Single-Channel Target Speech Extraction Utilizing Distance and Room Clues

Runwu Shi, Zirui Lin, Benjamin Yen, Jiang Wang, Ragib Amin Nihal, Kazuhiro Nakadai

Department of Systems and Control Engineering, Institute of Science Tokyo, Tokyo, Japan

{shirunwu, linzirui, wangjiang, ragib, nakadai}@ra.sc.e.titech.ac.jp, benjamin.yen@ieee.org

***Abstract*—This paper aims to achieve single-channel target speech extraction (TSE) in enclosures utilizing distance clues and room information. Recent works have verified the feasibility of distance clues for the TSE task, which can imply the sound source's direct-to-reverberation ratio (DRR) and thus can be utilized for speech separation and TSE systems. However, such distance clue is significantly influenced by the room's acoustic characteristics, such as dimension and reverberation time, making it challenging for TSE systems that rely solely on distance clues to generalize across a variety of different rooms. To solve this, we suggest providing room environmental information (room dimensions and reverberation time) for distance-based TSE for better generalization capabilities. Especially, we propose a distance and environment-based TSE model in the time-frequency (TF) domain with learnable distance and room embedding. Results on both simulated and real collected datasets demonstrate its feasibility. Demonstration materials are available at https://runwushi.github.io/distance-room-demo-page/.**

***Index Terms*—Target speech extraction, distance-based sound separation, single-channel.**

I. Introduction

Target speech extraction (TSE) is the task that utilizes speaker-related information as auxiliary clues to extract target speech from an audio mixture consisting of multiple speakers. Such speaker clues can be categorized into physiological and spatial information about the speaker. Currently, physiological clues dominate research in TSE, which includes enrolled voice [1]–[3], facial and lip movement [4], and even gestures while speaking [5]. Another speaker clue is spatial-related information, mainly referring to the direction of arrival (DOA), which has been explored in the early stages of TSE. Specifically, studies use microphone arrays to enhance speech from the speaker's direction [6], [7], as well as neural TSE systems [8], [9]. Direction clues offer several advantages over physiological clues. For example, direction clues do not require ideal speaker-related information, which may not always be readily available, and relying on direction clues circumvents the need to handle sensitive biometric data such as the target speaker's voice [10]. However, direction clues present challenges because they are less deterministic and are less effective in separating speakers with the same direction.

Despite these limitations, we argue that there is still room for further exploration of spatial-based TSE, particularly considering that another spatial clue: the target speaker's distance, has been underdeveloped in TSE for a considerable period. Unlike DOA-based TSE, which requires DOA estimations

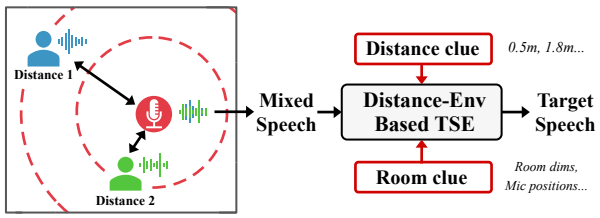

Fig. 1. Target speech extraction using distance and room clues.

from a microphone array, distance-based methods can be realized in single-channel, providing more flexibility. It should be noted that distance-based methods are often designed for reverberant enclosures since distance clues rely on the transfer properties between the speaker and the microphone in enclosed spaces, where the direct-to-reverberation ratio (DRR) decreases as the target-to-microphone distance increases. To explain, the late reverberation component is less affected by distance, while the direct and early components are inversely proportional to it [11]–[13].

Few studies have used distance information in TSE or speech separation (SS). The first attempt is Distance Based Sound Separation (DSS) [13], in which a basic recurrent neural network (RNN) model is used to separate the single-channel mixed audio into near and far groups according to a static distance threshold such as 1.5m. To make such threshold-based methods more flexible, [14] proposes a multichannel region based sound extraction system, in which DOA and distance clues are utilized simultaneously. Furthermore, this study also proposed a dynamic distance threshold to control the boundary between the near and far groups directly. Following DSS, [15] also adopts the idea of a static distance threshold for TSE. This method contains a speaker encoder that only outputs the speaker embedding within the distance threshold, thereby the distance clue is indirectly utilized, implying this method still classifies as enrolled voice-based TSE.

Unlike previous approaches, [16] introduced a distance-based TSE that relies solely on distance clues for single-channel TSE. This model outputs the speaker located near the query distance rather than separating the mixture into near and far groups. However, distance clues are highly sensitive to room acoustic parameters such as microphone placement, room dimensions, and reverberation time. Since the model does not consider these factors, its ability to generalize across

different environments is significantly limited. Moreover, its performance in real-world scenarios has yet to be verified.

To address this, we explicitly incorporate room acoustic characteristics into the distance-based TSE system by designing a TF domain model that utilizes learnable distance and room embeddings. To validate this approach, we built both simulated and real-world datasets, including a location-wise real room impulse response (RIR) dataset and real-world speech recordings. The experimental results demonstrate the feasibility of our method. The remainder of this paper is organized as follows: Section 2 describes the proposed method, Section 3 presents the datasets and experimental results, and Section 4 concludes the paper.

II. Methodology

A. Problem Formulation

Assuming K speakers in an enclosed room, and denote $s_k(t)$, $x_k(t)$ and $h_k(t)$ as the kth speaker's anechoic speech, reverberant speech, and RIR, respectively. The reverberant speech $x_k(t)$ can be formulated as

$$x_k(t) = s_k(t) * h_k(t), \tag{1}$$

where $*$ represents the convolution operator. For a single-channel microphone in the room. The collected signal mixture can be represented as

$$y = \sum_{k=1}^{K} x_k(t). \tag{2}$$

For the proposed distance-based TSE system, it is essential to extract the target speech given a query distance and identify the presence or absence of speakers at the same time. The extraction process can be depicted as

$$y \xrightarrow{d_q} \sum_{k} x_k(t), k \text{ s.t. } |d_k - d_q| \leq r_{spk}. \tag{3}$$

where the d_q is the query distance, d_k means the relative distance between the microphone and the kth speaker, and r_{spk} represents the speaker distance range centered with the ground truth speaker distance. Thus the boundary between the presence and the absence will be r_{spk}. Given a query distance d_q, the model outputs the target speech if d_q falls within a single speaker's range and aggregates speech from all speakers when d_q overlaps multiple ranges. Moreover, the model needs to output zero if the query distance d_q falls out of the range r_{spk}, depicted as

$$y \xrightarrow{d_q} 0, \forall k, \ |d_k - d_q| > r_{spk}. \tag{4}$$

B. Proposed Model

The proposed TSE model follows the clue encoder combined with speech extraction module strategy in the TF domain [7]. The overall structure is shown in Fig 2 (a). The input feature of the model is the concatenated real and

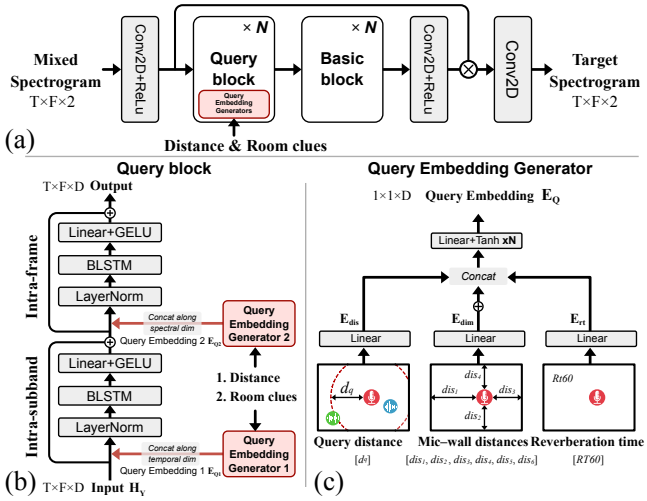

Fig. 2. Flowchart of the proposed method showing (a) Overall structure, (b) Structure of Query block, (c) Structure of Query embedding generator.

imaginary components $\mathbf{Y} \in \mathbb{R}^{2\times T\times F}$ obtained through short-time Fourier transform (STFT), which is then processed by the encoder consisting of the initial 2D convolution layer with 3×3 kernel and the global layer normalization followed by the rectified linear unit (ReLu) activation function to obtain the non-negative D-dimensional TF embeddings $\mathbf{H_Y} \in \mathbb{R}^{D\times T\times F}$ [2], [17]. The TF embeddings $\mathbf{H_Y}$ are fed into the Query and Basic block stacks, followed by a 2D convolution layer with 3×3 kernel and a ReLu activation to output the target speech mask element multiplied with $\mathbf{H_Y}$. Finally, a 2D convolution layer with 3×3 kernel is used to map the TF embeddings into real and imaginary components, and the obtained complex spectrogram is converted back into time-domain using the inverse STFT. This section introduces these modules.

1) Query block: The Query block contains two Query embedding generators (QEG) for learnable query embeddings $\mathbf{E_{Q1}}$ and $\mathbf{E_{Q2}}$, learned from query distance and room features, as shown in Fig 2 (b). These embeddings are fused with the intermediate representation in both temporal and spectral domains, which can be regarded as retrieving the learned query embeddings in all TF bins among intra-subband and intra-frame directions. In detail, for the intra-subband process, the input TF embedding $\mathbf{H_Y}$ is first concatenated with the query embedding $\mathbf{E_{Q1}}$ along the temporal axis. This combined embedding is passed through Layer Normalization (LN) along its last dimension and subsequently processed by a Bidirectional Long Short-Term Memory (BLSTM) layer. Finally, the last concatenated dimension in the hidden state of the BLSTM is cropped to match the input dimension and then passed through a linear layer that is activated using a Gaussian error linear unit (GELU) and connected using residual. The intra-frame fusion follows the same procedure, but concatenates the query embedding $\mathbf{E_{Q2}}$ along the spectral dimension, and applies the BLSTM across subbands in each frame.

2) Query embedding generator: The QEG converts the raw query distance and room information into useful query embeddings. The structure of QEG is presented in Fig 2 (c).

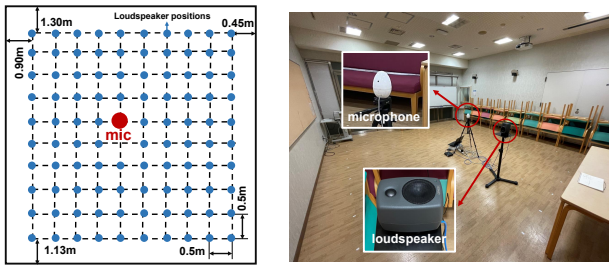

Fig. 3. RIR recording locations and environment.

The used clues are raw query distance d_q, and room clues including microphone-wall distances sequence $Dis_{mw} = [dis_1, dis_2, ..., dis_6]$, and reverberation time $RT60$, each of which corresponds to one linear layer and outputs three embeddings E_{dis}, E_{dim}, and E_{rt}, these embeddings are concatenated and fed into three linear layers with tanh activation function for final query embedding $\mathbf{E_Q}$ with D-dimension [14]. We use the distance from the microphone to surrounding walls as features because it represents the room dimensions and microphone placement simultaneously, and it has been shown that the sound field transfer function can reflect the structure of the surrounding space, e.g., the distance to the nearest wall [18]. The E_{dis} embedding is calculated by summing all individual mic-wall distance embeddings, and this ambiguous representation shows better generalization performance.

3) Basic block: The Basic block adopts the same structure as the Query block but omits the fusion step, applying a single RNN across all frames and subbands, similar to DPRNN [19].

C. Loss function

Two kinds of loss functions are used for the model to output both target speech and zero speech conditioned on the query distance. Given the target speech $\mathbf{x}_s$ and the estimated speech $\hat{\mathbf{x}_s}$, we adopt the signal-to-distortion ratio (SDR) as the loss function when the speakers exist at the query distance,

$$L_{active}(\mathbf{x}_s, \hat{\mathbf{x}_s}) = 10\log_{10}\left(\frac{\|\mathbf{x}_s\|^2}{\|\mathbf{x}_s - \hat{\mathbf{x}_s}\|^2 + \tau\|\mathbf{x}_s\|^2}\right) \quad (5)$$

where the τ is the soft threshold to control the upper bound of the loss and set to 10^{-3}. The scale invariant SDR (SI-SDR) [20] is not considered since we wish the model could handle inactive source conditions in which the model would output zero values [21]. For the condition that no speaker is near the query distance, the adopted loss function is L_0 [21], [22],

$$L_0(\mathbf{y}, \hat{\mathbf{x}_s}) = 10\log_{10}\left(\|\hat{\mathbf{x}_s}\|^2 + \tau^{\text{inactive}}\|\mathbf{y}\|^2\right), \quad (6)$$

where $\mathbf{y}$ is the mix speech, and the τ^{inactive} represents the soft threshold which is set to 10^{-2}.

III. Experiment

A. Dataset Generation

For model training, the mixed speech is formed by superposing multiple reverberant signals, each generated by convolving an anechoic speech signal with a location-specific RIR, and we consider both simulated and realistic RIRs. Additionally, real-world speech is also recorded for further validation.

Simulated RIR dataset: We used the randomized image method for RIR generation [23]. Specifically, considering different room settings and microphone placement significantly influenced the transfer function characteristics and the model performance, we constructed two sub-simulated RIR datasets to assess the model's generalization capabilities.

*1) **Sim1** One room with fixed microphone position:* This dataset contains a $7 \times 8 \times 3$ meters room with a fixed microphone at $3.5 \times 4 \times 1.1$ meters. The reverberation time (RT60) is set to 0.2s. Speaker positions are randomly initialized at least 0.5 meters from the walls, with heights ranging from 1.2 to 2.0 meters. 10,000 RIRs were generated within 0.2 to 5.0 meters from the microphone.

*2) **Sim2** 1,000 rooms:* This simulated dataset contains 1,000 randomly generalized rooms between sizes $4 \times 5 \times 2.5$ meters to $8 \times 10 \times 3.0$ meters. The RT60 is randomly selected between 0.2s to 0.5s. Each room has one microphone position, each corresponding to 500 speaker positions, resulting in 490,325 RIR samples. To ensure a more balanced distribution of samples across varying distances, the generation process is organized into distance bands, e.g., 0-0.5 meters, so the RIRs for farther speakers may be missing in smaller rooms.

Real recorded dataset: Open source datasets of location-dependent RIRs or real recordings are scarce. To make the evaluation more realistic, we collected real RIRs and speech samples in a conference room with a reverberation time of 0.6 seconds. The room size is $5.9 \times 6.9 \times 2.9$ meters.

*1) **Real recorded RIR**:* We collect real RIRs in a meeting room using the sine sweep method and follow the processing from [24]. We use an eight-channel microphone array to record the RIR at 98 locations, as shown in Fig 3, resulting in 784 single-channel RIR segments. This dataset is publicly released for community research. [1]

*2) **Real-world recorded speech**:* We collect real speech signals in the same room. Specifically, we fix the microphone and play speech signals from the LibriSpeech dataset at 14 distinct locations with different distances. For each location, we play ten speech segments from the same speaker. Since the speech is recorded separately, supervised training can be realized by overlapping different recorded speech.

Speech and sample generation: For the Sim1, Sim2, and real recorded RIR datasets, the speech is adopted from the Libri-Light [25] dataset, and 128, 48, and 64 speakers were used for training, validation, and testing. Simulated RIR datasets were split into training, validation, and testing datasets with a ratio of 0.9:0.02:0.08. Each utterance with a length of 4s is convolved from randomly chosen speech and RIR, and the root mean square energy is randomly adjusted from -25 to -20 dB to simulate the louder voices of farther speakers.

B. Training Process and Configuration

The model contains 4 Query blocks and 4 Basic blocks, the dimension size D and the hidden state size of LSTM is set to 64. The three linear layers in each QEG contain 96, 64, and

[1] https://github.com/RunwuShi/GridRIRDataset

TABLE I
RESULTS OF TWO SPEAKER MIXTURES ON SIM1 AND SIM2. SDR AND SDRI REPORT **NON-OVERLAP**/OVERLAP SPEAKER CONDITIONS.

Training Dataset	Clue	SDR↑ (dB)	SDRi↑ (dB)	PESQ↑	L_0 ↓	No/o ratio
Sim1	Dis	11.95/24.48	11.99/-5.52	2.98	4.56	20.40%
Sim2	Dis	6.71/18.07	6.62/-11.93	2.34	20.35	20.30%
	Dis+Dim+Rt	7.96/19.03	8.08/-10.97	2.54	12.52	20.20%
	Dis+Rt	7.46/19.35	7.45/-10.65	2.46	16.04	20.08%
	Dis+Dim	7.13/19.99	7.17/-10.01	2.42	16.09	20.44%

TABLE II
GENERALIZATION PERFORMANCE ON REAL COLLECTED RIR DATASET.

Training Dataset	Clue	SDR↑ (dB)	SDRi↑ (dB)	PESQ↑	L_0 ↓	No/o ratio
Sim1	Dis	0.71/5.32	0.74/-24.68	1.56	14.29	38.56%
Sim2	Dis	3.06/10.32	3.00/-19.68	1.95	21.37	40.64%
	Dis+Dim+Rt	4.50/9.46	4.42/-20.54	2.11	14.30	39.06%
	Dis+Rt	4.01/10.40	4.02/-19.60	2.04	19.52	40.54%
	Dis+Dim	3.35/7.32	3.40/-22.68	1.94	16.13	38.62%
Finetune	Dis	8.20/11.42	8.06/-18.58	2.57	14.43	8.84%

64 units. A frame length of 32 ms with 16 ms frameshift is used in STFT. The r_{spk} centered with ground truth speaker distance is set to 0.5 meters. For training, we use the Adam [26] optimizer, and gradient clipping is set to a maximum norm of 5. The batch size is set to 14. The initial learning rate is 0.001, and if no lower loss is found in 10 epochs, the learning rate is reduced to 80%. Training has 400 epochs with 25% inactive samples corresponding to zero output.

IV. RESULTS AND DISCUSSION

A. Results on Simulated RIR Dataset

The evaluation metric needs to consider both the presence and absence of speakers at the query distance. For the presence condition, we use SDR, SDR improvement (SDRi), and Perceptual Evaluation of Speech Quality (PESQ) to verify speech quality and scale accuracy. For the absence condition, we use the inactive loss L_0 as the metric. Each test generates 1,000 utterances randomly, and the results of each dataset are obtained by testing five times and expressed as the mean. Specifically, for the presence condition of existing multiple speakers near the query distance, the model will output the overlapped speech, and we report the SDR and SDRi under the non-overlapped and overlapped speaker conditions separately. We also present the ratio of non-overlap/overlap samples.

Table I presents the results of distance-based TSE under conditions with and without adding room information as the clues, and we test both models trained on two speaker mixtures from the Sim1 and Sim2 datasets. Sim1 scenario can be regarded as the most idealized experiment setting, in which the model is trained and tested in the same room with a fixed microphone position setup, and this performance should be the best performance of such distance-based TSE methods. For the Sim1 dataset containing one room, the TSE using only distance clue achieves an SDR of 11.95 dB under non-overlapped speaker conditions. For the Sim2 dataset containing 1,000 rooms, we evaluate the performance of distance-based TSE methods using various additional room-related clues, which utilize room configuration (Dis+Dim), reverberation time (Dis+Rt), and a combination of both as auxiliary clues (Dis+Dim+Rt). The result presents the model using room configuration and reverberation time achieves an SDR of up to 7.96 dB. Moreover, lower L_0 indicates that the use of room clues effectively improves recognition of non-existent speakers, which is essential for speaker distance estimation [11], [12], [16]. The results also show that the reverberation information improves the quality of the extracted speech more than the microphone-wall distances. Overall, the different rooms introduced by Sim2 have a great performance decay relative to the one room of Sim1, suggesting the necessity of additional room information.

TABLE III
PERFORMANCE OF FINETUNED MODEL ON REAL RECORDED SPEECH

Finetune Type	SDR ↑ (dB)	SDRi ↑ (dB)	SI-SDR ↑ (dB)	SI-SDRi ↑ (dB)	PESQ↑
All location	7.24	7.22	5.19	5.17	2.46
New location	7.34	7.31	5.57	5.55	2.45

B. Performance on Real Recorded RIR and Speech Dataset

This section first evaluates the generalization ability of several distance-based TSEs on the real collected RIR datasets. The results are presented in Table II. For distance clue-only methods, the Sim2-trained model achieves an SDR of 3.06 dB, outperforming the model trained on Sim1, which contains one room. For the methods with additional room clues, the SDR can be further improved to 4.50 dB. We also finetune the model trained on Sim2 by 100 epochs using 70% of the real RIR dataset and set r_{spk} to 0.1m for more accurate extraction. The SDR can be further enhanced to 8.20 dB. These results demonstrate the effectiveness of the proposed method in real RIR scenarios.

To verify the effectiveness of the distance-based TSE methods on real-world speech, we also finetune the model on real recorded reverberant speech segments. Finetuning is necessary because both simulated and realistic RIRs differ from actual reverberant conditions, especially given the sensitivity of distance cues to environmental factors. Specifically, we compare two strategies, in the first strategy, we use all of the locations to finetune the model, where 50% of the audio is used with finetune and the other 50% is used for testing, and in the second strategy, we use all of the audio from the 7 locations for finetune and use the audio from the others locations for testing. The finetune process contains 10 epochs with a learning rate of 0.0001. The results are presented in Table III. The distance-based TSE can achieve an SDR of 7.34 dB under the second fintune strategy.

TABLE IV
COMPARISON OF DISTANCE AND ENROLLED VOICE BASED TSE ON SIM2

Method	Params ($\times 10^6$)	SDR ↑ (dB)	SDRi ↑ (dB)	SI-SDR ↑ (dB)	SI-SDRi ↑ (dB)	PESQ↑
Dis	1.25	6.71	6.62	5.46	5.37	2.34
Dis+Dim+Rt	1.29	7.96	8.08	6.94	7.05	2.54
SpEx+ [27]	11.15	–	–	3.59	3.39	1.85
CIENet [1]	2.67	–	–	9.19	8.90	2.72

C. Performance Comparison with Different Methods

As shown in Table IV, the proposed method is compared with other methods on the Sim2 dataset. We reproduce two advanced enrolled speech-based TSE models, a time-domain TSE model SpEx+ [27] and a TF-domain TSE model CIENet-mDPRNN [1], for which we used 8 seconds of target speaker voice as the clue for both models and the training objective and evaluation metrics follow the original papers. We adopt the official implementation of SpEx+ and implement CIENet ourselves. For a fair comparison, we test the distance-based methods where each query distance corresponds to only one speaker. The results demonstrate the feasibility of distance-based TSE. However, due to the heterogeneity between the distance clue and speech, extracting speech is inherently more challenging than using the enrolled voice.

V. CONCLUSION

We proposed to utilize distance and room information as clues for the TSE system, and we built realistic datasets to verify the proposed method. The results demonstrate the use of such room information enhances the performance and generalization ability of distance-based TSEs, and the performance on real recorded speech also proves the potential of this task. Future work will focus on more real-world scenarios.

REFERENCES

[1] X. Yang, C. Bao, J. Zhou, and X. Chen, "Target Speaker Extraction by Directly Exploiting Contextual Information in the Time-Frequency Domain," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 476–10 480.

[2] X. Yang, C. Bao, and X. Chen, "Coarse-to-Fine Target Speaker Extraction Based on Contextual Information Exploitation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3795–3810, 2024.

[3] L. Yang, W. Liu, L. Tan, J. Yang, and H.-G. Moon, "Target Speaker Extraction with Ultra-Short Reference Speech by VE-VE Framework," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[4] Z. Mu and X. Yang, "Separate in the Speech Chain: Cross-Modal Conditional Audio-Visual Target Speech Extraction," 2024.

[5] Z. Pan, X. Qian, and H. Li, "Speaker Extraction With Co-Speech Gestures Cue," *IEEE Signal Processing Letters*, vol. 29, pp. 1467–1471, 2022.

[6] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, 1985.

[7] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, "Neural Target Speech Extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[8] R. Gu, L. Chen, S.-X. Zhang, J. Zheng, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Neural Spatial Filter: Target Speaker Speech Separation Assisted with Directional Information," in *Interspeech 2019*. ISCA, 2019, pp. 4290–4294.

[9] 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.

[10] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, and J.-F. Bonastre, "Adversarial disentanglement of speaker representation for attribute-driven privacy preservation," *arXiv preprint arXiv:2012.04454*, 2020.

[11] S. S. Kushwaha, I. R. Roman, M. Fuentes, and J. P. Bello, "Sound Source Distance Estimation in Diverse and Dynamic Acoustic Conditions," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023, pp. 1–5.

[12] M. Neri, A. Politis, D. A. Krause, M. Carli, and T. Virtanen, "Speaker Distance Estimation in Enclosures From Single-Channel Audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2242–2254, 2024.

[13] K. Patterson, K. Wilson, S. Wisdom, and J. R. Hershey, "Distance-Based Sound Separation," in *Interspeech 2022*. ISCA, 2022, pp. 901–905.

[14] R. Gu and Y. Luo, "Rezero: Region-customizable sound extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[15] J. Lin, P. Wang, H. Dinkel, J. Chen, Z. Wu, Z. Yan, Y. Wang, J. Zhang, and Y. Wang, "Focus on the sound around you: Monaural target speaker extraction via distance and speaker information," *arXiv preprint arXiv:2306.16241*, 2023.

[16] R. Shi, B. Yen, and K. Nakadai, "Distance based single-channel target speech extraction," *arXiv preprint arXiv:2412.20144*, 2024.

[17] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[18] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3165–3177, 2022.

[19] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.

[20] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[21] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki, "Soundbeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 121–136, 2022.

[22] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the Fuss about Free Universal Sound Separation Data?" in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 186–190.

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

[24] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio engineering society convention 122*. Audio Engineering Society, 2007.

[25] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673.

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

[27] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," *arXiv preprint arXiv:2005.04686*, 2020.