LOCATA CHALLENGE: A DEEP NEURAL NETWORKS-BASED REGRESSION APPROACH FOR DIRECTION-OF-ARRIVAL ESTIMATION

Junhyeong Pak and Jong Won Shin

School of Electrical Engineering and Computer Science,
Gwangju Institute of Science and Technology, Gwangju, Republic of Korea

ABSTRACT

Source localization algorithm plays a crucial role for a wide range of applications in audio signal processing that exploit multiple microphones. The accuracy of a localization algorithm usually depends on several types of spatial information, such as interchannel phase differences (IPDs) of a sound signal captured by multichannel microphone, but this information can be easily degraded by background noise and reverberations. Until now, no regression model based on deep neural networks (DNNs) has been reported that deals with the phase difference in the time-frequency (TF) domain, as the phase difference has no specific pattern that can be utilized in deep learning. In this paper, we introduce the *phase difference with an artificial structure* (PDAS). The pattern in the PDAS is generated artificially, but it can be helpful for a DNN regression model. By exploiting the PDAS, the proposed DNN-based localization approach reduces distortion caused by interference in a real-world environment, and achieves a significant improvement in localization accuracy. To determine the frame-by-frame directional angle of a sound source, we find a peak value of each distribution of direction-of-arrival (DoA) obtained from whole DoA estimates across specific TF components. Our experimental results show that the proposed method outperforms a baseline algorithm with respect to determining the source direction in on-line processing.

Index Terms— source localization, direction-of-arrival estimation, phase difference with artificial structure, deep neural networks

1. INTRODUCTION

Source localization algorithm is important for various types of speech processing, such as speech enhancement [1, 2, 3], recognition [4, 5], and so on. Since source localization algorithms usually consider spatial information, such as the interchannel phase difference (IPD) or interchannel level difference (ILD), their performance depends on spatial cues. However, the performance of localization algorithm can be easily degraded because the information is usually corrupted by background noises or reverberations.

Thus far, many studies on localization algorithms have focused on their abilities to deal with various spatial cues to determine the direction of sound sources [6]-[22]. In this context, the interchannel time difference [6] or IPD [7, 8] represents one of the most important types of spatial information for subband-based localization. Some approaches have been used to select reliable components of information in the time-frequency (TF) domain to determine the directions of sound sources [9, 10]. Furthermore, the selected information can be dealt with by several post-processing techniques, such as clustering and statistical fitting, in order to improve performance. Although the ILD can be employed as an information source in localization, its reliability tends to be seriously degraded according to the distance between the sound source and the microphone. In this regard, a previous study [11] presented a statistical approach to localization utilizing both the IPD and ILD as a complementary strategy. To overcome performance degradation caused by reverberations, direction-of-arrival (DoA) estimators based on onset detection [9, 12], in addition to zero-crossing rate analysis, have been utilized [12, 13]. Various techniques to select reliable spatial information in TF slots have also been described, for example, a sinusoidal approach [8], coherence tests [9, 14] estimation consistency [10], and signal-to-noise ratio (SNR) analysis [12, 13]. To determine the directions of sound sources, statistical approaches, including mixture models or k-means clustering, have been employed [9, 10, 14, 15, 16].

Over the last decade, impressive results related to speech processing have been derived using deep neural network (DNN). In [17, 18], a DNNs-based localization approach exploiting subspace features of well-established baselines, such as MUSIC [19] and ESPRIT [20], has reported according to advantages of deep learning. This approach utilizes eigenvectors and steering vectors as input data and the weights of the first hidden layer of a network. In [21, 22], the authors proposed a DNN classification method to determine the DoAs of sources. This tries to mimic how the human auditory system localizes sound sources and shows further improved performance when taking prior information about interference location. However, the performance of these conventional localization algorithms is hampered by interferences, as they usually use distorted spatial information without any effort related to noise reduction or dereverberation.

In this paper, we proposed a DNN regression model for source localization by considering phase difference. For deep learning, the input and target feature should contain specific patterns, but phase difference has no distinctive pattern in itself. Thus, phase difference may be inappropriate for DNNs-based localization algorithm. To overcome this, we investigate a novel approach of DNN-based localization utilizing the *phase difference with an artificial structure* (PDAS) instead of phase difference. We apply the PDAS in a DNN regression model to estimate all DoAs in whole TF slots, and find the peak value for the DoA distribution within specific temporal periods and frequency bands to determine the direction of the sound source in on-line processing. It is noted that the proposed algorithm can reduce distortions in spatial cues caused by background noises and reverberations while most baseline localization approaches only utilizes corrupted cues. The experimental results

This work was supported by GIST Research Institute (GRI) grant funded by the GIST 2018 and the Technology Innovation Program (10076583, Development of free-running speech recognition technologies for embedded robot system) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

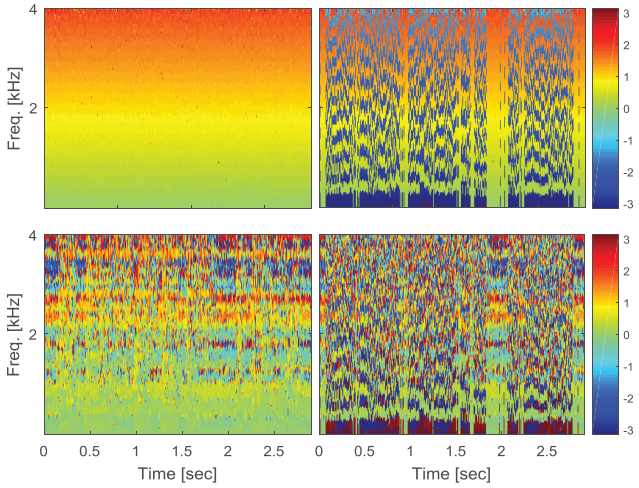

Figure 1: TF plots for (left) phase difference and (right) PDAS in the case of a sound source located in the azimuth direction of $\theta = 40^{\circ}$: (Top) "clean and dry" observation and (bottom) "noisy and reverberant" observation with *Babble* noise at 5 dB SNR and $RT_{60} = 0.2$ s.

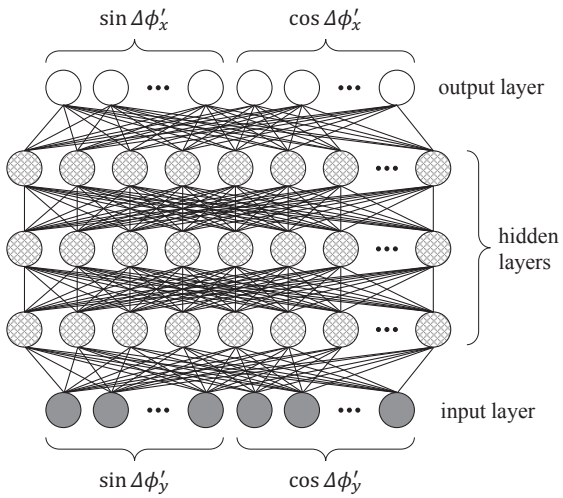

Figure 2: A neural network for the PDAS estimation. $\Delta\phi'_x$ and $\Delta\phi'_y$ represent the PDAS of the clean and interference-corrupted speech, respectively.

demonstrate the performance of the proposed method in terms of the root mean square error (RMSE) and standard deviation about the DoA estimates compared with the MUSIC algorithm [19] for recorded speech files under real-world condtions.

2. DNN-BASED DOA ESTIMATION UTILIZING PDAS

In this section, we describe how to generate the PDAS, which has helpful patterns for deep learning, and propose a novel DNN-based localization approach that exploits the PDAS. Assuming a far-field model, we consider a stereophonic speech signal captured by a dual microphone. For speech signals of stereo channels at temporal index n, $x_1(n)$ and $x_2(n)$, we describe the discrete Fourier transform (DFT) coefficients of the signals as follows:

$$X_1(k,m) = \sum_{n=0}^{N-1} x_1(ml+n)w(n)e^{-j\frac{2\pi}{N}kn}, \tag{1}$$

$$X_2(k,m) = \sum_{n=0}^{N-1} x_2(ml+n)w(n)e^{-j\frac{2\pi}{N}kn}, \tag{2}$$

where m, l and N denote the temporal index, frame shift, and DFT dimension, respectively. The phase difference $\Delta\phi_x(k,m)$ for k-th frequency bin in the m-th frame between the two DFT coefficients, $X_1(k,m)$ and $X_2(k,m)$, can be derived as below:

$$\Delta\phi_x(k,m) = \angle\big(X_2(k,m)\cdot X_1^*(k,m)\big), \tag{3}$$

where $\angle(\cdot)$ and $*$ represent the phase operator of complex coefficients and complex conjugate, respectively. The directional angle $\theta_x(k,m)$ is obtained from phase difference $\Delta\phi_x(k,m)$ by the geometric relationship between them.

$$\theta_x(k,m) = \sin^{-1}\left(\frac{c\cdot\Delta\phi_x(k,m)}{2\pi f d}\right), \tag{4}$$

where c, f, and d are the speed of sound, frequency according to k-th bin index, and distance between two microphones, respectively.

In our DNN-based localization algorithm, which can reduce distortion caused by background noises and reverberations, we utilize the geometric relationship between the phase difference and DoA described in (4). In general, the phase difference across whole TF bins has no specific pattern in itself even though it is obtained from a noiseless and anechoic conditions. Additionally, the phase difference is affected by interferences including background noises and reverberations in Figure 1. As a result, the phase difference of speech signals is unsuitable for deep learning. To overcome this obstacle, we utilize the PDAS instead of the phase difference in our regression model. Despite the artficial pattern of the PDAS, it can be useful in a DNN aimed at reducing interference. We define the PDAS $\Delta\phi'(k,m)$ when we have the pitch information in voiced region, as below:

$$\Delta\phi'(k,m) = \Delta\phi(k,m) + \frac{\pi}{\alpha}i, \tag{5}$$

where i represents the closest harmonic index to k and α is a tunable parameter, and we select α as unity, which can represent the simplest pattern of the PDAS for deep learning. The PEFAC algorithm in [23] is utilized for pitch estimation, and we expand its applications to both voiced and unvoiced regions. It is noted that the accurate estimation of the fundamental frequency by the PEFAC algorithm is not critical, as it has only a limited role in building the artificial pattern in the PDAS. Thus we can expand application of the PEFAC to generate artificial pattern for both voiced and unvoiced regions. Figure 1 shows the distinction between the PDAS and phase difference in cases of speech under "clean and dry" conditions and "noisy and reverberant" conditions.

In addition to using the PDAS to predict the direction of the sound source, there is one more consideration for deep learning. The mean square error (MSE) function is a well-known cost function in DNN regression-based applications, but it is practical only when a dataset contains continous values. To overcome this issue,

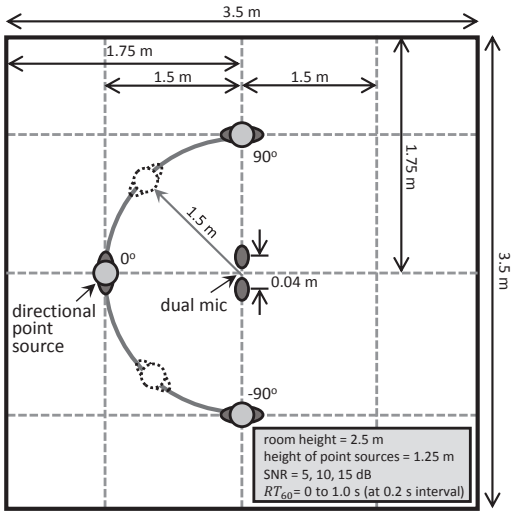

Figure 3: Simulated room configurations to generate a training dataset for the proposed DNN regression model.

we utilize the two sinusoidal versions (i.e., sine and cosine) of the PDAS. These are applicable in a DNN-based localization approach using the MSE as a cost funtion since their values of them are continuous in range of $[-1, 1]$.

We consider the sinusoidal versions of the PDAS as the input and output data of a DNN, as shown in Figure 2. Thus, the dimension of the converted PDAS is twice of that of the PDAS (i.e., $(K/2-1)\times 2$). The network has three hidden layers, and each layer has 2,048 nodes activated by the rectified linear unit. The linear unit is utilized as the activation function of the output layer.

Finally, the sinusoidal versions of the PDAS estimated by the DNN need to be changed into the PDAS estimate. By utilizing the `MATLAB` built-in function, $\arctan 2(\cdot)$, we can obtain the PDAS estimate $\Delta\hat{\phi}'_x(k,m)$ as an enhanced spatial cue by the proposed method. Similar to (5), the phase difference estimate $\Delta\hat{\phi}_x(k,m)$ can be finally obtained as below:

$$\Delta\hat{\phi}_x(k,m) = \Delta\hat{\phi}'_x(k,m) - \frac{\pi}{\alpha}i. \tag{6}$$

The DoA estimates for all the TF components can be obtained by plugging $\Delta\hat{\phi}_x(k,m)$ into (4). To determine the frame-by-frame directions of a sound source, the peak of distribution of all the DoA values within a specific time interval is calculated. It is noted that the proposed algorithm only utilizes the estimated DoAs except future values for on-line processing.

3. EXPERIMENTAL RESULTS

To verify the accuracy performance of the proposed algorithm in cases of recorded audio files in real-world environments, we utilize the LOCATA challenge dataset [24, 25]. We consider a wide range of training data related to background noise, reverberations, and one or two sound sources to ensure generalization of the DNN-based localization method. As shown in Figure 3, a small room (dimensions of $3.5\times 3.5\times 2.5$ m) simulated by the image method [26] is utilized for generation of the training data samples. Two microphones are located nearby the center of the room at (1.73 m, 1.75 m, 1.25 m) and (1.77 m, 1.75 m, 1.25 m), whereas a static speaker is placed 1.5 m from the center of the microphone array from -90° to 90° at 10° intervals.

Table 1: Comparison of the performance of the algorithms for the LOCATA development dataset *Task 1* measured by the *DICIT* array. Average azimuth estimate (Avg), standard deviation (SD), and average RMSE value in degrees obtained by the MUSIC (MU) [19] and the proposed algorithm (PR). O and X represent the case with and without, respectively, the inclusion of the perfect knowledge about speech presence during the performance assessment.

recording		1		2		3	
ground truth		58.0		-60.0		-21.3	
assessment with VAD		X	O	X	O	X	O
MU	Avg	39.6	48.6	-35.3	-42.7	**-22.0**	-27.3
	SD	24.4	20.3	29.1	21.6	21.3	17.5
	RMSE	24.2	17.8	26.0	21.1	16.3	14.0
PR	Avg	**60.2**	**60.3**	**-59.3**	**-59.3**	-22.8	**-22.8**
	SD	**2.6**	**3.0**	**0.8**	**0.9**	**1.6**	**1.8**
	RMSE	**2.9**	**3.2**	**0.9**	**0.9**	**1.9**	**2.0**

A number of speech files of the TIMIT database were mixed with three types of noise (*Babble, Factory*, and *Volvo*) obtained from the NOISEX-92 database to simulate noisy environment. The noises were generated as stereophonic signals as the speech wave files but as diffuse signals according to the standard of [27]. Three SNRs were used: 5, 10, and 15 dB. Both the speech and noise signals were sampled at 8 kHz, and a 16 ms length window was applied with the frame shift 8 ms. In addition, we considered several reverberations from 0 to 1.0 s at intervals of 0.2 s. Please note that we applied large variations of environmental conditions related to the number of sources, background noises, and reverberations in order to improve the generalization performance of the proposed method.

We applied the proposed method to both the development dataset and the evaluation dataset (i.e., the LOCATA challenge dataset [24, 25]), especially *Task 1* for a static source localization using static microphones array. The development dataset includes the ground truth angle and the perfect knowledge about speech presence for each recording, whereas the evaluation dataset does not contain this information. Note that prior information about speech presence regions is not exploited in localization. It is utilized only in analyzing how accurately the proposed method predicts the directional angle in the case of the development dataset. Since we utilize a dual microphone array with a distance of 4 cm for generating the training set, we can also extract stereophonic signals from the *DICIT* array-recorded speech wave files (i.e., two stereophonic wave files per each recording captured by the 6th, 7th and 7th, 9th channel of the *DICIT* array with a same distance). Figure 4 shows the localization results of the MUSIC algorithm [19] and those of the proposed method using real-time processing for the development dataset related to *Task 1*. As shown in Table 1, there is a significant difference between the estimated results at each temporal period in terms of the average DoA value, standard deviation, and average RMSE of the estimated azimuth in degrees. The average RMSE of the proposed method is much less than that of the MUSIC algorithm for all three recordings in *Task 1*. Furthermore, using the proposed method, the standard deviation of the DoA estimates is much less than that of the MUSIC algorithm. With regard to the prediction of the directional angle, this finding implies that the

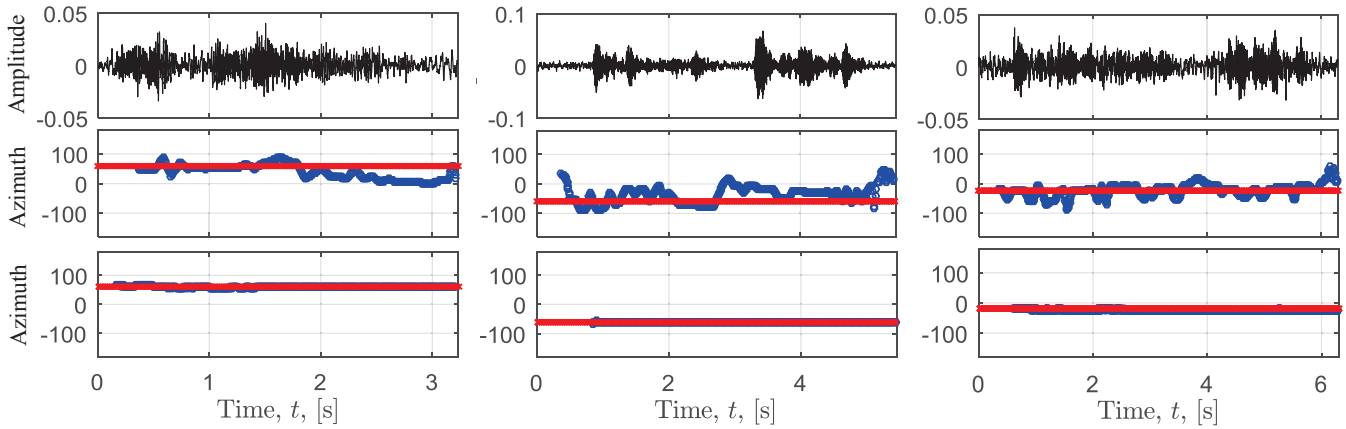

Figure 4: Plots of the estimated azimuth angles (blue marks) and ground truth angles (red marks) according to the given time index for three recordings of the LOCATA development dataset for *Task 1* measured by the *DICIT* array. From top to bottom: sound waveform, azimuth results of the MUSIC algorithm [19], and azimuth results of the proposed algorithm.

Table 2: Comparison of the performance of the algorithms for the LOCATA evaluation dataset for *Task 1* measured by the *DICIT* array. Average azimuth estimate and standard deviation in degrees obtained by the MUSIC [19] and proposed algorithm, without any information about speech presence region.

recording		1	2	3	4	5	6	7	8	9	10	11	12	13
MU	Avg	20.73	30.95	-39.72	27.20	-38.75	-6.51	35.43	-31.89	29.29	-31.68	31.69	-31.85	27.32
	SD	21.76	25.83	21.61	15.83	29.25	**22.93**	25.65	26.91	18.12	23.68	29.73	30.49	22.13
PR	Avg	23.23	52.24	-56.66	32.05	-38.48	-7.56	47.23	-53.36	22.98	-23.23	56.48	-60.52	37.43
	SD	**7.09**	**7.81**	**2.70**	**2.63**	**1.81**	23.45	**2.17**	**3.39**	**2.60**	**1.15**	**4.14**	**2.85**	**11.54**

proposed method is more accurate in real-time as compared with the baseline approach when a loudspeaker is located in a fixed direction as a static sound source. The results of the evaluation dataset for LOCATA *Task 1* are shown in Table 2. This dataset has no ground truth information about the direction of the static source in each recording. Thus, we present only the results of the average azimuth estimate and standard deviation for 13 recordings measured by the *DICIT* array. Similar to the results of the development dataset, the standard deviations of the estimated azimuth angles using the proposed method are much less than those of the MUSIC algorithm for almost all the recording cases in *Task 1*. Thus, we can conclude that the proposed method can estimate the directional angles of a sound source accurately in real-time processing as compared with the MUSIC algorithm, which is one of the most well-known localization algorithms. It is noted that neither the proposed method nor the baseline method considers the information about speech presence at any of the temporal moments. Therefore, the accuracy of the results generated by both algorithms can be enhanced by considering voice activity detection in each recording.

4. CONCLUSION

We investigated a novel DNN-based regression approach to predict the direction of a sound source. Focusing on the idea that a specific pattern of input and output features can be helpful for deep learning, we present the PDAS that can enable a DNN regression model to operate properly and predict the direction of sound sources accurately. The experimental results for the recorded data in real-world conditions show that the proposed method outperforms the baseline approach with respect to the accuracy of DoA estimation. The proposed method shows much smaller standard deviations for the estimated azimuth angles in real-time than those obtained by the MUSIC approach in both the development and evaluation datasets related to a static source localization task. Furthermore, for the development dataset, the RMSEs obtained using the proposed method are much less than those obtained using the baseline algorithm.

5. REFERENCES

[1] R. Talmon, I. Cohen, and S. Gannot, "Multichannel speech enhancement using convolutive transfer function approximation in reverberant environments," in *Proc. IEEE ICASSP*, 2009, pp. 3885-3888.

[2] Y. G. Jin, J. W. Shin, and N. S. Kim, "Spectro-temporal filtering for multichannel speech enhancement in short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 21, issue 3, pp. 352-355, Mar. 2014.

[3] Y. G. Jin, J. W. Shin, and N. S. Kim, "Decision-directed speech power spectral density matrix estimation for multichannel speech enhancement," *JASA Express Lett.*, vol. 141, issue 3, pp. EL228-EL233, Feb. 2017.

[4] A. Stolcke, "Making the most from multiple microphones in meeting recognition," in *Proc. IEEE ICASSP*, 2011, pp. 4992-4995.

[5] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T.

Nakatani, A. Nakamura, and J. Yamoto, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 499-513, Feb. 2012.

[6] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320-327, 1976.

[7] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Syst. Man Cybern. B*, vol. 34, no. 4, pp. 1736-1773, 2004.

[8] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1913-1928, 2010.

[9] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust DoA estimation of multiple speech sources," in *Proc. IEEE ICASSP*, 2014, pp. 2287-2291.

[10] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization using estimation consistency in the time-frequency domain," in *Proc. IEEE ICASSP*, 2017, pp. 516-520.

[11] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833-1847, Aug. 2007.

[12] S. -Y. Lee and H. -M. Park, "Multiple reverberant sound localization based on rigorous zero-crossing-based ITD selection," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 671-674, 2010.

[13] Y. -I. Kim and R. M. Kil, "Estimation of interaural time differences based on zero-crossings in noisy multisource environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 734-743, 2007.

[14] M. Cobos, J. J. Lopez, and D. Martinez, "Two-microphone multi-speaker localization based on a Laplacian mixture model," *Digit. Signal Process.*, vol. 21, no. 1, pp. 66-76, Jan. 2011.

[15] S. Araki, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. IEEE ICASSP*, 2009, pp. 41-44.

[16] C. Kim and K. K. Chin, "Sound source separation algorithm using phase difference and angle distribution modeling near the target," in *Proc. Interspeech*, 2015, pp. 751-755.

[17] R. Takeda and K. Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," in *Proc. IEEE ICASSP*, 2017, pp. 2217-2221.

[18] R. Takeda, Y. Kudo, K. Takashima, Y. Kitamura, and K. Komatani, "Unsupervised adaptation of neural networks for discriminative sound source localization with eliminative constraint," in *Proc. IEEE ICASSP*, 2018, pp. 3514-3518.

[19] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagation*, vol. AP-34, no. 3, pp. 276-280, 1986.

[20] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984-995, 1989.

[21] N. Ma and G. J. Brown, "Speech localisation in a multitalker mixture by humans and machines," in *Proc. Interspeech*, 2016, pp. 3359-3363.

[22] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 12, pp. 2444-2453, 2017.

[23] S. Gonzalez and M. Brookes, "PEFAC - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 518-530, Feb. 2014.

[24] IEEE-AASP Challenge on Acoustic Source LOCalization And TrAcking (LOCATA), 2018 [Online]. Available: `http://www.locata-challege.org`.

[25] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2018.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943-950, 1979.

[27] ETSI Standard EG 202 396-1 v1.2.2, "Speech processing, transmission and quality aspects (STQ): Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database," 2008.