南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

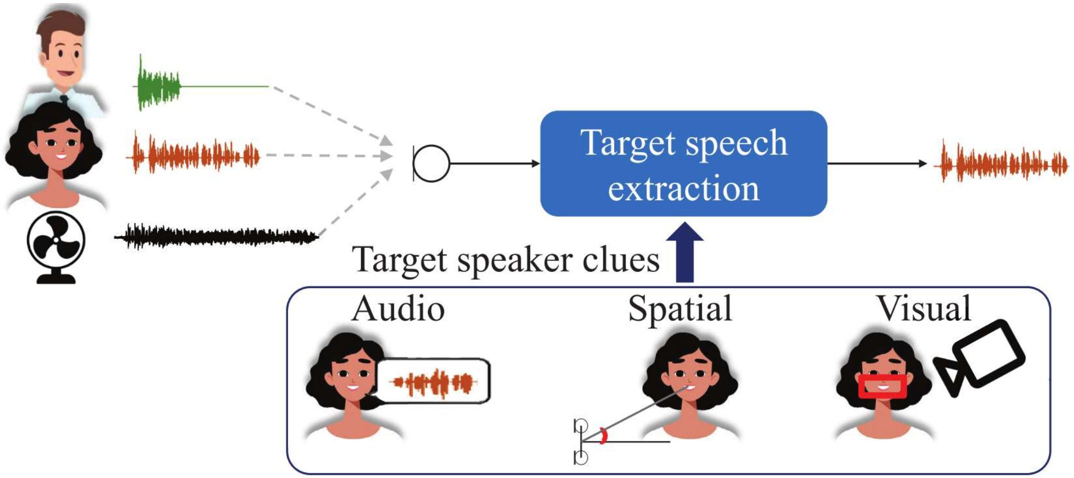
模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Typing to Listen at the Cocktail Party: Text-Guided Target Speaker Extraction

Xiang Hao, Jibin Wu, Jianwei Yu, Chenglin Xu, Kay Chen Tan *Fellow, IEEE*

arxiv 2023

Background



- 鸡尾酒会效应：人类在嘈杂环境中能够集中注意力于某一特定音源（如某个说话人）的能力。这种能力依赖于复杂的听觉处理机制，包括空间、语义和听觉线索的综合。
- 在计算机听觉领域，试图模仿这一效应的任务被称为**目标说话人提取（Target Speaker Extraction，TSE）**。
- 传统TSE方法通常采用以下线索来识别和提取目标说话人：
  - ➢ **声纹**：利用说话人的预录语音生成声学特征作为参考。但这种方法面临以下问题：
    a. 隐私问题：声纹采集需要用户的语音样本，这可能涉及隐私泄露。
    b. 质量和可用性：录音样本的质量（如背景噪声、录音设备的不同）会显著影响提取效果。
    c. 内部变异性：同一说话人在不同条件下（情绪、环境、距离等因素）的声音特征可能有很大差异。
  - ➢ **空间线索**：利用声音的方向或位置（如麦克风阵列）来分离目标说话人。这需要额外的硬件支持，在实际应用中受限。
  - ➢ **视觉线索**：如唇部同步特征，但在纯音频环境中不适用。

Introduction

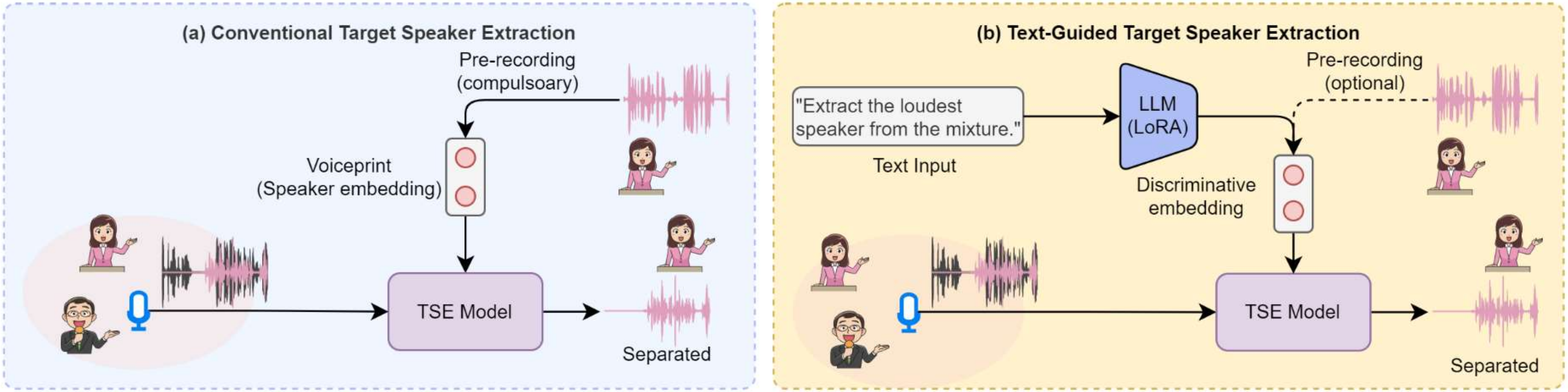


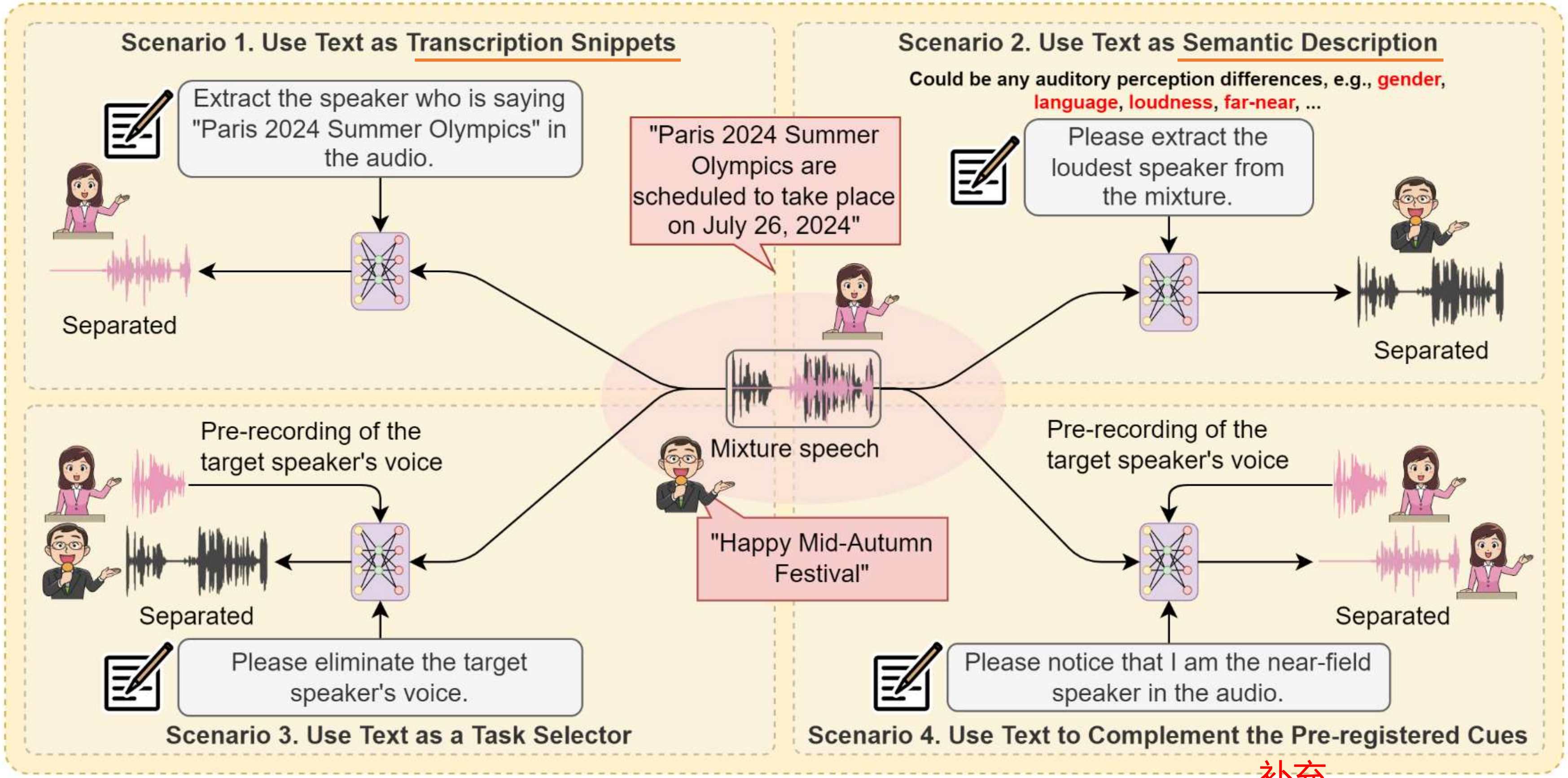
Fig. 1. Comparison between conventional TSE system and our proposed Text-Guided TSE system. The former relies on the pre-registered voiceprint of the target speaker as an extraction cue, while our system offers flexibility to incorporate text-based cues to facilitate target speaker extraction.

- 为了克服上述局限性，论文提出利用文本描述作为目标提取线索的创新方法。其背景如下：
  - 人类描述能力：人类可以通过**语义化描述**（如“提取说‘2024巴黎奥运会’的说话人”）有效区分目标说话人。
  - 大型语言模型（LLM）发展：近年来，基于深度学习的LLM（如LLaMA 2）在自然语言理解任务中表现优异，为将文本与语音任务结合提供了基础。
  - 隐私保护：文本描述通常不包含个人敏感信息，相较声纹线索更具**隐私友好性**。
  - 灵活性和鲁棒性：文本线索可以描述复杂的上下文信息（如“最响亮的声音”或“靠近麦克风的说话人”），提升模型的适应性和稳定性。

Introduction

内容

语义



Scenario 1. Use Text as Transcription Snippets
Extract the speaker who is saying "Paris 2024 Summer Olympics" in the audio.
Separated
Scenario 2. Use Text as Semantic Description
Could be any auditory perception differences, e.g., gender, language, loudness, far-near, ...
Please extract the loudest speaker from the mixture.
Separated
"Paris 2024 Summer Olympics are scheduled to take place on July 26, 2024"
Mixture speech
"Happy Mid-Autumn Festival"
Pre-recording of the target speaker's voice
Separated
Please eliminate the target speaker's voice.
Scenario 3. Use Text as a Task Selector
Pre-recording of the target speaker's voice
Separated
Please notice that I am the near-field speaker in the audio.
Scenario 4. Use Text to Complement the Pre-registered Cues

补充

Fig. 2. New application scenarios enabled by the proposed LLM-TSE model. The central part is a mixture audio sample where two speakers' voices overlap. The male speaker, although positioned at a greater distance from the microphone, has a voice with higher volume and is saying "Happy Mid-Autumn Festival". In contrast, the female speaker is nearer to the microphone but speaks in a quieter tone, delivering the message "Paris 2024 Summer Olympics are scheduled to take place on July 26, 2024". The illustration's four corners show the innovative application scenarios enabled by LLM-TSE.

Method

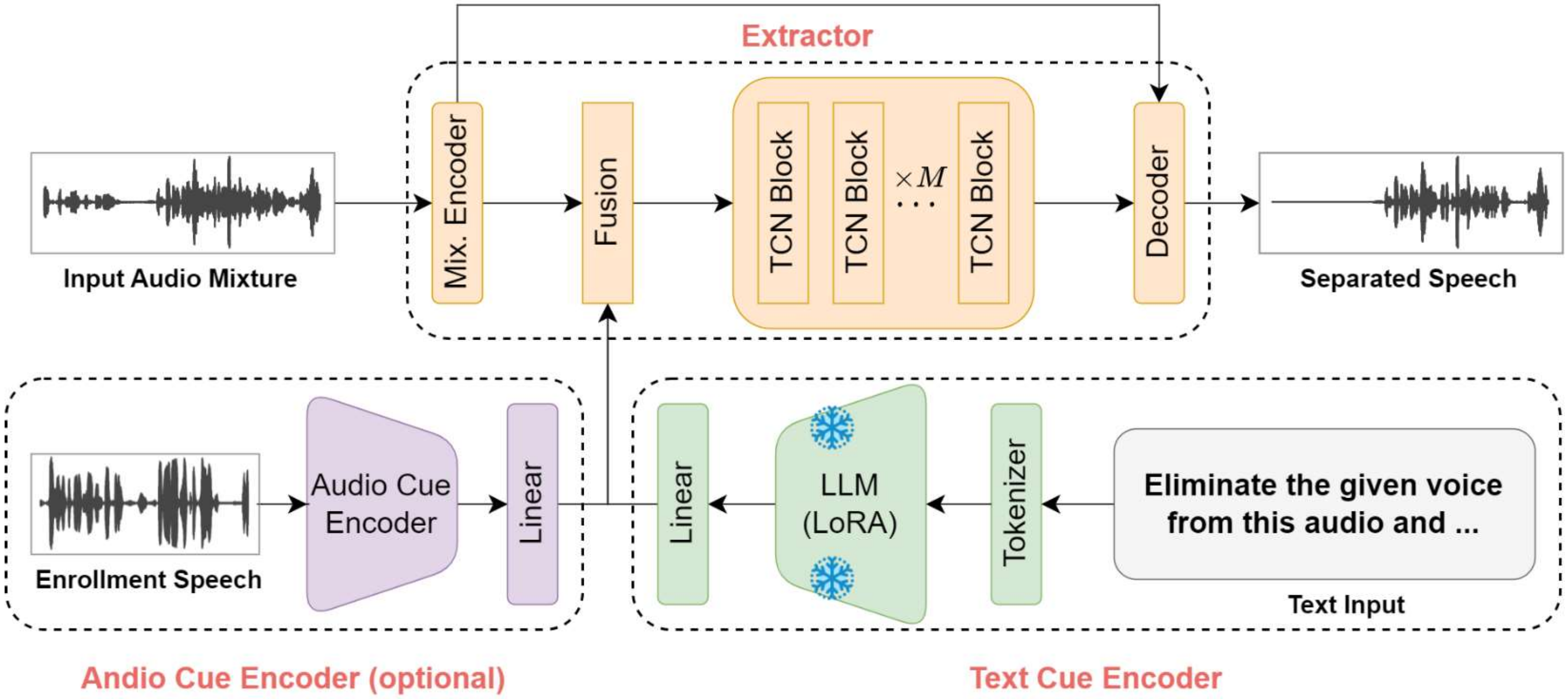


Fig. 3. Overview of the proposed LLM-TSE model architecture. We use LoRA [70] to fine-tune a small number of parameters of the LLM component.

Loss： Scale-Invariant Signal-to-Distortion Ratio （标度不变信号失真比）

$$\mathcal{L}^{\text{SI-SDR}} = -10\log_{10}\left(\frac{\|\frac{\hat{\mathbf{y}}^T\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}\|^2}{\|\frac{\hat{\mathbf{y}}^T\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}-\hat{\mathbf{y}}\|^2}\right).$$

编码-融合-提取-解码

TD-SpeakerBeam： M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara,T. Nakatani, and S. Araki, “Improving speaker discrimination of target speech extraction with time-domain SpeakerBeam,” Jan. 2020, 60 citations (Semantic Scholar/arXiv) [2023-02-14] arXiv:2001.08378 [cs,eess].

Experiments

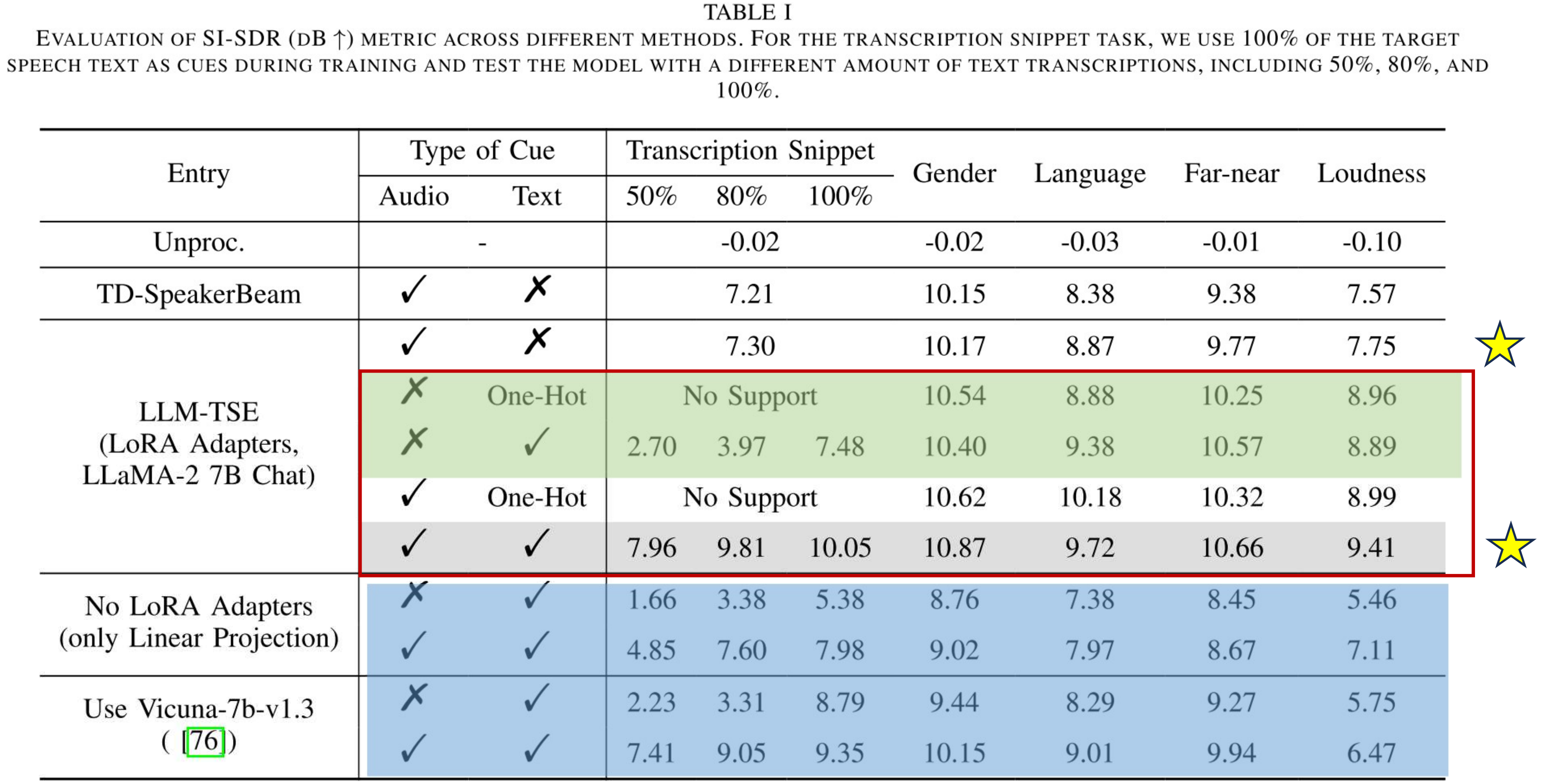


TABLE I

Evaluation of SI-SDR (dB ↑) metric across different methods. For the transcription snippet task, we use 100% of the target speech text as cues during training and test the model with a different amount of text transcriptions, including 50%, 80%, and 100%.

Entry	Type of Cue		Transcription Snippet			Gender	Language	Far-near	Loudness	
	Audio	Text	50%	80%	100%					
Unproc.	-		-0.02			-0.02	-0.03	-0.01	-0.10	
TD-SpeakerBeam	✓	✗	7.21			10.15	8.38	9.38	7.57	
LLM-TSE (LoRA Adapters, LLaMA-2 7B Chat)	✓	✗	7.30			10.17	8.87	9.77	7.75	☆
	✗	One-Hot	No Support			10.54	8.88	10.25	8.96	
	✗	✓	2.70	3.97	7.48	10.40	9.38	10.57	8.89	
	✓	One-Hot	No Support			10.62	10.18	10.32	8.99	
	✓	✓	7.96	9.81	10.05	10.87	9.72	10.66	9.41	☆
No LoRA Adapters (only Linear Projection)	✗	✓	1.66	3.38	5.38	8.76	7.38	8.45	5.46	
	✓	✓	4.85	7.60	7.98	9.02	7.97	8.67	7.11	
Use Vicuna-7b-v1.3 ([76])	✗	✓	2.23	3.31	8.79	9.44	8.29	9.27	5.75	
	✓	✓	7.41	9.05	9.35	10.15	9.01	9.94	6.47	

Efficacy of Using Input Text as Independent Cues 🟩

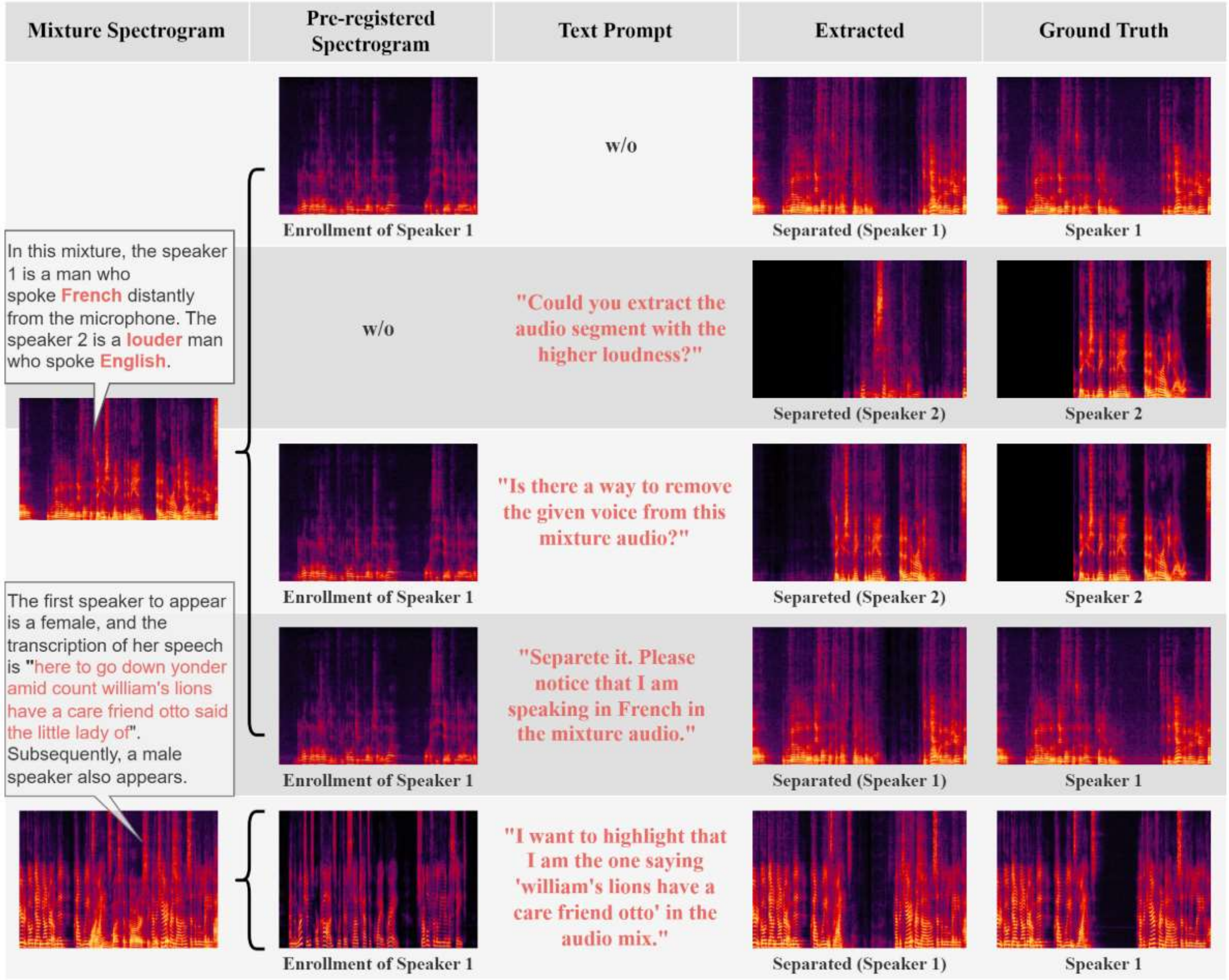
Compared with One-Hot System □

Efficacy of Using Input Text to Complement the Pre-registered Cues ☆

Ablation Studies on Text Encoder Selection 🟦

[76] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin,Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica,“Judging LLM-as-a-judge with MT-Bench and Chatbot Arena,” Jul.2023, arXiv:2306.05685 [cs].

Experiments



In this mixture, the speaker 1 is a man who spoke French distantly from the microphone. The speaker 2 is a louder man who spoke English.

"Could you extract the audio segment with the higher loudness?"

"Is there a way to remove the given voice from this mixture audio?"

"Separete it. Please notice that I am speaking in French in the mixture audio."

The first speaker to appear is a female, and the transcription of her speech is "here to go down yonder amid count william's lions have a care friend otto said the little lady of". Subsequently, a male speaker also appears.

"I want to highlight that I am the one saying 'william's lions have a care friend otto' in the audio mix."

Fig. 4. Samples generated from the proposed LLM-TSE model. The text box contains information about the input audio mixture. The term "w/o" indicates the absence of a certain input.

Efficacy of Using Input Text as Independent Cues

Efficacy of Using Input Text as Task Selector

Efficacy of Using Input Text to Complement the Pre-registered Cues

Thank you!