SSDQ: Target Speaker Extraction via Semantic and Spatial Dual Querying

Xinjia Zhu, Xinyuan Qian, *Senior Member, IEEE*, and Dong Liang

***Abstract*—Target Speaker Extraction (TSE) in real-world multi-speaker environments is highly challenging. Previous works have largely relied on pre-enrollment speech to extract the target speaker's voice. However, such methods are limited in spontaneous scenarios where pre-enrollment speech or spatial information is unavailable. To address this, we propose Semantic and Spatial Dual Querying (SSDQ), a unified framework that integrates natural language descriptions and region-based spatial queries to guide Target Speaker Extraction (TSE). SSDQ employs dual query encoders for semantic and spatial cues, fusing them into the audio stream via a FiLM-based interaction module. A novel Controllable Feature Wrapping (CFW) mechanism further enables a dynamic balance between speaker identity and acoustic clarity. We also introduce SS-Libri, a spatialized mixture dataset designed to benchmark dual-query systems. Extensive experiments demonstrate that SSDQ achieves superior extraction accuracy and robustness under challenging conditions, yielding the SI-SNRi of 19.63 dB, SNRi of 20.30 dB, PESQ of 1.83, and STOI of 0.26.**

***Index Terms*—Target speaker extraction, multi-modal, multi-channel target speaker extraction.**

I. Introduction

Speech Enhancement (SE) [1], [2], [3], [4] and Speech Separation (SS) [5], [6] are techniques that either improve the overall signal quality or separate all sources regardless of the speaker identity. In contrast, TSE focuses on isolating and extracting the speech of a specific target speaker from a mixture of overlapping audio sources. Because it does not require prior information about the number or identity of interfering speakers, TSE finds particular utility in applications like personalized voice assistants [7] and speaker-specific transcription [8], where the aim is to focus solely on a target speaker's voice.

Traditional TSE methods [9], [10], [11], [12], [13] often rely on audio cues to distinguish the target speaker from a mixture of sources, such as leveraging pre-recorded clean speech as auxiliary information to guide the extraction process. However, such reliance poses a major limitation, as clean reference audio is seldom available in real-world scenarios, particularly in AI speech assistant applications, where prior knowledge of the target speaker is typically unavailable. Other prevalent TSE methods [14], [15], [16] utilize visual cues, such as lip or gesture movements and facial features, to identify and isolate the target speaker. Although these methods provide valuable information, they also have limitations, as visual cues may not always be available or reliable in all scenarios.

By leveraging multi-channel microphone arrays or other spatially-aware sensors, methods [17], [18] can locate the positions of sounding objects. Indeed, spatial TSE methods [19], [20], [21] can also exploit the relative positions of speakers to help isolate the target voice. However, these methods still face challenges, particularly when the speaker's precise spatial location and identity are unknown. In parallel, the emergence of text-query based TSE methods [22], [23], [24] introduces a promising new modality for guiding speaker extraction via natural language descriptions. While such semantic guidance enhances flexibility and applicability in open-domain scenarios, these methods often lack the spatial resolution needed to disambiguate multiple candidates with similar textual traits. To the best of our knowledge, no existing work has jointly leveraged both spatial and semantic cues to guide target speaker extraction. Combining these complementary modalities enables accurate extraction, especially in complex auditory scenes involving multiple speakers with overlapping speech content.

To address these challenges, we propose SSDQ, a novel approach that integrates both semantic and spatial queries for TSE. As shown in Fig. 1, it utilizes dual queries: (i) a spatial query defines a precise region through a concrete angular range and (ii) a semantic query defines speaker attributes and a vague region (e.g. front-left). The contributions are as follows:

- We introduce SSDQ, a unified framework that jointly leverages semantic and spatial queries from multi-channel audio to enable robust target speaker extraction in complex multi-speaker scenarios.
- We propose CFW, a novel feature modulation technique that dynamically adjusts the model's attention between speaker identity and speech clarity.
- We design and collect SS-Libri, a synthetic spatialized mixture dataset that showcases the potential of dual-query systems to improve TSE.
- Experimental results demonstrate that SSDQ achieves superior performance, with the SI-SNRi of 19.63 dB, SNRi of 20.30 dB, PESQ of 1.83, and STOI of 0.26, outperforming existing methods in both intelligibility and perceptual quality.

Received 7 May 2025; revised 11 July 2025; accepted 18 July 2025. Date of publication 21 July 2025; date of current version 14 August 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62306029, in part by Shenzhen Longhua Science and Technology Innovation Special Funding Project (Industrial Sci-Tech Innovation Center of Low-Altitude Intelligent Networking) and Shenzhen Research Institute of Big Data under Grant K00120240007, and in part by Beijing Natural Science Foundation under Grant L233032. The associate editor coordinating the review of this article and approving it for publication was Prof. Huseyin Hacihabiboglu. *(Corresponding author: Xinyuan Qian.)*

Xinjia Zhu and Dong Liang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: xinjiazhu@nuaa.edu.cn; liangdong@nuaa.edu.cn).

Xinyuan Qian is with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: qianxy@ustb.edu.cn).

Demo, code and data are available at https://zhuxj821.github.io/SSDQ-project/.

Digital Object Identifier 10.1109/LSP.2025.3591408

1070-9908 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.
Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 23,2025 at 07:47:18 UTC from IEEE Xplore. Restrictions apply.

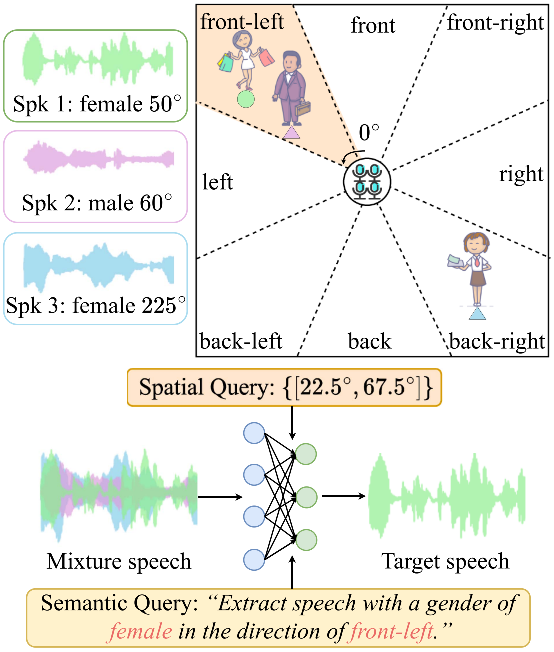

Fig. 1. Our proposed TSE via semantic and spatial dual querying. The spatial query specifies a spatial direction, while the semantic query defines speaker attributes and vague direction, allowing precise isolation of the target speaker (Spk: speaker).

II. Methods

A. Problem Formulation

Let $\mathbf{z}^m(\tau)$ denote the target speaker's speech, $\mathbf{s}_i^m(\tau)$ represent the interference speech signal from the i-th interfering speaker, and $\mathbf{n}^m$ be the noise signal, where $i \in \{1, 2, \ldots, C\}$ and C is the total number of interfering speakers. Here, m denotes the channel index, with multiple channels capturing spatial audio information. The mixture speech $\mathbf{y}^m(\tau)$ can be expressed as

$$\mathbf{y}^m(\tau) = \mathbf{h}_0^m * \mathbf{z}(\tau) + \sum_{i=1}^{C} \mathbf{h}_i^m * \mathbf{s}_i(\tau) + \mathbf{n}^m(\tau) \tag{1}$$

where $\mathbf{h}_0^m$ and $\mathbf{h}_i^m$ are the Room Impulse Response (RIR) vectors between the target speaker and interfering speakers. $*$ denotes the convolution operator. Specifically, TSE aims to extract the estimated $\hat{\mathbf{z}}(\tau)$ that approximates $\mathbf{z}(\tau)$ from $\mathbf{y}^m(\tau)$.

B. Overview

To extract the target speaker from spatialized speech mixtures, we propose a novel TSE framework, SSDQ. As shown in Fig. 2, SSDQ is guided by two complementary inputs: (i) a spatial query Q_{spa} represent a precise region through a concrete angular range, and (ii) a semantic query Q_{sem}, which encodes speaker attributes such as gender and vague spatial region. The input mixture $\mathbf{y}^0(\tau)$ is first encoded into frame-level features $\mathbf{y}_t$. Spatial features are extracted via Interaural Phase Difference (IPD) and Target Phase Difference (TPD) computation, guided by Q_{spa}, to produce a spatial embedding $\mathbf{c}_{\text{spa}}$. In parallel, the semantic query is encoded into a semantic embedding $\mathbf{c}_{\text{sem}}$. These are fused via a FiLM-based modulation to produce a multimodal representation $\mathbf{x}_t$. The fused features are processed by a DPRNN to model both intra-chunk and inter-chunk dependencies, resulting in intermediate output $\mathbf{m}_t$. A CFW module then integrates $\mathbf{m}_t$ and the original $\mathbf{y}_t$ using ResBlock1D and TCN layers, yielding enhanced features $\mathbf{w}_t$. These are combined with $\mathbf{y}_t$ to reconstruct the target speech $\hat{\mathbf{z}}(\tau)$.

C. Speech Encoder

The first channel of the multi-channel input, $\mathbf{y}^0(\tau)$, is encoded by a 1-D convolutional front-end to produce frame-level representations:

$$\mathbf{y}_t = Enc_{\text{speech}}(\mathbf{y}^0(\tau)) \tag{2}$$

where $Enc_{\text{speech}}(\cdot)$ denotes a temporal encoder with overlapping convolutional windows of kernel size l_k, and $t \in \{1, \ldots, \lfloor \frac{2(T-l_k)}{l_k} \rfloor + 1\}$ indexes the output frames.

D. Spatial Feature Calculation

Following [25], IPD is computed in the time domain using convolutional kernels approximating the short-time Fourier transform (STFT) filters. For microphone pair (u_1, u_2):

$$\text{IPD}_{nf}^{(u)} = \arctan\left(\frac{y^{u_1} * F_{nf}^{\text{re}}}{y^{u_1} * F_{nf}^{\text{im}}}\right) - \arctan\left(\frac{y^{u_2} * F_{nf}^{\text{re}}}{y^{u_2} * F_{nf}^{\text{im}}}\right) \tag{3}$$

where $y^{u_{(\cdot)}}$ denotes the waveform input, $*$ denotes the convolution operator, n is the time frame index, and f is the frequency bin. F_{nf}^{re} and F_{nf}^{im} are the real and imaginary kernels.

To compute TPD, we define $Q_{\text{spa}} = \{[\theta_s, \theta_e]\}$, where θ_s and θ_e represent the start and end azimuth angles (in degrees) that define the spatial sector of interest. Subsequently, we sample azimuth angles $\{\theta_k\}_{k=1}^{K}$ within this interval for directional reasoning and compute corresponding theoretical delays $\zeta^{(u)}(\theta_k)$. Following Rezero [26], these are embedded as phase patterns $\mathbf{e}_{\text{TPD}}^{(u)}[n]$, where $n \in [1, T]$. We then compute the cosine similarity between the IPD and TPD embeddings:

$$V^{(u)} = \frac{1}{T} \sum_{n=1}^{T} \left\langle \mathbf{e}_{\text{IPD}}^{(u)}[n], \mathbf{e}_{\text{TPD}}^{(u)}[n] \right\rangle \tag{4}$$

with $\mathbf{e}_{\text{IPD}}^{(u)}[n]$ is defined as: $\begin{bmatrix} \cos(\text{IPD}^{(u)}[n]) \\ \sin(\text{IPD}^{(u)}[n]) \end{bmatrix}$. The spatial embedding is averaged over all microphone pairs $\mathcal{U}$, and the spatial cue $\mathbf{c}_{spa}$ is computed as:

$$\mathbf{c}_{\text{spa}} = \text{Concat}(\text{IPD}, V), \quad V = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} V^{(u)} \tag{5}$$

E. Text Encoder

The semantic query Q_{sem} incorporates two key components: (i) speaker attributes (e.g., gender), and (ii) a vague spatial region (e.g., front-left or rear-right). This query is encoded into a semantic cue $\mathbf{c}_{\text{sem}}$ using a pretrained text encoder—specifically, the Contrastive Language-Audio Pretraining (CLAP) model [27], which is designed to align audio and textual modalities in a shared semantic space:

$$\mathbf{c}_{\text{sem}} = Enc_{\text{text}}(Q_{\text{sem}}) \tag{6}$$

where $Enc_{\text{text}}(\cdot)$ denotes the text encoder.

F. Fusion

The speech feature $\mathbf{y}_t$ is concatenated with the spatial cue $\mathbf{c}_{\text{spa}}$ to yield a fused embedding $\mathbf{x}_t$. To incorporate semantic

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 23,2025 at 07:47:18 UTC from IEEE Xplore. Restrictions apply.

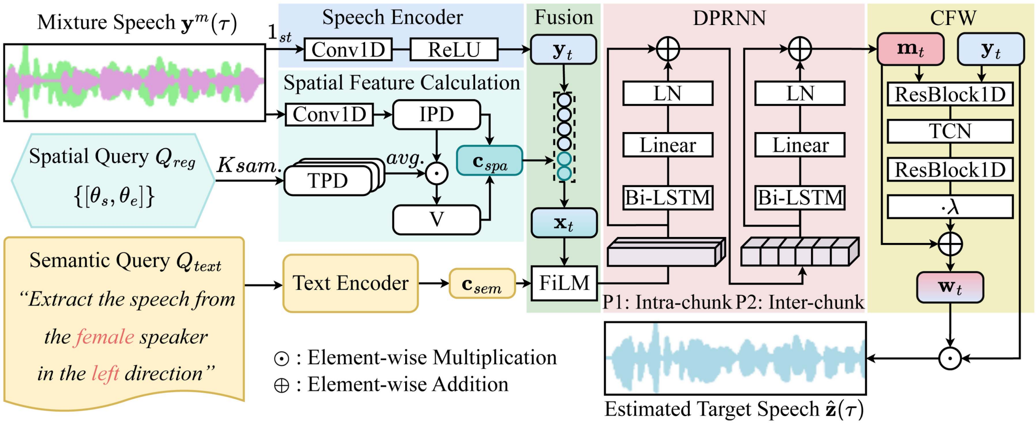

Fig. 2. Architecture of the proposed SSDQ network. The input mixture speech $\mathbf{y}^0(\tau)$ is first encoded into frame-level representations $\mathbf{y}_t$. A spatial query Q_{spa} guides the computation of a spatial cue vector $\mathbf{c}_{\text{spa}}$, while a semantic query Q_{sem} is embedded into a semantic cue $\mathbf{c}_{\text{sem}}$ via a pretrained text encoder. These cues are fused with $\mathbf{y}_t$ using Feature-wise Linear Modulation (FiLM) to produce a query-aware representation. A Dual-Path Recurrent Neural Network (DPRNN) models both intra- and inter-chunk dependencies, and the CFW module refines the output $\mathbf{m}_t$, generating a wrapped feature $\mathbf{w}_t$ that is added to $\mathbf{y}_t$. The final target speech $\hat{\mathbf{z}}(\tau)$ is reconstructed from the fused features.

information, we apply FiLM [28], where the semantic cue $\mathbf{c}_{\text{sem}}$ is used to modulate $\mathbf{x}_t$ through learned scaling and shifting:

$$\text{FiLM}(\mathbf{x}_t, \mathbf{c}_{\text{sem}}) = \gamma(\mathbf{c}_{\text{sem}}) \odot \mathbf{x}_t + \mu(\mathbf{c}_{\text{sem}}) \tag{7}$$

Here, $\gamma(\cdot)$ and $\mu(\cdot)$ are functions that generate scale and bias parameters conditioned on $\mathbf{c}_{\text{sem}}$.

G. Mask Estimator (DPRNN)

The mask $\mathbf{m}_t$ is estimated from the embedding FiLM $(\mathbf{x}_t, \mathbf{c}_{\text{sem}}) \in \mathbb{R}^{T\times D}$. The input is segmented into overlapping chunks and fed into a DPRNN. The intra-chunk stage captures short-term dependencies via a Bi-LSTM, followed by linear projection and Layer Normalization (LN). The inter-chunk stage models long-range temporal structure using another Bi-LSTM with LN.

H. CFW and Speech Decoder

Inspired by [29], we design a Controllable Feature Wrapping (CFW) module that refines an estimated mask $\mathbf{w}_t$ by integrating early speech features from the encoder output $\mathbf{y}_t$ with the estimated mask $\mathbf{m}_t$ which is the output of DPRNN. Structurally, CFW comprises a series of two ResBlock1D layers and a Temporal Convolutional Network (TCN), which together form the transformation function $\phi(\cdot)$. The wrapped feature is computed as:

$$\mathbf{w}_t = \mathbf{m}_t + \phi(\mathbf{y}_t, \mathbf{m}_t) \cdot \lambda \tag{8}$$

where $\lambda \in [0, 1]$ is a tunable scalar of reference features. A higher λ emphasizes speaker identity, while a lower λ prioritizes speech clarity and disentanglement.

The wrapped feature $\mathbf{w}_t$ is then used to generate the masked embedding via element-wise multiplication, and the final target speaker waveform $\hat{\mathbf{z}}(\tau)$ is reconstructed through an overlap-and-add (OnA) operation $\hat{\mathbf{z}}(\tau) = \text{OnA}(\mathbf{w}_t \odot \mathbf{y}_t)$.

III. Experiments and Results

A. SS-Libri Dataset

Existing datasets such as AudioCaps [30] have contributed significantly to audio-language research, but they are limited to single-channel audio and do not provide spatial multi-channel recordings with corresponding textual annotations. Due to the lack of publicly available datasets containing both spatial and semantic information, we generate our own synthetic data. Specifically, we construct a spatialized multi-channel dataset, SS-Libri, by augmenting the single-channel recordings from the LibriSpeech [31] corpus with simulated spatial and directional cues.

1) Synthetic Spatial Mixture: We use the train-clean-100 subset to generate training data, and dev-clean and test-clean subsets for validation and testing, respectively. Room impulse responses (RIRs) are simulated using the gpuRIR [32] toolkit. A circular four-microphone array with 5 cm radius is placed at the center of the room, with cardioid directivity. Using this setup, we generate 14,197, 1,329, and 1,231 multi-channel mixtures for the training, validation, and test sets of SS-Libri.

2) Generated Semantic Query: We leverage text queries that jointly describe speaker characteristics (e.g., gender) and coarse directional information (e.g., angular sector). Each semantic query Q_{sem} encodes these two components in natural language, such as: *"Extract speech with a gender of sex in the direction of sector."* We first manually design a small set of base query templates that cover key combinations of speaker attributes and direction-related expressions. These foundational templates are then expanded using ChatGPT-4o [33] to generate a diverse and natural set of text prompts.

B. Evaluation Metrics

We evaluate models using four widely adopted metrics: Scale-invariant Signal-to-Distortion Ratio improvement (SI-SDRi), Signal-to-Distortion Ratio improvement (SDRi), Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI).

C. Experimental Setup

We train separate models for each type of query input to ensure fair comparisons. During training, we adopt the SI-SDRi as the loss function to optimize the extracted speech quality. The implementation is based on PyTorch 2.0, and all experiments are conducted on a server equipped with two NVIDIA RTX 4090 GPUs. In our setting, the combination of semantic and

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 23,2025 at 07:47:18 UTC from IEEE Xplore. Restrictions apply.

TABLE I
PERFORMANCE EVALUATION OF TSE METHODS WITH DIFFERENT QUERIES ON SS-LIBRI DATASET

Query	Cue	Method	Trainable Params (M)	Runtime (s)	SI-SDRi (dB)↑	SDRi (dB)↑
Semantic Query	Speaker Attributes	AudioSep CLAP [22]	26.4	1.41	1.617	1.902
	Speaker Attributes	CLAPsep [24]	44.3	1.67	7.792	6.770
	Speaker Attributes	Ours (w/o Q_{spa})	3.9	0.58	14.790	15.581
	Vague Region	Ours (w/o Q_{spa})	3.9	0.58	-0.021	-0.013
	Speaker Attributes + Vague Region	Ours (w/o Q_{spa})	3.9	0.60	15.235	16.019
Spatial Query	Precise Region	Rezero [26]	12.3	1.33	9.088	9.678
	Precise Region	Ours (w/o Q_{sem})	3.9	0.62	11.254	12.171
Semantic Query + Spatial Query	Speaker Attributes + Speaker Region	**Ours (DualQuery)**	**3.9**	**0.67**	**19.634**	**20.302**

spatial queries is designed to uniquely identify the target speaker without ambiguity in each scenario. Specifically, there is only one target speaker that matches both the semantic and spatial conditions simultaneously, enabling precise target extraction.

D. Results

Table I presents a comparative evaluation of TSE methods using different query and cue types on the SS-Libri dataset. The evaluation focuses on four dimensions: model capacity (in terms of trainable parameters), runtime efficiency, and separation performance measured by SI-SDRi and SDRi. We define the speaker region as a spatial cue that includes both vague (e.g., "front-left") and precise (e.g., $[22.5°, 67.5°]$) spatial descriptors. Speaker attributes refer to identity-related cues such as gender or role (e.g., "man"). For semantic-only queries, our method achieves a substantial improvement over the state-of-the-art CLAPsep [24], reducing the parameter count from 44.3 M to just 3.9 M, while increasing SI-SDRi from 7.792 dB to 14.790 dB and SDRi from 6.770 dB to 15.581 dB. Notably, using vague regions alone leads to performance collapse (e.g., SI-SDRi: -0.021 dB), highlighting the insufficiency of imprecise spatial information. However, fusing vague region cues with speaker attributes mitigates this degradation and yields an SI-SDRi of 15.235 dB. Spatial-only queries using precise regions also perform well, with our model outperforming Rezero [26] despite having significantly fewer parameters (3.9 M vs. 12.3 M) and achieving higher SI-SDRi (11.254 dB vs. 9.088 dB). Combining both semantic and spatial cues (DualQuery) delivers the best performance, reaching 19.634 dB SI-SDRi and 20.302 dB SDRi, while maintaining a low parameter budget and fast runtime.

E. Ablation Studies

Fig. 3 presents the relationship between input SNR levels and four key evaluation metrics—SI-SDRi, SDRi, PESQ, and STOI—after applying min-max normalization to map all values into the range $[0, 1]$, enabling cross-metric comparison. As the SNR increases from -5 dB to +5 dB, all four metrics reflect consistent improvements in speech extraction quality under progressively cleaner acoustic conditions. Specifically, SI-SDRi and SDRi show the most rapid gains, indicating their high sensitivity to SNR variation. These results confirm that signal-to-noise conditions significantly influence all metrics.

Fig. 4 illustrates the variation of SI-SDRi and SDRi performance metrics under different λ settings in the CFW module. As shown in the figure, setting λ to 0 is equivalent to disabling the CFW module. Both metrics achieve their peak performance when λ is set to 0.75, with SDRi of 20.303 dB and SI-SDRi of 19.634 dB. Compared to the lowest-performing setting at $\lambda = 0$, this configuration improves SI-SDRi by 1.015 dB and SDRi by 0.93 dB. The result suggests that λ setting of 0.75 provides an effective balance between speaker identity and acoustic clarity in the CFW.

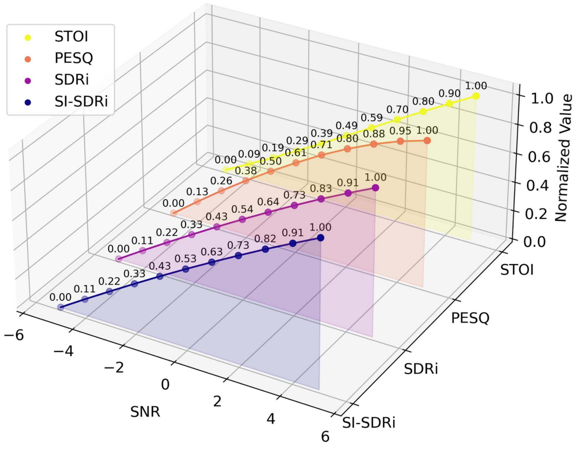

Fig. 3. Normalized metric differences vs. SNR.

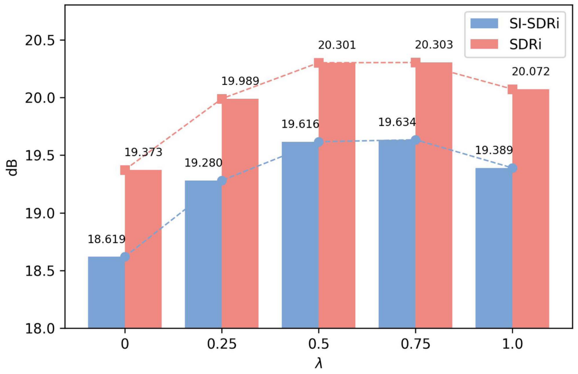

Fig. 4. Effect of λ on SI-SDRi and SDRi in the CFW module.

IV. CONCLUSION AND FUTURE WORK

In this work, we introduce SSDQ, a novel framework that integrates semantic and spatial cues via natural language and region-based queries for multi-channel target speaker extraction. While the current study focuses on fundamental speaker attributes (e.g., gender) and predefined spatial cues, it serves as a first step toward more expressive and flexible query-driven speech extraction. Future work will aim to broaden the semantic query space to include richer speaker characteristics, such as language and speaking style. We also plan to explore more diverse and fine-grained spatial descriptions to improve generalization and applicability in real-world environments.

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 23,2025 at 07:47:18 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proc. Interspeech*, 2013, vol. 2013, pp. 436–440.

[2] A. R. Yuliani, M. F. Amri, E. Suryawati, A. Ramdan, and H. F. Pardede, “Speech enhancement using deep learning methods: A review,” *Jurnal Elektronika dan Telekomunikasi*, vol. 21, no. 1, pp. 19–26, 2021.

[3] X. Gan, Z. Zheng, and Q. Zeng, “Speech enhancement algorithm based on wave-U-Net,” in *Proc. Int. Symp. Comput. Technol. Inf. Sci.*, 2023, pp. 433–437.

[4] D. Michelsanti et al., “An overview of deep-learning-based audio-visual speech enhancement and separation,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 1368–1396, 2021.

[5] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[6] J. Agrawal, M. Gupta, and H. Garg, “A review on speech separation in cocktail party environment: Challenges and approaches,” *Multimedia Tools Appl.*, vol. 82, no. 20, pp. 31035–31067, 2023.

[7] D. Pal, C. Arpnikanondt, and M. A. Razzaque, “Personal information disclosure via voice assistants: The personalization–privacy paradox,” *SN Comput. Sci.*, vol. 1, pp. 1–17, 2020.

[8] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, “I-Vector-based speaker adaptation of deep neural networks for french broadcast audio transcription,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6334–6338.

[9] C. Xu, W. Rao, E. S. Chng, and H. Li, “SpEx: Multi-scale time domain speaker extraction network,” in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, 2020, vol. 28, pp. 1370–1384.

[10] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, “SpEx : A complete time domain speaker extraction network,” in *Proc. Interspeech*, 2020, pp. 1406–1410.

[11] K. Žmolíková et al., “SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures,” *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 4, pp. 800–814, Aug. 2019.

[12] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, “Multi-stage speaker extraction with utterance and frame-level reference signals,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6109–6113.

[13] K. Liu, Z. Du, X. Wan, and H. Zhou, “X-SEPFORMER: End-to-end speaker extraction network with explicit optimization on speaker confusion,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[14] Z. Pan, R. Tao, C. Xu, and H. Li, “Muse: Multi-modal target speaker extraction with visual cues,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6678–6682.

[15] Z. Pan et al., “Scenario-aware audio-visual TF-Gridnet for target speech extraction,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2023, pp. 1–8.

[16] Z. Pan, X. Qian, and H. Li, “Speaker extraction with co-speech gestures cue,” *IEEE Signal Process. Lett.*, vol. 29, pp. 1467–1471, 2022.

[17] X. Qian, Q. Liu, J. Wang, and H. Li, “Three-dimensional speaker localization: Audio-refined visual scaling factor estimation,” *IEEE Signal Process. Lett.*, vol. 28, pp. 1405–1409, 2021.

[18] X. Qian, Z. Pan, Q. Zhang, K. Chen, and S. Lin, “GLMB 3D speaker tracking with video-assisted multi-channel audio optimization functions,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 8100–8104.

[19] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, “L-SpEx: Localized target speaker extraction,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7287–7291.

[20] M. Delcroix et al., “Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 691–695.

[21] J. Heitkaemper, T. Fehér, M. Freitag, and R. Haeb-Umbach, “A study on online source extraction in the presence of changing speaker positions,” in *Proc. Int. Conf. Stat. Lang. Speech Process.*, 2019, pp. 198–209.

[22] X. Liu et al., “Separate anything you describe[j],” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 33, pp. 458–471, 2024.

[23] X. Liu et al., “Separate what you describe: Language-queried audio source separation,” in *Proc. Interspeech*, 2022, pp. 436–440.

[24] H. Ma, Z. Peng, X. Li, M. Shao, X. Wu, and J. Liu, “CLAPSep: Leveraging contrastive pre-trained model for multi-modal query-conditioned target sound extraction,” in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, 2024, pp. 4945–4960.

[25] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, “Multi-modal multi-channel target speech separation,” *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 530–541, Mar. 2020.

[26] R. Gu and Y. Luo, “ReZero: Region-customizable sound extraction,” in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, 2024, pp. 2576–2589.

[27] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, 2023, pp. 1–5.

[28] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 3942–3951.

[29] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, “Exploiting diffusion prior for real-world image super-resolution,” *Int. J. Comput. Vis.*, vol. 132, no. 12, pp. 5929–5949, 2024.

[30] C. D. Kim, B. Kim, H. Lee, and G. Kim, “Audiocaps: Generating captions for audios in the wild,” in *Proc. 2019 Conf. North Amer. Chap. Assoc. Comput. Linguistics: Hum. Lang. Technol., Vol. 1 (Long Short Papers)*, 2019, pp. 119–132.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An asr corpus based on public domain audio books,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[32] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, “gpuRIR: A python library for room impulse response simulation with GPU acceleration,” *Multimedia Tools Appl.*, vol. 80, no. 4, pp. 5653–5671, 2021.

[33] A. Hurst et al., “GPT-4O system card,” 2024, *arXiv:2410.21276*.

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 23,2025 at 07:47:18 UTC from IEEE Xplore. Restrictions apply.