ONLINE DIRECTION OF ARRIVAL ESTIMATION BASED ON DEEP LEARNING

Qinglong Li, Xueliang Zhang and Hao Li*

Department of Computer Science, Inner Mongolia University, Hohhot, China, 010021

31509030@mail.imu.edu.cn, cszxl@imu.edu.cn, lihao.0214@163.com

ABSTRACT

Direction of arrival (DOA) estimation is an important topic in microphone array processing. Conventional methods work well in relatively clean conditions but suffer from noise and reverberation distortions. Recently, deep learning-based methods show the robustness to noise and reverberation. However, the performance is degraded rapidly or even model cannot work when microphone array structure changes. So it has to retrain the model with new data, which is a huge work. In this paper, we propose a supervised learning algorithm for DOA estimation combining convolutional neural network (CNN) and long short term memory (LSTM). Experimental results show that the proposed method can improve the accuracy significantly. In addition, due to an input feature design, the proposed method can adapt to a new microphone array conveniently only use a very small amount of data.

Index Terms— Direction of arrival estimation (DOA), convolutional neural network (CNN), long short term memory (LSTM)

1. INTRODUCTION

Robust DOA estimation is important for many applications such as robotics and beamforming [1]. However, accurate DOA estimation is very challenging when received speech signals are distorted by background noise and reverberation.

Many signal processing approaches are developed for DOA estimation in literature. These approaches can be generally divided into following categories: i) subspace based approaches such as the multiple signal classification (MUSIC) [2], ii) time difference of arrival (TDOA) based approaches that use the family of generalized cross correlation (GCC) methods [3, 4], iii) signal synchronization based approaches such as the steered response power with phase transform (SRP-PHAT) [5], and multichannel cross correlation coefficient (MCCC) [6], iv) model-based approaches such as maximum likelihood method [7]. In practice, these traditional methods generally suffer from either one or a combination of following problems: high computational cost, degradation in performance in presence of low signal to noise ratio (SNR) and/or heavy reverberation environment.

Recently, deep neural networks (DNNs) show their power on speech signal processing, e.g. speech separation [8], automatic speech recognition (ASR) [9] etc. Following this, DNNs are also used for the task of DOA estimation [10–12]. In [10], Takeda and Komatani employed a DNN with 7 hidden layers to predict the DOA with a discriminative training method. The input feature is eigenvectors of correlation matrices at each frequency bin. Experimental results showed that the method is sensitive to the reverberation. In [11], GCC was used as the feature of a one-hidden layer perceptron neural network. Results showed that it is robust to high level noises and strong reverberations. In [12], CNN-based method was proposed, in which phase component of short time Fourier transform (STFT) was used as feature of CNN. The CNN-based method showed robustness to noise and small perturbations in microphone positions. A problem for all above supervised-based methods is that the model has to be retrained overall when the structure of the microphone array changes, e.g. microphone number changing.

In this paper, we proposed a method combining CNN and LSTM to address the online DOA estimation in noisy and reverberant environments. Unlike the existing methods which mainly rely on the array geometry and a short signal observation, the proposed approach uses a special feature to learn the relationship between the received signals and the DOA. In addition, experimental results show that the proposed method is robust to the topologies of microphone array and the trained model can get a better performance on a new microphone array structure use only very few data for adaptation.

2. SYSTEM DESCRIPTION

2.1. DOA Estimation as a Classification Problem

The problem of DOA estimation based on DNN can be formulated as a J-class classification problem, where each class corresponds to a possible angle in a set $\Theta = \{\theta_1, ..., \theta_J\}$, and the estimated angle is given as the class with the highest posterior probability. The number of classes J depends on the array geometry as well as the resolution of the whole range. For example, the angle range of a circular microphone array is from 0° to 350° and the total number of

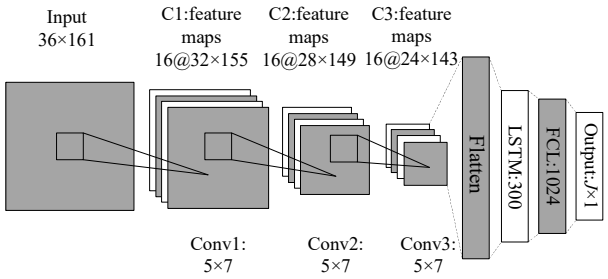

Fig. 1. Proposed model architecture

classes is $J = 36$, when the resolution is $10°$.

2.2. Model Architecture

After many trials, we select the best model, the architecture of the proposed model is shown in Fig. 1. The model consists of three convolutional layers, one LSTM layer and a full connection layer (FCL). The activation function of three convolutional layers and FCL is rectified linear units (ReLU) [13]. Softmax [14] is used as the final-layer activation function. Cross-entropy [15] is used as the loss function and the optimizer is Adam gradient [16]. A dropout procedure [17] with rate 0.5 is used for LSTM and full connection layer to avoid overfitting. Each convolutional layer has 16 local filters of size 5×7 and the number of nodes for LSTM and FCL is 300 and 1024, respectively. There are no subsampling layers after convolutional layers, in our experiments, inclusion of subsampling layers showed a slight decrease in performance. The weights and the bias of model are initialized randomly by uniform distribution.

2.3. Feature Extraction

Selecting a suitable feature is very important for learning algorithm. In this section, we will introduce our feature in details.

In time domain, the m-th microphone signal $y_m(k)$ with noise $v_m(k)$ is modeled as

$$y_m(k) = g_m * s(k) + v_m(k) \tag{1}$$

where $s(k)$ is pure signal, g_m is the channel impulse response from the source to microphone m $(m = 1, ..., M)$, where M is the number of microphones.

Every channel is transformed to frequency domain. In this research, a fixed 20-ms frame size is used with 50% overlap between frames. The discrete Fourier transform (DFT) is applied on every frame. The length of one frame is 320 and the number of frequency bins is 161, where sampling rate is 16kHz.

We calculate the GCC_{PHAT} between every two adjacent microphone y_m and y_{m+1} at location θ, as shown in Eq. (2)

$$GCC_{PHAT}(t, f, m, \theta) = \Psi_{y_m y_{m+1}}(t, f) e^{-j2\pi\tau_{m,\theta}} \tag{2}$$

where t is the time frame index, and f is the frequency bin index, $\Psi_{y_m y_{m+1}}$ is the generalized cross-spectrum defined as

$$\Psi_{y_m y_{m+1}}(t, f) = \vartheta(t, f) E\left[Y_m(t, f) Y^*_{m+1}(t, f)\right] \tag{3}$$

where

$$\vartheta(t, f) = \frac{1}{\sqrt{E\left[|Y_m(t, f)|^2\right] E\left[|Y_{m+1}(t, f)|^2\right]}} \tag{4}$$

is a weighting function to overcome the impact of the fluctuating levels of the speech source signal. E denotes expectation. Y_m is the spectrum at the m-th channel, Y_{m+1} indicates the first channel if $m = M$. The number of θ is 36 (from 0 degrees to 350 degrees with 10-degree resolution) and we generate $\tau_{m,\theta}$ of every location θ for every microphone pair.

Then, the input data can be extracted from GCC_{PHAT} by:

$$G(t, f, \theta) = \frac{1}{M} \sum_{m=1}^{M} GCC_{PHAT}(t, f, m, \theta) \tag{5}$$

Finally, we connect the $G(t, f, \theta)$ of every location θ in sequence as feature. Since the power spectrum peak contains information about DOA, we ignore the phase information and only use the power spectrum of the feature as input. So the size of one input frame is 36(directions) × 161(sub-bands) in this study. In [11], they calculate GCC_{PHAT} between every two channel, so the length of one input frame is $N \times C^2_M$(N is the dimension of GCC_{PHAT}).

3. EXPERIMENTS

3.1. Experimental Setup

The proposed method is evaluated on circular microphone arrays with various microphone numbers and array radius. The noisy and reverberant signals of microphone array are simulated as following steps. First, given microphone number and radius, we simulate a circular microphone array in a shoebox-shaped room with dimension 4m × 5m × 3m (length, width and height). The microphone array is placed at the center of the simulated room and its height is 1m. The target speaker is placed at one of the directions from $0°$ to $350°$ with $10°$ intervals, and the horizontal distance between the speaker and microphone array is fixed at 1.5m. The height of speakers is 1.7m. Next, we generate the room impulse responses (RIRs) using the Image method [18], and the source speeches are convolved with RIRs to form the reverberant microphone array input signals. The reverberation time (T_{60}) is 0.5s by setting the absorption coefficients. Finally, the additive noises are added to each microphone.

3.2. Dataset

For the training set, a 6-channel circular array with a radius of 0.035m is used. We randomly select 30 utterances from the RASC863 Mandarin Chinese database [19] for each source location. Each utterance is convolved with one RIR corresponding to its direction. So the number of source speeches is 1080 (30 utterances $\times$ 36 directions). Then uncorrelated white noise is added to each reverberant speech with SNR at 0 dB, 3 dB and 6 dB, respectively. The total number of training data is 3240 (1080 $\times$ 3 SNRs).

We also use the similar way to generate the test data which consists the matched condition, the radius-unmatched and the number-unmatched conditions. For the matched test set, we use the same array as the one in the training stage. For the radius-unmatched test set, the radius is 0.04m and 0.05m for 6-channel circular array, respectively. The number-unmatched test set is generated by using 4- and 5-microphone with radius of 0.035m. The simulated test data is generated by convolving 108 (36 directions$\times$1 utterances$\times$3 SNRs) utterances selected from RASC863 corpus.

Since we focus on online DOA estimation which predicts a direction for each frame, silence segments of target speech are labeled as silence which means no direction. Therefore, the size of the output target for each frame is 37 that corresponds to 36 directions and 1 no direction. We perform voice activity detection (VAD) [20] on clean speech to label the silence and voice part.

3.3. Evaluation Metrics

To evaluate the performance of DOA estimation, previous studies [11, 12] employed classification accuracy (the ratio of the number of correctly estimated frames to the total number of voice frames), which is rough for DOA estimation in some cases, however, a slightly inaccurate estimate of DOA is acceptable, such as beamforming. Here, we employ two measurements 1) voice decision error (VDE), which indicates the percentage of frames are misclassified in terms of voice and voiceless, 2) accuracy ratio (AR) [21] on the voice frames, i.e. DOA estimation is correct if the deviation of the estimated angle is with in $\pm 10^\circ$ of the truth angle. The VDE and AR are defined as follows:

$$VDE = \frac{N_{p\to n} + N_{n\to p}}{N},\ AR = \frac{N_{0.1}}{N_p} \tag{6}$$

where $N_{p\to n}$ and $N_{n\to p}$ indicate the number of frames misclassified as voice and voiceless, respectively. N represents the number of total frames in a sentence. $N_{0.1}$ denotes the number of frames with the DOA estimation deviation smaller than $\pm 10^\circ$ of the target. N_p is the number of all speech frames. Apparently, lower VDE and higher AR indicate better DOA estimation.

4. EVALUATION AND COMPARISON

4.1. Evaluation

The proposed deep-learning based approach is evaluated and compared to CNN-based classification method [12] for DOA estimation. CNN-based method directly uses phase component of the STFT coefficients of the received signals as input. And the target is 36 directions corresponding to the DOA classes. The model consists 3 convolutional layers, each containing 64 small filters of size 2$\times$2 and two full connection layers each has 512 units. The activation function of output layer is softmax and the others are ReLU.

Table 1 shows the results of two methods (CNN and proposed) in matched conditions. It can be seen that two methods have similar VDE (32.91% vs. 33.17% on average). However, the proposed method achieves much better results on AR (40.10% vs. 75.29%). This is mainly because we use LSTM to track the long-time dependence of the DOA information.

To evaluate the robustness of the proposed method to the topology of microphone array, we conduct two experiments, 1) fixing microphone number and changing radius; 2) changing microphone number. Table 2 shows the results in radius-unmatched conditions. It can be seen that when the radius of circular array changes to 0.04m, both methods have very similar results on both VDE and AR, compared with their results in training condition (radius is 0.035m). When the radius increases to 0.05m, VDE for both methods are similar to the matched condition and the proposed achieves even better results. Performance of both methods drops about

Table 1. DOA estimation performance in matched conditions

Radius (meters)	SNR (dB)	Type of Models			
		CNN		proposed	
		VDE	AR	VDE	AR
0.035	**0**	33.00	32.55	33.84	**67.06**
	3	32.87	38.73	33.74	**73.53**
	6	32.85	49.01	**31.93**	**85.29**
	Avg.	32.91	40.10	33.17	**75.29**

Table 2. DOA estimation performance in radius-unmatched conditions

Radius (meters)	SNR (dB)	Type of Models			
		CNN		proposed	
		VDE	AR	VDE	AR
0.04	**0**	33.02	31.77	34.52	**68.77**
	3	32.94	39.70	33.02	**75.24**
	6	32.90	46.03	**31.97**	**82.90**
	Avg.	32.95	39.17	33.17	**75.64**
0.05	**0**	33.09	28.15	**31.65**	**61.32**
	3	33.13	35.61	**30.70**	**73.08**
	6	33.10	41.42	**29.37**	**77.33**
	Avg.	33.11	35.06	**30.57**	**70.58**

Table 3. DOA estimation performance in number-unmatched conditions

Radius (meters)	SNR (dB)	Microphone Number			
		4		5	
		VDE	AR	VDE	AR
0.035	0	38.37	31.26	37.72	54.04
	3	39.70	46.68	35.36	69.56
	6	37.96	52.87	33.68	79.13
	Avg.	38.68	43.60	35.59	67.58

5% on AR. Apparently, the proposed method is still much better than the CNN method.

When the number of microphone changes, the CNN method cannot work because the size of the input feature doesn't match the one in training condition. But our method can still work by using the trained model. Table 3 shows the results of number-unmatched conditions where the new circular array has 4 and 5 microphones with 0.035m radius. It should be mentioned that in these cases, the distance between neighboring microphones is also changed. From Table 3, we can find that the performance of the system is getting worse as the number of microphones decreases. The next step is to do adaptation on the original model to observe whether the performance could be improved.

4.2. Adaptation

To do adaptation, we generate the noisy and reverberant speeches on new topologies of microphone array. Specifically, we generate only one utterance for each direction at each SNR. For a specific new microphone array, the total number of adaptation data is 108 (36 directions$\times$3 SNRs). During the adaptation, we only update the kernel weights of CNN and fix the others. The number of adaptation epoches is 4. After adaptation, the new model is evaluated on the test set.

Fig. 2 shows the VDE and AR on before and after adaptation of a 4-mic array with a 0.035m-radius. It can be seen that the performance gets better, i.e. VDE drops about 2.5% and AR increases about 7.6% on average. Fig. 3 shows the results on before and after adaptation of a 6-mic array with a 0.05m-radius. There is also a significant improvement after adaptation, i.e. VDE drops about 3.8% and AR increases about 10.2% on average. It means that the model can be efficiently adapted to a new microphone array.

Another question is that what the difference will be between the adaptive model and a newly trained model using a large amount of matched data. To answer this question, we use 3240 noisy and reverberant speeches which are generated on a 6-mic array with a 0.05m-radius to train a new model. Fig. 4 shows the performance comparison of the adaptive model and the newly trained model. We can find that the newly trained model achieves slightly better results than the adaptive model. By adaptation, the proposed model ensures the performance while greatly saving the time of retraining.

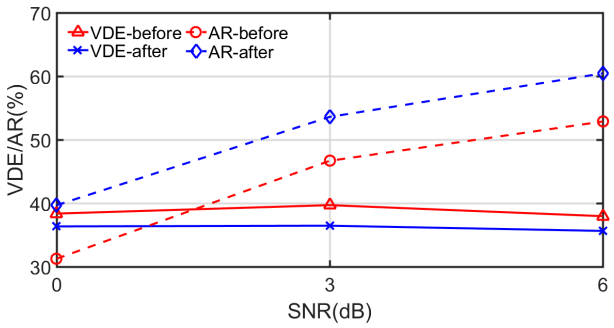

Fig. 2. The performance of 4-microphone array on before and after adaptation

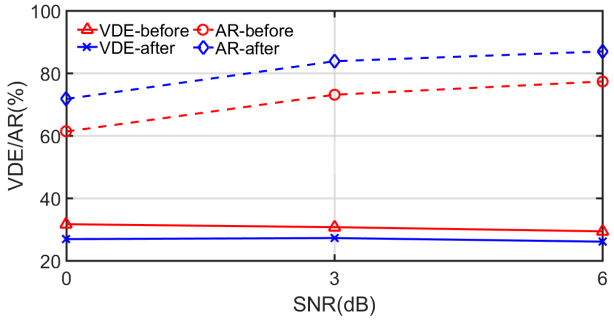

Fig. 3. The performance of 0.05m-radius array on before and after adaptation

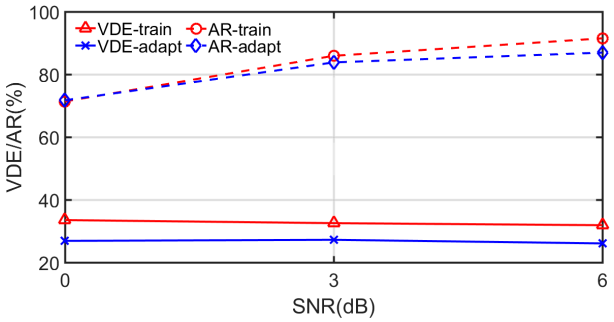

Fig. 4. The comparison of the adaptive model and the newly trained model

5. CONCLUSION

In this paper, we proposed a CNN-LSTM model to estimate DOA. Due to an input feature design, the proposed method can adapt to a new microphone array conveniently with a few data. In this research, the position of speakers is fixed, and future work involves the movements of position.

6. ACKNOWLEDGMENT

This research was supported by the China National Nature Science Foundation (No.61365006).

7. REFERENCES

[1] J.C. Chen, K. Yao, and R.E. Hudson, "Source localization and beamforming," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30–39, 2002.

[2] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, March 1986.

[3] Y.T. Huang, J. Benesty, G.W. Elko, and R.M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, November 2001.

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

[5] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1997, vol. 1, pp. 375–378.

[6] J. Benesty, J.D. Chen, and Y.T.Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Transactions on speech and audio processing*, vol. 12, no. 5, pp. 509–519, 2004.

[7] P. Stoica and K.C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 7, pp. 1132–1143, July 1990.

[8] J. Du, Y.H. Tu, Y. Xu, L.R. Dai, and C.H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. IEEE Intl. Conf. Signal Process.* IEEE, 2014, pp. 473–477.

[9] C. Weng, D. Yu, M.L. Seltzer, and J. Droppo, "Single-channel mixed speech recognition using deep neural networks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, vol. 18, pp. 5632–5636.

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

[11] X. Xiao, S.K. Zhao, X.H. Zhong, D.L. Jones, E.S. Chng, and H.Z. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[12] Chakrabarty S and Habets E A, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 1–3.

[13] V. Nair and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[14] R.A. Dunne and N.A. Campbell, "On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function," in *Proc. 8th Aust. Conf. on the Neural Networks, Melbourne, Australia*, 1997, pp. 181–185.

[15] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on information theory*, vol. 26, no. 1, pp. 26–37, 1980.

[16] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

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

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

[19] A.J. Li, Z.G. Yin, T.Q. Wang, Q. Fang, and F. Hu, "Rasc863-a chinese speech corpus with four regional accents," *ICSLT-o-COCOSDA, New Delhi, India*, 2004.

[20] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[21] B.S. Lee and P.W. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.