INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Robust DOA Estimation Based on Convolutional Neural Network and Time-Frequency Masking

Wangyou Zhang, Ying Zhou, Yanmin Qian

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
wyz-97@sjtu.edu.cn, zhouy49@sjtu.edu.cn, yanminqian@sjtu.edu.cn

Abstract

In the scenario with noise and reverberation, the performance of current methods for direction of arrival (DOA) estimation usually degrades significantly. Inspired by the success of time-frequency masking in speech enhancement and speech separation, this paper proposes new methods to better utilize time-frequency masking in convolution neural network to improve the robustness of localization. First a mask estimation network is developed to assist DOA estimation by either appending or multiplying the estimated masks to the original input feature. Then we further propose a multi-task learning architecture to optimize the mask and DOA estimation networks jointly, and two modes are designed and compared. Experiments show that all the proposed methods have better robustness and generalization in noisy and reverberant conditions compared to the conventional methods, and the multi-task methods have the best performance among all approaches.

Index Terms: source localization, direction-of-arrival estimation, convolutional neural networks, time-frequency masking, multi-task learning

1. Introduction

Sound source localization is a task that estimates the direction of arrival (DOA) of a speaker from received speech signal. The DOA estimation is essential of various applications, such as human-robot interaction and teleconferencing, and is also widely used in beamforming for speech enhancement [1, 2, 3, 4, 5]. Over the years, several conventional signal processing techniques have been developed for broadband DOA estimation, such as generalized cross correlation with phase transform (GCC-PHAT) [6], steered response power using the phase transform (SRP-PHAT) [7] and multiple signal classication (MUSIC) [8]. However, these methods rely on some ideal assumptions, for example the noise is white, the signal-to-noise ratio (SNR) is greater than 0dB [9] and each time-frequency bin is dominated by the direct component of a single sound source [10]. Thus their performance is unsatisfactory in the presence of noise and reverberation.

Recently, deep neural network (DNN) based methods [10, 11, 12, 13, 14, 15] have been proposed to improve the robustness of DOA estimation. These DNN approaches learn a mapping between signal features and a discretized DOA space directly, without those assumptions about the environment. Different features such as GCC vectors [11], the eigenvectors of the spatial correlation matrix which correspond to the noise subspace [12], the cosine-sine interchannel phase difference (CSIPD) features [10] as well as original phase spectrum [14] have been used as inputs of DNNs. Most of these features are derived in time-frequency (T-F) domain and they are still vulnerable to distortion since not all T-F bins are dominated by target speech.

This work aims to improve the robustness by eliminating most noise-dominated T-F bins in the feature to minimize the effects of noises and reverberation. In recent years, ideal ratio mask (IRM) has been proven effective in speech separation and speech enhancement [16, 17, 18], and there are also works involving IRM with conventional DOA estimation [19]. Inspired by the success of T-F masking in these works, we propose three kinds of convolution neural network (CNN) based methods with the idea of T-F masking for broadband DOA estimation. We note that there is recent work also using similar T-F masking based CNN (called *identifier*) for keyword-based speaker localization [10]. Our work is performed independently and we make more comprehensive development and analysis. Moreover we extend and propose new methods when integrating T-F masking for DOA.

The rest of this paper is organized as follows. Section 2 describes the DOA estimation problem. Section 3 introduces our proposed methods for broadband DOA estimation with CNN and T-F masking. The experimental results are discussed in Section 4, and conclusions are summarized in Section 5.

2. Problem Description

Suppose that the geometry of array is known and there is a single target source, the signal received in noisy and reverberant environments can be modelled in T-F domain as

$$\mathbf{Y}(t,f) = \mathbf{r}(f)S(t,f) + \mathbf{H}(t,f) + \mathbf{N}(t,f) \qquad (1)$$

where $\mathbf{Y}(t,f)$ denotes the received signal, $\mathbf{r}(f)S(t,f)$, $\mathbf{H}(t,f)$, and $\mathbf{N}(t,f)$ represent its direct, reverberation and noise components respectively. $S(t,f)$ is the signal received from the reference microphone, and $\mathbf{r}(f)$ is the relative transfer function which can be formulated as

$$\mathbf{r}(f) = [A_1(f)e^{-j2\pi f\tau_1}, ..., A_M(f)e^{-j2\pi f\tau_M}]^T \qquad (2)$$

where τ_i is the time difference of arrival (TDOA) between the two signals received from the ith and the reference microphones, $A_i(f)$ denotes the relative gain of the ith microphone, and M denotes the number of microphones. The true DOA information is contained in direct signal, and has a relationship with the TDOA of each microphone pair, which is reflected in phase part of the direct signal in T-F domain. Thus the phase information is the essence in the DOA estimation task.

The conventional SRP-PHAT algorithm uses the time delay between each microphone pair to build the target function which

*Yanmin Qian is the corresponding author.

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-3158

can be formulated as

$$F_{\text{SRP-PHAT}}(\boldsymbol{\tau}) = 2\pi \sum_{l=1}^{I} \sum_{k=1}^{I} \int_{-\infty}^{+\infty} \frac{\Gamma_{lk}(\omega)}{|\Gamma_{lk}(\omega)|} e^{j\omega(\tau_l - \tau_k)} d\omega \tag{3}$$

where $\Gamma_{lk}(\omega) = \mathcal{F}\left(E\left[y_l(t) y_k^*(t + \tau_l - \tau_k)\right]\right)$ is the cross-power density spectrum for two microphone signals $y_l(t)$ and $y_k(t)$, $\mathcal{F}(\cdot)$ denotes the Fourier transform, and $E[\cdot]$ is the expectation operator. The directions of the signal sources then correspond to the peak of the target function.

In the convolutional neural network (CNN) based framework, DOA estimation is generally formulated as an I-class classification problem, where I denotes the number of classes. Phase-related features are fed into the CNN and a mapping from input features to the corresponding DOA label is learned.

3. CNN Based DOA Estimation with T-F Masking

In this section, we first describe the baseline method using a simple CNN architecture in [14] for DOA, and then propose three new methods incorporating T-F masking.

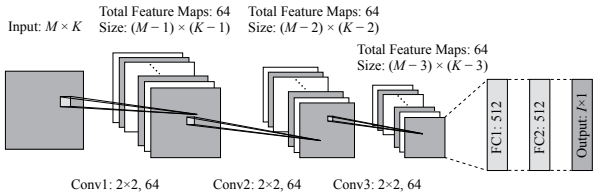


Figure 1: *Baseline CNN architecture for DOA.*

3.1. CNN for DOA

The architecture proposed in [14] is a CNN with 3 convolutional layers and 3 fully-connected layers. The input vector is the phase component of the STFT coefficient of the received signal at each microphone instead of explicitly extracted features. The output is an $I \times 1$ vector indicating the posterior probabilities of the I DOA classes. Each of the I classes corresponds to a discretized DOA value. The cross entropy loss function $\mathcal{L}_{\text{CE}}$ is used for training. In the inference phase, given a test microphone array signal, the posterior probability for each DOA class can be generated by the trained DOA estimator.

In this paper, we slightly change the size of several layers in the aforementioned architecture to build the baseline system, since the microphone array we use is a 6-mic circular array rather than the 4-mic uniform linear array (ULA) in [14]. The 6-mic circular array is chosen because it can receive more information from the sound source and resolve DOAs from 0° to 360°, while the 4-mic ULA can only detect DOAs from 0° to 180° due to its symmetrical directivity [20]. Thus, the number of class I is 72 and the discretized DOA space corresponds to a set $\Theta = \{0^\circ, 5^\circ, ..., 355^\circ\}$ in the baseline architecture. The final architecture of the baseline system is shown in Fig.1, where M denotes the number of microphones and K denotes the total number of frequency bins.

Although the basic CNN architecture has strong representation capability, its performance may still degrade significantly in noisy and highly reverberant environments. Moreover it usually needs data pre-processing such as voice activity detection (VAD) to eliminate non-speech frames, which may not be accurate and cannot eliminate the effects of noise in different frequency bins. So we propose three methods to improve the performance of the CNN-based method and they are described in the following sections.

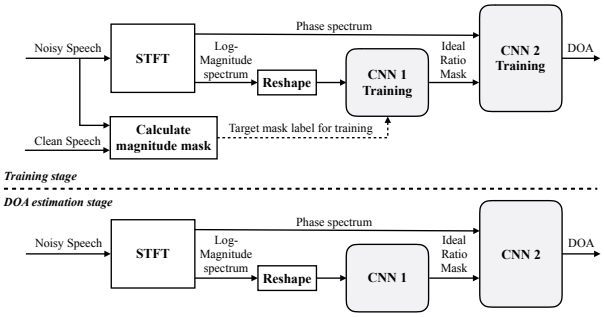


Figure 2: *Overview of the proposed mask-CNN system for DOA.*

3.2. CNN with Ideal Ratio Mask for DOA

To utilize the T-F masking, an intuitive idea is to train the mask estimation network in advance and then use the estimated mask to enhance the input features for the DOA network training. Mask model and DOA model are built separately: First we train the mask estimation network to derive a magnitude-related mask which represents the probability of each T-F bin being dominated by the target speech signal. Then we enhance the input features with the estimated mask and train the DOA estimation network with these new features. The overview of this proposed mask-CNN system is illustrated in Fig.2[1].

To enhance the input feature, we can simply append the mask to the 6-channel input as an additional feature. In addition, we have also tried multiplying the input by the mask to minimize the effects of noise-dominated T-F bins, and thus the mask is regarded as the weight of each T-F bin in input features. The performance of both approaches are evaluated in Section 4. In our experiments, the performance of the latter mode is slightly better than the former one.

The mask estimation network is also a CNN with the architecture proposed in [21], which is a regression model that maps noisy log-magnitude feature to the corresponding clean mask. The input vector consists of 11 consecutive frames (5 preceding and 5 following the current frame) of the log-magnitude spectrum of the received signal at each microphone, and the output is the estimated soft mask of the current frame. To compute the target mask label for each frame, we consider the ratio of parallel clean speech signal power spectrum and noisy signal power spectrum, which can be formulated as

$$\text{IRM} = \frac{S^2(t,f)}{S^2(t,f) + N^2(t,f)} \tag{4}$$

where $S(t,f)$ and $N(t,f)$ denote the magnitude spectrum of clean speech signal and noise signal at the t-th time frame and the f-th frequency bin respectively. The mean squared error (MSE) loss function $\mathcal{L}_{\text{MSE}}$ is used for training the mask estimation network.

3.3. Multi-task learning for DOA estimation

3.3.1. Standard multi-task learning

Since the mask and DOA estimation networks are trained separately in Section 3.2, the estimated mask may not completely

[1]The Reshape block represents frame expansion, which reconstructs the original log-magnitude spectrum to several 11-frame features.

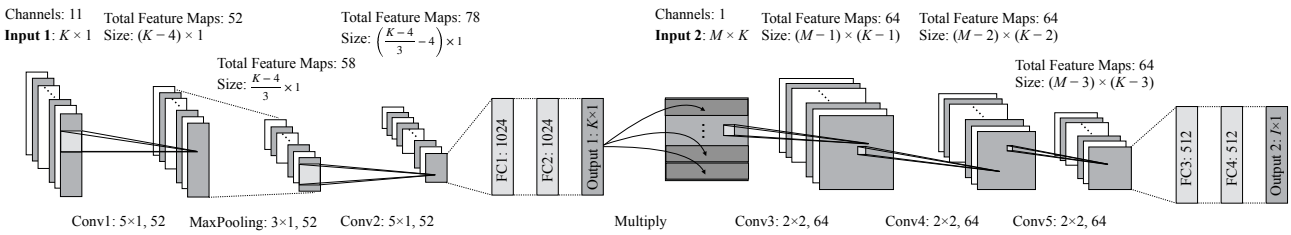


Figure 3: *The proposed multi-task learning architecture for DOA estimation.* Input 1 *is the 11-frame magnitude spectrum and* Input 2 *is the 1-frame phase spectrum.* Output 1 *is the estimated mask and* Output 2 *is the DOA classification result.*

match the task of DOA estimation. Thus we propose a multi-task learning architecture to alleviate the mismatch problem between these two modules, and the two networks are trained jointly which will force them to learn a more suitable mask for DOA task and accurate DOA estimation instantaneously. The whole architecture is illustrated in Fig.3.

There are two inputs and two outputs in this architecture. The first input is the log-magnitude spectrum which is fed into the T-F mask network, and the second input is the phase spectrum which is multiplied by the predicted mask outputs first and then fed into the DOA network. The two outputs are the estimated T-F mask and the DOA classification individually, which are used to calculate the losses for the optimization. The loss function for training is a combination of the mean squared error loss for mask estimation network and the cross entropy loss for DOA estimation network:

$$\mathcal{L}_{\text{multi}} = \alpha \mathcal{L}_{\text{MSE}} + (1-\alpha)\mathcal{L}_{\text{CE}} \tag{5}$$

where α is a constant and set to 0.01 in our experiments.

3.3.2. Pseudo multi-task learning

For a standard multi-task architecture, the losses of two tasks are both considered to optimize the two tasks instantaneously. However, if we only cares about the DOA estimation task, we can regard the other one as an auxiliary task and use the DOA classification loss to update the entire network. Thus we propose a pseudo multi-task learning architecture whose structure is the same as that illustrated in Fig.3, except that the training loss for the mask output is removed. Another motivation is that a magnitude-related mask may not be the best choice for DOA estimation task. Therefore we remove the explicit constraints on the mask estimation output so that the network can learn a mask that best matches the DOA estimation task, and we call this architecture pseudo multi-task learning. The loss function is the same as equation (5) and α is set to 0.

4. Experimental results

4.1. Experimental setup

In this paper, the proposed methods are evaluated using a 6-mic circular array with an inter-microphone distance of 3.6 cm. The input signals are transformed into time-frequency domain by short-time Fourier transform (STFT) with a sampling rate of 16 kHz, DFT length of 256 and 50% overlap. For each time frame, the Hanning window is applied. The DOA range is discretized to 72 classes with a 5° resolution.

To simulate different acoustic conditions, the room impulse responses (RIRs) are generated using the image method [22]. Two configurations are used for training and testing data generation, so that we can evaluate the methods under both matched and mismatched conditions. Two configurations are shown in Table 1 and Table 2.

Table 1: *Configuration for both training and testing data generation under the matched condition. All rooms are 2.5 m high.*

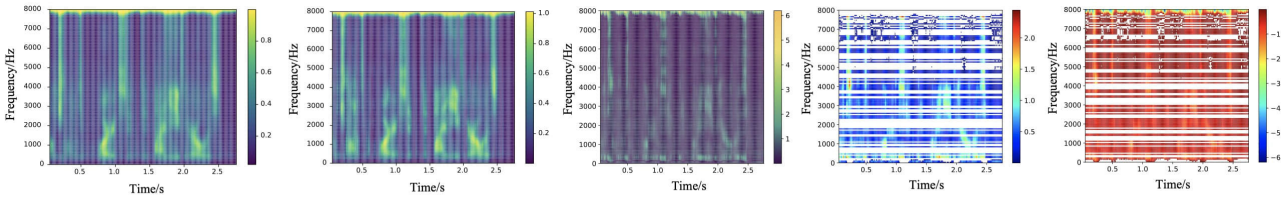
Simulated training data	
Signal	Speech signals from TIMIT Noise signals from CHiME 3
Room size	R1: 6×6 m, R2: 5×5 m
Array positions	arbitrary positions in each room
Source-array distance	1 m and 2 m for each position
RT_{60}	R1: 0.3 s, R2: 0.2 s
SNR	from -5dB to 20dB with 5dB interval

Table 2: *Configuration for testing data generation under the mismatched condition. All rooms are 3 m high.*

Simulated testing data	
Signal	Speech signals from TIMIT Noise signals from CHiME 3
Room size	R1: 7×6 m, R2: 8×8 m
Array positions	arbitrary positions in each room
Source-array distance	1.5 m for both rooms
RT_{60}	R1: 0.45 s, R2: 0.53 s
SNR	from -5dB to 20dB with 5dB interval

To generate the data for the matched condition, the configuration in Table 1 is used. We simulate 500 different array positions for every combination of room size, source-array distance and RT_{60} in Table 1 and generate 4000 RIRs in total. Then we choose 6300 clean utterances from TIMIT database, convolve them with the RIRs and add them with a noise randomly selected from the 3rd CHiME challenge database of noise [23]. In total, the data consists of 37,800 utterances, whose duration is around 32.3 hours. We randomly choose 6,000 and 7,800 utterances from these data as validation set and test set respectively, and the rest as the training data to evaluate the DOA estimation algorithm in the matched acoustic condition. In addition, the data generated with the configuration in Table 2 is utilized for another testing set in the mismatched acoustic condition, and there are 3,000 utterances and the duration is around 2.5 hours in the mismatched testing data.

For CNN training, the input log-magnitude features are all normalized to $[-1, 1]$ and the input phase features are all wrapped to $(-\pi, \pi]$. The mean squared error loss function is used for the mask estimation network while the DOA estimation network use cross entropy loss function. All CNNs are trained for 20 epochs with the Adam optimizer [24], a learning rate of



(a) Separately trained mask (b) Multi-task trained mask (c) The modulus of Pseudo multi-task trained mask (d) Positive part of pseudo multi-task trained mask (e) Negative part of pseudo multi-task trained mask

Figure 4: *Illustration of the estimated masks of proposed methods.*

0.00003 and a mini-batch size of 128. For each fully-connected layer, dropout [25] with rate 0.5 is used to avoid over-fitting.

Table 3: *Segment level accuracy in matched acoustic conditions*[2]

SNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
SRP-PHAT	0.471	0.543	0.629	0.689	0.739	0.760
MUSIC	0.599	0.661	0.715	0.710	0.733	0.710
basic CNN	0.786	0.847	0.894	0.934	0.953	0.968
mask+CNN	0.814	0.865	0.911	0.945	0.957	0.967
mask×CNN	0.826	0.875	0.919	0.950	0.963	0.970
multi-task	**0.833**	**0.882**	**0.923**	**0.955**	**0.964**	**0.973**
pse-multi-task	0.820	0.863	0.914	0.944	0.957	0.966

Table 4: *Segment level accuracy in mismatched acoustic conditions*

SNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
SRP-PHAT	0.412	0.527	0.624	0.684	0.708	0.731
MUSIC	0.534	0.665	0.759	0.795	0.800	0.797
basic CNN	0.616	0.709	0.778	0.829	**0.859**	**0.877**
mask+CNN	0.621	0.707	0.774	0.814	0.845	0.859
mask×CNN	0.634	0.715	0.775	0.809	0.836	0.855
multi-task	0.649	0.725	0.782	0.816	0.844	0.856
pse-multi-task	**0.658**	**0.734**	**0.789**	**0.833**	0.858	0.874

4.2. Results and discussion

4.2.1. Performance of DOA estimation

All testing data are divided into segments with the duration of 500 ms and the performance of DOA estimation methods are evaluated in terms of segment level accuracy. We consider a prediction is correct if the difference between the prediction and the true DOA is less than or equal to 5°.

First we evaluate the performance of different DOA estimation methods in matched acoustic conditions and the results are shown in Table 3. We can see that all CNN-based methods outperform the conventional methods (SRP-PHAT and MUSIC) and our proposed methods show higher accuracy than the basic CNN method. As expected, the result shows that T-F masking is a powerful method for improving robustness, especially in low SNR conditions. The multi-task method has the best performance in almost all SNR conditions, which confirms our previous assumption that a phase-related mask may better match the DOA estimation task.

Then we evaluate the generalization ability of those methods in mismatched acoustic conditions. In Table 4, we can see that the performance of all methods degrades due to the mismatch and the CNN-based methods still outperform SRP-PHAT and MUSIC methods in all cases. The newly proposed CNN predictors with T-F masking still work well in mismatched scenarios, and the improvement is obvious especially for the low SNR levels. This observation shows the good generalization of the proposed methods for DOA.

For the two types of multi-task learning, the pseudo multi-task method has good performance in both matched and mismatched acoustic conditions, although only using one loss for the model optimization.

4.2.2. Mask estimation comparison of different methods

To better understand the mask module in the proposed methods, we compare the different masks estimated by three proposed approaches, and the estimated masks are shown in Fig.4. As expected, the separately trained mask and the mask estimated by the multi-task network have very similar patterns, i.e. the magnitude-related mask. An interesting phenomenon is that in Fig.4 (c), the modulus of the mask estimated by the pseudo multi-task network also has a similar pattern. The difference is that the mask contains positive values as well as negative values and the two parts are shown separately in Fig.4 (d) and (e) respectively. We can see the mask pattern is disconnected but still exists, and the positive and negative parts are just complementary to each other on the frequency bands. This means that the network automatically learns a magnitude-related mask for DOA estimation even without explicit constraints or losses on the mask estimation. Thus a magnitude-related mask is suitable for DOA estimation task.

5. Conclusion

In this work, we propose three DOA estimation methods based on CNN and T-F masking, including integrating the T-F mask module for DOA separately and jointly optimizing T-F mask model with DOA simultaneously. The proposed methods are evaluated in both matched and mismatched conditions. They all show strong robustness in noisy and reverberant environments, and the multi-task joint learning method achieves much better performance than the basic CNN method and conventional methods under low-SNR conditions, without much performance degradation under high-SNR conditions. Further work will consider different input features for mask and DOA estimation and a better metric for mask estimation.

6. Acknowledgements

This work was supported by the China NSFC projects (No. 61603252 and No. U1736202). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

[2] *mask+CNN* and *mask×CNN* denote appending masks to the input and multiplying the input by masks respectively. *pse-multi-task* denotes the pseudo multi-task method in Section 3.3.2.

7. References

[1] M. Wölfel and J. McDonough, *Distant speech recognition*. John Wiley & Sons, 2009.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.

[3] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[4] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *arXiv preprint arXiv:1707.06527*, 2017.

[5] Y.-m. Qian, C. Weng, X.-k. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, 2018.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[7] A. Johansson, N. Grbic, and S. Nordholm, "Speaker localisation using the far-field srp-phat in conference telephony," in *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, 2002.

[8] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[9] W. He, P. Motlicek, and J.-M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," *Proc. Interspeech 2018*, pp. 312–316, 2018.

[10] S. Sivasankaran, E. Vincent, and D. Fohr, "Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment," in *Interspeech*, 2018.

[11] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2814–2818.

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

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

[14] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on*. IEEE, 2017, pp. 136–140.

[15] S. Chakrabarty and E. A. Habets, "Multi-speaker localization using convolutional neural network trained with noise," *arXiv preprint arXiv:1712.04276*, 2017.

[16] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1849–1858, 2014.

[17] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[18] Y. Zhou and Y. Qian, "Robust mask estimation by integrating neural network-based and clustering-based approaches for adaptive acoustic beamforming," in *Int Conf on Acoustics, Speech, and Signal Processing, in press. Google Scholar*, 2018.

[19] P. Pertilä and E. Cakir, "Robust direction estimation with convolutional neural networks based steered response power," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6125–6129.

[20] A. Karbasi and A. Sugiyama, "A doa estimation method for an arbitrary triangular microphone arrangement," in *European Signal Processing Conference 2006 (EUSIPCO 2006)*, no. LCAV-CONF-2009-019, 2006.

[21] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM), 2017 IEEE International Workshop of*. IEEE, 2017, pp. 1–5.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[23] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 504–511.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.