Time-Frequency-Bin-Wise Beamformer Selection and Masking for Speech Enhancement in Underdetermined Noisy Scenarios

Kouei Yamaoka*, Andreas Brendel†, Nobutaka Ono‡, Shoji Makino*,
Michael Buerger†, Takeshi Yamada*, and Walter Kellermann†

*University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki, 305-8577 Japan,
{yamaoka@mmlab.cs, maki@tara, takeshi@cs}.tsukuba.ac.jp
†Friedrich-Alexander University Erlangen-Nürnberg, 91058 Erlangen, Germany,
{andreas.brendel, michael.buerger, walter.kellermann}@FAU.de
‡Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo, 191-0065 Japan, onono@tmu.ac.jp

***Abstract*—In this paper, we present a speech enhancement method using two microphones for underdetermined situations. A conventional speech enhancement method for underdetermined situations is time-frequency masking, where speech is enhanced by multiplying zero or one to each time-frequency component appropriately. Extending this method, we switch multiple preconstructed beamformers at each time-frequency bin, each of which suppresses a particular interferer. This method can suppress an interferer even when both the target and an interferer are simultaneously active at a given time-frequency bin. As a switching criterion, selection of minimum value of the outputs of the all beamformers at each time-frequency bin is investigated. Additionally, another method using direction of arrival estimation is also investigated. In experiments, we confirmed that the proposed methods were superior to conventional time-frequency masking and fixed beamforming in the performance of speech enhancement.**

I. Introduction

Beamforming and blind source separation (BSS) [1] are commonly used in speech enhancement and can yield a good performance as long as a sufficient number of microphones are available. Automatic speech recognition can be improved by applying these methods (e.g., [2]). However, the capability of these microphone array methods to suppress multiple interferers depends on the number of microphones M. If there are N sound sources consisting of one target and $(N-1)$ interferers, we need the same number of microphones ($M = N$) to suppress all interferers by null steering. However, commonly-used small recording devices such as voice recorders often have only two microphones. Although several conventional methods such as time-frequency masking [3], [4], multichannel Wiener filtering [5], and the statistical modeling of observations using latent variables [6] can work well in underdetermined situations ($M < N$), they face a tradeoff between a low signal distortion and high noise reduction performance.

In this paper, we propose a new method of speech enhancement realizing a high performance by selecting one out of multiple pretrained beamformers as an extension of conventional speech enhancement based on time-frequency masking. If M microphones are available, a single beamformer can generally form $(M-1)$ nulls. This means that a single beamformer can suppress only one interferer in the two microphone case ($M = 2$). However, if we can construct $(N-1)$ beamformers, each suppressing one of the $(N-1)$ interferers, we can improve the speech enhancement performance by using a combination of these beamformers rather than a single beamformer.

In [7], the combination of multiple beamformers with different steering directions for audio zooming was considered. However, in this study, we combine multiple beamformers with the same steering direction (the same target) but different null directions. Speech enhancement by Wiener filtering and frequency-bin-wise combination of multiple fixed null beamformers using a square microphone array was proposed in [8], [9]. However, this method tends to generate target signal distortions. The reduction of mechanical noise, such as the sound of actuators and motors in a robot, by selecting the most suitable noise covariance matrix at each time-frequency bin for the computation of maximum signal-to-noise ratio (MaxSNR) beamformers has also been proposed [10]. This method requires the clustering of multichannel mechanical noise covariance matrices in a training phase under the assumption that the number of the patterns of the actuator is usually limited. In contrast to the methods presented above, we switch multiple signal-dependent beamformers at each time-frequency bin for underdetermined speech enhancement with a low amount of target distortions such as minimum variance distortionless response (MVDR) beamformers [11], [12].

In this paper, we propose two criteria for selecting a suitable beamformer from a set of previously trained beamformers, namely, minimum value selection (MIN), that is, the choice of the beamformer corresponding to the minimum absolute output among the outputs of these multiple beamformers, and its extension based on the direction of arrival (DOA) estimation. If a time-frequency bin is occupied by either the target and one of the interferers or by a single source, i.e., it fulfills the W-disjoint orthogonality (W-DO), MIN can suppress the interferer. The combination of MIN and DOA estimation can also suppress noise in a time-frequency bin when no target is present. We evaluate the performance of

ISBN 978-90-827970-1-5 © EURASIP 2018

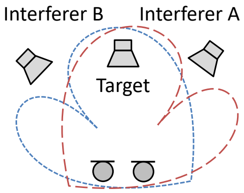


Fig. 1: Combination of two beamformers with a spatial null for each interferer

each proposed method and confirm its effectiveness.

II. Conventional linear beamforming

We model the microphone signals in the short-time Fourier transform (STFT) domain. Here, let $x_i(\omega, t)$ be the ith microphone signal at the angular frequency ω in the tth time frame. When two microphones observe one target and one interferer, we can perform conventional speech enhancement using, e.g., a MaxSNR beamformer [11], [13] or an MVDR beamformer [11], [12], which steers a spatial null in the direction of the interferer, as described by the following equations:

$$\begin{aligned} y(\omega,t) &= \mathbf{w}^{\mathrm{H}}(\omega)\mathbf{x}(\omega,t), && (1)\\ \mathbf{x}(\omega,t) &= \left[x_1(\omega,t), x_2(\omega,t)\right]^{\mathrm{T}}, && (2)\\ \mathbf{w}(\omega) &= \left[w_1(\omega), w_2(\omega)\right]^{\mathrm{T}}, && (3) \end{aligned}$$

where $y(\omega, t)$ is the output signal of the beamformer, $\mathbf{w}(\omega)$ denotes the spatial filter vector, $(\cdot)^{\mathrm{T}}$ denotes the transpose, and $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose. However, only $(M-1)$ sources can be suppressed using M microphones. Hence, the performance of linear speech enhancement may be degraded under an underdetermined situation with $M < N$, i.e., when we have fewer microphones than sound sources N.

III. Proposed speech enhancement method based on selection from multiple beamformers

Without loss of generality, we consider a situation with $M = 2$ microphones and $N = 3$ sound sources consisting of a target signal, interferer signal A, and interferer signal B (hereafter referred to as tgt, $\mathrm{i_A}$, and $\mathrm{i_B}$, respectively). In this situation, we cannot construct a null beamformer that suppresses both interferers. However, if only tgt and $\mathrm{i_A}$ are observed, we can construct beamformer A that suppresses $\mathrm{i_A}$ using a conventional beamforming method, and analogously beamformer B that suppresses $\mathrm{i_B}$ (see Fig. 1). Then, we obtain the following two output signals y_{A} and y_{B} from an observation $\mathbf{x}$ consisting of tgt, $\mathrm{i_A}$, and $\mathrm{i_B}$:

$$\begin{aligned} y_{\mathrm{A}}(\omega,t) &= \mathbf{w}_{\mathrm{A}}^{\mathrm{H}}(\omega)\mathbf{x}(\omega,t), && (4)\\ y_{\mathrm{B}}(\omega,t) &= \mathbf{w}_{\mathrm{B}}^{\mathrm{H}}(\omega)\mathbf{x}(\omega,t), && (5) \end{aligned}$$

where $\mathbf{w}_{\mathrm{A}}$ and $\mathbf{w}_{\mathrm{B}}$ are the spatial filters defining the beamformers A and B, respectively.

TABLE I: Combinations of the input signals and corresponding output signals

$\mathbf{x}$	y_{A}	y_{B}	y_{MIN}	y_{DOA}
tgt	tgt	tgt	tgt	tgt
$\mathrm{i_A}$	0^{*}	$\mathrm{i_A}$	0^{*}	0^{**}
$\mathrm{i_B}$	$\mathrm{i_B}$	0^{*}	0^{*}	0^{**}
tgt, $\mathrm{i_A}$	tgt	tgt, $\mathrm{i_A}$	tgt	tgt
tgt, $\mathrm{i_B}$	tgt, $\mathrm{i_B}$	tgt	tgt	tgt
$\mathrm{i_A}$, $\mathrm{i_B}$	$\mathrm{i_B}$	$\mathrm{i_A}$	$\mathrm{i_A}$ or $\mathrm{i_B}$	0^{**}
tgt, $\mathrm{i_A}$, $\mathrm{i_B}$	tgt, $\mathrm{i_B}$	tgt, $\mathrm{i_A}$	tgt, $\mathrm{i_A}$ or tgt, $\mathrm{i_B}$	tgt, $\mathrm{i_A}$ or tgt, $\mathrm{i_B}$

* suppressed ** masked

For every time-frequency bin of the observed signal $\mathbf{x}$, the dominant sound comprises the seven patterns listed in Table I with the corresponding beamformer outputs y_{A} and y_{B}, where the trivial case with no active sources is not considered here. If only tgt is dominant in $\mathbf{x}(\omega, t)$, both beamformers output tgt (see the second row of Table I). If only $\mathrm{i_A}$ is dominant, beamformer A suppresses the input signal and beamformer B outputs a slightly altered version of $\mathrm{i_A}$. At time-frequency bins consisting of tgt and $\mathrm{i_B}$, beamformer B outputs tgt only, whereas beamformer A outputs both tgt and $\mathrm{i_B}$.

In our proposed method, we perform speech enhancement by choosing the output from y_{A} and y_{B}, which means that we choose the best result from multiple beamformers in each time-frequency bin. Here, the important issue is how to select the appropriate beamformer. Under the assumption of W-DO, either tgt, $\mathrm{i_A}$, or $\mathrm{i_B}$ is active in a time-frequency bin. Then, the requirement for the selection strategy is that the output should be tgt if it is dominant, otherwise zero as an interferer is dominant in this case. In this paper, we propose two methods of selecting beamformers satisfying this requirement, which we call MIN and its extension DOA.

A. Minimum value selection of multiple beamformers

The magnitude of the output of beamformer A (B) is smaller than that of beamformer B (A) when $\mathrm{i_A}$ ($\mathrm{i_B}$) is dominant in the respective time-frequency bin. Thus, the following MIN selection criterion satisfies the requirement described above:

$$y_{\mathrm{MIN}}(\omega,t) = \begin{cases} y_{\mathrm{A}}(\omega,t) & \text{if } \ |y_{\mathrm{A}}(\omega,t)| \le |y_{\mathrm{B}}(\omega,t)|, \\ y_{\mathrm{B}}(\omega,t) & \text{otherwise.} \end{cases} \qquad (6)$$

This formula implies that MIN selects the output of the beamformer with the smaller magnitude. The rationale behind MIN is that it suppresses the interferer signal in a time-frequency bin consisting of tgt and one interferer signal under the assumption that the magnitude of tgt is smaller than the magnitude of the sum of tgt and the interferer signal, which is a valid assumption when considering statistically independent target and interferer signals. However, if a time-frequency bin contains $\mathrm{i_A}$ and $\mathrm{i_B}$, MIN will select $\mathrm{i_B}$ or $\mathrm{i_A}$ (see the seventh row of Table I). Thus, although the magnitude of the output of MIN is smaller than $\mathrm{i_A}$ and $\mathrm{i_B}$, one of the interferers remains.

MIN is similar to time-frequency masking, in which a soft or hard mask is applied to time-frequency bins dominated by interferers. Thus, masking methods require W-DO in general. In contrast, MIN selects the beamformer that suppresses an interferer in each time-frequency bin, i.e., it can also suppress noise in a time-frequency bin that contains both the target and one interferer. Therefore, the proposed method is applicable even if the assumption of W-DO is not satisfied. From the above, it can be concluded that the MIN-based selection is an extension of time-frequency masking overcoming the limitation of requiring W-DO.

ISBN 978-90-827970-1-5 © EURASIP 2018

B. Extension of minimum value selection by DOA estimation

Speech enhancement by MIN, which is a simple selection rule, is expected to show high speech enhancement performance. However, when a time-frequency bin contains multiple noise signals, it is not possible to suppress all of them. If there is no target in such a bin, the respective signal component should be suppressed completely similarly to time-frequency masking (see the last column of Table I), i.e.,

$$y_{\mathrm{DOA}}(\omega,t) = M(\omega,t) y_{\mathrm{MIN}}(\omega,t), \tag{7}$$

where $M(\omega,t)$ is a soft mask. To construct $M(\omega,t)$, we estimate the activity of the sound sources by evaluating a trained probabilistic model that describes the DOAs of the sound sources.

In this paper, we use a one-microphone-pair version of [14] for this task, which has been used as a baseline algorithm for the DOA estimation methods in [15] and [16]. Here, the relative phase ratios (RPRs) $\phi(\omega,t)$ of the observed microphone signals defined as

$$\phi(\omega,t) \ := \ \frac{x_2(\omega,t)}{x_1(\omega,t)} \cdot \frac{|x_1(\omega,t)|}{|x_2(\omega,t)|}, \tag{8}$$

are clustered by a Gaussian mixture model (GMM). The mean μ_k $(k=1,\ldots,K)$ of each complex Gaussian distribution $\mathcal{N}^c$ is an expected RPR associated with a DOA from a predefined grid, where K is the number of clusters/grid points

$$\mathcal{N}^c(\phi(\omega,t);\mu_k,\sigma^2) = \frac{1}{\pi\sigma^2}\exp\left\{-\frac{|\phi(\omega,t)-\mu_k|^2}{\sigma^2}\right\}, \tag{9}$$

where σ^2 is the variance common to all Gaussian components, which can be user-defined without performance loss [17]. Then, the mean of the Gaussian component with the maximum likelihood given the observation $\phi(\omega,t)$ is chosen as the time-frequency-bin-wise DOA estimate

$$\mathrm{DOA}^{\mathrm{L}}(\omega,t) \ = \ \underset{\mu_k}{\operatorname{argmax}} \ \mathcal{N}^c(\phi(\omega,t);\mu_k,\sigma^2). \tag{10}$$

Hereafter, we abbreviate the local, i.e., time-frequency-bin-wise DOA, as $\mathrm{DOA}^{\mathrm{L}}$, and the global, i.e., time-frame-wise DOA as $\mathrm{DOA}^{\mathrm{G}}$. Note that $\mathrm{DOA}^{\mathrm{L}}(\omega,t)$ is a RPR associated with the DOA corresponding to the time-frequency bin (ω,t). Then, a soft mask $M^{\mathrm{L}}(\omega,t)$ based on the local DOA estimates is computed as

$$M^{\mathrm{L}}(\omega,t) \ = \ \frac{\mathcal{N}^c(\phi(\omega,t);\mu_{k=\mathrm{target}},\sigma^2)}{\sum_{k=1}^{K}\mathcal{N}^c(\phi(\omega,t);\mu_k,\sigma^2)}, \tag{11}$$

where $\mu_{k=\mathrm{target}}$ is the mean corresponding to the target direction. This procedure, based on bin-wise estimates, is justified in time-frequency bins satisfying W-DO.

Next, source activity estimation (SAE) of each source is performed per time frame by averaging the local DOA estimates as follows:

$$\mathrm{SAE}_k(t) \ = \ \frac{1}{C}\sum_{\omega=1}^{C}\eta_k^{\mathrm{L}}(\omega,t), \tag{12}$$

$$\eta_k^{\mathrm{L}}(\omega,t) \ = \ \begin{cases} 1 & \text{if } \mathrm{DOA}^{\mathrm{L}}(\omega,t)=\mu_k, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

$$\boldsymbol{\eta}^{\mathrm{L}}(\omega,t) \ = \ [\eta_1^{\mathrm{L}}(\omega,t),\cdots,\eta_K^{\mathrm{L}}(\omega,t)]^{\mathrm{T}}, \tag{14}$$

where C is the number of frequency bins. Here, $\eta_k^{\mathrm{L}}(\omega,t)$ is a Boolean variable that indicates whether the sound source active in time-frequency bin (ω,t) belongs to cluster k. $\mathrm{SAE}_k(t)$ is smoothed by applying linear weighted moving average (LWMA) as

$$\mathrm{SAE}_k(t) \ \leftarrow \ \sum_{i=-T}^{T}\frac{(T+1-|i|)\mathrm{SAE}_k(t+i)}{T+1-|i|}, \tag{15}$$

where T is the number of time frames taken into account for LWMA. Then, the global DOA estimation, i.e., the estimation of a DOA associated with a time frame, is performed by applying a fixed threshold to the SAE estimates.

$$\eta_k^{\mathrm{G}}(t) \ = \ \begin{cases} 1 & \text{if } \mathrm{SAE}_k(t) > \text{threshold}, \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

$$M^{\mathrm{G}}(t) \ = \ \frac{\mathrm{SAE}_{k=\mathrm{target}}(t)}{\sum_{k=1}^{K}\mathrm{SAE}_k(t)}, \tag{17}$$

where $\eta_k^{\mathrm{G}}(t)$ and the soft mask $M(t)^{\mathrm{G}}$ are global counterparts of $\eta_k^{\mathrm{L}}(\omega,t)$ and $M^{\mathrm{L}}(\omega,t)$, respectively. The global DOA estimation works even if some time-frequency bins do not satisfy W-DO owing to the averaging of local DOA estimates.

Finally, the local DOA estimate is chosen if W-DO is fulfilled and the global DOA estimate otherwise, i.e.,

$$\eta_k(\omega,t) = \begin{cases} \eta_k^{\mathrm{L}}(\omega,t) & \text{if } \left\langle \boldsymbol{\eta}^{\mathrm{L}}(\omega,t), \boldsymbol{\eta}^{\mathrm{G}}(t)\right\rangle = 1, \\ \eta_k^{\mathrm{G}}(t) & \text{otherwise,} \end{cases} \tag{18}$$

where $\langle\cdot,\cdot\rangle$ is the standard inner product operator. Then, the soft mask M in (7) is constructed as

$$M(\omega,t) = \begin{cases} 1 & \text{if } \eta_{k=\mathrm{target}}(\omega,t)=1, \\ M^{\mathrm{L}}(\omega,t) & \text{if } \left\langle \boldsymbol{\eta}^{\mathrm{L}}(\omega,t), \boldsymbol{\eta}^{\mathrm{G}}(t)\right\rangle = 1, \\ M^{\mathrm{G}}(t) & \text{otherwise.} \end{cases} \tag{19}$$

IV. Experimental evaluation

A. Experimental conditions

To evaluate the effectiveness of our proposed method, we conducted experiments using the "Underdetermined-speech and music mixtures" test data from the community-based Signal Separation Evaluation Campaign (SiSEC) [18]. These data contain two mixtures with three male or female speakers. We conduct speech enhancement experiments for each speaker as a target sound. The experimental conditions are listed in Table II.

We used two signal-dependent beamformers to produce time-invariant filters: a MaxSNR beamformer [11], [13] and an MVDR beamformer [11], [12]. We gave the same target-active period and noise-active period as prior information for both beamformers to evaluate the beamformers on the same basis. Using these periods, we calculated the interference-plus-noise covariance matrix for the computation of both the beamformer filters and the target covariance matrix for the MaxSNR beamformer. Besides, we performed the eigenvalue decomposition for the spatial correlation matrix of the target-active period and used the eigenvector corresponding to the maximum eigenvalue as the estimate of the transfer function,

TABLE II: Experimental conditions

Number of microphones	2
Distance between microphones	5 cm
DOAs	40°, 80°, and 105°
Reverberation time	250 ms
Sampling rate	16 kHz
FFT frame length / shift	2048 / 512 samples
Training period	5 s
Test period	35 s (5 s × 7)

which is assumed to be prior information for the MVDR beamformer.

We evaluate the performance by comparing the results of three conventional speech enhancement methods and four algorithmic variants of the proposed method. As conventional methods, we evaluate MaxSNR_SOL and MVDR_SOL, which are the underdetermined speech enhancement methods using a single MaxSNR and an MVDR beamformer, respectively. We also investigate the performance of the degenerate unmixing estimation technique (DUET) [19] as a conventional method of time-frequency binary masking with a stereo microphone. We discuss the combinations of a MaxSNR and an MVDR beamformer with the MIN and DOA selection strategy and abbreviate these algorithmic variants as MaxSNR_MIN, MVDR_MIN, MaxSNR_DOA, and MVDR_DOA, respectively. We therefore require a target-active period and two noise-active periods (for $\mathrm{i_A}$ and $\mathrm{i_B}$ separately) for the computation of the beamformer filters.

For the DOA estimation, the variance was set to 10 for all Gaussians. Here, only the frequency bins corresponding to 1–4 kHz were considered as they contain most of the signals energy. For the SAE, we used nine time frames (384 ms) for smoothing by LWMA.

To validate the effectiveness of our proposed method, we investigate the performance of the considered algorithms in all seven combinations of the source signals, as shown in Table I. Here, all observed signal mixtures can be assumed to be sparse (and therefore W-DO is satisfied in most time-frequency bins) because the target and noise signals are speech. We use objective criteria, namely, the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR) [20] to quantify the results. A concise representation of the results is obtained by averaging these criteria over speakers for the period consisting of all speakers. Here, the reference signal is the source image, i.e., the noise-free reverberant speech signal.

B. Results and discussion

The results of speech enhancement using the algorithms discussed above are shown in Fig. 2 for the test period that includes three speakers. The conventional single beamformers can suppress only one interferer but generate a low amount of artificial noise. Therefore, they only show good performance for SAR. The proposed method shows a high performance in SDR and SIR, and is superior to DUET as well as to MaxSNR_SOL and MVDR_SOL. Moreover, the extension of MIN with DOA estimation results in a significant improvement of the SIR. Considering these results, it can be concluded that our proposed methods, particularly MVDR_MIN and MVDR_DOA, improve the speech enhancement performance.

The fact that the best performance is achieved by the MVDR beamformer in combination with MIN or its extension can be understood by investigating the properties of MIN and the two beamforming methods. MaxSNR only maximizes SNR and imposes no constraint on the target direction. Thus, the magnitude or phase of the target in y_A and y_B may be different, which is a serious problem when choosing different outputs in neighboring time-frequency bins. We show the selected beamformers using MVDR_MIN in Fig. 3(a) examplary. According to this figure, the selected beamformer switches frequently in the time-frequency plane. Hence, if the magnitude or phase of the target in the two outputs is different, it causes distortions of the target if different beamformer outputs are chosen in time-frequency bins close to each other, which is an effect similar to that occurring in time-frequency masking. On the other hand, since the MVDR beamformer outputs the target source without distortion, both the magnitude and the phase match. Thus, no distortion due to the switching of the beamformers occurs. This explains the higher SDR of MVDR_MIN (and _DOA) than of MaxSNR_MIN. Considering only the SIR performance, the MaxSNR beamformer, which does not impose a constraint on the target direction, may be superior to the MVDR beamformer. However, it results in a greater distortion. Thus, it can be concluded that MVDR_MIN and its extension realize a high performance with a low distortion of the target signal in underdetermined situations.

Now, we discuss the extension of MIN by DOA estimation. Figure 3(b) shows the selected beamformers and masked time-frequency bins. The main benefit of DOA estimation is the improvement of SIR by using a soft mask (see the red time-frequency bins in the figure). Although the SAR performance decreases because the masking generates artificial noise, the SDR performance is almost the same as that for MVDR_MIN. Thus, we can conclude that the extension with DOA estimation is effective for improving noise reduction performance.

V. Conclusions

In this paper, we have proposed a new method of speech enhancement for underdetermined scenarios and performed a performance evaluation for the 2-microphone case. This method achieves high speech enhancement performance by selection from multiple preconstructed beamformers and time-frequency masking. We proposed two methods for selecting the pretrained beamformers: MIN, which can suppress noise in the presence of one target and one interferer in each time-frequency bin, and its extension based on speaker activity detection from the spatial information of the sources. Both methods can be considered as extensions of time-frequency masking.

We demonstrated the effectiveness of the proposed method by performing experiments in a reverberant environment containing one target and two interferers. The MVDR beamformer combined with MIN and DOA estimation showed the highest performance with approximately 3 dB improvement in SDR and 7 dB improvement in SIR. This means that our proposed method is superior to the conventional methods we used for

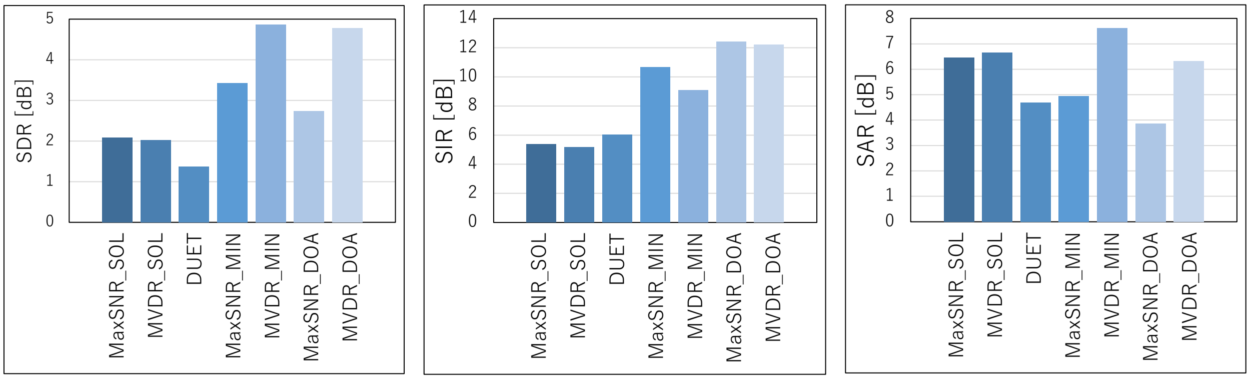
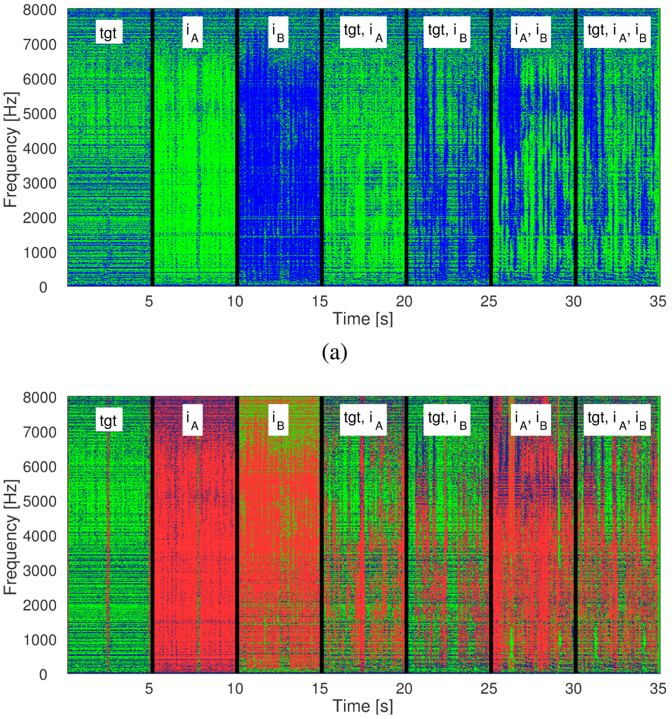


Fig. 2: Results of speech enhancement for the test period consisting of three speakers



(a)

(b)

Fig. 3: Selected beamformers in (a) MVDR_MIN and (b) MVDR_DOA for each test period. Green, beamformer A; blue, beamformer B; red, masked by (7)

benchmarking our results in terms of the speech enhancement performance.

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science (JSPS) under Grant 16H01735, SECOM Science and Technology Foundation, and the Tsukuba-DAAD Joint Research Program.

References

[1] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, Springer, 2007.

[2] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 4, pp. 780–793, Apr. 2017.

[3] O. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, 2010.

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

[6] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," *Proc. WASPAA*, pp. 147–150, Oct. 2007.

[7] N. Q. K. Duong, P. Berthet, S. Zabre, M. Kerdranvat, A. Ozerov, and L. Chevallier, "Audio zoom for smartphones based on multiple adaptive beamformers," *Proc. LVA/ICA*, pp. 121–130, Feb. 2017.

[8] S. Takada, S. Kanba, T. Ogawa, K. Akagiri, and T. Kobayashi, "Sound source separation using null-beamforming and spectral subtraction for mobile devices," *Proc. WASPAA*, pp. 30–33, 2007.

[9] T. Ogawa, S. Takada, K. Akagiri, and T. Kobayashi, "Speech enhancement using a square microphone array in the presence of directional and diffuse noise," *IEICE Trans. Fundamentals*, vol. E93-EA, no. 5, pp. 926–935, May 2010.

[10] M. Togami, T. Sumiyoshi, Y. Obuchi, Y. Kawaguchi, and H. Kokubo, "Beamforming array technique with clustered multichannel noise covariance matrix for mechanical noise reduction," *Proc. EUSIPCO*, Aug. 2010.

[11] H. L. Van Trees, *Optimum Array Processing*, John Wiley & Sons, 2002.

[12] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[13] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," *Proc. ICASSP*, vol. I, pp. 41–45, Apr. 2007.

[14] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 392–402, Feb. 2014.

[15] O. Schwartz, Y. Dorfan, M. Taseska, E. A. P. Habets, and S. Gannot, "DOA estimation in noisy environment with unknown noise power using EM algorithm," *Proc. HSCMA*, pp. 86–90, Mar. 2017.

[16] O. Schwartz, Y. Dorfan, E. A. P. Habets, and S. Gannot, "Multi-speaker DOA estimation in reverberation conditions using expectation-maximization," *Proc. IWAENC*, pp. 1–5, Sep. 2016.

[17] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1692–1703, Oct. 2015.

[18] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," *Proc. ICA'09*, 2009.

[19] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer, pp. 217–241, 2007.

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