UNSUPERVISED ADAPTATION OF NEURAL NETWORKS FOR DISCRIMINATIVE SOUND SOURCE LOCALIZATION WITH ELIMINATIVE CONSTRAINT

Ryu Takeda[1], *Yoshiki Kudo*[2], *Kazuki Takashima*[2], *Yoshifumi Kitamura*[2] *and Kazunori Komatani*[1]

[1]The Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan
[2]Research Institute of Electrical Communication, Tohoku University, Sendai, Japan

ABSTRACT

This paper describes an unsupervised adaptation method of deep neural networks (DNNs) regarding discriminative sound source localization (SSL). DNNs-based SSL and its unsupervised adaptation fail under different conditions from those during training. The estimations sometimes include incoherent unpredictable errors due to the NN's non-linearity. We propose an eliminative posterior probability constraint using a model-based SSL for unsupervised DNNs adaptation. This constraint forces the probability of "less possible candidates" to become zero to eliminate incoherent errors. The candidates are indicated by a model-based SSL method because it can estimate the azimuth of the sound source with moderate accuracy and explicit reasoning. As a result, the localization performance of adapted DNNs improved more than that of model-based SSL. Experimental results showed that our method improved localization correctness of 1D azimuth and 3D regions by a maximum of 13.3 and 5.9 points compared with the model-based SSL.

Index Terms— sound source localization, neural networks, unsupervised adaptation

1. INTRODUCTION

1.1. Motivation

Sound source localization (SSL) is a necessary function for autonomous systems, such as robots [1], because it enables them to detect sound and determine its location. Such systems are expected to work robustly in unknown environments. SSL that is based on a data-driven machine learning approach, including the use of deep neural networks (DNNs) [2, 3, 4, 5, 6, 7, 8], is compatible with such autonomous systems for three reasons: the number of microphones and their arrangement on systems may not be decided on the basis of SSL performance, this type of SSL automatically calibrates any configurations to systems, and systems can collect training data by themselves if necessary. Here, discriminative DNNs directly estimate short-time posterior probabilities of the "*position labels*" corresponding to the presence of sound source and 3D regions in space.

The adaptation of DNNs to unknown environments is essential because the performance of DNNs degrades under conditions that are different from those during training, and unknown patterns will inevitably appear for real-world use. DNNs are very sensitive to the "condition/environment" that includes source position, reverberation and signal-to-noise ratio (SNR). We previously tried an utterance-wise unsupervised adaptation of DNNs-based SSL using entropy cost function [9]. This method adapts some parameters of NNs to observed speech signals. Performance under unknown conditions improved slightly, and the empirical early stopping method [10] was required to avoid adaptation failures. Since the minimization of entropy is a necessary condition but not enough to satisfactorily improve performance, more specific adaptation criterion is required for further stable performance improvement.

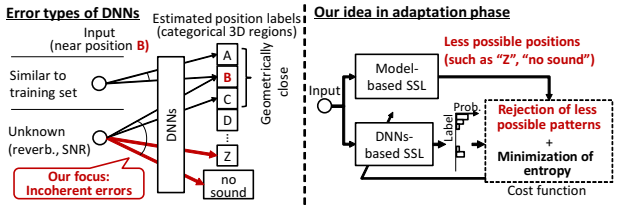

Fig. 1. Problems regarding DNNs-based SSL and our approach

The errors of DNN-based SSL under unknown conditions have been categorized into two types of error: a) close errors and b) incoherent errors (left-hand side of Fig. 1). The former means that the geometrical distance between the estimated source location and the ground truth is short. For example, the ground truth is 0° in azimuth, and the estimation is 10°. The latter pattern is caused by the non-linearity of DNNs or insufficient training data, which is difficult to explain rationally. For example, the ground truth is 90° in azimuth, but DNNs confidently estimate it as 15°. DNNs also tend to output "no sound" labels for unknown conditions or after failure to adapt. On the contrary, an advantage of traditional model-based SSL [11] is the explicit reasoning that provides for its estimations and the moderate performance under various conditions, while DNNs are specialized to the conditions similar to those encountered during training.

We propose an eliminative constraint to suppress suspicious posterior probability using a model-based SSL to avoid incoherent errors during adaptation (right-hand side of Fig. 1). The reduction of incoherent errors is one of the minimum requirements for creating stable behavior in DNNs. Our constraint forces the estimated posterior probability of "less possible candidates" to become zero. These candidates are indicated by a model-based SSL method that estimates the azimuth and the presence of a sound source in a more stable manner than DNNs under unknown conditions. Our method makes the adaptation procedure more stable and improves localization correctness both of 3D regions and 1D azimuth better than those of model-based SSL alone. *Our approach does not assume a specific DNNs configuration and can be applied even after the adaptation of the model-based SSL.* We conducted experiments to assess the performance of the adaptation method under *completely open conditions* compared with the training conditions in terms of microphone devices, reverberation, source positions, and SNR.

Our contributions to the SSL area are 1) the stable unsupervised adaptation method combined with model-based and DNNs-based SSL and 2) the analysis and discussion of DNNs behavior trained by data with over 10,000 position patterns.

1.2. Relation to Prior Work

Unsupervised adaptation of DNNs-based SSL, especially in 3D space, has not been studied much to the best of our knowledge while more NNs-based SSL methods have been proposed recently [5, 7, 12]. NNs-based SSL has been widely studied with narrow-band antennas [3], but the topic of adaptation has not been researched because array environments do not change dramatically. The azimuth, not the height or depth, is usually estimated because the far-field model is usually assumed.

DNNs used in automatic speech recognition (ASR) have several model adaptation methods that can avoid overfitting, mainly regarding speaker adaptation [13, 14]. They focus on speech signals instead of spatial information, which is important for SSL. These adaptation methods usually require several utterances (more than five) for adaptation while *one utterance or fewer is desirable for SSL*. Unsupervised adaptations are based on a statistical generative model and maximum likelihood estimation, such as constrained maximum likelihood linear regression (CMLLR) [15]. More popular *semi*-supervised approaches use a linear input network (LIN) [16] or a linear hidden network (LHN) [17]; A posterior of parameters is maximized for cost functions to avoid overfitting [18].

2. BASELINE: UNSUPERVISED ADAPTATION USING ENTROPY MINIMIZATION

This section is an overview of unsupervised adaptation of DNNs, as proposed in [9]. In this paper, all the variables in the models are represented in the short-time Fourier transformation (STFT) domain with frame index t and frequency-bin index w.

2.1. Posterior Probability Estimation using DNNs

The DNNs estimate the posterior probability $p(z|\boldsymbol{f})$ of discrete variable z with K location labels from input feature $\boldsymbol{f}$. The discrete labels z are defined by the system developer in accordance with the required resolution of the application. For example, the label "no-sound" represents a sound without a source (only noises), and the label "0°" represents a sound source located in the range of $[-2.5°, 2.5°]$ in azimuth. We assume that the label set is the combination of the patterns for depth, height, and azimuth (0°, 5°, ..., 355°) plus "no-sound". The DNNs are trained by using various data with different sound positions, reverberations, and so on.

The overview of the input features and DNNs' configuration is shown in Fig. 2. We used a set of noise-space eigenvectors as the input of DNNs [19, 6], and the last layer of DNNs was a soft-max function to represent the posterior probability of location labels. The DNNs can deal with non-speech signals and multiple sound sources [8]. Since the concept of this research is not strongly influenced by the configuration of DNNs, please see [9] for more detail.

We will explain the process to obtain the input features because they are also used in the model-based SSL. The process is as follows: 1) Calculate a correlation matrix $\boldsymbol{R}_w = \mathbb{E}[\boldsymbol{x}_w \boldsymbol{x}_w^H]$ of the observed signal vector $\boldsymbol{x}_w[t] = [x_{w,1}[t], ..., x_{w,N}[t]]^T$ from N microphones. 2) Apply eigenvalue decomposition (EVD) to $\boldsymbol{R}_w$ and sort eigen vectors in descending order of eigen values. 3) Select $N - M$ eigenvectors $\boldsymbol{e}_{w,j}$ in noise-space $\mathbb{S}_n$. We assume $M = 1$ to localize one sound source. 4) Extract the following feature: $\boldsymbol{f}$ is a set of eigenvectors $[\boldsymbol{e}_{w_l,M+1}^T, ..., \boldsymbol{e}_{w_l,N}^T, ..., \boldsymbol{e}_{w_h,M+1}^T, ..., \boldsymbol{e}_{w_h,N}^T]^T$ where w_l and w_h are respectively the lower and upper indices of the frequency bin used for localization. This feature is extracted every 110 milliseconds because of the mean operation for $\boldsymbol{R}_w$ (block-wise).

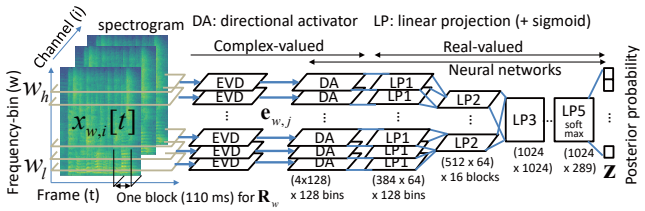

Fig. 2. Configuration of DNNs

2.2. Parameter Adaptation by Entropy Minimization

Some parameters of the DNNs are updated by minimizing the cost function in the adaptation phase. The following entropy J_e was used for the cost function in previous work because the minimum entropy (ambiguity) is one of the necessary conditions for the ideal state in terms of discrimination:

$$J_e(\boldsymbol{\Theta}) = \mathbb{E}[-\textstyle\sum_i p_i \log p_i], \quad \dfrac{\partial J}{\partial p_i}(\boldsymbol{\Theta}) = \mathbb{E}[-\log p_i - 1], \qquad (1)$$

where Θ represents the target parameters for updates in the DNNs and $p_i = p(i|\boldsymbol{f})$ represents the estimated probability of label i from the i-th output node z_i of the DNNs as shown in Fig. 2. The expectation means taking the average of the blocks used for adaptation, such as one utterance or fewer. By applying the chain rule and taking the partial derivatives of the output variables of each layer, we can calculate the gradient of the target parameters for the updates.

A linear input network (LIN) [16] was used for the parameters for adaptation. Each eigenvector in the input feature $\boldsymbol{f}$ is linearly transformed, and the update rules are as follows:

$$\hat{\boldsymbol{e}}_{w,i} = \boldsymbol{V}_w \boldsymbol{e}_{w,i} + \boldsymbol{b}_w, \ (i = M+1, ..., N), \qquad (2)$$

$$\boldsymbol{V}_w \leftarrow \boldsymbol{V}_w - \alpha \textstyle\sum_i \boldsymbol{\delta}_{w,i} \boldsymbol{e}_{w,i}^H, \quad \boldsymbol{b}_w \leftarrow \boldsymbol{b}_w - \alpha \textstyle\sum_i \boldsymbol{\delta}_{w,i}, \qquad (3)$$

where $\boldsymbol{V}_w \in \mathbb{C}^{N \times N}$ and $\boldsymbol{b}_w \in \mathbb{C}^N$ represent a complex-valued matrix and bias vector, respectively. $\boldsymbol{\delta}_{w,i}$ is a propagated error vector corresponding to each input vector. The weight $\boldsymbol{V}_w$ is shared at each frequency bin, and its initial value is an identity matrix. This transformation is expected to modify the error of the extracted eigenvectors caused by reverberation or noise. The early stopping technique [10] was applied to avoid over-fitting or trivial solutions.

3. PROPOSED METHOD

The overview of our adaptation process is shown in Fig. 3. First, the model-based SSL outputs localization scores for each azimuth label. Then, we select less possible candidates on the basis of these scores. Finally, the parameters of the DNNs are updated by minimizing the cost function on the basis of entropy with eliminative constraint expressed by weight $\boldsymbol{m}$ for such candidates. We explain the cost function, and then the calculation of the weight and model-based SSL.

3.1. Cost Function using Eliminative Constraint

We design the following new cost function J for adaptation:

$$J(\boldsymbol{\Theta}) = J_e(\boldsymbol{\Theta}) + \lambda \mathbb{E}[\textstyle\sum_i m_i p_i^2], \qquad (4)$$

where λ is a weight for a constraint term and m_i is a weight parameter of p_i. The second term is an eliminative constraint to reject incoherent (less possible) candidates. The weight m_i becomes 1 if the corresponding location label i is incoherent. $m_i = 0$ means that the label seems confident. For example, if the azimuth estimation

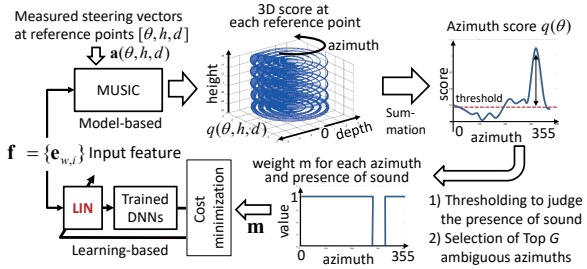

Fig. 3. Overview of our method

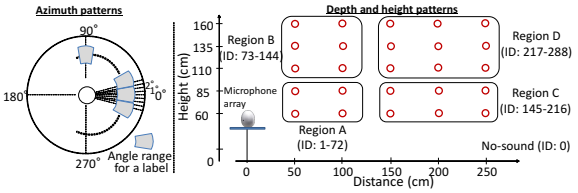

Fig. 4. 1) Locations for impulse responses. 2) Mapping IDs

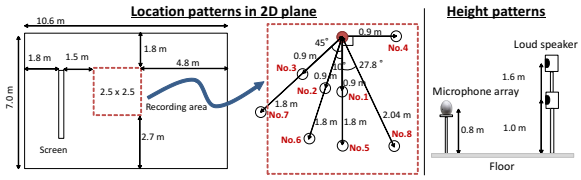

Fig. 5. Locations for test set. The mic-array and loud speaker are different from those used in Fig. 4.

is confident but its depth and height are not, the weight for these depth and height labels at the same azimuth becomes 0. Since the minimum of the second-term is zero, our constraint does not suffer entropy minimization.

3.2. Estimation of Less Possible Candidates based on MUSIC

We used a multiple signal classification (MUSIC) method as a model-based SSL. This method calculates a type of likelihood of source location $q(\theta, d, h)$ at azimuth θ, depth d, and height h. We represent the source location in space as $\boldsymbol{r} = [\theta, d, h]$. The advantages of MUSIC are the high directional resolution to distinguish the locations and the scalability for multiple sound sources.

The sound source position $\boldsymbol{r}$ can be estimated on the basis of the orthogonality between eigenvectors in noise-space $\boldsymbol{e}_{w,i} \in \mathbb{S}_n$ and the reference steering vectors $\boldsymbol{a}(\boldsymbol{r})$. The reference steering vectors can be obtained from the analytical transfer function model or the measured impulse responses at discrete points $\boldsymbol{r}_v = [\theta_v, d_v, h_v], (v = 1, ..., V)$. The MUSIC score $q_w(\boldsymbol{r})$ at frequency bin w is defined as

$$q_w(\boldsymbol{r}) \quad = \quad 1/(\textstyle\sum_{i=M+1}^{N} |\boldsymbol{a}_w^H(\boldsymbol{r})\boldsymbol{e}_{w,i}|/||\boldsymbol{a}_w(\boldsymbol{r})||). \qquad (5)$$

The function gives a high value when $\boldsymbol{r}$ is the true position of the source $\boldsymbol{r}_m$. The broad-band score q is calculated by averaging q_w from a lower bin w_l to an upper bin w_h. Then, we calculate the azimuth score $q(\theta_v) = \sum_{d,h} q(\theta_v, d, h)$ with summation over height and depth because the discrimination of depth and height is usually difficult if we cannot design a good microphone arrangement.

We estimate the weight $\boldsymbol{m}$ by using reliable azimuth scores. The default value of weight is zero. First, a threshold is given to the highest score with a parameter T_{th} to judge the presence of sound. If the score is less than T_{th}, the weight that corresponds to the sound presence label becomes 1. If not, we choose the top-G ambiguous candidates in descending order of score. The weight that corresponds to the location whose azimuth is equal to the G candidates remains 0, and the weight of others becomes 1. The DNNs search the best locations from among the restricted candidates whose weights are defined as 1 during adaptation.

4. EXPERIMENTS

4.1. Experimental Settings

Recording conditions: All speech data for training were generated using impulse responses recorded in real anechoic and reverberant rooms. The size of the reverberant room was 7.83 [m] × 5.87 [m] × 2.57 [m] (depth x width x height), and its reverberant time was about RT_{20} 700 [ms]. Four-channel impulse responses were recorded at 16 kHz by using microphones horizontally attached to egg-shaped

Table 1. Experiment Parameters

Number of sources	0 or 1 at each block
Training source	Speech of 48 males and 48 females
Test source	2 males and 2 females (**speaker open**)
Impulse responses for training	Anechoic room Reverberant room (RT_{20} 800 [ms])
Position patterns for training data	18,000 = 360 (azimuth) × 5 (depth) × 5 (height) × 2 (room)
Position labels	289 = 72 azimuths × 4 + no-sound
DNN Input / Output dim.	1536 ($\{\boldsymbol{e}_{i,w}\}_{i=2,3,4,w=21,...,148}$) / 289
DNN Middle Layer	Shown in Fig. 2

surface[1]. The resolution of the azimuth was 1° (360 directions), and there were 25 combinations of distance and height, as shown in Fig. 4. The speech signal data for the test set were recorded with a different loud speaker in a different reverberant room as shown in Fig. 5. There were 16 location patterns in total: eight positions in 2D-plane [No.1, ..., No.8] × two heights [1.0 m, 1.6 m].

Feature extraction: The STFT parameters were set to be the same for all experiments: the size of the Hamming window was 512 points (32 [ms]), and the shift size was 160 points (10 [ms]). The block size for calculating $\boldsymbol{R}_w$ was 11 (110 [ms]). The bandwidth used for the features was set to $[656 - 4625]$ [Hz], and 128 frequency bins from 21 to 148 in the bin-index were used for SSL.

Data for training and test sets: The speech data for the training came from 48 male and 48 female speakers using the Acoustical Society of Japan-Japanese Newspaper Article Sentences (ASJ-JNAS) corpora[2] (one hour in total). The data for the test came from two male and two female speakers and were different from the training data in the same corpora. There were fifteen utterances in total, and the content had phonetically balanced sentences. The training data were generated using all four-channel impulse responses, and we randomly selected 15% from them. Gaussian noise of 20 dB was added to the speech signals of the training set, and 0, 10 and 20 dB were added to the test set. The total number of labels was 289 and the resolution in azimuth for localization was 5°. The label ID "0" represents the "no sound source", and the others represent the source locations, i.e., IDs 1-72 for the azimuth in Region A, IDs 73-144, 145-216, and 217-288 for the azimuth in Regions B, C, and D in Fig.4. The correct location labels were added on the basis of voice activity of clean speech signals block-by-block (every 110 [ms]). Configurations are listed in Table 1.

[1]http://www.sifi.co.jp/system/modules/pico/index.php?content_id=39
[2]http://research.nii.ac.jp/src/JNAS.html

Table 2. Localization correctness in total (%)

Condition		Baselines			Proposed
Resolution	SNR (dB)	MUSIC	DNNs w/o adapt.	DNNs ($\lambda = 0$)	DNNs ($\lambda = 10$)
1D azimuth (73 labels)	0	50.7	34.8	29.3	**64.0**
	10	77.7	60.0	49.4	**83.1**
	20	88.1	82.2	83.6	**89.4**
3D regions (289 labels)	0	40.2	30.8	29.1	**44.2**
	10	50.1	41.7	40.8	**56.0**
	20	57.4	54.1	58.6	**61.6**

Configuration of DNNs (Fig. 2): The size of directional activators was 4×128 in each w-th sub-band in the DNNs. There were sixteen blocks of linear projection (LP2) to combine intermediate outputs. The network sizes [dim. of input, dim. of output] of the LP1, ..., and LP5 corresponded to 386×64, 512×64, 1024×1024, 1024×1024, and 1024×289, respectively. There were a total of 1536 dimensions of features for the DNN input, and 289 output dimensions to classify all labels ($K = 289$). All weight parameters were initialized by using a Gaussian distribution $N(0, 0.025)$. The cross-entropy was used as the cost function for training, and we stopped training after three epochs because there was no improvement in block-level correctness for the training set. The unsupervised adaptation was applied to each utterance (5 sec. on average), and we stopped the adaptation after 490 iterations. The G for weight calculation was 5, and this allowed a ± 12.5-degree error in azimuth. We checked the performance of several learning rates $\alpha = [0.5, 0.1, 0.05]$ and constraint weights $\lambda = [0, 5, 10, 20]$. The influence of these parameters will be also explained in results.

Configuration of MUSIC: There were 1,800 reference steering vectors calculated from anechoic impulse responses (72 azimuth $\times$ 5 depth $\times$ 5 height, i.e. $V = 1800$), as shown in Fig. 4. The score for the 3D regions is calculated by the summing score $q(\boldsymbol{r})$ over the positions in each region. The threshold T_{th} for judging the presence of sound was set to maximize the localization correctness of each SNR of the test set to show the performance limitations.

Evaluation criteria: We calculated the correctness of the test set classification at the block level. We allowed a ± 7.5-degree error in azimuth. For example, when the ground truth is 0°, the estimated locations at 355° and 5° are also considered to be correct. There was a total of 785 blocks for the test data per position, and the ratio of "no-sound" blocks for the test speech signals was 28.8%.

4.2. Results and Discussions

Table 2 summarizes the total 1D-azimuth and 3D-region localization correctness of MUSIC and each DNN in the test set. The *w/o adapt.* entries denote the results of DNNs without adaptation. The 20 dB SNR was the same condition as that in the training phase. Our proposed DNNs resulted in having the best performance in all cases. Our method outperformed MUSIC by 5.9 points at 10 dB regarding the 3D regions and 13.3 points at 0 dB SNR regarding the 1D azimuth. The larger $\lambda > 4$ seemed to improve performance, but $\lambda = 10$ and 20 had almost the same results. The azimuth correctness of our method is better than that of MUSIC, which gives top-G ambiguous candidates to DNNs. This fact indicates that the error patterns between the DNNs and MUSIC were different. The performance of the adapted DNNs without any constraint $\lambda = 0$ degraded as the SNR worsened. Note that *the 3D-region correctness of our method was worse than MUSIC in some positions* because of the small number of depth and height patterns compared with the number of azimuth patterns in the training data.

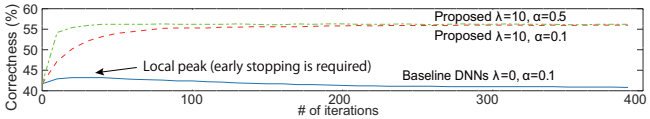

Fig. 6. Correctness vs. # of iterations for convergence comparison

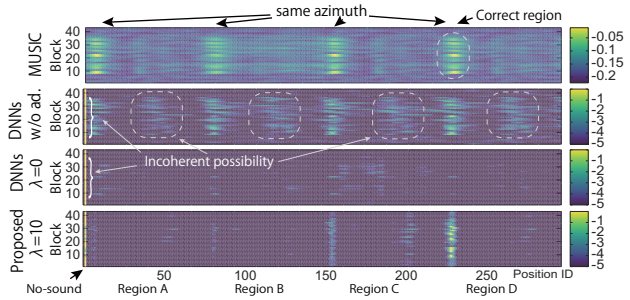

Fig. 7. MUSIC score and posterior probabilities of each DNN on $\log_{10}$ scale for 3D-region estimation

Figure 6 shows the relationship between the correctness for the 3D regions and the number of iterations during adaptation in the case of the 10 dB test set. The performance of the baseline DNNs with $\lambda = 0$ degraded after 30 iterations. This can be avoided by applying early stopping heuristically. On the other hand, our method stably converged to one of having high correctness, and a larger learning rate α accelerated the convergence speed. Therefore, hundreds of iterations are not required for adaptation.

Figure 7 shows an example of the 3D-region score of MUSIC and each DNN on a log scale. The vertical axis represents the number of blocks (corresponding to time) and the horizontal axis represents location IDs, shown in Fig. 4. The MUSIC score is a blur but does not contain incoherent peaks. The unadapted and adapted DNNs with $\lambda = 0$ include some incoherent peaks. Our constraint reduced these errors, and the DNNs could estimate clearer posterior probabilities compared with that of MUSIC.

4.3. Remaining Issues

The essential issues are performance dependency on locations, adaptation with shorter duration than one utterance, and an efficient and automatic real-world data-augmentation method. The first issue will be solved by the data-augmentation of position patterns based on generative model of arriving sound signals. The second issue must be solved regarding real-time processing and moving sources. The last issue is required to improve the potential performance of DNNs. Since autonomous systems, such as robots, can collect data automatically, such real-world data augmentation will be more important in the future. An alternative adaptation of model-based and DNNs-based SSL has the potential for further improvement.

5. CONCLUSION

We tackled the unsupervised adaptation of DNNs-based SSL regarding unknown conditions. We proposed an eliminative constraint of possibility based on a model-based SSL to suppress incoherent errors during adaptation. Experiments revealed that our adapted DNNs improved the localization correctness of 1D azimuth and 3D regions under unknown conditions compared to that of model-based SSL.

Acknowledgement This work was partly supported by JSPS KAKENHI Grant Numbers JP16H02869 and the Cooperative Research Project Program of the RIEC, Tohoku University.

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] 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.

[3] K.-L. Du, A.K.Y. Lai, K.K.M. Cheng, and M.N.S. Swamy, "Neural methods for antenna array signal processing: a review," *Signal Processing*, vol. 82, no. 4, pp. 547–561, 2002.

[4] 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.

[5] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 2814–2818.

[6] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 405–409.

[7] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *Proc. of Machine Learning for Signal Processing*, 2016, pp. 1–6.

[8] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *Proc. of IEEE Workshop on Spoken Language Technology*, 2016, pp. 603–609.

[9] R. Takeda and K. Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2217–2221.

[10] C.M. Bishop, *Pattern Recognition and Machine Learning*, Shpringer, 2006.

[11] 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.

[12] K. Nakadai N. Yalta and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[13] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7893–7897.

[14] T. Yoshioka, A. Ragni, and M. JF Gales, "Investigation of unsupervised adaptation of DNN acoustic models with filter bank input," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 6344–6348.

[15] S. P Rath, D. Povey, and K. Veselỳ, "Improved feature processing for deep neural networks.," in *Proc. of Interspeech*, 2013, pp. 109–113.

[16] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist speaker normalization and adaptation," in *Proc. of Eurospeech*, 1995, pp. 2183–2186.

[17] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. D. Mori, "Linear hidden transformations for adaptation of hybrid ann/hmm models," *Speech Communication*, vol. 49, no. 10, pp. 827–835, 2007.

[18] Z. Huang, S. M. Siniscalchito, I-F. Chen, W. Jiadong, and C.-H. Lee, "Maximum a posteriori adaptation of network parameters in deep models," in *Proc. of Interspeech*, 2015, pp. 1076–1080.

[19] D. Torrieri and K. Bakhru, "Simplification of the MUSIC algorithm using a neural network," in *Proc. of MILCOM*, 1996, vol. 3, pp. 873–876.