Hindawi
Complexity
Volume 2020, Article ID 5819624, 7 pages
https://doi.org/10.1155/2020/5819624

Research Article

An Adaptive Method Based on Multiscale Dilated Convolutional Network for Binaural Speech Source Localization

Lulu Wu,[1] **Hong Liu**,[1] **Bing Yang**,[1] **and Runwei Ding**[1,2]

[1]*Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Shenzhen 518055, China*
[2]*School of Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China*

Correspondence should be addressed to Hong Liu; hongliu@pku.edu.cn

Received 27 June 2020; Revised 16 November 2020; Accepted 24 November 2020; Published 30 December 2020

Academic Editor: Zhile Yang

Most binaural speech source localization models perform poorly in unprecedentedly noisy and reverberant situations. Here, this issue is approached by modelling a multiscale dilated convolutional neural network (CNN). The time-related crosscorrelation function (CCF) and energy-related interaural level differences (ILD) are preprocessed in separate branches of dilated convolutional network. The multiscale dilated CNN can encode discriminative representations for CCF and ILD, respectively. After encoding, the individual interaural representations are fused to map source direction. Furthermore, in order to improve the parameter adaptation, a novel semiadaptive entropy is proposed to train the network under directional constraints. Experimental results show the proposed method can adaptively locate speech sources in simulated noisy and reverberant environments.

1. Introduction

Speech source localization (SSL) attracts growing attention in the past decades. It is widely applied in human-robot interaction systems and video conference systems. Binaural speech source localization is a subtask of speech source localization, aiming at estimating the direction of arrival (DOA) of a speech source utilizing audio signals recorded by binaural microphones mounted in artificial ears of a dummy head [1]. The pipeline of binaural speech source localization contains two steps. Firstly, extracting interaural cues, i.e., interaural time differences (ITD) and interaural level differences (ILD) from received binaural signals [2–4]. With the inclusion of the dummy head, the frequency-dependent characteristics of spatial cues can be captured by the head-related transfer function (HRTF) [5, 6]. This frequency dependency motivates the use of time-frequency representations for binaural signals. A typical time-frequency representation for binaural signals is based on Gammatone filters which are usually used to simulate the peripheral processing of human auditory system [7–9]. The second step for DOA estimation is to apply geometric analysis technique [1] or off-line models [4, 7] to map interaural cues to sound source DOA. Over years, most methods were proposed to improve the performance of binaural SSL from two aspects: estimating robust interaural cues and improving the generalization of learning-based models.

Interaural time difference is the time delay corresponding to the maximum value of the crosscorrelation function of the left and right microphone signals. Interaural level difference is the logarithmic difference of the power energy between left and right microphone signals. However, in the noisy and reverberant environments, there would be additional peaks in the crosscorrelation function and power energy loss of the target speech source. The additional peaks and energy loss would lead to unreliable interaural cues estimation. In order to refine these unreliable interaural cues, the time-delay compensation method was proposed to align ILD and ITD [10], reverberation weighting method was proposed to suppress early and late reverberation [11], and echo-free onsets detection method was proposed to detect direct-path signals [12]. Since ITD is more robust at low frequencies (lower than 1.5 kHz) and ILD is more reliable at high frequencies [13], the Gammatone filters are usually used to filter the low and high frequencies. Karthik and Ghosh used Gammatone filters to preprocess the binaural

signals and mapped the frequency-dependent ITD to azimuths using ITD-azimuth templates [14]. May et al. modeled the ITD and ILD in sub-bands for every source direction using Gaussian mixture models (GMMs) [7]. In the scene with multiple activate speech sources, the time-frequency (TF) representation of binaural signal is also able to distinguish noise and speech source in different fragments. Christensen et al. investigated different TF weight estimation approaches for interaural cues [15]. Recently, deep neural network has shown significant performance of speech source localization against noise and reverberation, including time-frequency masking estimation [16] and multi-source localization [17]. Convolutional neural network (CNN) can be used to estimate broadband direction of arrival (DOA) of speech source using phase components [17] and to jointly locate and classify multiple speech sources [18]. Frequency-dependent deep neural network (DNN) and head movements can be exploited to detect multiple DOAs and identify front-back confusions [19]. However, training such a robust and well-generalized model requires a large number of various acoustic conditions. There are few studies that are proposed to improve the adaptability of a model to previously unseen conditions. Takeda and Komatani proposed a training scheme for unsupervised adaptation of DNNs' parameters using self-entropy and parameter selection [20], and Wang et al. proposed a data-efficient method based on DNN and clustering to improve binaural localization performance in the mismatched HRTF condition [21], but the localization performance still stays poor. In order to solve the off-grid problem, an off-grid BSSL method based on an off-grid wideband sparse Bayesian learning algorithm is proposed, which is only better than the state-of-the-art HRTF-based BSSL methods [22]. It remains challenging how to generalize the learning-based model and make it adaptively locate binaural signals in previously unseen and adverse acoustic conditions.

Here, we propose a multiscale dilated CNN-based method to further disentangle these issues. The cross-correlation function (CCF) and interaural level difference (ILD) are extracted from binaural signals as input features. In order to preserve the detailed spatial information, the CCF and ILD are separately preprocessed in different dilated CNNs with specific dilation factors. Afterwards, both encoded interaural representations of CCF and ILD are fused to learn crossdomain information. The crossdomain information encoded by multiscale dilated CNNs provides trade-off between small and large receptive fields for CCF and ILD features to better generalize the network in diverse acoustic conditions. In this network, a remaining problem is how to adapt network parameters to unseen acoustic conditions. Drawing on the research of unsupervised adaptation of network parameters [20], we also propose a semiadaptive entropy as the objective function. Different from self-entropy, the semiadaptive entropy includes the crossentropy part to improve the localization performance. Besides, a learning factor is used to weight the attention of cross-entropy and self-entropy.

In summary, our contributions are as follows:

(i) We propose a multiscale dilated CNN framework for binaural speech source localization, which effectively encodes crosscorrelation function and interaural level difference features from different dilation factors.

(ii) We propose a semiadaptive entropy for CNN's parameter adaptation. Experimental results demonstrate that multiscale dilated CNN trained with semiadaptive entropy achieves significant improvements over regular DNN and CNN in noisy and reverberant acoustic environments.

2. Multiscale Dilated CNN

Suppose that there is only one target speaker, the received binaural signals can be formulated by convolving speech signal and head-related impulse responses (HRIR) in the time domain as

$$y_m(n) = s(n) \otimes h_m(n) + v_m(n),\ m \in \{l, r\}, \tag{1}$$

where the symbol $\otimes$ represents convolution operation, m represents the binaural microphone index, l and r refer to the left and right microphones, n is the index of time frame, $s(n)$ denotes the speech signal, and $h_m(n)$ denotes the head-related impulse response. In order to resemble the frequency selectivity of the human cochlea, binaural signals are decomposed into 32 auditory channels using a fourth-order Gammatone filter bank [23]. The centre frequencies of Gammatone filters are logarithmically equally spaced on the equivalent rectangular bandwidth scale between 80 Hz and 8 kHz. After filtering binaural signals, the crosscorrelation function is computed between the left and right signals in each frequency sub-band independently. The CCF is further normalized by the autocorrelation of the left and right signals. The CCF is formulated as a function of time delay τ:

$$ccf_{n,k}(\tau) = \frac{R_{lr}(n, k, \tau)}{\sqrt{R_{ll}(n, k, 0) R_{rr}(n, k, 0)}}, \tag{2}$$

where $R_{lr}(n, k, \tau)$ denotes the crosscorrelation between left and right signals and k is the index of frequency sub-band. $R_{ll}(n, k, 0)$ and $R_{ll}(n, k, 0)$ denote the autocorrelation of left and right signals at $\tau = 0$, respectively. Generally, the diameter of artificial ears of the dummy head is about 15–17 cm. According to the sound propagation speed, the arrival time difference between two ears can be estimated within ± 1.1 ms. In the realistic conditions, considering the head shadowing effect, the maximum time delay is set to 2 ms. For example, the crosscorrelation function of binaural signals sampled at 16 kHz within a range of centre delays ± 2 ms forms a matrix CCF with size of 32×65. The other interaural cue ILD is energy difference in logarithmic between binaural signals, which is formulated as follows:

$$ild_{n,k} = 10 \log_{10} \frac{\sum_{t \in \mathbb{S}\{n\}} y_r^2(n, k, t)}{\sum_{t \in \mathbb{S}\{n\}} y_l^2(n, k, t)}, \tag{3}$$

where $\mathbb{S}\{n\}$ denotes the set of a series of sample indexes t in the nth frame. Since the binaural signals are framed into short and stable speech signals, there would be nonenergy frames. These nonenergy frames would be disregarded. The interaural level difference of binaural signals forms a vector ILD with size of 32×1 in all frequency sub-bands.

2.1. Network Architecture. SSL can be regarded as a direction classification task based on CNN. By dilating dense convolutional kernels with zeros, dilated CNN can operate on a coarser receptive field and show robust performance for voice activity detection in noisy environments [24]. Therefore, the dilated CNN is considered in our network to encode robust interaural features. The schematic diagram of the proposed multiscale dilated CNN is depicted in Figure 1. Two examples of dilated kernels with kernel size of 3 are shown in the upper right side of Figure 1. The number of zero cells between adjacent cells depends on the dilation factor (DF). Black blocks denote the parameter of convolutional kernels to activate corresponding input cells, while white blocks denote zeros to keep input cells inactivated. The number of zeros between two activated cells is DF–1.

In binaural speech source localization, the CCF and ILD reflect time-related and energy-related physical information, respectively. In our method, separated branches of multiscale dilated CNN are designed to better capture independent interaural characteristics according to their physical meanings. The branch for CCF consists of two parallel dilated CNNs, one of which stacks two dilated CNN layers with DF = 2 (i.e., dilation-2 CNN) and the other branch stacks two dilated CNN layers with DF = 5 (i.e., dilation-5 CNN). This multiscale dilated CNN is designed to locate the azimuths of binaural signals in the frontal hemifield with range of $[-90^\circ, 90^\circ]$. Taking 37 azimuths spaced at a step of 5° as examples, 65 samples of time delays of CCF are exactly twice the number of DOAs. The DOA of a signal is estimated by considering the maximum of crosscorrelation and the surrounding values of this maximum in a kernel. In reality, adjacent DOAs within some angular distances are also considered. With this in mind, we implicitly include the tolerance errors of 5° and 10° by setting dilation factors to 2 and 5. The kernels with dilation factors 2 and 5 describe the tolerances ranging in $[0^\circ, 10^\circ]$. Here, dilation factor 4 is not included since it can be obtained by moving kernels with dilation factor 2 twice. The other branch for ILD consists of only one layer of dilated CNN with dilation factor 2. All CNN layers employ 64 kernels to double expand frequency bands and are activated by rectified linear unit activation function and a dropout probability 0.5. The max-pooling layers are added after each dilation-2 CNN to reduce parameters but are excluded in dilation-5 CNN to preserve details. Finally, all interaural representations are fused in a fully connected layer with 128 neurons and followed by an output layer with Softmax activation function. The aforementioned parameters are sufficiently evaluated in experiments.

2.2. Semiadaptive Entropy. As mentioned before, adjacent azimuths within some tolerances can be considered correct. Additionally, due to the intermittence of speech, weak-speech frames are inevitably dominated by noise. In this section, we propose a semiadaptive entropy to train multiscale dilated CNN. In most regression tasks, the Kullback–Leibler divergence (KLD) is widely used to measure the similarity between two probability distributions. In this paper, the probability distributions refer to the true DOA and the estimated DOA in binaural speech source localization. The KLD can be formulated as a sum of the "truth" entropy and the soft crossentropy:

$$D_{KL}(q \,||\, p) = \sum_i (q_i \cdot \log q_i - q_i \cdot \log p_i), \tag{4}$$

where q_i and p_i denote the probabilities of the true DOA and the estimated ith azimuth, respectively. The DOA probability of a silent or noise-dominant frame is assumed to be uniformly distributed on I azimuths. With this assumption, the "truth" entropy of KLD is substituted by a uniform entropy. Besides, a learning factor λ is applied to balance the crossentropy and the uniform entropy:

$$\mathcal{J} = -(1-\lambda)\mathbb{E}\left[\sum_i q_i \log p_i\right] - \lambda\mathbb{E}\left[\sum_i \frac{1}{I}\log p_i\right], \tag{5}$$

where $\mathbb{E}[\cdot]$ means averaging over training samples. Under directional constraint ($\lambda \neq 1$), the network is able to fine-tune parameters under diverse acoustic conditions. The ADADELTA [25] algorithm is used to minimize the loss function. Training process would be stopped if no lower error appears on the validation set within last 3 epochs. The azimuth probability $\mathcal{P}(\theta)$ of a received signal block consisting of contextual frames is produced by averaging frame-level azimuth probabilities. The target DOA is estimated by maximizing $\mathcal{P}(\theta)$.

3. Experiments and Discussion

3.1. Experimental Setup. The proposed method is evaluated using a binaural setup in simulated acoustic conditions, including signal-to-noise ratio (SNR), noise types, and reverberation time. Acoustic conditions are summarized in Table 1. Speaking sources are positioned in the frontal plane between -90° and 90° with a step of 5°, i.e., 37 directions, and their elevations are the same as the receiver's. Based on the binaural signal formulation, the head-related impulse response (HRIR) from the KEMAR dataset [26] are convolved with speech recordings from TIMIT dataset [27]. To simulate the noisy conditions, six kinds of common noises from the NOISEX-92 dataset [28] are properly truncated and added to each microphone signal based on the same SNR. Each noise is processed as diffuse noise by summing all the directional noises generated by convolving the noise and HRIR at 37 uncorrelated directions. To simulate the reverberant conditions, an enclosure of $(10 \times 6 \times 3)$ m is simulated using the Roomsim toolbox [29] based on the image method [30]. All surfaces in this room are equally reverberant. A dummy head indexed by Subject_021 from the CIPIC dataset [31] is placed at the centre position. The source-to-sensor distance is 1.5 m. The binaural room impulse responses yielded by this reverberant setup are convolved with testing speech

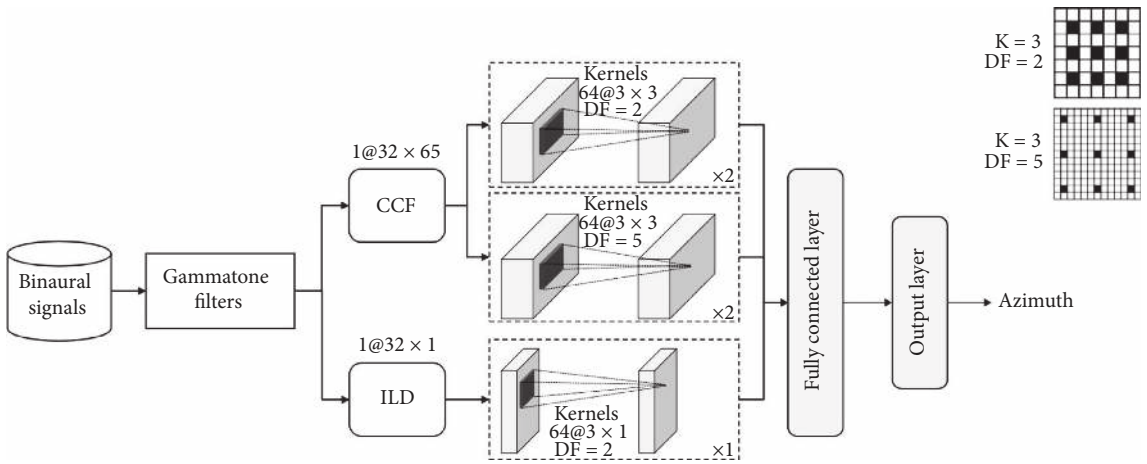


Figure 1: Schematic diagram of the multiscale dilated CNN.

Table 1: Configuration of training and testing sets.

	Training set	Testing set
KEMAR HRIRs	Anechoic HRIRs	Anechoic HRIRs with headphone AKGK271 MK II
TIMIT speech recordings	10 males and 10 females	Other 3 males and 3 females
Source-to-sensor distance	0.5 m, 1 m, 2 m, 3 m	1 m, 1.5 m
Noise types	Babble, destroyerops and factory1	White, m109 and f16
SNRs	−20 dB: 15 : 25 dB	−10 dB: 10 : 30 dB
Reverberation time $\overline{RT}_{60}$	None	0.1 s, 0.3 s, 0.5 s
Direct-to-reverberant ratio (DRR)	None	−1.44 dB, −2.02 dB, −2.58 dB
Number of binaural mixtures	52369 noise-free and noisy signals and 5819 for validation set	936 for each kind of noise and SNR, and 1221 reverberant signals

recordings to generate a reverberant data set. All binaural speech mixtures are sampled at 16 kHz and framed by a Hamming window of 512 samples with a shift of 256 samples. A signal block contains 20 contextual frames, equivalent to a segment with 336 ms duration. The localization performance is measured in terms of the localization accuracy, which considers an estimated DOA is correct if the estimated DOA is within 5° away from the true DOA.

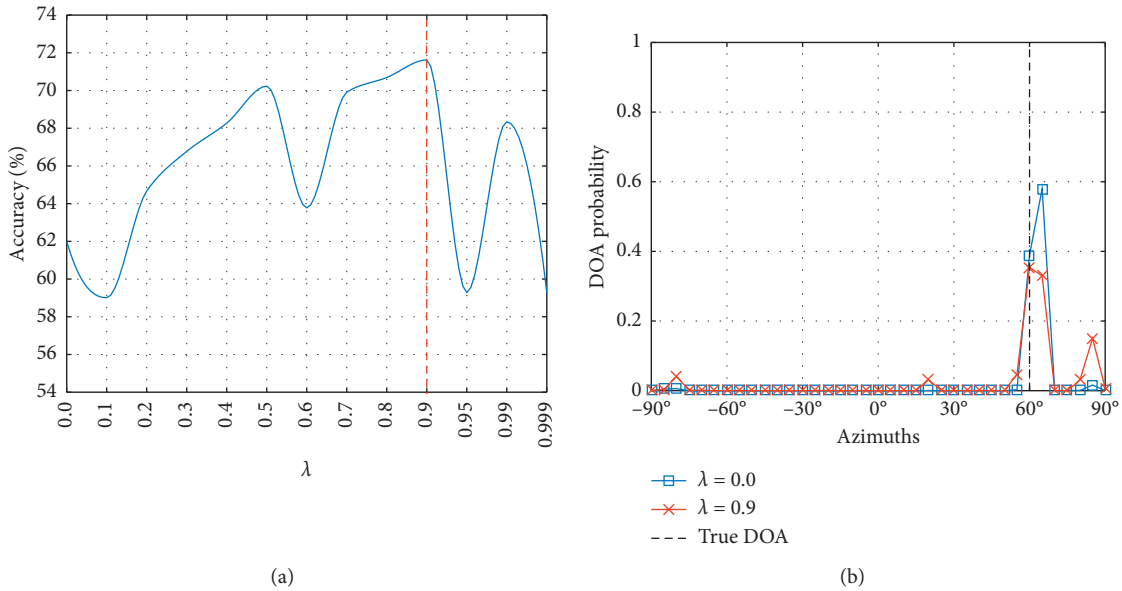
3.2. Influence of Learning Factor. The adaptability of our network is influenced by the learning factor so that the value of λ needs to be evaluated to maximize the adaptability. Note that the semiadaptive entropy lacks directional information when $\lambda = 1$; hence, the maximum value of λ is set to 0.999. The minimum value of λ is set to 0; thus, the semiadaptive entropy becomes crossentropy. In experiments, our network is trained with different learning factors ranging from 0 to 0.999 and λ is determined by evaluating the localization accuracy on the validation set under noisy conditions with −20 dB SNR. Figure 2(a) shows the localization performance with different λ. There are three local maxima in Figure 2(a) with different learning factors $\lambda = 0.5, 0.9$ and 0.99, respectively. During the ADADELTA [25] updating algorithm, the learning rate is automatically updated using accumulated gradient:

$$E\left[\Delta x^2\right]_t = \rho E\left[\Delta x^2\right]_{t-1} + (1-\rho)\Delta x_t^2. \tag{6}$$

The formulation of our semiadaptive entropy also looks like the form of this accumulated gradient. The gradient of each term of the semiadaptive entropy can be calculated separately and the accumulated gradient becomes

$$E\left[g^2\right]_t = (1-\lambda)\rho E\left[g_1^2\right]_{t-1} + (1-\lambda)(1-\rho)g_{1t}^2 + \lambda\rho E\left[g_2^2\right]_{t-1} + \lambda(1-\rho)g_{2t}^2, \tag{7}$$

where g_{1t} and g_{2t} represent the gradient of the crossentropy and the uniform entropy, respectively. Here, λ is also a hyperparameter and serves as a momentum factor to control the learning rate. Therefore, the model could fall into different local maxima or saddle points during updating iteration. Through sufficient validation, the λ is set to 0.9 with the best performance, indicating relatively high adaptability of this network in noisy environments. A DOA probability of a binaural signal in -10 dB SNR condition is depicted in Figure 2(b). The real DOA of the signal is 60°, but it gets wrong DOA of 65° when the network is trained with $\lambda = 0$. Red curve shows the wrong DOA probability is reduced when training network with $\lambda = 0.9$. In addition, due to the effect of the uniform entropy, the azimuths far away from the



(a) (b)

Figure 2: (a) Evaluation of the learning factor λ. (b) DOA probabilities for a noise-dominant frame positioned at 60°.

true DOA may have nonzero probabilities. It is demonstrated that semiadaptive entropy can effectively improve the adaptability of the network.

3.3. Evaluation of Binaural SSL. Our method is compared with two baseline network-based methods, i.e., multilayer perceptron (MLP) [8] and frequency-dependent DNN [19], and the network architecture is also evaluated in ablation studies:

Regular CNN: the regular CNN is used in our architecture instead of dilated CNN

Dilation-2 CNN: the CCF and ILD are fed into separate branches of dilated CNN as in the proposed architecture, but the CCF branch only stacks two layers of dilation-2 CNN

Dilation-5 CNN: the CCF and ILD are fed into separate branches of dilated CNN as in the proposed architecture, but the CCF branch only stacks two layers of dilation-5 CNN

Cascaded DCNN: the dilation-2 CNN and dilation-5 CNN are cascaded in the CCF branch rather than parallel.

Localization accuracies of these methods are shown in Table 2 (in noisy scenes) and Table 3 (in noisy and reverberant scenes). In Table 3, the symbol "-/-" means no additive noise. In noisy conditions, MLP outperforms the frequency-dependent DNN in low-SNR conditions, which is because the ITD and ILD are estimated on the whole signal block rather than short frames. Compared with the results of DNN, the CNN-based methods improve the average accuracy by 2% to 6%. The reason is that adjacent frequency bands can provide mutual information for each other rather than independent frequency bands. In reverberant conditions, the dilation-5 CNN outperforms the others since the remote information is equally important to the mutual information in cross sub-bands, where the remote information includes the interaural features in direct path, early and late reverberation. The dilated CNN with relatively larger receptive fields can capture more remote information at a time. Due to the complementarity of different dilated kernels, the multiscale dilated CNN trained with $\lambda = 0.9$ performs well in noisy conditions but slightly worse than dilation-5 CNN in reverberant conditions. It makes sense the fusion of multiscale dilated CNN learns an automatic trade-off between small and large dilated kernels in noisy and reverberant conditions. Furthermore, we also demonstrate the importance of the semiadaptive entropy. Compared with crossentropy, the network trained with semiadaptive entropy improves the localization accuracy by nearly 10% in strong noisy scenes and 4.62% on average in reverberant scenes.

4. Conclusions

In this work, we proposed an adaptive binaural SSL method based on multiscale dilated CNN. The separate dilated CNN can encode discriminative representations of CCF and ILD features. By synchronously operating on the inputs, the dilation-2 CNN and dilation-5 CNN complemented each other in noisy and reverberant conditions. Additionally, we derived a semiadaptive entropy from the Kullback–Leibler divergence to adaptively train the network under directional constraints. Training with a high value of the learning factor, the multiscale dilated CNN can generalize well in previously unseen scenes. Experimental results have demonstrated the superiority of this method when compared with other baseline methods and single-scale networks in adverse scenarios.

Table 2: Localization accuracy (%) of different approaches in additively noisy environments.

SNR	—	−10 dB			0 dB			10 dB			20 dB			30 dB		
Noise	Avg.	White	F16	M109	White	F16	M109	White	F16	M109	White	F16	M109	White	F16	M109
MLP [8]	83.77	62.93	53.21	67.41	72.65	71.37	82.37	81.62	86.75	95.51	89.42	96.26	99.15	98.18	99.89	99.89
DNN [19]	82.56	43.16	38.25	53.42	70.30	57.16	86.32	97.65	92.31	100.0	99.89	99.89	100.0	100.0	100.0	100.0
Regular CNN	84.65	54.38	41.99	65.60	73.61	69.76	86.43	89.96	90.81	99.04	98.72	99.79	99.79	99.89	100.0	100.0
Dilation-2 CNN	87.46	45.30	54.17	75.75	70.94	77.67	97.76	97.33	93.59	99.68	99.89	99.89	100.0	100.0	100.0	100.0
Dilation-5 CNN	90.14	62.61	54.17	80.34	83.55	75.85	99.15	97.86	98.61	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Cascaded DCNN	89.62	57.05	54.38	87.61	76.92	76.50	99.25	95.51	97.22	100.0	99.79	100.0	100.0	100.0	100.0	100.0
Ours $\lambda = 0$	89.34	59.83	47.54	78.63	84.19	74.47	98.61	98.40	98.50	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Ours $\lambda = 0.9$	91.85	68.16	56.62	90.06	86.00	80.24	99.36	98.61	98.72	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 3: Localization accuracy (%) of different approaches in the noisy and reverberant scenes.

RT_{60}/DRR	—	0.1 s/−1.44 dB		0.3 s/−2.02 dB		0.5 s/−2.58 dB	
Noise/SNR	Avg.	-/-	White/15 dB	-/-	White/15 dB	-/-	White/15 dB
MLP [8]	28.87	43.24	24.46	33.42	24.19	23.84	24.05
DNN [19]	67.69	92.14	78.11	74.94	53.51	63.81	43.65
Regular CNN	61.40	85.26	79.73	58.23	52.16	49.40	43.65
Dilation-2 CNN	57.69	77.15	75.41	56.02	50.14	43.74	43.65
Dilation-5 CNN	84.03	94.59	89.46	92.14	75.95	86.62	65.41
Cascaded DCNN	73.16	91.15	77.84	84.52	56.62	79.25	49.59
Ours $\lambda = 0$	78.86	93.12	87.97	83.78	71.08	76.50	60.68
Ours $\lambda = 0.9$	83.48	94.59	89.05	90.66	77.70	85.08	63.81

Data Availability

All the data are open and its source is already stated in our paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by National Natural Science Foundation of China (nos. 61673030 and U1613209) and National Natural Science Foundation of Shenzhen (no. JCYJ20190808182209321).

References

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA, USA, 1997.

[2] 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.

[3] X. Wu, D. Talagala, W. Zhang, and T. Abhayapala, "Spatial feature learning for robust binaural sound source localization using a composite feature vector," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6320–6324, Shanghai, China, March 2016.

[4] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1503–1512, 2012.

[5] M. Zohourian, R. Martin, and N. Madhu, "New insights into the role of the head radius in model-based binaural speaker localization," in *Proceedings of IEEE European Signal Processing Conference*, pp. 221–225, Kos Island, Greece, August 2017.

[6] G. Bill, "HRTF measurements of a KEMAR dummy-head microphone," *MIR Media Lab. Perceptual Computing-Technical Report*, vol. 280, pp. 1–7, 1994.

[7] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[8] K. Youssef, S. Argentieri, and J. Zarader, "A binaural sound source localization method using auditive cues and vision," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 217–220, Kyoto, Japan, March 2012.

[9] N. Ma, G. Brown, and J. Gonzalez, "Exploiting top-down source models to improve binaural localisation of multiple sources in reverberant environments," in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 160–164, Dresden, Germany, September 2015.

[10] J. Zhang and H. Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter," *IEEE Transactions on Signal Processing*, vol. 63, no. 18, pp. 4771–4783, 2015.

[11] C. Pang, H. Liu, J. Zhang, and X. Li, "Binaural sound localization based on reverberation weighting and generalized parametric mapping," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1618–1632, 2017.

[12] J. Huang, N. Ohnishi, and N. Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Transactions on Instrumentation and Measurement*, vol. 46, no. 4, pp. 842–846, 1997.

[13] L. Jeffress, "A place theory of sound localization," *IEEE Journal of Comparative and Physiological Psychology*, vol. 61, pp. 468–486, 1947.

[14] G. R. Karthik and P. K. Ghosh, "Binaural speech source localization using template matching of interaural time difference patterns," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5164–5168, Calgary, Canada, April 2018.

[15] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to licalising multiple speakers in reverberant environments," in *Proccedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4593–4596, Taipei, Taiwan, April 2009.

[16] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 178–188, 2019.

[17] S. Chakrabarty and E. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 136–140, New Paltz, NY, USA, October 2017.

[18] W. He, P. Motlicek, and J.-M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 312–316, Hyderabad, Telangana, September 2018.

[19] 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/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017.

[20] R. Takeda and K. Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2217–2221, New Orleans, LA, USA, March 2017.

[21] J. Wang, J. Wang, K. Qian et al., "Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 4, 2020.

[22] J. Ding, J. Li, C. Zheng, and X. Li, "Wideband sparse Bayesian learning for off-grid binaural sound source localization," *Signal Processing (SP)*, vol. 166, Article ID 107250, 2019.

[23] D. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, Piscataway, NJ, USA, 2006.

[24] S.-Y. Chang, B. Li, G. Simko et al., "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5549–5553, Calgary, Canada, April 2018.

[25] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," arXiv preprint arXiv:1212.5701, 2012.

[26] H. Wierstorf, M. Geier, and S. Spors, "A free database of head-related impulse response measurements in the horizontal plane with multiple distances," in *Proceedings of Audio Engineering Society Convention (AES)*, London, UK, May 2011.

[27] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, and D. Pallett, "Darpa TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," Technical Report n 93, NASA STI/Recon, Washington, DC, USA, 1993.

[28] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[29] D. Campbell, K. Palomaki, and G. Brown, "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching," *Computing and Information Systems (CIS)*, vol. 9, no. 3, pp. 48–51, 2005.

[30] 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.

[31] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 99–102, New Paltz, NY, USA, October 2001.