SUPERVISED SOURCE ENHANCEMENT COMPOSED OF NONNEGATIVE AUTO-ENCODERS AND COMPLEMENTARITY SUBTRACTION

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

A method for constructing deep neural networks (DNNs) for accurate supervised source enhancement is proposed. Attempts were made in previous studies to estimate the power spectral densities (PSDs) of sound sources, which are used to estimate Wiener filters for source enhancement, from the output of multiple beamformings using DNNs. Although performance improved, it was not possible to guarantee accurate PSD estimation since the trained DNNs were treated as black boxes. The proposed DNN construction method uses non-negative auto-encoders and complementarity subtraction. This study also reveals that auto-encoders whose weights are non-negative correspond to non-negative matrix factorization (NMF), which decomposes source PSDs into non-negative spectral bases and their activations. It further introduces a complementarity subtraction method for estimating PSDs accurately. Through several experiments, it was confirmed that the signal-to-interference plus noise ratio improved by approximately 12 dB for datasets captured in various noisy/reverberant rooms.

Index Terms— microphone array, beamforming, Wiener filtering, deep learning, auto-encoder, non-negative matrix factorization

1. INTRODUCTION

Microphone array signal processing [1, 2] has been used for emphasizing a target source in noisy environments for various applications such as speech recognition-based car-navigation control, noise-canceling headsets for hands-free communication in very noisy factories, sports game sound enhancement for broadcast audio rendering [3], and clearly picking up distant sound sources [4, 5].

Classical array-signal-processing techniques use spatial cues included in the phase/amplitude differences between microphones to identify a target source. Beamforming [1, 2] is a simple way to achieve this, but in practical applications, Wiener post-filtering is often also applied to the beamforming output to further improve source enhancement. Various studies have attempted to estimate the Wiener filter by analyzing microphone observation [6]-[10]. Estimating the power spectral density (PSD) of a target and noise sources is a common strategy since the Wiener filter can be calculated directly from the estimated PSDs. In the PSD-estimation-in-beamspace method [11, 12], PSDs of multiple beamformings are modeled as a simple linear mixture of source PSDs, assuming that the source signals are uncorrelated with each other. By solving the mixture using a linear de-mixing matrix, PSDs of sound sources in the angles of interest are calculated.

Apart from spatial cues, the spectral characteristics of sources may also be used to segregate a target source from other sources [3, 13]. To this end, several previous studies have investigated supervised source enhancement. The application of deep neural networks (DNNs) is a recent hot topic in audio signal processing [14]-[17]. A DNN is known to be a powerful tool for describing the nonlinear relationships between two different pieces of information, i.e., input and output feature vectors [18]-[20], and its sophisticated network parameters can be obtained through back-propagation optimization [21] using a huge number of datasets. Our recent studies [5, 22] attempted to incorporate DNNs into the PSD-estimation-in-beamspace method, describing the mappings from the PSDs of multiple beamformings into those of sound sources by using nonlinear DNNs. Experiments showed that the signal-to-interference plus noise ratio (SINR) could be improved; however, the accuracy of PSD estimation with the DNN mapping function could not be guaranteed since the network structure was treated as a black box.

To accurately estimate a Wiener post-filter, we propose a method of constructing *explainable* neural networks composed of a non-negative auto-encoder and complementarity subtraction. Non-negative matrix factorization (NMF) is a well-known model for decomposing a spectrogram into spectral bases and their activations [23]-[27]. Some of the relationships between NMF and DNNs have been discussed [28]. To incorporate the NMF-based decomposition model into DNNs, we found that constructing an auto-encoder [29] while constraining the non-negative network weights may be appropriate. The PSD of targets/noise can be estimated accurately by complementarily subtracting non-negative auto-encoder outputs [30] when they model the spectral characteristics of target/noise sources independently.

This paper is organized as follows. We explain the conventional supervised source separation method using DNNs in Sec. 2. In Sec. 3, we introduce the proposed neural network construction method. After evaluating the proposed method through several experiments in Sec. 4, we conclude this paper in Sec. 5.

2. CONVENTIONAL METHOD

2.1. Wiener filtering-based source enhancement

Assume that an M (≥ 2)-sensor microphone array and K sound sources are placed in an acoustic field. The direction of a target source is assumed to be given, whereas any information about other sources, i.e., interfering noise, are unknown a priori. The transfer functions between the K sound sources and M microphones are denoted as $\mathbf{A}_\omega \in \mathbb{C}^{M\times K}$, where ω denotes the frequency index. When the time-frame index is described as τ, K source signals are denoted as $\mathbf{s}_{\omega,\tau} \in \mathbb{C}^{K\times 1}$, and background noise at the microphones is denoted as $\mathbf{n}_{\omega,\tau} \in \mathbb{C}^{M\times 1}$. The M observed signals $\mathbf{x}_{\omega,\tau} \in \mathbb{C}^{M\times 1}$ are modeled in the time-frequency domain as

$$\mathbf{x}_{\omega,\tau} = \mathbf{A}_\omega \mathbf{s}_{\omega,\tau} + \mathbf{n}_{\omega,\tau}. \quad (1)$$

When the filter of a beamforming for emphasizing the target source is represented as $\mathbf{h}_\omega \in \mathbb{C}^{M\times 1}$, the output signal of the beamformer $Y_{\omega,\tau}$ is given by

$$Y_{\omega,\tau} = \mathbf{h}_\omega^{\mathrm{H}} \mathbf{x}_{\omega,\tau}, \tag{2}$$

where $^{\mathrm{H}}$ denotes the Hermitian transpose. As an implementation of $\mathbf{h}_\omega$, the minimum variance distortion-less response (MVDR) method [31] may be used. To boost noise reduction, the Wiener post-filter $V_{\omega,\tau}$ is multiplied by $Y_{\omega,\tau}$ as

$$Z_{\omega,\tau} = V_{\omega,\tau} Y_{\omega,\tau}. \tag{3}$$

To achieve effective source enhancement, $V_{\omega,\tau}$ needs to be accurately estimated by analyzing $\mathbf{x}_{\omega,\tau}$.

2.2. Wiener post-filter estimation using DNNs

In the PSD-estimation-in-beamspace method [11, 12], $L\,(\geq 2)$ beamformers are used for acoustic field analysis. One beamformer is designed to point its mainlobe to the target sound source, whereas the rest are focused on other sources. Assuming that source signals are mutually uncorrelated in every frequency band, the relationships between the PSDs of beamforming outputs $\mathbf{\Phi}_{\mathrm{Y}_\omega} \in \mathbb{R}^{L\times 1}$ and those of source signals grouped into N angular regions $\mathbf{\Phi}_{\mathrm{G}_\omega} \in \mathbb{R}^{N\times 1}$ are approximated with a linear mixture model using a matrix of the beamformers' gain $\mathbf{D}_\omega \in \mathbb{R}^{L\times N}$,

$$\mathbf{\Phi}_{\mathrm{Y}_\omega} \approx \mathbf{D}_\omega \mathbf{\Phi}_{\mathrm{G}_\omega}. \tag{4}$$

When the sparseness of sound sources in the time-frequency domain can be assumed, equality will hold for the relationship between the *instantaneous* PSD of the beamformers' output and sound sources at frame τ

$$\mathbf{\Phi}_{\mathrm{Y}_{\omega,\tau}} = \mathbf{D}_\omega \mathbf{\Phi}_{\mathrm{G}_{\omega,\tau}}. \tag{5}$$

By solving the inverse problem linearly, as in (6), the power spectra of grouped sound sources are estimated by

$$\hat{\mathbf{\Phi}}_{\mathrm{G}_{\omega,\tau}} = (\mathbf{D}_\omega)^{\dagger}\, \mathbf{\Phi}_{\mathrm{Y}_{\omega,\tau}}, \tag{6}$$

where † denotes the matrix inverse operator when $L{=}N$; otherwise, it denotes the pseudo-inverse. Although the details are omitted, the Wiener post-filter is calculated after estimating the power spectra of target $\phi_{\mathrm{S}_{\omega,\tau}}$ and that of noise $\phi_{\mathrm{N}_{\omega,\tau}}$ as

$$V_{\omega,\tau} = \frac{\hat{\phi}_{\mathrm{S}_{\omega,\tau}}}{\hat{\phi}_{\mathrm{S}_{\omega,\tau}} + \hat{\phi}_{\mathrm{N}_{\omega,\tau}}}. \tag{7}$$

Since the sparseness assumption may not always hold in practice, which would cause errors to appear in the estimated Wiener post-filter, modeling the relationship as a nonlinear mapping as

$$\{\hat{\phi}_{\mathrm{S}_{\omega,\tau}}, \hat{\phi}_{\mathrm{N}_{\omega,\tau}}\} \leftarrow \mathcal{M}\left(\mathbf{\Phi}_{\mathrm{Y}_{\omega,\tau}}\right), \tag{8}$$

may provide a better estimate of the PSDs, where $\mathcal{M}(\cdot)$ represents a nonlinear mapping function. This was implemented using DNNs in our previous studies [5, 22]. In those studies, a large number of datasets composed of inputs $\Phi_{\mathrm{Y}_{\omega,\tau}}$ and outputs $\{\phi_{\mathrm{S}_{\omega,\tau}}, \phi_{\mathrm{N}_{\omega,\tau}}\}$ were given as supervisors, and the network parameters were optimized through back propagation [21].

In these previous studies [5, 22], the estimated DNN mapping function was treated as a black box because the network size, i.e., the number of layers and nodes, and input/output features were heuristically determined, network parameters were randomly initialized, and features were normalized to follow a normal distribution. Thus, it was difficult to trace whether the optimized DNN mapping function could provide an accurate estimate of the Wiener post-filter.

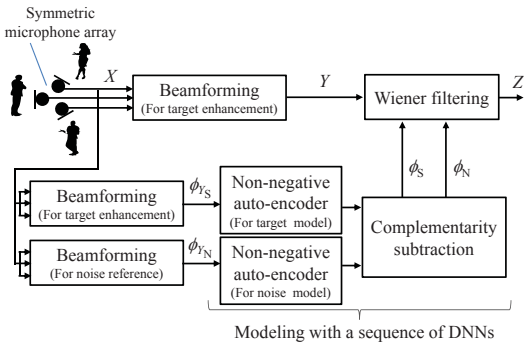

Fig. 1. Processing flow of proposed method

3. PROPOSED METHOD

To construct DNNs for accurately estimating Wiener post-filters, we propose a method of constructing explainable neural networks based on PSD estimation. Figure 1 shows the processing flow of the proposed method. The constructed neural networks are composed of non-negative auto-encoders to model the spectral characteristics of target/noise sources, as explained in Sec. 3.1, and complementarity subtraction for estimating the power spectra of target/noise sources, as explained in Sec. 3.2. The supplemental information for network optimization is discussed in Sec. 3.3.

3.1. Modeling of target/noise-source spectral characteristics using nonnegative auto-encoders

We now explain our neural network construction method for modeling the spectral characteristics of target/noise sources. The NMF is a well-known source-separation technique and is specifically applied to single-channel input signals. In the NMF framework, a mixture spectrogram $\mathbf{S} \in \mathbb{R}^{\Omega\times\Upsilon}$ is decomposed into non-negative spectral bases $\mathbf{B} \in \mathbb{R}^{\Omega\times\beta}$ and their non-negative activations $\mathbf{A} \in \mathbb{R}^{\beta\times\Upsilon}$,

$$\mathbf{S} = \mathbf{B}\mathbf{A}, \tag{9}$$

where Υ, Ω, and β are the number of time-frames, that of frequency bands, and that of spectral bases, respectively. Although there are many types of norm standards [23, 26], $\mathbf{A}$ and $\mathbf{B}$ are iteratively optimized to satisfy (9).

Similar to the NMF framework, the idea of decomposing a spectrogram into spectral bases and activations can be represented using neural networks. The simplest way to achieve this idea is to construct a 3-layer non-negative auto-encoder. To model the spectral characteristics of target/noise sources independently, 2-layered input vectors are organized as

$$\mathbf{q}_{\mathrm{S}}^{(1)} = [\bar{\phi}_{Y_{\mathrm{S}_{1,\tau}}}, \ldots, \bar{\phi}_{Y_{\mathrm{S}_{\Omega,\tau}}}]^{\mathrm{T}}, \tag{10}$$

$$\mathbf{q}_{\mathrm{N}}^{(1)} = [\bar{\phi}_{Y_{\mathrm{N}_{1,\tau}}}, \ldots, \bar{\phi}_{Y_{\mathrm{N}_{\Omega,\tau}}}]^{\mathrm{T}}, \tag{11}$$

where $^{\mathrm{T}}$ denotes transposition. The term $\mathbf{q}_{\mathrm{N}}^{(1)} \in \mathbb{R}^{\Omega\times 1}$ is used for modeling target sources, which is composed of the output PSD of a beamformer that focuses on the target source angle $\bar{\phi}_{Y_{\mathrm{S}_{\omega,\tau}}}$. Likewise, $\mathbf{q}_{\mathrm{N}}^{(1)} \in \mathbb{R}^{\Omega\times 1}$ is used for modeling noise sources, which is composed of the output PSD of beamformers $\bar{\phi}_{\mathbf{Y}_{\mathrm{N}_{\omega,\tau}}}$. Note that the elements in $\mathbf{q}_{\mathrm{S}}^{(1)} \in \mathbb{R}^{\Omega\times 1}$ and $\mathbf{q}_{\mathrm{N}}^{(1)} \in \mathbb{R}^{\Omega\times 1}$ are normalized by the beamformers' gain. Although more than one noise source may be assumed in the noise-source model, we only consider a single noise model in this paper to investigate noise reduction performance with

the simplest implementation. Thus $\bar{\phi}_{Y_{\mathrm{N}_{\omega,\tau}}}$ is calculated by averaging $L-1$ noise-reference beamforming outputs. When the n-th layer input vector is denoted as $\mathbf{q}^{(n)}$ as a general form, it is calculated recursively as

$$\mathbf{u}^{(n)} = \mathbf{W}^{(n)}\mathbf{q}^{(n-1)} + \mathbf{b}^{(n)}, \tag{12}$$

$$\mathbf{q}^{(n)} = f^{(n)}\left(\mathbf{u}^{(n)}\right), \tag{13}$$

where $\mathbf{W}^{(n)}$ and $\mathbf{b}^{(n)}$ denote the weight matrix and bias vector of the n-th layer, respectively. The network parameter $\mathbf{p}$ is composed of an N-layer of weight matrices and bias vectors. As an activation function $f^{(n)}(\cdot)$, the rectified linear unit (ReLU) is used because it makes network outputs non-negative

$$f^{(n)}(u) = \max(0, u). \tag{14}$$

By regarding $\{\mathbf{W}_{\mathrm{S}}^{(2)}, \mathbf{W}_{\mathrm{N}}^{(2)}\} \in \mathbb{R}^{\beta\times\Omega}$ as β sets of spectral bases of target/noise sources, its elements are constrained to be non-negative as follows:

$$W_{\mathrm{S},i,j}^{(2)} \geq 0 \qquad (\forall\, i, j), \tag{15}$$

$$W_{\mathrm{N},i,j}^{(2)} \geq 0 \qquad (\forall\, i, j), \tag{16}$$

where $W_{\mathrm{S},i,j}^{(n)}$ denotes the (i,j)-th elements of $\mathbf{W}_{\mathrm{S}}^{(n)}$. Then, the second-layer input for the target model $\mathbf{q}_{\mathrm{S}}^{(2)} \in \mathbb{R}^{\beta\times 1}$ and noise model $\mathbf{q}_{\mathrm{N}}^{(2)} \in \mathbb{R}^{\beta\times 1}$ can be regarded as non-negative activations corresponding to each spectral basis at a time-frame. By replacing the weight matrix in the 3rd layer with the transposed weight matrix in the 2nd layer, as in (17) and (18), the spectrogram is reconstructed since activations $\{\mathbf{q}_{\mathrm{S}}^{(2)}, \mathbf{q}_{\mathrm{N}}^{(2)}\}$ and spectral bases $\{\mathbf{W}_{\mathrm{S}}^{(3)}, \mathbf{W}_{\mathrm{N}}^{(3)}\}$ are multiplied.

$$\mathbf{W}_{\mathrm{S}}^{(3)} = \mathbf{W}_{\mathrm{S}}^{(2)\mathrm{T}}, \tag{17}$$

$$\mathbf{W}_{\mathrm{N}}^{(3)} = \mathbf{W}_{\mathrm{N}}^{(2)\mathrm{T}} \tag{18}$$

Here, $\{\mathbf{b}_{\mathrm{S}}^{(3)}, \mathbf{b}_{\mathrm{N}}^{(3)}\}$ should be replaced by zero-vectors. The reconstruction model in (19) and (20) is equal to the single time-frame NMF model in (9).

$$\mathbf{q}_{\mathrm{S}}^{(3)} = f^{(3)}\left(\mathbf{W}_{\mathrm{S}}^{(3)}\mathbf{q}_{\mathrm{S}}^{(2)} + \mathbf{b}_{\mathrm{S}}^{(3)}\right) = \mathbf{W}_{\mathrm{S}}^{(3)}\mathbf{q}_{\mathrm{S}}^{(2)}, \tag{19}$$

$$\mathbf{q}_{\mathrm{N}}^{(3)} = f^{(3)}\left(\mathbf{W}_{\mathrm{N}}^{(3)}\mathbf{q}_{\mathrm{N}}^{(2)} + \mathbf{b}_{\mathrm{N}}^{(3)}\right) = \mathbf{W}_{\mathrm{N}}^{(3)}\mathbf{q}_{\mathrm{N}}^{(2)} \tag{20}$$

Through nonnegative auto-encoders, the 3rd-layer inputs $\mathbf{q}^{(3)} = [\mathbf{q}_{\mathrm{S}}^{(3)\mathrm{T}}, \mathbf{q}_{\mathrm{N}}^{(3)\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2\Omega\times 1}$ may be pre-enhanced power spectra of target/noise sources. The neural networks whose structure is defined in (17) and (18) are called auto-encoders [29]. Thus, constructing a 3-layer nonnegative auto-encoder corresponds to modeling the spectral characteristics of target/noise sources with the NMF model.

3.2. Complementarity subtraction for estimating power spectra of target/noise sources

Through 3-layer non-negative auto-encoders, a de-noising effect may be somewhat obtained. We believe a more noise-reducing unit is needed to accurately estimate the power spectra of target/noise sources. In our previous studies on the PSD-estimation-in-beamspace method [30], we discussed the roles of $(\mathbf{D}_\omega)^\dagger$ defined in (6). In the beamforming output, although the target source is emphasized, its interferences still remain as residual noise. The same phenomenon may be applicable in de-noising auto-encoder outputs. We model 3rd-layer inputs by simply adding the ideal power spectra of target source $\mathbf{d}_{\mathrm{S}} \in \mathbb{R}^{\Omega\times 1}$ and that of noise source $\mathbf{d}_{\mathrm{N}} \in \mathbb{R}^{\Omega\times 1}$ as follows:

$$\mathbf{q}_{\mathrm{S}}^{(3)} \approx \mathbf{d}_{\mathrm{S}} + \mathbf{\Gamma}_{\mathrm{S}}\,\mathbf{d}_{\mathrm{N}}, \tag{21}$$

$$\mathbf{q}_{\mathrm{N}}^{(3)} \approx \mathbf{d}_{\mathrm{N}} + \mathbf{\Gamma}_{\mathrm{N}}\,\mathbf{d}_{\mathrm{S}}, \tag{22}$$

where

$$\mathbf{d}_{\mathrm{S}} = [\phi_{\mathrm{S}_{1,\tau}}, \ldots, \phi_{\mathrm{S}_{\Omega,\tau}}]^{\mathrm{T}}, \tag{23}$$

$$\mathbf{d}_{\mathrm{N}} = [\phi_{\mathrm{N}_{1,\tau}}, \ldots, \phi_{\mathrm{N}_{\Omega,\tau}}]^{\mathrm{T}}, \tag{24}$$

$$\mathbf{\Gamma}_{\mathrm{S}} = \mathrm{diag}\{[\gamma_{\mathrm{S},1}, \ldots, \gamma_{\mathrm{S},\Omega}]\}, \tag{25}$$

$$\mathbf{\Gamma}_{\mathrm{N}} = \mathrm{diag}\{[\gamma_{\mathrm{N},1}, \ldots, \gamma_{\mathrm{N},\Omega}]\}. \tag{26}$$

Here, $\mathbf{\Gamma}_{\mathrm{S}}$ and $\mathbf{\Gamma}_{\mathrm{N}}$ are composed of interference-remaining rates for each frequency band. They are satisfied with $0 < \gamma_{\mathrm{S},\omega} < 1$ and $0 < \gamma_{\mathrm{N},\omega} < 1$.

Although the details on the theory proposed in our previous study [30] are omitted, the inverse relationships of (21) and (22) can be modeled approximately if the remaining weights are sufficiently small.

$$\mathbf{d}_{\mathrm{S}} \approx \mathbf{q}_{\mathrm{S}}^{(3)} - \mathbf{\Gamma}_{\mathrm{S}}\,\mathbf{q}_{\mathrm{N}}^{(3)}, \tag{27}$$

$$\mathbf{d}_{\mathrm{N}} \approx \mathbf{q}_{\mathrm{N}}^{(3)} - \mathbf{\Gamma}_{\mathrm{N}}\,\mathbf{q}_{\mathrm{S}}^{(3)} \tag{28}$$

Since the complementarity subtraction in (27) and (28) is described by a matrix form, it can be embedded in a sequence of DNNs as

$$\mathbf{W}_{\mathrm{Init}}^{(4)} = \left[\begin{array}{c|c} \boldsymbol{I}_\Omega & -\mathbf{\Gamma}_{\mathrm{S}} \\ \hline -\mathbf{\Gamma}_{\mathrm{N}} & \boldsymbol{I}_\Omega \end{array}\right], \tag{29}$$

where $\mathbf{W}_{\mathrm{Init}}^{(4)} \in \mathbb{R}^{2\Omega\times 2\Omega}$ denotes the initial value of that weight matrix, and the structure in (29) is iteratively updated through back propagation optimization.

Supplementarily, capturing sound using a regular circular microphone array might be better for accurately estimating the power spectra of target/noise sources. When such a symmetric microphone array is used, it enables us to make $\mathbf{D}_\omega$ independently of the target direction. Then, the noise reduction processes in (27) and (28) are expected to be optimized independently of the target arrival direction.

3.3. Implementation of network parameter optimization

We now briefly summarize the optimization process of $\mathbf{p}$.

[Step 1] The spectral characteristics of target/noise sources are independently modeled with non-negative auto-encoders. To initialize the spectral bases, the gain-normalized beamforming output power created when only target/noise sources are inputted (unmixed datasets) is calculated. By applying k-means clustering, β types of different spectral characteristics are selected, and they are used as an initial value of $\{\mathbf{W}_{\mathrm{S}}^{(2)}, \mathbf{W}_{\mathrm{N}}^{(2)}\}$. After inserting an unmixed dataset into $\{\mathbf{q}_{\mathrm{S}}^{(1)}, \mathbf{q}_{\mathrm{N}}^{(1)}\}$ and $\{\mathbf{q}_{\mathrm{S}}^{(3)}, \mathbf{q}_{\mathrm{N}}^{(3)}\}$, back propagation is applied while constraining network parameters, as in (15)-(18). By replacing $\{\mathbf{q}_{\mathrm{S}}^{(1)}, \mathbf{q}_{\mathrm{N}}^{(1)}\}$ with the beamforming output power when noisy observed signals are inputted (mixed datasets), 3-layer denoising non-negative auto-encoders are constructed as a target/noise source model.

[Step 2] Complementarity subtraction is embedded behind denoising non-negative auto-encoders. After initializing the weight matrix, as in (29), and $\mathbf{b}^{(4)}$ is replaced with a zero-vector, the network parameters of only the 4-th layer are pre-trained through back propagation. Since $\mathbf{q}^{(4)}$ is expected to be composed of the power spectra of target/noise sources, the cost function for network optimization is designed by

Table 1. Parameters used in experiments

# of microphones, M	3
Sampling rate	16 kHz
FFT length	8 ms
# of frequency bands, Ω	50 (ERB scale [32])
# of beamformings, L	3
# of layers, N	4
# of spectral bases, β	320
# of nodes, J_n	J_1: 2Ω, J_2: 2β, J_3: 2Ω, J_4: 2Ω
# of arrival directions of target	5 (0, 45, 90, 135, 180 deg)
# of background noise level	5 (-10, -5, 0, 5, 10 dB)
# of trials for each condition	100 (training), 20 (evaluation)
# of frames for each signal	499 (4.0 sec)
# of training datasets	1,247,500 (=5*5*100*499)
# of evaluation datasets	249,500 (=5*5*20*499)

Table 2. Evaluation of noise-reduction performances (SINRimp: SINR improvement, BF: Beamforming, WF: Wiener filtering)

Background noise level [dB]	-10	-5	0	5	10
Input SINR [dB]	-1.7	-2.2	-3.8	-6.7	-10.6
SINRimp (BF) [dB]	3.2	3.4	4.5	7.0	10.7
SINRimp (Conv. DNN-WF) [dB]	3.8	4.0	4.9	7.2	10.9
SINRimp (Prop. DNN-WF) [dB]	12.3	11.3	12.0	11.8	13.3
SINRimp (Ideal WF) [dB]	17.7	14.7	14.8	14.1	15.0

$$E(\mathbf{p}) = \frac{1}{2}\left\langle \| \mathbf{q}_{\text{Ideal}}^{(4)} - \mathbf{q}^{(4)} \|^2 \right\rangle, \tag{30}$$

where supervised output features are defined by $\mathbf{q}_{\text{Ideal}}^{(4)} = [\mathbf{d}_{\text{S}}^{\text{T}}, \mathbf{d}_{\text{N}}^{\text{T}}]^{\text{T}}$.

[Step 3] The $N = 4$ layers of DNNs are optimized through back propagation. After the optimization of $\mathbf{p}$, the power spectra of target/noise sources can be estimated frame-by-frame using the optimized DNNs. By applying the Wiener filter, as in (3), and inverse-FFT, enhanced signals are obtained, as shown in Fig. 1.

4. EXPERIMENTS

4.1. Experimental conditions

We conducted experiments to investigate the effectiveness of the proposed method. The regular circular array was composed of $M=3$ omni-directional microphones that were configured as a regular triangle with a diameter of 0.046 m. The array was placed in 8 different situations (= near a wall/center of 4 reverberant rooms), and recorded room impulse responses, and background noise (e.g., office and exhibition hall noise) were reproduced from 6 loudspeakers placed in the corners of the rooms. The target speech was assumed to arrive from one of five different directions (45-degrees interval), and less than two pieces of interference speech were placed in the other arrival direction. The source signals were randomly selected from 4,000 male/female sentences. After adjusting the background noise level to the target speech from -10 to 10 dB, the array-observed signals were simulated. In total, 2.8 hours ($\approx 5 \times 5 \times 100 \times 4$ seconds) of observation signals were used for training, and 0.6 hours ($\approx 5 \times 5 \times 20 \times 4$ seconds) of different signals were used for evaluation. The other parameters are listed in Table 1.

The neural networks were constructed by modeling the target/noise sources with β=320 spectral bases for each auto-encoder. By using optimized $\mathbf{p}$, the PSDs of the target/noise sources were estimated and Wiener filtering was applied to the fixed beamforming output (Prop. DNN-WF). For comparison, neural networks of the same size were used, whose input features were normalized to follow a zero-mean normal distribution, the final layer activation function was replaced by a linear function, and network parameters were initialized with normal auto-encoders (Conv. DNN-WF). We also calculated the output performances when Wiener filter was ideally designed to investigate the upper boundary of performance.

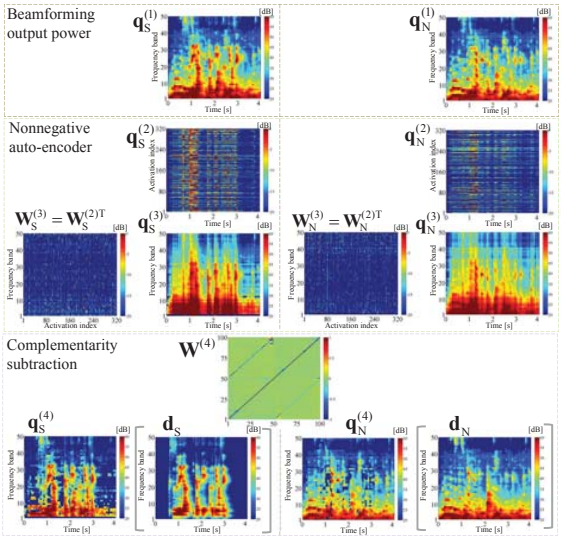

Fig. 2. Visualization of input/hidden/output layer and some network parameters when applying proposed method

4.2. Experimental results

The experimental results are listed in Table 2. The SINR improvement (= output SINR - input SINR) was calculated for the output of MVDR beamforming (BF), Conv. DNN-WF, Prop. DNN-WF, and ideal WF. Although SINR improvement varied corresponding to the input SINR with conventional methods, it remained around 12 dB almost independently of the input SINR with the proposed method. Since SINR improvement was around 15 dB when ideal Wiener filter was applied, we believe that our method will work well in various noisy environments.

Input/hidden/output layers and some network parameters are illustrated in Fig. 2. By de-noising non-negative auto-encoders, the target/noise sources seemed to enhance to some degree. It was confirmed that $\mathbf{W}^{(4)}$ works for complementarity subtracting the auto-encoder outputs for each frequency band. Since the estimated PSDs were similar to the ideal value, complementarity subtraction may be important for accurate PSD estimation.

5. CONCLUSION

A method for constructing deep neural networks composed of non-negative auto-encoders and complementarity subtraction is proposed for supervised source separation. To model the spectral characteristics of target/noise sources, non-negative auto-encoders were applied to gain normalized beamforming outputs. After de-noising through non-negative auto-encoders, complementarity subtraction was applied to boost the accuracy of power spectra estimation. By calculating Wiener filters frame-by-frame, output signals were obtained. Through several experiments, we confirmed that SINR improved by around 12 dB with the proposed method.

We will work on implementing a convolutional model into our proposed method since effective cues for estimating Wiener filters may be obtained from the dependency between time frames.

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] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and K. Ohmuro, "Integrated approach of feature extraction and sound source enhancement based on maximization of mutual information", *in Proc. ICASSP 2016*, pp. 186–190, 2016.

[4] K. Niwa, Y. Hioka, and K. Kobayashi, "Optimal microphone array observation for clear recording of distant sound sources", *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 24(10), pp. 1785–1795, 2016.

[5] K. Niwa, Y. Koizumi, T. Kawase, K. Kobayashi, and Y. Hioka, "Pinpoint extraction of distant sound source based on DNN mapping from multiple beamforming outputs to prior SNR", *in Proc. ICASSP 2016*, pp. 435–439, 2016.

[6] 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.

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

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

[9] 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.

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

[11] 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.

[12] 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.

[13] T. Kawase, K. Niwa, M. Fujimoto, N. Kamado, K. Kobayashi, S. Araki, and T. Nakatani, "Real-time integration of statistical model-based speech enhancement with unsupervised noise PSD estimation using microphone array", *in Proc. ICASSP 2016*, pp. 604–608, 2016.

[14] 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.

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

[16] 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.

[17] 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.

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

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

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

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

[22] T. Kawase, K. Niwa, K. Kobayashi, and Y. Hioka, "Application of neural network to source PSD estimation for Wiener filter based array sound source enhancement", *IWAENC 2016* (in press).

[23] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization", *Nature*, vol. 401, pp. 788–791, 1999.

[24] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. "Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation", Wiley, 2009.

[25] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription", *in Proc. WASPAA 2003*, pp. 177?180, 2003.

[26] C. Fevotte, N. Bertin, and J-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis", *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[27] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization", *in Proc. ICASSP 2012*, pp. 261–264, 2012.

[28] M. Kim and P. Smaragdis, "Adaptive denoising autoencoders: a fine-tuning scheme to learn from test mixtures", *in Proc. LVA/ICA 2015*, 2015

[29] G. E. Hinton, R. R. Salakhutdinoy, "Reducing the dimensionality of data with neural networks", *Science*, vol. 313(5786), pp. 504-507, 2015.

[30] K. Niwa, T. Kawase, K. Kobayashi, and Y. Hioka, "PSD estimation in beamspace using property of M-matrix", *IWAENC 2016* (in press).

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

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