パラ言語・非言語情報の記述文をクエリとした目的音声抽出 *

◎関 健太郎，伊藤信貴，山内一輝，岡本悠希，山岡洸瑛，齋藤佑樹（東大），
高道慎之介（東大／慶大），猿渡洋（東大）

1 はじめに

目的音声抽出（Target speech extraction: TSE）は，雑音や複数話者が同時に話す環境において，特定の話者の音声のみを抽出する技術であり，音声認識，補聴器，音声対話システムなど，さまざまな用途に応用される [1–3]．TSE では，複数の重なった音声の中から，対象となる目的音声を識別して抽出するという課題が存在する．

従来の TSE システムは，目的音声を指定するための手がかりとして，登録音声から抽出した声紋や [1,2]，空間的手がかり [4]，視覚的手がかり [5] などを活用してきた [3]．声紋を用いる手法は目的音声と同一話者の音声を事前に登録し，その音声特徴に基づいて目的音声を識別する．この方法は追加機器を必要としないが，プライバシーに関する懸念がある上，登録音声と目的音声の間に感情や抑揚といった発話スタイルの違いが存在する場合に性能が劣化する [6]．空間的手がかりに基づく手法は，マイクロフォンアレイを用いた話者の追跡を必要とする．視覚的手がかりを用いる手法は，ビデオカメラを用いるが，照明条件が悪い環境では性能が劣化する．これらの限界は，現実世界で利用可能な柔軟かつ頑健な TSE システムの必要性を示している．

近年，音源分離の文脈において，言語クエリによる音源分離（LASS: Language-queried Audio Source Separation）という，言語クエリを用いて音源を抽出する新たな枠組みが検討されている．従来のカテゴリラベルに基づく枠組みとは異なり，LASS は直感的かつ柔軟性に富んだ枠組みとして注目されている [7]．この LASS の枠組みは，従来の TSE 手法が抱えるプライバシーの問題や，マイクロフォンアレイやビデオカメラといった専用機器の必要性を解消する利点を有している．しかし，従来の LASS 手法では “speech（音声）” のような一般的な記述文を採用しているため [7]，混合音声の中で目的音声を識別できず，TSE への適用が困難である．

近年，LASS の枠組みに基づいて，言語クエリによる TSE の実現を目指す研究が進められている．音声に含まれる情報は言語情報，パラ言語情報，非言語情報に分類される [8]．既存の言語クエリを用いた TSE 手法は言語的情報に焦点を当てており，パラ言語・非言語情報の利用は限られた特徴のみに注目していた [9–12]．しかし，言語情報は騒音下では認識が困難になるため，信頼性が低下する．また，パラ言語情報は異なる話者間で共通しやすいため，目的音声を指定する安定した手がかりとはならない．これに対し，話者の個性を表す非言語情報は，他者と識別しやすく，時間的にも比較的安定しているため，TSE において有利であると考えられる．

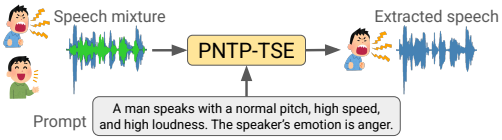

Fig. 1: Overview of PNTP-TSE.

本研究では，Fig. 1 に示すように，パラ言語・非言語情報の記述文をクエリとして利用する新たな TSE タスクである Para-linguistic and Non-linguistic Text Prompts-based TSE (PNTP-TSE) を提案する．これらの情報は言語情報に比べて雑音環境下でも埋もれにくいため [13], PNTP-TSE は雑音環境下でも有効と期待される．さらに，PNTP-TSE は非言語情報を活用することで，発話スタイルのようなパラ言語情報が似ている音声が混合している状況においても効果的である．また，PNTP-TSE はポストプロダクション（事後編集）に適しており，音響エンジニアが混合音を確認してプロンプトを作成することが可能である．加えて，本研究では PNTP-TSE を含む多様な言語クエリ TSE を支援するための新たなデータセット “PromptTSE” を構築・公開する．さらに，PNTP-TSE におけるベースライン手法を開発し，実験評価を行った．実験の結果，PNTP-TSE は，本研究の実験条件において声紋に基づく TSE を上回る性能を達成することが示された．

2 PNTP-TSE: タスク定義

観測信号 $y[t] \in \mathbb{R}$ は，以下のように表される：

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

ここで，$y[t]$，$x_s[t]$，$x_i[t]$，および $n[t] \in \mathbb{R}$ はそれぞれ，混合音声，目的音声，干渉音声，雑音を表す．t は離散時間インデックス，s は目的音声の話者インデックス，i は話者のインデックス，$\mathcal{S}$ は混合に含まれる話者のインデックス集合である．なお，$\mathcal{S}$ はそのサイズも含めて未知であるものとする．

TSE タスクは所与の手がかり $\boldsymbol{c}_s$ を用いて $y[t]$ から目的音声を抽出する課題であり，以下のように定式化される：

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

* Target Speech Extraction Based on Para-linguistic and Non-linguistic Descriptions, by Kentaro Seki, Nobutaka Ito, Kazuki Yamauchi, Yuki Okamoto, Kouei Yamaoka, Yuki Saito (The University of Tokyo), Shinnosuke Takamichi (The University of Tokyo/Keio University), Hiroshi Saruwatari (The University of Tokyo).

ここで，$\hat{x}_s[t]$ は目的音声の推定値，TSE$(\cdot;\boldsymbol{\theta})$ はパラメータ $\boldsymbol{\theta}$ により定義される TSE システムを表す．手がかり $\boldsymbol{c}_s$ は混合音声中から目的音声を識別可能にする情報であり，ユーザが提供するパラ言語・非言語情報の記述文に対応する．

3 PromptTSE: コーパスの説明

3.1 概要

我々は，目的音声に関するパラ言語・非言語情報の記述文と，混合音声および目的音声の組を構成要素とするデータセット “PromptTSE” を構築・提供する．記述文はパラ言語情報（発話スタイル・感情）と非言語情報（性別・話者性）を記述する．さらに，本データセットは声紋に基づく TSE 及び言語情報に基づく TSE で利用するために，登録音声および書き起こしが付与されている．

3.2 元となる音声データセット

PromptTSE の構築に使用した音声データは，LibriTTS [14] および EARS [15] に由来する．5 秒未満の音声データは除外し，1 発話しか持たない話者も除外した．これにより，すべての話者に少なくとも 2 発話を含めるようにした．

3.2.1 LibriTTS

LibriTTS はテキスト音声合成のためにオーディオブックから構成されたデータセットであり，発話単位に分割されており，各発話には話者 ID および書き起こしが付与されている．目的音声をノイズが少ないものに限定するため，LibriTTS の clean サブセットを利用し，信号対雑音比（Signal-to-noise ratio: SNR）が 20 dB 以上の音声のみを用いた．

3.2.2 EARS

LibriTTS は中立的な感情表現を中心としているため，EARS の感情音声データを用いることで感情表現の多様性を確保した．EARS は，音声強調および残響除去のために無響室で収録された高品質なデータセットである．我々は，EARS に含まれる音声データのうち，感情ラベルが付与された書き起こし付き発話のみを目的音声として利用した．各発話には書き起こし，話者 ID，22 種類の感情ラベルが付与されている．各話者・感情の組み合わせごとに 1 発話が存在し，感情分布が均等になるよう設計されている．

3.3 混合音声と目的音声

混合音声は，3.2 節で述べたデータの中から異なる 2 発話の発話をそれぞれ目的音声および干渉音声として選定し，それらを $[-5, 5]$ dB の範囲でランダムに決定された SNR で混合することで作成した．干渉音声以外の雑音は追加していない．各発話に対して，その発話を目的音声とする混合音声が 1 つ作成された．目的音声と干渉音声の組み合わせは同一コーパス内または異なるコーパス間のどちらでも可能とし，例えば目的音声が LibriTTS，干渉音声が EARS といった組み合わせも含まれる．また，混合音声サンプルは元のコーパスにおける分割に従って Train/Val/Test のサブセットに分割されており，各サブセット間に話者の重複は存在しない．PromptTSE の全体的な特性（総時間，話者数，発話数）は，Table 1 に要約されている．

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	1086	37	37	1160
	EARS	99	2	6	107
	Sum	1185	39	43	1267
Number of utterances	LibriTTS	68 876	2560	2365	73 801
	EARS	2275	46	138	2459
	Sum	71 151	2606	2503	76 260

3.4 プロンプト

Table 2 に示すように，記述文は以下のフォーマットで作成された：“A [subject] speaks with a [pitch-class] pitch, [speed-class] speed, and [loudness-class] loudness. [Additional-information]”．

パラ言語情報: 発話スタイルの記述文として，ピッチ，話速，音量に関するテキスト記述を採用した．各目的音声について，ピッチ（平均 F0），話速（1 秒あたりの音節数），音量（フルスケール基準のラウドネス単位）を算出した．これらの値は，それぞれ “very low”, “low”, “middle”, “high”, “very high” の 5 つのカテゴリに分類した．これらのカテゴリは，値の下位 10%，次の 20%，中央の 40%，次の 20%，および上位 10%に対応している．なお，LibriTTS-P [16] は LibriTTS に対して発話スタイルの記述文を提供しているが，EARS を組み込むことでクラス分布が変化したため，本研究では EARS を含めた新しい全体分布に基づく再割り当てを行った．

さらに，EARS の発話に対しては，感情ラベルを用いたパラ言語的情報のプロンプト（例: “The speaker’s emotion is [emotion label].”）を生成し，前述の形式の “[Additional-information]” として利用した．

非言語情報: 非言語情報のうち性別については，“[subject]” によって記述した．“[subject]” は性別タグ（male または female）に基づいて “woman” または “man” とし，タグのない発話に対しては “person” とした．

さらに，LibriTTS-P には話者性に関するテキスト記述が含まれており，これを追加情報として用いた．LibriTTS-P では，3 名の専門アノテータが “calm” や “fluent” など 40 種以上の印象語を，強度（例：“slightly”，“very”）とともに付与したプロンプトを提供している．このアノテーションにより，多面的な話者性の記述が実現される．我々は各発話に対してラ

Table 2: Examples of prompts 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.

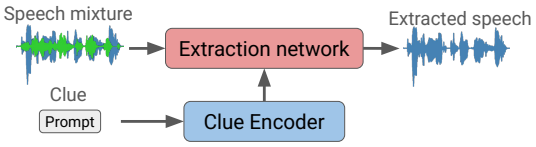

Fig. 2: Baseline model for PNTP-TSE.

ンダムに 1 つの話者性記述文を選定し，これを用いて非言語情報についての記述文を作成した．

3.5 追加の目的音声情報の割り当て

声紋に基づく TSE で用いるために，各混合発話に対し，同一話者による別の目的音声を登録音声として割り当てた．この登録音声の平均時間は 9.76 秒である．また，言語情報に基づく TSE のために，元の音声コーパスで付与された書き起こしを割り当てた．

4 ベースライン手法

本研究では，LASS の先行研究 [7, 12] に基づいて PNTP-TSE のベースライン手法を開発した．Fig. 2 に示すように，このベースライン手法は，分離ネットワーク（Extraction network）と手がかりエンコーダ（Clue encoder）の 2 つのネットワークから構成される．

手がかり（Clue）にはユーザーが提供する記述文（Prompt）$\boldsymbol{c}_s$ を利用し，事前学習された CLAP（Contrastive language-audio pretraining）モデル [17][1] によって固定次元の埋め込みベクトル $\boldsymbol{z}$ を抽出する．このモデルのパラメータは学習時に更新せず固定した．

分離ネットワークには，ResUNet [18] を 30 層に拡張した ResUNet30 [12] を用いる．このネットワークは，手がかりエンコーダによって得られた埋め込み $\boldsymbol{z}$ で条件付けることによって目的音声を抽出する．

5 実験

5.1 実験条件

比較手法: 本研究では，ベースライン手法における分離ネットワークの条件付けに用いる $\boldsymbol{z}$ を変更することで，手がかりの種類ごとの性能差を評価した．すべての手法において，手がかりエンコーダのパラメータは学習中固定されている．

従来手法として，声紋に基づく手法を採用し，この方法では音声から話者特徴量を抽出するモデルを手がかりエンコーダに用いた．“Voiceprint” では，同一話者による登録音声を WavLMForXVector モデル[2] に入力し，x-vector [19] を抽出して $\boldsymbol{z}$ とした．“Oracle” では，目的音声そのものを用いて x-vector を生成した．いずれの条件においても，抽出された x-vector は記述文埋め込みの代わりに分離ネットワークの条件付けとして使用された．

言語クエリベースの TSE では，4 節で述べた手法を用い，記述文の種類を変更して評価を行った．“PNTP-TSE” では，3.4 節で構築した記述文を使用した．要素別検証として，“PNTP-TSE” から gender, pitch, speed, loudness のいずれかを除いた記述文での学習・評価も実施し，それぞれ “PNTP-TSE w/o gender” などと呼称した．さらに，“Gender”, “Pitch”, “Speed”, “Loudness” では単一属性のみを含む記述文を用いた．また，“Rephrased” 条件では，Test セットにおいて ChatGPT [20] を用いて言い換えた記述文を使用し，表現の違いに対するモデルの頑健性を調査した．

モデルと学習: 分離ネットワークには ResUNet30 を用いた．入力・出力は 16 kHz でサンプリングされたモノラル音声であり，長さは 10 秒に正規化された．すべての手法において埋め込み $\boldsymbol{z}$ は 512 次元のベクトルであり，手がかりエンコーダのパラメタは固定した．学習は PromptTSE コーパスの Train セットを用い，最適化手法には AdamW [21]（初期学習率 0.001，10000 ステップのウォームアップ，エポック数 100，バッチサイズ 16）を用いた．損失関数には目的音声と抽出された音声との間の波形領域での L1 損失を使用した．

評価指標: 学習済みモデルは，学習時と同条件の下で PromptTSE コーパスの Test セットを用いて評価した．評価には，scale-invariant signal-to-distortion ratio (SI-SDR) [22]，perceptual evaluation of speech quality (PESQ) [23]，extended short-time objective intelligibility (ESTOI) [24]，deep noise suppression mean opinion score (DNSMOS) [25] の 4 指標を用いた．

5.2 結果

声紋に基づく TSE との比較: 表 3 に結果を示す．“PNTP-TSE” は “Oracle” には及ばなかったが，すべての評価指標において “Voiceprint” を上回った．これは，“Voiceprint” は目的音声と登録音声との間の発話スタイルの差によって性能が低下するためと考えられる [6]．この仮説を確認するため，目的音声と登録音声のピッチおよび話速のクラス一致率を調べ

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

[2] 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	$\mathbf{1.52 \pm 0.35}$	$\mathbf{0.720 \pm 0.16}$	$\mathbf{3.01 \pm 0.27}$

た．その結果，ピッチは 66.8%，話速は 36.9% であり，両者が一致した割合は 24.9% に留まった．このことから，実用上は目的音声と登録音声の発話スタイルが異なることが多く，“Voiceprint” は性能が低下することが確認された．実用的な場面では “Oracle” で用いたようなノイズのない目的音声は得られないため，“PNTP-TSE” が “Voiceprint” を上回る性能を達成したことは実応用における有用性を示している．

要素別検証: “Gender”，“Pitch”，“Speed”，“Loudness” はいずれも “PNTP-TSE” より明確に性能が低く，単一の属性のみで話者を一意に識別するには不十分であることが示された．特に “Speed” は “Speech mixture” を下回る SI-SDR となり，目的音声の識別に失敗していた．これは，話速が系列的な特徴であり，フレーム単位でのモデリングが困難であるためと考えられる．また，“Gender” はクラス数が 2 と少なく，他の属性より識別能力が劣ることが示唆される．一方で “PNTP-TSE w/o Speed” は “PNTP-TSE” よりも性能が低下しており，話速は補助的手がかりとして有用であることが示される．“PNTP-TSE w/o Gender” は “PNTP-TSE” とほぼ同等の性能を示した．これは性別が話者性やピッチと相関を持つためと考えられる．また，“Rephrased” は “PNTP-TSE” より低い性能であったことから，学習時には記述文として多様な表現を取り入れる必要性が示唆される．

6 結論

本研究では，パラ言語・非言語情報の記述文を手がかりとして用いる，LASS ベースの新たな TSE タスク “PNTP-TSE” を提案した．本タスクの定義に加え，ベースラインモデルおよび対応するデータセットである PromptTSE を提供した．従来手法である声紋に基づく TSE 手法が発話スタイルの違いにより性能劣化を起こすのに対し，提案する PNTP-TSE はこの問題を克服することを実験的検証により確認した．PNTP-TSE は，記述文が不正確な場合に性能が劣化する可能性があり，これは今後の課題である．

謝辞: 本研究は科研費 24KJ0860 及び立石科学技術振興財団 研究助成 (S) の助成を受け実施した．

参考文献

[1] Quan Wang *et al.*, in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 2728–2732.

[2] Kateřina Žmolíková *et al.*, *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[3] Katerina Zmolikova *et al.*, *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[4] Rongzhi Gu *et al.*, in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 4290–4294.

[5] Triantafyllos Afouras *et al.*, in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 3244–3248.

[6] Zifeng Zhao *et al.*, in *Proc. INTERSPEECH*, Incheon, Korea, 2022, pp. 5333–5337.

[7] Xubo Liu *et al.*, in *Proc. INTERSPEECH*, Incheon, Korea, pp. 1801–1805.

[8] Hiroya Fujisaki, in *Proc. Speech Prosody – International Conference on Speech Prosody 2004*, Nara, Japan, 2004.

[9] Ziyang Jiang *et al.*, *arXiv preprint arXiv:2411.03109*, 2024.

[10] Xiang Hao *et al.*, *arXiv preprint arXiv:2310.07284*, 2023.

[11] Mingyue Huo *et al.*, *arXiv preprint arXiv:2501.09169*, 2025.

[12] Danilo de Oliveira *et al.*, in *Proc. IWAENC*, Aalborg, Denmark, 2024, pp. 90–94.

[13] Miho Kawamura Zhi Zhu *et al.*, *Acoustic Science and Technology*, vol. 76, no. 6, pp. 317–326, 2020.

[14] Heiga Zen *et al.*, in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1526–1530.

[15] Julius Richter *et al.*, in *Proc. INTERSPEECH*, Kos, Greek, 2024, pp. 4873–4877.

[16] Masaya Kawamura *et al.*, in *Proc. INTERSPEECH*, Kos, Greek, 2024, pp. 1850–1854.

[17] Yusong Wu *et al.*, in *Proc. ICASSP*. Rhodes, Greek: IEEE, 2023, pp. 1–5.

[18] Qiuqiang Kong *et al.*, in *Proc. ISMIR*, Online, 2021, pp. 342–349.

[19] David Snyder *et al.*, in *Proc. ICASSP*, Calgary, AB, Canada, 2018, pp. 5329–5333.

[20] OpenAI, https://chat.openai.com, 2025, accessed: 2025-05-16.

[21] Ilya Loshchilov *et al.*, in *Proc. ICLR*, New Orleans, Louisiana, United States, 2019.

[22] Jonathan Le Roux *et al.*, in *Proc. ICASSP*, Brighton, United Kingdom, 2019, pp. 626–630.

[23] Antony W Rix *et al.*, in *Proc. ICASSP*, vol. 2, Salt Lake City, Utah, United States, 2001, pp. 749–752.

[24] Jesper Jensen *et al.*, *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[25] Chandan KA Reddy *et al.*, in *Proc. ICASSP*. Online: IEEE, 2021, pp. 6493–6497.