2019 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP2019)
Honolulu, Hawaii, USA, March 4-7, 2019

Performance Evaluation of Time-Frequency-Bin-Wise Switching Beamformer in Reverberant Environments

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
{yamaoka@mmlab.cs, maki@tara, takeshi@cs}.tsukuba.ac.jp

[2]Tokyo Metropolitan University
6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan
onono@tmu.ac.jp

Abstract

In this paper, we investigate the performance of the time-frequency-bin-wise switching (TFS) beamformer in reverberant environments. The TFS beamformer is a beamforming technique we previously proposed for underdetermined situations. A conventional speech enhancement method for underdetermined situations is time-frequency masking, which assumes that all sources are W-disjoint orthogonal. On the other hand, the assumption of the TFS beamformer is that only the interferer signals satisfy the W-disjoint orthogonality (W-DO), which relaxes the limitation of conventional time-frequency masking. However, long reverberant environments may cause the breakdown of W-DO. In this study, we therefore conducted experiments on underdetermined speech enhancement in reverberant environments to evaluate the effectiveness of the TFS beamformer. We confirmed that it was superior to conventional time-frequency masking in terms of the performance of speech enhancement regardless of the reverberation time.

1. 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]). The performance of these methods using a microphone array, however, decreases in underdetermined situations in which we have fewer microphones M than sound sources N. Recently, small recording devices such as voice recorders and smartphones have become common. These devices often have only two microphones, and therefore two-channel processing should be more convenient. Although several conventional methods such as time-frequency masking [3, 4], multichannel Wiener filtering [5], and a virtual microphone technique [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 speech enhancement in underdetermined situations realizing high performance with low signal distortion.

We previously proposed the time-frequency-bin-wise switching (TFS) beamformer [7] as an extension of conventional speech enhancement based on time-frequency masking, which uses multiple preconstructed beamformer filters. If M microphones are available, a single beamformer can generally form $M - 1$ spatial 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 using two microphones, 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 (see Fig. 1). Therefore, this method enhances a speech signal by multiplying it by the best beamformer filter to suppress interferers in each time-frequency bin rather than multiplying it by a scalar as in time-frequency masking.

In [8], the combination of multiple beamformers with different steering directions for audio zooming was considered. However, in the present 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 and motors in a robot, by selecting the most suitable noise covariance matrix in each time-frequency bin for the computation of maximum signal-to-noise ratio (MaxSNR) beamformer filters 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 patterns of the actuator 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 time-frequency bin for underdetermined speech enhancement with no target distortion.

In this paper, we evaluate the performance of the TFS beamformer in reverberant environments. One of the advantages of this method is that W-disjoint orthogonality (W-DO) between interferer signals is required instead of that between target and interferer signals. That is, this method relaxes the limitation of conventional time-frequency masking assuming W-DO, which means that only one source can dominate a time-frequency bin. Since long reverberant environments may break this assumption, the performance of conventional time-frequency masking may be degraded. On the other hand, our proposed method should guarantee a certain performance in such environments. Therefore, we study the effect of reverberation on the TFS beamformer by comparing the speech enhancement performance in various reverberant environments.

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

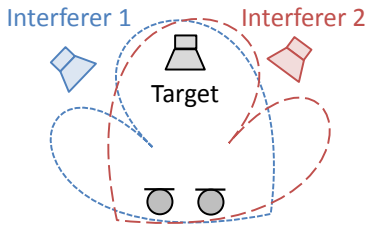


Figure 1: Combination of two beamformers with a spatial null for each interferer in an underdetermined situation with $M = 2$ and $N = 3$

2. Conventional linear beamformer

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 angular frequency ω in the tth time frame. When M microphones observe N sound sources consisting of one target and $N-1$ interferers in determined situations ($M = N$), we can perform conventional speech enhancement using a beamformer, such as an MVDR beamformer, which steers a spatial null in the direction of the interferer, as described by the following equations:

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

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

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

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 MVDR beamformer can enhance the target signal with a distortionless response. However, only $M-1$ interferers can be suppressed using M microphones. The performance of linear speech enhancement may therefore be degraded in underdetermined situations with $M < N$, i.e., when we have fewer microphones M than sound sources N.

3. Proposed TFS beamformer

Without loss of generality, we consider a situation with $M = 2$ microphones and N sound sources consisting of one target and $N-1$ interferers. In this situation, we cannot construct a null beamformer that suppresses all interferers simultaneously. However, if only the target and the kth interferer are observed ($k = 1, \ldots, K$ and $K = N-1$), we can construct the kth beamformer, which suppresses only the kth interferer using a conventional beamforming method, and the same is true for the other beamformers and interferers. Then, we obtain the following output signal $y_k(\omega, t)$ for each beamformer from an observation $\boldsymbol{x}(\omega, t)$ consisting of N sound sources:

$$y_k(\omega, t) = \boldsymbol{w}_k^{\mathrm{h}}(\omega)\boldsymbol{x}(\omega, t), \tag{4}$$

where $\boldsymbol{w}_k(\omega)$ is the spatial filter defining beamformer k. Then, the TFS beamformer [7] basically uses a combination of these beamformer filters for underdetermined speech enhancement as follows:

$$y(\omega, t) = \sum_{k=1}^{K} m_k(\omega, t)\boldsymbol{w}_k^{\mathrm{h}}(\omega)\boldsymbol{x}(\omega, t), \tag{5}$$

$$m_k(\omega, t) = \begin{cases} 1 & \text{if } |\boldsymbol{w}_k^{\mathrm{h}}(\omega)\boldsymbol{x}(\omega, t)|^2 \leq |\boldsymbol{w}_{k'}^{\mathrm{h}}(\omega)\boldsymbol{x}(\omega, t)|^2 \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where $k' = 1, \ldots, K$ and $k' \neq k$, and $m_k(\omega, t)$ is a time-frequency binary mask that takes a value of one if $\boldsymbol{w}_k(\omega)$ is the best beamformer and zero otherwise. Note that the proposed TFS beamforming (5) and (6) is in complete agreement with that for the conventional beamforming (1) in the determined case ($N = 2$ and, thus, $K = 1$).

This method has several advantages 1) We can use any conventional beamformer to construct the spatial filter $\boldsymbol{w}_k$, such as an MVDR beamformer, MaxSNR beamformer [12, 14], and also a fixed beamformer. 2) No target distortion due to the switching of the beamformers occurs if we use appropriate beamformers such as the MVDR beamformer. Here, if both the magnitude and phase of the output signal $y_k(\omega, t)$ of all beamformers match, the beamformers can be considered to be appropriate. 3) W-DO between interferer signals is required instead of that between target and interferer signals, that is, this method relaxes the limitation of conventional time-frequency (TF) masking assuming W-DO. Because of advantage 3), the proposed method should be robust against reverberation.

3.1 Extension of TFS beamformer using time-frequency masking as postprocessing

Speech enhancement by the TFS beamformer shows high speech enhancement performance in an underdetermined noisy environment. 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,

$$y_{\mathrm{post}}(\omega, t) = M(\omega, t)y(\omega, t), \tag{7}$$

where $M(\omega, t)$ is a time-frequency mask. We previously proposed the construction of $M(\omega, t)$ for the TFS beamformer by estimating the activity of the sound sources. Here, we estimated the activity by evaluating a trained probabilistic model that describes the directions of arrival (DOAs) of the sound sources [15].

4. Experimental evaluation

To evaluate the performance of our proposed TFS beamformer in reverberant environments, we conducted an experiment using observed signals that are convolutive mixtures of impulse responses simulated by a room impulse response generator [16]. The layout of the sound source and microphones is shown in Fig. 2. The fast Fourier transform (FFT) frame lengths used in the experiment and the other experimental conditions are listed in Tables 1 and 2, respectively.

For the proposed method, we combined the MVDR beamformer and provided the relative transfer function (RTF) and interferer-wise-active periods as prior information for the

Table 1: FFT frame lengths used in the experiment

	120 ms	310 ms	780 ms	
MVDR	1024	2048	4096	[samples]
DUET	2048	2048	2048	[samples]
TFS	1024	4096	16384	[samples]

Table 2: Experimental conditions

Number of microphones M	2
Number of sound sources N	4
Distance between microphones	4 cm
Reverberation time	120, 310, and 780 ms
Sampling rate	16 kHz
FFT frame shift	1 / 4 overlap
Training period	5 s
Test period	10 s

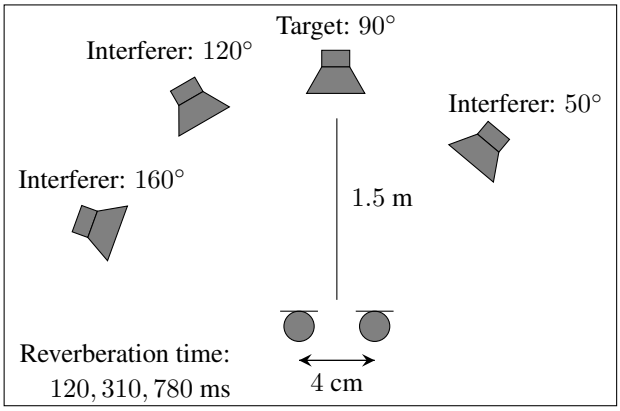


Figure 2: Layout of sound source and microphones in the experiment

multiple preconstructed beamformers. Using these periods, we calculated the interference-plus-noise covariance matrices for the computation of the spatial filters.

We evaluated the performance of our proposed method by comparison with the results of the following two conventional methods: MVDR, which is underdetermined speech enhancement with a single MVDR beamformer, i.e., conventional beamforming and the degenerate unmixing estimation technique (DUET) [17] as an example of conventional time-frequency binary masking with a stereo microphone. We used Japanese male/female and English male/female speech as the target signals, whose DOA was 90°, and three speeches as the interferer signals, whose DOAs were 50°, 120°, and 160°. The signal-to-noise ratio (SNR) between the target signal and each interferer signal was set to 0 dB. We used the best window length for each method and each reverberant environment in terms of the signal-to-distortion ratio (SDR) performance (see Table 1).

We used objective criteria, namely, the SDR, signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR) [18], to quantify the results. A concise representation of the results is obtained by averaging these criteria over speakers. Here, the reference signal is the source image, i.e., the noise-free reverberant speech signal.

4.1 Results and discussion

The improvements of SDR, SIR, and SAR are shown in Fig. 3 for each reverberation time. Figure 4 is a histogram showing the proportion of the sources that are simultaneously active in each time-frequency bin. In this figure, we consider that source $s_i(\omega, t)$ $(i = 1, \ldots, 4)$ is active when it has an amplitude greater than $\max|s_i(\omega, t)|/10$ for all i at each frequency [19]. If two or more sources are active, W-DO is not satisfied.

Regarding the speech enhancement performance in Fig. 3, the MVDR can suppress only one interferer; thus, it shows good performance only for SAR. Since W-DO is satisfied in almost all time-frequency bins with a reverberation time of 120 ms, the DUET shows good performance. However, mixed signals with a long reverberation tend not to satisfy W-DO as shown in Fig. 4. The improvements of the performance with the DUET are therefore significantly decreased in the case of long reverberation. On the other hand, our proposed method can work well in such cases. Considering these results, it can be concluded that our proposed method improves the speech enhancement performance in reverberant noisy environments.

Next, we discuss the relationship between the proposed TFS beamformer and the reverberation time. As the reverberation time increases, the improvement of each criterion for the proposed method decreases. However, the decreases are small, which shows its robustness against reverberation. As shown in Fig. 4(b), there are two sources in approximately 20% of the time-frequency bins. Speech enhancement by time-frequency masking, such as by the DUET, is theoretically impossible in such bins. On the other hand, when two microphones are available, the proposed method can suppress an interferer even if two sound sources exist. Note that if two interferers exist, postprocessing by time-frequency masking is necessary [15]. Moreover, focusing on Fig. 4(c), there are multiple sound sources in a large number of time-frequency bins. When two sound sources are simultaneously active, the proposed method can work well as described above. Even though the proposed method cannot suppress all interferers at the same time, at least one of the interferers can be suppressed, whereas time-frequency masking suppresses all sources including the target speech or does not suppress any sources. Thus, a certain improvement is guaranteed for the proposed method.

5. Conclusions

In this study, we have evaluated the performance of the proposed time-frequency-bin-wise switching (TFS) beamformer in reverberant environments. This method has an advantage that W-DO between interferer signals is required instead of that between target and interferer signals, which means that this method relaxes the limitation of conventional time-frequency masking assuming W-DO.

We demonstrated the effectiveness of the proposed method by performing an experiment on speech enhancement in various reverberant environments. The proposed method showed high performance regardless of the reverberation time and was superior to the conventional methods used for comparison in terms of the speech enhancement performance.

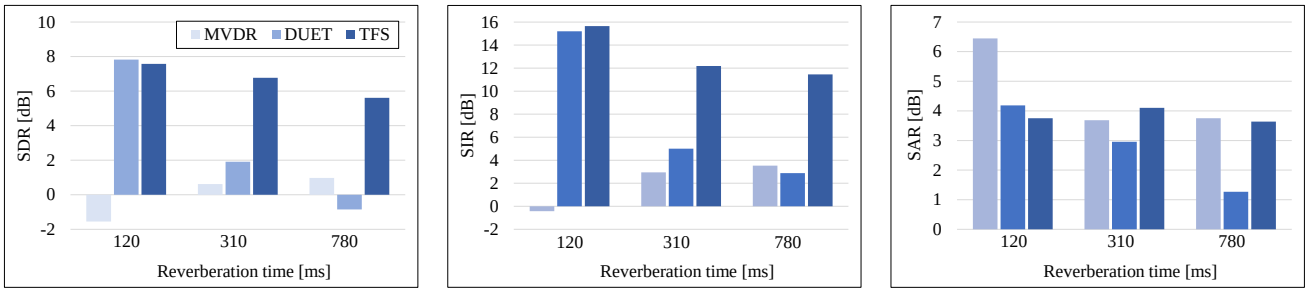
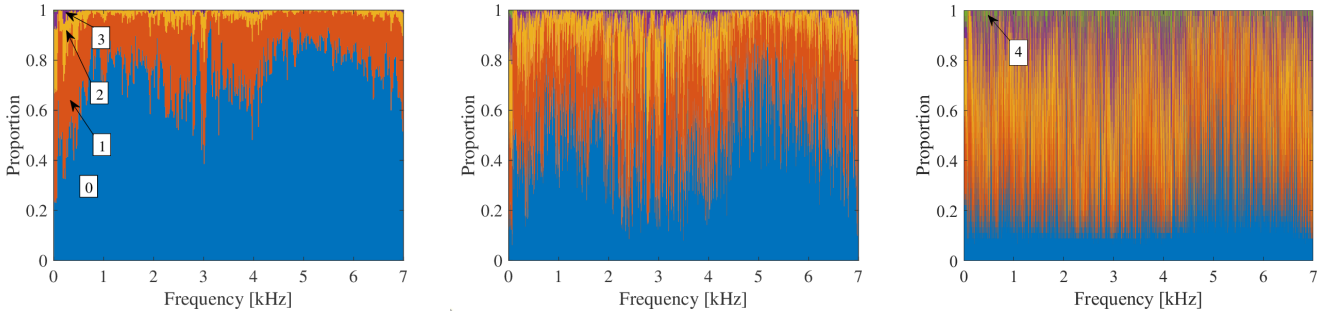


Figure 3: Results of speech enhancement for different reverberation times



(a) 120 ms (b) 310 ms (c) 780 ms

Figure 4: Proportion of simultaneously active source in each time-frequency bin for different reverberation times

References

[1] S. Makino *et al.*, *Blind Speech Separation*, Springer, 2007.

[2] T. Higuchi *et al.*, "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] O. Yılmaz *et al.*, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] H. Sawada *et al.*, "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 *et al.*, "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] H. Katahira *et al.*, "Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, pp. 1–8, 2016.

[7] K. Yamaoka *et al.*, "Time-frequency-bin-wise beamformer selection and masking for speech enhancement in underdetermined noisy scenarios," *in Proc. EUSIPCO*, pp. 1596–1600, 2018.

[8] N. Q. K. Duong *et al.*, "Audio zoom for smartphones based on multiple adaptive beamformers," *in Proc. LVA/ICA*, pp. 121–130, 2017.

[9] S. Takada *et al.*, "Sound source separation using null-beamforming and spectral subtraction for mobile devices," *in Proc. WASPAA*, pp. 30–33, 2007.

[10] T. Ogawa *et al.*, "Speech enhancement using a square microphone array in the presence of directional and diffuse noise," *IEICE Trans. Fundam.*, vol. E93-EA, no. 5, pp. 926–935, 2010.

[11] M. Togami *et al.*, "Beamforming array technique with clustered multichannel noise covariance matrix for mechanical noise reduction," *in 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," *in Proc. IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[14] S. Araki *et al.*, "Blind speech separation in a meeting situation with maximum SNR beamformers," *in Proc. ICASSP*, pp. 41–45, Apr. 2007.

[15] K. Yamaoka *et al.*, "Performance evaluation of nonlinear speech enhancement based on virtual increase of channels in reverberant environments," *in Proc. EUSIPCO*, pp. 2388–2392, 2017.

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

[17] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino *et al.*, Eds., pp. 217–241, Springer, 2007.

[18] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] A. Blin *et al.*, "Underdetermined blind separation of convolutive mixtures of speech using time-frequency mask and mixing matrix estimation," *IEICE Trans. Fundam.*, vol. 88-A, no. 7, pp. 1693–1700, 2005.