Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Robust Interaural Time Difference Estimation Based on Convolutional Neural Network

Hong Liu, Peipei Yuan, Bing Yang and Lulu Wu

Abstract— **This paper proposes a novel cross correlation function (CCF) extraction method based on convolutional neural network for time difference of arrival (TDOA) estimation or further direction of arrival (DOA) estimation. CNN is utilized to learn the relationship between the cross correlation localization features and the pre-processed waveform signal which may include not only the source signal but also the background noise and reverberation. In contrast to many previous sound source localization approaches, the proposed method focuses on the spatial feature extraction. Two kind of outputs, grouped or encoded CCF, are designed to capture the implicit tendency of location information. The experimental results demonstrate that the proposed method outperforms the conventional TDOA estimation methods under environments with different levels of noise and reverberation.**

I. INTRODUCTION

Sound source localization (SSL) is to determine the azimuth in the horizontal plane, the elevation in the vertical plane or the distance of a sound source, employing the signals received by several microphones. As a nature and effective way for human-robot interaction (HRI), binaural SSL using a pair of microphones equipped on both sides of robot head become more and more important for several decades.

Plenty of methods have been proposed for binaural sound source localization, which mainly consist of three steps. First, the binaural cues such as interaural time difference (ITD) and interaural level difference (ILD) are extracted from the received signals in the time-frequency (TF) domain. ITD mainly describes the time difference of a sound source arriving at two ears, while ILD describes the level difference [1]. Second, the off-line training is executed to build the relationship between the spatial features and the source location. The statistical models including Gaussian mixture model [2], [3], deep neural networks (DNNs) [4] and convolutional neural networks (CNNs) [5] are popular in this stage. Third, online frame-wise or TF bin-wise localization can be achieved via the trained statistical model.

Since ITD carries rich spatial information and TDOA-based SSL approach can be conducted in real time, the ITD extraction become more important for robust binaural SSL. Generalized cross correlation (GCC) methods combining with phase transform (PHAT) [6], Roth [7], smoothed coherence transform (SCOT) [8] processors are the classical approaches to estimate the TDOA between microphone channels. However, the performance may degrades seriously in the presence of low SNR noise signal and high reverberation.

Therefore, several approaches have been proposed to improve the robustness of SSL through obtain more accurate localization feature. TF mask guided SSL methods aim to combine TF mask with the generalized cross correlation-phase transform (GCC-PHAT) or steered-response SNR [9], [10]. It can derive accurate results by emphasizing more reliable TF bins. The sinusoidal tracks were utilized to represent the voiced speech which is sparse in the frequency (spectrum) domain [1]. The main idea of aforementioned methods to eliminate the effect of reverberation and noise is based on the W-disjoint orthogonal (WDO) assumption [11]. Inter-channel Phase Difference (IPD) enhancement based on CNN is proposed to restore the contaminated IPD directly extracted from the input signals [12].

Due to the aforementioned reasons, a novel cross correlation extraction based on CNN is proposed. It directly extract the localization features of direct-path signal from the input signals in essence. Recently, deep learning based SSL algorithms have been shown to give the state-of-art performance in the DOA estimation and TF mask estimation. In this study, the cross correlation function between left and right ear signal is realized by the local convolution operation of CNN. The experimental results with various levels of noise and room configurations demonstrate that the proposed method can extract accurate features for robust sound source localization.

The rest of the this paper is organized as follows. Section II introduces the proposed binaural localization feature extraction method in detail. Section III describes the evaluation framework and experimental setup. The experimental results with various acoustic environments are illustrated in Section IV. Finally, conclusions are given in Section V.

II. CNN-BASED ITD ESTIMATION

Fig. 1 illustrates the schematic diagram of the proposed system for binaural sound source localization. During training, clean speech signal is spatialized by binaural room impulse response (BRIR). Then noise signal is added to the received signal. After divided into several frequency bands by the gammatone filter bank, the mixture signals in the time domain are fed to the CNN, named generalized cross correlation net (GCCNet), in order to learn the feature

*This work is supported by National Natural Science Foundation of China (NSFC No.61673030, U1613209), Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No. ZDSYS201703031405467), National Engineering Laboratory for Video Technology - Shenzhen Division.

H. Liu, P. Yuan, B. Yang and L. Wu are with the Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, Shenzhen 518055, CHINA. hongliu@pku.edu.cn, peipeiyuan@pku.edu.cn, bingyang@sz.pku.edu.cn and luluwu@sz.pku.edu.cn

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

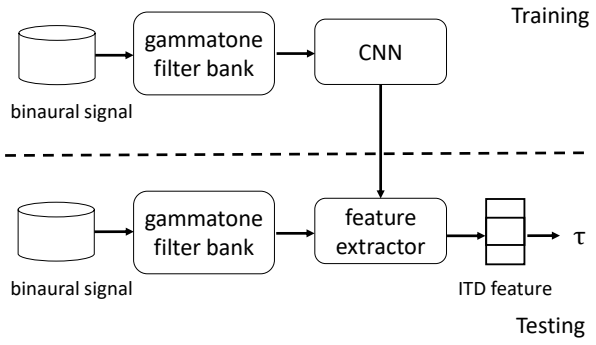

Fig. 1. Schematic diagram of the proposed generalized cross correlation CNN (GCCNet) system in training (top) and testing (bottom) phases.

extraction ability from the noisy and reverberant signal. During testing, binaural signals are first processed by the filter bank the same way as the training stage. Then the filtered signals are utilized as input of the GCCNet. ITD is obtained by picking the peak of the prediction of GCCNet. Furthermore, ITD can be utilized for DOA estimation.

A. Pre-processing of Binaural Signals

The received binaural signals $x_i(n)$ which contain noise and reverberation can be modeled as

$$x_i(n) = s(n) \circledast h_i(n) + v_i(n), \forall i = l, r \tag{1}$$

where $\circledast$ denotes the convolution operation, n and i represent the time sampling point, the index of left channel l and right channel r, $s(n)$ denotes the sound signal emitted by the source, $h_i(n)$ represents the impulse response between the source and ears and $v_i(n)$ is the additive noise signal.

The binaural signals are first analysed by an auditory front-end pre-processing, which is a bank of 32 gammatone filters with center frequencies spaced uniformly on the equivalent rectangular bandwidth (ERB) scale from 80 Hz to 8 kHz, which almost covers the entire speech spectrum [13]. After processed by the gammatone filter bank and a following half-wave rectification, binaural signals are further enframed by a window of 20 ms with a frame shift of 10 ms. The signal sampling rate is 44.1 kHz in this work, the length of one time frame is 882 time points.

There are two kind of data usually exploited as input feature for learning-based SSL approaches. The first one is inter-channel features such as ITD, ILD, cross correlation function (CCF) [2], [4], [12], [14] or the subspace-based features [15], [16]. The second one is the more original information of the signal such as the real and imaginary parts of short time fourier transform (STFT) [17], the ear signals in the time domain [18] and so on. We take the binaural signals in the time domain as the input of CNN. The signal sampling rate is 44.1 kHZ, the final input data are generated with a dimension of $32 \times 882 \times 2$ (frequency bands $\times$ time frame points $\times$ channels).

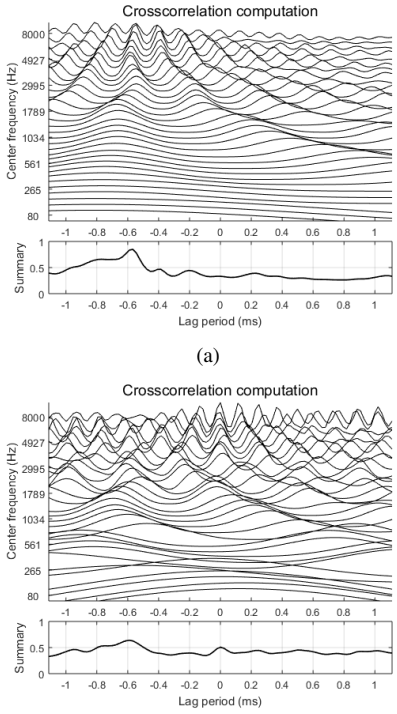

Fig. 2. Cross correlation function (CCF) feature of 32 filter channels extracted from binaural signals where the source is located at azimuth -65° and elevation 0°. (a) clean sound signal. (b) noisy signal with signal-to-noise ratio (SNR) at -5 dB.

B. CNN-based Feature Extraction

The cross correlation function (CCF) of binaural signals is computed as

$$R_{x_l x_r}(\tau) = E\left[x_l(n) x_r(n-\tau)\right], \tag{2}$$

where $E[\cdot]$ denotes expectation. TDOA estimation τ is obtained when $R_{x_l x_r}$ achieves the maximum [6]. However, the cross correlation values may spread or smeared and make it difficult to distinguish peaks in the presence of interference. As shown in Fig. 2, it can be observed that the peak in the summation for clean sound signal is sharper and smoother than the one for noisy signal. And there may appear fake peaks leading to wrong TDOA.

According to specific training targets, the output data is divided into several groups along the frequency axis and then fed into different sets of FC layers. The proposed two kinds of output are describe as follows in detail.

Grouped CCF: For CCF computed from the clean binaural source signals as the training target, N frequency bands are considered as a group corresponding to one FC layers set. It means that 32 frequency bands are divided into several groups and all bands within a group shares the CNN parameters. The cross correlation values fluctuate violently as the frequency increase and adjacent frequency bands can be regarded as having the similar characteristics.

Encoded CCF: Since the degree of fluctuation of CCF presented in the high frequency bands makes it difficult to learn the relationship as Fig. 2 shown. Although the CCF

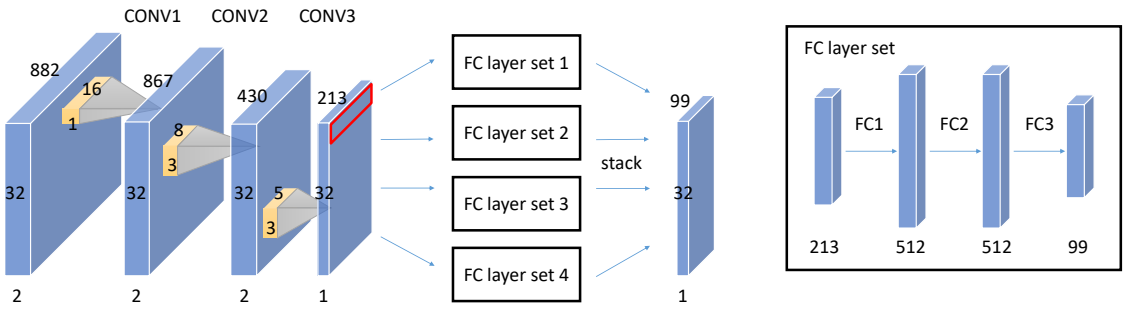

Fig. 3. The architecture of GCCNet-Grouped.

contains more spatial information than ITD, the position of the main peak in CCF is the most important information. Furthermore, the role of the aforementioned weights of GCC is to compensate for the presence or absence of signal power [6]. The Gaussian encoding [17] of the main peak of CCF is proposed instead of regarding the whole CCF as training target. The desired output of each frequency band are the maximum of Gaussian functions centered around the main peak of CCF as

$$\hat{R}_j = e^{(L_j - L_{mp})^2/\sigma^2}, \tag{3}$$

where L_j denotes the discrete time lag at index j, L_{mp} represents the lag corresponding to main peak. Here, σ controls the width of the Gaussian distribution, which is formulated as

$$\sigma(k) = f(k), \tag{4}$$

where $f(k)$ is the frequency-dependent function describing the trend of reliability of the main peak as the frequency increases, and we set $f(k)$ to $e^{-0.1k+3.8}$ in this work. The width of Gaussian functions controlled by σ is narrower than the true main peak, which makes it more robust in the presence of noise and reverberation.

The CNN consists of an input layer, three convolutional layers (CONV), three fully connected (FC) layers and an output layer as illustrated in Fig. 3. Three convolutional layers perform the cross correlation along the frequency bins and time sample points axis, which aim to realize the cross correlation between different channels. The kernel size of the first convolutional layer is set to 1×16. The second convolutional layer has a 2-D kernel of size 3×8 and the third one has a 2-D kernels of size 3×5. Each convolutional layer is followed by rectified linear (ReLU) activation function [19]. The number of nodes of each FC layer is set to 512, 512 and 99 respectively. The last fully connected layer is followed by sigmoid activation function.

C. Final TDOA Estimation

At last, all outputs of the corresponding FC layer sets are stacked together and form a 32×99 dimensional output. The TDOA estimation is obtained by picking the peak of the CCF summation. We can also estimate TF-wise TDOA from each frequency band.

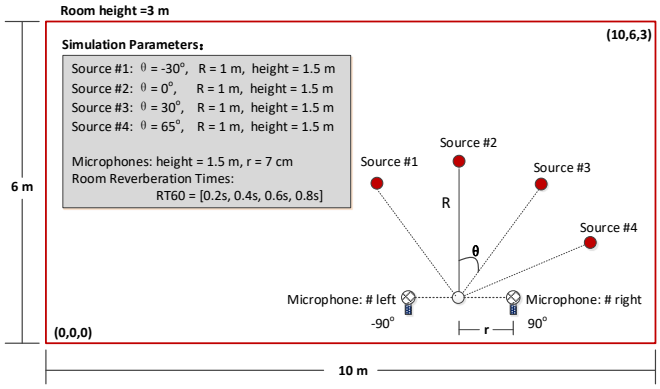

Fig. 4. Simulated scene and parameters of acoustic environments for room1. The average radius of heads in the CIPIC HRTF database is 7 cm approximately.

III. EVALUATION

A. The Dataset

To evaluate the performance of the proposed system, binaural signals are simulated by convolving clean speech signals with the head-related impulse responses (HRIRs) as Eq. (1) shown. There are 25 azimuths ranging from $[-80°, -65°, -55°, -45° : 5° : 45°, 55°, 65°, 80°]$ ($0°$ locates at the middle front of the head) defined in interaural-polar coordinates. HRTFs are measured for 45 different subjects including 27 males, 16 females and KEMAR with large and small pinnae in the CIPIC HRTF database [20]. In this work, the subject #21 (i.e., Kemar head) is selected to generate spatial information received by binaural microphones with 25 azimuths and one elevation ($0°$ in constant), which means the sources from the frontal plane are only considered in the following experiments.

Room impulse responses for four rooms were also simulated via Roomsim toolbox [21] based on image method. The simulated scene and detailed parameters of room size, the distance between source and head, the head location in each room and reverberation time (RT_{60}) are listed in Table I and illustrated in Fig. 4, respectively.

The speech recordings from TIMIT database [22] are taken as the source signals. The training dataset contains 20 sentences randomly selected from the TIMIT training data,

TABLE I
ROOM CONFIGURATION FOR TRAINING AND TEST DATASET

Dataset	Room	W (m)	L (m)	H (m)	R (m)	Center of head (m)	RT_{60} (sec)
training, validation	room1	10	6	3	1	(5, 3, 1.5)	0.2, 0.4, 0.6
	room2	6	5	4.5	1	(3, 2.5, 1.2)	0.2, 0.4, 0.6
	room3	6	4	3	1	(2, 2, 1.2)	0.2, 0.4, 0.6
test	room4	5.5	8	4	1	(3.5, 1.5, 1.2)	0.2, 0.4, 0.6, 0.8

while the validation dataset and test dataset both contain 7 sentences randomly selected from the TIMIT test data respectively. After convolving each sentence with all 25 HRIRs, the babble noise from NOISEX-92 database [23] was regarded as the interference, which was added to the audio signals. For the training and validation dataset, the noisy signals were generated with SNRs ranging from [0:10:30] dB, while the SNRs for test dataset belong to the range of [-5:10:25] dB.

B. Experimental Setup

The feature extraction ability of convolutional neural networks are trained to be frame-wise. For the CCF block as training target, the whole CCF is divided into 4 parts along the frequency and fed into 4 fully connected layer sets respectively, which provides a good tradeoff between frequency resolutions and computational cost.

TABLE II
PARAMETERS OF SIGNAL USED IN EXPERIMENTS.

Parameter	Value
Sampling frequency	44.1 kHz
Window type	hanning
Frame length (FFT length)	882 points (20 ms)
Frame overlap	441 points (10 ms)
Lag boundary	99 points ([-1.1, 1.1] ms)
Frequency-related function	$f(k) = e^{-0.1k+3.8}$

In this work, the feature extraction CNN were trained under two kinds of acoustic conditions:

- using anechoic signals in the training set;
- using noisy and reverberant signals in the training set (i.e., multi-conditional training, MCT).

For all experiments, one batch is composed of 128 frames started from 11-th frame per signal, the learning rate is set to 0.003 initially and divided by 3 every 10 epochs until the performance of validation set is no longer improved. The Adam optimizer is utilized to minimize the mean absolute error (MAE) during training.

C. Methods for Comparison

Two TDOA estimation algorithms are compared with the proposed approach. The first one is the classical cross correlation approach [6] without specific weight. The disturbed binaural signals obey the same pre-processing before extraction, where the mixed binaural signals were first processed by a gammatone filter bank and a half-wave rectifier. Then cross correlation of 32 frequencies with a lag range of $\pm$1.1 ms was performed. This method is denoted as GCC. The second method is DNN-based TDOA estimation with integrated time-frequency masking [24]. Furthermore, the model with implicit mask training procedure is selected for comparison due to its outstanding performance. The number of hidden units in LSTM is set to 160. The number of discrete time delay, DFT bins and mel-frequency bands is set to 99, 882 and 30, respectively. Both Long Short-Term memory (LSTM) in the whole DNN architecture contain 3 hidden layers. During training, the early stop and the learning rate attenuation strategies are as same as the training procedure of proposed method. We call this approach TDOA-Mask.

Two proposed models for comparison are listed as follows:

- GCCNet-Grouped. The proposed feature extraction CNN with the CCF separated into blocks as the training target.
- GCCNet-Encoded. The proposed cross correlation extraction CNN with the Gaussian function encoding the main peak of CCF as the training target.

The root-mean-square error (RMSE) of time delay of arrival was measured by

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\tau}(t) - \tau(t))^2}, \tag{5}$$

where t denotes the time frame index, T is the frame number, $\hat{\tau}$ denotes the TDOA estimated from mixture signal in each frame and τ denotes the true TDOA (experiments over full frequency bands) or TDOA estimated from anechoic single source signal (experiments across frequency sub-bands).

IV. EXPERIMENTS AND DISCUSSION

A. Anechoic Training

To evaluate the performance of feature extraction ability for all comparison methods, the RMSE of the final TDOA estimation across full frequency bands is measured. It can be seen from Table III that both GCCNet-Group and GCCNet-Encoded models have capacity to extract more accurate cross correlation feature than the GCC and TDOA-Mask do. Especially, the role of Gaussian function encoding can be demonstrated by the experimental results that GCCNet-Encoded model performs the best.

The performance of TDOA estimation for each frequency band is also shown in Table IV. We estimate the TDOA from the peak of each frequency band and then compute the RMSE over all TF-wise results. It can be proved that

TABLE III

THE RMSE OF TDOA ESTIMATION (MS) OVER FULL FREQUENCY BANDS FOR MODELS TRAINED IN ANECHOIC ENVIRONMENT.

	SNR (dB)				RT_{60} (s)			
	-5	5	15	25	0.2	0.4	0.6	0.8
GCC	0.295	0.260	0.248	0.246	0.224	0.274	0.338	0.350
TDOA-Mask	0.299	0.278	0.249	0.244	0.241	0.276	0.324	0.331
GCCNet-Grouped	0.221	0.220	0.214	**0.208**	0.195	0.229	0.275	0.284
GCCNet-Encoded	**0.189**	**0.200**	**0.207**	0.210	**0.182**	**0.207**	**0.242**	**0.255**

TABLE IV

THE RMSE OF TDOA ESTIMATION (MS) ACROSS FREQUENCY SUB-BANDS FOR MODELS TRAINED IN ANECHOIC ENVIRONMENT.

	SNR (dB)				RT_{60} (s)			
	-5	5	15	25	0.2	0.4	0.6	0.8
GCC	0.246	0.239	0.239	0.205	0.236	0.239	0.237	0.236
GCCNet-Grouped	0.226	0.227	0.227	0.227	0.226	0.223	0.222	0.220
GCCNet-Encoded	**0.151**	**0.150**	**0.150**	**0.149**	**0.149**	**0.150**	**0.152**	**0.154**

TABLE V

THE RMSE OF TDOA ESTIMATION (MS) OVER FULL FREQUENCY BANDS FOR MODELS TRAINED IN MULTI-CONDITIONAL ENVIRONMENT.

	SNR (dB)				RT_{60} (s)			
	-5	5	15	25	0.2	0.4	0.6	0.8
GCC	0.295	0.260	0.248	0.246	0.224	0.274	0.338	0.350
TDOA-Mask	**0.089**	**0.069**	**0.051**	**0.032**	**0.058**	**0.063**	**0.073**	**0.069**
GCCNet-Grouped	0.185	0.195	0.200	0.203	0.179	0.205	0.225	0.233
GCCNet-Encoded	0.171	0.180	0.188	0.192	0.173	0.190	0.205	0.215

TABLE VI

THE RMSE OF TDOA ESTIMATION (MS) ACROSS FREQUENCY SUB-BANDS FOR MODELS TRAINED IN MULTI-CONDITIONAL ENVIRONMENT.

	SNR (dB)				RT_{60} (s)			
	-5	5	15	25	0.2	0.4	0.6	0.8
GCC	0.246	0.239	0.225	0.205	0.236	0.239	0.237	0.236
GCCNet-Grouped	0.237	0.235	0.235	0.233	0.237	0.240	0.240	0.240
GCCNet-Encoded	**0.134**	**0.136**	**0.139**	**0.142**	**0.136**	**0.137**	**0.140**	**0.143**

the TF-wise localization features make the TDOA or DOA estimation results more robust, which implicitly indicates that the proposed method can be extended to the multiple sound source localization. The GCCNet-Encoded TDOA estimation model exhibits the strongest robustness especially in the low SNR and high reverberation environment. This makes sense as the accurate time delay information is mostly encoded by the Gaussian functions for each frequency band.

B. Multi-conditional Training

The performance of different methods using multi-conditional training procedure are listed in Table V. These methods are tested in environments with various levels of background noise and reverberation time. The performance of TDOA-Mask is outstanding. The possible reason is that the architecture of TDOA-Mask makes it possible to directly generate the TDOA from the masked GCC-PHAT. Meanwhile, our models only produce the CCF feature. Then TDOA are calculated by picking the peak. It also can be observed that the performance of the two CNN model are comparable and are both better than the results presented in Table III, the RMSE of GCCNet-Encoded ranges from 0.171 to 0.215 ms in Table V, while from 0.182 to 0.255 ms in Table III.

Similarly, Table VI presents the RMSE of TDOA estimation across each frequency sub-band. As can be seen, the proposed algorithm GCCNet-Encoded lead to large improvements over GCCNet-Grouped and baseline method, 0.134 to 0.143 ms vs. 0.233 to 0.240 ms and 0.205 to 0.246 ms.

C. Localization in Realistic Environment

The proposed method is also evaluated in realistic environment. The binaural signals are recorded by KU100 dummy head. In Fig. 5, the dummy head is placed in the center of an office and there are two microphones equipped within the "ears" for signal collection. The office environment is of dimensions (6 m $\times$ 5 m $\times$ 3 m). The RT_{60} is 0.3 s approximately due to the material of the walls, the floor

Fig. 5. Sound localization in realistic environment.

and the roof of the room. The SNR is around 20 dB. The distance between the sound source and the dummy head is fixed at 1 m. The azimuth and elevation configuration of real recordings is as same as the one of simulated data.

In Table VII, our models trained using anechoic data perform well and the ones trained using multi-conditional data are comparable with TDOA-Mask. It also demonstrate that the proposed methods relies less on the type of training data.

Training Procedure	Anechoic	MCT
GCC	0.500	
TDOA-Mask	0.411	**0.257**
GCCNet-Grouped	0.367	0.315
GCCNet-Encoded	**0.316**	0.279

TABLE VII
THE RMSE OF TDOA ESTIMATION IN REALISTIC ENVIRONMENT.

V. CONCLUSIONS

This paper presents a novel localization feature extraction approach of TDOA estimation or for further DOA estimation. Two kinds of training target, CCF-grouped and CCF-encoded, are designed for model training. The proposed feature extraction CNN was able to directly extract the accurate spatial features from the filtered waveform binaural signals. Experimental results demonstrate that the encoding procedure of the main peak of CCF performs the best over almost all acoustic environments. In this work, the role of the CNN can be viewed as localization feature extractor for each time frequency bin. We will extend this work for multiple sound source localization in the future.

REFERENCES

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

[2] 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.

[3] 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.

[4] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2122–2131, 2018.

[5] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 8–21, 2019.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[7] P. R. Roth, "Effective measurements using digital signal analysis," *IEEE Spectrum*, vol. 8, no. 4, pp. 62–70, 1971.

[8] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *IEEE*, vol. 61, no. 10, pp. 1497–1498, 1973.

[9] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *INTERSPEECH*, 2018, pp. 322–326.

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

[11] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[12] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1335–1345, 2019.

[13] D. Wang and G. J. Brown, "Computational auditory scene analysis: Principles, algorithms, and applications," pp. 199–199, 2006.

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

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

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

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

[18] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 451–455.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[20] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001, pp. 99–102.

[21] D. Campbell, K. Palomaki, and G. Brown, "A MATLAB simulation of shoebox room acoustics for use in research and teaching," *Computing and Information Systems Journal*, vol. 3, pp. 48–51, 2005.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "The DARPA TIMIT acoustic-phonetic continous speech corpus," *National Institute of Standards and Technology*, 1993.

[23] A. Varga and H. J. 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.

[24] P. Pertila and M. Parviainen, "Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 436–440.