Locate and Beamform: Two-dimensional Locating All-neural Beamformer for Multi-channel Speech Separation

Yanjie Fu[1], Meng Ge[2,*], Honglong Wang[1], Nan Li[1], Haoran Yin[1],
Longbiao Wang[1,*], Gaoyan Zhang[1], Jianwu Dang[1], Chengyun Deng[3], Fei Wang[3]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]Beijing Xiaoju Technology Co., Ltd., Beijing, China

`fuyanjie@tju.edu.cn, gemeng@tju.edu.cn, longbiao_wang@tju.edu.cn`

Abstract

Recently, stunning improvements on multi-channel speech separation have been achieved by neural beamformers when direction information is available. However, most of them neglect to utilize speaker's 2-dimensional (2D) location cues contained in mixture signal, which limits the performance when two sources come from close directions. In this paper, we propose an end-to-end beamforming network for 2D location guided speech separation merely given mixture signal. It first estimates discriminable direction and 2D location cues, which imply directions the sources come from in multi views of microphones and their 2D coordinates. These cues are then integrated into location-aware neural beamformer, thus allowing accurate reconstruction of two sources' speech signals. Experiments show that our proposed model not only achieves a comprehensive decent improvement compared to baseline systems, but avoids inferior performance on spatial overlapping cases.

Index Terms: Speech separation, neural beamforming, 2D sound source localization, spatial overlapping issue

1. Introduction

Speech separation aims to extract all individual speech signals from the observed mixture speech, which is a high-demand front-end technology for various real-world applications, such as speech recognition [1, 2] and speaker verification [3, 4].

For quite some time, multi-channel speech separation solutions have attracted much research attention due to the benefit of spatial information in the microphone array setup. For example, deep learning based beamformers [5, 6, 7] are well studied to estimate speech and noise covariance matrices and beamforming weights in an end-to-end fashion. The spatial discrimination facilitates separation when the azimuth difference between speakers is large. Nevertheless, most multi-channel approaches fail on separation when two sources' directions are less than $15°$ away from each other. This is the so-called *spatial overlapping* or *spatial ambiguity* issue [8].

To retain practicality on small azimuth difference cases, some multi-channel separation systems introduce extra speaker cue [9, 10] from an enrolled reference utterance, visual cue [11] from face tracking and lip movements, or direction cue [5, 6, 7] from an additional visual detection system. Such systems require wide-angle camera, speaker embedding model or lip reading network, which might be more resource-consuming.

In blind source separation applications, above mentioning cues are not available and only observed mixture signal can be leveraged to process. Previous works [12, 13, 14] make efforts to estimate direction-of-arrival (DOA) for each source to guide the estimation of individual sources' speech signals. They rely

on the assumption that all microphones can be approximated to have the same DOA when distance between source and microphone array is large. However, this assumption does not hold when source locates in near-field of a large-spacing microphone array. Moreover, obtaining only one DOA is not sufficient for 2D localization. Thus we offer a 2-observer solution rather than using the centroid of array as reference point, which means determining source's two DOAs at two outermost microphones of a linear array. This enables us to obtain source's (x, y) coordinates on a 2D horizontal plane through triangulation and leverage these 2D location cues to provide extra discrimination, thus can better handle small azimuth difference cases.

In this work, to explicitly guide separation task with 2D location cues and relieve performance decline problem under spatial overlapping conditions, we propose **L**ocate **a**nd **B**eamform Network (**LaBNet**), a 2D location-aware neural beamforming framework. Specifically, we design 2D Locator, aiming at estimating source's two DOAs at two outermost microphones as well as corresponding direction embeddings. 2D coordinates of each source are then calculated with the estimated 2 DOAs. In this way, LaBNet can explicitly utilize spectral, spatial, direction and location features with 2D Locator and location-aware neural beamformer to prompt the difference between speakers from all perspectives. Experimental results show that both overall speech separation performance and the lower bound on spatial overlapping cases are substantially improved compared to existing state-of-the-art methods.

2. Generalized RNN-beamformer Baseline

Consider the input mixture waveform $\mathbf{y} \in \mathbb{R}^{L \times M}$, where L is denoted as the number of audio samples and M as the number of audio channels, the corresponding time-frequency representations after STFT can be formulated as

$$Y(t, f) = S(t, f) + I(t, f) \tag{1}$$

where S stands for the target speech and I represents the sum of interfering speakers' speech and noise (if any).

GRNN-BF proposed in [7] contains two modules, complex ratio filter (cRF) [15] estimator and RNN beamformer. The cRFs are predicted by cRF estimator and then used to calculate the target speech and interference covariance matrices. The RNN beamformer accepts the target speech and interference covariance matrices and generates frame-level beamforming weights for the target speaker.

For each T-F (time-frequency) unit of mixture, the estimated cRF applied to its neighbouring units following

$$\hat{\mathbf{S}}_{\text{cRF}_i}(t, f) = \sum_{\tau_1=-K}^{\tau_1=K} \sum_{\tau_2=-K}^{\tau_2=K} \text{cRF}_{\mathbf{S}_i}(t, f, \tau_1, \tau_2) \\ *\mathbf{Y}(t + \tau_1, f + \tau_2) \tag{2}$$
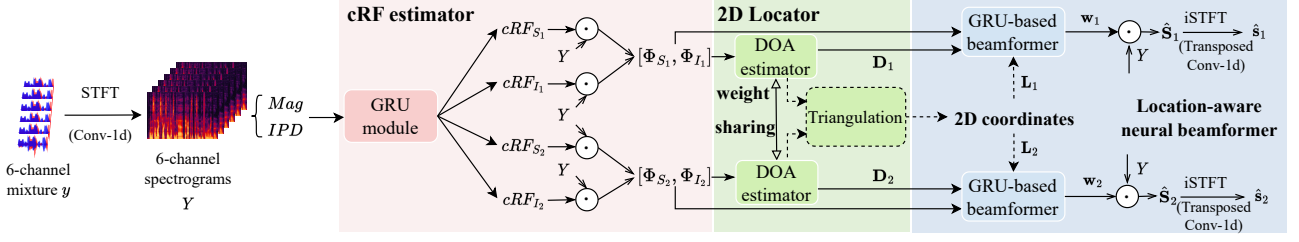

Figure 1: *Architecture of our proposed LaBNet. cRF estimator (red part) produces cRFs to obtain speech and interference covariance matrices for both speakers as input of following modules. 2D Locator (green part) is introduced for explicit discrimination guidance to neural beamformer with predicted direction and location embeddings. Location-aware neural beamformer (blue part) predicts beamforming weights for each source. All components in our framework are trainable, which allows backpropagation all around.*

where $\hat{\mathbf{S}}_{\text{cRF}_i}(t, f)$ indicates the i-th speaker's speech estimated by $\text{cRF}_{\mathbf{S}_i}$, and K defines the number of neighboring T-F bins. Likewise, the i-th speaker's interference noise $\hat{\mathbf{I}}_{\text{cRF}_i}(t, f)$ can be obtained by $\text{cRF}_{\mathbf{I}_i}$ in a similar manner. Then, the frame-wise i-th speaker's speech covariance matrix can be calculated with the estimated target speech and its conjugate transpose by a layer normalization [16] shown in Eq. (3). Another layer normalization is applied for computing speaker i's corresponding interference noise covariance matrix $\boldsymbol{\Phi}_{\mathbf{I}_i}(t, f)$.

$$\boldsymbol{\Phi}_{\mathbf{S}_i}(t, f) = \text{LayerNorm}(\hat{\mathbf{S}}_{\text{cRF}_i}(t, f)\hat{\mathbf{S}}_{\text{cRF}_i}^{\text{H}}(t, f)) \quad (3)$$

GRNN-BF believes that a better beamformer solution can be directly learned from the speech and noise covariance matrices by neural network. Thus only one unified RNN model is applied to predict the frame-level beamforming weights directly from the covariance matrices, which can be formulated as

$$\mathbf{w}_i(t, f) = \mathbf{GRU}\left([\boldsymbol{\Phi}_{\mathbf{S}_i}(t, f), \boldsymbol{\Phi}_{\mathbf{I}_i}(t, f)]\right) \quad (4)$$

Here, $[\cdot, \cdot]$ denotes the concatenation operation of the i-th speaker's speech and interference noise covariance matrices. Next, the STFT representation of mixture $Y(t, f)$ is beamformed with the estimated beamforming weights $\mathbf{w}_i(t, f)$ to obtain the estimate of the i-th target speaker's speech spectrogram $\hat{\mathbf{S}}_{\text{GRNN-BF}_i}(t, f)$, which can be formulated as

$$\hat{\mathbf{S}}_{\text{GRNN-BF}_i}(t, f) = \mathbf{w}_i^{\text{H}}(t, f)\mathbf{Y}(t, f) \quad (5)$$

Finally, the separated time-domain waveform of the i-th speaker's speech $\hat{s}_i$ can be converted from the beamformed spectrum $\hat{\mathbf{S}}_{\text{GRNN-BF}_i}(t, f)$ using iSTFT. All of the covariance matrices and beamforming weights are complex-valued. The real and imaginary parts of these complex-valued tensors are concatenated throughout the network.

3. LaBNet Architecture

GRNN-BF [7] requires prior known DOA of target speaker, which is calculated according to the location of target speaker in the video view captured by a wide-angle camera, thus does not suit applications where only observed mixture signal is available. Meanwhile, the DOA is roughly estimated and its DOA estimator cannot be jointly optimized with neural beamformer, which limits the upper bound performance of the system.

In this section, we introduce LaBNet to utilize discriminative 2D direction and location information merely from the given mixture signal (i.e., there is no requirement for providing any extra auxiliary information in our system), thus alleviating the performance degradation caused by spatial overlapping issue [8] without sacrificing the overall separation performance.

3.1. Overview of LaBNet

Fig.1 shows the detailed architecture of our proposed LaBNet, which consists of cRF estimator, 2D locator and location-aware neural beamformer. Specifically, the magnitude of reference channel and interaural phase difference (IPD) features of five microphone pairs are extracted from multi-channel mixture STFT. Unlike GRNN-BF [7], the directional feature captured by camera is not available here and we pass magnitude rather than log power spectrum (LPS) along with IPD to cRF estimator. We replace dialated Conv-1D blocks with a GRU module in cRF estimator due to consideration of saving GPU memory usage. After obtaining cRFs for each speaker's speech and interference, multi-channel target speech and corresponding interference can be computed with Eq. (2). Then, speech and interference covariance matrices are calculated with two layer normalization [16] shown in Eq. (3). The next step is to predict beamforming weights from the covariance matrices. This work proposes to introduce a tiny 2D locator (0.9 M) to offer direction and location discrimination for GRU-based beamformer, even on closely located cases, to learn more accurate beamforming weights and exhibit more effective speech separation capability.

3.2. 2D Locator

DOA estimator. For each source, the DOA estimator accepts its speech and interference covariance matrices and output the frame-level spatial spectrum(s) [17] related to its DOA(s). The likelihood-based spatial spectrum is based on the assumption that the probability of source being at each individual angle follows Gaussian distribution that maximizes at ground truth DOA. Formulaically, we adopt a 210-dimensional vector $p_i^{(n)}(\theta)$ for each time frame to include the probabilities of one source locating at 210 individual directions (the azimuth θ ranges from $-15°$ to $195°$):

$$p_i^{(n)}(\theta) = e^{-d(\theta, \theta')^2/\sigma^2} \quad (6)$$

where θ' is the ground truth DOA of one source, σ is a predefined constant that controls the width of the Gaussian function and $d(\cdot, \cdot)$ denotes the angular distance. Note that $n \in \{1, ..., N\}$ and when N is set to 1, only one DOA at the centroid of array is estimated; when $N = 2$, 2 DOAs w.r.t. 2 outermost microphones will be estimated. The DOA estimator is composed of two convolutional layers and a GRU module. The first convolutional layer projects covariance matrices into the DOA space to obtain $210 \times N$-dimensional direction embedding $\mathbf{D}_i$ for each T-F unit. The second one convolves along time and DOA axes and aggregates features across all frequency bins to generate initial spatial spectrums for each time
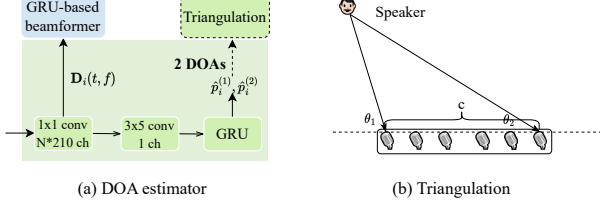
(a) DOA estimator (b) Triangulation

Figure 2: *The components of DOA estimator and the diagram of triangulation. Dashed line means optional operation.*

frame. The following GRU module learns temporal context information among all time frames to polish the frame-level spatial spectrum before outputing it. We select mean squared error to measure the disparity between the estimated and the ideal frame-level spatial spectrum for the i-th speaker $\hat{p}_i^{(n)}$ and $p_i^{(n)}$ at n-th observer:

$$\mathcal{L}_{\text{DOA}_i} = \sum_{n=1}^{N} ||\hat{p}_i^{(n)} - p_i^{(n)}||^2 \qquad (7)$$

Triangulation. Given prior known interval between 2 outermost microphones c and the estimated 2 DOAs regarding them θ_1, θ_2, the distances between source and outside microphones can be given by Law of Sines (Fig.2 (b)). We then compute 2D coordinates (x, y) which represent source's relative location to the reference microphone.

3.3. Location-aware neural beamformer

Two additional inputs are taken by our proposed location-aware neural beamformer besides speech and noise covariance matrices compared to GRNN-BF: 1) DOA estimator's intermediate output, the direction embedding $\mathbf{D}_i$, that indicates from which azimuth direction the speech comes in each T-F bin. 2) the location embedding $\mathbf{L}_i$, which is the repeats of source's frame-level 2D coordinates $(x, y)_i$ along the frequency dimension, indicating each source's location in each T-F unit.

The estimated direction and location embeddings for each speaker, $\mathbf{D}_i(t, f)$ and $\mathbf{L}_i(t, f)$ are concatenated with speech and interference covariance matrices of corresponding speaker along feature dimension and respectively fed into two parallel beamformers to generate the beamforming weights for each speaker following Eq. (8) and thus can more accurately separate the two speaker's speech.

$$\mathbf{w}_i(t, f) = \mathbf{GRU}\left([\boldsymbol{\Phi}_{\mathbf{S}_i}(t, f), \boldsymbol{\Phi}_{\mathbf{I}_i}(t, f), \mathbf{D}_i(t, f), \mathbf{L}_i(t, f)]\right) \qquad (8)$$

In this way, the 2D Locator explicitly performs an auxiliary task for blind source separation. The introduced direction and location embeddings just compensate the less discriminative spatial information in covariance matrices when sources come from adjacent directions, thus enabling the neural beamformer to regain discrimination even on closely located cases.

3.4. End-to-end Training

Given the mixture $\mathbf{y}$, target speech $\mathbf{s}_i$, estimated target speech $\hat{\mathbf{s}}_i$, weighted source-to-distortion ratio (wSDR) [18] can be computed as follows: $\mathcal{L}_{\text{wSDR}_i}(\mathbf{y}, \mathbf{s}_i, \hat{\mathbf{s}}_i) = \gamma \mathcal{L}_{\text{SDR}}(\mathbf{s}_i, \hat{\mathbf{s}}_i) + (1 - \gamma)\mathcal{L}_{\text{SDR}}(\mathbf{y} - \mathbf{s}_i, \mathbf{y} - \hat{\mathbf{s}}_i)$, where $\gamma = ||\mathbf{s}_i||^2/(||\mathbf{s}_i||^2 + ||\mathbf{y} - \mathbf{s}_i||^2)$, $\mathcal{L}_{\text{SDR}}(\mathbf{x}, \hat{\mathbf{x}}) = -\langle \mathbf{x}, \hat{\mathbf{x}} \rangle/(||\mathbf{x}|| ||\hat{\mathbf{x}}||)$. The wSDR losses of both speakers are added up as speech separation loss. For optimizing the whole system, we use a multi-task loss as training objective:

$$\mathcal{L}_{\text{LaBNet}} = \alpha * \sum_{i=1}^{2} \mathcal{L}_{\text{DOA}_i} + \beta * \sum_{i=1}^{2} \mathcal{L}_{\text{wSDR}_i} \qquad (9)$$

where α and β are the weights of DOA estimation task and speech separation task, respectively. We adopt azimuth sorting training strategy [19], sorting all sources' labels in ascending order by DOAs. In this way, permutations of DOAs and speech signals are not mismatched and LaBNet can be clear about with which branch to estimate which source during training.

4. Experiments and Analysis

4.1. Dataset

We follow the simulation procedure in [20]. We simulate 6-channel reverberant speech from LibriSpeech corpus [21] with a linear microphone array with spacings of 4 cm, 4 cm, 12 cm, 4 cm, 4 cm. We use pyroomacoustics [22] to randomly simulate RIRs of 50, 10 and 10 different rooms for train, validation and test dataset split. The sizes (length, width, height) of rooms are ranged from 4m, 3m, 2.5m to 12m, 9m, 5m. The RT60s are sampled between 0.3s and 0.8s. The distance between source and microphone array is set to [0.5m, 8m]. After simulation, we randomly select two source signals of different speakers from the same room to mix and truncate to 4 seconds long segments. The distance between two sources is set to at least 1 meter due to social distancing. Our dataset contains 40,000 utterances (44 hours) for training, 5000 for validation, 3000 for testing. To evaluate ASR performance, we simulate 1000 more utterances which are not truncated and therefore aligned with label text. Examples with azimuth difference of $< 15°$, $15 - 45°$, $45 - 90°$ and $> 90°$ account for 16%, 26%, 32% and 26% in the dataset.

4.2. Experimental setup

In order to train on a single GeForce RTX 3090, we conduct all our experiments with a batch size of 4. When training LaBNet, we use one warm-up epoch and Adam optimizer [26]. The initial learning rate is set to 1e-4 and the max norm of the gradients is set to 3. For the first 10 epochs, LaBNet is trained with emphasis on DOA estimation ($\alpha = 5$, $\beta = 1$). And for rest epochs, α and β are set to 1 and 10, respectively. For ablation study, we train GRNN-BF with only separation loss after removing 2D Locator of LaBNet. We apply uPIT strategy [27] when training all systems except LaBNet which is trained with azimuth sorting training strategy. The model achieves the best performance on validation set is chosen after 40 epochs. STFT is conducted with 512-point FFT along 32ms Hamming window with 50% stride, and the output feature dimension is 257. The STFT and iSTFT are implemented with fixed convolutional encoder and decoder [11]. The cosIPDs [28] are computed from five pairs between the first microphone and the rest microphones. The cRF estimator is a 2-layer uni-directional GRU with 500 hidden units followed by 4 FC layers with ReLU activation [29]. The GRU module in DOA estimator is a 2-layer uni-directional GRU with 210 hidden units and every neural beamformer contains a FC layer and a 2-layer uni-directional GRU with 300 hidden units followed by a FC layer. The size of cRF $K \times K$ is set to 3×3. The constant σ in spatial spectrum coding is set to 8. We use pretrained Conformer-CTC Large to evaluate the ASR performance. See https://huggingface.co/nvidia/stt_en_conformer_ctc_large.

4.3. Results and discussion

The reverberant speech of each source is taken as reference signal to measure SI-SDR [30], PESQ and WER of the separated speech. DOA estimation performance is also tested with accuracy and mean absolute error for LaBNet. Azimuth sorting

Table 1: *Experimental results on our simulated dataset. SI-SDR (dB), PESQ and WER (%) are evaluted under different azimuth difference ranges between speakers. Bold face indicates the best performance.*

Systems / Metrics	Features & Setup	SI-SDR (dB) ↑ / PESQ ↑ / WER (%) ↓				
		$< 15°$	$15 - 45°$	$45 - 90°$	$> 90°$	Avg.
Conv-TasNet [23]	monaural waveform	4.35/1.42/72.6	4.46/1.42/69.8	4.62/1.43/68.8	4.75/1.44/68.5	4.58/1.43/69.2
FaSNet-TAC [24]	multi-channel waveform	3.80/1.34/73.5	6.17/1.49/62.4	7.80/1.63/49.0	8.37/1.68/36.7	6.99/1.56/67.1
NB-LSTM [25]	multi-channel STFT	2.61/1.36/79.1	6.56/1.71/57.3	9.57/2.09/28.7	10.65/2.29/11.9	8.13/1.95/32.3
GRNN-BF [7]	Mag, IPD	4.94/1.57/46.3	8.03/1.94/25.3	9.55/2.13/24.4	9.92/2.18/13.1	8.62/2.02/29.9
GRNN-BF-Large	Mag, IPD	5.18/1.61/45.9	8.35/2.01/25.1	10.11/2.24/23.0	10.55/2.31/12.5	9.10/2.11/26.9
LaBNet	Mag, IPD, Estimate 1 DOA	6.28/1.78/44.4	9.73/2.30/24.1	11.42/2.55/18.9	11.96/2.68/11.3	10.44/2.40/25.2
	Mag, IPD, Estimate 2 DOAs	6.42/1.79/40.1	9.80/2.28/23.3	11.53/2.59/16.2	11.88/2.64/11.6	10.46/2.43/22.6
	Mag, IPD, Estimate 2 DOAs & 2D coordinates	**6.62/1.82/39.0**	**10.01/2.34/21.5**	**11.73/2.63/15.7**	**12.11/2.71/10.6**	**10.69/2.47/20.4**

Table 2: *Accuracy (%) (error $< 5°$ as correct) and MAE (°) results reflecting frame-level DOA estimation performance.*

Model	Accuracy (%) ↑ / MAE (°) ↓				
	$< 15°$	$15 - 45°$	$45 - 90°$	$> 90°$	Avg.
DOAnet [31]	59.9/5.2	75.3/4.4	73.8/6.3	61.2/9.6	68.9/5.5
LaBNet	88.7/2.7	93.1/2.0	94.3/2.0	88.5/3.0	91.7/2.4

training strategy is only applied in training stage as azimuth order is not given when inferencing and all results are the ones computed from the permutation with the highest SI-SDR score.

4.3.1. Overall speech separation performance

As shown in Table 1, our proposed LaBNet significantly outperforms previous baselines: NB-LSTM [25] and GRNN-BF [7] in frequency domain, Conv-TasNet [23] and FaSNet-TAC [24] in time domain. Compared to GRNN-BF, LaBNet obtains 2.07 dB absolute improvement on SI-SDR (i.e., 10.69 dB vs. 8.62 dB). In metric of PESQ and WER, LaBNet achieves 0.45 and 9.5% absolute improvements over GRNN-BF.

Spatial overlapping issue. Although the average performance of all multi-channel models is substantially better than that of single-channel model Conv-TasNet, results of category $< 15°$ (azimuth difference between sources less than $15°$) reflect the limitation of multi-channel methods [24, 25] on spatial overlapping cases. LaBNet, however, mitigate this issue with a decent average SI-SDR of 6.62 dB and PESQ of 1.82, which manifests the inclusiveness of our system.

4.3.2. Ablation study

We conduct ablation study on LaBNet about whether to introduce 2D Locator, the number of DOAs to be estimated and whether to estimate sources' 2D coordinates. Note that GRNN-BF (14.6M) with 2D Locator is equivalent to LaBNet (15.5M). For a fair comparison, we scale GRNN-BF to a comparable model size (15.7M) by adding GRU layers, termed as GRNN-BF-Large. Additionally estimating 1 and 2 DOAs elevate overall performance to 10.44 dB/2.4/25.2% and 10.46dB/2.43/22.26% from 9.1 dB/2.11/26.9%. Adding the estimation of 2D coordinates brings 0.2 dB further improvement on SI-SDR. Thereby, LaBNet gains 1.6 dB SI-SDR, 0.36 PESQ and 6.5 % WER absolute improvements over GRNN-BF-Large without increasing model size. This verifies the beneficial guidance from direction and location embeddings to neural beamformer for separation task.

Table 2 compares the accuracy and MAE results of DOAnet [31] and LaBNet. Considering that $5°$ is an admissible error [32], the DOA estimation performance of LaBNet is quite out-
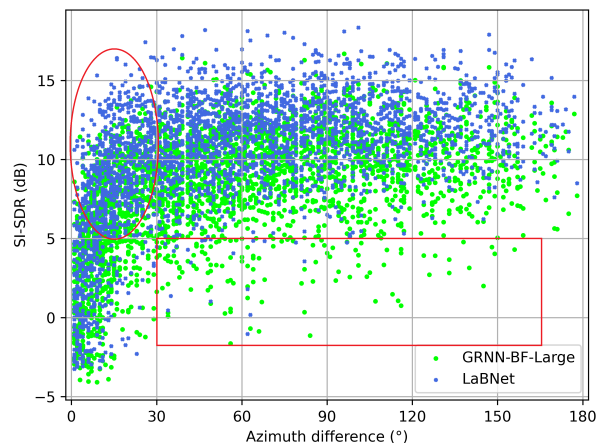

Figure 3: *The scatter plot of SI-SDR distribution under different azimuth differences between sources. Results of GRNN-BF and LaBNet are reported on testset.*

standing (average angular error of $2.36°$), even on spatial overlapping cases. By separately performing DOA estimation for each individual source rather than for all sources in the mixture once, DOAs are precisely predicted. This confirms that DOA estimation and separation in LaBNet benefit each other.

Fig.3 illustrates distributions of SI-SDR score of GRNN-BF-Large and LaBNet on testset. It can be observed that: (1) In terms of $< 30°$ cases, LaBNet produces less "bad cases" with low or negative SI-SDR and elevates more examples to higher than 5 dB SI-SDR (red oval). (2) 2D Location cue does provide extra discrimination for separation, even on large azimuth difference cases, which almost eliminates examples whose SI-SDR is less than 5 dB (red rectangle). This favors backend ASR task a lot. (3) The SI-SDR distribution estimated by LaBNet is more compact with higher mean but lower variance, which suits ASR systems that are sensitive to low SI-SDR cases.

5. Conclusion and Future Work

In this work, we design a novel multi-channel input and multiple outputs 2D location-aware beamforming network, aiming at providing additional discrimination with direction and location cues. By introducing a light-weight 2D locator (0.9M), our proposed system performs stunningly on both speech separation and DOA estimation tasks. Future work will explore an array geometry agnostic model and consider environmental noise.

6. Acknowledgements

7. References

[1] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.

[2] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

[3] W. Rao, C. Xu, E. S. Chng, and H. Li, "Target Speaker Extraction for Multi-Talker Speaker Verification," in *Proc. Interspeech 2019*, 2019, pp. 1273–1277.

[4] C. Xu, W. Rao, J. Wu, and H. Li, "Target speaker verification with selective auditory attention for single and multi-talker speech," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 29, pp. 2696–2709, 2021.

[5] Y. Xu, M. Yu, S.-X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu, "Neural Spatio-Temporal Beamformer for Target Speech Separation," in *Proc. Interspeech 2020*, 2020, pp. 56–60.

[6] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6089–6093.

[7] Y. Xu, Z. Zhang, M. Yu, S.-X. Zhang, and D. Yu, "Generalized Spatio-Temporal RNN Beamformer for Target Speech Separation," in *Proc. Interspeech 2021*, 2021, pp. 3076–3080.

[8] L. Chen, M. Yu, D. Su, and D. Yu, "Multi-band pit and model integration for improved multi-channel speech separation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 705–709.

[9] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[10] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "L-spex: Localized target speaker extraction," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7287–7291.

[11] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.

[12] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot, "Multi-microphone speaker separation based on deep doa estimation," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[13] A. Aroudi and S. Braun, "Dbnet: Doa-driven beamforming network for end-to-end reverberant sound source separation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 211–215.

[14] M. Hafsati, K. Bentounes, and R. Marxer, "Blind speech separation through direction of arrival estimation using deep neural networks with a flexibility on the number of speakers," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, 2022, pp. 1–5.

[15] W. Mack and E. A. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2019.

[16] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[17] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 74–79.

[18] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[19] H. Taherian, K. Tan, and D. Wang, "Location-based training for multi-channel talker-independent speaker separation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 696–700.

[20] H. Yin, M. Ge, Y. Fu, G. Zhang, L. Wang, L. Zhang, L. Qiu, and J. Dang, "MIMO-DoANet: Multi-channel Input and Multiple Outputs DoA Network with Unknown Number of Sound Sources," in *Proc. Interspeech 2022*, 2022, pp. 891–895.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[22] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 351–355.

[23] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6394–6398.

[25] C. Quan and X. Li, "Multi-channel narrow-band deep speech separation with full-band permutation invariant training," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 541–545.

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

[27] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[28] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5.

[29] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[30] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[31] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1462–1466.

[32] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.