# TIME-FREQUENCY-BIN-WISE SWITCHING OF MINIMUM VARIANCE DISTORTIONLESS RESPONSE BEAMFORMER FOR UNDERDETERMINED SITUATIONS

*Kouei Yamaoka*[1]*, Nobutaka Ono*[2]*, Shoji Makino*[1]*, and Takeshi Yamada*[1]

[1]University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8577, Japan
E-mail: {yamaoka@mmlab.cs, maki@tara, takeshi@cs}.tsukuba.ac.jp
[2]Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan
E-mail: onono@tmu.ac.jp

## ABSTRACT

In this paper, we present a speech enhancement method using two microphones in underdetermined situations. Time-frequency (TF) binary masking is a conventional method of enhancing speech in underdetermined situations by appropriately multiplying each TF component by zero or one. Extending this method, we previously proposed a new method called the time-frequency-bin-wise switching (TFS) beamformer. In this method, we switch multiple preconstructed beamformers in each TF bin, each of which suppresses a particular interferer. However, this method requires the pre-estimation of beamformer filter coefficients using the target-active period and interferer-wise-active periods as the prior information. In this paper, to overcome this limitation, we formulate the switching and construction of spatial filters as a joint optimization problem, which can be understood from two viewpoints: the clustering of the most dominant interferer signal in each TF bin and the construction of a minimum variance distortionless response beamformer using such bins. In an experiment, we confirmed that the proposed method was superior to conventional TF masking and fixed beamforming during speech enhancement regardless of the direction of interferers.

***Index Terms—*** beamforming, time-frequency masking, speech enhancement, underdetermined situation, nonlinear signal processing

## 1. INTRODUCTION

Beamforming and blind source separation [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 a 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 and smartphones often have only two microphones. Although several conventional methods such as time-frequency (TF) 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 low signal distortion and high noise reduction performance. Therefore, the purpose of this study is to develop a new method of underdetermined speech enhancement realizing high performance with low signal distortion.

We have proposed the time-frequency-bin-wise switching (TFS) beamformer [7] as an extension of conventional speech enhancement based on TF binary masking, which uses multiple preconstructed beamformers. 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 a two-microphone case. 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. On the basis of this idea, this method enhances speech by multiplying by the best beamformer filter to suppress interferers in each TF bin rather than multiplying by a scalar as in TF masking.

In [8], 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 the frequency-bin-wise combination of multiple fixed null beamformers using a square microphone array was proposed in [9, 10]. However, this method tends to distort the target signal. The reduction of mechanical noise, such as the sound of actuators in a robot, by selecting the most suitable noise covariance matrix in each TF bin to compute maximum signal-to-noise ratio (MaxSNR) beamformers has also been proposed [11]. This method requires the clustering of multichannel mechanical noise covariance matrices in a training phase under the assumption that the number of actuator patterns is usually limited. In contrast to the methods presented above, we switch multiple signal-dependent beamformers, such as minimum variance distortionless response (MVDR) beamformers [12, 13], in each TF bin for underdetermined speech enhancement without target distortion.

In this paper, we propose an MVDR beamformer-based TFS beamformer that requires the same prior information as a conventional MVDR beamformer, i.e., the relative transfer function (RTF) of the target source, whereas we previously proposed a TFS beamformer [7] that had the limitations of requiring the target-active period and interferer-wise-active periods. In our proposed method, we formulate the selection of the best beamformer and the construction of MVDR filters

This work was supported by JSPS under Grant 16H01735, and SECOM Science and Technology Foundation.

as a joint optimization problem. This method can suppress $N-1$ interferers in the TF plane in underdetermined situations by assuming that there are $M-1$ interferers in a TF bin rather than W-disjoint orthogonality (W-DO) [3, 14].

## 2. CONVENTIONAL MVDR BEAMFORMER

We model the microphone signals in the short-time Fourier transform (STFT) domain. Here, let $x_i(\omega,t)$ be the $i$th microphone signal at the angular frequency $\omega$ in the $t$th time frame. When $M$ microphones observe one target and $N-1$ interferers in determined situations ($M=N$), we can perform conventional speech enhancement using an MVDR beamformer [12, 13], which steers a spatial null in the direction of an interferer, as described by

$$y(\omega,t) = \boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t), \tag{1}$$

$$\boldsymbol{x}(\omega,t) = [x_1(\omega,t)\cdots x_M(\omega,t)]^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{w}(\omega) = [w_1(\omega)\cdots w_M(\omega)]^{\mathrm{T}}, \tag{3}$$

where $y(\omega,t)$ is the output signal of the beamformer, $\boldsymbol{w}(\omega)$ denotes the spatial filter vector, $(\cdot)^{\mathrm{T}}$ denotes the transpose, and $(\cdot)^{\mathrm{H}}$ denotes the Hermitian transpose. The filter $\boldsymbol{w}(\omega)$ is constructed by solving the following optimization problem:

$$\min_{\boldsymbol{w}} \sum_{\omega} \mathbb{E}\left[|\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t)|^2\right] \quad \text{s.t.} \quad \boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{a}(\omega) = 1, \tag{4}$$

where $\mathbb{E}[\cdot]$ is the expectation operator and $\boldsymbol{a}(\omega)$ is the RTF of target, which is defined as the ratio of the acoustic transfer functions $\boldsymbol{h}(\omega) = [h_1(\omega)\cdots h_M(\omega)]^{\mathrm{T}}$ from the target source to the microphone array.

$$\boldsymbol{a}(\omega) = \left[1 \;\; \frac{h_2(\omega)}{h_1(\omega)} \;\; \cdots \;\; \frac{h_M(\omega)}{h_1(\omega)}\right]^{\mathrm{T}} \tag{5}$$

The cost function $\mathcal{J}_{\mathrm{c}}$ is

$$\mathcal{J}_{\mathrm{c}} = \sum_{\omega}\left\{\mathbb{E}\left[|\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t)|^2\right] + 2\mathrm{Re}[\lambda^*(\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{a}(\omega)-1)]\right\}, \tag{6}$$

where $\mathrm{Re}[\cdot]$ takes the real part of the input argument, and $\lambda^*$ is the complex-valued Lagrange multiplier in the method of the Lagrange multiplier. The closed-form solution is

$$\boldsymbol{w}(\omega) = \frac{\boldsymbol{\Phi}(\omega)^{-1}\boldsymbol{a}(\omega)}{\boldsymbol{a}^{\mathrm{H}}(\omega)\boldsymbol{\Phi}(\omega)^{-1}\boldsymbol{a}(\omega)}, \tag{7}$$

$$\boldsymbol{\Phi}(\omega) = \mathbb{E}[\boldsymbol{x}(\omega,t)\boldsymbol{x}(\omega,t)^{\mathrm{H}}]. \tag{8}$$

The MVDR beamformer can enhance the target signal with a distortionless response. However, only $M-1$ interferers can be suppressed, and thus the performance may be degraded in an underdetermined situation with $M<N$. Here, note that when $\boldsymbol{x}$ is composed of only interferers, the beamforming is called MVDR beamformer, and when it includes the target, the beamforming is called the minimum power distortionless response (MPDR) beamformer. The MVDR beamformers that we used for the proposed method in the following can be replaced by MPDR beamformers.

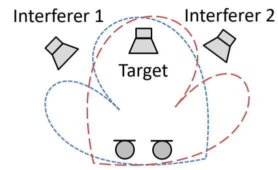



**Fig. 1**: Combination of two beamformers with a spatial null for each interferer in a situation with $M=2$ and $N=3$

## 3. TFS OF MVDR BEAMFORMERS

### 3.1. Conventional TFS beamformers

Without loss of generality, we consider a situation with $M=2$ microphones and $N=3$ sound sources consisting of a target signal and interferer signals 1 and 2. In this situation, we cannot construct a null beamformer that suppresses both interferers. Here, if only the target and interferer $k$ are observed ($k=1,2$), we can construct the beamformer $k$ with a spatial filter $\boldsymbol{w}_k$ that suppresses only the interferer $k$ (see Fig. 1). Then, we obtain the following two output signals $y_k(\omega,t)$ from the observation $\boldsymbol{x}(\omega,t)$ consisting of three sources:

$$y_k(\omega,t) = \boldsymbol{w}_k^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t). \tag{9}$$

Then, we perform speech enhancement as follows:

$$y(\omega,t) = \begin{cases} y_1(\omega,t) & \text{if } |y_1(\omega,t)| \leq |y_2(\omega,t)|, \\ y_2(\omega,t) & \text{otherwise.} \end{cases} \tag{10}$$

This means that we choose the best spatial filter to suppress interferers in each TF bin, which we call the TFS technique. We refer to this selection rule as the minimum value selection (MIN), and this beamformer as the TFS beamformer.

When we have the precise RTF, the powers of the target in $|y_k(\omega,t)|$ ideally match owing to the constraint in (4). Therefore, the comparison of $|y_k|$ is equal to that of the powers of interferers. The power of $y$ composed of $y_k(\omega,t)$, which has minimum power, is thus minimum. In accordance with the above theory, this method requires the precise RTF, which is the key problem as in the other speech enhancement techniques.

This method has several advantages: 1) We can use any conventional beamformer to construct the spatial filter, e.g., a MaxSNR beamformer [12, 15]. 2) W-DO between interferers is required instead of that between the target and the interferers, that is, this method relaxes the limitation of conventional TF masking assuming W-DO. 3) No target distortion due to the TFS of beamformers occurs if we use appropriate beamformers such as the MVDR beamformer. Here, if both the magnitude and phase of the output signal of all beamformers match, the beamformers can be considered to be appropriate.

In [7], we concluded that the TFS of the MVDR beamformer shows good speech enhancement performance. However, since this method requires the pre-estimation of each spatial filter, the target-active period and interferer-$k$-active periods are required separately as the prior information. This severely limits the practicality in acoustic environments.

### 3.2. Proposed TFS of MVDR beamformer

Although one of the advantages of the TFS beamformer is the availability of an arbitrary beamforming technique, we here focus on the MVDR beamformer. By reformulating the TFS technique and MVDR beamformer as a joint optimization problem, we can overcome the serious limitation of the conventional TFS beamformer.

Without loss of generality, we consider a two-microphone case ($M = 2$). The optimization problem of the MVDR beamformer based on the TFS technique is

$$\min_{\boldsymbol{w},m} \sum_{k=1}^{K} \sum_{\omega} \mathbb{E}\left[|m_k(\omega,t)\boldsymbol{w}_k^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t)|^2\right] \quad \text{s.t.} \quad \boldsymbol{w}_k^{\mathrm{H}}(\omega)\boldsymbol{a}(\omega) = 1, \tag{11}$$

where $K = N-1$ is the number of spatial filters and $m_k(\omega,t)$ is a TF binary mask that takes a value of one if $\boldsymbol{w}_k(\omega)$ is used and zero otherwise. Note that if $N = 2$ and, thus, $K = 1$ (i.e., a determined case), this optimization problem is equal to the conventional one in (4). Using the method of the Lagrange multiplier, we obtain the cost function $\mathcal{J}_{\mathrm{p}}$ as

$$\mathcal{J}_{\mathrm{p}} = \sum_{k=1}^{K} \sum_{\omega} \left\{\mathbb{E}\left[|m_k(\omega,t)\boldsymbol{w}_k^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t)|^2\right] + 2\mathrm{Re}[\lambda_k^*(\boldsymbol{w}_k^{\mathrm{H}}(\omega)\boldsymbol{a}(\omega) - 1)]\right\}, \tag{12}$$

where $\lambda_k^*$ is the $k$th complex-valued Lagrange multiplier. Since the minimization of $\mathcal{J}_{\mathrm{p}}$ is a joint optimization problem, it is difficult to optimize both $\boldsymbol{w}_k$ and $m_k$ simultaneously, whereas it is straightforward to optimize them alternately.

With $\boldsymbol{w}_k$ fixed, the cost function regarding $m_k$ is

$$\mathcal{J}_{\mathrm{p}}(m_k) = \sum_{k=1}^{K} \sum_{\omega} \mathbb{E}\left[|m_k(\omega,t)\boldsymbol{w}_k^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t)|^2\right]. \tag{13}$$

Thus,

$$m_k(\omega,t) = \begin{cases} 1 & \text{if } |\boldsymbol{w}_k^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t)|^2 \leq |\boldsymbol{w}_{k'}^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t)|^2 \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where $k' = 1, \dots, K$ and $k' \neq k$. This optimization means that we choose the best spatial filter in each TF bin; that is, we perform TFS of the MVDR beamformer.

Next, with $m_k$ fixed, the cost function regarding $\boldsymbol{w}_k$ is equal to (12), that is, $\mathcal{J}_{\mathrm{p}}(\boldsymbol{w}_k) = \mathcal{J}_{\mathrm{p}}$. Focusing on the TF bins where the $k$th filter is used (i.e., $m_k(\omega,t) = 1$), $\mathcal{J}_{\mathrm{p}}(\boldsymbol{w}_k)$ is equal to the conventional one (6); thus, this optimization problem has the following closed-form solution:

$$\boldsymbol{w}_k(\omega) = \frac{\boldsymbol{\Phi}_k(\omega)^{-1}\boldsymbol{a}(\omega)}{\boldsymbol{a}^{\mathrm{H}}(\omega)\boldsymbol{\Phi}_k(\omega)^{-1}\boldsymbol{a}(\omega)}, \tag{15}$$

$$\boldsymbol{\Phi}_k(\omega) = \mathbb{E}\left[(m_k(\omega,t)\boldsymbol{x}(\omega,t))\,(m_k(\omega,t)\boldsymbol{x}(\omega,t))^{\mathrm{H}}\right]. \tag{16}$$

Using the above equations, we iteratively update $\boldsymbol{w}_k$ and $m_k$. For the initialization, $\boldsymbol{w}_k$ or $m_k$ can be computed by conventionally constructing a spatial filter, such as a null beamformer [12], or a TF mask, such as a degenerate unmixing estimation technique (DUET) [16], respectively. We can avoid the permutation problem by using an initial value, that is, filter $\boldsymbol{w}_k$ suppresses interferer $k$ in every frequency bin, whereas this does not hold when we use a random value. Finally, speech enhancement is performed using

$$y_k(\omega,t) = m_k(\omega,t)\boldsymbol{w}_k^{\mathrm{H}}(\omega)\boldsymbol{x}(\omega,t), \tag{17}$$

$$y(\omega,t) = \sum_{k=1}^{K} y_k(\omega,t), \tag{18}$$

where $y_k$ ideally contains a part of the target and only the $k$th interferer is suppressed by $\boldsymbol{w}_k$, and $y$ contains the completely restored target and suppressed interferers. Note again that all of the formulas presented above, including the computation of a spatial filter (15) and an enhanced signal (17) (18), are in complete agreement with those for the conventional MVDR beamformer in the determined situation.

This joint optimization problem can be understood from two viewpoints: the clustering of the most dominant interferer in each TF bin, which is the computation of $m_k$, and the construction of a MVDR beamformer $\boldsymbol{w}_k$ using such bins.

**Table 1**: Experimental conditions

| | |
|---|---:|
| Number of microphones $M$ | 2 |
| Number of sound sources $N$ | 3 or 4 |
| Distance between microphones | 4 cm |
| Reverberation time | 300 ms |
| Sampling rate | 8 kHz |
| FFT frame length / shift | 1024 / 256 samples |
| Test period | 5 s |

## 4. EXPERIMENTAL EVALUATION

### 4.1. Experimental conditions

To evaluate the effectiveness of our proposed method, we conducted an experiment using observed signals that are convolutive mixtures of impulse responses simulated by a room impulse response generator [17]. The experimental conditions are listed in Table 1.

We evaluated the performance of our proposed method by comparing with the results of the following three conventional methods: MVDR, which is underdetermined speech enhancement with a single MVDR beamformer; DUET [16] as an example of TF binary masking with a stereo microphone; and the conventional TFS beamformer previously proposed in [7]. Here, we combined the MVDR beamformer by MIN (see (10)). This method requires the RTF from the target source to the microphones and interferer-wise-active periods.

For the proposed method, we used the same RTF and a null beamformer [12] to initialize the spatial filters $\boldsymbol{w}_k$, which enhances the target and steers a spatial null in a random direction. We set the interval of the random direction to be at least 20° and evaluated five different initial directions of null steering. We updated $m_k$ and $\boldsymbol{w}_k$ ten times iteratively using (14) and (15), respectively.

To verify the effectiveness of the proposed method, we prepared six interferers A to F, whose directions of arrival (DOAs) were 20°, 40°, 60°, 110°, 130°, and 150°, respectively. We used nine combinations consisting of one interferer

**Table 2**: Results of speech enhancement for different noise sets (A, 20°; B, 40°; C, 60°; D, 110°; E, 130°; F, 150°). TFS w/o PE, Proposed TFS-MVDR without pre-estimation (PE) of beamformer; TFS w/ PE, Conventional TFS-MVDR with PE

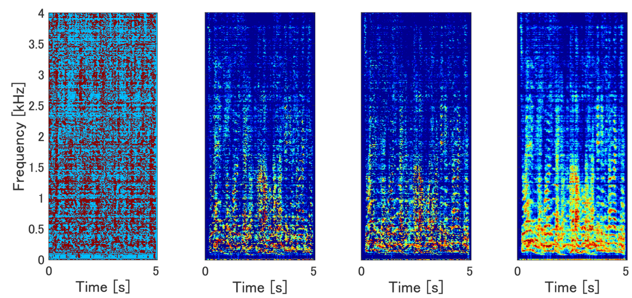
(a) SDR improvement [dB]

| Method | Noise set | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AD | AE | AF | BD | BE | BF | CD | CE | CF | Ave. | ACE | BDF | CDF | Ave. |
| MVDR | -0.75 | -0.41 | -1.22 | 0.76 | 1.27 | -0.05 | -0.32 | 0.01 | -0.90 | -0.18 | -0.57 | 0.00 | 0.02 | -0.18 |
| DUET | 2.15 | 2.63 | 3.24 | 2.49 | 2.63 | 2.93 | 0.39 | 0.61 | 1.29 | 2.04 | 1.64 | 2.35 | 1.14 | 1.71 |
| TFS w/o PE | **4.97** | **5.37** | **4.95** | **4.99** | **4.99** | **5.44** | **3.69** | **4.02** | **3.72** | **4.68** | **3.80** | **4.04** | **3.21** | **3.68** |
| TFS w/ PE | 5.79 | 6.51 | 6.41 | 5.85 | 6.26 | 6.04 | 4.26 | 4.75 | 4.85 | 5.64 | 5.10 | 5.37 | 4.33 | 4.93 |

(b) SIR improvement [dB]

| Method | AD | AE | AF | BD | BE | BF | CD | CE | CF | Ave. | ACE | BDF | CDF | Ave. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MVDR | 0.18 | 0.82 | 0.65 | 1.59 | 2.37 | 1.15 | 0.19 | 0.80 | 0.44 | 0.91 | 0.47 | 1.30 | 0.76 | 0.84 |
| DUET | 4.02 | 2.63 | 4.82 | 5.11 | 2.63 | 4.68 | 2.14 | 1.84 | 2.87 | 3.41 | 3.56 | 4.61 | 3.43 | 3.87 |
| TFS w/o PE | **8.71** | **8.93** | **8.39** | **8.73** | **8.42** | **9.53** | **6.41** | **6.59** | **6.75** | **8.05** | **7.28** | **8.45** | **6.57** | **7.43** |
| TFS w/ PE | 8.62 | 9.71 | 9.56 | 7.44 | 9.47 | 9.47 | 6.04 | 7.03 | 7.25 | 8.29 | 7.72 | 8.63 | 6.65 | 7.67 |

from interferers A to C and another from interferers D to F as noise signals. Additionally, we investigated the performance using the noise sets ACE, BDF, and CDE as a case of including three interferers. For each of the cases, we set the number of spatial filters $K$ to $N-1$. As the target signal, we used Japanese male/female and English male/female speech (i.e., there were four types of target signal), whose DOA was 90°. The SNR between the target signal and each interferer signal was set to 0 dB. We used objective criteria, namely, the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) [18] to quantify the results. A concise representation of the results was obtained by averaging these criteria over the target speech and the initialization of the spatial filters. Here, the reference signal was the source image, i.e., the noise-free reverberant speech signal.




(a) (b) (c) (d)

**Fig. 2**: Results of speech enhancement by the proposed TFS-MVDR. (a): TF mask $m_k$ indicating the selected beamformer (blue: $k=1$, red: $k=2$), (b) and (c): $y_1$ and $y_2$ (see (17)), respectively, (d): enhanced speech $y$ ($y = y_1 + y_2$, see (18))

### 4.2. Results and discussion

The SDR and SIR are shown in Table 2 for each noise set. The MVDR, which has the signal-dependent spatial filter $\boldsymbol{w}(\omega)$, can suppress only one interferer in each frequency bin; thus, it failed in speech enhancement. The proposed TFS-MVDR shows a performance superior to that of DUET regardless of the interferer DOA. Additionally, the performance is close to that of conventional TFS-MVDR even though the proposed TFS-MVDR does not require the pre-estimation, which the conventional TFS-MVDR requires. Considering these results, it can be concluded that our proposed TFS-MVDR improves the speech enhancement performance in an underdetermined noisy environment.

Figures 2(a)–(d) show examples of the results of speech enhancement by the proposed TFS-MVDR. As shown in Fig. 2(a), the selected beamformer switches frequently in the TF plane. However, the proposed TFS-MVDR basically satisfies the linear constraint in (11); thus, no distortion due to the switching of the beamformers occurs. The enhanced signals $y_1$ and $y_2$ are shown in Figs. 2(b) and (c), respectively, and the final output $y$ of the proposed method, which is the sum of $y_1$ and $y_2$ (see (18)), is shown in Fig. 2(d).

The set of colored TF bins in Fig. 2(a) contains interferer $k$ ($k=1$ (blue), $k=2$ (red)) regardless of the target. Therefore, each of the sets is composed of the target and only the interferer $k$. The beamforming using a signal represented by each of these sets is thus performed under the determined condition. Here, note that the output signal $y_k$ has only a part of the target because it is computed using the set of TF bins whose $m_k(\omega,t)=1$. Moreover, $y_k$ and each of the other enhanced signals $y_{k'}$ are completely disjointed [14]. The enhanced signal $y$, which is the sum of $y_k$, thus suppresses both the interferers and the target is restored completely.

In this experiment, we used time-invariant spatial filters $\boldsymbol{w}_k(\omega)$, whereas the MVDR beamformer can construct a time-variant spatial filter $\boldsymbol{w}(\omega,t)$. Thus, our proposed method can also do this. With time-variant spatial filters $\boldsymbol{w}_k(\omega,t)$, we expect that the proposed method will work well in time-varying environments.

## 5. CONCLUSIONS

In this paper, we have proposed TFS of the MVDR beamformer as a new method of underdetermined speech enhancement using two microphones, which is an extension of the conventional MVDR beamformer. This method is also an extension of conventional TF masking and can be considered as a combination of beamforming and TF masking.

We demonstrated the effectiveness of the proposed method by performing an experiment in a reverberant environment. The proposed method showed high performance regardless of the interferer DOA and was always superior to the conventional methods used for comparison in terms of speech enhancement performance.

## 6. 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, 2017.

[3] Ö. 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, "Underdetermined 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] K. Yamaoka, A. Brendel, N. Ono, S. Makino, M. Buerger, T. Yamada, and W. Kellermann, "Time-frequency-bin-wise beamformer selection and masking for speech enhancement in underdetermined noisy scenarios," *Proc. EUSIPCO*, pp. 1596–1600, 2018.

[8] 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, 2017.

[9] 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.

[10] 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, 2010.

[11] 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*, pp. 741–745, 2010.

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

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

[14] A. Jourjine, S. Rickard, and Ö. Yılmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," *Proc. ICASSP*, pp. 2985–2988, 2000.

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

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

[17] E. A. P. Habets, "Room impulse response (RIR) generator," Available at: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator, 2008.

[18] 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.