Distant Noise Reduction Based on Multi-delay Noise Model Using Distributed Microphone Array

Yuma Koizumi, Shoichiro Saito, Suehiro Shimauchi, Kazunori Kobayashi, and Noboru Harada
NTT Media Intelligence Laboratories, NTT Corporation, Tokyo, Japan

***Abstract*—We propose a novel framework for reducing distant noise by using a distributed microphone array; reducing noise propagated from a far distance in real-time. Previous studies have revealed that a distributed microphone array with an instantaneous mixing assumption can effectively reduce noise when the target and noise sources are significantly far apart. However, in distant noise reduction, the target and noise sources are not usually instantaneously mixed because the reverberation- and propagation-time from the noise sources to a microphone is longer than the short-time Fourier transform (STFT) length. To express reverberation- and propagation-parameters, we introduce a multi-delay noise model that represents the reverberation-time as a convolution of the transfer-function-gains and the noise sources and the propagation-time as time-frame delays. These parameters are estimated on the basis of the maximum a posteriori (MAP) estimation. Experimental results show that the proposed method outperformed conventional methods in several performance measurements and could reduce distant noise propagated from more than 100 m away in a real-environment.**

***Index Terms*—Distant noise reduction, distributed microphone array, MAP estimation, and transfer function.**

I. Introduction

Noise reduction has been used as a front-end technique of various practical applications such as automatic speech recognition [1], [2]. Recently, it has been applied to emerging applications such as anomaly detection in sound for detecting faulty equipment in a factory [3] and immersive audio field representation for sports broadcasting [4], [5]. Since these emerging applications are used in a large-scale space, noise sources often distribute far from a microphone. For example at a baseball game, cheering-noise in the outfield stands is propagated more than 100 m away from the main-microphone, which is placed close to the home-base to record ball hitting/catching sounds and the umpire's voice. In this study, we aim to build a novel framework to reduce distant noise and demonstrate that our framework reduces noise propagated from far away in real-time.

Microphone arrays are commonly used to reduce noise. Traditional techniques use densely arranged microphones and have focused on the both of the difference of the amplitude and the phase spectrum [6]–[8]. Since these methods are based on the rigorous physical modeling of the wave, practical-use cases in complicated reverberation environments have not been adequately investigated. Meanwhile, distributed microphone array techniques are investigated for noise reduction in real-environments [9]–[13]. These techniques use the microphones close to each source, and a time-frequency (T-F) mask is calculated only from the amplitude-spectrum of the observed signals. The latter approach would be more robust in real environments because the rigorous physical model is not needed and noise is reduced based on only the characteristic of the observed signals.

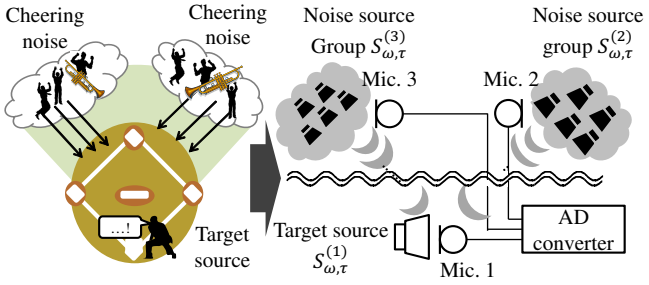

Fig. 1. Observation model of distant noise sources using distributed microphone array.

An assumption of distributed microphone array techniques is that the observation can be modeled as the instantaneous mixing of each source in the time-frequency domain [11]–[13]. This assumption has often been satisfied when a small-scale space is used such as a conference room. However, in distant noise reduction, this assumption is not valid for the following reasons. When the reverberation-time, *i.e.*, the impulse response length, is longer than the short-time Fourier transform (STFT) length, the reverberation cannot be expressed by multiplying a transfer-function to a sound source in a single STFT-frame. In addition, when propagation-time is longer than the STFT length, a sound source observed by each microphone distributes in different STFT-frames. If the impulse response and the propagation-time are known, the above problem can be solved by using a sufficiently long STFT length. However, the impulse response and the propagation-time are unknown in practice.

In this study, we propose a framework to reduce distant noise by using a distributed microphone array. First, we introduce a multi-delay noise model on the basis of the strategy of the "Multidelay block frequency domain adaptive filter" (MDF) [14], to express long reverberation- and propagation-time while using a short STFT length. In the multi-delay noise model, long reverberation-time is represented as a convolution of the transfer-function-gains and the noise sources, and the propagation-time is compensated by frame-shift represented as time-frame delays. Thus, the parameters of long reverberation- and propagation-time to be estimated become transfer-function-gains and time-frame-delays. Then, we ex-

ISBN 978-90-827970-1-5

tend the multi-delay noise model to a probabilistic model and estimate these parameters on the basis of the maximum a posteriori (MAP) estimation.

The rest of this paper is organized as follows. Section II briefly introduces the conventional noise reduction. Then, in Section III, a multi-delay noise model and a distant noise reduction framework are proposed. After investigating the performance of the proposed method in Section IV, we conclude this paper in Section V.

II. Conventional method

A. Noise reduction using time-frequency mask

Let us consider the problem of estimating a target source $S^{(1)}_{\omega,\tau}$ from an observed signal recorded using M microphones. We define the microphone placed close to the target source as microphone number $m = 1$, and its observation is written as

$$X^{(1)}_{\omega,\tau} = S^{(1)}_{\omega,\tau} + N^{(1)}_{\omega,\tau}, \tag{1}$$

where $N^{(1)}_{\omega,\tau}$ is the noise propagated to the microphone $m = 1$ consisted of $I-1$ noise sources as $S^{(i)}_{\omega,\tau}$ $(i \in \{2, ..., I\})$, $\omega \in \{1, 2, ..., \Omega\}$ and $\tau \in \{1, 2, ..., T\}$ denote the frequency and time indices, respectively. To simplify of notation, we assume $S^{(1)}_{\omega,\tau}$ includes the transfer function from the target source to the microphone $m = 1$. Hereafter, we call the microphone $m = 1$ the "*main-microphone*."

In noise reduction using T-F masks, the output signal $\hat{S}_{\omega,\tau}$ is obtained by multiplying a T-F mask to $X^{(1)}_{\omega,\tau}$ as

$$\hat{S}_{\omega,\tau} = G_{\omega,\tau} X^{(1)}_{\omega,\tau}, \tag{2}$$

where $G_{\omega,\tau}$ is a T-F mask such as the ideal-ratio-mask (IRM) [15] defined as

$$G_{\omega,\tau} = \frac{|S^{(1)}_{\omega,\tau}|}{|S^{(1)}_{\omega,\tau}| + |N^{(1)}_{\omega,\tau}|} \approx \frac{|X^{(1)}_{\omega,\tau}| - |N^{(1)}_{\omega,\tau}|}{|X^{(1)}_{\omega,\tau}|}. \tag{3}$$

From (3), to calculate $G_{\omega,\tau}$, we need to estimate the amplitude-spectrum of noise $|N^{(1)}_{\omega,\tau}|$ from the observed signals.

B. T-F mask design based on instantaneous mixing

Typical noise reduction techniques represent $X^{(m)}_{\omega,\tau}$ as instantaneous mixing of I sound sources, thus $N^{(1)}_{\omega,\tau}$ can be represented as $N^{(1)}_{\omega,\tau} = \sum_{i=2}^{I} A^{(1,i)}_{\omega} S^{(i)}_{\omega,\tau}$, where $A^{(1,i)}_{\omega}$ is the transfer-function from i-th source to the target microphone. Moreover, some literature on distributed microphone arrays assumes the additivity of the amplitude-spectrum [11]–[13]. Thus, the amplitude spectrum of the observation of the target microphone can be expressed by the product sum of the amplitude-spectrum omitting the phase as

$$|X^{(1)}_{\omega,\tau}| \approx |S^{(1)}_{\omega,\tau}| + \sum_{i=2}^{I} |A^{(1,i)}_{\omega}||S^{(i)}_{\omega,\tau}|. \tag{4}$$

Therefore, to design a T-F mask, we need to estimate $|A^{(1,i)}_{\omega}|$ and $|S^{(i)}_{\omega,\tau}|$. To estimate these parameters, previous studies using distributed microphone arrays [11]–[13] have adopted a power-spectrum-density (PSD)-estimation-in-beamspace method [16] and/or the transfer-function-gain non-negative matrix factorization (NMF)

III. Proposed method

When the target source and the noise sources are far apart, they are not instantaneously mixed. In the following sections, guided by the strategy of MDF, the observed signal of distant noise sources is represented in a multi-delay noise model in Section III-A. Then, the multi-delay noise model is extended to a probabilistic model in Section III-B and the parameter estimation procedure is detailed in Section III-C.

A. Multi-delay noise model for distant noise reduction

In this section, we model an observed signal of distant noise sources on the basis of the following assumptions:

1) The positions of each source and microphone are fixed.
2) Densely placed noise sources are regarded together as one noise source group $S^{(i)}_{\omega,\tau}$ as shown in Fig. 1.
3) The number of noise groups is $I-1 = M-1$, and each microphone is placed close to each noise group. Then, we assume that $|X^{(m)}_{\omega,\tau}| \approx |S^{(m)}_{\omega,\tau}|$ holds for microphone $m = \{2, ..., M\}$.

First, to model long reverberation- and propagation-time, we adopt the strategy of the MDF [14]. In the MDF, to express a long reverberation-time, the transfer-function is separated into $(K+1)$-blocks, and the observed signal is expressed as the convolution of the transfer-functions and each source. To represent a propagation-time, we extend the MDF by using a time-frame-delay D_m, and $N^{(1)}_{\omega,\tau}$ is expressed as

$$N^{(1)}_{\omega,\tau} = \sum_{m=2}^{M} \sum_{k=0}^{K} A^{(1,m)}_{\omega,k} S^{(m)}_{\omega,\tau-D_m-k}, \tag{5}$$

where $A^{(1,m)}_{\omega,k}$ is k-th block transfer-function from m-th noise source group to the main-microphone, and $D_m \in \mathbb{N}_+$ is a time-frame-delay corresponding to the propagation-time from the m-th noise source group to the main-microphone. Here we assume the additivity of the amplitude-spectrum likewise (4). Then, by replacing $|S^{(m)}_{\omega,\tau}|$ to $|X^{(m)}_{\omega,\tau}|$ in accordance with assumption 3), (5) can be rewritten as

$$|\hat{N}^{(1)}_{\omega,\tau}| \approx \sum_{m=2}^{M} \sum_{k=0}^{K} a^{(m)}_{\omega,k} |X^{(m)}_{\omega,\tau-D_m-k}|, \tag{6}$$

where $a^{(m)}_{\omega,k} = |A^{(1,m)}_{\omega,k}|$. Hereafter, we call (6) a "*multi-delay noise model*." The unknown parameters in the multi-delay noise model are transfer-function-gains and time-frame-delays

$$\mathbf{a} := \{a^{(m)}_{\omega,k} | m = 2, ..., M, \omega = 1, ..., \Omega, k = 0, ..., K\},$$
$$\mathbf{D} := \{D_m | m = 2, ..., M\}.$$

If these parameters have been estimated in advance, distant noise can be reduced by (3) and (6) as shown in Fig. 2. Thus, we define the problem of reducing distant noise as a problem of estimating $\Theta := \{\mathbf{a}, \mathbf{D}\}$.

ISBN 978-90-827970-1-5 © EURASIP 2018

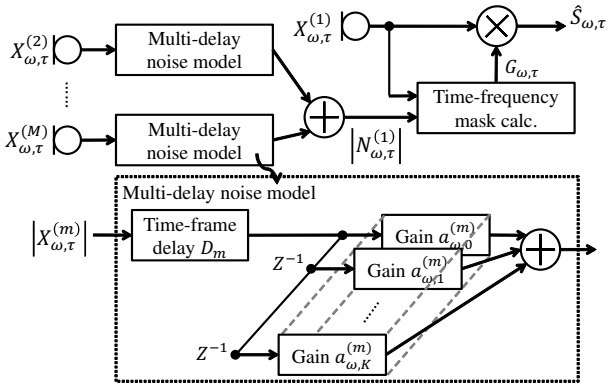

Fig. 2. Multi-delay noise model and distant noise reduction framework.

B. Probabilistic extension of multi-delay noise model

To estimate Θ from training data in advance, we adopt a machine learning approach. We now extend the multi-delay noise model as a probabilistic model, and Θ is trained so as to maximize an objective function.

Since the transfer-function-gains and time-frame-delay are physical variables intrinsically, estimation accuracy would be improved by designing a probabilistic model to reflect its physical characteristics. Thus, we define prior distributions of each parameter to incorporate the physical characteristics to a probabilistic model. Then, the posterior probability of Θ is used as the objective function, and Θ is estimated as

$$\Theta \leftarrow \arg\max_{\Theta} \mathcal{J}(\Theta), \tag{7}$$

$$\begin{aligned}\mathcal{J}(\Theta) &= \ln p(\Theta|\mathbf{X}) = \ln \frac{p(\mathbf{X}^{(1)}|\Theta,\mathbf{X}^R)p(\Theta)p(\mathbf{X}^R)}{p(\mathbf{X})}, \\ &\stackrel{c}{=} \ln p(\mathbf{X}^{(1)}|\Theta,\mathbf{X}^R) + \ln p(\Theta),\end{aligned} \tag{8}$$

where $\mathbf{X} := \{\mathbf{X}^{(m)}|m = 1,...,M\}$, $\mathbf{X}^R := \{\mathbf{X}^{(m)}|m = 2,...,M\}$, and $\mathbf{X}^{(m)}$ includes $|X^{(m)}_{\omega,\tau}|$ for all ω and τ. Additionally, since any prior information of $\mathbf{X}$ and $\mathbf{X}^R$ cannot be obtained, we use the uninformative prior for $p(\mathbf{X})$ and $p(\mathbf{X}^R)$, and omit it for simplify of the optimization. Here we assume $\mathbf{a}$ and $\mathbf{D}$ are independent, and then (8) can be written as

$$\mathcal{J}(\Theta) = \ln p(\mathbf{X}^{(1)}|\mathbf{a},\mathbf{D},\mathbf{X}^R) + \ln p(\mathbf{a}) + \ln p(\mathbf{D}). \tag{9}$$

To calculate (9), we model each distribution. First, the target source is assumed to be a temporally sparse event such as a hitting sound and the umpire's voice. Then, since $|X^{(1)}_{\omega,\tau}| = |N^{(1)}_{\omega,\tau}|$ holds for most time-frames, we now assume the following probabilistic model

$$\ln|X^{(1)}_{\omega,\tau}| = \ln|\hat{N}^{(1)}_{\omega,\tau}| + \epsilon_{\omega,\tau}. \tag{10}$$

Assuming the error value $\epsilon_{\omega,\tau}$ follows the Gaussian distribution $\mathcal{N}\left(\epsilon_{\omega,\tau}\left|0,\sigma^2\right.\right)$, the likelihood function of an observed signal is written as

$$p(\mathbf{X}^{(1)}|\mathbf{a},\mathbf{D},\mathbf{X}^R) := \prod_{\omega=1}^{\Omega}\prod_{\tau=1}^{T}\mathcal{N}\left(\ln|X^{(1)}_{\omega,\tau}|\left|\ln|\hat{N}^{(1)}_{\omega,\tau}|,\sigma^2\right.\right). \tag{11}$$

Next, we define the prior distribution of $\mathbf{a}$. Since transfer-function-gains are non-negative continuous variables, its prior distribution should be modeled to satisfy its constraint. As an implementation to satisfy the constraint, $p(\mathbf{a})$ is defined by an exponential distribution as

$$p(\mathbf{a}) := \prod_{\omega=1}^{\Omega}\prod_{m=2}^{M}\prod_{k=0}^{K}\frac{1}{\alpha_{\omega,k}}\exp\left\{-\frac{a^{(m)}_{\omega,k}}{\alpha_{\omega,k}}\right\}, \tag{12}$$

where $a^{(m)}_{\omega,k} \geq 0$. Since the transfer-function-gain has a physical characteristic of exponentially decaying [17], the hyperparameter $\alpha_{\omega,k}$ is set as $(k+1)^{-1}$. Finally, we define the prior distribution of $\mathbf{D}$. When the distance between the main-microphone and m-th microphone ϕ_m [m] can be estimated approximately, the time-frame-delay can also be estimated approximately from ϕ_m as

$$Q_m = \text{floor}\left\{\frac{\phi_m}{C}\cdot\frac{f_s}{f_{\text{shift}}}\right\}, \tag{13}$$

where C [m/s] is the sound velocity, f_s [sample/sec] is the sampling rate, f_{shift} [sample/frame] is the shift length of STFT, and floor$\{\cdot\}$ is the flooring function. In addition, D_m is a non-negative integer. Therefore, we define $p(\mathbf{D})$ as a Poisson distribution with parameter Q_m calculated from ϕ_m as

$$p(\mathbf{D}) := \prod_{m=2}^{M}\frac{Q_m^{D_m}\exp\{-Q_m\}}{D_m!}, \tag{14}$$

where ! denotes the factorial operator.

C. Parameter estimation procedure

It is difficult to obtain the global optima analytically because this problem is a simultaneous optimization for non-negative continuous and integer variables $a^{(m)}_{\omega,k}$ and D_m. To obtain a local optima, we adopt an alternately repeat optimization of a proximal gradient method for $a^{(m)}_{\omega,k}$ and grid-search method for D_m as shown in **Algorithm 1** . Specifically, first $a^{(m)}_{\omega,k}$ is optimized by repeating the following procedure I_{ter} times by a proximal gradient method. Next, D_m is optimized so as to maximize $\mathcal{J}(\Theta)$ by a grid-search algorithm.

To simply calculate $\nabla a^{(m)}_{\omega,k}\mathcal{J}(\Theta)$, we use $\sigma^2 = 1$ in (11). Then, the log-likelihood function and the log prior distribution of $\mathbf{a}$ are written as

$$\ln p(\mathbf{X}^{(1)}|\mathbf{a},\mathbf{D},\mathbf{X}^R) \stackrel{c}{=} \sum_{\omega=1}^{\Omega}\sum_{\tau=1}^{T}-\frac{1}{2}\left(\ln\frac{|X^{(1)}_{\omega,\tau}|}{|\hat{N}^{(1)}_{\omega,\tau}|}\right)^2, \tag{15}$$

$$\ln p(\mathbf{a}) \stackrel{c}{=} \sum_{\omega=1}^{\Omega}\sum_{m=2}^{M}\sum_{k=0}^{K}-\frac{a^{(m)}_{\omega,k}}{\alpha_{\omega,k}}. \tag{16}$$

Thus, $\nabla a^{(m)}_{\omega,k}\mathcal{J}(\Theta)$ is calculated as

$$\nabla a^{(m)}_{\omega,k}\mathcal{J}(\Theta) = \sum_{t=1}^{T}\frac{\psi^{(m,D_m,k)}_{\omega,\tau}\ln\left(\frac{|X^{(1)}_{\omega,\tau}|}{|\hat{N}^{(1)}_{\omega,\tau}|}\right)}{|\hat{N}^{(1)}_{\omega,\tau}|} - \frac{1}{\alpha_{\omega,k}}, \tag{17}$$

where we write $\psi^{(m,D_m,k)}_{\omega,\tau} = |X^{(m)}_{\omega,\tau-D_m-k}|$ due to limitations of space. Moreover, to adjust the $\alpha_{\omega,k}$ in accordance with the

ISBN 978-90-827970-1-5 © EURASIP 2018

amplitude-level of the observation on each microphone, we calculate $\alpha_{\omega,k} = \gamma_\omega (k+1)^{-1}$ where

$$\gamma_\omega = \frac{1}{\sum_{k=0}^{K}(k+1)^{-1}} \cdot \frac{1}{T}\sum_{\tau=1}^{T} \frac{|X^{(1)}_{\omega,\tau}|}{\sum_{m=2}^{M}|X^{(m)}_{\omega,\tau-Q_m-k}|}. \quad (18)$$

Algorithm 1 Optimization for multi-delay noise model. λ is step-size of gradient method.

Input: $\mathbf{X}$
Output: $\mathbf{a}, \mathbf{D}$
Initialize $a^{(m)}_{\omega,k} = \alpha_{\omega,k}$ and $D_m = Q_m$.
while *until algorithm convergence* **do**
  repeat
    $a^m_{\omega,k} \leftarrow a^m_{\omega,k} + \lambda \nabla_{a^m_{\omega,k}} \mathcal{J}(\Theta)$ for all m, ω, k.
    $a^m_{\omega,k} \leftarrow \max(0, a^m_{\omega,k})$ for all m, ω, k.
  until I_{ter} times
  $D_m \leftarrow \arg\max_{D_m} \mathcal{J}(\Theta)$ for all m.
end while

IV. Experiments

A. Experimental condition

We conducted objective experiments and a verification experiment to evaluate the performance of the proposed method (`Prop`). For comparison, we used a noise reduction method using a distributed microphone array [12] (`Conv`). Although this conventional method assumes the amplitude-spectra of noise sources are unknown, we model the noise as $|N^{(1)}_{\omega,\tau}| \approx \sum_{m=2}^{M} |A^{(1,m)}_{\omega}||X^{(m)}_{\omega,\tau}|$ and estimate only transfer-function-gain. To evaluate the validity of the multi-delay noise model (6), the conventional method with the ground-truth of the time-frame-delay $\mathcal{D}_m$ (`Conv-FD`) was also compared, *i.e.*, $|N^{(1)}_{\omega,\tau}| \approx \sum_{m=2}^{M} |A^{(1,m)}_{\omega}||X^{(m)}_{\omega,\tau-\mathcal{D}_m}|$.

Since it is difficult to collect training/test data in a large-scale space, the objective experiments were conducted in a simulation environment as shown in Fig. 3. To simulate a large-scale room, the "room impulse response (RIR) generator" [18] was used. The parameters for the RIR simulation were as follows: the sound velocity was 340.0 [m/s], the sampling rate was 16 kHz, the reverberation-time (RT_{60}) was 1.0 [s], the reflection order was 10, and the microphone type was omnidirectional. Each noise group was propagated from two loudspeakers: one emitted a vocal source, and the other emitted a drum source. Ten pieces of music from "The Mixing Secret Dataset 100 (MSD100)" [19] were used as the training dataset of noise sources. As the test datasets, a Japanese speech database consisting of 200 utterances spoken by 2 males and 2 females from the ATR Japanese speech database was used for a target source dataset, and 6 pieces of music in MSD100 were used for noise sources. The noisy signals were formed by mixing RIR convolved speech utterances with the RIR convolved noises at signal-to-noise ratio (SNR) levels of -12, -6, and 0 dB.

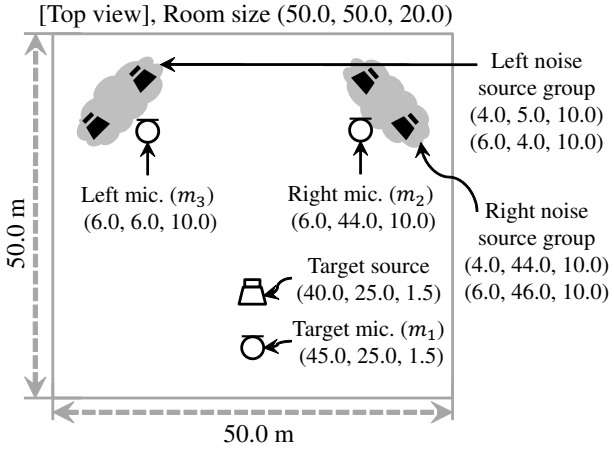

Fig. 3. Arrangement of simulated microphones and sound sources. Each (x, y, z) coordinate [m] denotes position of microphone and sound source.

The frame size of the STFT was 512 samples (= 32 ms), and the frame was shifted by 256 samples (= 16 ms). Approximated distances were set to $\phi_2 = \phi_3 = 40$ m (*i.e.*, $Q_m = 7$). Other parameters were defined practically as $\lambda = 10^{-5}$, $K = 10$, and $I_{\text{ter}} = 20$. The ground-truth of the time-frame-delay was $\mathcal{D}_m = 8$.

B. Objective experiments

The proposed method was compared with the conventional methods in terms of distant noise reduction by using three objective measurements: the signal-to-distortion ratio (SDR), the short-time intelligibility measure (STOI) [20], and the perceptual evaluation of speech quality (PESQ). The "BSS-Eval toolbox [21]" was used to calculate the SDR.

Table I shows the evaluation results. All scores of the proposed method were always higher than those of the conventional methods irrespective of the input SNR conditions. Since `Conv-FD`, which is an instantaneous mixing model involving ground-truth time-frame-delay $\mathcal{D}_m$, has higher scores than `Conv` in all measurements, it is necessary to consider time-frame-delay in distant noise reduction. In addition, all scores of `Prop`, which models the distant noise by convolutional mixing, were higher than those of `Conv-FD`, which models it by instantaneous mixing. According to these results, the proposed multi-delay noise model and noise reduction framework effectively reduced distant noise.

C. Verification experiment in real-environment

To test whether the proposed method works in a real-environment, we tested the proposed method in a baseball stadium. The target sources were ball hitting/catching sounds and the umpire's voice that came from close to the home-base. The noise was cheering noise in left- and right-outfield stands including cheering voice/whistle/drum noise. We placed three microphones close to the home-base $m = 1$ and in both outfield stands $m = 2$ and $m = 3$.

Fig. 4 (a) and (b) show spectrograms of $X^{(1)}_{\omega,\tau}$ and $X^{(3)}_{\omega,\tau}$, respectively. As we can see in these spectrograms, a time-frame-

ISBN 978-90-827970-1-5

TABLE I
EXPERIMENTAL RESULTS (AVERAGE ± STANDARD DEVIATION).

Method	SDR [dB]	STOI [%]	PESQ
Input SNR: -12 dB			
Conv	-6.62 ± 2.12	42.02 ± 6.09	1.04 ± 0.55
Conv-FD	-0.80 ± 2.24	53.81 ± 6.37	1.39 ± 0.33
Prop	$\mathbf{2.88 \pm 2.36}$	$\mathbf{60.64 \pm 6.70}$	$\mathbf{1.57 \pm 0.21}$
Input SNR: -6 dB			
Conv	-0.03 ± 2.15	57.98 ± 6.51	1.45 ± 0.34
Conv-FD	5.40 ± 2.11	69.03 ± 6.78	1.81 ± 0.18
Prop	$\mathbf{7.86 \pm 1.98}$	$\mathbf{74.31 \pm 7.04}$	$\mathbf{2.06 \pm 0.17}$
Input SNR: 0 dB			
Conv	6.16 ± 1.52	72.97 ± 7.20	1.91 ± 0.20
Conv-FD	10.45 ± 1.75	80.90 ± 7.46	2.25 ± 0.16
Prop	$\mathbf{11.51 \pm 1.59}$	$\mathbf{83.88 \pm 7.34}$	$\mathbf{2.51 \pm 0.15}$

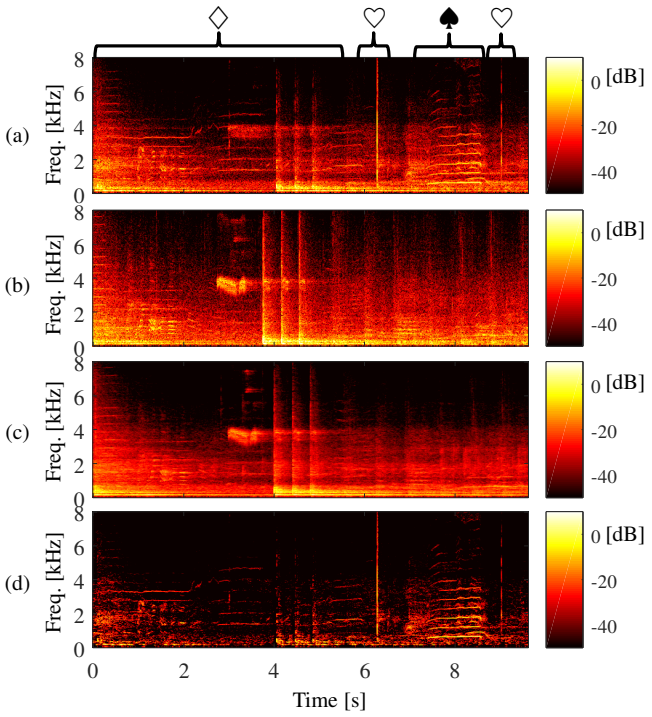

Fig. 4. Spectrograms of (a) observed signal at home-base $X^{(1)}_{\omega,\tau}$, (b) observed signal at left-outfield stand $X^{(3)}_{\omega,\tau}$, (c) estimated noise $\hat{N}^{(1)}_{\omega,\tau}$, and (d) estimated target source $\hat{S}_{\omega,\tau}$. Each event ♢, ♡, and ♠ denotes cheering voice/whistle/drum noise, catching sound, and umpire's voice, respectively.

delay and reverberation occurred, thus instantaneous mixing cannot be assumed. Fig. 4 (c) and (d) show the estimated noise $\hat{N}^{(1)}_{\omega,\tau}$ and output $\hat{S}_{\omega,\tau}$, respectively. The former shows that long reverberation- and propagation-time are adjusted by the multi-delay noise model, and the latter shows that the distant noise is reduced. Results of this verification experiment suggest that the proposed method effectively reduces distant noise under practical conditions.

V. CONCLUSIONS

In this study, we proposed a framework to reduce distant noise by using a distributed microphone array. First, we introduced distant noise by a multi-delay noise model, to represent long reverberation- and propagation-time. Then, the model was extended as a probabilistic model, and its parameters were estimated on the basis of the maximum a posterior (MAP) estimation. Experimental results showed that the proposed method outperformed conventional methods in several performance measurements and could reduce distant noise propagated from more than 100 m away in a real-environment. Thus, it can be concluded that the proposed method is effective for distant noise reduction.

REFERENCES

[1] T. Yoshioka, N. Ito, M. Delcreix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Dabian, M. Espi, T. Higuchi, A. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,," in *Proc.* ASRU, 2015.

[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 Trans. ASLP, 2017.

[3] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing Acoustic Feature Extractor for Anomalous Sound Detection Based on Neyman-Pearson Lemma," in *Proc.* EUSIPCO, 2017.

[4] R. Oldfield, B. Shirley, and J. Spille, "Object-based Audio for Interactive Football Broadcast," Multimed. Tools & Appl., 2015.

[5] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and H. Ohmuro, "Informative Acoustic Feature Selection to Maximize Mutual Information for Collecting Target Sources," IEEE Trans. ASLP, 2017.

[6] J. G. Ryan, R. A. Goubran, "Near-field beamforming for microphone arrays," *in Proc.* ICASSP, 1997.

[7] Y. R. Zheng, R. A. Goubran and M. El-Tanany, "Robust near-field adaptive beamforming with distance discrimination," IEEE Trans. SAP, 2004.

[8] E. Fisher and B. Rafaely, "Near-field spherical microphone array processing with radial filtering," IEEE Trans. ASLP, 2011.

[9] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *Proc.* WASPAA, 2009.

[10] S. Miyabe, N. Ono, and S. Makino, "Optimizing frame analysis with non-integer shift for sampling mismatch compensation of long recording," in *Proc.* WASPAA, 2013.

[11] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with non-negative matrix factorization for asynchronous distributed recording," in *Proc.* IWAENC, 2014.

[12] T. Kako, K. Niwa, K. Kobayashi, and H. Ohmuro, "Wiener filter design by estimating sensitivities between distributed asynchronous microphones and sound sources," in *Proc.* WASPAA, 2015.

[13] Y. Matsui, S. Makino, N. Ono, and T. Yamada, "Multiple Far Noise Suppression in a Real Environment Using Transfer-Function-Gain NMF," in *Proc.* EUSIPCO, 2017.

[14] J. S. Soo and K. K. Pang, "Multidelay Block Frequency Domain Adaptive Filter," IEEE Trans. ASSP, 1990.

[15] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015.

[16] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," IEEE Trans. ASLP, 2013.

[17] S. Makino and Y. Kaneda, "Exponentially Weighted Step-Size Projection Algorithm for Acoustic Echo Cancellers," IEICE Trans. Fundamentals 1992.

[18] E. A. P. Habets, "Room impulse response generator," https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator/ (Jan. 2018, accessed).

[19] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 Signal Separation Evaluation Campaign," in *Proc.* LVA/ICA, 2015.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. ASLP, 2011.

[21] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Trans. ASLP, 2006.

ISBN 978-90-827970-1-5 © EURASIP 2018