SOUND SOURCE LOCALIZATION BASED ON DEEP NEURAL NETWORKS WITH DIRECTIONAL ACTIVATE FUNCTION EXPLOITING PHASE INFORMATION

Ryu Takeda and Kazunori Komatani

The Institute of Scientific and Industrial Research, Osaka University
8-1, Mihogaoka, Ibaraki, Osaka 567-0047, Japan

ABSTRACT

This paper describes sound source localization (SSL) based on deep neural networks (DNNs) using discriminative training. A naïve DNNs for SSL can be configured as follows. Input is the frequency-domain feature used in other SSL methods, and the structure of DNNs is a fully-connected network using real numbers. The training fails because its network structure loses two important properties, i.e., the orthogonality of sub-bands and the intensity- and time-information saved in complex numbers. We solved these two problems by 1) integrating directional information at each sub-band hierarchically, and 2) designing a directional activator that could treat the complex numbers at each sub-band. Our experiments indicated that our method outperformed the naïve DNN-based SSL by 20 points in terms of the block-level accuracy.

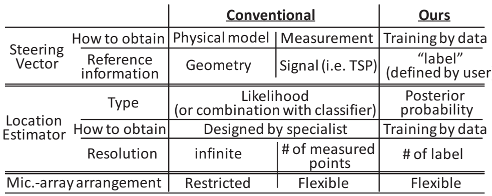
Index Terms— Sound source localization, Deep Neural Networks, Frequency domain, Discriminative training

1. INTRODUCTION

Sound source localization (SSL) is the most fundamental function for autonomous robots (or systems) [1] because it enables them to **detect sound events** and to **recognize sound locations**. These two kinds of awareness are essential for robots to start actions and to determine whether they should react to events or not. The two main difficulties with SSL on robots are: 1) restrictions on the position and the number of microphones and 2) complicated acoustic properties that depend on their bodies. The SSL on robots should be able to overcome these two difficulties.

The conventional approaches to SSL in the frequency domain obtain "steering vectors" (SVs) by using physical models [2, 3, 4, 5] or measurements [6] (Fig.1). The SVs are representations of the intensity- and time-difference between microphones from reference points in space to robots, and are used in the localization process. Here, the SVs are usually complex numbers to treat intensity and time (phase) information simultaneously. The former calculates the SVs analytically by using geometrical information, and achieves high-resolution SSL under special microphone arrangements. The latter can be applied to any microphone arrangements because it measures actual SVs at each reference point by using reference signals, such as a Time-stretched pulse (TSP). Although the latter approach resolves the two difficulties, the location estimator based on likelihood has various parameters and optimal parameters vary by the distance and the height of reference points.

Our approach is entirely based on the discriminative **machine learning** from obtaining the SVs to learning of the location estimator. This approach estimates the posterior probability of sound location directly without thresholding parameters. Since all parameters are optimized for each robot, the accuracy of localization is expected to be improved from that of previous methods. The various training data can be recorded by the robot or generated by using a statistical generative model. Note that it only requires the observed sound signals and the correct "labels" that a developer has designed for various applications. Such labels may not only include points in space, "30° from front", but also rough labels such as "Far in front".



		Conventional		**Ours**
Steering Vector	How to obtain	Physical model	Measurement	Training by data
	Reference information	Geometry	Signal (i.e. TSP)	"label" (defined by user)
Location Estimator	Type	Likelihood (or combination with classifier)		Posterior probability
	How to obtain	Designed by specialist		Training by data
	Resolution	infinite	# of measured points	# of label
Mic.-array arrangement		Restricted	Flexible	Flexible

Fig. 1. Approaches for sound source localization

We propose two techniques to apply deep neural networks (DNNs) to SSL in the frequency domain: 1) a hierarchical integration of directional information, and 2) a novel directional activator that can deal with complex numbers. Here, the directional activator is like an expression of SVs in DNNs, and it can utilize both of intensity and phase information. The activator is designed based on the orthogonality used in Multiple Signal Classification (MUSIC) [7]. Therefore, we adopt the features used in MUSIC as the input of DNNs. First, the *directional image* of real numbers is calculated by directional activators at each sub-band. Then, these directional images are hierarchically integrated step by step. The experiments reveal the robustness of DNNs in terms of the speaker. The analysis of obtained DNNs' parameters will contribute to applying DNNs to other frequency-domain signal processing.

Another applicable structure of DNNs is a fully-connected networks, and it fails in the case of the frequency-domain SSL. This is because each sub-band in the frequency domain is usually orthogonalized, and fully-connected networks destroy such a meaningful orthogonal structure. The input of DNNs in automatic speech recognition [8, 9, 10, 11] and speech enhancement area [12, 13, 14], are usually features calculated from the power spectrum. Since they are correlated at neighboring sub-bands, fully-connected networks work well as a speech feature extractor.

The DNNs with real numbers also fails due to the loss of phase information, and the importance of phase information is mentioned in [15]. Here, two solutions for complex number have been proposed: 1) complex-valued NNs (CVNNs) [15, 16, 17] and 2) real-valued feature with DNNs [18]. Some of them uses likelihood calculated from CVNNs, and others uses binaural features for the input of NNs at each sub-band. The probabilistic aspect of CVNNs is not discussed because its output is complex value. Therefore, their techniques cannot be applied directly to our situation of multi-channel SSL and posterior probability estimation.

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

2. FUNDAMENTAL METHODS

This section introduces the principle of MUSIC-based SSL and DNNs, and the problem with naïve DNN-based SSL. Hereafter, all sound signals have been analyzed by short-time Fourier transformation (STFT) and all variables in models are represented in the STFT domain with frame index t and frequency-bin index w [19].

2.1. Sound Source Localization based on MUSIC

The sound arrival process from M $(M < N)$ sound sources to the sound signals $\boldsymbol{x}_w[t] = [x_{w,1}[t], ..., x_{w,N}[t]]^T$ received at N microphones embedded on a robot are modeled as a linear time-invariant system. The observed vector $\boldsymbol{x}_w[t]$ is represented as

$$\boldsymbol{x}_w[t] \quad = \quad \sum_{m=1}^{M} \boldsymbol{a}_w(\boldsymbol{r}_m) s_{w,m}[t] + \boldsymbol{n}_w[t], \tag{1}$$

where $s_{w,m}[t]$ represents a m-th source sound signal and $\boldsymbol{n}_w = [n_{w,1}[t], ..., n_{w,N}[t]]^T$ is a noise signal vector. The $\boldsymbol{a}_w(\mathbf{r}) = [a_{w,1}(\mathbf{r}), ..., a_{w,N}(\mathbf{r})]^T$ is an SV that represents the transfer function from the reference sound position, $\boldsymbol{r}$, to each microphone. In other words, this vector includes the **intensity- and time-difference information of a signal among the microphones**. MUSIC uses the orthogonality of eigenvectors of the correlation matrix $\boldsymbol{R}_w = \mathbb{E}[\boldsymbol{x}_w[t]\boldsymbol{x}_w^H[t]]$. Here, the notation, $\cdot^H$, denotes the Hermitian transpose, and $\mathbb{E}[\cdot]$ means an expectation operator.

The linear space spanning correlation matrix $\boldsymbol{R}_w$ can be divided into two orthogonal sub-spaces: the signal space, $\mathbb{S}_s$, and the noise space, $\mathbb{S}_n$. The eigenvectors and eigenvalues of $\boldsymbol{R}_w$ are obtained by applying eigenvalue decomposition (EVD); $\boldsymbol{E}_w = [\boldsymbol{e}_{w,1}, ..., \boldsymbol{e}_{w,N}] \in \mathbb{C}^{N \times N}$ for the former and $\boldsymbol{\Lambda}_w = \text{diag}[\lambda_{w,1}, ..., \lambda_{w,N}]$ for the latter. The eigenvalues are sorted in descending order. Here, $\boldsymbol{e}_{w,i} \in \mathbb{C}^N$ $(i = 1, ..., M)$ corresponds to a basis set of signal space $\mathbb{S}_s$ and $\boldsymbol{e}_{w,j} \in \mathbb{C}^N$ $(j = M + 1, ..., N)$ corresponds to that of noise space $\mathbb{S}_n$. This means that $\boldsymbol{a}_w^H(\boldsymbol{r}_m)\boldsymbol{e}_{w,i} = 0$ $(\boldsymbol{e}_{w,i} \in \mathbb{S}_n)$ holds over the correct sound positions, $\boldsymbol{r}_m$ $(m = 1, ..., M)$. Note that these eigenvectors have already been *normalized* in terms of features. The actual estimator using this orthogonality can be seen in [20].

2.2. Model and Learning of Neural Networks

The structure of NNs is defined recursively on the layer index, l. The input vector, $\mathbf{x}_l = [x_{l,1}, ..., x_{l,N_l}]^T \in \mathbb{R}^{N_l}$, is projected into output vector $\mathbf{x}_{l+1} = [x_{l+1,1}, ..., x_{l+1,N_{l+1}}]^T \in \mathbb{R}^{N_{l+1}}$ by arbitrary function $\mathbf{f}_l$. The final output of the L-th layer can be recursively described for $l = 0, ..., L-1$ given the initial input vector, $\mathbf{x}_0$.

$$\mathbf{x}_{l+1} \quad = \quad \mathbf{f}_l(\mathbf{x}_l; \boldsymbol{\theta}_l) \tag{2}$$

where $\boldsymbol{\theta}_l$ is a parameter set of $\mathbf{f}_l$. There are several types for the function, $\mathbf{f}_l$. For example, the affine transformation, $\mathbf{W}_l\mathbf{x}_l + \mathbf{b}_l$, is used to represent network links, and the sigmoid function, $1/(1 + \exp(-x_{l,i}))$, is used to express the activation of each vector.

Back propagation is applied to optimize the parameters, $\boldsymbol{\theta}_l$, by using the training data set. Given the cost function, E, and supervisory signal vector $\mathbf{r} = [r_1, ..., r_{N_L}]^T \in \mathbb{R}^{N_L}$, the parameter update rules can also be recursively described. After the initial error vector, $\boldsymbol{\epsilon}_L = \left(\frac{\partial E}{\partial \mathbf{x}}(\mathbf{r}, \mathbf{x}_L)\right)$, is calculated, we update each parameter for $l = L-1, ..., 0$ with a learning parameter η as:

$$\boldsymbol{\epsilon}_l = \frac{\partial \mathbf{f}_l^T}{\partial \mathbf{x}}(\mathbf{x}_l)\boldsymbol{\epsilon}_{l+1}, \quad \boldsymbol{\theta}_l \leftarrow \boldsymbol{\theta}_l - \eta \frac{\partial \mathbf{f}_l^T}{\partial \boldsymbol{\theta}_l}(\mathbf{x}_l)\boldsymbol{\epsilon}_{l+1}. \tag{3}$$

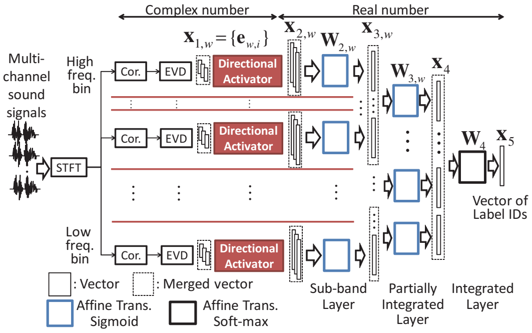


Fig. 2. Network structure of our DNN-based SSL

2.3. Training of DNN-based SSL and Its Problems

Since MUSIC estimates sound locations by using eigenvectors, the main role of DNNs for SSL is obtaining the mapping from eigenvectors $\mathbf{e}_{w,i}$ to the probability, $p_k(k = 1, ..., K)$, of reference positions $\mathbf{r}_k(k = 1, ..., K)$ or labels. The naïve configuration of DNNs is a fully-connected network with real numbers that is often used in the speech recognition and speech enhancement area. Here, complex numbers are considered to be two dimensional real numbers. However, the training of this configuration for DNNs does not work well and results in insufficient accuracy.

The two problems with these types of DNNs is loss of: 1) orthogonality of sub-bands and 2) intensity- and time-information in complex numbers. Our features at each sub-band are almost orthogonalized by FFT unlike those of speech recognition or speech enhancement in the power spectrum domain. Therefore, applying DNNs with fully-connected network and real numbers to our features wastes the structural information of each value, especially time information, which is important in SSL.

3. NETWORK CONFIGURATION FOR SOUND SOURCE LOCALIZATION

This section explains the hierarchical structure and complex-number activator to solve problems in training DNNs. First, the network architecture is explained, and then the details on the activator that was designed are provided. Note that complex-number networks can be expressed by real-number networks with special structures. Therefore, complex numbers are just used as mathematical expressions.

3.1. Network Architecture for Frequency Domain Processing

Our proposed network architecture is based on a hierarchical structure among sub-bands, and there is an overview of this in Fig. 2. The process can be divided into two phases: 1) the extraction of a *directional image* and 2) the propagation and integration of a directional image. Here, the directional image is an activation pattern that differs according to the SVs of sound sources.

The directional image is extracted by using the orthogonality of the input eigenvectors and the *SVs* in DNNs, which is the same as MUSIC does. First, the input signals are analyzed by STFT and the correlation matrices $\mathbf{R}_w$ are calculated at each frequency bin w. Then, EVD is applied and we obtain the eigenvectors, $\mathbf{e}_{w,i}$. These eigenvectors are treated as the input vector, $\mathbf{x}_{1,w} = [\mathbf{e}_{w,2}^T, ..., \mathbf{e}_{w,N}^T]^T$, at frequency bin w. We calculate the *directional image* $\mathbf{x}_{2,w}$ from eigenvectors at each w, whose details are explained in the next subsection.

The *directional images* are integrated in three hierarchical steps using the ordinal network structure based on the affine transformation, sigmoid, and soft-max function. This is because the directional image at neighboring sub-bands has a correlation to some extent. The first step is the integration at each sub-band, and the *sub-band layer* at each sub-band outputs new directional image. The second step is the integration among sub-bands, and the input of the *partially integrated layer* connects directional images from several sub-bands. For example, when the output of the sub-band layer at w is noted by $\mathbf{y}_{2,w}$, the input of the l-th partially integrated layer is $\mathbf{x}_{3,l} = [\mathbf{y}_{2,w_l}^T, ..., \mathbf{y}_{2,w_h}^T]^T$. Here, w_l and w_h represent the lower and upper index for the integration. These layers also output directional images. The last step is the integration of the outputs from the sub-integrated layer, and we call it as *integrated layer*. Its inputs, $\mathbf{x}_4$, have the same structure as the partially integrated layer.

3.2. Model and Training of Directional Activators

Work that remained was the modeling and training of directional activators that output the directional image. We designed the activators using the orthogonality of eigenvectors and SV used in MUSIC. The DNNs learn these directional activators that work like the SVs through discriminative training.

We define the directional activators by using latent vectors $\mathbf{a}_j(||\mathbf{a}_j|| = 1)$ that are expected to behave as the SVs. These activators are based on the following inner product that can simultaneously measure intensity and time-differences.

$$\mathbf{f}_w(\mathbf{x}) = [f(\mathbf{x}; \mathbf{a}_{w,1}), ..., f(\mathbf{x}; \mathbf{a}_{w,N})]^T, \;\; f(\mathbf{x}; \mathbf{a}) = 1 - \frac{|\mathbf{a}^H \mathbf{x}|}{||\mathbf{x}||}. \tag{4}$$

If the latent vector, $\mathbf{a}_k$, corresponds to a true SV of position $\mathbf{r}_k$ in an ideal case, the correct directional activator returns 1 because the eigenvectors and correct SV are orthogonal. The directional image is defined by the connected outputs of all activators with all eigenvectors in noise space, $\mathbf{x}_{2,w} = [\mathbf{f}_w(\mathbf{e}_{w,2})^T, ..., \mathbf{f}_w(\mathbf{e}_{w,N})^T]^T$. After this, we will summarize the parameters into a matrix representation, $\mathbf{A}_w = [\mathbf{a}_{w,1}, ..., \mathbf{a}_{w,N}]$, at frequency-bin w. This activation process is similar to the combination of the linear projection by $\mathbf{A}_w$ and the activation by absolute function in the matrix formation.

The propagation error and update rule of the parameters of directional activators are obtained by calculating gradients.

$$\frac{\partial \mathbf{f}_w^H}{\partial \mathbf{x}}(\mathbf{x})\boldsymbol{\epsilon}_{w,l+1} = -\frac{1}{|\mathbf{x}|}\left(\mathbf{A}_w \mathrm{diag}[\boldsymbol{\theta}_w] - \frac{\mathbf{x}}{|\mathbf{x}|}\boldsymbol{\phi}_w^T\right)\boldsymbol{\epsilon}_{w,l+1}, \tag{5}$$

$$\frac{\partial \mathbf{f}_w^H}{\partial \mathbf{A}_w}(\mathbf{x})\boldsymbol{\epsilon}_{w,l+1} = -\frac{\mathbf{x}}{|\mathbf{x}|}\left(\boldsymbol{\theta}_w^H + \mathbf{A}_w \mathrm{diag}[\boldsymbol{\phi}_w]\right)\mathrm{diag}[\boldsymbol{\epsilon}_{w,l+1}] \tag{6}$$

Here, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ represent the phase and similarity vector defined as:

$$\boldsymbol{\theta}_w = \left[\frac{\mathbf{a}_{w,1}^H \mathbf{x}}{|\mathbf{a}_{w,1}^H \mathbf{x}|}, ..., \frac{\mathbf{a}_{w,N}^H \mathbf{x}}{|\mathbf{a}_{w,N}^H \mathbf{x}|}\right]^T, \boldsymbol{\phi}_w = \left[\frac{|\mathbf{a}_{w,1}^H \mathbf{x}|}{||\mathbf{x}||}, ..., \frac{|\mathbf{a}_{w,N}^H \mathbf{x}|}{||\mathbf{x}||}\right]^T \tag{7}$$

After the parameters are updated, the norm of each activation vector $\mathbf{a}_k$ is normalized to 1. The Eq. (5) is not used in this paper because there are no parameters before the layers of directional activators.

4. EXPERIMENTS

4.1. Experimental setups

Recording conditions: All speech data were generated by using impulse responses recorded in a real environment. Four-channel impulse responses were recorded at 16 kHz in both an anechoic room and a reverberant room with an RT_{20} of 640 [ms] by using microphones embedded on a humanoid NAO [21]. RT_{20} means the reverberation time. We denoted the loudspeaker positions as (distance [cm] and height [cm]). The combinations of patterns for recording impulse responses were set to (30,30), (90,30) and (90, 90) to take into consideration situations in which people talked to the robot from different distances and heights. The resolution of the directional angle was 5° (72 directions), as shown in Fig. 3. There were a total of 216 recorded impulse responses.

Feature extraction: The parameters for STFT were set to be the same for all the experiments: the size of the Hamming window was 256 points (16 [ms]) and the shift size was 80 points (5 [ms]). The block size for calculating $\boldsymbol{R}_w$ was 40 (200 [ms]). The bandwidth used for features was set to $[750 - 4750]$ [Hz] and 64 frequency-bins were used for SSL. These configurations are listed in Tab.1.

Data for training and test set: The speech data for training came from 49 male and 49 female speakers in the Acoustical Society of Japan-Japanese Newspaper Article Sentences (ASJ-JNAS) corpora[1], and one hour of data was used. The data for test came from one male and one female speaker, which was different from the training data in the same corpora. There was an average of seven utterances per speaker, and the content was phonetically balanced sentences. The training and test set generated by using impulse responses included speech signals from a combination of 72 directions (5° intervals) × 3 positions × speaker patterns. After the speech signal was generated, we added white noise of 0, 20 and 40 dB to check the robustness of each method. The total number of labels was 217. The label ID "0" represents "no sound source", the others represent source locations; IDs 1-72 for the azimuth 0-355° at (30, 30), IDs 73-144 for the azimuth at (90, 30) and IDs 145-216 for the azimuth at (90, 90), respectively. The correct labels are added based on voice activity block by block (every 200 [ms]).

Configuration of DNNs: The configuration of naïve DNNs was that of $L = 7$ layers with 1024 hidden nodes. There were four dimensions and 216 directional activators in the w-th sub-band in our DNNs. The 216 was the same number of recorded impulse responses for the analysis. There were eight blocks in the partially-integrated layer and integrated layer. The network sizes of the sub-band, partially-integrated and integrated layer corresponded to 217×648, 217×1736, and 217×1736. There were a total of 960 dimensions of features for DNN input. The output dimensions were 217 to classify all labels. The directional activators are initialized at random and their norms are normalized to 1. The initial weight of $\mathbf{W}_{2,w}$ was like an identity matrix; the element of the i-th row and i-th column was 1, and others were Gaussian noise with variance 0.0001. The initial weights of the partially-integrated and integrated layers were obtained by connecting eight such weights. These initial parameters empirically enable us to interpret the trained parameters easily. The cross-entropy was used as the cost function E.

Evaluation criteria: We calculated the accuracy of classification at the block-level. The three methods we compared were are naïve DNN-based SSL, the proposed SSL, the basic MUSIC used in [20], the Bartlett's and Capon's beamformer [22]. The broadband spacial/MUSIC spectrum is calculated by summing the narrowband spectra. We chose a best threshold of the broadband spatial/MUSIC spectrum for each test set for the discrimination of source existence. **Note that this criterion is not based on the geometrical distance**. We checked robustness against 1) speaker, 2) SNR of white noise and 3) reverberation. We prepared two kinds of data under different conditions, i.e., those in an anechoic and a reverberant room.

[1] http://research.nii.ac.jp/src/JNAS.html

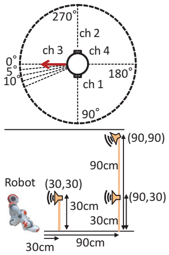


Fig. 3. Positions

Table 1. Parameters of experiment

Parameter	Value
Number of sources	0 or 1 at each block
Noise signal	Gaussian
Training	49 males, 49 females
Test	1 male and 1 female **(speaker open)**
Sampling frequency	16kHz
Frame length and shift	16 ms and 5 ms
Block size	200 ms (40 frames)
Bandwidth ($[W_l \;\; W_h]$)	[570 4750] Hz

Table 2. Block-level accuracy of classification at each SNR (%).

	Training	Test					
	Anechoic	Anechoic			Reverberant		
Method	SNR (dB)	40	20	0	40	20	0
Bartlett	—	80.9	80.5	60.7	45.7	45.7	45.7
Capon	—	81.2	80.7	60.8	45.7	47.0	45.7
MUSIC	—	77.3	83.8	68.9	45.8	51.4	48.8
Naïve DNNs	40	67.8	48.3	45.6	45.8	45.6	45.6
	20	53.3	65.3	47.0	37.9	47.7	45.7
	0	22.0	32.0	50.2	0.8	22.5	45.3
Ours	40	**89.3**	55.3	45.6	45.7	45.6	45.6
	20	57.8	**86.9**	53.6	45.9	**51.6**	45.8
	0	54.4	62.3	**74.2**	19.0	37.3	48.1

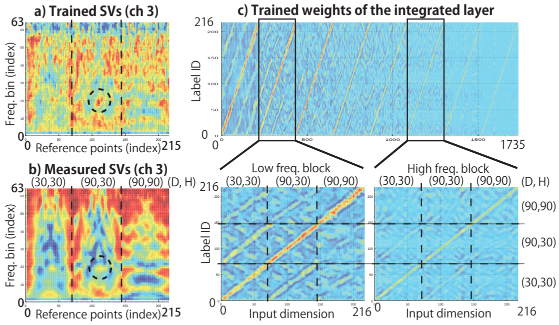


Fig. 4. The images of the smoothed absolute SVs ((a), (b)), and the smoothed network weight $\mathbf{W}_4$ (c). The notation (D, H) represents the combination of the distance and height [cm].

4.2. Results and Analysis

The accuracy of each method is summarized in Table 2. Here, *Anechoic* in *Test* means an environmentally closed test, and *Reverberant* means an environmentally open test ($\mathrm{RT}_{20} = 640$[ms]). Note that the percentage of "no sound source" blocks is 45.6%, and the total number of blocks in the test set is 108432.

First, the maximum accuracies of Bartlett, Capons and MUSIC in the anechoic room were 80.6, 81.9 and 83.8%, respectively. Since the optimum threshold parameters of these three methods vary at each distance and height, it is difficult to perform best with one threshold parameter. The main reason of the low performance in the reverberant room is that the peak of the estimator moved slightly from the correct location.

Second, our method outperformed the naïve DNNs in all cases. The accuracy of naïve DNNs was a maximum of 67.8% at an SNR of 40 dB in the anechoic room, and the accuracy is less than that of MUSIC. On the other hands, if we use our DNNs trained by data matched SNR with test set, its performance becomes better than that of MUSIC. Accuracies of all methods in the reverberant room decrease compared with those in the anechoic room. If we train DNNs by using reverberant and multi-SNR data, its performance will improve. Our preliminary result showed that the training of DNNs with multi-SNR data succeeded. Therefore, the multi-condition training (MTC) is a key technique for the further improvement. The main problem will be how we generate various kinds of reverberant and noisy data appropriate for each robot (or system).

Figure 4 have the images of smoothed directional activators (=SVs) and network weight $\mathbf{W}_4$ obtained with the SNRs of 0 and 40 dB training sets. The images of a) and b) are the trained and the measured SVs of microphone channel 3. Although the trained SVs look like the measured one, the some regions differ from those of the measured one (circular dashed line). This indicates that the SVs are adapted to robots through training. We can see the importance of each frequency block and filter pattern of them from the image c). Since the power of speech signal corresponds to high-freq block is lower than that of low-freq. block, we can understand that the DNNs automatically weight information from low-freq. block. It is interesting that the striped patterns become detailed at high-freq. block. These filter patterns work as the integrating and smoothing filter of the directional images obtained from each block.

4.3. Remained Issues

Since the robustness against speaker was confirmed. we should improve the robustness against 1) reverberation, 2) non-Gaussian noise, 3) unknown direction and 4) the number of sound sources. The common approaches for these robustness, especially for 1) and 2), are MCT based on data generation and DNNs' configuration. The construction of the **generative model for training** is an important topic for our machine learning approach. The generation of reverberant speech signals will be the key technique for localization in reverberant environment. The optimization of DNNs' parameters, such as dimensions of each weight and activator, should also be investigated. Other matrix decomposition or sound source separation methods may also be applied instead of EVD.

The robustness for 3) and 4) will require well-designed configurations of DNNs. The DNNs used in this paper cannot localize the unknown direction that did not appear in the training set. We need to design I) the structure of DNNs, and II) the output labels and its probability used in the cross-entropy because they must associate the unknown direction with the known directions included in the training set. We also need to design the output probability to deal with multiple source situations in addition to the MCT with several sound sources. The output label will include not only one source case but also two source case, such as "0 and 60 degree". In such case, DNNs may obtain a directional activator that reacts when there are two sound sources at specific locations.

5. CONCLUSION

We proposed SSL based on DNNs that works in the frequency domain. The key ideas to realize DNNs-based SSL are 1) constructing a hierarchical network structure that integrated sub-band information step by step and 2) designing a novel directional activator that could treat complex numbers. Experiments demonstrated that our method outperformed the naïve DNN-based SSL.

Future work mainly involves improving robustness against the reverberation and the number of sound sources. Moreover, a suitable configuration of DNNs for SSL should be researched more because it seriously affects performance.

Acknowledgment This work was partly supported by JSPS KAKENHI Grant Number 15K16051 and by SCOPE of MIC.

6. REFERENCES

[1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. of 17 th National Conf. on Artificial Intelligence*, 2000, pp. 832–839.

[2] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[3] B. D. Rao and K. V. S. Hari, "Performance analysis of Root-MUSIC," *IEEE Trans. Signal Processing*, vol. 37, no. 12, pp. 1939–1949, 1989.

[4] A. Parthasarathy, S. Kataria, L. Kumar, and R. M. Hegde, "Representation and modeling of spherical harmonics manifold for source localization," in *Proc. of ICASSP*, 2015, pp. 26–30.

[5] M. J. Taghizadeh, S. Haghighatshoar, A. Asaei, P. N. Garner, and H. Bourlard, "Robust microphone placement for source localization from noisy distance measurements," in *Proc. of ICASSP*, 2015, pp. 2579–2583.

[6] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environment," in *Proc. of IROS*, 2009, pp. 664–669.

[7] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Anttenas and Propagation*, vol. AP-32, no. 3, pp. 276–280, 1986.

[8] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transaction," in *Proc. of ASRU*, 2011, pp. 24–29.

[9] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using contex-dependent deep neural network," in *Proc. of Interspeech*, 2011, pp. 437–440.

[10] G. Hinton, L. Deng, D. Yu, G. E. Geroge, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and others, "Deep neural networks for acuostic modelling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[11] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocaburary speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 20, no. 6, pp. 82–97, 2012.

[12] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270–279, 2013.

[13] Y. Xu, J. Du, L.R. Dai, and C.H. Lee, "An experimental study on speech enhancement based on deep neural networks," in *IEEE Signal Processing Letters*, 2014, vol. 21, pp. 65–68.

[14] S. Nie, H. Zhang, X. L. Zhang, and W. Liu, "Deep stacking networks with time series for speech separation," in *Proc. of ICASSP*, 2014, pp. 6667–6671.

[15] W.-H. Yang and K.-K. Chan P.-R. Chang, "Complex-valued neural-network for direction-of-arrival estimation," *Electronics Letters*, vol. 30, no. 7, pp. 574–575, 1994.

[16] H. Tsuzuki, M. Kugler, S. Kuroyanagi, and A. Iwata, "An approach for sound source localization by complex-valued neural network," *IEICE Trans. on Information and Systems*, vol. 96, no. 10, pp. 2257–2265, 2013.

[17] K. Terabayashi, R. Natsuaki, and A. Hirose, "Ultrawideband direction-of-arrival estimation using complex-valued spatiotemporal neural networks," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 25, no. 9, pp. 1727–1732, 2014.

[18] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *Proc. of Interspeech*, 2015, pp. 3302–3306.

[19] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. of ICASSP*, 2008, pp. 85–88.

[20] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and implementation of selectable sound separation on the texai telepresence system using HARK," in *Proc. of ICRA*, 2011, pp. 2130–2137.

[21] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. O. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic design of Nao humanoid," in *Proc of ICRA*, 2009, pp. 769–774.

[22] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.