Emphasis of Sound Sources Using Nonnegative Matrix Factorization Based on Stereo Localizations

Kento Yoshida and Yasue Mitsukura

Keio University
3-14-1 Hiyoshi, Kohoku, Yokohama, Kanagawa 223-8522, Japan
Phone/FAX: +81-45-566-1718
E-mail: yoshida@mitsu.sd.keio.ac.jp, mitsukura@sd.keio.ac.jp

Abstract

In this study, we emphasized sound sources using nonnegative matrix factorization (NMF) with the likelihood of the directions obtained from the estimated localizations. NMF is one of the most commonly used techniques for the separation of music signals. The emphasis of sound sources by NMF has several problems: the uncertainty of the decomposition of the bases corresponding to the sources and the necessity to group the obtained bases according to the sources. These problems have to be solved simultaneously to improve the separation performance. In order to simultaneously solve these problems, we proposed a novel NMF incorporating the cost based on the likelihood of the localizations. According to the experimental results, we could improve the separation performance by using our proposed method.

1. Introduction

The emphasis or separation of sound sources is widely studied as a basic technique in many applications related to sound. For example, in music, remixing instrument sounds as desired by users or secondary fabrication can be implemented by using the separated sounds.

Various methods for the emphasis or separation of sound sources have been proposed. In studies on speech signals, the separation based on localizations is performed by assuming the W-disjoint orthogonality (W-DO) of sources in the time-frequency (T-F) domain [1]. The localization is the arrival direction of a source signal relative to an observation channel. This method realizes the separation by making a binary mask that moves the T-F components in a certain direction. This method is, however, not suitable for the separation of music signals due to the failure of W-DO; music components are prone to superpose with each other in the T-F domain.

Nonnegative matrix factorization (NMF) [2] is one of the most commonly used techniques for the separation of music signals. NMF is a technique used to decompose a matrix consisting of nonnegative components, such as a power spectrum, into some bases, which show the frequent patterns in the matrix, and activations, which show how the corresponding bases appear in the matrix. Music signals can be regarded as the superposition of a finite number of music notes from multiple sources. The separation based on NMF is performed by estimating bases that correspond to the spectra of notes by assuming the superposition of the spectra in the T-F domain. Several problems occur in the emphasis or separation based on NMF.

One is the uncertainty of the decomposition of the bases corresponding to the sources. NMF has a tendency to represent co-occurring components by a single basis. On the other hand, the spectrum of a note has a harmonic structure with peaks at frequencies that are integer multiples of the fundamental frequency. The decomposition of the bases corresponding to notes is expected by regarding the components at these frequencies as co-occurring ones. However, when components from multiple sources always co-occur, these components are expressed by one basis, resulting in the failure of separation.

The necessity to group the obtained bases according to the sources is another problem. Various studies have been conducted in order to automatically perform the grouping by using the characteristics of the obtained bases or activations [3], [4]. However, it is considered that the performance of the grouping is highly dependent on the accuracy of the decomposition of the bases. It is necessary to simultaneously solve various problems to realize the separation based on NMF.

In a previous attempt to simultaneously solve the problems of NMF, Takeda, et al. proposed a framework of multichannel complex NMF incorporating the concept of localizations obtained from the observed signals at multiple channels, assuming W-DO of the sources [5].

In this study, we separate music signals using NMF with the likelihood related to the directions of T-F components obtained from the estimated localizations. In the proposed method, we incorporate a cost based on the likelihood into NMF. We aim to decompose the T-F components from different directions into different bases. The automatic grouping
of the bases is also tackled simultaneously by introducing the concept of the localizations of the bases. We use stereo information in order to obtain the likelihood in this paper. Actually, we can apply the proposed method to single-channel signals by using another likelihood. The proposed method also has flexibility in how to determine the likelihood.

2. Proposed Method

2.1 Likelihood of the localizations

The localizations are estimated using stereo signals in this paper. A short-time Fourier transform (STFT) is applied to the stereo signals to obtain T-F representations, that is, spectrograms. From the phase differences between the corresponding T-F components on the spectrograms, the time delay \( \tau_{f,t} \) is calculated as follows:

\[
\tau_{f,t} = \frac{\theta_{f,t}}{\omega_{f,t}}
\]

where \( \theta \) is the phase difference and \( \omega \) is the frequency. The subscripts show the \( f \)th frequency bin and the \( t \)th frame component in the spectrogram. The localization \( \phi_{f,t} \) is estimated as follows:

\[
\phi_{f,t} = \sin^{-1}\left(\frac{ct_{f,t}}{d}\right)
\]

where \( c \) is the velocity of sound and \( d \) is the distance between the observation channels.

Using the estimated localizations, we determine the likelihood regarding the directions. We employ the following Gaussian distribution in this study.

\[
K_{\text{gauss}}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)
\]

The likelihood \( p_{f,t}(\phi) \) is obtained as follows:

\[
p_{f,t}(\phi) = \frac{1}{\kappa} K_{\text{gauss}}\left(\frac{\phi - \phi_{f,t}}{\kappa}\right)
\]

where \( \kappa \) is the bandwidth parameter. \( \kappa \) is usually determined experimentally.

Here, we also determine the candidate localizations of the sources. The total distribution composed of all likelihoods is described as

\[
p_{\text{total}}(\phi) = \frac{1}{N} \sum_{f,t} p_{f,t}(\phi)
\]

where \( N \) is the total number of T-F components in the spectrogram. The candidate localization \( \hat{\phi}_i \) is determined as the localization that maximize \( p_{\text{total}}(\phi) \).

\[
\hat{\phi}_{\text{candidate}} = \{ \hat{\phi}_i \}
\]

2.2 NMF process

NMF determines bases and activations that minimize the distance between the original nonnegative matrix and the matrix reconstructed from the estimated bases and activations. Several distance functions exist. We employ Itakura-Saito (IS) divergence as a distance function and the power spectrogram as a nonnegative matrix in this paper. The distance function based on IS divergence between two variables \( y \) and \( x \) is described as follows:

\[
d_{\text{IS}}(y, x) = \frac{y}{x} - \log\frac{y}{x} - 1
\]

where the value becomes zero when \( x \) is equal to \( y \). IS divergence depends only on the ratio of \( y \) to \( x \). It is known that IS divergence is suitable for the decomposition of sound signals that have scale differences between the frequency bands. We aim to minimize the cost function using the likelihood in addition to the distance function. This problem is described as follows:

\[
\min \left\{ \sum_{f,t} \frac{y_{f,t}}{x_{f,t}} \log \frac{y_{f,t}}{x_{f,t}} - 1 \right\} + \beta \sum_{k} \left(1 - \hat{p}_{f,t}(\phi_k)\right) w_{f,t} h_{k,t}
\]

where \( w \) and \( h \) are the basis and activation components, respectively, \( k \) indicates the \( k \)th basis, \( y \) is the original nonnegative component, and \( x \) is the reconstructed component given by the following equation.

\[
x_{f,t} = \sum_{k} w_{f,t} h_{k,t}
\]

\( \hat{p}_{f,t}(\phi) \) is the normalized likelihood:

\[
\hat{p}_{f,t}(\phi) = \frac{1}{M_{p,t}} p_{f,t}(\phi)
\]

where

\[
M_{p,t} = \max p_{f,t}(\phi)
\]

\( \phi_k \) is the localization of the \( k \)th basis. In order to obtain \( \phi_k \), we define the distribution of the \( k \)th basis \( p_k(\phi) \) as follows:

\[
p_k(\phi) = \frac{w_{f,t} h_{k,t}}{\sum_{f,t} w_{f,t} h_{k,t} p_{f,t}(\phi)}
\]

The provisional localization \( \hat{\phi}_k \) is determined as the localization that maximizes \( p_k(\phi) \) as follows:

\[
\hat{\phi}_k = \arg \max_{\phi} p_k(\phi)
\]
The localization \( \phi_k \) is determined by assigning \( \hat{\phi}_k \) to the closest candidate localization as follows.

\[
\phi_k = \arg \min_{\phi_i} |\hat{\phi}_k - \hat{\phi}_i| \tag{14}
\]

\( \beta \) in Eq. (8) is a parameter indicating the balance between the first term and the second term. The first term in Eq. (8) is the distance based on IS divergence between the original and reconstructed matrices. The second term is the cost regarding the likelihood that we defined. The cost decreases as the localizations of the bases and the T-F components on the reconstructed matrices become closer.

By solving Eq. (8), we obtain the following updating formulas for the basis and activation components.

\[
w_{f,k} \leftarrow w_{f,k} \frac{\sum_i y_{f,i} h_{k,i}}{\sum_k w_{f,k} h_{k,i}^2} \tag{15}
\]

\[
h_{k,t} \leftarrow h_{k,t} \sqrt{\frac{\sum_f \left( w_{f,k} h_{k,t} + \beta \left( 1 - \hat{\rho}_{f,j}(\phi_k) \right) y_{f,t} \right)^2}{\sum_k w_{f,k} h_{k,t}^2}} \tag{16}
\]

\( \phi_k \) is also updated by Eqs. (12)-(14) in an iterative process.

2.3 Reconstruction of the separated spectrogram

We group bases according to \( \phi_k \) after the final update. We determine the \( i \)-th set of basis indices \( \Gamma_i \) as follows:

\[
\Gamma_i = \{ k | \phi_k = \hat{\phi}_i \} \tag{17}
\]

The \( i \)-th reconstructed matrix \( \hat{y}_i \) is estimated as

\[
\hat{y}_i^{f,j} = \frac{\sum_{k\in \Gamma_i} w_{f,k} h_{k,i}}{\sum_k w_{f,k} h_{k,i}^2} y_{f,j} \tag{18}
\]

We can obtain the separated signals by applying an inverse STFT to the reconstructed spectrogram using the original phase information.

3. Experiments

We conducted an experiment on the separation of sound sources. We used stereo music signals made by MIDI software as experimental signals. The signals consisted of three sources whose localizations differed from each other; sources 1, 2 and 3 had left, center and right localizations, respectively. The localizations are roughly set in the software.

The scores of the sources are shown in Fig. 1. Due to the co-occurrence of sound components and the superposition in the T-F domain, it is considered to be difficult to separate the sources using ordinary NMF.

We compared the separation performances of ordinary NMF and the proposed method. The grouping of bases for ordinary NMF was performed to estimate the localizations of the obtained bases as in the proposed method. We employed two sound quality indices for the evaluation of the estimated signals: the source-to-distortion ratio (SDR) and the source-to-interference ratio (SIR) [6]. SDR and SIR indicate the degrees of comprehensive distortion and the distortion by the interference, respectively.

The experimental conditions or parameters are shown in Table 1.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Figure 1: Scores of sources}
\]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1: Experimental conditions**

<table>
<thead>
<tr>
<th>Signal</th>
<th>Sampling rate</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>44.1 kHz</td>
<td>4 s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STFT</th>
<th>Window size</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50 ms</td>
<td>25 ms</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>c</th>
<th>d</th>
<th>( \kappa )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>340 m/s</td>
<td>0.75 cm</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NMF</th>
<th>Number of bases</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>

The scores of the sources are shown in Fig. 1. Due to the co-occurrence of sound components and the superposition in the T-F domain, it is considered to be difficult to separate the sources using ordinary NMF.

We compared the separation performances of ordinary NMF and the proposed method. The grouping of bases for ordinary NMF was performed to estimate the localizations of the obtained bases as in the proposed method. We employed two sound quality indices for the evaluation of the estimated signals: the source-to-distortion ratio (SDR) and the source-to-interference ratio (SIR) [6]. SDR and SIR indicate the degrees of comprehensive distortion and the distortion by the interference, respectively.

The experimental conditions or parameters are shown in Table 1. \( d \) is determined as the distance where aliasing does not occur for the sampling rate. \( \kappa \) is determined experimentally so that the number of peaks in the total distribution becomes equal to the number of sources. The other conditions are also determined experimentally.
4. Results

Table 2 shows the separation performances expressed as the indices SDR and SIR (dB) along with the values before the separation. A higher value indicates a superior performance. According to the results, the performance is particularly poor for the separation of source 3 by ordinary NMF. It is considered to be difficult to separate source 3 from the other sources, especially from source 2, due to the co-occurrence and superposition. However, the separation performance for source 3 was greatly improved using the proposed method. Other performances also improved. The greatest improvement was almost +12 dB for the SDR of source 3. It is considered that our method performed better estimation due to the simultaneous decomposition of the bases and grouping.

5. Conclusions

In this paper, we proposed a novel emphasis algorithm using NMF with the likelihood related to the directions of T-F components obtained from the estimated localizations. According to the experimental results, we could greatly improve the separation performance, by up to +12 dB for SDR, and simultaneously solve the problems of NMF.

We consider that the separation performance is dependent on the determination of the likelihood. We would like to attempt to determine a better likelihood and improve the separation performance in future works. We should also compare the proposed method with multichannel complex NMF.

References


