MASK-BASED MVDR BEAMFORMER FOR NOISY MULTISOURCE ENVIRONMENTS: INTRODUCTION OF TIME-VARYING SPATIAL COVARIANCE MODEL

Yuki Kubo[†,⋆] *Tomohiro Nakatani*[†] *Marc Delcroix*[†] *Keisuke Kinoshita*[†] *Shoko Araki*[†]

[†] NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
[⋆] The University of Tokyo, Tokyo, Japan

ABSTRACT

This paper proposes a method for designing a time-varying minimum variance distortionless response (MVDR) beamformer using time-frequency masks, with the aim of improving speech enhancement in noisy multi-speaker environments. A key to successful beamforming is to estimate accurately a time-varying spatial covariance matrix (SCM) for noise composed of both stationary diffuse noise and highly time-varying speech. For this purpose, we introduce a stochastic model that can represent the time-varying characteristics of the noise SCM, and derive a method for estimating a time-varying noise SCM based on the model. Experiments show that the proposed method can substantially improve the performance of the beamformer in terms of automatic speech recognition (ASR) accuracy and source-to-distortion ratio compared with a conventional time-invariant MVDR beamformer.

Index Terms— Beamforming, Time-frequency mask, Time-varying processing, Speech enhancement, ASR

1. INTRODUCTION

Recently, a mask-based MVDR beamformer has been extensively studied as a frontend for far-field ASR [1, 2, 3, 4, 5]. This beamformer is advantageous for improving the performance of a neural network-based ASR backend because it can reduce noise while precisely maintaining the spectral shape of target speech. With this approach, the masks are used to indicate the probability of speech (or noise) dominating individual time-frequency (TF) bins [6, 7, 8], and to estimate the spatial covariance matrices (SCMs) of the target speech (or the noise). The effectiveness of this approach has been shown in tasks ranging from medium vocabulary far-field ASR [3, 4, 5] to large vocabulary meeting transcription [9].

The accurate estimation of an SCM for noise is crucial when designing an MVDR beamformer. Conventional approaches estimate the noise SCM by averaging or smoothing the spatial characteristics of the noise over a rather long time duration, assuming that the noise SCM is time invariant or slowly time-varying [3, 4, 5, 10, 11, 12]. However, when the noise to be reduced includes both stationary diffuse noise and highly time-varying sounds such as speech, the noise SCM becomes time-varying even within a short time duration of the order of a few hundreds of milliseconds. Accordingly, the conventional approaches cannot precisely estimate the noise SCM, and the noise reduction performance of the MVDR beamformer is rather limited.

In this paper, we discuss a method for designing a time-varying MVDR beamformer that enhances target speech in noisy multi-speaker environments by offline processing. Since speech is one of the noise sources to be reduced, it is desirable to estimate the time-varying noise SCM with high time resolution of the order of a few tens of milliseconds. For this purpose, we introduce a stochastic model that can represent the time-varying characteristics of the noise SCM. In the model, a complex inverse Wishart mixture model (cIWMM) is used as the prior distribution of the noise SCM [13], where rapid change in the noise SCM is represented by its time-varying mixture weights, which are determined based on masks that indicate which noise source dominates which TF bin. A maximum a posteriori estimation (MAP) method is derived to obtain the time-varying noise SCM.

We conducted simulation experiments on the enhancement of target speakers from a set of multichannel sound mixtures, each of which contained three simultaneous speakers and diffuse babble noise. Two different types of masks, oracle masks and estimated masks, were examined, where the oracle masks are the power ratios of the target speech to the noise in the simulation data. With both masks, the proposed method achieved significantly better word error rates (WERs) and source-to-distortion ratios (SDRs) [14] for the enhanced signals than a time-invariant MVDR beamformer. Furthermore, WERs obtained with the proposed method decreased when we used a shorter time block, with the maximum error reduction rate being 20.3 % compared with the time-invariant beamformer.

In the remainder of this paper, Section 2 summarizes related work. Sections 3 and 4, respectively, describe the conventional mask-based MVDR beamformer and the proposed time-varying MVDR beamformer. Experimental results and concluding remarks can be found in Sections 5 and 6, respectively.

2. RELATED WORK

Researchers have reported that the MVDR beamformer is effective in improving the performance of far-field ASR [4, 10]. However, little work has been done on whether or not it is advantageous to make the beamformer time-varying.

Certain online beamforming techniques enable us to recursively update the noise SCM in a frame-by-frame manner [11, 15]. However, the goal of these techniques is to achieve reliable online beamforming with a performance level similar to that obtained with offline processing. Therefore, a rather small forgetting factor is adopted for the recursive update, resulting in the noise SCM being greatly smoothed over a long duration.

Full-rank spatial covariance analysis (FCA), which was proposed for underdetermined blind source separation [16], models time-varying SCMs by employing the weighted sum of time-invariant full-rank spatial covariance matrices. This model is closely related to the cIWMM used in this paper, but it has never before been employed to design time-varying MVDR beamformer. The same is true of a complex Gaussian mixture model (cGMM) [3], which was proposed for mask-based speech enhancement.

3. MASK-BASED BEAMFORMING

This section gives a brief overview of mask-based MVDR beamforming. Here, masks indicate the probability of each source dominating each TF bin. Many techniques have been proposed for estimating the masks, e.g., techniques using cGMM [3], neural networks [4, 5, 17, 18, 19], or combination of the two [20, 21]. Throughout this paper, we assume that masks are estimated with certain existing methods, and focus only on how to design the MVDR beamformer based on the estimated masks.

3.1. Estimation of a steering vector using estimated masks

Suppose that a signal captured by I microphones contains a target speech signal and certain additive noise, which is composed of one or more noise sources. The captured signal is denoted as $\boldsymbol{x}_{tf} = (x_{tf,1}, \ldots, x_{tf,i}, \ldots, x_{tf,I})^\top \in \mathbb{C}^I$ in the STFT domain, where $t = 1, \ldots, T$, $f = 1, \ldots, F$, and $i = 1, \ldots, I$ denote the indices of time frames, frequency bins and microphones, respectively. The superscript $\top$ denotes transpose.

With the mask-based MVDR beamformer, the estimated masks are used to estimate an SCM for each source, separately, in the captured signal. Let $j = 1, \ldots, J$ represent the index of a source, where $j = v$ and $j \neq v$, respectively, correspond to the target speech and the other noise sources, and $\lambda_{tf}^{(j)} \in \mathbb{R}$ is the estimated mask that satisfies $\lambda_{tf}^{(j)} \geq 0$ and $\sum_j \lambda_{tf}^{(j)} = 1$. Then the SCM for each source j is estimated as

$$\hat{\mathcal{R}}_f^{(j)} = \frac{1}{\sum_t \lambda_{tf}^{(j)}} \sum_t \lambda_{tf}^{(j)} \boldsymbol{x}_{tf} \boldsymbol{x}_{tf}^{\mathrm{H}}, \tag{1}$$

where the superscript H denotes Hermitian transpose.

A steering vector, $\boldsymbol{h}_f$, that contains the room transfer functions from the target speaker to the microphones, is estimated using generalized eigenvalue decomposition with noise covariance whitening [22, 23] as

$$\boldsymbol{h}_f = \hat{\mathcal{R}}_f^{(n)} \mathrm{Ev}\{(\hat{\mathcal{R}}_f^{(n)})^{-1} \hat{\mathcal{R}}_f^{(v)}\}. \tag{2}$$

Here, $\mathrm{Ev}\{\cdot\}$ is a function for extracting an eigenvector corresponding to the maximum eigenvalue. $\hat{\mathcal{R}}_f^{(v)}$ and $\hat{\mathcal{R}}_f^{(n)}$ are, respectively, the estimated SCMs of the target speech and the noise, where $\hat{\mathcal{R}}_f^{(n)}$ is equal to the sum of the SCMs of all the noise sources and calculated as

$$\hat{\mathcal{R}}_f^{(n)} = \sum_{j \neq v} \hat{\mathcal{R}}_f^{(j)}. \tag{3}$$

3.2. MVDR beamforming

Given the steering vector, $\boldsymbol{h}_f$, an MVDR beamformer is estimated as a filter, $\boldsymbol{w}_f \in \mathbb{C}^I$, that minimizes the power of the noise, i.e., $|\boldsymbol{w}_f^{\mathrm{H}} \hat{\mathcal{R}}_f^{(n)} \boldsymbol{w}_f|^2$, with a distortionless constraint on the target speech direction, i.e., $\boldsymbol{w}_f^{\mathrm{H}} \boldsymbol{h}_f = 1$, as follows:

$$\boldsymbol{w}_f = \frac{(\hat{\mathcal{R}}_f^{(n)})^{-1} \boldsymbol{h}_f}{\boldsymbol{h}_f^{\mathrm{H}} (\hat{\mathcal{R}}_f^{(n)})^{-1} \boldsymbol{h}_f}. \tag{4}$$

Finally, we obtain the estimate of the target speech, $\hat{s}_{tf}$, by multiplying the filter, with the captured signal as follows:

$$\hat{s}_{tf} = \boldsymbol{w}_f^{\mathrm{H}} \boldsymbol{x}_{tf}. \tag{5}$$

4. PROPOSED METHOD

4.1. Need for time-varying SCM

In eq. (3), the noise SCM is obtained by mixing the time-invariant SCMs of all the noise sources, calculated by eq. (1), with equal time-independent contributions. However, when the noise includes a highly time-varying sound source, such as speech, the contribution of the SCM from that source to the noise SCM should also be time-varying. As a consequence, the noise SCM obtained with eq. (3) contains large mismatches with the true noise SCM at each short time block, and this substantially lowers the upper performance limit of the MVDR beamformer.

To overcome the above limit, we propose a way of estimating the time-varying noise SCM based on the estimated mask. Assuming that the captured signals are divided into consecutive short time blocks, denoted by B_k for $k = 1, 2, \cdots$, we derive, in the following, a way of estimating the noise SCM, $\hat{\mathcal{R}}_{k,f}^{(n)}$ at each short time block, which then yields the resultant time-varying MVDR beamformer as follows:

$$\boldsymbol{w}_{k,f} = \frac{(\hat{\mathcal{R}}_{k,f}^{(n)})^{-1} \boldsymbol{h}_f}{\boldsymbol{h}_f^{\mathrm{H}} (\hat{\mathcal{R}}_{k,f}^{(n)})^{-1} \boldsymbol{h}_f}. \tag{6}$$

4.2. Model-based time-varying SCM estimation

This section presents a stochastic model for the noise SCM, and a way of calculating $\hat{\mathcal{R}}_{k,f}^{(n)}$ based on the MAP estimation.

4.2.1. Model of SCMs

We first assume that the captured signal, $\boldsymbol{x}_{tf}$ ($t \in B_k$), at each TF bin is categorized into target speech or noise according to the masks, and modeled by a complex Gaussian mixture model defined as

$$p(\boldsymbol{x}_{tf}|\Theta) = \lambda_{tf}^{(v)} \mathcal{N}_c(\boldsymbol{x}_{tf}; \boldsymbol{0}, \mathcal{R}_{k,f}^{(v)}) + \lambda_{tf}^{(n)} \mathcal{N}_c(\boldsymbol{x}_{tf}; \boldsymbol{0}, \mathcal{R}_{k,f}^{(n)}). \tag{7}$$

where $\Theta = \left\{\mathcal{R}_{k,f}^{(v)}, \mathcal{R}_{k,f}^{(n)}\right\}$ is a set of unknown SCMs for the target speech and the noise. Here, $\lambda_{tf}^{(n)} = \sum_{j \neq v} \lambda_{tf}^{(j)}$ is the sum of the masks over all the noise sources.

Next, as a conjugate prior of the complex Gaussian distribution, we introduce a complex inverse Wishart distribution (cIWD) for the prior of the target speech SCM. The cIWD is defined as

$$\mathcal{IW}(\mathcal{R}; \Psi, \nu) := \frac{(\det \Psi)^\nu \exp\left(-\mathrm{tr}(\Psi \mathcal{R}^{-1})\right)}{\pi^{I(I-1)/2} \prod_{i=1}^{I} \Gamma(\nu - I + 1) (\det \mathcal{R})^{\nu+I}} \tag{8}$$

on a set of all Hermitian positive-definite matrices of size $I \times I$. Ψ and ν are hyperparameters of a cIWD, respectively, denoting a scale parameter and a degree-of-freedom parameter. In contrast, considering that the noise is composed of more than one noise source, we adopt cIWMM as the prior of the noise SCM, associating each element distribution with each noise source. Then, the two priors are written, respectively, as

$$p(\mathcal{R}_{k,f}^{(v)}) = \mathcal{IW}(\mathcal{R}_{k,f}^{(v)}; \Psi_f^{(v)}, \nu_f^{(v)}), \tag{9}$$

$$p(\mathcal{R}_{k,f}^{(n)}) = \sum_{j \neq v} \mu_{k,f}^{(j)} \mathcal{IW}(\mathcal{R}_{k,f}^{(n)}; \Psi_f^{(j)}, \nu_f^{(j)}), \tag{10}$$

where $\Psi_f^{(j)}$ and $\nu_f^{(j)}$ are hyperparameters of a cIWD of the jth source, and $\mu_{k,f}^{(j)}$ is the mixture weight of the jth noise source.

The time-varying characteristics of the noise SCM are modeled in eq. (10). In order to model the time-varying contribution of each noise source, we determine the mixture weight, $\mu_{k,f}^{(j)}$, at each short time block B_k as the ratio of the sum of the masks for the jth source to that for all the sources as follows:

$$\mu_{k,f}^{(j)} = \frac{\sum_{t \in B_k} \lambda_{tf}^{(j)}}{\sum_{t \in B_k, j' \neq v} \lambda_{tf}^{(j')}}. \tag{11}$$

On the other hand, to make the estimation reliable, we determine $\Psi_f^{(j)}$ by offline processing, as a time-invariant SCM of the jth source estimated by eq. (1) multiplied with a constant $\nu_f^{(j)} - I$ as

$$\Psi_f^{(j)} = \frac{\nu_f^{(j)} - I}{\sum_t \lambda_{tf}^{(j)}} \sum_t \lambda_{tf}^{(j)} \boldsymbol{x}_{tf} \boldsymbol{x}_{tf}^{\mathrm{H}}. \tag{12}$$

4.2.2. MAP estimation

Letting $\mathcal{X}_{kf} = \{\boldsymbol{x}_{tf} | t \in B_k\}$ be a set of captured signals in block k, the MAP function, $\mathcal{L}(\Theta)$, to be maximized becomes

$$\begin{aligned} \mathcal{L}(\Theta) &= \log p(\mathcal{X}_{kf}|\Theta) + \log p(\Theta) \qquad (13) \\ &= \sum_{f,k} \sum_{t \in B_k} \log \Big[\lambda_{tf}^{(v)} \mathcal{N}_c(\boldsymbol{x}_{tf}; \boldsymbol{0}, \mathcal{R}_{k,f}^{(v)}) \\ &\quad + \lambda_{tf}^{(n)} \mathcal{N}_c(\boldsymbol{x}_{tf}; \boldsymbol{0}, \mathcal{R}_{k,f}^{(n)}) \Big] \\ &\quad + \sum_{f,k} \log \mathcal{IW}(\mathcal{R}_{k,f}^{(v)}; \Psi_f^{(v)}, \nu_f^{(v)}) \\ &\quad + \sum_{f,k} \log \sum_{j \neq v} \mu_{k,f}^{(j)} \mathcal{IW}(\mathcal{R}_{k,f}^{(n)}; \Psi_f^{(j)}, \nu_f^{(j)}). \qquad (14) \end{aligned}$$

Because it is difficult to maximize the MAP function analytically, we instead maximize a minorizer function $Q(\Theta)$ [24] that satisfies

$$\begin{aligned} \mathcal{L}(\Theta) \geq & \sum_{f,k} \sum_{t \in B_k} \Big\{ \lambda_{tf}^{(v)} \log \mathcal{N}_c(\boldsymbol{x}_{tf}; \boldsymbol{0}, \mathcal{R}_{k,f}^{(v)}) \\ & + \lambda_{tf}^{(n)} \log \mathcal{N}_c(\boldsymbol{x}_{tf}; \boldsymbol{0}, \mathcal{R}_{k,f}^{(n)}) \Big\} \\ & + \sum_{f,k} \log \mathcal{IW}(\mathcal{R}_{k,f}^{(v)}; \Psi_f^{(v)}, \nu_f^{(v)}) \\ & + \sum_{f,k} \sum_{j \neq v} \mu_{k,f}^{(j)} \log \mathcal{IW}(\mathcal{R}_{k,f}^{(n)}; \Psi_f^{(j)}, \nu_f^{(j)}) =: Q(\Theta). \end{aligned} \tag{15}$$

By differentiating $Q(\Theta)$ with respect to $\mathcal{R}_{k,f}^{(n)}$ and setting it at zero, we obtain

$$\hat{\mathcal{R}}_{k,f}^{(n)} = \frac{\sum_{t \in B_k} \lambda_{tf}^{(n)} \boldsymbol{x}_{tf} \boldsymbol{x}_{tf}^{\mathrm{H}} + \sum_{j \neq v} \mu_{k,f}^{(j)} \Psi_f^{(j)}}{\sum_{t \in B_k} \lambda_{tf}^{(n)} + \sum_{j \neq v} \mu_{k,f}^{(j)} (\nu_f^{(j)} + I)}. \tag{16}$$

Eq. (16) is the solution of the MAP estimation.

Figure 1 illustrates the processing flow at each frequency f. First, as preprocessing, the time-invariant prior of each noise source is calculated with eq. (12) using a whole utterance, and then the time-varying noise SCM is calculated with eqs. (11) and (16).

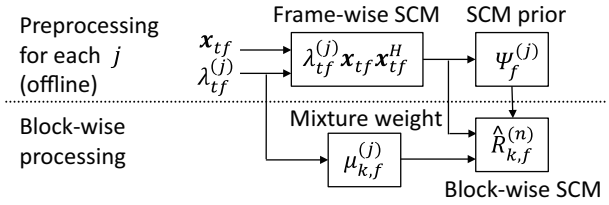

Fig. 1. Processing flow of proposed method.

4.3. Simple alternative formulation

Even when estimated masks contain only one noise class, we can still obtain time-varying noise SCMs as a result of the MAP estimation. This can be a simple alternative formulation of the proposed method. The MAP solution of this formulation becomes

$$\hat{\mathcal{R}}_{k,f}^{(n)} = \frac{\sum_{t \in B_k} \lambda_{tf}^{(n)} \boldsymbol{x}_{tf} \boldsymbol{x}_{tf}^{\mathrm{H}} + \Psi_f^{(n)}}{\sum_{t \in B_k} \lambda_{tf}^{(n)} + \nu_f^{(n)} + I}, \tag{17}$$

where $\Psi_f^{(n)}$ and $\nu_f^{(n)}$ are hyperparameters of the cIWD. In the experiments, we determine $\Psi_f^{(n)}$ using eq. (12) and replacing $\nu_f^{(j)}$ and $\lambda_{tf}^{(j)}$ with $\nu_f^{(n)}$ and $\lambda_{tf}^{(n)}$.

5. EXPERIMENTS

To evaluate the efficacy of the proposed method, in the following we compared its performance with that of conventional beamformers in terms of ASR accuracy and SDR.

5.1. Dataset

A set of simulated sound mixtures were prepared for the experiments. Each mixture contained three simultaneous speech utterances and diffuse babble noise. All three utterances in each mixture were extracted from the Wall Street Journal (WSJ) corpus [25]. One of them was used as a target to be enhanced, and the other two were used as jammers to be reduced. While each mixture included a whole utterance of the target, it contained only beginning parts (2 secs) of the jammers to simulate the short speech overlap that often occurs in real conversation. We convolved each utterance with impulse responses measured using 4 microphones under the recording condition shown in Fig. 2. The target was randomly placed at A or B, and the two jammers were randomly placed at two locations of 1, 2 or 3. The signal-to-noise ratio (SNR) between the target and the babble noise was set at 5 dB, and the SNRs between the target and the jammers were randomly set at -5, 0, or 5 dB. 7138 and 1640 mixtures were prepared as training and development sets, respectively.

5.2. Methods to be compared and evaluation metrics

We evaluated two beamformers for the proposed method, hereafter denoted by TV1 and TV2, that calculate the time-varying noise SCMs, respectively, with eqs. (16) and (17). Short time blocks with sizes of 1, 2, 4, 8, 16, 24, and 32 frames were used for the time-varying estimation. STFT was performed using a 64 msec Hanning window with a 16 msec shift. We compared TV1 and TV2 with a conventional time-invariant MVDR beamformer that calculated time-invariant noise SCMs with eq. (3), and hereafter denoted by TIV. We further compared TV1 and TV2 with the conventional mask-based online MVDR beamformer proposed in [3], hereafter

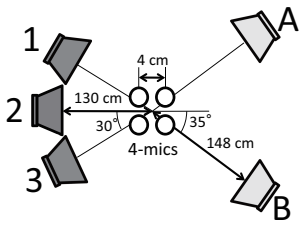

Fig. 2. Recording conditions.

denoted by TVconv. To make the comparison fair, for TVconv, we initialized the noise SCM as that obtained with TIV and estimated the steering vector with eq. (2), while we did not use the SCM model.

We took the WERs and SDRs obtained with TIV as the baseline, and evaluated the word error reduction rates (WERRs) and source-to-distortion ratio improvements (SDRimps) obtained with TV1 and TV2 as the relative improvement from the TIV baseline. For the WER evaluation, we adopted an ASR system developed for CHiME-4 [26], which was composed of a TDNN acoustic model (AM) trained with lattice-free MMI, and a trigram language model [27, 28]. The AM was trained on the above-mentioned training set. The WER and SDR of the development set with no beamforming were 24.8 % and -2.28 dB, respectively.

5.3. Evaluation using oracle masks

We first evaluated the beamformers using oracle masks to exclude the influence of mask estimation errors from the evaluation. The oracle masks were given by the power ratios of each source to the captured signal at each TF bin. Figure 3 shows the WERRs and SDRimps obtained with TV1 and TV2 relative to the TIV baseline (WER: 8.66 %, SDR: 5.49 dB), setting $\nu = 20$ and 40 for the cIWD in eq. (8). The figure clearly shows that the two proposed methods provided significantly improved WERs and SDRs compared with TIV, and that TV1 was consistently better than TV2. Furthermore, the improvement in the WERs and SDRs was larger in most of the cases as the block size became smaller. In particular, TV1 achieved the best WERR when the block size was 1 frame. Note that without the noise SCM prior, the WERRs obtained with TV1 were 7.7, 2.4, and -547.8 % with block sizes of 32, 16, and 1 frames, respectively. This also indicates the significant importance of the noise SCM prior.

We then evaluated the conventional online beamformer, TVconv, using a block size of 4 frames. The best WERR obtained with TVconv was 6.6 % when we set the forgetting factor at 0.1. In comparison, TV1 and TV2 achieved much better WERRs, i.e., 17.9 % and 13.4 %, under the same condition.

5.4. Evaluation using estimated masks

Next, we evaluated the beamformers using estimated masks to take mask estimation errors into account. The masks were estimated with a method that integrates a neural network and a cGMM, denoted as NNcGMM [20]. In the experiments, we found that a permutation problem [29] remains in the masks estimated with NNcGMM, and greatly affects the performance of the beamformers. For comparison, we also evaluated NNcGMM after correcting the permutation alignment of the masks using oracle masks. NNcGMMs with and without the oracle permutation (OP) alignment are denoted by NNcGMM w/ and w/o OP.

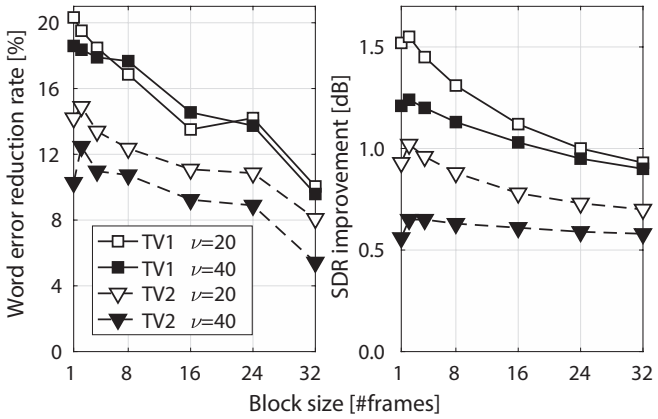

Fig. 3. WERRs and SDRimps obtained with TV1 and TV2 compared with TIV baseline (WER: 8.66 %, SDR: 5.49 dB).

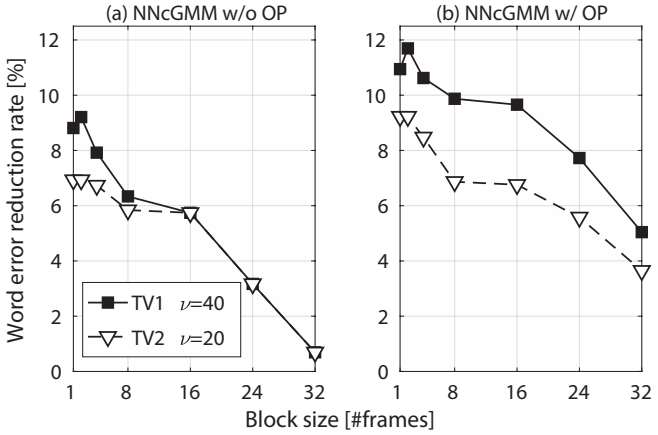

Fig. 4. WERRs obtained with NNcGMMs w/o and w/ OP compared with the TIV baseline (10.1 % w/o OP, 9.32 % w/ OP).

WERRs obtained with TV1 and TV2 relative to the TIV baseline (10.1 % w/o OP, 9.32 % w/ OP) are shown in Fig. 4, where we set $\nu = 40$ for TV1 and 20 for TV2 as the best parameters for the respective methods. Again, TV1 and TV2 were significantly better than TIV. Furthermore, NNcGMM w/ OP was much better than NNcGMM w/o OP. This suggests that accurate permutation alignment is important for the time-varying MVDR beamformer.

6. CONCLUDING REMARKS

This paper proposed a method for designing a mask-based time-varying MVDR beamformer for speech enhancement in noisy multi-speaker environments. The proposed method models the prior distribution of a time-varying noise SCM with a cIWMM, and estimates the SCM based on the MAP estimation. We conducted simulation experiments using oracle masks and estimated masks, and the proposed method achieved significantly better WERs and SDRs than two conventional mask-based MVDR beamformers, namely a time-invariant beamformer and an online beamformer. Furthermore, we confirmed that the performance of the proposed method improved in most cases as we reduced the size of a short time block. This confirms that the proposed method can reliably estimate the noise SCM with a high time resolution of the order of a few tens of milliseconds.

7. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] S. Araki, M. Okada, T. Higuchi, A. Ogawa, and T. Nakatani, "Spatial correlation model based observation vector clustering and MVDR beamforming for meeting recognition," *Proc. ICASSP 2016*, pp. 385–389, 2016.

[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," *Proc. IEEE ASRU-2015*, pp. 436–443, 2015.

[11] 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.

[12] 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.

[13] J. Azcarreta, N. Ito, S. Araki, and T. Nakatani, "Permutation-free CGMM: complex Gaussian mixture model with inverse Wishart mixture model based spatial prior for permutation-free source separation and source counting," *Proc. IEEE ICASSP-2018*, pp. 51–55, 2018.

[14] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[15] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, 2011.

[16] 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.

[17] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *Proc. IEEE ICASSP 2017*, pp. 31–35, 2016.

[18] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE Trans. ASLP*, pp. 1901–1913, 2017.

[19] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," *Proc. ICASSP 2018*, pp. 5064–5068, 2018.

[20] 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.

[21] 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.

[22] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," *Proc. IEEE ICASSP-2017*, pp. 681–685, 2017.

[23] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," *Proc. IEEE ICASSP 2017*, pp. 544–548, 2015.

[24] K. Lange, *MM Optimization Algorithms*, Society for Industrial and Applied Mathematics, 2016.

[25] 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.

[26] 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 and Language*, vol. 46, pp. 535–557, 2017.

[27] "CHiME4 advanced baseline," `https://github.com/kaldi-asr/kaldi/blob/master/egs/chime4/s5_1ch`.

[28] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," *arXiv: 1803.10109*, 2018.

[29] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.