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.

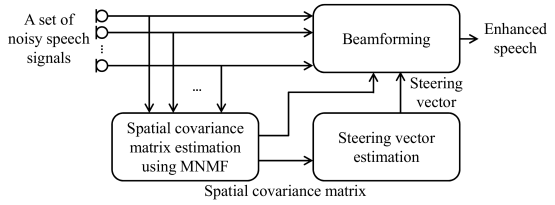

Fig. 1. The proposed approach to unsupervised speech enhancement based on a variant of beamforming that calculates the spatial covariance matrices of speech and noise from the corresponding spectrograms obtained by MNMF.

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.

Estimation method \ Propagation process	Speech |rank-1 Noise |full-rank	Speech |full-rank Noise |full-rank
Maximum Likelihood Estimation	MVDR in Eqs. (3) & (4)	- -
Maximum A Posteriori Estimation	Rank-1 MWF in Eqs. (5) & (6)	Full-rank MWF in Eqs. (8) & (9)

spectrogram of each source as the product of a basis matrix (a set of basis spectra) and an activation matrix (a set of temporal activations) as in nonnegative matrix factorization (NMF). Since each TF bin is decomposed into a sum of all sources (e.g., speech and noise) with phase information, more accurate spatial covariance matrices can be obtained than DNN- or GMM-based mask estimation methods [7–12] that directly calculate the spatial covariance matrix of speech from noisy TF bins without any decomposition. The proposed unsupervised speech enhancement is expected to work even in environments where there are no matched training data. Since there are only a few studies of speech enhancement based on BSS [19–21], we further investigate integration of MNMF [16] and various types of beamforming in a wide variety of conditions. More specifically, we test time-invariant and time-varying versions of MVDR beamforming [1], rank-1 MWF beamforming, and full-rank MWF [5].

2. BEAMFORMING

We use three beamforming methods, namely, MVDR, rank-1 MWF, and full-rank MWF. We review these beamforming methods from two perspectives: the propagation process of each source and the estimation method of the filter. First, we define two kinds of propagation processes from each sound source to a microphone array. The first one, which we call the *rank-1 propagation process*, is modeled by using a single steering vector for each source. It considers mainly a direct wave. The second one, which we call the *full-rank propagation process*, considers a more complicated propagation process. It is modeled with a full-rank spatial covariance matrix. Second, we consider two estimation methods to obtain a beamforming filter. One estimation method is the maximum a posteriori estimation method, in which a target speech signal is assumed to follow a Gaussian distribution. The other method is the maximum likelihood estimation method, in which the assumption on a target speech signal is not considered. Table 1 summarizes the relationship among these beamforming methods.

These beamforming methods are performed in the short-time Fourier transform (STFT) domain. Let $\boldsymbol{x}_{ft} \in \mathbb{C}^M$ be an observed signal picked up with an M-channel microphone array in frequency bin f and time frame t. These methods apply a linear filter $\boldsymbol{w}_{ft} \in \mathbb{C}^M$ to the multichannel observed signal in order to produce an enhanced target signal $y_{ft} \in \mathbb{C}$:

$$y_{ft} = \boldsymbol{w}_{ft}^H \boldsymbol{x}_{ft}. \tag{1}$$

Before explaining each beamforming method, we declare the notation on steering vectors and the spatial covariance matrices as shown in Table 2.

In MVDR beamforming [1], a multichannel observed signal $\boldsymbol{x}_{ft}$ is assumed as follows:

$$\boldsymbol{x}_{ft} = \boldsymbol{p}_f s_{ft} + \boldsymbol{n}_{ft}, \tag{2}$$

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

Signal	Target speech	Noise	Source l
Steering vector	$\boldsymbol{p}$	-	$\boldsymbol{r}_l$
Spatial covariance matrix	$\boldsymbol{P}$	$\boldsymbol{Q}$	$\boldsymbol{R}_l$

where $s_{ft} \in \mathbb{C}$ is a single target speech signal, $\boldsymbol{p}_f \in \mathbb{C}^M$ is a steering vector of the target speech, and $\boldsymbol{n}_{ft} \in \mathbb{C}^M$ is noise signal. $\boldsymbol{s}_{ft} = \boldsymbol{p}_f s_{ft} \in \mathbb{C}^M$ 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 $\boldsymbol{0} \in \mathbb{C}^M$ and a full-rank spatial covariance matrix of the noise $\boldsymbol{Q}_f \in \mathbb{C}^{M \times M}$, and the noise is assumed to propagate in the full-rank propagation process. The observed signal also follows a Gaussian distribution: $\boldsymbol{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{p}_f s_{ft}, \boldsymbol{Q}_f)$. 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:

$$\boldsymbol{w}_f^{\mathrm{MVDR}} = \frac{\boldsymbol{Q}_f^{-1}\boldsymbol{p}_f}{\boldsymbol{p}_f^H \boldsymbol{Q}_f^{-1}\boldsymbol{p}_f}, \tag{3}$$

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

$$\boldsymbol{w}_{ft}^{\mathrm{MVDR}} = \frac{\boldsymbol{Q}_{ft}^{-1}\boldsymbol{p}_f}{\boldsymbol{p}_f^H \boldsymbol{Q}_{ft}^{-1}\boldsymbol{p}_f}. \tag{4}$$

With the same assumption on the propagation as in the MVDR beamforming, rank-1 MWF beamforming assumes the target speech follows a Gaussian distribution: $s_{ft} \sim \mathcal{N}_{\mathbb{C}}(0, \phi_f)$. 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 $p(s_{ft}|\boldsymbol{x}_{ft})$:

$$\boldsymbol{w}_f^{\mathrm{r1MWF}} = \frac{\boldsymbol{Q}_f^{-1}\boldsymbol{p}_f}{\boldsymbol{p}_f^H \boldsymbol{Q}_f^{-1}\boldsymbol{p}_f + \phi_f^{-1}}, \tag{5}$$

$$\boldsymbol{w}_{ft}^{\mathrm{r1MWF}} = \frac{\boldsymbol{Q}_{ft}^{-1}\boldsymbol{p}_f}{\boldsymbol{p}_f^H \boldsymbol{Q}_{ft}^{-1}\boldsymbol{p}_f + \phi_f^{-1}}. \tag{6}$$

When the assumption is not used (or the variance of the target source $\phi_f \to \infty$ 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:

$$\boldsymbol{x}_{ft} = \sum_{l=1}^{L} \boldsymbol{x}_{ftl}, \tag{7}$$

where $\boldsymbol{x}_{ftl} \in \mathbb{C}^M$ is an source image from a source l picked up with an M-channel microphone array. The l-th source image $\boldsymbol{x}_{ftl}$ follows a Gaussian distribution with a mean of $\boldsymbol{0}$ and a full-rank spatial covariance matrix $\boldsymbol{R}_{fl} \in \mathbb{C}^{M \times M}$. We regard $l = 1$ is the index for the target speech and the rest are noise. The target speech image $\boldsymbol{s}_{ft} = \boldsymbol{x}_{ft1}$ follows a Gaussian distribution with a full-rank spatial covariance matrix of the target speech $\boldsymbol{P}_f = \boldsymbol{R}_{f1} \in \mathbb{C}^{M \times M}$:

$\boldsymbol{s}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{0}, \boldsymbol{P}_f)$. The noise signal $\boldsymbol{n}_{ft} = \sum_{l \neq 1} \boldsymbol{x}_{ftl}$ follows a Gaussian distribution with the full-rank spatial covariance matrix $\boldsymbol{Q}_f = \sum_{l \neq 1} \boldsymbol{R}_{fl}$. Both of the target speech and noise are assumed to propagate in the full-rank propagation process. The observed signal $\boldsymbol{x}_{ft}$ also follows a Gaussian distribution: $\boldsymbol{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_{ft}, \boldsymbol{Q}_f)$. Full-rank MWF is also obtained using maximum a posteriori estimation that maximizes $p(\boldsymbol{s}_{ft}|\boldsymbol{x}_{ft})$:

$$\boldsymbol{w}_f^{\mathrm{frMWF}} = (\boldsymbol{P}_f + \boldsymbol{Q}_f)^{-1} \boldsymbol{P}_f \boldsymbol{u}, \tag{8}$$

$$\boldsymbol{w}_{ft}^{\mathrm{frMWF}} = (\boldsymbol{P}_{ft} + \boldsymbol{Q}_{ft})^{-1} \boldsymbol{P}_{ft} \boldsymbol{u}, \tag{9}$$

where $\boldsymbol{u} \in \mathbb{C}^M$ 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 $\boldsymbol{X}$ into two smaller nonnegative matrix pairs, $\boldsymbol{B}$ and $\boldsymbol{C}$. In speech signal processing, a set of frequency spectra is identified by the basis matrix $\boldsymbol{B}$ along with a set of temporal activation represented by the activation matrix $\boldsymbol{C}$.

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 $\boldsymbol{X}_{ft} = \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^H \in \mathbb{C}^{M \times M}$. 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 $\boldsymbol{H}_{fl} \in \mathbb{C}^{M \times M}$ that models the spatial property of the l-th sound source at the frequency bin f. Let $z_{lk} \in [0, 1]$ indicate whether the k-th NMF basis belongs to the l-th cluster ($z_{lk} = 1$) or not ($z_{lk} = 0$). MNMF employs a factorization model as follows:

$$\hat{\boldsymbol{X}}_{ft} = \sum_{k=1}^{K} \left(\sum_{l=1}^{L} \boldsymbol{H}_{fl} z_{lk} \right) b_{fk} c_{kt}, \tag{10}$$

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

We estimate optimal $\boldsymbol{H}_{fl}$, z_{lk}, b_{fk} and c_{kt} with the factorization model in Eq. (10). The MNMF algorithm to minimize the following IS divergence between the given matrix $\boldsymbol{X}_{ft}$ and its factorization model was derived as the form of multiplicative update formulas by Sawada et al. [16].

$$D_{IS}(\boldsymbol{X}, \{\boldsymbol{H}, \boldsymbol{Z}, \boldsymbol{B}, \boldsymbol{C}\}) = \sum_{f=1}^{F} \sum_{t=1}^{T} d_{IS}(\boldsymbol{X}_{ft}, \hat{\boldsymbol{X}}_{ft}). \tag{11}$$

To calculate the beamforming filters, we need to compute the spatial covariance matrices of speech and noise, $\boldsymbol{P}$ and $\boldsymbol{Q}$. 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:

$$\boldsymbol{P}_{ft} = \sum_{k=1}^{K} \boldsymbol{H}_{f1} z_{1k} b_{fk} c_{kt}, \tag{12}$$

$$\boldsymbol{Q}_{ft} = \sum_{k=1}^{K} \left(\sum_{l=2}^{L} \boldsymbol{H}_{fl} z_{lk} \right) b_{fk} c_{kt}, \tag{13}$$

$$\boldsymbol{P}_f = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{P}_{ft}, \tag{14}$$

$$\boldsymbol{Q}_f = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{Q}_{ft}. \tag{15}$$

3.2. Initialization of MNMF

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

3.3. Steering vector estimation

The steering vector $\boldsymbol{p}_f$ is approximated as the principal component of the spatial covariance matrix of the target speech $\boldsymbol{P}_f$:

$$\boldsymbol{p}_f = \mathcal{PE}\{\boldsymbol{P}_f\}, \tag{16}$$

where $\mathcal{PE}\{\cdot\}$ represents the principal eigenvector of a matrix.

The rank-1 MWF beamforming filter requires the variance of the target speech ϕ_f in Eqs. (5) and (6). By approximating the spatial covariance matrix $\boldsymbol{P}_f$ as $\phi_f \boldsymbol{p}_f \boldsymbol{p}_f^H$, we obtain the ϕ_f as follows:

$$\phi_f \simeq \frac{\|\boldsymbol{P}_f\|}{\|\boldsymbol{p}_f \boldsymbol{p}_f^H\|}, \tag{17}$$

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.

Sample frequency	16 kHz
Frame length	64 ms
Frame overlap	10 ms
Number of microphones	5
Number of expected sources	5
Number of bases	25
Number of updates	200

background noise. Each utterance consists of six channels, from which we used five by eliminating channel 2 facing the opposite direction. There are four different types of noise environments, namely, bus, street, cafe, and pedestrian area. We evaluated the ASR performance with word error rate (WER) using the real noisy evaluation set consisting of 1320 utterances ("et05_real_noisy"). We trained a DNN-HMM acoustic model [24, 25] using the training set described above. It had four hidden layers with 2k rectified linear units (ReLUs) [26] and a softmax output layer with 2k nodes. Its input is a 1,320-dimensional feature vector consisting of 11 frames of 40-channel log Mel-scale filterbank (lmfb) outputs and their delta and acceleration coefficients. Mean and variance normalization was applied to input vectors. Dropout [27] and batch normalization [28] were used in the training of all hidden layers. The language model was the standard WSJ 5k trigram LM. The Kaldi WFST decoder [29] was used for decoding.

The other ASR task is an in-house test set ("noisy_JNAS"), which consists of 200 sentences from the Japanese newspaper article sentence (JNAS) [30] corpus spoken by five male speakers in a crowded cafeteria. The utterances were recorded with a five-channel microphone array. For realistic scenario for distant ASR systems, we constructed a hemispherical array with micro-electro-mechanical systems (MEMS) microphone elements, which are increasingly being used in commercial products. The distance between the speakers and the array was set to be around 1m. The DNN-HMM acoustic model was also trained using multi-condition data, in which the noise data of the CHiME-3 were added to the original clean speech data of the JNAS. It had six hidden layers with 2048 sigmoidal nodes and an output layer with 3k nodes. A trigram language model was also trained using the JNAS. We used the Julius decoder [31]. The noisy JNAS test set has a lot of different characteristics from the CHiME-3 test set, as it was recorded in a new noisy environment, and the microphone type and geometry were also different.

We performed MNMF with the configurations shown in Table 3. Each beamforming method constructs its filter from the same spatial covariance matrices estimated by MNMF to ensure that randomness on the initial values does not affect experimental results. We used Beamformit [14] as a baseline for comparison. We also trained a feed-forward DNN for mask estimation using the IBM as the target. We also tried to use bidirectional long short-term memory (BLSTM) [9], but the feed-forward DNN slightly outperformed BLSTM in our preliminary experiments, thus we show the results obtained with the feed-forward DNN. The DNN structure is the same as the acoustic model for the CHiME-3 ASR task, except that the input feature is a 1,110-dimensional feature vector consisting of 11 frames of static 100-dimensional lmfb outputs, and the output is a 201-dimensional frequency mask. The DNN was trained using the CHiME-3 data set to generate TF masks for MVDR and GEV beamforming (DNNm-MVDR and GEV) and used in both CHiME-3 and noisy JNAS evaluation sets.

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

Methods	Time	Eq.	CHiME-3	noisy JNAS
Not Enhance	-	-	22.39	41.34
Beamformit	Invariant	-	15.60	35.28
DNNm-MVDR	Invariant	-	11.51	16.59
DNNm-GEV	Invariant	-	**11.02**	12.40
MNMF-MVDR	Invariant	(3)	12.63	11.58
	Varying	(4)	12.61	11.68
MNMF-r1MWF	Invariant	(5)	12.61	11.79
	Varying	(6)	12.46	**10.85**
MNMF-frMWF	Invariant	(8)	12.89	11.24
	Varying	(9)	12.70	11.24

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 methods which were trained using the matched data to the test environment, the proposed methods showed consistently high performance without prior learning.

The noisy JNAS ASR task was conducted assuming there are no data for retraining. Time-varying MNMF-r1MWF achieved the 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 beamforming was the most effective in combination with the proposed MNMF-based estimation method. 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 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

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[2] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. of IEEE ASRU*, 2015, pp. 504–511.

[3] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. of IEEE ASRU*, 2015, pp. 436–443.

[4] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. ASLP*, vol. 12, no. 6, pp. 561–571, 2004.

[5] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. ASLP*, vol. 18, no. 2, pp. 260–276, 2010.

[6] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. ASLP*, vol. 15, no. 5, pp. 1529–1539, 2007.

[7] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. of IEEE ICASSP*, 2016, pp. 5210–5214.

[8] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, "Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming," in *Proc. of IEEE ICASSP*, 2017, pp. 286–290.

[9] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. of IEEE ICASSP*, 2016, pp. 196–200.

[10] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks." in *Proc. of Interspeech*, 2016, pp. 1981–1985.

[11] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *Proc. of IEEE ICASSP*, 2017, pp. 3246–3250.

[12] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proc. of ICML*, vol. 70, 2017, pp. 2632–2641.

[13] B. Loesch and B. Yang, "Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions." in *Proc. of LVA/ICA*, 2010, pp. 41–48.

[14] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. ASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[15] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[16] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. ASLP*, vol. 21, no. 5, pp. 971–982, 2013.

[17] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, 2010.

[18] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[19] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. ASLP*, vol. 18, no. 7, pp. 1830–1840, 2010.

[20] M. Mimura, Y. Bando, K. Shimada, S. Sakai, K. Yoshii, and T. Kawahara, "Combined multi-channel NMF-based robust beamforming for noisy speech recognition," in *Proc. of Interspeech*, 2017, pp. 2451–2455.

[21] Y. Tachioka, T. Narita, I. Miura, T. Uramoto, N. Monta, S. Uenohara, K. Furuya, S. Watanabe, and J. Le Roux, "Coupled initialization of multi-channel non-negative matrix factorization based on spatial and spectral information," in *Proc. of Interspeech*, 2017, pp. 2461–2465.

[22] V. A. Barroso and J. M. Moura, "Maximum likelihood beamforming in the presence of outliers," in *Proc. of IEEE ICASSP*, 1991, pp. 1409–1412.

[23] G. C. Carter, "Coherence and time delay estimation," *Proc. of IEEE*, vol. 75, no. 2, pp. 236–255, 1987.

[24] A. R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 14–22, 2012.

[25] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. ASLP*, vol. 20, no. 1, pp. 30–42, 2012.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. of ICML*, 2010, pp. 807–814.

[27] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of ICML*, 2015, pp. 448–456.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[30] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *JASJ(E)*, vol. 20, no. 3, pp. 199–206, 1999.

[31] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," in *Proc. of Eurospeech*, 2001, pp. 1691–1694.