SOFT MASKING BASED ADAPTATION FOR TIME-FREQUENCY BEAMFORMERS UNDER REVERBERANT AND BACKGROUND NOISE ENVIRONMENTS

Yohei Kawaguchi and Masahito Togami

Central Research Laboratory, Hitachi, Ltd.
1-280, Higashi-koigakubo Kokubunji-shi, Tokyo 185-8601, Japan
{yohei.kawaguchi.xk, masahito.togami.fe}@hitachi.com

ABSTRACT

We propose a new method to adapt a beamformer under reverberant and background noise environments. Several adaptation approaches have two stages consisting of time-frequency masking and an update of the beamformer. However, in these approaches, the adaptation error is large under such environments because the sparseness assumption does not hold. We focus on the premise that the adaptation error can be reduced by avoiding the degradation caused by the overlap between sources in time-frequency bins. Therefore, we derive a formula to update the beamformer in cases when overlap exists by using soft masking. The proposed method controls adaptation with soft masking to avoid the degradation caused by the overlap and reduces the adaptation error. Experimental results under a reverberant and background noise environment indicate that the proposed method improves the performance.

1. INTRODUCTION

Noise reduction techniques based on a microphone array have been intensively studied in recent years due to their many applications, for example, in hands-free speech recognition and in teleconference systems. These techniques are categorized into two classes: time-frequency domain masking approaches and beamforming approaches. The former class can be used in both underdetermined cases and determined cases. The latter class can be used in the determined cases.

Time-frequency domain masking approaches such as binary masking [1] separate the desired signal by Direction of Arrival (DOA) estimation for each time-frequency bin. These approaches make use of the sparseness assumption, which involves the characteristics of speech where only one source or zero sources are active in each time-frequency bin. Therefore, the performance degrades significantly under practical reverberant or background noise environments because the reflection and noise components result in overlap between sources in each bin, and thus, the sparseness assumption is not true. Under such environments, binary masking leads to musical noise as well as to degraded performance. We can alleviate this problem by using several soft masking approaches [2, 3] by modeling the overlap. These approaches do not solve the problem completely, but they improve the performance in practical environments.

Beamforming approaches such as Linearly-Constrained Minimum Variance (LCMV) [4], can potentially reduce interference while keeping the desired signal undistorted if we know the transfer characteristics such as the steering vector and correlation matrix of sound sources. Therefore, we need to adapt the beamformer to these transfer characteristics. To adapt the beamformer, we need to know which components are those of the desired source or the interference in the input signals. Several methods employ binary masking and beamforming in the time-frequency domain [5, 6, 7] to solve this problem. These binary masking and beamforming approaches (BM-BF) use binary masking in the adaptation process and beamforming in the final separation. In the adaptation process, we separate each source roughly using binary masking and calculate the beamformer from the roughly separated signals. In the final separation, we separate the desired signal by beamforming. BM-BF overcomes the difficulty of adaptation of beamforming. Moreover, BM-BF does not suffer from musical noise. However, the performance of BM-BF is degraded by reverberation and background noise because it employs the sparseness assumption of binary masking.

In this paper, we propose a method to adapt the beamformer in order to improve the performance in reverberant and background noise environments. The development of this method was motivated by the robustness of soft masking against reverberation and background noise. We focus on the premise that the adaptation error is reduced by avoiding the degradation caused by the overlap of sources in time-frequency bins. We model the overlap and formulate an updating formula of the beamformer for cases when overlap exists by using soft masking. This soft masking and beamforming approach (SM-BF) improves the performance in reverberant and background noise environments because it controls the averaging weight based on the probability that the desired component is active, which can be calculated from the modeled overlap. In section 5, our experimental results under a reverberant and background noise environment show that the proposed method improves the performance.

2. PROBLEM STATEMENTS AND NOTATION

We observe N sources with M microphones in a reverberant and background noise environment. This situation is modeled by the convolutive mixing model

$$x_m(t) = \sum_{i=0}^{N-1} \sum_{l=0}^{\infty} a_{m,i}(l) s_i(t-l) + d_m(t), \qquad (1)$$

where $x_m(t)$ is the signal observed by the m-th microphone, t is the time index, $s_i(t)$ is the signal of the i-th source, $a_{m,i}(t)$

 ISSN 2076-1465

represents the impulse response from the i-th source to the m-th microphone, and $d_m(t)$ is the background noise of the m-th microphone. Now, we define the 0-th source as the desired source and the i-th source ($i \neq 0$) as the interference. The desired source exists in a given area of source directions Λ_{des}, and the interference sources do not exist in Λ_{des}. This paper employs a time-frequency domain approach. Using a short-time Fourier transform (STFT), the time-frequency representation is given by

$$\mathbf{X}(f,\tau) = \sum_{i=0}^{N-1} \mathbf{A}_i(f) S_i(f,\tau) + \mathbf{D}(f,\tau), \tag{2}$$

where f is the frequency bin index, τ is the time-frame index of the STFT, $\mathbf{X}(f,\tau) = [X_0,\cdots,X_{M-1}]^T$, $X_m(f,\tau)$ is the STFT of $x_m(t)$, $\mathbf{A}_i(f) = [A_{0,i},\cdots,A_{M-1,i}]^T$, $A_{m,i}(f)$ is the STFT of $a_{m,i}(t)$, $S_i(f,\tau)$ is the STFT of $s_i(t)$, $\mathbf{D}(f,\tau) = [D_0,\cdots,D_{M-1}]^T$, and $D_m(f)$ is the STFT of $d_m(t)$. Our goal is to obtain the estimate of $S_0(f,\tau)$, i.e., $Y(f,\tau)$ by calculation from $\mathbf{X}(f,\tau)$. This paper employs beamforming

$$Y(f,\tau) = \mathbf{W}^H(f)\mathbf{X}(f,\tau), \tag{3}$$

and we present an adaptation method of the beamformer $\mathbf{W} = [W_0,\cdots,W_{M-1}]^T$.

We explain the sparseness of sources in the time-frequency domain. The sparseness assumption means that for the i-th source, the power of which is the maximum in (f,τ) (source i is active in (f,τ)), the power of $\mathbf{N}_i(f,\tau) \triangleq \sum_{k=0,k\neq i}^{N-1} \mathbf{A}_k(f)S_k(f,\tau) + \mathbf{D}(f,\tau)$ is sufficiently small, i.e.,

$$\mathbf{X}(f,\tau) \approx \mathbf{A}_i(f)S_i(f,\tau). \tag{4}$$

This assumption is widely employed for solving the underdetermined problem. However, the assumption cannot hold under reverberant and background noise environments.

3. BINARY MASKING-BASED ADAPTATION CONTROL

We show a conventional adaptation control method of BM-BF [5]. First, we estimate the direction index of each time-frequency $j(f,\tau)$ by using a DOA estimation method in each time-frequency, for example, SPIRE [8] or MDSBF [5]. Next, we separate the desired component and the interference roughly by using binary masking.

$$\hat{\mathbf{X}}_{\mathrm{des}}(f,\tau) = \begin{cases} \mathbf{X}(f,\tau) & \text{if} \quad j(f,\tau) \in \Lambda_{\mathrm{des}}, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$$\hat{\mathbf{X}}_{\mathrm{int}}(f,\tau) = \begin{cases} 0 & \text{if} \quad j(f,\tau) \in \Lambda_{\mathrm{des}}, \\ \mathbf{X}(f,\tau) & \text{otherwise.} \end{cases} \tag{6}$$

Then we estimate the steering vector of the desired source and the correlation matrix of the interference R on the sparseness assumption.

$$\hat{\mathbf{A}}_0(f) = \left\langle \frac{\hat{\mathbf{X}}_{\mathrm{des}}(f,\tau)\left|\hat{X}_{\mathrm{des0}}(f,\tau)\right|}{|\mathbf{X}_{\mathrm{des}}(f,\tau)|\,\hat{X}_{\mathrm{des0}}(f,\tau)} \right\rangle \tag{7}$$

$$R(f) = \left\langle \mathbf{X}_{\mathrm{int}}(f,\tau)\mathbf{X}_{\mathrm{int}}^H(f,\tau) \right\rangle \tag{8}$$

where $\hat{\mathbf{A}}_0$ is the estimated steering vector of the desired source, R is the estimated correlation matrix of the interference, $\hat{X}_{\mathrm{des0}}$ is the signal of the 0-th microphone of $\hat{\mathbf{X}}_{\mathrm{des}}$, and $\langle\cdot\rangle$ is the time average operator. Finally, we calculate the Frost beamformer [4],

$$\mathbf{W}(f) = \frac{R^{-1}(f)\hat{\mathbf{A}}_0(f)}{\hat{\mathbf{A}}_0(f)^H R^{-1}(f)\hat{\mathbf{A}}_0(f)}. \tag{9}$$

Under reverberant and background noise environments, the performance of binary masking is degraded because of the overlap in the pre-adaptation process, and then the beamformer $\mathbf{W}(f)$ also degrades.

4. SOFT MASKING-BASED ADAPTATION CONTROL

We model the overlap that leads to the degradation in the adaptation process. We do not discard the sparseness assumption completely, and assume a Gaussian distribution for $\mathbf{N}_i(f,\tau)$ such that the i-th source is active, i.e., $\mathcal{N}(\mathbf{0},\sigma^2(f)\Sigma(f))$ where $\sigma^2(f)$ is a variance parameter and $\Sigma(f)$ is a correlation matrix divided by $\sigma^2(f)$ in the reverberant sound field model [9],

$$\Sigma(f) = \begin{pmatrix} 1 & \gamma_{0,1} & \cdots & \gamma_{0,M-1} \\ \gamma_{1,0} & 1 & \cdots & \gamma_{1,M-1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{M-1,0} & \gamma_{M-1,1} & \cdots & 1 \end{pmatrix} \tag{10}$$

where $\gamma_{m,n} = \mathrm{sinc}(2\pi f d_{m,n}/c)$ where $d_{m,n}$ is the distance between the m-th microphone and the n-th microphone, and c is the sound velocity. Therefore, the pdf of $\mathbf{X}(f,\tau)$ given that i is active is given by:

$$\begin{aligned} & p(\mathbf{X}(f,\tau)|\sigma^2(f),\mathbf{A}_i(f),S_i(f,\tau)) \\ = \; & \frac{1}{(2\pi\sigma^2)^{M/2}\,|\Sigma|^{1/2}} \\ & \exp\left\{ -\frac{1}{2\sigma^2}\left(\mathbf{X}-\mathbf{A}_iS_i\right)^H\Sigma^{-1}\left(\mathbf{X}-\mathbf{A}_iS_i\right) \right\} \end{aligned} \tag{11}$$

We need to estimate $\mathbf{A}_i(f)$ from the input signals $\mathbf{X}(f,\tau)$ to adapt the beamformer by (9). This is equivalent to the GMM parameter estimation with missing data because we do not know which source i is active. We can employ the EM algorithm for maximum likelihood estimation with missing data like this. The Q function $Q(\Theta;\Theta^{(t)})$ of the EM algorithm is given by:

$$\begin{cases} Q(\Theta;\Theta^{(t)}) = \displaystyle\sum_{f,\tau,i} \mu_i^{(t)}(f,\tau)\log r_i p(\mathbf{X}|i,\Theta) & (12) \\ \mu_i^{(t)}(f,\tau) = \dfrac{r_i^{(t)} p\left(\mathbf{X}|i,\Theta^{(t)}\right)}{\sum_{k=0}^{N-1} r_k^{(t)} p\left(\mathbf{X}|k,\Theta^{(t)}\right)} & (13) \\ \displaystyle\sum_{i=0}^{N-1} r_i^{(t)} = 1 & (14) \end{cases}$$

where (t) represents the number of iterations,

$$\mathbf{\Theta} = (\sigma^2, (S_0, \cdots, S_N), (\mathbf{A}_0, \cdots, \mathbf{A}_N), r_i), \tag{15}$$

and r_i is an a priori probability that the i-th source is active. We can estimate $\mathbf{A}_i$ by using the Lagrange multiplier method, namely $\frac{\partial J}{\partial \mathbf{A}_i^H} = 0$, where $J = Q + \lambda(\sum_{i=0}^{N} r_i - 1)$. The estimated $\mathbf{A}_i$ with normalized norm and phase is given by:

$$\begin{aligned} \hat{\mathbf{A}}_i &\sim \sum_{\tau} \mu_i(f,\tau) A_{0,i}^* S_i^* \mathbf{X}(f,\tau) \\ &\approx \sum_{\tau} \mu_i(f,\tau) \frac{\mathbf{X}(f,\tau)\,|X_0(f,\tau)|}{|\mathbf{X}(f,\tau)|\,X_0(f,\tau)}, \end{aligned} \tag{16}$$

where $X_0(f,\tau)$ is the signal of the 0-th microphone of $\mathbf{X}(f,\tau)$. Therefore, $\hat{\mathbf{A}}_0(f)$ and $R(f)$ are updated for each block of T frames as follows:

$$\hat{\mathbf{A}}_0(f) = \left\langle \mu_0(f,\tau) \frac{\mathbf{X}(f,\tau)\,|X_0(f,\tau)|}{|\mathbf{X}(f,\tau)|\,X_0(f,\tau)} \right\rangle \tag{17}$$

$$R(f) = \left\langle (1-\mu_0(f,\tau))^2 \mathbf{X}(f,\tau)\mathbf{X}^H(f,\tau) \right\rangle \tag{18}$$

According to (17) and (18), we can use $\mu_i(f,\tau)$ as the averaging weight in the adaptation process of $\hat{\mathbf{A}}_0(f)$ and $R(f)$, and can compute $\mathbf{W}$ from these estimated $\hat{\mathbf{A}}_0$ and R according to (9).

To calculate (17) and (18), we need $\mu_0(f,\tau)$. We iterate EM steps, i.e., we estimate $\mu_i(f,\tau)$ with (13) and update the other parameters of $\mathbf{\Theta}$ as follows:

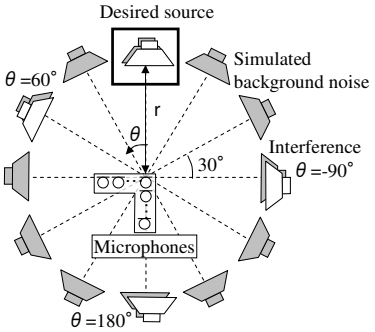
$$\hat{S}_i = \frac{\hat{\mathbf{A}}_i^H \Sigma^{-1} \mathbf{X}(f,\tau)}{\hat{\mathbf{A}}_i^H \Sigma^{-1} \hat{\mathbf{A}}_i} \tag{19}$$

$$r_i^{(t+1)} = \frac{\sum_{f,\tau} \mu_i^{(t)}(f,\tau)}{\sum_{f,\tau,i} \mu_i^{(t)}(f,\tau)} \tag{20}$$

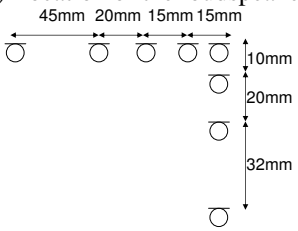
$$\begin{aligned} (\sigma^2)^{(t+1)}(f) &= \frac{1}{MT} \sum_{\tau,i} \mu_i^{(t)}(f,\tau) \left(\mathbf{X} - \hat{\mathbf{A}}_i \hat{S}_i\right)^H \\ &\quad \times \Sigma^{-1} \left(\mathbf{X} - \hat{\mathbf{A}}_i \hat{S}_i\right) \end{aligned} \tag{21}$$

As a result, in each block for T frames, the proposed method is applied to calculate (19), iterate (13), (20), and (21) until convergence, and finally to update (17), (18), and (9).

The iteration process of the proposed method is essentially equivalent to an extension to $M > 2$ of 2ch EM algorithm-based soft masking [3]. Therefore, we can call the proposed method the soft masking-based adaptation control method or the soft masking and beamforming approach (SM-BF). If the sparseness assumption is true, $\mu_i(f,\tau)$ is near 0 or 1; i.e., the proposed method is equivalent to BM-BF explained in section 3. Otherwise, we use $\mu_i(f,\tau)$ as the averaging weight in the adaptation of $\hat{\mathbf{A}}_0(f)$ and $R(f)$, and avoid the estimation error of $\hat{\mathbf{A}}_0(f)$ and $R(f)$ in time-frequency bins where the power of $\mathbf{N}_i$ is large, and also avoid the degradation of the beamformer.



(a) Location of the loudspeakers



(b) Microphone array

Fig. 1. Experimental setup. Here, r stands for the distance between the loudspeakers and the microphone array, and θ is the direction of interference.

5. EXPERIMENTAL RESULTS

We evaluated the performance of the proposed method under reverberant and background noise environments. The signals for evaluation were simulated by convolution of source signals with the impulse responses which were recorded for each location of the loudspeaker in a reverberant room. The reverberation time was about 400 ms. The impulse responses and source signals were recorded at a sampling rate of 48 kHz and then downsampled to 16 kHz. The desired source signal was human speech. The length of the signals was 18 s. The simulation was done with the setup illustrated in Fig. 1(a), using a microphone array consisting of eight microphones, as shown in Fig. 1(b). The desired source was configured in front, and the interference was configured at rm distance from the microphone array and the azimuth θ, where $r \in \{1\text{m}, 3\text{m}\}$ and $\theta \in \{180^\circ, -90^\circ, 60^\circ\}$. Λ_{des} was set to be $[-30^\circ, 30^\circ]$ We made the background noise by mixing all the signals for each direction. The desired source for the interference power ratio was set to be about 0 dB, and the desired source for the background noise power ratio was set to be about 10 dB. The number of sources N that the proposed method used was set at five.

First, in Fig. 2, we show an example of phase error of $\phi_{0,1}(f)$ for the adaptation approach based on binary masking and the proposed one based on soft masking. $\phi_{m,n}(f)$ is the m-th row, n-th column component of the estimated correlation matrix $R(f)$. "BM-BF" is the adaptation approach based on binary masking that is explained in section 3, and "SM-BF" is the proposed adaptation approach based on soft masking. The error is the difference between the phase of $\phi_{0,1}(f)$ estimated by adaptation of "BM-BF" or "SM-BF" and the ideal phase of $\phi_{0,1}(f)$ estimated by adaptation of

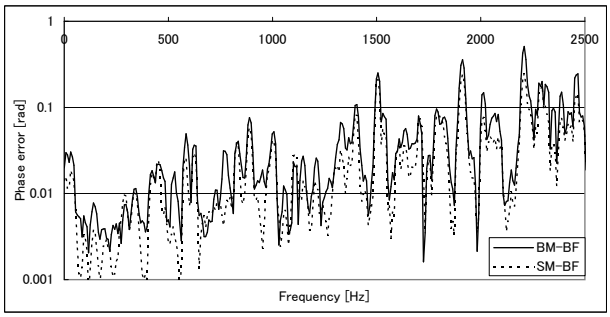


Fig. 2. Examples of phase error of $\phi_{0,1}(f)$ for BM-BF and SM-BF, where $\phi_{0,1}(f)$ is the 0-th row, 1-th column component of the estimated correlation matrix $R(f)$.

$R(f) = \langle \mathbf{X}(f,\tau)\mathbf{X}^H(f,\tau) \rangle$ under the condition where only the interference and the background noise exist during the whole time. This result shows that the estimation error of the correlation matrix of "SM-BF" is lower than that of "BM-BF" under reverberant and background noise environments.

Next, we compared the separation performance of the proposed method with two other methods: the binary masking and beamforming approach and the soft masking approach [3]. The measurements were Noise Reduction Rate (NRR) and Perceptual Evaluation of Speech Quality (PESQ) [10]. $NRR = 10\log_{10} \frac{\langle (x_0(t) - x_{\mathrm{des0}}(t))^2 \rangle}{\langle (y(t) - x_{\mathrm{des0}}(t))^2 \rangle}$, where y is the output signal of the final separation by the beamformer, and x_{des0} is the desired component in the input signal of the 0th microphone. Fig. 3 shows example waveforms and spectrograms of the output signals. Tab. 1 gives the NRR and PESQ of each method. "BM-BF" is the binary masking and beamforming approach. "SM" is the extension to $M > 2$ of soft masking [3]. "SM-BF" is the soft masking and beamforming approach.

In Fig. 3, the residual noise of "SM-BF" was less than that of "BM-BF." NRR of "SM-BF" was higher than that of "BM-BF" under most conditions, "SM-BF" outperformed "BM-BF" under all the conditions in PESQ. The performance of "BM-BF" was low under the noisy and reverberant environment of these experiments. In particular, "BM-BF" suffered from reverberation when the distance was large. Based on these results, we can infer that the proposed method reduced the estimation error of the correlation matrix and the steering vector, and the separation performance of the proposed method was also superior to that of "BM-BF" under reverberant and background noise environments. In Fig. 3, the residual noise of "SM" was less than that of "BM-BF" and "SM-BF." However, as the spectrogram of "SM" shows, the output signals of "SM" contained some musical noise and were more distorted than "BM-BF" and "SM-BF." NRR of "SM-BF" was higher than or equaled to that of "SM" under all the conditions, and "SM-BF" outperformed "SM" under all the conditions both in PESQ. The musical noise and the distortion are essential problems of time-frequency domain masking approaches under background noise and reverberant environments. We could not find any musical noise in the output signals of "BM-BF" or "SM-BF" because the final separation process of these methods is beamforming.

6. CONCLUSION

We proposed a method to adapt a beamformer under reverberant and background noise environments. We focused on the premise that the adaptation error can be reduced by avoiding the degradation caused by the overlap between sources in time-frequency bins under reverberant and background noise environments, and we derived a formula to update the beamformer in cases when overlap exists by using soft masking. The proposed method controls adaptation with soft masking to avoid the degradation caused by the overlap and reduces the adaptation error of the correlation matrix and the steering vector. Our experimental results under a reverberant and background noise environment showed that the proposed method outperforms other conventional methods.

7. REFERENCES

[1] 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, July 2004.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind sparse source separation with spatially smoothed time-frequency masking," in *Proc. IWAENC2006*, Sept. 2006.

[3] Y. Izumi, N. Ono, and S. Sagayama, "Sparsness-based 2ch bss using the em algorithm in reverberant environment," in *Proc. WASPAA2007*, Oct. 2007, pp. 147–150.

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

[5] M. Togami, Y. Obuchi, and A. Amano, "Automatic speech recognition of human-symbiotic robot emiew," in *Human-Robot Interaction*, Nilanjan Sarkar, Ed., pp. 395–404. I-tech Education and Publishing, 2007.

[6] J. Cermak, S. Araki, H. Sawada, and S. Makino, "Blind source separation based on beamformer array and time frequency binary masking," in *Proc. ICASSP2007*, Apr. 2007, pp. I–145–I–148.

[7] M. Kuhne, R. Togneri, and S. Nordholm, "Adaptive beamforming and soft missing data decoding for robust speech recognition in reverberant environments," in *Proc. InterSpeech2008*, Sept. 2008, pp. 976–979.

[8] M. Togami, T. Sumiyoshi, and A. Amano, "Stepwise phase difference restoration method for sound source localization using multiple microphone pairs," in *Proc. ICASSP2007*, Apr. 2007, pp. I–117–I–120.

[9] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelmann, and M. C. Thompson, "Measurement of correlation coefficients in reverberant sound fields," *Journal of Acoustical Society of America*, vol. 27, no. 6, pp. 1072–1077, 1955.

[10] ITUT P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, 2001.

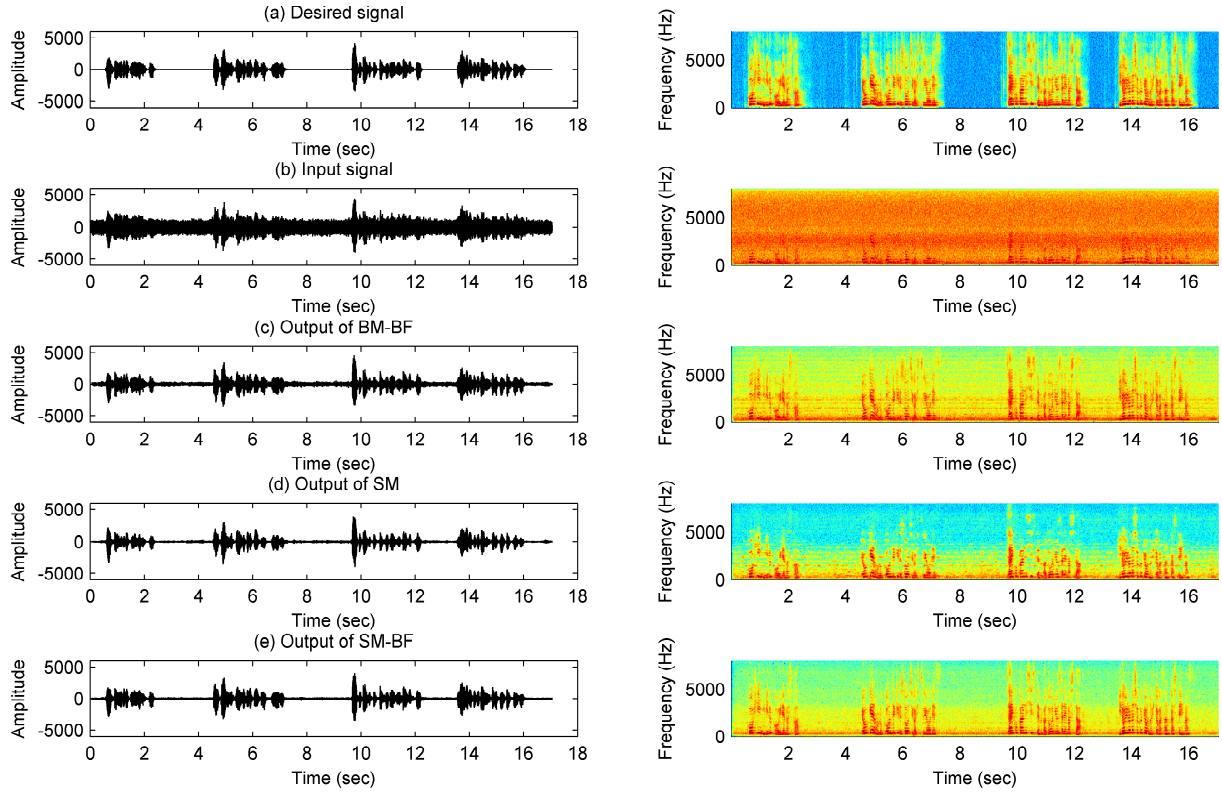


Fig. 3. Example waveforms and spectrograms.

Table 1. Evaluation results.

(a) NRR (dB)

Interference			Method		
	r	θ	BM-BF	SM	SM-BF
White noise	1m	180°	5.5	6.3	8.7
		−90°	5.5	6.1	9.2
		60°	6.6	5.2	6.0
	3m	180°	3.3	5.0	6.0
		−90°	3.4	3.5	4.6
		60°	3.9	3.5	5.2
Babble noise	1m	180°	4.1	5.8	6.9
		−90°	2.1	4.6	5.6
		60°	3.3	5.2	6.7
	3m	180°	2.7	4.3	4.3
		−90°	1.4	3.6	3.6
		60°	1.4	3.3	4.0

(b) PESQ

Interference			Method			
	r	θ	Input	BM-BF	SM	SM-BF
White noise	1m	180°	1.78	2.29	2.38	2.60
		−90°	1.77	2.40	2.43	2.68
		60°	1.83	2.47	2.42	2.60
	3m	180°	1.86	2.09	2.12	2.22
		−90°	1.81	2.08	2.00	2.29
		60°	1.86	2.15	2.12	2.29
Babble noise	1m	180°	1.65	2.33	2.41	2.55
		−90°	1.64	2.33	2.25	2.56
		60°	1.63	2.35	2.29	2.57
	3m	180°	1.71	2.11	2.21	2.25
		−90°	1.70	2.11	2.14	2.26
		60°	1.71	2.10	2.08	2.32