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

Noise Suppression Using Beamformer and Transfer-function-gain Nonnegative Matrix Factorization with Distributed Stereo Microphones

Yutaro Matsui[1], *Shoji Makino*[1], *Nobutaka Ono*[2], *Takeshi Yamada*[1]

[1]University of Tsukuba
1–1–1 Tennodai, Tsukuba, Ibaraki, Japan
E-mail: {y.matsui@mmlab.cs,maki@tara,takeshi@cs}.tsukuba.ac.jp

[2]Tokyo Metropolitan University
6–6 Asahigaoka, Hino, Tokyo, Japan
E-mail: onono@tmu.ac.jp

Abstract

In this paper, we introduce a novel approach to noise suppression using distributed recording devices. We focus on situations where multiple devices having stereo microphones are available for recording. We employ noise suppression based on phase information before applying transfer-function-gain nonnegative matrix factorization (NMF). Experiments using impulse responses measured in a meeting room showed that the proposed method outperformed the conventional methods using transfer-function-gain NMF in terms of the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR).

1. Introduction

An asynchronous distributed microphone array is one of the frameworks of array signal processing to perform speech enhancement or noise suppression. Since we can use portable recording devices such as voice recorders, smartphones or laptops, it is easy to construct an asynchronous distributed microphone array. However, the performance of the array signal-processing approaches based on phase information is degraded owing to phase drift between observed signals recorded by an asynchronous microphone array [1, 2]. On the other hand, the approaches utilizing amplitude information for noise suppression can still work effectively. One powerful framework of such approaches is the noise suppression method using transfer-function-gain nonnegative matrix factorization (NMF) [3–5].

When we use voice recorders, smartphones or laptops as recording devices in our daily lives, we can obtain two-channel signals from one device if it has stereo microphones. We can apply array signal processing based on phase information such as a beamformer to the signals from one device because they are synchronous. Thus, we propose a new approach that applies noise suppression based on phase information to the signals recorded by each microphone before applying transfer-function-gain NMF to all the signals recorded by an asynchronous microphone array. Employing the noise suppression based on phase information as the preprocessing is considered effective because it can improve the signal-to-noise ratio (SNR) of the input signals of transfer-function-gain NMF. In the experiments using impulse responses, the proposed method outperformed the conventional methods of noise suppression using transfer-function-gain NMF in terms of the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR).

2. Conventional method

Let us assume that one target speech signal and $K-1$ interfering speech signals are captured by M recording devices, each having stereo microphones. The observed signals are expressed by

$$\boldsymbol{X}(f) = [\boldsymbol{x}(f,1), \cdots, \boldsymbol{x}(f,2M)]^{\mathrm{T}}, \tag{1}$$

$$\boldsymbol{x}(f,m') = [x(1,f,m'), \cdots, x(N,f,m')]^{\mathrm{T}}, \tag{2}$$

where $x(n,f,m')$ denotes the short-time Fourier transform (STFT) coefficient of the observed signal obtained by the m'th microphone ($1 \leq m' \leq 2M$) at time n ($1 \leq n \leq N$) and frequency f ($1 \leq f \leq F$). The observed signals captured by the mth recording device are denoted by $\boldsymbol{x}(f, 2m-1)$ and $\boldsymbol{x}(f,2m)$. The superscript T denotes non-conjugate transposition. In the following, we first describe the conventional method of noise suppression using transfer-function-gain NMF.

By assuming the time-invariance of the transfer function gain, namely, the amplitude of the transfer function, we express the mixture model in the amplitude domain by

$$\bar{\boldsymbol{X}}(f) \approx \bar{\boldsymbol{A}}(f)\bar{\boldsymbol{S}}(f), \tag{3}$$

$$\bar{\boldsymbol{A}}(f) = \left[\bar{\boldsymbol{a}}_{\mathbf{S}}(f), \bar{\boldsymbol{a}}_{\mathbf{N}_1}(f), \cdots, \bar{\boldsymbol{a}}_{\mathbf{N}_{K-1}}(f)\right], \tag{4}$$

$$\bar{\boldsymbol{a}}_j(f) = [\bar{a}_j(f,1), \cdots, \bar{a}_j(f,2M)]^{\mathrm{T}} \quad (j = \mathbf{S}, \mathbf{N}_k), \tag{5}$$

$$\bar{\boldsymbol{S}}(f) = \left[\bar{\boldsymbol{s}}_{\mathbf{S}}(f), \bar{\boldsymbol{s}}_{\mathbf{N}_1}(f), \cdots, \bar{\boldsymbol{s}}_{\mathbf{N}_{K-1}}(f)\right]^{\mathrm{T}}, \tag{6}$$

$$\bar{\boldsymbol{s}}_j(f) = [\bar{s}_j(1,f), \cdots, \bar{s}_j(N,f)]^{\mathrm{T}} \quad (j = \mathbf{S}, \mathbf{N}_k), \tag{7}$$

where $\bar{a}_j(f,m')$ is the transfer function gain between source j and the m'th microphone at frequency f, and $\bar{s}_j(n,f)$ is the activation of the source j at time n and frequency f, $\bar{\cdot}$ denotes the amplitude of the element. The bold characters $\mathbf{S}$ and

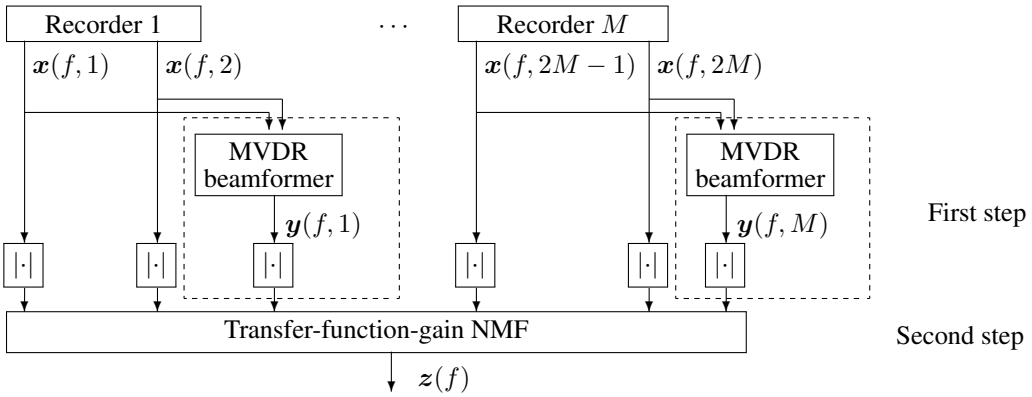


Figure 1: Processing flow of proposed noise suppression method

$\mathbf{N}_k$ denote the target speech and the kth interfering speech ($1 \leq k \leq K-1$), respectively. We can obtain the matrices of the transfer function gain $\bar{\boldsymbol{A}}(f)$ and activations $\bar{\boldsymbol{S}}(f)$ of each source by applying transfer-function-gain NMF. When the m'th microphone is placed closest to the target source, the enhanced target signal $z(n,f)$ is obtained as follws by Wiener filtering:

$$z(n,f) = \lambda(n,f,m')x(n,f,m'), \tag{8}$$

$$\lambda(n,f,m') = \frac{\tilde{x}^2_{\mathbf{S}}(n,f,m')}{\tilde{x}^2_{\mathbf{S}}(n,f,m') + \sum_k \tilde{x}^2_{\mathbf{N}_k}(n,f,m')}, \tag{9}$$

$$\tilde{x}_{\mathbf{S}}(n,f,m') = \tilde{a}_{\mathbf{S}}(f,m')\tilde{s}_{\mathbf{S}}(n,f), \tag{10}$$

$$\tilde{x}_{\mathbf{N}_k}(n,f,m') = \tilde{a}_{\mathbf{N}_k}(f,m')\tilde{s}_{\mathbf{N}_k}(n,f), \tag{11}$$

where $\tilde{\cdot}$ denotes the component estimated by transfer-function-gain NMF, and $\lambda(n,f,m')$ is the Wiener filter that enhances the target signal. (See [4, 5] for details of the noise suppression using transfer-function-gain NMF.)

3. Proposed noise suppression

Figure 1 shows the processing flow of the proposed noise suppression method. The proposed method consists of two steps. In the first step, we apply noise suppression based on phase information to the synchronous signals captured by each recording device. Specifically, in this paper, a maximum variance distortionless response (MVDR) beamformer [6, 7] is employed. The signal of the mth recording device enhanced with an MVDR beamformer, $\boldsymbol{y}(f,m)$, is expressed by

$$y(n,f,m) = \boldsymbol{w}^{\mathrm{H}}(f)\left[x(n,f,2m-1), x(n,f,2m)\right]^{\mathrm{T}}, \tag{12}$$

$$\boldsymbol{y}(f,m) = \left[y(1,f,m), \cdots, y(N,f,m)\right]^{\mathrm{T}}, \tag{13}$$

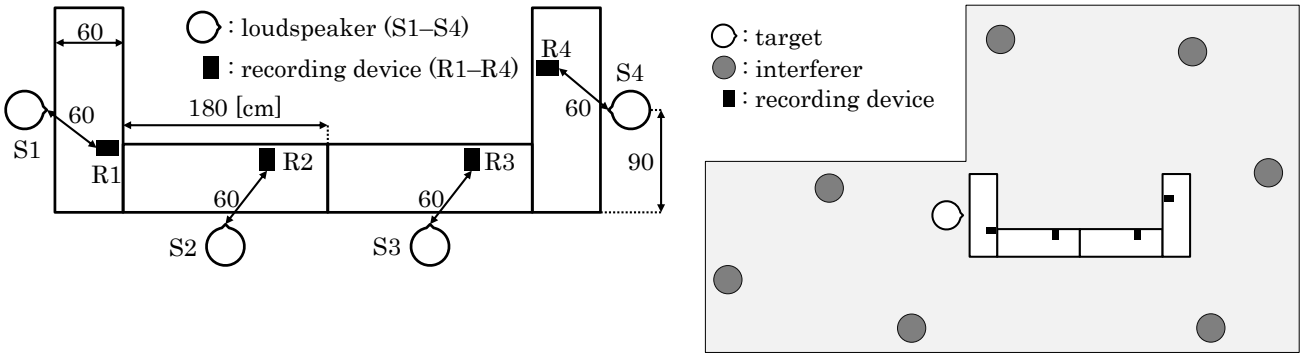
$$\boldsymbol{w}(f) = \left[w(f,2m-1), w(f,2m)\right]^{\mathrm{T}}, \tag{14}$$

where $\boldsymbol{w}(f)$ is the spatial filter estimated by the MVDR beamformer, and the superscript H denotes conjugate transposition. The spatial filter $\boldsymbol{w}(f)$ is computed to minimize the response of nontarget signals such that the target signal is never distorted. We consider that the MVDR beamformer is effective for preprocessing in transfer-function-gain NMF because it can enhance the target signal with a distortionless response. In the second step, we obtain the enhanced target signal by applying transfer-function-gain NMF. The input signals of transfer-function-gain NMF $\bar{\boldsymbol{X}}'(f)$ are expressed by

$$\bar{\boldsymbol{X}}'(f) = \underbrace{\left[\bar{\boldsymbol{x}}(f,1), \cdots, \bar{\boldsymbol{x}}(f,2M), \bar{\boldsymbol{y}}(f,1), \cdots, \bar{\boldsymbol{y}}(f,M)\right]^{\mathrm{T}}}_{3M \text{ channel}}. \tag{15}$$

The amplitude spectrograms of the signals enhanced by beamformers, $\bar{\boldsymbol{y}}(f,m)$, are utilized as the additional input signals in transfer-function-gain NMF. In the proposed approach, we obtain the target signal by applying Wiener filtering (described in the previous section) to $\boldsymbol{y}(f,m)$, where the mth recording device is placed closest to the target source.

When the signals enhanced by beamformers are utilized as the inputs of transfer-function-gain NMF, the SNR of input signals is higher than that when only the signals captured by the recording devices are utilized. Therefore, the proposed approach is expected to improve the performance of noise suppression. Note that both the noise suppression approaches using transfer-function-gain NMF and an MVDR beamformer require single-source sections of the target and interferer signals; hence, the preprocessing of the MVDR beamformer requires no additional information.



Setting A: Interferer in near field

Setting B: Interferer in far field

Figure 2: Arrangement of loudspeakers and recording devices

4. Experimental evaluation

We conducted experiments on noise suppression to evaluate the performance of our proposed approach. The mixture signals were generated by convolving clean speech [8] with the impulse responses measured in a real environment.

4.1 Experimental condition

In this experiment, we recorded time-stretched pulse (TSP) signals in a real environment to generate impulse responses. Figure 2 illustrates the arrangement of the loudspeakers and recording devices. We set up two situations where interferers were placed in the near field (Setting A) and far field (Setting B). For the far-field situation, we applied the model of diffuse noise [5]. In the experiments with Setting A, MVDR beamformers were applied to enhance the closest source to each recording device, regardless of whether the closest source was the target or interferer. The enhanced signals of S1, S2, S3, and S4 were obtained by Wiener filtering of the mixture signals of R1, R2, R3, and R4, respectively. In the experiments with Setting B, MVDR beamformers were applied to enhance the target signal, and the enhanced signal was obtained from the observed signals recorded by device R1, which was placed closest to the target source.

We evaluated the performance of two methods, namely, supervised transfer-function-gain NMF (SNMF) described in [4, 5] and the proposed method (Proposed) in terms of the SDR and SIR [9]. In addition, we evaluated MVDR beamformers applied to the signals recording by the device placed closest to the target source (MVDR) to examine how the performance of the MVDR beamformers affected the performance of the proposed method. Note that an MVDR beamformer constructed by stereo microphones can steer a spatial null in only one direction, which means that each MVDR beamformer suppresses the noise arriving from one direction. Table 1 shows the sampling frequency of each recording device. The other hyperparameters were set as shown in Table 2.

Table 1: Sampling frequency of each recording device

Device a	16,000 Hz
Device b	16,001 Hz
Device c	16,002 Hz
Device d	16,003 Hz

Table 2: Experimental conditions

Frame length of STFT	4,096
Frame overlap	50%
Signal length for noise suppression	10 s
Signal length for supervised training	10 s
Number of NMF iterations	50
Reverberation time (T_{60})	0.6 s

4.2 Evaluation result

Figure 3 and 4 summarize the SDRs and SIRs obtained in the experiments with Setting A and Setting B, respectively. The scores named "Unproc" denote the SDR and SIR of the mixture signals without processing. The results showed that the proposed approach outperformed the conventional noise suppression using transfer-function-gain NMF under both the near-field interferer and the far-field diffuse noise conditions. Moreover, the greater the improvement of the SDR by the MVDR beamformer, the better the noise suppression performance of the proposed method. This indicates that the proposed method improved the performance of noise suppression because the SNR of the input signals was increased by applying the MVDR beamformer to the mixture signals. These results confirmed that the proposed method is capable of improving the noise suppression performance for mixture signals observed by an asynchronous distributed microphone array consisting of recording devices having stereo microphones.

5. Conclusion

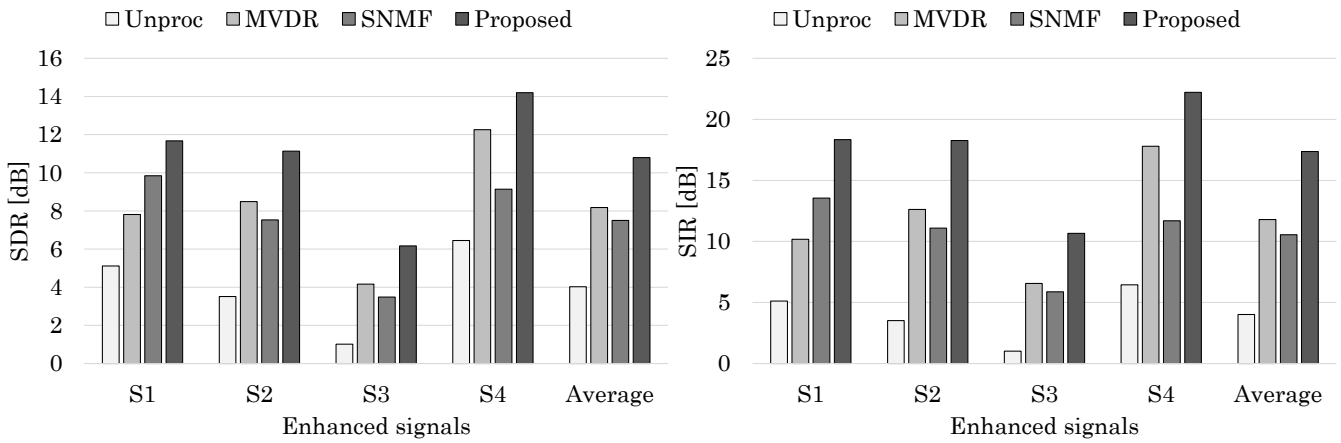


Figure 3: SDR and SIR of Setting A

In this paper, we proposed a new approach to noise suppression using distributed recording devices with stereo microphones. The critical idea is to apply an MVDR beamformer to the synchronous stereo signals captured by each recording device to improve the SNR of the input signals of transfer-function-gain NMF so that transfer-function-gain NMF can more accurately estimate the signals. The signals enhanced by MVDR beamformers are utilized for transfer-function-gain NMF as extra input signals. Experiments using the mixture signals generated by convolving the impulse responses measured in a meeting room showed that the proposed method greatly outperformed the conventional noise suppression using transfer-function-gain NMF.

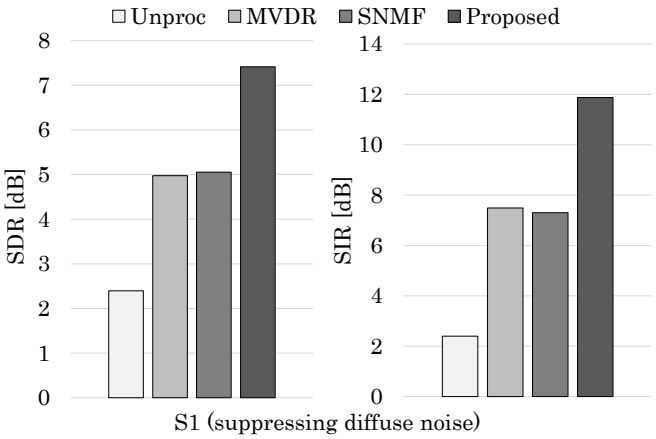


Figure 4: SDR and SIR of Setting B

Acknowledgment

This work was partially supported by SECOM Science and Technology Foundation and the Japan Society for the Promotion of Science (JSPS) KAKENHI through Grant-in-Aid for Scientific Research under Grant 16H01735.

References

[1] E. Robledo-Arnuncio, T. S. Wada, and B.-H. Juang, "On dealing with sampling rate mismatches in blind source separation and acoustic echo cancellation," *Proc. WASPAA*, pp. 34–37, Oct. 2007.

[2] Z. Liu, "Sound source separation with distributed microphone arrays in the presence of clock synchronization errors," *Proc. IWAENC*, pp. 1–4, Sept. 2008.

[3] M. Togami, Y. Kawaguchi, H. Kokubo, and Y. Obuchi, "Acoustic echo suppressor with multichannel semi-blind non-negative matrix factorization," *Proc. APSIPA*, pp. 522–525, Dec. 2010.

[4] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," *Proc. IWAENC*, pp. 204–208, Sept. 2014.

[5] Y. Murase, H. Chiba, N. Ono, S. Miyabe, T. Yamada, and S. Makino, "Diffuse noise suppression with asynchronous microphone array based on amplitude additivity model," *Proc. APSIPA*, pp. 599–603, Dec. 2015.

[6] H. L. Van Trees, Ed., *Optimum Array Processing*, Wiley, May 2002.

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

[8] NTT Advanced Technology Corporation, "Multilingual speech database," 1994.

[9] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, July 2006.