PINPOINT EXTRACTION OF DISTANT SOUND SOURCE BASED ON DNN MAPPING FROM MULTIPLE BEAMFORMING OUTPUTS TO PRIOR SNR

Kenta Niwa[1]*, Yuma Koizumi*[1]*, Tomoko Kawase*[1]*, Kazunori Kobayashi*[1]*, and Yusuke Hioka*[2]

[1]: NTT Media Intelligence Laboratories, NTT Corporation, Japan
[2]: Department of Mechanical Engineering, University of Auckland, New Zealand

ABSTRACT

We propose a method for estimating the prior signal-to-noise ratio (SNR), which is used for calculating the Wiener filter for distant sound source extraction, from output signals of beamforming using statistical mapping based on the deep neural network (DNN). Since informative features to estimate the prior SNR are included in multiple beamforming outputs, the SNR can be accurately estimated by this mapping using the DNN. The proposed method was applied to a large microphone array, the design of which was optimized to form effective directivity patterns to extract distant sound sources. Experimental results proved that the target source was clearly extracted with the proposed method.

Index Terms— source enhancement, microphone array, beamforming, signal-to-noise ratio (SNR), deep neural network (DNN)

1. INTRODUCTION

Signal processing using microphone arrays [1, 2] has been used to extract a target source in noisy environments. Most studies on microphone arrays have been focused on problems in which the target sound source is located within a range of a few meters from the microphone array. However, there are quite a few practical situations in which the target source needs to be zoomed in from a remote location, such as the recording of an athlete's voice on a playing field in a stadium. The ultimate goal with this study was to extract the target source surrounded by various noise sources from a remote location.

Most conventional studies on microphone arrays have been focused on the design of beamforming; however, several researchers have focused on array structures, e.g., minimum redundancy [1] and rigid spherical array [10]–[13], for improving the performance of the sound source enhancement. We previously investigated the properties of the optimum spatial correlation matrix [1, 2] to segregate sound sources. The findings from this study were implemented in a microphone array specifically designed to follow [14]–[16]. With this specific microphone array, the effective cues for segregating the target source from other noise sources are included in the observed signals; thus, sharp directivity to enhance the target source can be formed over a broad frequency range, even if an ordinary beamforming method is adopted. However, the noise reduction capability of the previous method is limited; thus, it is still difficult to extract the target source if it is severely contaminated by noise sources.

Applying the Wiener filter as a post-filter of beamforming is also effective for boosting noise reduction performance [3]–[9]. To calculate the Wiener filter, it is necessary to estimate the SNR at the beamforming output. The power spectral density (PSD) estimation method using multiple beamforming outputs (e.g. PSD estimation in beamspace [8, 9]) was developed to estimate the PSD of both the target source and noise for calculating the signal-to-noise ratio (SNR) of microphone array observation. However, the required prior SNR for calculating the Wiener filter should be that at the beamforming output but not at the microphone array observation. Thus, an alternative method for estimating the prior SNR is required to achieve better sound source enhancement. To this end, we propose a method that uses the DNN [17]–[23], which maps the output signals of the multiple beamforming outputs to the prior SNR. Thanks to the recent progress in research, the DNN can provide an accurate mapping between two pieces of information if informative features are available as its input. For this study, we attempted to use the output of multiple beamforming outputs as the input of the DNN and obtain a more accurate prior SNR for calculating the Wiener filter. The Wiener filter was then applied to the output of a beamformer applied to the microphone array with the optimal design of its spatial correlation matrix.

2. MICROPHONE ARRAY FOR INCREASING MUTUAL INFORMATION OF MIMO

2.1. Observation model

Let us assume that K source signals are observed using M microphones. This situation is regarded as a multiple-input and multiple-output (MIMO) system. When the transfer function between the k-th source and m-th microphone is denoted as $A_{m,k,\omega}$, the observed signals $\mathbf{x}_{\omega,\tau}$ are expressed by

$$\mathbf{x}_{\omega,\tau} = \mathbf{A}_{\omega}\mathbf{s}_{\omega,\tau} + \mathbf{n}_{\omega,\tau}, \tag{1}$$

where each vector or matrix in (1) is defined as

$$\mathbf{x}_{\omega,\tau} = [X_{1,\omega,\tau}, \ldots, X_{M,\omega,\tau}]^{\mathrm{T}}, \tag{2}$$
$$\mathbf{A}_{\omega} = [\mathbf{a}_{1,\omega}, \ldots, \mathbf{a}_{K,\omega}], \tag{3}$$
$$\mathbf{a}_{k,\omega} = [A_{1,k,\omega}, \ldots, A_{M,k,\omega}]^{\mathrm{T}}, \tag{4}$$
$$\mathbf{s}_{\omega,\tau} = [S_{1,\omega,\tau}, \ldots, S_{K,\omega,\tau}]^{\mathrm{T}}, \tag{5}$$
$$\mathbf{n}_{\omega,\tau} = [N_{1,\omega,\tau}, \ldots, N_{M,\omega,\tau}]^{\mathrm{T}}. \tag{6}$$

with the transposition being denoted as $^{\mathrm{T}}$. The k-th source signal is represented as $S_{k,\omega,\tau}$ and the background noise received by the m-th microphone is denoted as $N_{m,\omega,\tau}$, where ω and τ denote the index of frequency and frame, respectively.

The sound sources and noises are assumed to be uncorrelated;

$$\mathbf{R}_{\mathrm{S},\omega} = \left\langle \mathbf{s}_{\omega,\tau}\mathbf{s}_{\omega,\tau}^{\mathrm{H}} \right\rangle = \sigma_{\mathrm{S},\omega}^2\mathbf{I}_K, \tag{7}$$

$$\mathbf{R}_{\mathrm{N},\omega} = \left\langle \mathbf{n}_{\omega,\tau}\mathbf{n}_{\omega,\tau}^{\mathrm{H}} \right\rangle = \sigma_{\mathrm{N},\omega}^2\mathbf{I}_M, \tag{8}$$

where $\langle\cdot\rangle$ and $^{\mathrm{H}}$ denote the expectation and Hermitian conjugate, respectively. Then, the spatial correlation matrix [1, 2] is given by

$$\mathbf{R}_{\mathrm{X},\omega}=\left\langle \mathbf{x}_{\omega,\tau}\mathbf{x}_{\omega,\tau}^{\mathrm{H}}\right\rangle=\mathbf{A}_{\omega}\mathbf{R}_{\mathrm{S},\omega}\mathbf{A}_{\omega}^{\mathrm{H}}+\mathbf{R}_{\mathrm{N},\omega}=\sigma_{\mathrm{S},\omega}^{2}\mathbf{R}_{\mathrm{A},\omega}+\sigma_{\mathrm{N},\omega}^{2}\mathbf{I}_{M}, \tag{9}$$

where $\mathbf{R}_{\mathrm{A},\omega}$ is composed of the received power $\sigma_{\mathrm{A},\omega}^{2}$, assumed to be normalized a priori, and the cross-correlation between microphones $\Gamma_{i,j,\omega}$, which is expressed as

$$\mathbf{R}_{\mathrm{A},\omega}=\mathbf{A}_{\omega}\mathbf{A}_{\omega}^{\mathrm{H}}=\begin{bmatrix}\sigma_{\mathrm{A},\omega}^{2} & \Gamma_{1,2,\omega} & \cdots & \Gamma_{1,M,\omega}\\ \Gamma_{2,1,\omega} & \sigma_{\mathrm{A},\omega}^{2} & \cdots & \Gamma_{2,M,\omega}\\ \vdots & \vdots & \ddots & \vdots\\ \Gamma_{M,1,\omega} & \Gamma_{M,2,\omega} & \cdots & \sigma_{\mathrm{A},\omega}^{2}\end{bmatrix}. \tag{10}$$

2.2. Array structure design to increase mutual information of MIMO using parabolic reflectors

In our previous study [16], we (i) derived a model of the optimum structure of the spatial correlation matrix to segregate source signals and (ii) tested the performance of sound source enhancement by building a microphone array that follows the model. To measure how much information about $\mathbf{s}_{\omega,\tau}$ is included in $\mathbf{x}_{\omega,\tau}$, the mutual information between $\mathbf{s}_{\omega,\tau}$ and $\mathbf{x}_{\omega,\tau}$, $I_{\mathbf{s};\mathbf{x}}$ was defined as

$$I_{\mathbf{s};\mathbf{x}}=H_{\mathbf{s}}-H_{\mathbf{s}|\mathbf{x}}, \tag{11}$$

where $H_{\mathbf{s}}$ and $H_{\mathbf{s}|\mathbf{x}}$ denote the entropy of the transmitted information and propagation loss, respectively. If the structure of $\mathbf{A}_{\omega}$ is irregular or the received background noise level is substantially high, $H_{\mathbf{s}|\mathbf{x}}$ will increase. To investigate the structure of the spatial correlation matrix which maximizes $I_{\mathbf{s};\mathbf{x}}$, the channel capacity of the MIMO system, denoted as C_{ω}, is calculated as [24, 25]

$$C_{\omega}=\max\{I_{\mathbf{s};\mathbf{x}}\}=\log_{2}\det\left(\frac{\sigma_{\mathrm{S},\omega}^{2}}{\sigma_{\mathrm{N},\omega}^{2}}\mathbf{R}_{\mathrm{A},\omega}+\mathbf{I}_{M}\right). \tag{12}$$

By applying the eigenvalue decomposition to $\mathbf{R}_{\mathrm{A},\omega}$, C_{ω} is expressed by [24, 25]

$$C_{\omega}=\log_{2}\prod_{m=1}^{M}\left(\frac{\sigma_{\mathrm{S},\omega}^{2}}{\sigma_{\mathrm{N},\omega}^{2}}\Lambda_{m,\omega}+1\right), \tag{13}$$

where $\Lambda_{m,\omega}$ denotes the m-th eigenvalue of $\mathbf{R}_{\mathrm{A},\omega}$. In our previous study [16], we proved that C_{ω} is maximized if signals are observed to homogenize the eigenvalues;

$$\Lambda_{1,\omega}=,\ldots,=\Lambda_{M,\omega}. \tag{14}$$

The eigenvalues are homogenized, as in (14), and the array observations are then de-correlated as

$$\lim_{\Gamma_{i,j,\omega}\to 0}\mathbf{R}_{\mathrm{A},\omega}\to\sigma_{\mathrm{A},\omega}^{2}\mathbf{I}_{M}. \tag{15}$$

If $I_{\mathbf{s};\mathbf{x}}$ is increased, effective clues for segregating sound sources will be included in the observation signals.

As a microphone array implementation for increasing $I_{\mathbf{s};\mathbf{x}}$, we previously developed a microphone array whose microphones are semi-optimally placed in front of parabolic reflectors [16], as shown in Fig. 1. When a sound source is located in front of a parabolic reflector, the reflected waves pass through an area around a focal point of the reflector. Due to the reflector, even a small perturbation of the microphone position will drastically change the amplitude and phase of the received signal. Thus, we assumed that $I_{\mathbf{s};\mathbf{x}}$ would increase by optimizing the microphone arrangement. We constructed a microphone array composed of 12 parabolic reflectors and $M=96$ omnidirectional microphones, the details of which are explained in our previous study [16]. To increase $I_{\mathbf{s};\mathbf{x}}$, eight microphones were placed in front of each reflector.

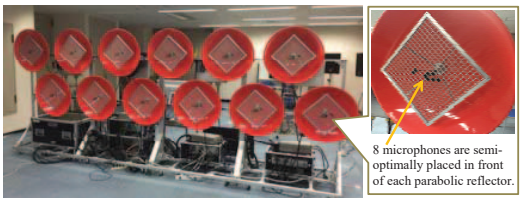

Fig. 1. Array structure to increase mutual information of MIMO using parabolic reflectors (4.0 m (W) × 1.5 m (H) × 1.0 m (D))

2.3. Source enhancement using beamforming

When a signal is observed from the array discussed in Sec. 2.2, it is difficult to analytically derive the transfer functions between the sound sources and microphones. Thus, we designed beamforming filters using pre-measured room impulse responses (RIRs). When the minimum variance distortionless response (MVDR) method [26] is used, filter coefficients for emphasizing the sound source arriving from the i-th position are calculated by

$$\mathbf{w}_{i,\omega}=\frac{\mathbf{R}_{\mathrm{A},\omega}^{-1}\mathbf{a}_{i,\omega}}{\mathbf{a}_{i,\omega}^{\mathrm{H}}\mathbf{R}_{\mathrm{A},\omega}^{-1}\mathbf{a}_{i,\omega}}. \tag{16}$$

After multiplying $\mathbf{w}_{i,\omega}$ by the observed signals, the i-th enhanced output signal is calculated as

$$Y_{i,\omega,\tau}=\mathbf{w}_{i,\omega}^{\mathrm{H}}\mathbf{x}_{\omega,\tau}. \tag{17}$$

Even when a beamforming is applied, as in (17), sharp directivity can be formed over a broad range of frequencies [16]. Although the SNR improves, especially in a high frequency range, some residual noise still remains in the beamforming output, especially when the noise level is significantly higher than that of the target speech. To clearly extract a user-pointed sound source, it is necessary to improve the source enhancement processing.

3. DNN-BASED PRIOR SNR ESTIMATION USING MULTIPLE BEAMFORMING OUTPUTS

3.1. Wiener filter design using prior SNR

To extract the target source arriving from the i-th source position, we apply the Wiener filter to the beamforming output using

$$Z_{i,\omega,\tau}=G_{i,\omega,\tau}Y_{i,\omega,\tau}. \tag{18}$$

The Wiener filter $G_{i,\omega,\tau}$ is calculated by

$$G_{i,\omega,\tau}=\frac{10^{\left(\xi_{i,\omega,\tau}/10\right)}}{1+10^{\left(\xi_{i,\omega,\tau}/10\right)}}, \tag{19}$$

where $\xi_{i,\omega,\tau}$ [dB] denotes the SNR at the beamforming output.

The PSD-estimation-in-beamspace method is effective for estimating the SNR at the observation point, which differs from $\xi_{i,\omega,\tau}$ [8, 9]. By using $L\ (\geq 2)$ beamforming output PSDs, as in (20), and the response sensitivities, as in (21), the PSDs of the target sound source and those of other noise can be estimated individually,

$$\phi_{\mathrm{Y},i,l,\omega}=\left\langle\left|Y_{\rho(i,l),\omega,\tau}\right|^{2}\right\rangle, \tag{20}$$

$$D_{l,k,\omega}=\left|\mathbf{w}_{l,\omega}^{\mathrm{H}}\mathbf{a}_{k,\omega}\right|^{2}, \tag{21}$$

where $\rho(i,l)$ denotes the focus position index of the l-th beamforming when the sound source arriving from the i-th position is the target. Since the SNR at the observation point could be estimated from multiple beamforming outputs $|Y_{\rho(i,l),\omega,\tau}|^{2}$, informative features to estimate $\xi_{i,\omega,\tau}$ could also be included in the PSDs.

3.2. DNN mapping from beamforming outputs to prior SNR

Our proposed DNN-based mapping method converts multiple beamforming outputs $|Y_{\rho(i,l),\omega,\tau}|^2$ to $\xi_{i,\omega,\tau}$. The DNN is a state-of-the-art statistical approach that has been attracting interest in many engineering fields recently. If informative features for estimating the desired SNR are included in multiple beamforming outputs, with the DNN algorithm, the network parameters between the $|Y_{\rho(i,l),\omega,\tau}|^2$ and $\xi_{i,\omega,\tau}$, which is supervised a priori, are optimized automatically.

Let the following feature vector composed of log-scaled multiple beamforming outputs be set to the input layer of the DNN with N-layers,

$$\mathbf{q}_{\Omega_j}^{(1)} = 10\log_{10}\{[|Y_{\rho(i,1),\Omega_1,\tau}|^2,\ldots,|Y_{\rho(i,1),\Omega_O,\tau}|^2,\ldots \\ \ldots,|Y_{\rho(i,L),\Omega_1,\tau}|^2,\ldots,|Y_{\rho(i,L),\Omega_O,\tau}|^2]\}^{\mathrm{T}}, \quad (22)$$

where Ω_j is the index of the frequency band divided into O equivalent rectangular bandwidth (ERB) scale [27]. The element number of $\mathbf{q}_{\Omega_j}^{(1)}$ is $L \times O$. This shrinkage is introduced to reduce the number of procedures for network parameter optimization. Given that the network parameter $\mathbf{p}_{\Omega_j}$ includes both $\mathbf{P}_{\Omega_j}^{(2)},\ldots,\mathbf{P}_{\Omega_j}^{(N)}$ and $\mathbf{b}_{\Omega_j}^{(2)},\ldots,\mathbf{b}_{\Omega_j}^{(N)}$, $\mathbf{u}_{\Omega_j}^{(n)}$ and $\mathbf{q}_{\Omega_j}^{(n)}$ are calculated by a recursive update for $N-1$ times expressed by

$$\mathbf{u}_{\Omega_j}^{(n)} = \mathbf{P}_{\Omega_j}^{(n)}\mathbf{q}_{\Omega_j}^{(n-1)} + \mathbf{b}_{\Omega_j}^{(n)}, \quad (23)$$

$$\mathbf{q}_{\Omega_j}^{(n)} = \mathbf{f}_{\Omega_j}^{(n)}\left(\mathbf{u}_{\Omega_j}^{(n)}\right). \quad (24)$$

Note that the network parameter was prepared for each frequency band independently. Assuming the number of nodes in the n-layer is denoted as J_n, the vectors and matrices in (23) and (24) are defined as

$$\mathbf{u}_{\Omega_j}^{(n)} = \left[u_{1,\Omega_j}^{(n)},\ldots,u_{J_n,\Omega_j}^{(n)}\right]^{\mathrm{T}}, \quad (25)$$

$$\mathbf{q}_{\Omega_j}^{(n)} = \left[q_{1,\Omega_j}^{(n)},\ldots,q_{J_n,\Omega_j}^{(n)}\right]^{\mathrm{T}}, \quad (26)$$

$$\mathbf{P}_{\Omega_j}^{(n)} = \begin{bmatrix} P_{1,1,\Omega_j}^{(n)} & \cdots & P_{1,J_{n-1},\Omega_j}^{(n)} \\ \vdots & \ddots & \vdots \\ P_{J_n,1,\Omega_j}^{(n)} & \cdots & P_{J_n,J_{n-1},\Omega_j}^{(n)} \end{bmatrix}, \quad (27)$$

$$\mathbf{b}_{\Omega_j}^{(n)} = \left[b_{1,\Omega_j}^{(n)},\ldots,b_{J_n,\Omega_j}^{(n)}\right]^{\mathrm{T}}, \quad (28)$$

$$\mathbf{f}^{(n)}\left(\mathbf{u}_{\Omega_j}^{(n)}\right) = \left[f^{(n)}\left(u_{1,\Omega_j}^{(n)}\right),\ldots,f^{(n)}\left(u_{J_n,\Omega_j}^{(n)}\right)\right]^{\mathrm{T}}. \quad (29)$$

For the activation function $f^{(n)}(\cdot)$, either a sigmoid function $(n=2,\cdots,N-1)$ or an identity function $(n=N)$ is used;

$$f(u) = \begin{cases} 1/(1+\exp(-u)) & (n=2,\ldots,N-1) \\ u & (n=N) \end{cases}. \quad (30)$$

Given that the number of nodes in the output layer is $J_N = 1$, the estimated SNR is obtained from the network parameter $\mathbf{p}_{\Omega_j}$;

$$\hat{\xi}_{i,\Omega_j,\tau} = q_{1,\Omega_j}^{(N)}. \quad (31)$$

After extending $\hat{\xi}_{i,\Omega_j,\tau}$ into the linear frequency scale, the Wiener filter to extract the target source is designed, as in (19).

Thanks to the progress in machine learning, the deep belief network (DBN) [17] is effective for setting appropriate initial values of $\mathbf{p}_{\Omega_j}$. In this study, we used the contrastive divergence [18, 19] to specify an appropriate amount of update for the network parameters of each layer. After initializing $\mathbf{p}_{\Omega_j}$, the network parameters

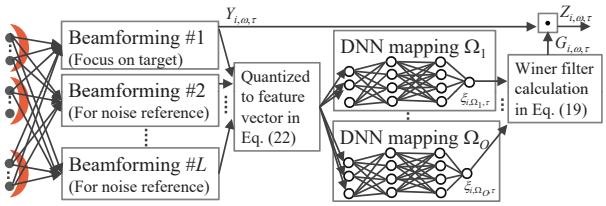

Fig. 2. Procedure of proposed method

were optimized based on the back propagation [28] to minimize the estimation error, which is defined by

$$E(\mathbf{p}_{\Omega_j}) = \frac{1}{2}\sum_{d=1}^{D} \| \xi_{i,\Omega_j,\tau} - \hat{\xi}_{i,\Omega_j,\tau} \|^2, \quad (32)$$

where D denotes the total number of training datasets composed of both multiple beamforming outputs $\mathbf{q}_{\Omega_j}^{(1)}$ and the true SNR $\xi_{i,\Omega_j,\tau}$ as the supervisor. The procedures in (23) and (24) applied to D samples can be represented as a matrix form given by

$$\mathbf{U}_{\Omega_j}^{(n)} = \mathbf{P}_{\Omega_j}^{(n)}\mathbf{Q}_{\Omega_j}^{(n-1)} + \mathbf{b}_{\Omega_j}^{(n)}\mathbf{1}_D^{\mathrm{T}}, \quad (33)$$

$$\mathbf{Q}_{\Omega_j}^{(n)} = \mathbf{f}_{\Omega_j}^{(n)}(\mathbf{U}_{\Omega_j}^{(n)}), \quad (34)$$

where

$$\mathbf{U}_{\Omega_j}^{(n)} = [\mathbf{u}_{\Omega_j,1}^{(n)},\ldots,\mathbf{u}_{\Omega_j,D}^{(n)}], \quad (35)$$

$$\mathbf{Q}_{\Omega_j}^{(n)} = [\mathbf{q}_{\Omega_j,1}^{(n)},\ldots,\mathbf{q}_{\Omega_j,D}^{(n)}]. \quad (36)$$

The gradient of the network parameters is recursively calculated from the output layer $(n=N)$ towards the input layer $(n=1)$. Given that $\mathbf{\Xi}_{\Omega_j} = [\xi_{\Omega_j,1},\ldots,\xi_{\Omega_j,D}]$, the gradient at the n-th layer $\mathbf{\Delta}_{\Omega_j}^{(n)}$ is derived by

$$\mathbf{\Delta}_{\Omega_j}^{(n)} = \begin{cases} \mathbf{f}^{(n)\prime}(\mathbf{U}_{\Omega_j}^{(n)}) \odot (\mathbf{P}_{\Omega_j}^{(n+1)\mathrm{T}}\mathbf{\Delta}_{\Omega_j}^{(n+1)}) & (n=2,\cdots,N-1) \\ \mathbf{\Xi}_{\Omega_j} - \mathbf{Q}_{\Omega_j}^{(n)} & (n=N) \end{cases} \quad (37)$$

where $\odot$ denotes an element-wise product of matrices. The gradient of the error functions is derived by

$$\partial\mathbf{P}_{\Omega_j}^{(n)} = \frac{1}{D}\mathbf{\Delta}_{\Omega_j}^{(n)}\mathbf{Q}_{\Omega_j}^{(n-1)\mathrm{T}}, \quad (38)$$

$$\partial\mathbf{b}_{\Omega_j}^{(n)} = \frac{1}{D}\mathbf{\Delta}_{\Omega_j}^{(n)}\mathbf{1}_D^{\mathrm{T}}. \quad (39)$$

Finally, the network parameters are updated as

$$\mathbf{P}_{\Omega_j}^{(n)} \leftarrow \mathbf{P}_{\Omega_j}^{(n)} + \Delta\mathbf{P}_{\Omega_j}^{(n)}, \quad (40)$$

$$\mathbf{b}_{\Omega_j}^{(n)} \leftarrow \mathbf{b}_{\Omega_j}^{(n)} + \Delta\mathbf{b}_{\Omega_j}^{(n)}, \quad (41)$$

where the perturbations for each update are calculated by

$$\Delta\mathbf{P}_{\Omega_j}^{(n)} = \mu\Delta\mathbf{P}_{\Omega_j}^{(n)*} - \epsilon\left(\partial\mathbf{P}_{\Omega_j}^{(n)} + \lambda\mathbf{P}_{\Omega_j}^{(n)}\right), \quad (42)$$

$$\Delta\mathbf{b}_{\Omega_j}^{(n)} = \mu\Delta\mathbf{b}_{\Omega_j}^{(n)*} - \epsilon\partial\mathbf{b}_{\Omega_j}^{(n)}, \quad (43)$$

Here, $\Delta\mathbf{P}_{\Omega_j}^{(n)*}$ and $\Delta\mathbf{b}_{\Omega_j}^{(n)*}$ are the perturbations of the previous update, ϵ is the learning rate, and μ and λ are the momentum coefficient and weight decay, respectively.

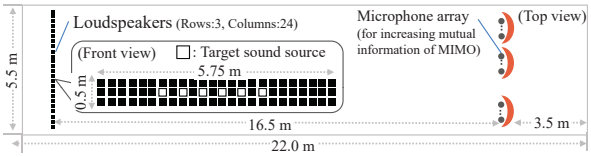

Fig. 3. Experimental setup

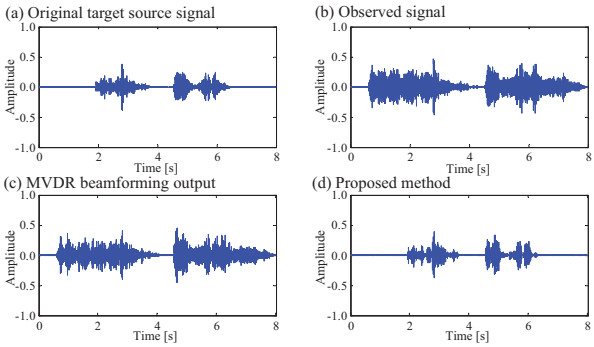

Fig. 4. Output waveform when $K = 5$ sources are used. (a) Original target source signal, (b) observed signal, (c) MVDR beeamforming that focuses on target, and (d) output signal with proposed method

4. EXPERIMENTS

4.1. Experimental conditions

We investigated the performance of the proposed method by evaluating both prior SNR estimation error and the SNR of the output signal. As a comparison method, we applied MVDR beamforming, as in (17). To design beamforming filters, RIRs from 72 loudspeakers to $M = 96$ microphones were measured in prior, as shown in Fig. 3. The distance between the center of the array and a sound source was 16.5 m, and the distance between loudspeakers was 0.25 m. To design the MVDR filter, impulse responses of 8 ms arriving from the direct sound were used. For evaluation, the positions of the target source were limited to 6, as shown in Fig. 3. We designed $L = 3$ beamforming filters for each target position. The focus point of each beamforming was the target position ($l = 1$), 1.25 m left/right from the target position ($l = 2, 3$), respectively. To investigate the relationships between the number of sound sources and the output SNR, K was varied from 2 to 5. We randomly placed $K - 1$ interference noise sources at the positions at which RIRs were measured. In total, 50 trials for sound source arrangement were executed for each combination of K and target position. A total of 24 types of male/female speech signals were used as the source signals.

By shrinking L=3 beamforming outputs into an ERB-scale feature vector, as in (22), $O = 57$ types of network parameters were optimized, which were not varied with the target source position. The total number of the training/test dataset was D=598800. The speech signals and positions of the interference sources were different among the training and test datasets (i.e. open test). The number of layers for the DNN was set to N=4. The other parameters used in the experiment are summarized in Table 1.

4.2. Experimental results

To evaluate the proposed prior estimation method, we calculated the averaged absolute error $|\xi_{i,\Omega_j,\tau} - \hat{\xi}_{i,\Omega_j,\tau}|$ for each frequency band,

Table 1. Parameters used in experiments

# of microphones, M	96
Sampling rate	16 kHz
FFT length	16 ms
# of frequency bands, O	57 (ERB scale)
# of beamformings, L	3
# of measured impulse responses	72 (Rows: 3, Columns: 24)
# of layers, N	4
# of nodes, J_n	J_1:171, J_2:220, J_3:220, J_4:1
Learning coefficient, ϵ	0.005, 0.0025, 0.0001
Iteration number	30 (for each ϵ)
Momentum coefficient, μ	0.5 (first 3), 0.9 (after 4)
Decay weight, λ	0.0002
# of target source positions	6
# of noise sources patterns	4 (K=2,3,4,5)
# of source position arrangements	50 (for each target position)
# of frames for each speech	499 (8.0 sec)
# of training datasets, D	598800 (=6*4*50*499)
# of evaluation datasets (open test)	598800 (=6*4*50*499)

Table 2. Prior SNR estimation error with proposed method

Frequency [kHz]	0.5	1.0	2.0	4.0	7.5
Prior SNR estimation error [dB]	8.8	6.4	5.4	5.0	2.1

Table 3. Evaluation of output SNR

Number of sound sources	K=2	K=3	K=4	K=5
SNR (observed point) [dB]	0.6	-2.6	-4.5	-5.7
SNR (beamforming) [dB]	6.8	4.4	3.3	2.7
SNR (proposed method) [dB]	34.7	30.8	28.7	27.3

and the results are listed in Table 2. The estimation error decreased, especially at high frequencies. Since the sharp directivity could be formed as frequency increased with our array [16], informative cues to estimate the prior SNR could be included in the multiple beamforming outputs.

After estimating prior SNR for each frequency band, we designed the Wiener filter and applied it, as in (19). Fig. 4 shows waveform examples when $K = 5$ sound sources were used. The target source was positioned third from the left, as shown in Fig. 3. With the proposed method, the output signal was almost the same waveform as the original signal. Table 3 shows the relationships between K and the averaged SNR of the output signal of MVDR beamforming and the proposed method. From these results, SNR improved by about 25 dB compared with MVDR beamforming, independently of K. Thus, it was confirmed that the proposed method is effective for extracting the target source even when it was positioned at a remote location.

5. CONCLUSION

We proposed a DNN-based mapping method from multiple beamforming outputs to the prior SNR at the beamforming output. By learning the relationships between the multiple beamforming outputs and prior SNR beforehand, the DNN parameters were optimized. By using estimated prior SNR, the Wiener filter was generated and applied to the beamfoming output. Through experiments, the estimation error on the prior SNR was sufficiently low and it was enable to pick up the target source positioned at a remote location.

For future work, we will investigate the environmental robustness of the pre-learned DNN parameters and apply another structure for microphone array observations.

6. REFERENCES

[1] H. L. V. Trees, *Optimum array processing*, Wiley-Interscience (Part IV ed.), 2002.

[2] D. H. Johnson and D. E. Dudgeon, *Array processing: concepts and techniques*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[3] C. Marro, Y. Mahieux, K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. on Speech and Audio Proc.*, no. 6, pp. 240–259, 1998.

[4] T. Wolff and M. Buck, "A generalized view on microphone array postfilters," *in Proc. IWAENC 2010*, 2010.

[5] R. Zelinski, "A microphone array with adaptive post–filtering for noise reduction in reverberant rooms," *in Proc. ICASSP 1988*, 5, 2578–2581, 1988.

[6] I. A. McCowan, H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. on Audio, Speech, and Language Proc.*, no. 11, pp. 709–716, 2003.

[7] K. U. Simmer, J. Bitzer, and C. Marro, *Microphone arrays: signal processing techniques and applications*, chapter 3, pp. 39–60, Springer, 1 edition, 2001.

[8] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 21, pp. 1240–1250, 2013.

[9] K. Niwa, Y. Hioka, and K. Kobayashi, "Post-filter design for speech enhancement in various noisy environments," *in Proc. IWAENC 2014*, pp. 36–40, 2014.

[10] J. Meyer, and G. W. Elko, "A spherical microphone array for spatial sound recordings," *J. Acoust. Soc. Am.* , vol. 111, Issue 5, 2346, 2002.

[11] B. Rafaely, "Open–sphere designs for spherical microphone arrays," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 15, no. 2, pp. 727–732, 2007.

[12] H. Morgenstern and B. Rafaely, "Analysis of acoustics MIMO systems in enclosed sound fields," *in Proc. ICASSP 2012*, pp. 209–212, 2012.

[13] T. D. Abhayapala, and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," *in Proc. ICASSP2002*, vol. II, pp. 1949–1952, 2002.

[14] K. Niwa, Y. Hioka, K. Furuya, and Y. Haneda, "Diffused sensing for sharp directive beamforming," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 21, pp. 2346–2355, 2013.

[15] K. Niwa, Y. Hioka, K. Kobayashi, K. Furuya, and Y. Haneda, "Evaluation of microphone array based on diffused sensing with various filter design methods", *in Proc. EUSIPCO 2013*(156974157), 2013.

[16] K. Niwa, T. Kako, and K. Kobayashi, "Microphone array for increasing mutual information between sound sources and observation signals," *ICASSP 2015*, pp. 534–538, 2015.

[17] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1544, 2006.

[18] Y. Bengio, "Learning deep architecture for ai," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[19] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 771–800, 2002.

[20] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends pm far field multiple microphones based speech recognitions," *in Proc. ICASSP2014*, pp. 5579–5582, 2014.

[21] S. Renals and P. Swietojanski, "Neural networks for distant speech recognition," *in Proc. HSCMA2014*, 2014.

[22] S. Araki, T. Hayashi, M. Delcroix, M Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," *in Proc. ICASSP2015*, pp. 116–120, 2015.

[23] W. Zheng, Y. Zou, and C. Ritz, "Spectral mask estimation using deep neural networks for inter-sensor data ratio model based robust DOA estimation," *in Proc. ICASSP2015*, pp. 325–329, 2015.

[24] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 41–59, 1996.

[25] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multi-element antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, 1998.

[26] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *in Proc. IEEE*, vol. 60, pp. 926–935, 1972.

[27] B. C. J. Moore, An Introduction to the Psychology of Hearing, Fifth Edition, Academic Press.

[28] D. E. Rumelhart and J. Mcclelland, "Parallel distributed processing: explorations in the microstructure of cognition," *MIT Press*, 1986.