UNSUPERVISED BEAMFORMING BASED ON MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION FOR NOISY SPEECH RECOGNITION

Kazuki Shimada, Yoshiaki Bando, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, Tatsuya Kawahara

Kyoto University, School of Informatics, Sakyo-ku, Kyoto, 606-8501 Japan
{shimada, bando, mimura, itoyama, yoshii, kawahara}@sap.ist.i.kyoto-u.ac.jp

ABSTRACT
This paper presents unsupervised multichannel speech enhancement for noisy speech recognition. Time-frequency (TF) mask estimation has actively been studied for estimating the steering vectors and spatial covariance matrices of speech and noise used for beamforming. The state-of-the-art approach to mask estimation is to use deep neural networks (DNNs) for classifying the TF bins of observed signals into speech and noise. Such a supervised approach, however, does not work well in an unknown environment. To accurately estimate the spatial covariance matrices in an unsupervised manner, we perform blind source separation (BSS) based on multichannel nonnegative matrix factorization (MNMF) for decomposing each TF bin into the components of speech and the other sources (noise). To clarify a suitable type of beamforming for MNMF, we tested both time-invariant and time-varying versions of the minimum variance distortionless response (MVDR) beamforming in addition to standard multichannel Wiener filtering (MWF). The experimental results showed that our MNMF-based beamforming approach outperformed the state-of-the-art DNN-based beamforming method in unknown environments that do not match the training data.

Index Terms— Noisy speech recognition, speech enhancement, multichannel nonnegative matrix factorization, beamforming

1. INTRODUCTION
Multichannel speech enhancement based on beamforming has intensively been studied for automatic speech recognition (ASR) in real environments. Using beamforming techniques, we can emphasize target speech coming from one direction and suppress noise coming from the other directions [1]. Recent competitions such as CHiME Challenge [2] showed the effectiveness of beamforming as preprocessing for ASR in adverse noisy conditions [3]. There are several variants of beamforming methods such as minimum variance distortionless response (MVDR) beamforming [1], generalized sidelobe canceller (GSC) [4], multichannel Wiener filtering (MWF) [5], and generalized eigenvalue (GEV) beamforming [6]. To use these methods formulated in the time-frequency (TF) domain, it is necessary to calculate linear filters based on the steering vectors and spatial covariance matrices of speech and noise [7–12].

A great deal of effort has been devoted to estimating the steering vectors and spatial covariance matrices corresponding to speech and noise. Conventional methods based on steered response power phase transform (SRP-PHAT) [13] and weighted delay-and-sum beamforming [14] are insufficient for ASR in real environments [2]. Recently, TF masking has been shown to improve the performance of ASR [7–12]. This approach is based on the assumption that each bin of an observed spectrogram is exclusively classified into two categories (i.e., speech and noise) [7–12]. The spatial covariance matrices of the target speech and noise can be calculated from the classified TF bins [7–12]. The steering vector of the target speech is then obtained by calculating the principal component of the spatial covariance matrix [7–9]. One approach to such binary classification is to use an unsupervised method based on complex Gaussian mixture models (CGMMs) [7]. In recent years, the most popular method is to use deep neural networks (DNNs) for estimating TF masks without using phase information [8–12]. To train DNNs in a supervised manner, however, sufficiently many pairs of noisy spectrograms and ideal binary masks (IBMs) are required.

One of the problems of the conventional mask estimation is that the phase information of each source is not adequately dealt with although the phase information plays an essential role in various kinds of multichannel audio signal processing. Another major problem of the DNN-based mask estimation is that the performance of ASR in unknown environments that are not covered by the training data would be considerably degraded because DNNs easily overfit to the training data. Several studies related to CHiME Challenge [15] suggested that multi-condition DNN training with various kinds of noise data mitigate the problem, but it is still an open question whether it is robust even when a microphone array with different frequency characteristics is used in unseen noisy environments. This calls for an unsupervised method that can estimate the phase of speech and noise.

In this paper we propose several variants of unsupervised speech enhancement that estimate both the spatial covariance matrices of speech and noise for beamforming by using a blind source separation (BSS) method called multichannel nonnegative matrix factorization (MNMF) [16–18] (Fig. 1). Given complex spectrograms of multichannel mixture signal, MNMF can estimate the spatial covariance matrices of individual sources as well as approximating the power
Table 1. Relationship between MVDR and MWF.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Likelihood Estimation</td>
<td>MVDR</td>
<td>in Eqs. (3) &amp; (4)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Maximum A Posteriori Estimation</td>
<td>Rank-1 MWF</td>
<td>in Eqs. (5) &amp; (6)</td>
<td>Full-rank MWF</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2. Notation on steering vector and spatial covariance matrix.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Target speech</th>
<th>Noise</th>
<th>Source l</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steering vector</td>
<td>p</td>
<td>r_l</td>
<td>l</td>
</tr>
<tr>
<td>Spatial covariance matrix</td>
<td>P</td>
<td>Q</td>
<td>R_l</td>
</tr>
</tbody>
</table>

where \( s_{ft} \in \mathbb{C} \) is a single target speech signal, \( p_f \in \mathbb{C}^M \) is a steering vector of the target speech, and \( n_{ft} \in \mathbb{C}^M \) is noise signal. \( s_{ft} = p_f s_{ft} \in \mathbb{C}^M \) is a target speech image, and the target speech is assumed to propagate in the rank-1 propagation process. The noise signal follows a Gaussian distribution with a mean of \( 0 \in \mathbb{C}^M \) and a full-rank spatial covariance matrix of the noise \( Q_f \in \mathbb{C}^{M \times M} \), and the noise is assumed to propagate in the full-rank propagation process. The observed signal also follows a Gaussian distribution: \( x_{ft} \sim \mathcal{N}(p_f s_{ft}, Q_f) \). Then, the residual noise is minimized with the constraint such that any signals from the target speech direction remain distortionless. It leads to the following widely-used time-invariant MVDR beamforming filter:

\[
\mathbf{w}_{ft}^{\text{MVDR}} = \frac{Q_f^{-1} p_f}{p_f^T Q_f^{-1} p_f},
\]

which is also obtained by maximizing the likelihood function \( p(x_{ft}|s_{ft}) \) without using an assumption on the target speech [22]. Furthermore, we can assume that the spatial covariance matrix is time-varying (\( Q_f \rightarrow Q_{ft} \)). Accordingly, the time-varying MVDR beamforming filter is obtained:

\[
\mathbf{w}_{ft}^{\text{MVDR}} = \frac{Q_{ft}^{-1} p_f}{p_f^T Q_{ft}^{-1} p_f}.
\]

With the same assumption on the propagation as in the MVDR beamforming, rank-1 MWF beamforming assumes the target speech follows a Gaussian distribution: \( s_{ft} \sim \mathcal{N}(0, \phi_f) \). Let \( \phi_f \in \mathbb{C} \) be a variance of the target source. Accordingly, the rank-1 MWF beamforming filter is obtained by maximum a posteriori estimation that maximizes \( p(s_{ft}|x_{ft}) \):

\[
\mathbf{w}_{ft}^{\text{1MWF}} = \frac{Q_{ft}^{-1} p_f}{p_f^T Q_{ft}^{-1} p_f + \phi_f^{-1}},
\]

\[
\mathbf{w}_{ft}^{\text{1MWF}} = \frac{Q_{ft}^{-1} p_f}{p_f^T Q_{ft}^{-1} p_f + \phi_f^{-1}}.
\]

When the assumption is not used (or the variance of the target source \( \phi_f \rightarrow \infty \) in Eqs. (5) and (6)), the rank-1 MWF beamforming is equivalent to the MVDR beamforming [5].

We can also use full-rank MWF for spatial filtering. A multichannel observed signal is assumed to be as follows:

\[
x_{ft} = \sum_{l=1}^{L} x_{ftl},
\]

where \( x_{ftl} \in \mathbb{C}^M \) is an source image from a source \( l \) picked up with an \( M \)-channel microphone array. The \( l \)-th source image \( x_{ftl} \) follows a Gaussian distribution with a mean of \( 0 \) and a full-rank spatial covariance matrix \( R_{ftl} \in \mathbb{C}^{M \times M} \). We regard \( l = 1 \) is the index for the target speech and the rest are noise. The target speech image \( x_{ft1} = x_{ftl} \) follows a Gaussian distribution with a full-rank spatial covariance matrix of the target speech \( P_f = R_{f1} \in \mathbb{C}^{M \times M} \);
\( s_{ft} \sim \mathcal{N}(0, P_f) \). The noise signal \( n_{ft} = \sum_{l\neq f} x_{fl} \) follows a Gaussian distribution with the full-rank spatial covariance matrix \( Q_f = \sum_{l \neq f} R_{fl} \). Both of the target speech and noise are assumed to propagate in the full-rank propagation process. The observed signal \( x_{ft} \) also follows a Gaussian distribution: \( x_{ft} \sim \mathcal{N}(s_{ft}, Q_f) \).

Full-rank MWF is also obtained using maximum a posteriori estimation that maximizes \( p(s_{ft} | x_{ft}) \):

\[
\begin{align*}
    w^\text{HMWF}_f &= (P_f + Q_f)^{-1} P_f u, \quad (8) \\
    w^\text{HMWF}_f &= (P_f + Q_f)^{-1} P_f u, \quad (9)
\end{align*}
\]

where \( u \in \mathbb{C}^M \) is an \( M \)-dimensional unit vector whose elements are all 0 except the element corresponding to the reference channel. It is fixed to the first channel in this study. Full-rank MWF is directly constructed from the spatial covariance matrices which maintain information on the scale and reverberation [5].

### 3. Proposed Method

For effective beamforming, it is important to estimate the spatial covariance matrices of speech and noise accurately. Recently, DNN-based TF mask estimation has been widely used for spatial covariance matrix estimation [7–12]. This section describes the proposed unsupervised estimation method based on a multichannel source separation framework which preserves phase information of each source unlike the conventional mask-based methods.

#### 3.1. Spatial Covariance Matrix Estimation Based on MNMF

MNMF is a source separation method based on a factorization model [16]. The model is a multichannel extension of NMF, which decomposes a given nonnegative matrix \( X \) into two smaller nonnegative matrix pairs, \( B \) and \( C \). In speech signal processing, a set of frequency spectra is identified by the basis matrix \( B \) along with a set of temporal activation represented by the activation matrix \( C \).

It is necessary for source separation with an \( M \)-channel microphone array to consider the spatial propagation process. MNMF treats the observed signal as an Hermitian positive semi-definite matrix \( X_{fl} = x_{fl} x_{fl}^H \in \mathbb{C}^{M \times M} \). The diagonal components of the matrix represent values of power of the \( M \) channels, and the off-diagonal components represent correlations between the channels. MNMF introduces a matrix \( H_{fl} \in \mathbb{C}^{M \times M} \) that models the spatial property of the \( l \)-th sound source at the frequency bin \( f \). Let \( z_{ik} \in [0, 1] \) indicate whether the \( k \)-th NMF basis belongs to the \( l \)-th cluster \( (z_{ik} = 1) \) or not \( (z_{ik} = 0) \). MNMF employs a factorization model as follows:

\[
    \hat{X}_{fl} = \sum_{k=1}^{K} \left( \sum_{l=1}^{L} H_{fl} z_{ik} \right) b_{jk} c_{kl}, \quad (10)
\]

where \( b_{jk} \in \mathbb{R}_+ \) and \( c_{kl} \in \mathbb{R}_+ \) are the basis and activation, and they represent the low-rank structure of the sound source. MNMF factorizes a hierarchically structured matrix \( X \) into a product of \( (HZ)B \) and \( C \), where \( \circ \) represents the Hadamard product.

We estimate optimal \( H_{fl} \), \( z_{ik} \), \( b_{jk} \), and \( c_{kl} \) with the factorization model in Eq. (10). The MNMF algorithm to minimize the following IS divergence between the given matrix \( X_{fl} \) and its factorization model was derived as the form of multiplicative update formulas by Sawada et al. [16].

\[
    D_{is}(X; \{H, Z, B, C\}) = \sum_{l=1}^{F} \sum_{k=1}^{T} d_{is}(X_{fl}, \hat{X}_{fl}). \quad (11)
\]

To calculate the beamforming filters, we need to compute the spatial covariance matrices of speech and noise, \( P \) and \( Q \). We assume that the sound source \( l = 1 \) given the special initial value described in the following Section 3.2 is the target speech. Hence, we can define the matrices as follows:

\[
\begin{align*}
    P_f &= \frac{1}{T} \sum_{t=1}^{T} H_{ft} z_{1k} b_{fk} c_{kl}, \quad (12) \\
    Q_f &= \frac{1}{T} \sum_{l=1}^{T} P_{fl}, \quad (13) \\
    P &= \frac{1}{T} \sum_{t=1}^{T} P_{ft}, \quad (14) \\
    Q &= \frac{1}{T} \sum_{l=1}^{T} Q_{ft}. \quad (15)
\end{align*}
\]

#### 3.2. Initialization of MNMF

The performance of MNMF heavily depends on the initial value of the matrix \( H_{ft} \) [21]. To initialize \( H_{ft} \) effectively, we used rank-1 MNMF [18] and the cross-spectrum method [23]. Rank-1 MNMF, which has the same structure as MNMF, approximates \( H_{ft} \) as a rank-1 matrix, and is less sensitive to initial values. This matrix is represented by an outer product of a steering vector of the source \( l \), \( r_{fl} \in \mathbb{C}^{M} \): \( H_{ft} = r_{fl} r_{fl}^H \). Rank-1 MNMF estimates a steering vector of each sound source in an unsupervised manner [18]. We initialize \( H_{ft} \) using these estimated steering vectors. For further stability, the steering vector corresponding to the target speech \( r_{f1} \) is initialized with the cross-spectrum method [23]. Other parameters are initialized with random values.

#### 3.3. Steering Vector Estimation

The steering vector \( p_j \) is approximated as the principal component of the spatial covariance matrix of the target speech \( P_f \):

\[
    p_j = \mathcal{P} \mathcal{E} \{ P_f \}, \quad (16)
\]

where \( \mathcal{P} \mathcal{E} \{ \cdot \} \) represents the principal eigenvector of a matrix.

The rank-1 MWF beamforming filter requires the variance of the target speech \( \phi_j \) in Eqs. (5) and (6). By approximating the spatial covariance matrix \( P_f \) as \( \phi_j p_j p_j^H \), we obtain the \( \phi_j \) as follows:

\[
    \phi_j \simeq \frac{\| P_f \|}{\| p_j p_j^H \|}, \quad (17)
\]

where \( \| \cdot \| \) represents the matrix norm of a matrix.

### 4. Experimental Evaluation

We evaluated the proposed methods through speech recognition experiments in real noisy environments. Two distinct ASR tasks were used. One is from the third CHiME Challenge [2] where plenty of training data are available, and the other is a task using our in-house data which are new to models trained with the CHiME data set.

#### 4.1. Experimental Settings

In the third CHiME Challenge [2], the noisy training set consists of 1,600 real noisy utterances and 7,138 simulated noisy utterances generated by artificially mixing the clean WSJ0 training set with
Table 3. MNMF experimental conditions.

<table>
<thead>
<tr>
<th>Sample frequency</th>
<th>16 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame length</td>
<td>64 ms</td>
</tr>
<tr>
<td>Frame overlap</td>
<td>10 ms</td>
</tr>
<tr>
<td>Number of microphones</td>
<td>5</td>
</tr>
<tr>
<td>Number of expected sources</td>
<td>5</td>
</tr>
<tr>
<td>Number of bases</td>
<td>25</td>
</tr>
<tr>
<td>Number of updates</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 4. Speech recognition performances (WER) on the CHiME-3 ASR task and the noisy JNAS ASR task.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Time</th>
<th>Eq.</th>
<th>CHiME-3</th>
<th>noisy JNAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Enhance</td>
<td>-</td>
<td>-</td>
<td>22.39</td>
<td>41.34</td>
</tr>
<tr>
<td>Beamformit</td>
<td>Invariant</td>
<td>-</td>
<td>15.60</td>
<td>35.28</td>
</tr>
<tr>
<td>DNNm-MVDR</td>
<td>Invariant</td>
<td>-</td>
<td>11.51</td>
<td>16.59</td>
</tr>
<tr>
<td>DNNm-GEV</td>
<td>Invariant</td>
<td>-</td>
<td>11.02</td>
<td>12.40</td>
</tr>
<tr>
<td>MNMF-MVDR</td>
<td>Invariant (3)</td>
<td>12.63</td>
<td>11.58</td>
<td></td>
</tr>
<tr>
<td>MNMF-r1MWF</td>
<td>Varying (4)</td>
<td>12.61</td>
<td>11.68</td>
<td></td>
</tr>
<tr>
<td>MNMF-r2MWF</td>
<td>Invariant (5)</td>
<td>12.61</td>
<td>11.79</td>
<td></td>
</tr>
<tr>
<td>MNMF-r2MWF</td>
<td>Varying (6)</td>
<td>12.46</td>
<td>10.85</td>
<td></td>
</tr>
</tbody>
</table>

4.2. Experimental results

The speech recognition performances in WER are listed in Table 4. In the CHiME-3 ASR task, the best-proposed method, time-varying rank-1 MWF with MNMF-based estimation (MNMF-r1MWF), achieved a 3.14 points lower WER than Beamformit. Although our methods did not achieve comparable WERs to DNN-based beamforming, they were trained using the matched data to the test environment, and therefore achieved high performance with prior learning.

The noisy JNAS ASR task was conducted assuming there are no data for retraining. Time-varying MNMF-r1MWF achieved a WER of 10.85%, which is 12.5% relative improvement from state-of-the-art DNNm-GEV. The noisy JNAS task was much different from the CHiME-3 task in terms of microphone setups and noise environments. Compared with our methods, the DNN-based beamforming performance deteriorated significantly in the unknown recording condition. In contrast, the proposed methods using MNMF maintained the high performance on both tasks.

Rank-1 MWF assumes that the propagation process of the target speech considers mainly a direct wave. The assumption is adequate in open spaces or large rooms, and can be a reason why rank-1 MWF showed the best performance in our data set recorded in a large cafeteria. Compared with MVDR, rank-1 MWF incorporates the scale of the target speech, which may have resulted in improved ASR performance. Use of a time-varying noise spatial covariance matrix yielded a further improvement since the spatial covariance matrix was estimated for every time frame with the proposed MNMF-based estimation method. It suggests that tracking non-stationary noise is important in beamforming.

5. CONCLUSION

We presented unsupervised speech enhancement methods based on integration of beamforming and MNMF. The proposed methods use MNMF to estimate the spatial covariance matrices of speech and noise with preserving their phase information without using any supervised training and then generate an enhanced speech signal with beamforming. The experimental results in real-recording ASR tasks demonstrated that the proposed methods are more robust in an unknown environment than the state-of-the-art beamforming method with DNN-based mask estimation. We plan to develop an online version of the proposed method.
6. REFERENCES


