INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Combined Multi-channel NMF-based Robust Beamforming for Noisy Speech Recognition

Masato Mimura[1], *Yoshiaki Bando*[1], *Kazuki Shimada*[1], *Shinsuke Sakai*[1], *Kazuyoshi Yoshii*[1 2], *Tatsuya Kawahara*[1]

[1]Kyoto University, School of Informatics, Sakyo-ku, Kyoto606-8501, Japan
[2]RIKEN Center for Advanced Intelligence Project (AIP), Chuo-ku, Tokyo103-0027, Japan
{mimura, yoshiaki, shimada, sakai, yoshii, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

We propose a novel acoustic beamforming method using blind source separation (BSS) techniques based on non-negative matrix factorization (NMF). In conventional mask-based approaches, hard or soft masks are estimated and beamforming is performed using speech and noise spatial covariance matrices calculated from masked noisy observations, but the phase information of the target speech is not adequately preserved. In the proposed method, we perform complex-domain source separation based on multi-channel NMF with rank-1 spatial model (rank-1 MNMF) to obtain a speech spatial covariance matrix for estimating a steering vector for the target speech utilizing the separated speech observation in each time-frequency bin. This accurate steering vector estimation is effectively combined with our novel noise mask prediction method using multi-channel robust NMF (MRNMF) to construct a Maximum Likelihood (ML) beamformer that achieved a better speech recognition performance than a state-of-the-art DNN-based beamformer with no environment-specific training. Superiority of the phase preserving source separation to real-valued masks in beamforming is also confirmed through ASR experiments.

Index Terms: beamforming, blind source separation, multi-channel non-negative matrix factorization, noisy speech recognition

1. Introduction

Speech reverberation and additive noise adversely influence the speech recognition accuracy when the user is distant from the microphone, and there are increasing demands and expectations for the robust distant automatic speech recognition (ASR) systems. Various research efforts have been made in line with this understanding and the reported results for recent open evaluations such as the Reverb Challenge [1] and the CHiME challenge [2] clearly show that multi-channel signal processing is particularly effective for an acceptable distant ASR performance in very adverse noisy conditions.

Acoustic beamforming [3][4] is a promising approach for this multi-channel speech enhancement front-end for ASR. In earlier attempts at beamforming for ASR, it was typically performed based on a roughly determined voice activity detection (VAD) results and the time differences of arrival (TDOA) estimated using geometry information [5][6], but they did not achieve ASR performances good enough for noisy speech in real situations. Recently, a number of works adopting another approach based on time-frequency masks have been reported showing impressive enhancement performances [7][8][9][10][11][12].

In this paper, we propose a natural extension to this mask-based beamforming. Our contribution is twofold. First, we reformulate the conventional mask-based beamforming framework so that the steering vector for the target speech is calculated using a time-frequency representation of speech preserving phase extracted via a complex-domain source separation technique. For this purpose, we adopt a multi-channel NMF employing rank-1 spatial model (rank-1 MNMF) [13] that separates out the target signal with little distortion, which is an advantage for ASR. Its rank-1 constraint also matches the characteristics of beamforming which is basically performed using a single steering vector. Secondly, we show that our novel multi-channel source separation technique called multi-channel robust NMF (MRNMF) [14] can be applied to robust estimation of noise spatial covariance matrices, as well as to the initialization for rank-1 MNMF estimation that yields robust separation results for speech source. These techniques resulted in a powerful beamformer which achieved a better ASR performance than a state-of-the-art DNN-based beamformer with no environment-specific training.

2. Acoustic beamforming

We perform speech enhancement in the short-time Fourier transform (STFT) domain. The data model considered in this paper is as follows:

$$\begin{aligned} y_{ft,m} &= g_{f,m}s_{ft} + v_{ft,m} \\ &= x_{ft,m} + v_{ft,m}, \end{aligned} \quad (1)$$

where $y_{ft,m}$, $x_{ft,m}$ and $v_{ft,m} \in \mathbb{C}$ are the noisy speech, speech and noise observations at the m-th microphone. t ($1 \leq t \leq T$) and f ($1 \leq f \leq F$) are the indices for time frame and frequency bin. s_{ft} is the single source signal and $g_{f,m} \in \mathbb{C}$ is the finite inpulse response of the recording environment, and the vector defined as $[g_{f,1}, g_{f,2}, ..., g_{f,M}]^{\mathrm{T}} \in \mathbb{C}^M$ is called the steering vector.

The speech enhancement via beamforming is performed by applying a time-invariant linear filter $\boldsymbol{h}_f \in \mathbb{C}^M$ to the noisy speech observations as:

$$\begin{aligned} z_{ft} &= \boldsymbol{h}_f^{\mathrm{H}} \boldsymbol{y}_{ft} \\ &= \boldsymbol{h}_f^{\mathrm{H}} \boldsymbol{x}_{ft} + \boldsymbol{h}_f^{\mathrm{H}} \boldsymbol{v}_{ft}, \end{aligned} \quad (2)$$

where $\boldsymbol{y}_{ft}$ is defined as $\boldsymbol{y}_{ft} = [y_{ft,1}, y_{ft,2}, ..., y_{ft,M}]^{\mathrm{T}} \in \mathbb{C}^M$. $\boldsymbol{x}_{ft}$ and $\boldsymbol{v}_{ft}$ are also defined in the same way. z_{ft} is the enhanced signal. In this paper, we adopt two specific types of beamformers, namely ML and MV beamformers [15] to evaluate our framework. The proposed method can also be applied to a variety of methods including the Generalized Eigenvector (GEV) beamformer [16], which is not presented in this paper due to space limitation.

http://dx.doi.org/10.21437/Interspeech.2017-642

2.1. Maximum likelihood (ML) beamformer

The ML beamformer[1] [15] is obtained by maximizing the following likelihood function:

$$p(\boldsymbol{y}_{ft}|s_{ft}) = \frac{1}{\det(\pi \boldsymbol{K}_f)} \exp(-(\boldsymbol{y}_{ft} - \boldsymbol{g}_f s_{ft})^{\mathrm{H}} \boldsymbol{K}_f^{-1}(\boldsymbol{y}_{ft} - \boldsymbol{g}_f s_{ft})), \quad (3)$$

on the assumption that the noise conforms to Gaussian distribution where $\boldsymbol{K}_f$ is the spatial covariance matrix of noise calculated as $\boldsymbol{K}_f = \frac{1}{T}\sum_t \boldsymbol{v}_{ft}\boldsymbol{v}_{ft}^{\mathrm{H}}$. The value of s_{ft} that maximizes the likelihood function is:

$$\hat{s}_{ft} = \frac{\boldsymbol{g}_f^{\mathrm{H}} \boldsymbol{K}_f^{-1} \boldsymbol{y}_{ft}}{\boldsymbol{g}_f^{\mathrm{H}} \boldsymbol{K}_f^{-1} \boldsymbol{g}_f} \quad (4)$$

and this leads to the following filter coefficients:

$$\boldsymbol{h}_f^{(ML)} = \frac{\boldsymbol{K}_f^{-1}\boldsymbol{g}_f}{\boldsymbol{g}_f^{\mathrm{H}}\boldsymbol{K}_f^{-1}\boldsymbol{g}_f}. \quad (5)$$

For performing the ML beamforming, both of $\boldsymbol{K}_f$ and $\boldsymbol{g}_f$ need to be estimated from the noisy observations in some way.

2.2. Minimum variance (MV) beamformer

Another version of beamforming method called MV beamformer [15] can be performed even when the noise covariance matrix $\boldsymbol{K}_f$ is not available. It minimizes the power (variance) of the filtered noisy speech under the constraint that the target speech signal remains distortionless as:

$$\boldsymbol{h}_f^{(MV)} = \arg\min_{\boldsymbol{h}_f} \boldsymbol{h}_f \boldsymbol{R}_f \boldsymbol{h}_f^{\mathrm{H}} \quad \text{s.t. } \boldsymbol{h}_f^{\mathrm{H}}\boldsymbol{g}_f = 1, \quad (6)$$

where $\boldsymbol{R}_f$ is the spatial covariance matrix of noisy speech calculated as $\boldsymbol{R}_f = \frac{1}{T}\sum_t \boldsymbol{y}_{ft}\boldsymbol{y}_{ft}^{\mathrm{H}}$. This leads to the following filter coefficients:

$$\boldsymbol{h}_f^{(MV)} = \frac{\boldsymbol{R}_f^{-1}\boldsymbol{g}_f}{\boldsymbol{g}_f^{\mathrm{H}}\boldsymbol{R}_f^{-1}\boldsymbol{g}_f}. \quad (7)$$

2.3. Mask-based beamforming

When a time-frequency mask is available [9][11][12], the noise covariance matrix $\boldsymbol{K}_f$ can be estimated by accumulating time-frequency bins clustered to be noise as:

$$\hat{\boldsymbol{K}}_f = \frac{1}{\sum_t M_{ft}^{(noise)}} \sum_t M_{ft}^{(noise)} \boldsymbol{y}_{ft}\boldsymbol{y}_{ft}^{\mathrm{H}}, \quad (8)$$

where the real-valued noise mask $M_{ft}^{(noise)}$ shared among all channels represents the probability that the time-frequency bin is dominated by noise. We can also find a steering vector $\hat{\boldsymbol{g}}_f$ by performing eigenvalue decomposition to the speech covariance matrix $\boldsymbol{J}_f = \frac{1}{T}\sum_t \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^{\mathrm{H}}$ and picking up the eigenvector with the largest eigenvalue. $\boldsymbol{J}_f$ is also estimated using speech mask $M_{ft}^{(speech)}$:

$$\hat{\boldsymbol{J}}_f = \frac{1}{\sum_t M_{ft}^{(speech)}} \sum_t M_{ft}^{(speech)} \boldsymbol{y}_{ft}\boldsymbol{y}_{ft}^{\mathrm{H}}. \quad (9)$$

We can construct an ML beamformer with both of $\boldsymbol{K}_f$ and $\boldsymbol{g}_f$ using formula (5), and an MV beamformer with $\boldsymbol{g}_f$ using (7).

[1]ML beamformer is also referred to as MVDR in the literature, e.g., [8],[11], whereas MV beamformer in 2.2 has traditionally been referred to as MVDR ([3], [17]) and sometimes MPDR ([18]). We follow the definitions in [15].

3. Proposed method

3.1. Steering vector estimation using rank-1 MNMF

As described in the previous section, we need to estimate the steering vector for target speech for both of ML and MV beamformers and it critically influences the performance of speech recognition backend as will be demonstrated in the next section. In this paper, we estimate the steering vector for target speech source utilizing a source separation technique called rank-1 MNMF [13]. We modify (1) and assume that the observation consists of one speech source and $N-1$ noise sources as:

$$\boldsymbol{y}_{ft} = \boldsymbol{G}_f \boldsymbol{s}_{ft}, \quad (10)$$

where one element of $\boldsymbol{s}_{ft} = [s_{ft,1}, s_{ft,2}, ..., s_{ft,N}]^{\mathrm{T}} \in \mathbb{C}^N$ is the speech source, and the N columns in the $M \times N$ matrix $\boldsymbol{G}_f = [\boldsymbol{g}_{f,1}, ..., \boldsymbol{g}_{f,N}]$ are the steering vectors for speech and noise sources. In rank-1 MNMF, this observation is modeled to conform to a zero-mean complex Gaussian distribution as:

$$\boldsymbol{y}_{ft} \sim \mathcal{N}_c(\boldsymbol{y}|\boldsymbol{0}, \boldsymbol{R}_{ft}), \quad (11)$$

where $\boldsymbol{R}_{ft} = \sum_n \boldsymbol{g}_{f,n}\boldsymbol{g}_{f,n}^{\mathrm{H}} r_{ft,n}$ and $r_{ft,n} = \sum_k \lambda_{nk} b_{fk} c_{kt}$. $\boldsymbol{b}_k\ (= [b_{1k}, ..., b_{Fk}]^{\mathrm{T}})$, $k = 1, ..., K$, are called bases and constitute representative spectral patterns that are regarded as activated by c_{kt} that represents gain for basis k at time t. The latent variable λ_{nk} indicates whether the basis k belongs to the source n or not. We estimate the model parameters λ_{nk}, b_{fk}, c_{kt}, and $\boldsymbol{G}_f$ by minimizing the negative log-likelihood

$$L(\theta; \mathcal{D}) = \sum_t \sum_f -\log \mathcal{N}_c(\boldsymbol{y}_{ft}|\boldsymbol{0}, \boldsymbol{R}_{ft}), \quad (12)$$

where $\theta = \{\lambda_{nk}, b_{fk}, c_{kt}, \boldsymbol{W}_f (\stackrel{\mathrm{def}}{=} \boldsymbol{G}_f^{-1}) | n = 1, ..., N,\ f = 1, ..., F,\ t = 1, ..., T,\ k = 1, ..., K\}$ and $\mathcal{D} = \{\boldsymbol{y}_{ft} | f = 1, ..., F,\ t = 1, ..., T\}$. We set the number of sources, N, equal to M for convinience to obtain the model parameters within the rank-1 NMF framework [13].

The minimization of this likelihood function is performed by a repetition of auxiliary function-based update rules for spatial part of the model parameters (i.e. $\boldsymbol{W}_f$) and multiplicative update rules for source subset (i.e. $\lambda_{nk}, b_{fk}, c_{kt}$) of the model parameters. Readers are referred to [13] for the detail. Once we obtain $\hat{\boldsymbol{W}}_f$, the estimate of $\boldsymbol{W}_f$, we can recover the speech and noise signals from N sources as

$$\hat{s}_{ft} = \hat{\boldsymbol{W}}_f \boldsymbol{y}_{ft}. \quad (13)$$

After determining the source index n_s as speech source, e.g., by choosing the one with the largest energy, we can obtain the contribution to the observation from speech source as:

$$\hat{\boldsymbol{x}}_{ft} = \hat{\boldsymbol{g}}_{f,n_s} \hat{s}_{ft,n_s}, \quad (14)$$

where $\hat{\boldsymbol{g}}_{f,n_s}$ is the n_s-th column of $\hat{\boldsymbol{W}}_f^{-1}$, and $\hat{s}_{ft,n_s}$ is the n_s-th element of the N-vector $\boldsymbol{s}_{ft}$. The speech covariance matrix can be computed as

$$\hat{\boldsymbol{J}}_f = \frac{1}{T}\sum_t \hat{\boldsymbol{x}}_{ft}\hat{\boldsymbol{x}}_{ft}^{\mathrm{H}} \quad (15)$$

3.2. Binary mask estimation using MRNMF

Since background noise spectrum in different time frames are usually highly correlated with each other, it can be assumed to lie in a low-rank subspace. On the other hand, human voices

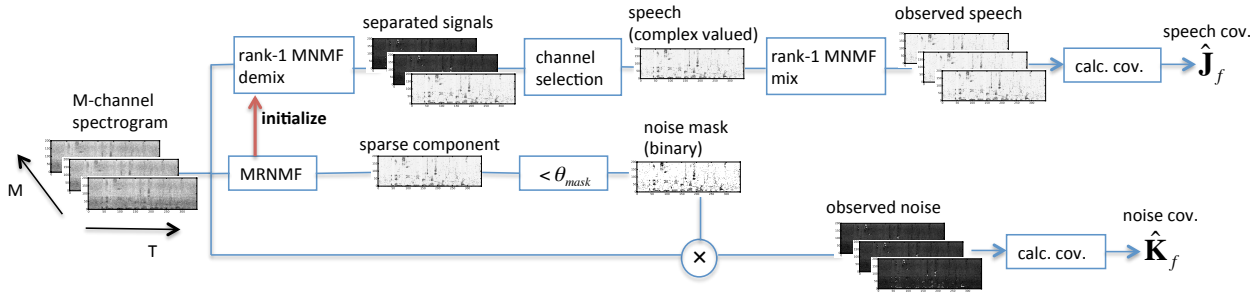

Figure 1: *Run-time algorithm for estimating noise and speech spatial covariance matrices based on combined BSS*

have more variation and are relatively sparse in the spectral domain. In multi-channel robust NMF (MRNMF) [14], the amplitude spectrogram of noisy speech observation is decomposed into channel-wise low-rank spectrograms $l_{ft,m} \in \mathbb{R}^+$ and a sparse spectrogram shared among all channels $s_{ft} \in \mathbb{R}^+$ as follows:[2]

$$|y_{ft,m}| \simeq l_{ft,m} + a_{t,m} s_{ft}, \tag{16}$$

where $a_{t,m} \in \mathbb{R}^+$ is the gain of the sparse component. This model is formulated as a unified Bayesian model and the estimation of noise and speech components is performed by a variational Bayesian (VB) inference. While the speech component s_{ft} is induced to be sparse by introducing gamma priors with the Jeffreys' hyperpriors, the noise component $l_{ft,m}$ is forced to be low-rank exploiting the Bayesian NMF decomposition framework [19] as:

$$l_{ft,m} \simeq \sum_k b_{fk,m} c_{kt,m}, \tag{17}$$

where $b_{fk,m}$ and $c_{kt,m}$ are the elements of the $F \times K$ NMF basis matrix $\boldsymbol{B}_m$ and $K \times T$ activation matrix $\boldsymbol{C}_m$. More details including an effective variational Bayesian inference algorithm are found in [14]. MRNMF has an attractive characteristics for robust speech recognition. It gives robust estimation for speech signal in very adverse conditions using only sparseness criterion without any prior knowledge on the testing conditions, even when the microphone array is partially occluded. Different from rank-1 MNMF or the complex GMM-based method proposed in [9], the speech signal is extracted as the sparse component without any external decision criteria in the MRNMF framework. Because the current version of MRNMF is performed in the amplitude domain, however, the output speech signal is not free from serious distortion and is not appropriate for input to back-end ASR. On the other hand, we can obtain a robust estimation of noise mask by thresholding the sparse component as:

$$M_{ft}^{(noise)} = \begin{cases} 1 & s_{ft} < \theta_{mask} \\ 0 & otherwise. \end{cases} \tag{18}$$

An example of the estimated noise mask is presented in Fig. 2. We define the threshold θ_{mask} as:

$$\theta_{mask} = 0.01 \cdot \frac{1}{FTM} \sum_{f,t,m} |y_{ft,m}|. \tag{19}$$

[2] $\mathbb{R}^+$ means the set of non-negative real numbers.

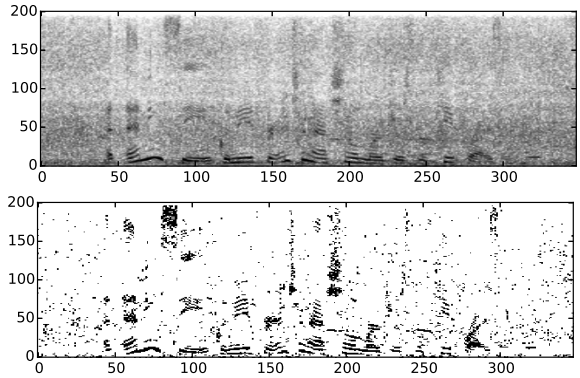

Figure 2: *An example of amplitude spectrum (top) and the estimated binary mask (bottom) with MRNMF for real noisy speech*

3.3. Combined BSS-based ML beamformer

Although rank-1 MNMF can estimate an accurate steering vector, it has three disadvantages. First, it cannot give a full-rank noise spatial covariance matirx required in ML beamforming. Secondly, we need some external criteria for choosing the speech source. Thirdly, because rank1-NMF is not perfectly free from permutation problem, speech signal can be separated into multiple sources, which can cause a serious performance degradation in subsequent beamforming and ASR.

These problems inherent to rank-1 MNMF can be solved by combining it with MRNMF. The first problem is solved by using the covariance matrix calculated with noise mask (17). Both of the second and third problems are solved by using the sparse component estimated using MRNMF for initializing rank-1 MNMF estimation. More precisely, a column in the mixing matrix $\boldsymbol{G}_f$ in rank-1 MNMF, which corresponds to the source we want to assign speech signal, is initialized with the steering vector estimated from the speech covariance matrix calculated using the MRNMF mask.

Based on the above discussion, we can construct a powerful ML beamformer by combining rank-1 MNMF and MRNMF without any training data. The run-time algorithm for estimating the speech and noise spatial covariance matrices in the proposed combined BSS-based framework is depicted in Fig. 1.

4. Experimental evaluation

We evaluated the proposed methods through the ASR task of 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 noise backgrounds. Each utterance consists of six channels from which we used five by eliminating channel 2 facing

Table 1: *Performance of proposed methods combined with multi-cond. acoustic model back-end (WER(%))*

					et05_real_noisy	ID
no enhancement					23.39	(1)
Beamformit					15.60	(2)
rank-1 MNMF (no beamforming)					15.35	(3)
	$\hat{\boldsymbol{J}}_f$	$\hat{\boldsymbol{K}}_f$	beamformer	needs training		
real valued mask-based	MRNMF	MRNMF	ML	no	12.99	(4)
	DNN	DNN	ML	yes	11.51	(5)
complex domain separation-based	rank-1 MNMF	-	MV	no	13.11	(6)
	rank-1 MNMF	MRNMF	ML	no	11.82	(7)
	MRNMF + rank-1 MNMF	MRNMF	ML	no	10.94	(8)

the opposite direction. There are four different types of noisy environments, namely, bus, street, cafe, and pedestrian area [2]. We trained a DNN-HMM acoustic model [20][21] using the training set described above. It has four hidden layers with 2k rectified linear units (ReLUs) [22] and a softmax output layer with 2k nodes. 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 is used as input. Dropout [23] and batch normalization [24] is used for training of all hidden layers. For decoding, we used the Kaldi WFST decoder [25]. The language model is the standard WSJ 5k trigram LM. We used the real noisy evaluation set ("et05_real_noisy") consisting of 1,320 utterances for evaluating the methods.

We used Beamformit [26] as a baseline bermformer against which we compared our method. We also trained a feed-forward DNN for mask prediction [11][12] using the ideal binary mask (IBM) [11] as target[3]. The DNN structure is the same as the acoustic model described above, except that the input feature is the 1,110-dimensional feature vector consisting of 11 frames of static 100-dimensional lmfb outputs and the output is $F(= 201)$-dimensional mask.

The experimental results are presented in Table 1. As shown in row (2), the baseline delay-and-sum beamformer (Beamformit) successfully reduced the WER. The WER obtained using rank-1 MNMF output directly without beamforming is shown in row (3). Speech channel selection is done by simply picking up the one with the highest power in this case. While it already gave a comparable average WER to Beamformit, we observed ASR errors due to permutation in some utterances.

In all beamforming experiments below, we used the same setting for acoustic signal processing: the sampling rate of audio signal is 16kHz, the window length and frame shift for short-time Fourier transform is 25ms (400 samples) and 10ms, which is the same as those for lmfb computation for ASR.

4.1. Mask-based beamforming using MRNMF

In the MRNMF-based system, both of the speech and noise covariance matrices are estimated using MRNMF-based masks. The speech mask is calculated as $M_{t,f}^{(speech)} = 1 - M_{t,f}^{(noise)}$. As shown in row (4) of Table 1, the MRNMF-based ML beamformer significantly outperformed Beamformit and rank-1 MNMF without beamforming, confirming that MRNMF-based robust mask estimation is effective for beamforming. However, it did not achieve a comparable performance to the ML beamformer constructed using a state-of-the-art DNN-based mask prediction (row (5)). Note that both of the MRNMF and DNN-based beamforming evaluated here are performed using real-valued masks.

[3]In [11] and [12], mask prediction was conducted using biderectional LSTMs, but the feed-forward DNNs using spliced lmfb features as input slightly outperformed biderectional LSTMs in our preliminary experiments and we show only the results obtained with DNNs here.

4.2. Complex domain source separation-based beamforming

We evaluated the proposed system combining complex domain source separation based on rank-1 spatial model (rank-1 NMF) and MRNMF.

First, from row (6), we can see that the accurate steering vector for the target speech estimated using rank-1 MNMF achieved a comparable performance to the MRNMF-based ML beamformer, even with MV beamformer that does not use the noise covariance. Interestingly, this result is much better than the WER obtained with the direct output of rank-1 MNMF (row (3)), suggesting that beamforming may not be drastically damaged by permutation errors.

When this steering vector by rank-1 MNMF is combined with the noise covariance matrix estimated using MRNMF (row (7)), the WER was significantly reduced from row (4), where the mask-based method was used for obtaining the target speech steering vector, suggesting the advantage of the phase-preserving source separation in steering vector estimation. Moreover, rank-1 MNMF estimated using the sparse component by MRNMF for initialization gave further significant improvement (row (8)), suggesting that the MRNMF output was a good initializer for rank-1 MNMF. This proposed system outperformed the DNN-based beamformer (row (5)) without any environment-specific training for mask generation.

5. Conclusion

We have proposed a novel acoustic beamforming method that utilizes rank-1 MNMF for accurately estimating a steering vector for the target speech and the multi-channel robust NMF for effectively estimating the noise mask as well as robustly initializing rank-1 MNMF to construct a powerful ML beamformer. We demonstrated the effectiveness of the proposed methods through noisy speech recognition experiments.

We are also interested to see how other techniques works as well, such as IVA [27] and full-rank MNMF [28] for source separation and Complex Gaussian Mixture Model [9] for mask prediction. Promising future work includes an extension of the current version of MRNMF to allow complex-valued inputs which would estimate an accurate full rank noise covariance and the steering vector directly without using real-valued masks in an unified framework. Since the proposed approach to beamforming is based on a robust source separation method, it will potentially be extended to distant ASR for multiple speakers.

6. References

[1] K.Kinoshita, M.Delcroix, T.Yoshioka, T.Nakatani, E.Habets, R.Haeb-Umbach, V.Leutnant, A.Sehr, W.Kellermann, R.Maas, S.Gannot, and B.Raj, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.

[2] J.Barker, R.Marxer, E.Vincent, and S.Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, 2015.

[3] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[4] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[5] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[6] X. Mestre and M. A. Lagunas, "On diagonal loading for minimum variance beamformers," in *Proc. ISSPIT*, 2003, pp. 459–462.

[7] D. H. T. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an Expectation Maximization framework," in *Proc. ICASSP*, 2010, pp. 241–244.

[8] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel MMSE-based framework for speech separation and noise reduction," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 21, no. 9, pp. 1913–1928, 2013.

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

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

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

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

[13] 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. Audio, Speech, & Language Proc.*, vol. 24, no. 9, pp. 1622–1637, 2016.

[14] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, and H. G. Okuno, "Variational Bayesian multi-channel robust NMF for human-voice enhancement with a deformable and partially-occluded microphone array," in *Proc. EUSIPCO*, 2016, pp. 1018–1022.

[15] D.H.Johnson and D.E.Dudgeon, *Array Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[16] E. Warsitz and R. Haeb-Umbach, "Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 15, no. 5, pp. 1529–1539, 2007.

[17] J.Benesty, J.Chen, Y.Huang, and J.Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 15, no. 3, pp. 1053–1065, 2007.

[18] H. L. V. Trees, *Detection, estimation, and modulation theory, part IV, Optimum Array Processing*. New York: Wiley, 2002.

[19] A. T. Cemgil, "Bayesian Inference for Nonnegative Matrix Factorisation Models," *Computational Intelligence and Neuroscience*, vol. 2009, no. 4, pp. 1–17, 2009.

[20] A.Mohamed, G.Dahl, and G.Hinton, "Acoustic modelling using deep belief networks," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 20, no. 1, pp. 14–22, 2012.

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

[22] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of ICML*, 2010, pp. 807–814.

[23] N.Srivastava, G.Hinton, A.Krizhevsky, I.Sutskever, and R.Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[24] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of ICML*, 2015, pp. 448–456.

[25] D.Povey, A.Ghoshal, G.Boulianne, L.Burget, O.Glembek, N.Goel, M.Hannemann, P.Motlicek, Y.Qian, P.Schwarz, J.Silovsky, G.Stemmer, and K.Vesely, "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011, pp. 1–4.

[26] X.Anguera, C.Wooters, and J.Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech & Language Process.*, vol. 15, no. 7, pp. 2011–2023, 2007.

[27] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, & Language Proc.*, vol. 15, no. 1, pp. 70–79, 2007.

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