ONLINE INTEGRATION OF DNN-BASED AND SPATIAL CLUSTERING-BASED MASK ESTIMATION FOR ROBUST MVDR BEAMFORMING

Yutaro Matsui[1,2], *Tomohiro Nakatani*[1], *Marc Delcroix*[1], *Keisuke Kinoshita*[1],
Nobutaka Ito[1], *Shoko Araki*[1], *Shoji Makino*[2]

[1]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[2]University of Tsukuba, 1–1–1 Tennodai, Tsukuba, Ibaraki, Japan
y.matsui@mmlab.cs.tsukuba.ac.jp, nakatani.tomohiro@lab.ntt.co.jp

ABSTRACT

This paper discusses the online estimation of time-frequency masks, which enables us to perform mask-based beamforming by online processing for robust automatic speech recognition (ASR). Two approaches to online mask estimation have been separately developed for this purpose. One is based on a deep neural network (DNN), which exploits the spectral features of the signal. The other is based on spatial clustering (SC), which exploits the spatial features of the signal. This paper proposes a new method that integrates the two online estimation approaches to further improve online mask estimation by exploiting the advantages of both approaches. Experiments using the real data of the CHiME-3 multichannel noisy speech corpus show that the proposed method greatly outperforms the conventional approaches in terms of improving the word error rate (WER).

Index Terms— Beamforming, time-frequency masking, online processing, DNN, spatial clustering

1. INTRODUCTION

When our daily speech is captured using distant microphones, various types of ambient noise are mixed with the captured signals, and severely degrade the ASR performance. To solve this problem, beamforming is being extensively studied as a noise reduction frontend for ASR. Delay-and-sum beamforming, minimum variance distortionless response (MVDR) beamforming, and maximum signal-to-noise ratio beamforming are often employed [1–3], and have been shown to improve the ASR performance in tasks ranging from medium vocabulary distant speech recognition [4] to large vocabulary meeting transcription [5, 6].

Accurate estimation of captured signal characteristics, such as the spatial covariance matrices of the speech and the noise, is crucial for performing effective beamforming. For this purpose, researchers have recently proposed time-frequency mask-based beamforming approaches [7–11]. The central idea is to leverage the spectral sparsity of speech signals by using time-frequency masks that represent the probability of speech (or noise) dominating the corresponding time-frequency points [12–14]. Then, the spatial covariance matrices of the speech and the noise are estimated solely from the time-frequency masks and the captured signal, and used for constructing beamformers.

The two main techniques proposed for mask estimation are the DNN-based and SC-based approaches. With the DNN-based approach [10, 11, 15], a DNN is trained in advance on training data so that it can estimate masks from the spectral features of a noisy speech signal. Then, the trained network is used to estimate the masks for test data. The SC-based approach [7, 9, 16–18], on the other hand, requires no prior training on the training data (for offline processing), and can estimate masks from the test data in an unsupervised learning manner. On the assumption that the spatial features of speech and noise have different distributions, this approach finds these two distributions based on the clustering of the spatial features, and the masks are estimated as the posteriors of each cluster at the corresponding time-frequency points. For these conventional approaches, both offline and online processing [9, 19–22] have already been formulated by researchers.

A joint optimization technique that integrates the above two approaches has also been proposed for offline processing [23, 24]. In [23], given a joint likelihood function for spectral features and spatial features, a mask estimation method was derived based on the expectation-maximization (EM) algorithm. With the method, the initial masks are first estimated based solely on a DNN, and then utilized by the SC-based mask estimation as the time-frequency dependent prior of the dominant sources to estimate the integrated masks. Experiments showed that the integrated approach was more effective than the two conventional mask estimation approaches.

This paper extends the above joint optimization approach to online mask estimation by integrating the two conventional online mask estimation approaches. This extension is important, for example, to provide frontend for low latency distant speech recognition, such as ASR in smart speakers. In the integrated framework, while the two conventional approaches perform online mask estimation in parallel, the integration is conducted at each time segment in a way similar to that used for the offline integration framework. In addition, this paper presents an effective way of initializing the model parameters of SC for online processing without prior knowledge of the speaker location, which has been difficult with conventional SC-based online mask estimation. In the experiments, an MVDR beamformer is performed based on the estimated masks and the result is evaluated using the real data of the CHiME-3 multichannel noisy speech corpus [25]. The proposed method is shown to greatly outperform the two conventional online approaches.

In the remainder of this paper, two conventional online mask estimation methods are described in Section 2, and the proposed method is presented in Section 3. Sections 4 and 5, respectively, provide experimental results and concluding remarks.

2. CONVENTIONAL ONLINE MASK ESTIMATION

Let us assume that a single speech signal is captured by multiple microphones with certain additive diffuse noise. Then, the goal of this paper is to present a method for estimating masks at individual time-

978-1-5386-8151-0/18/$31.00 ©2018 IEEE

frequency (TF) points in the short-time Fourier transform (STFT) domain that represents the probability of speech (or noise) dominating the respective TF points. With accurate mask estimates, we can design an effective beamformer to suppress the noise [9].

In the following, we first describe the two conventional online mask estimation methods, one based on SC [9] and the other based on DNN [19]. Then, in the next section, we present our proposed method, which integrates the two methods.

2.1. Spatial clustering-based online mask estimation

Figure 1 illustrates the processing flow of SC-based online mask estimation. In the flow, spatial features, $\boldsymbol{X}_{n,f}$, are first extracted from the captured signal at individual TF points as

$$\boldsymbol{y}_{n,f} = \left[y_{n,f,1}, \cdots, y_{n,f,M}\right]^T, \tag{1}$$

$$\boldsymbol{X}_{n,f} = \frac{\boldsymbol{y}_{n,f}}{\|\boldsymbol{y}_{n,f}\|}, \tag{2}$$

where $y_{n,f,m}$ is an STFT of the captured signal obtained by the m-th microphone ($1 \leq m \leq N_m$) at time n ($1 \leq n \leq N_t$) and frequency f ($0 \leq f \leq N_f$), $\boldsymbol{y}_{n,f}$ is a vector containing $y_{n,f,m}$ for all the microphones, T denotes non-conjugate transposition, and $\|\cdot\|$ denotes the Euclidean norm. If we assume that each TF point is dominated by either speech or noise according to the spectral sparsity of speech, $\boldsymbol{X}_{n,f}$ at each TF point has a spatial characteristic approximately corresponding to that of the dominant source. Then, by clustering the TF points based on the spatial features, we can obtain two clusters, one dominated by speech and the other dominated by noise. We can then estimate the masks using these clusters as a basis.

For the clustering, a complex angular central Gaussian mixture model (cACGMM) has been shown to be effective for modeling the distribution of $\boldsymbol{X}_{n,f}$ [18,26]. The model is defined at each frequency f as

$$p(\boldsymbol{X}_{n,f}; \theta^{\mathrm{SC}}) = \sum_{d=0}^{1} w_f^d \mathcal{A}(\boldsymbol{X}_{n,f} | d_{n,f} = d; \theta^{\mathrm{SC}}), \tag{3}$$

$$\mathcal{A}(\boldsymbol{X} | d_{n,f} = d; \theta^{\mathrm{SC}}) = \frac{(N_m - 1)!}{2\pi^{N_m} \det \mathbf{R}_f^d} \frac{1}{(\boldsymbol{X}^H (\mathbf{R}_f^d)^{-1} \boldsymbol{X})^{N_m}}, \tag{4}$$

where a random variable $d_{n,f}$, referred to as a dominant source index, indicates whether speech or noise dominates each TF point ($d_{n,f} = 0$ for noise and 1 for speech), w_f^d is the mixture weight, which corresponds to the prior of $d_{n,f}$, i.e., $w_f^d = p(d_{n,f} = d)$, and $\mathcal{A}(\boldsymbol{X}|d; \theta^{\mathrm{SC}})$ is the conditional distribution of $\boldsymbol{X}$ given d, defined as the complex angular central Gaussian (cACG) distribution. $\mathbf{R}_f^d$ is a shape parameter of a cACG distribution, which is characterized by an $N_m \times N_m$ positive definite Hermitian matrix and roughly corresponds to the spatial covariance matrix of a signal, and H denotes conjugate transposition. With this model, clustering is performed by estimating a set of model parameters, $\theta^{\mathrm{SC}} = \{\{w_f^d\}, \{\mathbf{R}_f^d\}\}$, based on maximum likelihood estimation. An efficient estimation method can be derived based on the EM algorithm. Then, the masks, $M_{n,f}^{d,\mathrm{SC}}$, are estimated as the posterior probability of individual TF points being dominated by the noise ($d = 0$) and by the speech ($d = 1$), i.e., $M_{n,f}^{d,\mathrm{SC}} = p(d_{n,f} = d | \boldsymbol{X}_{n,f}; \theta^{\mathrm{SC}})$.

Online mask estimation method can be formulated based on the online EM algorithm [9]. With the method, the captured signals are separated into short segments of the order of hundreds of milliseconds, which are referred to as minibatches, and the EM algorithm is performed for each minibatch one after another. Hereafter, each minibatch is denoted by $\mathcal{B}_l$ where l is the index of the minibatch, and is assumed to be composed of a set of time frames included in the corresponding segment. To accumulate the information extracted from each minibatch, the model parameters, $\{\mathbf{R}_{f,l-1}^d\}$, estimated in a minibatch, $\mathcal{B}_{l-1}$, are passed to the next minibatch, $\mathcal{B}_l$, and used for the optimization with a certain forgetting factor. (See [9] for the details of the online estimation.)

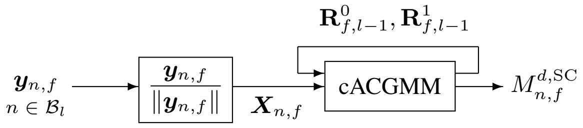
$\mathbf{R}_{f,l-1}^0, \mathbf{R}_{f,l-1}^1$
$\boldsymbol{y}_{n,f}$
$n \in \mathcal{B}_l$
$\frac{\boldsymbol{y}_{n,f}}{\|\boldsymbol{y}_{n,f}\|}$
$\boldsymbol{X}_{n,f}$
cACGMM
$M_{n,f}^{d,\mathrm{SC}}$

Fig. 1. Processing flow of SC-based online mask estimation

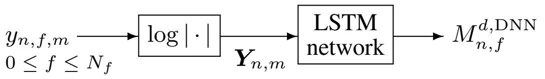

Fig. 2. Processing flow of DNN-based online mask estimation

Note that the initial values of $\mathbf{R}_{f,0}^d$ have to be set appropriately for the first minibatch, $\mathcal{B}_1$, in the above online estimation because otherwise the convergence is not very fast and it is very difficult to solve the permutation problem between speech and noise in an online processing manner. This requires the speaker location to be roughly fixed in advance and the initial values of $\mathbf{R}_{f,0}^1$ to be given for the location, which greatly limits the applicability of the online algorithm. For example, it was assumed for the CHiME-3 challenge that the speaker speaks while facing a tablet equipped with a set of microphones.

2.2. DNN-based online mask estimation

Figure 2 illustrates the processing flow of the DNN-based mask estimation. A bi-directional long short-term memory (BLSTM) network is widely used for DNN-based mask estimation [10], however, it cannot be applied to online processing due to its backward propagation path. Instead, by dropping the backward path from the BLSTM network, we obtain a long short-term memory (LSTM) network that can be applied to online mask estimation [19]. The LSTM network receives the spectral features of the signal, defined in eq. (6), and estimates masks in a frame-by-frame processing manner.

$$Y_{n,f,m} = \log|y_{n,f,m}| \tag{5}$$

$$\boldsymbol{Y}_{n,m} = \left[Y_{n,0,m}, \cdots, Y_{n,N_f,m}\right] \tag{6}$$

The LSTM network is composed of an LSTM layer followed by three feed-forward neural network layers. To train the network, we adopted the same scheme as used in [10], except that batch normalization was not performed as it is unsuitable for online processing. The desired output is a concatenation of two types of ideal binary masks, one for speech and the other for noise. The ideal binary masks for speech (or those for noise) take 1 when the TF points are dominated by the speech (or noise) and take 0 otherwise. Accordingly, the trained LSTM network outputs two different masks, $M_{n,f}^{d,\mathrm{DNN}}$ for $d = 0$ and 1, respectively, for the noise and the speech.

Note that, to obtain ideal binary masks, only simulated data can be used as the training data, from which we can extract the microphone images of speech and noise separately. So, the LSTM network performance often deteriorates when estimating masks for real data that are recorded in real acoustic environments because of the mismatch between simulated and real data.

3. PROPOSED ONLINE MASK ESTIMATION

For the integration of the two online mask estimation methods, we use the same scheme as one adopted for an offline integration framework [23]. For this purpose, we first introduce the following joint likelihood function.

$$\mathcal{L}(\theta^{\mathrm{SC}}) = p(\mathcal{X}_l, \mathcal{Y}_l; \theta^{\mathrm{DNN}}, \theta^{\mathrm{SC}}),$$

where $\mathcal{X}_l$ and $\mathcal{Y}_l$ are sets of $\boldsymbol{X}_{n,f}$ and $\boldsymbol{Y}_{n,m}$ s.t. $n \in \mathcal{B}_{l'}$ and $1 \leq l' \leq l$, and θ^{DNN} is a set of parameters for the LSTM network that are assumed to be fixed by pre-training. Then, with proper assumptions as regards conditional independence over time-frequency points that are commonly used in the conventional approaches [27], the above function is rewritten, disregarding constant terms, as

$$\mathcal{L}(\theta^{\mathrm{SC}}) = \prod_{n,f} \sum_{d} p(\boldsymbol{X}_{n,f} | d_{n,f} = d; \theta^{\mathrm{SC}}) p(d_{n,f} = d | \mathcal{Y}_l; \theta^{\mathrm{DNN}}),$$

where $p(\boldsymbol{X}_{n,f} | d_{n,f} = d; , \theta^{\mathrm{SC}})$ corresponds to the cACG distribution in eq. (3) and $p(d_{n,f} = d | \mathcal{Y}_l; \theta^{\mathrm{DNN}})$ can be interpreted as the masks $M_{n,f}^{d,\mathrm{DNN}}$ obtained by the LSTM network. With this interpretation, the above equation can be viewed as a variation of cACGMM with a time-frequency dependent mixture weight, defined as $M_{n,f}^{d,\mathrm{DNN}}$. Accordingly, the EM algorithm can be applied in the same way as SC-based online mask estimation [9] for the estimation of θ^{SC} and the integrated masks, $M_{n,f}^{d,\mathrm{INT}}$. In particular, because $M_{n,f}^{d,\mathrm{DNN}}$ is obtained by online processing, we can also formulate an online EM algorithm for the integrated method.

3.1. Integrated online mask estimation

Figure 3 shows the processing flow of the proposed online mask estimation method. In the flow, the integrated mask, $M_{n,f}^{d,\mathrm{INT}}$, in each minibatch $\mathcal{B}_l$ is estimated as follows:

1. The LSTM network first receives the spectral features, and estimates the masks, $M_{n,f}^{d,\mathrm{DNN}}$ in the minibatch $\mathcal{B}_l$.
2. The cACGMM receives the spatial features, $\boldsymbol{X}_{n,f}$, the shape parameters, $\mathbf{R}_{f,l-1}^{d}$, estimated in the previous minibatch, and $M_{n,f}^{d,\mathrm{DNN}}$. It then sets the initial values of the integrated masks, $M_{n,f}^{d,\mathrm{INT}}$, the shape parameters, $\mathbf{R}_{f,l}^{d}$, and the cumulative sum of the masks, $\Lambda_{f,l}^{d}$ as
$$\begin{aligned} M_{n,f}^{d,\mathrm{INT}} &= M_{n,f}^{d,\mathrm{DNN}}, \\ \mathbf{R}_{f,l}^{d} &= \mathbf{R}_{f,l-1}^{d}, \\ \Lambda_{f,l}^{d} &= \Lambda_{f,l-1}^{d} + \sum_{n \in \mathcal{B}_l} M_{n,f}^{d,\mathrm{DNN}}. \end{aligned}$$
3. The cACGMM updates $\mathbf{R}_{f,l}^{d}$ and $M_{n,f}^{d,\mathrm{INT}}$ by using EM iterations as shown below, and outputs both of them.

   (a) In M-step:
$$\mathbf{R}_{f,l}^{d} = \frac{\Lambda_{f,l-1}^{d}}{\Lambda_{f,l}^{d}} \mathbf{R}_{f,l-1}^{d} + \frac{1}{\Lambda_{f,l}^{d}} \mathbf{R}_{f,\mathrm{new}}^{d}$$

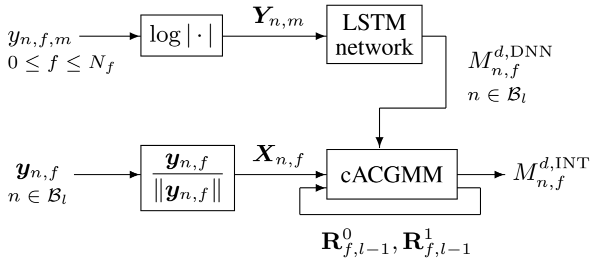

Fig. 3. Processing flow of integrated online mask estimation

   where
$$\mathbf{R}_{f,\mathrm{new}}^{d} = N_m \sum_{n \in \mathcal{B}_l} \frac{M_{n,f}^{d,\mathrm{INT}} \boldsymbol{X}_{n,f} \boldsymbol{X}_{n,f}^{H}}{\boldsymbol{X}_{n,f}^{H} \left(\mathbf{R}_{f,l}^{d}\right)^{-1} \boldsymbol{X}_{n,f}}$$

   (b) In E-step:
$$M_{n,f}^{d,\mathrm{INT}} = \frac{M_{n,f}^{d,\mathrm{DNN}} \mathcal{A}\left(\boldsymbol{X}_{n,f} | d; \theta^{\mathrm{SC}}\right)}{\sum_{d} M_{n,f}^{d,\mathrm{DNN}} \mathcal{A}\left(\boldsymbol{X}_{n,f} | d; \theta^{\mathrm{SC}}\right)}$$

In our experiments, we set the number of EM iterations at one for each minibatch.

3.2. Initialization of shape parameters

In the above online processing, the shape parameters, $\mathbf{R}_{f,0}^{d}$, and the cumulative sum of the masks, $\Lambda_{f,0}^{d}$, must be initialized at the beginning. It is natural to set $\Lambda_{f,0}^{d} = 0$ and to obtain $\mathbf{R}_{f,0}^{0}$ for noise from the test data when speech is absent. In contrast, the effective initialization of $\mathbf{R}_{f,0}^{1}$ for speech is the key to successful online estimation. This paper prepares the following three methods to achieve this, and compares them experimentally in the next section.

- **PreTrained:** $\mathbf{R}_{f,0}^{1}$ are trained in advance using certain training data that contain utterances from a pre-determined speaker location without additive noise. This method is identical to that used for the SC-based online mask estimation [9].
- **NoPrior:** $\mathbf{R}_{f,0}^{1}$ are set as identity matrices.
- **PostTrained:** $\mathbf{R}_{f,0}^{1}$ are initialized in the same way as NoPrior. However, $\mathbf{R}_{f,l}^{d}$ are not used for estimating $M_{n,f}^{d,\mathrm{INT}}$ at early minibatches until they have been reliably trained using the test data. With the early minibatches, $M_{n,f}^{d,\mathrm{INT}}$ is substituted with $M_{n,f}^{d,\mathrm{DNN}}$. $\mathbf{R}_{f,l}^{d}$ are determined to be reliable when $\Lambda_{f,l}^{1}$ exceeds a certain predetermined threshold because $\Lambda_{f,l}^{1}$ indicates the number of TF points that are dominated by speech and included in current and past minibatches.

Note that no prior training on $\mathbf{R}_{f,0}^{d}$ is required for NoPrior or PostTrained, and this is beneficial in terms of widening the applicability of SC-based online mask estimation.

4. EXPERIMENTAL EVALUATION

We conducted ASR experiments using the CHiME-3 Speech Separation and Recognition Challenge corpus [25] to evaluate the noise reduction performance of our proposed approach. The corpus was created by using a six-channel microphone array attached to a tablet

Table 1. The number of utterance data in the CHiME-3 corpus.

	real data	simulated (simu) data
Training set	1600	7138
Development (dev) set	1640	1640
Evaluation (eval) set	1320	1320

Table 2. Experimental conditions.

Sampling frequency	16 kHz
Frame length	64 ms
Frame overlap	75%
Window function	Hanning
Number of EM iterations	1
Number of microphones	6

device. The recordings were obtained in four different noisy environments, i.e., public transport (Bus), cafe (Caf), pedestrian area (Ped), and street junction (Str), and they feature several male and female speakers uttering sentences from the Wall Street Journal (WSJ) [28]. The corpus is divided into three individual subsets as Table 1, each containing both real data and simulated (simu) data.

For the evaluation, we employed a mask-based MVDR beamformer that was used for evaluating offline/online SC-based mask estimation methods in [9, 23], and used the speech recognizer described in [29], which was referred to as 1-pass SI and based on a multi-condition convolutional neural network (CNN) acoustic model and a recurrent neural network (RNN) language model.

4.1. Evaluation of online processing methods

We compared three online mask estimation methods, i.e., the DNN-based method using the LSTM network (LSTM), the SC-based method using cACGMM (cACGMM), and our proposed method, which integrates the two methods (Proposed). We used "PreTrained" shape parameters for the initialization of both cACGMM and Proposed. $\mathbf{R}^1_{f,0}$ was trained on speech signals recorded in a quiet booth and $\mathbf{R}^0_{f,0}$ was trained on signals that only contained noise. Both data are included in the CHiME-3 corpus. Also, based on the experimental setting adopted in [9], we set the size of the first minibatch at 500 ms and that of succeeding minibatches at 250 ms to ensure that the first minibatch contained the target speech signal. Other hyperparameters were set as shown in Table 2.

Table 3 summarizes the WERs obtained in the experiments. In the table, the proposed method outperformed cACGMM for both the real and simu data, and outperformed LSTM for the real data. Although LSTM was slightly better than the proposed method for the simu data, this is because there are almost no mismatches between the training and test sets in the simu data. Note that the performance on the simu data was not considered an official performance metric in the CHiME-3 challenge. Hereafter, we only focus on the performance on the real data.

Table 3. WERs (%) obtained using different online mask estimators. cACGMM and Proposed both used "PreTrained" shape parameters for shape parameter initialization. Note that the WERs (%) obtained when we performed offline processing with the BLSTM network, cACGMM, and the integrated approach were 7.34, 8.28, and 7.10, respectively, for real data of the eval set.

		dev			eval	
	Ave	simu	real	Ave	simu	real
No frontend	8.62	8.24	9.01	12.89	10.17	15.60
LSTM	5.33	**5.33**	5.33	8.14	**6.59**	9.69
cACGMM	5.60	5.98	5.21	9.20	8.86	9.53
Proposed	**4.98**	**5.33**	**4.63**	**7.35**	6.81	**7.89**

Table 4. WERs (%) obtained using online mask estimation methods, LSTM and two proposed methods (NoPrior and PostTrained). No method requires prior training of the shape parameters.

	Set	Bus	Caf	Ped	Str	Ave
LSTM	dev	6.34	5.32	4.69	4.96	5.33
NoPrior		6.06	4.76	4.59	**4.59**	5.00
PostTrained		**5.96**	**4.68**	**4.48**	4.62	**4.94**
LSTM	eval	13.48	9.81	8.00	7.45	9.69
NoPrior		12.15	**8.42**	**6.86**	**6.65**	8.52
PostTrained		**11.09**	8.48	7.29	**6.65**	**8.38**

4.2. Evaluation of shape parameter initialization methods

Next, we compared three online estimation methods that do not require pre-trained shape parameters of the cACGMM, i.e., LSTM, the proposed method with NoPrior (NoPrior), and that with PostTrained (PostTrained). With both NoPrior and PostTrained, the shape parameters are initialized as the identity matrices. Whereas NoPrior uses the integrated masks from the beginning of the online estimation, PostTrained started using them only after the cumulative sum of the masks obtained by LSTM exceeded a certain predetermined threshold. In the experiment, the threshold was set at 1.5 because it achieved the best WER for the dev set. Other hyperparameters are the same as those used for the experiments in section 4.1. Note that the cACGMM is not included in this comparison because it cannot perform online mask estimation without prior training.

Table 4 summarizes the WERs obtained in the experiments. We can see that the two proposed methods without any prior training of the shape parameters consistently outperformed LSTM. This means that the integration could successfully improve the accuracy of the LSTM-based online mask estimation by incorporating the SC-based online mask estimation with no prior training. When we compare the two proposed methods, PostTrained was slightly better than NoPrior.

5. CONCLUDING REMARKS

This paper proposed a new online mask estimation technique that integrates two conventional online mask estimation methods, one using an LSTM network and the other using a cACGMM-based SC. The integration is formulated based on a joint likelihood function defined on two different features of the signal, namely the spectral and spatial features, and is performed by optimizing the joint function using the online EM algorithm. The proposed method can improve the mask estimation accuracy by exploiting the two different features simultaneously, and it enables SC-based mask estimation to realize online processing with no prior training on the shape parameters of the cACGMM. Experiments using real data from the CHiME-3 multichannel noisy speech corpus showed that the proposed method greatly outperformed the conventional approaches in terms of WER reduction when it was combined with mask-based MVDR beamforming.

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] 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.

[3] 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, 2007.

[4] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, “Strategies for distant speech recognition in reverberant environments,” *EURASIP J. Adv. Signal Process*, vol. Article ID 2015:60, doi:10.1186/s13634-015-0245-7, 2015.

[5] P. Swietojanski, A. Ghoshal, and S. Renals, “Hybrid acoustic models for distant multichannel large vocabulary speech recognition,” *Proc. IEEE ASRU-2013*, pp. 285–290, 2015.

[6] T. Yoshioka, X. Chen, and M. J. F. Gales, “Impact of single-microphone dereverberation on DNN-based meeting transcription systems,” *Proc. IEEE ICASSP-2014*, pp. 5527–5531, 2014.

[7] D. H. Tran Vu and R. Haeb-Umbach, “Blind speech separation employing directional statistics in an expectation maximization framework,” *Proc. IEEE ICASSP-2010*, pp. 241–244, 2010.

[8] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, “A multichannel MMSE-based framework for speech source separation and noise reduction,” *IEEE Trans. ASLP*, 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,” *Proc. IEEE ICASSP-2016*, pp. 5210–5214, 2016.

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

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

[12] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[13] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, pp. 181–197. 2005.

[14] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” *Signal Processing*, vol. doi:10.1016/j.sigpro.2007.02.003, 2007.

[15] Y. Wang and D. Wang, “Towards scaling up classification-based speech separation,” *IEEE Trans. ASLP*, vol. 21, no. 7, pp. 1381–1390, 2013.

[16] M. I. Mandel, D. P. W. Ellis, and T. Jebara, “An EM algorithm for localizing multiple sound: Sources in reverberant environments,” in *Proc. the 2006 Conference on Advances in Neural Information Processing Systems*, pp. 953–960. MIT Press, 2011.

[17] H. Sawada, S. Araki, and S. Makino, “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” *IEEE Trans. ASLP*, vol. 19, no. 3, pp. 516–527, 2011.

[18] N. Ito, S. Araki, and T. Nakatani, “Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing,” *Proc. EUSIPCO-2016*, pp. 1153–1157, 2016.

[19] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, T. Morita, and T. Nakatani, “Online speech extraction and recognition from monaural speech mixtures,” *Spring Meeting of Acoustical Society of Japan*, pp. 97–98, 2018.

[20] J. Heymann, M. Bacchiani, and T. N. Sainath, “Performance of mask based statistical beamforming in a smart home scenario,” *Proc. IEEE ICASSP-2018*, pp. 6722–6726, 2018.

[21] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, “Exploring practical aspects of neural mask-based beamforming for far-field speech recognition,” *Proc. IEEE ICASSP-2018*, pp. 6697–6701, 2018.

[22] T. Higuchi, K. Kinoshita, N. Ito, S. Karita, and T. Nakatani, “Frame-by-frame closed-form update for mask-based adaptive MVDR beamforming,” *Proc. IEEE ICASSP-2018*, pp. 531–535, 2018.

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

[24] L. Drude and R. Haeb-Umbach, “Tight integration of spatial and spectral features for BSS with deep clustering embeddings,” *Proc. Interspeech-2017*, pp. 2650–2654, 2017.

[25] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” *Proc. IEEE ASRU-2015*, pp. 504–511, 2015.

[26] J. T. Kent, “Data analysis for shapes and images,” *Journal of Statistical Planning and Inference*, vol. 57, no. 2, pp. 181–193, 1997.

[27] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujimoto, “Dominance based integration of spatial and spectral features for speech enhancement,” *IEEE Trans. ASLP*, vol. 21, no. 12, pp. 2516–2531, 2013.

[28] D. B. Paul and J. M. Baker, “The design for the Wall Street Journal-based CSR corpus,” *HLT-91 Proc. the Workshop on Speech and Natural Language*, pp. 357–362, 1992.

[29] 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,” *Proc. IEEE ASRU-2015*, pp. 436–443, 2015.