These peaks are input into the extended Gaussian process (SPGP+HS [21]) to estimate the PSE.

In the analysis stage, we use supervised NMF, which has an optimum basis matrix obtained by a random search. As described in the previous section, a spectrum with a certain fundamental frequency can be randomly generated from a PSE (Fig. 4). By carrying out this process for each instrument and fundamental frequency, we can obtain a candidate for the basis matrix. Then, a distance-based evaluation value (score) is calculated for this candidate to obtain the active matrix using supervised NMF. By repeating these steps several times, the optimum basis matrix with the smallest score can be found. Finally, we analyze the spectrogram of the test signal using supervised NMF with the optimum basis matrix.

#### 3.1. Training Stage

**3.1.1. Spectral peak extraction**

The first step in training a PSE is to extract spectral peaks, which are used for the next process (see the next subsection), from well-prepared training data. Each of the acoustic signals contains only the needed musical sources of the instrumental category. The various sources do not sound at the same time. In this paper, 12 half-tone sources sound in sequence every octave. Applying NMF to the amplitude spectrogram $V (\in \mathbb{R}^{F \times T})$ of the signal, $V$ is approximately decomposed into the product of a basis matrix $W (\in \mathbb{R}^{F \times R})$ and an activity matrix $H (\in \mathbb{R}^{R \times T})$ as follows:

$$V \approx WH$$  

$$\forall i,j,k, W_{ij} \geq 0, \quad H_{jk} \geq 0,$$

where $F$, $T$ and $R$ are the number of frequency bins, the time, and the number of bases, respectively (here, $F=4,097$, $T=172$, and $R=12$). $W_{ij}$ (or $H_{jk}$) indicates the element at the $ith$ ($j$th) row and $jth$ ($k$th) column in $W$ ($H$).

$W$ and $H$ can be obtained by iteratively calculating update rules based on Kullback-Leibler divergence. The update rules for each matrix element are

$$W_{ij} \leftarrow W_{ij} \frac{\sum_k H_{jk} V_{ik} / (WH)_{ik}}{\sum_k H_{jk}}$$  

$$H_{jk} \leftarrow H_{jk} \frac{\sum_i W_{ij} V_{ik} / (WH)_{ik}}{\sum_i W_{ij}}.$$
SPGP+HS fits the variances of the amplitude of the spectral peaks better than the standard GP. Thus, we use SPGP+HS for the PSE estimation in our approach, where the amplitude fluctuation of various frequencies is considered to be a feature of the instrumental category.

The PSE indicates the probability that the spectral envelope is generated. A GP-based method such as SPGP+HS provides an easy model for representing such two-dimensional probabilities, utilizing the mean and variance contours. However, the model allows negative values in the generated spectral envelope since the model is based on the Gaussian distribution. In our approach, we take zero values when we obtain negative values as described in the following subsection.

3.2. Analysis Stage

3.2.1. Generation of spectra based on PSE

A spectral envelope $e_c \in \mathbb{R}^{F \times 1}$ for category $c$ is randomly sampled from the PSE $E_c(\mu_c, \sigma^2)$, as described in Sect. 2. Then, a spectrum $p_{c,v} \in \mathbb{R}^{F \times 1}$ specified with a fundamental frequency $v$ along envelope $e_c$ can be calculated using Eq. (8).


given the fundamental frequency $v$ along envelope $e_c$, we take zero values when we obtain negative values as zero by taking the maximum function

$$p_{c,v} = \max(e_c, 0) \cdot \Psi_v(f),$$

where $\max(\cdot, \cdot)$ denotes an element-wise maximum function and $0$ is a zero vector with dimensions of $F \times 1$. $\Psi_v(f)$ is a comb filter with a fundamental frequency $v$, calculated as follows:

$$\Psi_v(f) = \sum_l \exp\left\{-\frac{(f - v \cdot l)^2}{2\lambda_0^2}\right\},$$

where $l$ is the index of the Gaussian components and $\lambda_0$ is a hyper-parameter used to determine the kurtosis of each component (here, $\lambda_0 = 5$).

Using the above procedure, we can obtain the intended basis matrix $\tilde{W} \in \mathbb{R}^{F \times K}$ by concatenating $K$ randomly generated spectra for various categories and fundamental frequencies as follows:

$$\tilde{W} = \{p_{c_1,v_1}, p_{c_1,v_2}, \ldots, p_{c_1,v_K}, p_{c_2,v_1}, \ldots, p_{c_N,v_K}\},$$

where $C$ is the number of instrumental categories.

3.2.2. Sparse NMF and basis matrix optimization

In the analysis stage, our aim is to find the optimum NMF matrices $\tilde{W}$ and $\tilde{H}$ for a test acoustic signal. Given an amplitude spectrogram $X \in \mathbb{R}^{F \times T}$ of a test signal and a randomly generated basis matrix candidate $\tilde{W}$, the activity matrix $\tilde{H}$ can be obtained by applying supervised sparse NMF [22], minimizing the following cost function $\mathcal{F}$:

$$\mathcal{F}(\tilde{W}, \tilde{H}) = d(X, \tilde{W} \tilde{H}) + \lambda \|\tilde{H}\|_1 \quad \text{s.t.} \quad \tilde{H} \geq 0.$$
In our experiments, we set generation and evaluation procedures, as shown in Fig. 7.

The first term on the right-hand side is the Kullback-Leibler divergence between $X$ and $\hat{W}\hat{H}$, which measures the distance between the input and the reconstructed spectrogram. The second sparsity term penalizes the nonzero elements in $\hat{H}$ using the $L_1$ norm with an experimentally determined weighting-sparsity parameter $\lambda$ (in our experiments, $\lambda = 5$). The activity matrix $\hat{H}$ that minimizes Eq. (11) can be estimated by iteratively calculating the following update rule:

$$\hat{H} \leftarrow \hat{H} \odot (\hat{W}^T(X \odot (\hat{W}\hat{H}))) \odot (\hat{W}^T 1 + \lambda),$$  \hspace{1cm} (12)

where $\odot$ and $\odot$ denote element-wise multiplication and division, respectively. The vector 1 is an all-one vector with dimensions of $F \times T$.

Since $\hat{W}$ determines $\hat{H}$ in this calculation, $\hat{H}$ can be considered to be a function of $\hat{W}$. If $\hat{W}$ contains better (more suitable) spectra for the test signal, the cost function $\mathcal{F}(\hat{W}, \hat{H})$ must become smaller. Therefore, the minimization of the cost function can be used as a procedure for finding the optimum basis matrix $\hat{W}$. The optimum matrices are finally obtained by repeating the generation and evaluation procedures, as shown in Fig. 7.

In our experiments, we set $S = 10$. Since the row vectors in the activity matrix are arranged in order of a musical scale, the estimated activity matrix has a piano-roll-like representation.

4. EXPERIMENTS

4.1. Single-Instrument Music Analysis

4.1.1. Setup

To evaluate our proposed method, “wav-to-mid” tests were conducted. First, we used three songs (listed in Table 1) from the RWC music database [23] that contain only the sound of a single instrument to investigate how the PSE works in the case of unknown instrumental models. Each song has a different number of notes (‘Notes’ in the table), and is played only by the piano (originally played by the instruments in the brackets; ‘Pf,’ ‘Fl,’ and ‘Gt’ indicate piano, flute, and guitar, respectively). Because it makes no sense to use the whole song, in which similar phrases are repeated, each MIDI file contains only one phrase (‘Bars’ indicates the interval of the used bars). Acoustic signals used in the tests were obtained by internally recording each MIDI file at a 16 kHz sampling rate using the instrumental sounds (models) listed in Tables 2 and 3.

<table>
<thead>
<tr>
<th>No.</th>
<th>Catalog</th>
<th>Time (T)</th>
<th>Notes</th>
<th>Bars</th>
<th>Inst.</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>RM-C043</td>
<td>0:29 (895)</td>
<td>128</td>
<td>2–10</td>
<td>Pf(Pf, Fl)</td>
</tr>
<tr>
<td>#2</td>
<td>RM-J003</td>
<td>0:29 (922)</td>
<td>192</td>
<td>17–27</td>
<td>Pf</td>
</tr>
<tr>
<td>#3</td>
<td>RM-J009</td>
<td>0:32 (1004)</td>
<td>134</td>
<td>2–17</td>
<td>Pf(Gt)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Model name</th>
<th>Module</th>
<th>Program no. (or name)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piano</td>
<td>‘Piano 0’</td>
<td>blt. Garage Band</td>
<td>Grand Piano</td>
</tr>
<tr>
<td></td>
<td>‘Piano 1’</td>
<td>EXS24 mkII (blt. Logic)</td>
<td>Yamaha Grand Piano</td>
</tr>
<tr>
<td></td>
<td>‘Piano 2’</td>
<td>Roland SD-20</td>
<td>0 (Grand Piano)</td>
</tr>
<tr>
<td></td>
<td>‘Piano 3’</td>
<td>Roland SD-20</td>
<td>1 (Bright Piano)</td>
</tr>
<tr>
<td>Flute</td>
<td>‘Flute 0’</td>
<td>blt. Garage Band</td>
<td>Pop Flute</td>
</tr>
<tr>
<td></td>
<td>‘Flute 1’</td>
<td>Roland SD-20</td>
<td>73 (Flute)</td>
</tr>
<tr>
<td></td>
<td>‘Flute 2’</td>
<td>Roland SD-20</td>
<td>72 (Piccolo)</td>
</tr>
<tr>
<td></td>
<td>‘Flute 3’</td>
<td>Roland SD-20</td>
<td>74 (Recorder)</td>
</tr>
<tr>
<td>Strings</td>
<td>‘Strings 0’</td>
<td>blt. Garage Band</td>
<td>Strings Ensemble</td>
</tr>
<tr>
<td></td>
<td>‘Strings 1’</td>
<td>EXS24 mkII (blt. Logic)</td>
<td>Strings Ensemble</td>
</tr>
<tr>
<td></td>
<td>‘Strings 2’</td>
<td>Roland SD-20</td>
<td>48 (Strings)</td>
</tr>
<tr>
<td></td>
<td>‘Strings 3’</td>
<td>Roland SD-20</td>
<td>40 (Violin)</td>
</tr>
</tbody>
</table>
The accuracy was calculated as the ratio between the number of correct notes and the total number of notes. This is expressed as follows:

\[ \text{acc} = \frac{N_{\text{hit}} - (N_{\text{ins}} + N_{\text{del}})}{N_{\text{total}}} \times 100. \]  

For the PSE (‘PSE(open)’) estimation, we prepared MIDI recording of three different piano models that were not used in the test signal (e.g., ‘Piano 1,’ ‘Piano 2,’ and ‘Piano 3’ for the ‘Set 0’ experiment, listed in Table 2) in O2–O6 (On denotes a set of 12 half-tone notes in the nth octave \{C_n, C\#_n, \ldots, B_n\} in MIDI). Because in these experiments, we wanted to examine whether on estimated notes are estimated by the method in [2]. Then, MIDI data is generated by combining the information on W and H. For the two comparative methods based on supervised NMF, the systems store exemplars of individual notes of the piano (12 \times 5 = 60 notes from C2 to B6 in total). The difference is that sp-NMF(closed) stores the same model exemplars as in the experiment, whereas sp-NMF(open) stores the other model exemplars (‘Piano 1,’ ‘Piano 2,’ and ‘Piano 3’ for ‘Set 0,’ for example). In the unsupervised approach, after estimating W and H, fundamental frequencies corresponding to each column vector of W are estimated by the method in [2]. Then, MIDI data is generated by combining the information on W and H. For the two comparative methods based on supervised NMF, the systems store exemplars of individual notes of the piano (12 \times 5 = 60 notes from C2 to B6 in total). The difference is that sp-NMF(closed) stores the same model exemplars as in the experiment, whereas sp-NMF(open) stores the other model exemplars (‘Piano 1,’ ‘Piano 2,’ and ‘Piano 3’ for ‘Set 0,’ for example). The number of bases R equals the number of half-tones multiplied by the number of models (R = 60 \times 3 = 180 for the open test and R = 60 for the closed test).

Each of the methods was systematically evaluated in terms of the accuracy rate of its MIDI outputs, which were obtained by binarizing the activity matrix with a threshold \( \eta \). The accuracy was calculated as the ratio between the number of error notes (insertion errors \( N_{\text{ins}} \) and deletion errors \( N_{\text{del}} \)) and the total number of notes \( N_{\text{total}} \) as follows:

\[ \text{acc} = \frac{N_{\text{hit}} - (N_{\text{ins}} + N_{\text{del}})}{N_{\text{total}}} \times 100. \]  

Because the onset time and the duration of each sound source are not necessarily correct in the above binarizing process, we permitted the duration to differ and the onset time to shift \( \tau \) seconds (in this paper, \( \tau = 0.3 \)). The threshold \( \eta \) was determined by taking the best accuracy for each method and changing the value from 1.5 to 2.5 in steps of 0.1.

### 4.1.2. Results and discussion

Figure 8 shows the average accuracy rate of each model for each model set in five trials. Error bars in the figure indicate the maximum and minimum values among the trials. Concerning the results of the conventional methods, if the system encounters exactly the same sounds as those in the test signal, it yields a high performance (sp-NMF(closed)). However, if the system does not encounter the same sounds, the accuracy deteriorates (sp-NMF(open) and un-NMF). Meanwhile, our approach (PSE(open)) performed well even when the system does not learn the sounds of the test data in any song, maintaining accuracies that are quite close to the source-encountering method (PSE(closed)). The favorable results are due to the assumption that the PSE is a model-invariant feature and once it is estimated, it can be used to estimate the spectral envelopes of unknown models. Among our approaches, PSE(closed) performed slightly better than PSE(open) overall owing to the fact that it is closed. However, the accuracy of PSE(closed) for song #3 in ‘Set 0’ was, for example, lower than that of sp-NMF(closed) even though PSE(closed) is a closed model. We consider that this is due to the limitation of our method, which is based on a probabilistic model.

An analysis example for PSE(open) and PSE(closed) in ‘Set 0’ is shown in Figs. 9(a) and 9(b), respectively, when song #1 is analyzed (Fig. 10). Almost all the notes were estimated correctly with both approaches, but some notes were mistaken as octave-different notes, especially in PSE(open). This is because PSE(closed) captures the harmonics structure more accurately than PSE(open) and it has fewer errors related to octave differences.

### 4.2. Multiple-Instrument Music Analysis

#### 4.2.1. Setup

Next, we discuss multiple-instrument analysis experiments. In these experiments, we found that our approach is efficient for the analysis of multiple-instrument music signals. These experiments are more difficult than the previous experiments because we have to estimate instrumental category labels for each note as well as the pitches and onset time. Unsupervised NMF (un-NMF) does not provide the labels; therefore, it was not compared in these experiments.

The three songs used for the experiments include piano (‘PF’), flute (‘FI’), or strings (‘St’), as shown in Table 4.
Similar to the single-instrument experiments, we conducted four experiments, each of which used different models for the test acoustic signals as listed in Tables 2 and 3. For example, in the ‘Set 0’ experiment, the models ‘Piano 0,’ ‘Flute 0,’ and ‘Strings 0’ were used in the test signals and closed methods (PSE(closed) and sp-NMF(closed)), and the other models (‘Piano 1,’ ‘Piano 2,’ ‘Piano 3,’ ‘Flute 1,’ ‘Flute 2,’ ‘Flute 3,’ ‘Strings 1,’ ‘Strings 2,’ and ‘Strings 3’) were used for training the models in the open methods (PSE(open) and sp-NMF(open)). We also list which notes are included in each song in Table 5. In our experiments, this pitch range (octave) information is

\[
\begin{array}{cccccccc}
\text{No.} & \text{Catalog} & \text{Time (T)} & \text{Notes} & \text{Bars} & \text{Instruments} \\
\hline
\#4 & RM-C043 & 0:29 (895) & 128 & 2–10 & Pf, Fl \\
\#5 & RM-C046 & 0:21 (661) & 361 & 12–23 & Pf, Fl \\
\#6 & RM-C020 & 1:00 (1860) & 485 & 3–30 & Pf, Fl(Vl), St(Vc) \\
\end{array}
\]
known for each method; we generate a basis matrix within
the octave range that covers all pitches shown in Table 5
for each song. For example, the bases in the octave range
\(O_2\)–\(O_4\) (36 bases) and in the octave range
\(O_4\)–\(O_5\) (24 bases) are used for the piano and flute for song #4,
respectively. In this case, the number of bases for sp-
NMF(open) becomes
\[ R = \frac{3}{3}(36 + 24) = 180 \]
in total because three different models are used, and the number
of bases for sp-NMF(closed) is \( R = 36 + 24 = 60 \). In our
approach (PSE(open) and PSE(closed)), we generated \( R = 36 + 24 = 60 \) bases for the analysis of song #4. The other
experimental conditions were the same as in the previous
experiments (e.g., we repeated our method with \( S = 10 \) five
times and computed the mean, maximum, and minimum
accuracies).

### 4.2.2. Results and discussion

Figure 11 summarizes the MIDI-conversion results for
each method, where multiple-instrument sounds were
included in the analyzed music. Similar to the results
of single-instrument analysis, the accuracies of sp-NMF
seriously decrease for unknown models. On the other
hand, our PSE method maintains high accuracy even for
open tests. The reason why our approach obtained highly
accurate results is the same as that for the case of single-
instrument analysis: the PSE feature is model-invariant.

Figure 12 illustrates the analysis results (estimated \( \hat{H} \))
of song #4, which includes piano and flute sounds in
‘Set 0,’ comparing our approach (a) and sp-NMF (b),
which are both ‘closed.’ The analysis results for sp-NMF
are similar to the correct piano roll (Fig. 13) except for a
few errors, since exactly the same sound sources as those
used in the test are given in the basis matrix. In our
approach, there were more errors than with sp-NMF; some
piano notes were taken to be the flute and vice versa. One
of the reasons for this is the limitation of our approach,
which is based on an optimum search algorithm that finds
the optimum matrices to minimize the cost function \( \mathcal{F} \),
as shown in Fig. 7. The optimum matrices, which minimize
the cost function, do not always provide the highest
accuracy. From Fig. 14, we see that the accuracy does not
change greatly even as the number of samplings increases.

Nonetheless, our approach can partially separate piano
and flute notes because the estimated PSEs of a piano and
flute are roughly distinguishable from each other (as shown
in Fig. 15), which contributes to the separation of sources.

### 5. CONCLUSION

In this paper, we proposed an algorithm for monaural
musical source decomposition and multiple-pitch estima-

<table>
<thead>
<tr>
<th>No.</th>
<th>Catalog</th>
<th>Piano</th>
<th>Flute</th>
<th>Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>#4</td>
<td>RM-C043</td>
<td>D2–A7</td>
<td>D4–A7</td>
<td>—</td>
</tr>
<tr>
<td>#5</td>
<td>RM-C046</td>
<td>F2–C6</td>
<td>F4–A7</td>
<td>—</td>
</tr>
<tr>
<td>#6</td>
<td>RM-C020</td>
<td>E1–B4</td>
<td>F3–G7</td>
<td>F3–B4</td>
</tr>
</tbody>
</table>

Fig. 11 Analysis accuracies for each method for multiple-instrument songs.
tion. The method categorizes several spectral envelopes for each musical category, inspired by the invariance of spectral fluctuation in a category. This categorized envelope, called the probabilistic spectral envelope (PSE), has the characteristic of being able to absorb differences between models, pitches, modules, and so forth. The PSE consists of a mean envelope and variance envelope that can be simultaneously estimated by SPGP+HS regression as described in this paper. In the analysis stage, an iterative procedure combining sparse-constrained supervised NMF with PSE-distributed random sampling was employed to search for optimum NMF matrices, which were then used for NMF-based signal analysis.

The simulation experiments using MIDI sources show that the proposed method is robust to instrumental model changes. Since the results depend on random values, however, future research will include designing a direct optimum search method, such as maximum likelihood (ML) or maximum a posteriori (MAP) estimations.

The PSE presented in this paper allows negative values due to the effect of the variance; this might not be a suitable model when dealing with an NMF framework, where the matrices have non-negative values. We will also investigate other models to represent the probabilities of spectral envelopes with positive values.

REFERENCES


Toru Nakashika received his B.E. and M.E. degrees in computer science from Kobe University in 2009 and 2011, respectively, after which he continued his research as a doctoral student. From September 2011 to August 2012 he studied at INSA de Lyon in France. He is currently a 3rd-year doctoral student at Kobe University. His research interests include speech, music, and image signal processing and pattern recognition. He is a member of IEEE, ISCA, and ASJ.

Tetsuya Takiguchi received his B.S. degree in applied mathematics from Okayama University of Science, Okayama, Japan, in 1994, and his M.E. and Dr. of Engineering degrees in information science from Nara Institute of Science and Technology, Nara, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a researcher at IBM Research, Tokyo Research Laboratory, Kanagawa, Japan. He is currently an associate professor at Kobe University. His research interests include acoustic and image signal processing and pattern recognition. He received the Awaysa Award from the Acoustical Society of Japan in 2002. He is a member of IEEE, IPSJ, and ASJ.

Yasu o Ariki received his B.E., M.E., and Ph.D. degrees in information science from Kyoto University in 1974, 1976, and 1979, respectively. He was an assistant professor at Kyoto University from 1980 to 1990 and stayed at Edinburgh University as a visiting academic from 1987 to 1990. From 1990 to 1992 he was an associate professor and from 1992 to 2003 a professor at Ryukoku University. Since 2003 he has been a professor at Kobe University. He is mainly engaged in speech and image recognition and is interested in information retrieval and databases. He is a member of IEEE, IPSJ, JSIA, ITE, and IIEEJ.