Advance Publication Cover Page

Acoustical Science and Technology

Language-queried target speech extraction using para-linguistic and non-linguistic prompts

Kentaro Seki, Nobutaka Ito, Kazuki Yamauchi, Yuki Okamoto, Kouei Yamaoka, Yuki Saito,
Shinnosuke Takamichi, Hiroshi Saruwatari

J-STAGE Advance Publication on the web 24 June 2025
doi:10.1250/ast.e25.27

Acoust. Sci. & Tech.

ACOUSTICAL LETTER

Language-queried Target Speech Extraction Using Para-linguistic and Non-linguistic Prompts

Kentaro Seki[1*], Nobutaka Ito[2], Kazuki Yamauchi[1], Yuki Okamoto[1], Kouei Yamaoka[1], Yuki Saito[1], Shinnosuke Takamichi[1,3], Hiroshi Saruwatari[1].

[1] *Graduate School of Information Science and Technology, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 133–8656 Japan*
[2] *Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa-shi, Chiba 277-8561 Japan*
[3] *Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa 223-8522 Japan*

Abstract: This paper proposes a new language-queried target speech extraction (TSE) task called para-linguistic and non-linguistic text prompts-based TSE (PNTP-TSE), which uses text prompts that describe para-linguistic and non-linguistic information. This framework addresses the limitations of conventional TSE methods, such as privacy concerns in voiceprint-based systems and dependency on dedicated microphone arrays or video cameras. To support this framework, we construct and provide a new dataset, PromptTSE, which is specifically designed to facilitate various types of language-queried TSE, including PNTP-TSE. We develop a baseline method for PNTP-TSE and conduct experimental evaluations. The experimental results show that PNTP-TSE overcomes the performance degradation issue of voiceprint-based systems caused by the gap in speaking style between enrollment speech and target speech.

Keywords: Target speech extraction, Language-queried audio source separation, Non-linguistic information, Para-linguistic information

1. Introduction

Target speech extraction (TSE) is a technique of extracting only a specific target speaker's speech in possibly noisy, multi-speaker environments. This technique is crucial as a front-end for various applications, such as automatic speech recognition, hearing aids, and spoken dialogue systems [1–3]. TSE faces the challenges of identifying and extracting the target speech from multiple overlapping voices.

Conventional TSE systems have advanced by leveraging voiceprints [1,2], spatial clues [4], and visual clues [5] to identify the target speaker. Voiceprint-based methods do not require additional equipment, but they pose privacy concerns and degrade under variability in speaking style, such as emotion and intonation [6]. Spatial-clue-based methods require tracking for moving speakers using a microphone array, and visual-clue-based methods use a video camera and degrade in poor lighting. These limitations underscore the need for flexible, robust TSE systems suitable for real-world use.

Recently, in the context of sound source separation, a language-queried audio source separation (LASS) framework has emerged as an innovative approach to extracting audio sources using language queries. Unlike a conventional label-based framework constrained by predefined categories, LASS is gaining attention for its intuitive, versatile, and flexible framework [7]. The LASS framework offers notable advantages in TSE, addressing privacy concerns and requirement of dedicated microphone arrays or video cameras associated with conventional TSE approaches. However, conventional LASS methods use generic prompts such as "speech" [7, 8], which severely limits the application range of TSE. Thus, the capability of LASS to identify individual speakers in mixture, a key aspect of TSE, remains largely unexplored.

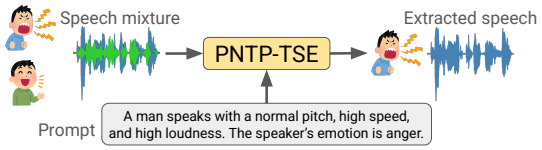

Fig. 1 Overview of PNTP-TSE.

Recently, a few studies explored extensions of LASS to language-queried TSE. Speech contains linguistic, para-linguistic, and non-linguistic information [9], and these studies have focused on specific aspects of this information as clues. Prior studies have primarily focused on either linguistic or para-linguistic information, while the use of non-linguistic information has been limited to gender [10–13]. However, linguistic information can be unreliable in noisy environments where transcription becomes difficult. Para-linguistic information, while useful, is often shared among different speakers and may vary over time, limiting its effectiveness as a consistent clue. In contrast, non-linguistic information that represents speaker identity is less likely to be shared by different speakers and relatively time-invariant, making it advantageous for TSE.

In this study, we propose a novel task, *para-linguistic and non-linguistic text prompts-based TSE (PNTP-TSE)*, which leverages text prompts describing para-linguistic and non-linguistic information, as shown in Figure 1. Since these types of information are less likely to be masked in noisy environments compared

* seki-kentaro922@g.ecc.u-tokyo.ac.jp

with linguistic content [14], PNTP-TSE is effective for noisy environments. Moreover, PNTP-TSE leverages non-linguistic information, making it effective even when overlapping speakers share similar paralinguistic information including speaking styles. PNTP-TSE suits post-production, where sound engineers can craft prompts after listening to the mixture. In addition, we construct and release *PromptTSE*, a novel dataset supporting comprehensive language-queried TSE*. Furthermore, we develop a baseline method for PNTP-TSE. Experimental results demonstrate that our baseline method significantly outperforms voiceprint-based methods.

2. PNTP-TSE: Task description

The observed signal $y[t] \in \mathbb{R}$ is represented as follows:

$$y[t] = x_s[t] + \sum_{i \neq s, i \in \mathcal{S}} x_i[t] + n[t], \tag{1}$$

where $y[t]$, $x_s[t]$, $x_i[t]$, and $n[t] \in \mathbb{R}$ denote the mixture, the target speech, an interfering speech, and the noise, respectively. Here, t represents the discrete-time index, s is the index of the target speaker, i indexes the speakers, and $\mathcal{S}$ is the set of the speaker indices in the mixture. Note that $\mathcal{S}$ is treated as unknown, including its size.

The TSE problem involves extracting the target speech from $y[t]$ using a given clue, $\boldsymbol{c}_s$, as follows:

$$\hat{x}_s[t] = \mathrm{TSE}(y[t], \boldsymbol{c}_s; \boldsymbol{\theta}), \tag{2}$$

where $\hat{x}_s[t]$ is an estimate of the target speech, and $\mathrm{TSE}(\cdot; \boldsymbol{\theta})$ represents a TSE system parameterized by $\boldsymbol{\theta}$. The clue $\boldsymbol{c}_s$ corresponds to the text prompts provided by the user, enabling the target speech to be identified.

3. PromptTSE: Corpus description

3.1. Overview

We construct and provide the PromptTSE dataset, which consists of tuples of mixture, target clean speech, and prompts that describe para-linguistic and non-linguistic information about the target speech. Specifically, these prompts describe speaker identity, emotional expressions, and speaking styles. Additionally, the dataset contains enrollment speech and transcriptions, which makes it useful for TSE methods based on voiceprints and linguistic information as well.

3.2. Original speech datasets

The original speech data used to construct our dataset were from LibriTTS [15] and EARS [16]. Speech data shorter than 5 seconds were excluded, and speakers with only one utterance were removed, ensuring at least two utterances per speaker.

LibriTTS: LibriTTS [15] is a dataset derived from audiobooks for text-to-speech synthesis, characterized by segmentation into utterances and the removal of too noisy or long utterances. Each utterance is paired with a speaker ID and transcription. To ensure the target speech remains acoustically clean, we utilized only the clean subset of LibriTTS, consisting of utterances with a signal-to-noise ratio (SNR) of at least 20 dB.

EARS: Since LibriTTS primarily consists of speech with neutral emotional expressions whose diversity is limited, we incorporated emotional speech data from EARS [16]. EARS is a high-quality dataset recorded in an anechoic chamber for speech enhancement and dereverberation. We selected emotional utterances with transcriptions†. These utterances are paired with transcriptions, speaker IDs, and emotion labels, which cover 22 emotion categories such as anger and sadness. Each emotion label appears with equal frequency. There is one utterance for each combination of speaker and emotion, ensuring a balanced distribution of emotions. For the class distribution of other paralinguistic information, please refer to Section 3.4.

Table 1 Overall characteristics of PromptTSE for each subset, shown both individually for each corpus and combined across corpora. The partitioning follows that of target clean utterances.

		Train	Val	Test	**All**
Total duration in hours	LibriTTS	185.34	6.62	6.80	198.75
	EARS	7.23	0.14	0.47	7.83
	Sum	192.57	6.76	7.27	206.58
Number of speakers	LibriTTS	1,086	37	37	1,160
	EARS	99	2	6	107
	Sum	1,185	39	43	1,267
Number of utterances	LibriTTS	68,876	2,560	2,365	73,801
	EARS	2,275	46	138	2,459
	Sum	71,151	2,606	2,503	76,260

3.3. Mixtures and target clean speech

Mixtures were created by selecting two utterances from different speakers from the original datasets as the target and interfering speech, and mixing them at a randomly determined SNR within the range of $[-5, 5]$ dB, and we did not add any noise other than the interfering speech. For each original speech utterance, exactly one mixture was created, using that original utterance as the target speech. Utterances were allowed to be paired both within the same corpus and across different corpora, enabling combinations such as a target utterance from LibriTTS and an interfering one from EARS. Additionally, the mixture samples were divided into training, validation, and test subsets according to the partitions of the original corpus. These subsets were designed to ensure no speaker overlap among the training, validation, and test subsets. Table 1 summarizes the overall characteristics of the PromptTSE dataset.

3.4. Prompts

As exemplified in Table 2, we generated prompts in the following format: "A [subject] speaks with a [pitch-class] pitch, [speed-class] speed, and [loudness-class] loudness. [Additional-information]."

Para-linguistic information: As the physical at-

* `https://github.com/sarulab-speech/PromptTSE`

† Their filenames start with "emo_[emotion-label]_sentences."

Table 2 Examples of prompts assigned in our corpus. The utterance ID is a newly assigned one in PromptTSE.

Utterance ID	Original corpus	Text description
train_023143	LibriTTS	A man speaks with a low pitch, very low speed, and very low loudness. The speaker's identity can be described as very masculine, very adult-like, slightly thick, slightly muffled, cool, sincere, slightly kind.
train_006880	EARS	A woman speaks with a normal pitch, normal speed, and normal loudness. The speaker's emotion is fear.

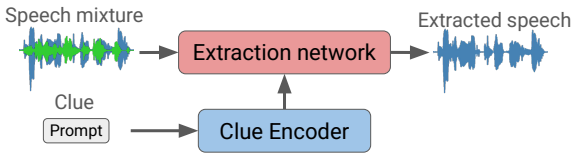

Fig. 2 Baseline model for PNTP-TSE.

tributes of the speaking style, we adopted textual descriptions of speaking speed, pitch, and loudness. For each target clean speech utterance, we calculated pitch (average F0), speaking speed (syllables per second), and loudness (loudness units relative to full scale [17]). The calculated values were then classified into five categories: very low, low, middle, high, and very high. These categories corresponded to the bottom 10%, the next 20%, the middle 40%, the next 20%, and the top 10% of the values. Note that while LibriTTS-P [18] provided speaking style prompts for LibriTTS [15], we reassigned these prompts because incorporating EARS [16] altered the distribution of class assignments.

Non-linguistic Information: The "[subject]" in the prompt was set as "woman" or "man" on the basis of the gender tag (male or female), and "person" for utterances without gender tags.

Additionally, LibriTTS-P contains textual descriptions related to speaker identity, which we also incorporated as additional information. LibriTTS-P includes prompts annotated by three expert annotators using over 40 impression words, such as "calm" and "fluent," along with intensity levels (e.g., "slightly," "very"). This annotation framework allows for a broad and subjective representation of speaker identity. We randomly selected one speaker identity prompt for each utterance.

Furthermore, for speech from EARS, we generated prompts using emotion labels as para-linguistic information in the following format: "The speaker's emotion is [emotion label]." This served as the "[Additional-information]" in the prompt format described earlier.

3.5. Additional clue assignments

For voiceprint-based TSE, each mixture utterance was assigned a different clean speech utterance spoken by the same target speaker as enrollment speech, which had an average duration of 9.76 seconds. Additionally, for transcription-based TSE, the transcriptions from the original speech corpora were provided.

4. Baseline method

Based on prior LASS-based source separation methods [9, 15], our baseline system adopted the same architecture, which consisted of a clue encoder and a separation network. As shown in Figure 2, this baseline method consists of two main components: a clue encoder and an extraction network.

The clue encoder employs a pretrained contrastive language-audio pretraining (CLAP) model [19][‡] to extract a fixed-dimensional embedding $\boldsymbol{z}$ from the variable-length natural language prompt $\boldsymbol{c}_s$. During training, its parameters were fixed.

The separation network is ResUNet30 [13], which is the ResUNet [20] backbone with 30 layers. This network extracts the target speaker by conditioning on the embedding $\boldsymbol{z}$ provided by the clue encoder.

5. Experiments

5.1. Experimental conditions

Compared methods: We conducted experiments by varying the conditioning $\boldsymbol{z}$ of the baseline method's separation network. Our experiments aimed to compare performance across different clue types, rather than attain state-of-the-art results. For all methods, the clue encoder was fixed during training.

To compare our approach with existing voiceprint-based ones, we replaced the prompt encoder with a speaker embedding extractor using audio clues. In the "Voiceprint" condition, an enrollment utterance from the same speaker was input into a pretrained WavLM-ForXVector[§] to extract an x-vector [21]. In the "Oracle" condition, the target clean speech itself was used to generate the x-vector. In both cases, the extracted x-vector replaced the prompt embedding to condition the separation network.

For language-queried TSE, we used the method described in Section 4 and varied the types of prompts. "PNTP-TSE" utilized the prompts constructed in Section 3.4. As an ablation study, we evaluated prompts in which descriptions of gender, pitch, speed, or loudness were removed from "PNTP-TSE," referring to them as "PNTP-TSE w/o gender," "PNTP-TSE w/o pitch," "PNTP-TSE w/o speed," and "PNTP-TSE w/o loudness," respectively. Additionally, we examined prompts that contained only a single attribute—gender, pitch, speed, or loudness—naming them "Gender," "Pitch," "Speed," and "Loudness," respectively. Under the "Rephrased" condition, the same model as "PNTP-TSE" was evaluated with rephrased prompts for the test set to investigate the model's sensitivity to variations in prompt expressions. These rephrased prompts were generated using ChatGPT [22] and are included in

[‡] https://huggingface.co/lukewys/laion_clap/blob/main/music_speech_audioset_epoch_15_esc_89.98.pt

[§] https://huggingface.co/microsoft/wavlm-base-plus-sv

Table 3 Performance comparison of TSE methods with different clues on the test set of PromptTSE. The values are presented as the mean with standard deviation. The values highlighted in bold are the largest mean for each modality.

Clue	Clue modality	SI-SDR (dB)	PESQ	ESTOI	DNSMOS
Speech mixture	-	-0.54 ± 3.01	1.16 ± 0.12	0.571 ± 0.11	2.65 ± 0.32
Voiceprint	Audio	6.75 ± 8.31	1.45 ± 0.33	0.688 ± 0.19	2.98 ± 0.27
Oracle	Audio	$\mathbf{8.94 \pm 6.14}$	$\mathbf{1.54 \pm 0.35}$	$\mathbf{0.735 \pm 0.14}$	$\mathbf{3.02 \pm 0.26}$
Gender	Language	3.25 ± 11.25	1.41 ± 0.36	0.616 ± 0.24	2.96 ± 0.28
Pitch	Language	4.55 ± 10.42	1.42 ± 0.35	0.650 ± 0.22	2.96 ± 0.28
Speed	Language	-1.30 ± 11.23	1.27 ± 0.28	0.532 ± 0.23	2.88 ± 0.29
Loudness	Language	4.12 ± 10.63	1.40 ± 0.32	0.636 ± 0.22	2.97 ± 0.27
PNTP-TSE w/o gender	Language	$\mathbf{8.43 \pm 6.87}$	1.51 ± 0.35	0.719 ± 0.16	3.00 ± 0.27
PNTP-TSE w/o pitch	Language	7.19 ± 8.00	1.47 ± 0.35	0.699 ± 0.18	2.98 ± 0.28
PNTP-TSE w/o speed	Language	8.29 ± 6.42	1.49 ± 0.33	0.719 ± 0.15	2.99 ± 0.27
PNTP-TSE w/o loudness	Language	7.19 ± 8.00	1.47 ± 0.35	0.699 ± 0.18	2.98 ± 0.28
Rephrased	Language	5.62 ± 9.56	1.43 ± 0.33	0.670 ± 0.21	2.99 ± 0.27
PNTP-TSE	Language	8.38 ± 6.97	$\mathbf{1.52 \pm 0.35}$	$\mathbf{0.720 \pm 0.16}$	$\mathbf{3.01 \pm 0.27}$

the PromptTSE dataset.

Model and training: As described in Section 4, we adopted ResUNet30 [13] as the separation network. The input and output were monaural audio sampled at 16 kHz, standardized their duration to 10 seconds. For all methods, the clue embedding $\boldsymbol{z}$ was a 512-dimensional vector. The clue encoder was fixed, whereas ResUNet30 was trained from scratch. Training was performed using the training subset of our corpus. We utilized the AdamW optimizer [23] with an initial learning rate of 0.001, a 10,000-step warm-up, 100 epochs, and a batch size of 16. A waveform-based L1 loss between the target and the extracted speech was employed.

Evaluation: The trained models were evaluated on the test subset of the corpus under the same conditions as during training. We used the following metrics: scale-invariant signal-to-distortion ratio (SI-SDR) [24], perceptual evaluation of speech quality (PESQ) [25], extended short-time objective intelligibility (ESTOI) [26], and deep noise suppression mean opinion score (DNSMOS) [27].

5.2. Results

Comparison with voiceprint-based TSE: Table 3 shows the results. Although "PNTP-TSE" did not surpass "Oracle," it outperformed "Voiceprint" across all metrics on average. This is likely because the performance of voiceprint-based TSE degrades with variations in the speaking style of each utterance [6]. To verify this hypothesis, we examined the matching rates of pitch and speed classes between the target and the enrollment speech. The pitch class matched in 66.8% of cases, the speed class matched in 36.9% of cases, and both classes matched in only 24.9% of cases. These results indicate that, in practice, the speaking style varies across utterances. Since oracle clean speech is not available in practical scenarios, the fact that "PNTP-TSE" outperformed "Voiceprint" demonstrates its practical applicability in real-world situations. The relatively small SI-SDR improvement of "PNTP-TSE" compared to previous studies [10–12] is likely due to the limited amount of training data.

Ablation study: "Gender," "Pitch," "Speed," and "Loudness" all resulted in a significantly lower performance than "PNTP-TSE" on average, indicating that using only one of these attributes is insufficient for uniquely identifying speakers. In particular, "Speed" achieves a lower SI-SDR than "Speech mixture," indicating a failure to identify the target speech. This can be attributed to the difficulty of modeling speed, which is a sequential feature rather than a frame-level one. Additionally, the results show that "Gender," having only two classes, was less effective for speaker identification than "Pitch" or "Loudness."

On the other hand, "PNTP-TSE w/o Speed" performed worse than "PNTP-TSE," suggesting that speed is useful as auxiliary information. Furthermore, "PNTP-TSE w/o Gender" achieved nearly the same results as "PNTP-TSE." This indicates that gender information can be substituted with other factors such as speaker identity (e.g., descriptions such as "a masculine voice") or pitch. "Rephrased" underperformed "PNTP-TSE," suggesting the need for diverse phrasings during training to improve robustness to domain shifts in textual clue.

6. Conclusion

We proposed a novel task called PNTP-TSE, which is LASS-based TSE utilizing para-linguistic and non-linguistic textual descriptions as clues. We defined this task and provided the baseline models and the PromptTSE dataset. Through experimental validation, we demonstrated that PNTP-TSE overcomes the limitations of existing voiceprint-based TSE approaches under our experimental conditions. PNTP-TSE is expected to degrade in performance when the prompt contains incorrect or misleading information, which remains a limitation of the current study. Future work thus includes investigating the robustness of PNTP-TSE to errors in prompt descriptions, as well as conducting ablation studies on emotional clues and speaker identity attributes.

Acknowledgement

This work was supported by JSPS KAKENHI 24KJ0860 and Research Grant S of the Tateisi Science and Technology Foundation.

References

[1] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 2728–2732.

[2] 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.

[3] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černockỳ, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[4] 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 *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 4290–4294.

[5] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 3244–3248.

[6] Z. Zhao, D. Yang, R. Gu, H. Zhang, and Y. Zou, "Target confusion in end-to-end speaker extraction: Analysis and approaches," in *Proc. INTERSPEECH*, Incheon, Korea, 2022, pp. 5333–5337.

[7] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate what you describe: Language-queried audio source separation," in *Proc. INTERSPEECH*, Incheon, Korea, pp. 1801–1805.

[8] H.-W. Dong, N. Takahashi, Y. Mitsufuji, J. McAuley, and T. Berg-Kirkpatrick, "CLIPSep: Learning text-queried sound separation with noisy unlabeled videos," in *Proc. ICLR*, 2023.

[9] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in *Proc. Speech Prosody – International Conference on Speech Prosody 2004*, Nara, Japan, 2004.

[10] Z. Jiang, X. Qian, J. Lei, Z. Pan, W. Xue, and X.-c. Yin, "pTSE-T: Presentation target speaker extraction using unaligned text cues," *arXiv preprint arXiv:2411.03109*, 2024.

[11] X. Hao, J. Wu, J. Yu, C. Xu, and K. C. Tan, "Typing to listen at the cocktail party: Text-guided target speaker extraction," *arXiv preprint arXiv:2310.07284*, 2023.

[12] M. Huo, A. Jain, C. P. Huynh, F. Kong, P. Wang, Z. Liu, and V. Bhat, "Beyond speaker identity: Text guided target speech extraction," *arXiv preprint arXiv:2501.09169*, 2025.

[13] D. de Oliveira, E. Grinstein, P. A. Naylor, and T. Gerkmann, "LASER: Language-queried speech enhancer," in *Proc. IWAENC*, Aalborg, Denmark, 2024, pp. 90–94.

[14] M. K. Zhi Zhu and M. Unoki, "Study on the perception of nonlinguistic information of noise-vocoded speech under noise and/or reverberation conditions," *Acoustic Science and Technology*, vol. 76, no. 6, pp. 317–326, 2020.

[15] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1526–1530.

[16] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Proc. INTERSPEECH*, Kos, Greek, 2024, pp. 4873–4877.

[17] B. Series, "Algorithms to measure audio programme loudness and true-peak audio level," in *International Telecommunication Union Radiocommunication Assembly*, 2011.

[18] M. Kawamura, R. Yamamoto, Y. Shirahata, T. Hasumi, and K. Tachibana, "LibriTTS-P: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning," in *Proc. INTERSPEECH*, Kos, Greek, 2024, pp. 1850–1854.

[19] 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. ICASSP*. Rhodes, Greek: IEEE, 2023, pp. 1–5.

[20] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep ResUNet for music source separation," in *Proc. ISMIR*, Online, 2021, pp. 342–349.

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, Calgary, AB, Canada, 2018, pp. 5329–5333.

[22] OpenAI, "ChatGPT 4o," https://chat.openai.com, 2025, accessed: 2025-05-16.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, New Orleans, Louisiana, United States, 2019.

[24] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR — Half-baked or Well done?" in *Proc. ICASSP*, Brighton, United Kingdom, 2019, pp. 626–630.

[25] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, Salt Lake City, Utah, United States, 2001, pp. 749–752.

[26] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[27] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*. Online: IEEE, 2021, pp. 6493–6497.