Microphone Array Speech Separation Algorithm based on DNN

Chaoyan Wu, Lin Zhou, Xijin Chen and Liyuan Chen
Southeast University, Nanjing, China
E-mail: wucy@seu.edu.cn; Linzhou@seu.edu.cn; chenxijin@seu.edu.cn; 450095871@qq.com

***Abstract*—Microphone array speech separation in high reverberant and low signal-to-noise ratio (SNR) environments, is still a challenge. This paper utilizes a deep neural network (DNN) and spatial information of a microphone array to train ideal binary masks and ideal ratio masks of target signals, then utilizes these masks to separate the target speech signals in a noisy and reverberant environment. In sound localization, SRP-PHAT is robust to noise and reverberation, while the SRP-PHAT computation is time-consuming in real application. To alleviate this problem, GSRP-PHAT which combines SRP-PHAT with Gammatone filter banks is proposed. GSRP-PHAT implying spatial information is applied as the input feature for the DNN. The simulation results show that the proposed algorithm achieves better performance in terms of source-to-distortion ratio, source-to-interference ratio, and short-time objective intelligibility in low signal to noise ratio (SNR) and high reverberant environments for omnidirectional speech separation. Also, the proposed method can still maintain the performance under the untrained conditions.**

I. INTRODUCTION

Speech separation is to separate target signal from the mixed signal with noise and reverberation, which is widely applied in various scenarios, such as intelligent assistant and hearing aids [1]. Multi-channel speech separation methods with spatial characteristics are superior to monaural methods [2]. Also, in speech separation, deep neural networks (DNNs) is utilized to estimate the spectrum masks or directly map to spectrum or waveform of the clean speech [3, 4].

Discriminative features are essential for the mask estimation and spectrum mapping in separation. However, traditional features, such as the logarithm power spectrum (LPS) and the Mel-frequency cepstral coefficient (MFCC), are lack of spatial information, which is insufficient for microphone array signals. Time difference of arrival (TDOA) can be conveniently inferred by a generalized cross-correlation (GCC) function in array signals processing [5]. Among various weight functions used to calculate GCC, phase transform (PHAT) achieves excellent results in noisy environments. Steered-response power phase transform (SRP-PHAT) is a classical approach that utilizes GCC weighted by PHAT to obtain a robust TDOA [6]. However, this method is time consuming and impractical in real applications. Some methods, such as LEMSalg [7] and GS [8], have been introduced to solve this problem. Unfortunately, the speech quality obtained using the aforementioned methods is inferior to SRP-PHAT. Besides, some improvements of SRP-PHAT have been proposed. SVD-PHAT is a real-time method for SRP-PHAT based on the fast-singular value decomposition, and the calculation of SVD-PHAT can be divide into two parts: off-line part calculating for the microphones array only, and online part calculating for received signals and executing search algorithm [9]. Hoang Do [10] proposed a real-time computation method for SRP-PHAT utilizing the stochastic region contraction.

Human's hearing system has the exceptional ability to extract interested sound source in chaotic environments. Inspired by this auditory property, computational auditory scene analysis (CASA) was proposed [11]. CASA simulates the perception process of human auditory system in two stages. The first stage is segmentation, dividing original speech signal into Time-Frequency (TF) units. And the second stage is grouping, aggregating the TF units which are from the same sound source.

Supervised learning is frequently applied in CASA for speech separation. Jiang [12] combines DNN with the binaural separation task, training Inter-aural time difference (ITD), Inter-aural level difference (ILD) and Gammatone-frequency cepstral coefficient features in each TF units to obtain the estimated ideal binary mask (IBM). The simulation result shows that DNN has generalization to spatial structure of source signals. Erdogan [13] associates masks with beamforming. The algorithm firstly estimates monaural masks for each microphone, then separates signals with minimum variance distortion less response—one of optimization criteria of adaptive beamforming.

Training target is an indispensable part of the network. Training targets can be speech's waveform [14-16], masks and parameters of filters or beam-formers [17-20]. Among these targets, Ideal ratio mask (IRM) and phase sensitive mask (PSM) are widely used. Though IRM is the essentially optimal mask and generally attains a higher source-to-distortion ratio (SDR) [21], IRM is inferior to PSM in terms of speech intelligibility, such as short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) [22]. PSM is a real number, while complex IRM (cIRM) has complex part which contains amplitude and phase information [23]. However, cIRM is hard to be exploited through common DNN [24].

Inspired by the recent progress, this paper proposes a microphone array speech separation algorithm based on DNN and improved SRP-PHAT. In our studies, the modified SRP-PHAT is proposed to reduce computational complexity. Also, these features contain the spatial information of SRP-PHAT and frequency information of Gammatone filter banks. Two

different masks, IBM and IRM are introduced as the DNN training target. Simulation is conducted in the noisy and reverberant environment. The results indicate that the proposed algorithm effectively separates target signal, even in the high reverberant environment and untrained conditions.

The remainder of this paper is organized as follows. Section II presents an overview of the proposed array speech separation system, including feature extraction and analysis. The structure of DNN is described in section III. Section IV presents simulation results and analysis. The conclusion is drawn in section V.

II. System Overview

This paper introduces a modified SRP-PHAT named GSRP-PHAT as the training feature of DNN to estimate the masks. GSRP-PHAT combines SRP-PHAT with Gammatone filter banks. Here in our study, IBM and IRM are selected as the training targets. The dimension of each training target is 37, including masks of the noise and the signals from 36 azimuths. The GSRP-PHAT is to estimate the IBM and IRM for each TF unit through DNN, then these masks are utilized to separate and reconstruct the target speech signals.

The structure of the proposed algorithm is shown in the Fig.1.

Assuming that the number of microphones and speakers are M and N respectively, the received signal of the mth microphone in a noisy and reverberant environment can be expressed as:

$$x_m(t) = \sum_{p=1}^{N} [a_{pm}\delta(t-\tau_{pm}) + h_{pm}(t)] * s_p(t) + v_m(t), m = 1,2,...,M \quad (1)$$

where terms a_{pm} and τ_{pm} denote attenuation coefficient and rectilinear propagation latency from the pth speaker to the mth microphone respectively. $h_{pm}(t)$ represents the impulse response of the reverberation. $s_p(t)$ is the source signal of the pth speaker. $v_m(t)$ represents the white noise received by the mth microphone and it is assumed to be uncorrelated with speech signals. Symbol '*' stands for linear convolution operation.

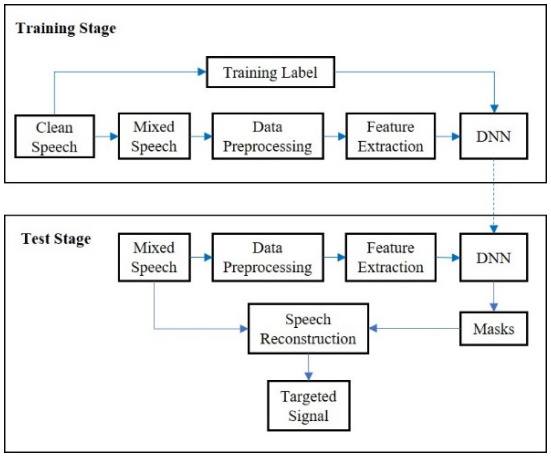

Fig. 1 The block diagram of the proposed algorithm

A. Feature Extraction

After preprocessing, the received signal from each microphone has been separated into frames. The traditional SRP-PHAT feature of each TF unit is formulated as:

$$SRP\text{-}PHAT_{k,l}(\theta) = 2\pi\sum_{u=1}^{N}\sum_{v=1}^{N}\int_{-\infty}^{+\infty}\frac{X_u^l(k,\omega)X_v^l(k,\omega)^*}{|X_u^l(k,\omega)X_v^l(k,\omega)^*|}e^{j\omega\tau(\theta,u,v)}d\omega \quad (2)$$

where k and l denote the indexes of the kth frame and the lth subband respectively. $X_u^l(k,\omega)$ and $X_v^l(k,\omega)$ denote the spectrum of temporal signals through the lth Gammatone filter. Superscript '*' stands for conjugation operation. $\tau(\theta,u,v)$ represents the rectilinear propagation latency between the uth and vth microphone given the azimuth of the sound source θ, which can be expressed as (3) if the microphone array is circular and the radius of the array is R.

$$\tau(\theta,u,v) = \frac{R\cos(\varphi_u - \theta) - R\cos(\varphi_v - \theta)}{c} \quad (3)$$

where φ_u and φ_v represent the azimuths of the microphones relative to the center of the array respectively. Term c represents acoustic velocity.

According to (2), amplitude suppression introduced by Gammatone filter is removed because of PHAT [25]. This problem severely degrades the performance of Gammatone filter banks. To alleviate this problem, this paper proposed GSRP-PHAT which combines SRP-PHAT with Gammatone filter banks.

The GSRP-PHAT feature contains both frequency information of Gammatone filter banks and phase information of SRP-PHAT, which are be formulated as follows:

$$GSRP\text{-}PHAT_{k,l}(\theta) = 2\pi\sum_{u=1}^{N}\sum_{v=1}^{N}\int_{\omega_1^l}^{\omega_2^l}\frac{X_u(k,\omega)X_v(k,\omega)^*}{|X_u(k,\omega)X_v(k,\omega)^*|}e^{j\omega\tau(\theta,u,v)}d\omega \quad (4)$$

where the frequency range of the lth Gammatone filter is from ω_1^l to ω_2^l, the corresponding amplitude range is [-20, 0]dB. $X_u(k,\omega)$ is the spectrum of kth frame.

The advantages of the GSRP-PHAT are described as follows:

1. Speech signals of all microphones do not need to be filtered by Gammatone filter, according to (4). Assuming the length of signal at each channel is L, the length of each filter is F, and the number of Gammatone filter is G. Then, the reduction of computation are $MLFG$ multiplications and $M\times G\times max(F,L)$ additions.

2. The frequency range of SRP-PHAT in (2) is larger than that of GSRP-PHAT in (4), which reduces the computation of GSRP-PHAT extraction.

In this paper, each frame is divided into 32 subbands corresponding 32 GSRP-PHAT values, and GSRP-PHAT has 360 dimensions.

B. Analysis of GSRP-PHAT

GSRP-PHAT are analyzed in certain acoustic environment to verify the effectiveness. A planar and uniform circular array consisting of six omnidirectional microphones is selected. Clean speech of a male and a female, are randomly chosen from TIMIT corpus, and the speech sources are located on the azimuth of 60 and 120 respectively. Also, Gaussian white noise is added to each microphone and SNR is 20dB. GSRP-PHATs of various subbands are extracted for each frame, which are displayed in Fig. 2.

In Fig. 2, brightness represents the amplitude. Abscissa is the frames index, and ordinate represents azimuth. It can be seen from the Fig. 2 that the highest value appears around degree of 60 and 120 in most frames. Moreover, GSRP-PHAT changes in different subbands, which indicates that in the same TF unit, there is usually only one sound source. To testify this phenomenon, IRM is calculated for each TF unit shown in Table 1. From the Table 1, GSRP-PHAT conforms to IRM label, demonstrating that GSRP-PHAT is able to distinguish energy distribution at various azimuths and TF units.

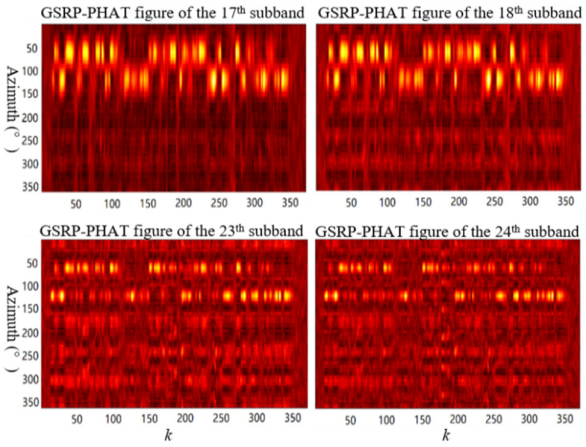

Fig. 2 GSRP-PHATs of different subbands

Table 1 IRM label of the simulated signal

Subband	60°	120°	Noise
17	0.6539	0.2619	0.0841
18	0.7255	0.1650	0.1095
23	0.0098	0.9843	0.0059
24	0.0221	0.9730	0.0049

III. DNN Based Speech Separation

A. Training Targets

In this paper, IBM and IRM are applied as training targets respectively, and the corresponding algorithms are called DNN-IBM and DNN-IRM. In our algorithm, intervals of azimuths is set to 10, and ***IBM*** is a vector with 37 dimensions, representing masks on noise and signals from 36 azimuths. IBM vector for target source can be described as ***IBM***=(IBM_0, IBM_1,…, IBM_{36}), where IBM_0 indicates the mask on the noise and is defined as:

$$IBM_0(k,l)=\begin{cases}1, & SNR(k,l)<0\\ 0, & otherwise\end{cases} \tag{5}$$

where (k,l) denotes the TF unit of the kth frame and the lth subband.

The remainder scalars of IBM vector indicate masks on the assumed sound source at nth azimuth and can be formulate as:

$$IBM_n(k,l)=\begin{cases}1, & s_n^2(k,l)\ge \max(s_i^2(k,l),v^2(k,l)),i=1,...,36\\ 0, & otherwise\end{cases} \tag{6}$$

where $s_n^2(k,l)$ and $s_i^2(k,l)$ are the energy of the assumed sound source at nth azimuth and the rest azimuths respectively. $v^2(k,l)$ is the noise energy of TF unit. The loss function of IBM-DNN algorithm is

$$J_{IBM}=-\sum_{n=0}^{36} IBM_n \ln(IBM_n') \tag{7}$$

where IBM_n is the ideal value, and IBM_n' is the output of the network.

Similar to IBM, IRM of each TF unit is a 37-dimentional vector which can be calculated as follows:

$$IRM_n(k,l)=\begin{cases}\left[\dfrac{s_n^2(k,l)}{\sum_{i=1}^{36}s_i^2(k,l)+v^2(k,l)}\right]^{1/2}, & n=1,2,...,36\\ \left[\dfrac{v^2(k,l)}{\sum_{i=1}^{36}s_i^2(k,l)+v^2(k,l)}\right]^{1/2}, & n=0\end{cases} \tag{8}$$

The loss function of IRM-DNN algorithm is

$$J_{IRM}=\frac{1}{2}\left\|\boldsymbol{IRM}-\boldsymbol{IRM}'\right\|^2 \tag{9}$$

where ***IRM***=(IRM_0, IRM_1,…,IRM_{36}) is the ideal value, and ***IRM***' is the output of the network.

B. Speech Reconstruction

Since the speech signals have time correlation, the moving average filter is applied to masks, which is expressed as follows:

$$\overline{\boldsymbol{M}}(k,l)=\frac{\sum_{k'=-d}^{d}\boldsymbol{M}(k+k',l)}{2d+1} \tag{10}$$

where $\boldsymbol{M}(k,l)$ stands for ***IBM***(k,l) or ***IRM***(k,l). d denotes the number of moving frames. In this paper, d is set to 2.

Target signal is reconstructed by the mixed signal and the masks as follows:

$$s_n(k,l)=\overline{M}_n(k,l)x(k,l) \tag{11}$$

where $x(k,l)$ denotes the mixed signal, $\overline{M}_n(k,l)$ is the mask at nth azimuth.

C. *Architecture of DNN*

The architecture of DNN is shown in Fig. 3. Five hidden layers with 512 neurons are utilized. Batch Normalization (BN) is utilized and the LReLU is applied as the activation function.

D. *Network Training*

Kaiming initialization and adaptive moment estimation optimizer are utilized to train the network. Initial learning rate is set to 0.001. Validation set is employed to avoid over fitting. During the training, the loss function of the validation set is calculated at the end of each iteration. And if it does not decline for the first time, the learning rate will be reduced to 1/10 of the original. And the training stage will stop when the loss function of validation set doesn't decline for the second time.

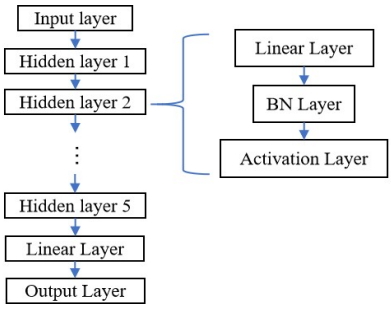

Fig. 3 Architecture of DNN

IV. Simulation and Result Analysis

A. *Simulation Setup*

The microphone array is comprised of six omnidirectional microphones uniformly arranged along a circle as shown as in Fig. 4. Radius of the array is 10 cm. The array horizontally locates in the center of a cube-shape chamber with the size of 7m×6m×3m.

Clean speech signals are randomly selected from TIMIT corpus. The room impulse response is generated by Image method [26], the clean speech signal convolves with room impulse response to generate reverberant environment. Gaussian white noise is added and is uncorrelated with signals.

There are 9 acoustic environments simulated in both training and testing stage, including reverberation time RT60 (0, 200ms and 600ms) and SNRs (0 10dB and 20dB). For the testing, besides above-mentioned conditions, other SNRs (3dB, 5dB, 7dB 9dB and 15dB) with higher RT60 (800ms) are included to investigate the generalization. Here, T0, T200 and T600 mean that the reverberation time RT60 are 0ms, 200ms and 600ms respectively.

Fig. 4 Arrangement of microphone array

SDR, source-to-interference ratio (SIR) and STOI are treated as the performance measures. SDR estimates general distortion of the signal. SIR assesses the effects of interference signals on the target signal. STOI, ranging between 0 and 1, quantifies the intelligibility of the speech.

B. *Evaluation and Analysis*

First, we evaluated the performance of the algorithm in the same acoustic environments as the training stage. The metrics for different algorithms is shown in Fig. 5, Fig. 6 and Fig. 7.

According to Fig. 5, in high SNR and low reverberation conditions, DNN-IBM has a slight advantage over SDR. However, with the increase of reverberation, DNN-IRM performance are comparable with DNN-IBM. According to Fig. 6, DNN-IBM exhibits a poor performance on SIR, because the energy in each TF unit is from different sound source signals. When the locations of different sound source signals are closer, IBM labels will distribute interference signals' energy to the target signal. According to Fig. 7, in

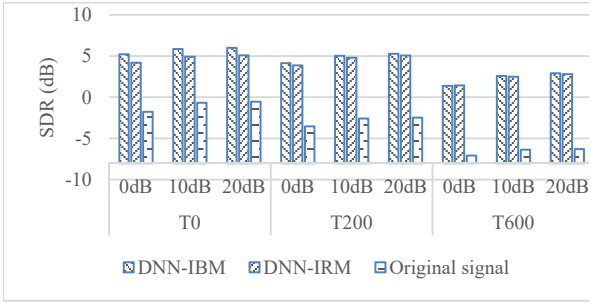

Fig. 5 SDR comparison of different algorithms

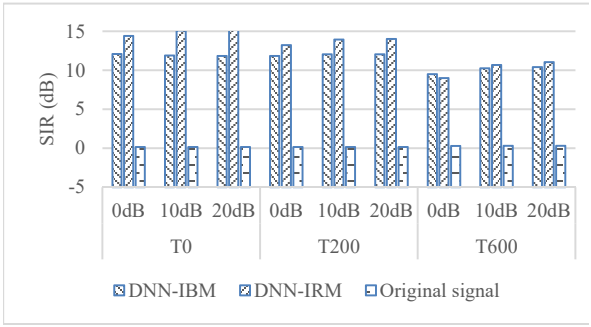

Fig. 6 SIR comparison of different algorithms

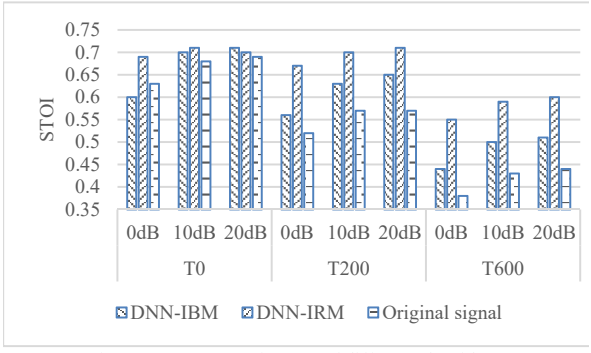

Fig. 7 STOI comparison on of different algorithms

environments with no reverberation and high SNR, DNN-IRM is slightly inferior to DNN-IBM. However, with the increase of reverberation and noise, DNN-IRM has better STOI than DNN-IBM.

According to Fig. 5, Fig. 6 and Fig. 7, in different reverberation environments, noise has weak effect on above-mentioned three metrics, indicating that the proposed algorithm is robust to the noise. The reasons are described as follows:

1. Application of Gammatone filter in CASA brings more robust features to distinguish noise and speech in TF unites.
2. IBM/IRM has addition noise label to suppress noise.

The performance of the DNN-IRM is further evaluated in the environments with SNRs are different from those of the training stage. The reverberation time is 800ms. The results are shown in Table 2.

From the Table 2, these three performance measures, SIR, SDR and STOI, change smoothly. Although the SNR and reverberation time in the testing stage differ from that of the training stage, there is no obvious performance degradation, which means that even if the testing acoustic environment does not match the training acoustic environment, the proposed algorithm can still achieve good separation result, and maintain stable speech intelligibility.

Also, we compare the IRM-DNN performance in matched environment and unmatched environment, which are shown in Fig. 8, Fig. 9 and Fig. 10.

Table 2 The performance of IRM-DNN in 800ms reverberation environments

SNR(dB)	SDR(dB)	SIR(dB)	STOI
0	0.495	7.316	0.52
3	0.921	8.170	0.53
5	1.168	8.650	0.54
7	1.378	9.009	0.55
9	1.549	9.302	0.55
10	1.635	9.429	0.55
15	1.878	9.839	0.56
20	1.951	9.967	0.56

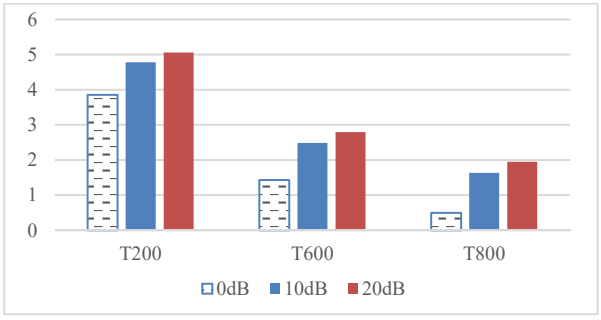

Fig. 8 SDR of DNN-IRM in 200ms, 600ms and 800ms reverberation and 0dB, 10dB and 20dB SNR environments

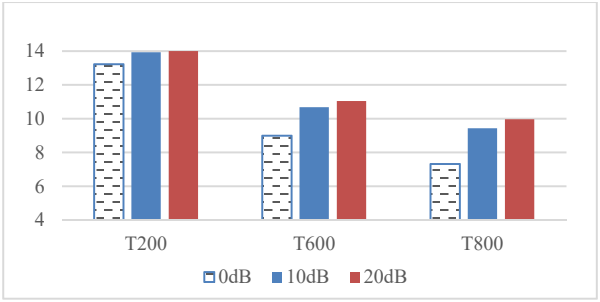

Fig. 9 SIR of DNN-IRM in 200ms, 600ms and 800ms reverberation and 0dB, 10dB and 20dB SNR environments

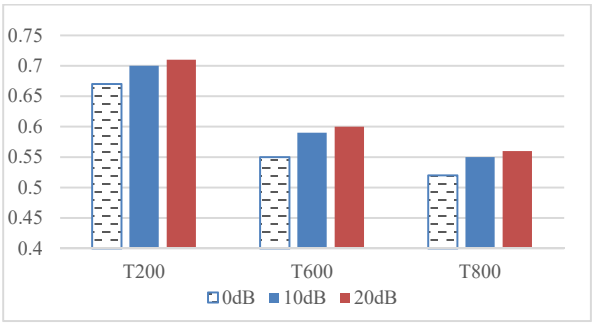

Fig. 10 STOI of DNN-IRM in 200ms, 600ms and 800ms reverberation and 0dB, 10dB and 20dB SNR environments

According to Fig. 8, Fig. 9 and Fig. 10, the performance degradation trend is reasonable from T200 to T800, which indicates that DNN-IRM can still maintain the good performance under the unmatched environments. The results demonstrate the generalization of the proposed algorithm.

V. Conclusion

This paper proposes a microphone array speech separation based on DNN and GSRP-PHAT. The GSRP-PHAT is the modified version of SRP-PHAT, which combines spatial information of SRP-PHAT and frequency information of Gammatone filter. The utilization of GSRP-PHAT obviously reduce the computational amount. The stimulation results indicate that the proposed algorithm based on DNN and GSRP-PHAT can achieves high separate performance and generalization in noisy and reverberant environments.

References

[1] J. Park and S. Kim, “Noise cancellation based on voice activity detection using spectral variation for speech recognition in smart home devices,” Intelligent Automation & Soft Computing, vol. 26, no.1, pp. 149–159, 2020.

[2] L. Pfeifenberger, T. Schrank, M. Zohrer, M. Hagmüller and F. Pernkopf, “Multi-channel speech processing architectures for noise robust speech recognition: 3rd chime challenge results,” Proc. IEEE ASRU, Scottsdale, AZ, USA, pp. 452–459, 2015.

[3] Y. Wang and D. Wang, “Boosting classification based speech separation using temporal dynamics”, INTERSPEECH 2012, Portland, OR, USA, pp.1528–1531, 2012.

[4] L. Zhou, S. Lu, Q. Zhong, Y. Chen, Y. Tang et al., “Binaural speech separation algorithm based on long and short time

memory networks," Computers, Materials & Continua, vol. 63, no. 3, pp. 1373–1386, 2020.
[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 4, pp. 320-327, August 1976, doi: 10.1109/TASSP.1976.1162830.
[6] X. Zhao, S. Chen, L. Zhou and Y. Chen, "Sound source localization based on srp-phat spatial spectrum and deep neural network," Computers, Materials & Continua, vol. 64, no. 1, pp. 253–271, 2020.
[7] H. F. Silverman, Ying Yu, J. M. Sachar and W. R. Patterson, "Performance of real-time source-location estimators for a large-aperture microphone array," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 4, pp. 593–606, 2005.
[8] M. D. Gillette and H. F. Silverman, "A linear closed-form algorithm for source localization from time-differences of arrival," IEEE Signal Processing Letters, vol. 15, pp. 1–4, 2008.
[9] F. Grondin and J. Glass, "SVD-PHAT: A Fast Sound Source Localization Method," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 4140-4144, doi: 10.1109/ICASSP.2019.8683253.
[10] H. Do, H. F. Silverman and Y. Yu, "A Real-Time SRP-PHAT Source Location Implementation using Stochastic Region Contraction(SRC) on a Large-Aperture Microphone Array," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI, 2007, pp. I-121-I-124, doi: 10.1109/ICASSP.2007.366631.
[11] Wang D L, Brown G J. Computational auditory scene analysis: Principles, algorithms, and applications [J]. IEEE Transactions on Neural Networks, 2008, 19(1): 199-199.
[12] Jiang Y, Wang D L, Liu R S, et al. Binaural classification for reverberant speech segregation using deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 2112-2121.
[13] Erdogan H, Hershey J R, Watanabe S, et al. Improved mvdr beamforming using single-channel mask prediction networks[C]. Interspeech. 2016: 1981-1985.
[14] R. Gu et al., "Enhancing End-to-End Multi-Channel Speech Separation Via Spatial Feature Learning," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7319-7323, doi: 10.1109/ICASSP40776.2020.9053092.
[15] S. Maiti and M. I. Mandel, "Speech Denoising by Parametric Resynthesis," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 6995-6999, doi: 10.1109/ICASSP.2019.8683130.
[16] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256-1266, Aug. 2019, doi: 10.1109/TASLP.2019.2915167.
[17] J. Heymann, L. Drude and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 196-200, doi: 10.1109/ICASSP.2016.7471664.
[18] X. Xiao et al., "Deep beamforming networks for multi-channel speech recognition," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 5745-5749, doi: 10.1109/ICASSP.2016.7472778.
[19] M. Pariente, S. Cornell, A. Deleforge and E. Vincent, "Filterbank Design for End-to-end Speech Separation," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6364-6368, doi: 10.1109/ICASSP40776.2020.9053038.
[20] T. Higuchi, N. Ito, T. Yoshioka and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 5210-5214, doi: 10.1109/ICASSP.2016.7472671.
[21] J. L. Roux, S. Wisdom, H. Erdogan and J. R. Hershey, "SDR – half-baked or well done?" IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, pp. 626–630, 2019.
[22] Z. Wang, X. Wang, X. Li, Q. Fu and Y. Yan, "Oracle performance investigation of the ideal masks," IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), Xi'an, China, pp. 1–5, 2016.
[23] D. Wang and C. Bao, "An Ideal Wiener Filter Correction-based cIRM Speech Enhancement Method Using Deep Neural Networks with Skip Connections," 2018 14th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 2018, pp. 270-275, doi: 10.1109/ICSP.2018.8652281.Cai W, Wang S, Wu Z. Accelerated steered response power method for sound source localization using orthogonal linear array[J]. Applied Acoustics, 2010, 71(2): 134-139.
[24] D. Yin, C. Luo, Z. Xiong and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," Advancement of Artificial Intelligence (AAAI), New York, USA, pp. 9458–9465, 2020.
[25] M. Pariente, S. Cornell, A. Deleforge and E. Vincent, "Filterbank Design for End-to-end Speech Separation," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 6364–6368, 2020.
[26] J. B. Alien 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, April, 1979.