Say Who You Want to Hear: Leveraging TTS Style Embeddings for Text-Guided Speech Extraction

Akam Rahimi, Triantafyllos Afouras, Andrew Zisserman

VGG, Department of Engineering Science, University of Oxford, UK

{akam, afourast, az}@robots.ox.ac.uk

Abstract

We introduce TextSep, a novel single-channel speech separation framework that leverages free-form textual description of a speaker's voice to guide separation from noisy multi-speaker audio mixtures, without relying on enrolment audio, images, or video. Building on advances in text-to-speech (TTS), we invert the Parler-TTS pipeline to extract rich style embeddings from the earliest cross-modal layer, enabling speech separation directly from natural language descriptions. Our main contributions are: (1) Curating a large pair of text description and clean-audio pairs (2) identifying and utilizing the projected key vectors of Parler-TTS as effective style embeddings via a lightweight wrapper; (3) integrating these embeddings into a transformer based architecture as prefix tokens and through FiLM modulation of encoder activations; and (4) demonstrating that TextSep achieves competitive performance on synthetic benchmarks, without requiring any reference audio or visual cues.

Index Terms: speech separation, target speaker extraction

1. Introduction

Isolating a single speaker's voice in multi-speaker environments, the classic "cocktail party problem", is a key challenge for hearing aids. Traditional Target Speech Extraction (TSE) systems address this using enrolment audio, face images, or video to identify the target speaker. However, these cues are often impractical: enrolment audio may simply be unavailable or degraded, video requires extra hardware, and they all raise privacy concerns.

This has led to interest in more accessible cues, such as natural language descriptions. These are user-friendly, privacy-safe embeddings that capture rich semantic and paralinguistic details (e.g., gender, tone, accent), making them ideal for speaker identification in real-world applications.

We introduce TextSep, a model that extracts a target speaker's voice using only free-form natural language (e.g., "the man with a calm voice and British accent"). TextSep directly conditions on these descriptions, offering a practical, intuitive alternative for hearing aids where explicit enrolment data is unavailable.

Inspired by recent advances in Text-to-Speech (TTS), particularly the use of natural language prompts to control speaking style, we leverage Parler-TTS [1] to extract style embeddings from user-provided descriptions. We repurpose these style embeddings to condition our separation model, allowing it to extract the target speaker's voice based solely on natural language input.

A major challenge is the lack of large, richly annotated datasets with free-form descriptions. Existing datasets are limited to coarse labels like gender or emotion, which restricts generalization.

To address this, we release a new dataset of 1 million (description, clean-speech) pairs, built from MultiVSR [2] audio segments. Our pipeline combines acoustic features from DataSpeech [3], high-level speaker traits from DeSTA2 [4], and a language model to generate diverse, natural descriptions. This scalable approach supports better training and evaluation of language-guided TSE systems. Our contributions:

- Release a large-scale benchmark of 1M (description, speech) pairs for natural-language-guided TSE.
- Show that Parler-TTS style embeddings are effective for conditioning separation models.
- Demonstrate that TextSep performs within 0.6 dB of enrolment-based systems on synthetic benchmarks.

2. Related work

Current approaches to target speech extraction (TSE) relied on explicit speaker cues, such as enrolment audio or video, to resolve permutation ambiguity in multi-speaker mixtures [5, 6, 7, 8, 9, 10]. While effective, these methods face practical limitations in many real-world scenarios, particularly when such cues are unavailable or raise privacy concerns.

Recent research has explored the use of natural language as a conditioning signal for source separation, spanning applications from general audio events and music to, increasingly, speech itself [11, 12, 13, 14]. LASS-Net [13] demonstrated that textual prompts could be used to extract arbitrary sounds from mixtures via a shared audio-text embedding space. More recently, several works have extended this paradigm to speech.

LLM-TSE [11] pioneered the integration of large language models, such as LLaMA-2, to interpret free-form text cues for TSE, enabling extraction or suppression of speakers based solely on descriptive prompts. StyleTSE [12] leveraged textual descriptions of speaking style and emotion, alongside or instead of reference audio even when speakers have similar acoustic characteristics. ConceptBeam [15] and contextual speech extraction approaches have explored topic-based and dialogue-driven cues, respectively.

Parallel progress in text-to-speech (TTS) has shown that natural language descriptions can control speaker identity, style, and channel conditions at synthesis time. Recent TTS models [1, 16, 17], learn rich acoustic representations from large-scale, text-annotated corpora, demonstrating fine-grained control via text. These advances suggest that TTS-derived embeddings can serve as powerful conditioning vectors for TSE, bridging the gap between descriptive language and audio generation.

3. Method

This section details **TextSep**, our text-guided speech-separation network. We first outline the overall architecture in section 3.1,

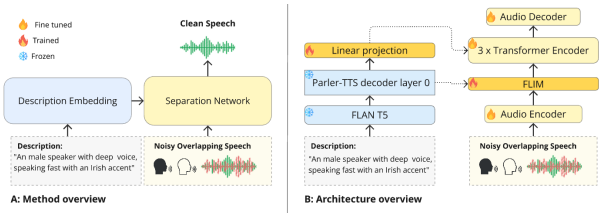

Figure 1: *TextSep overview and architecture.* ***A) Method overview.*** *A free-form description is turned into a style embedding and fed, together with the noisy mixture, into a separation network that outputs the clean target speech.* ***B) Architecture details.*** *The description is processed by a frozen FLAN-T5 encoder; its representations pass through the first cross-attention layer of Parler-TTS, whose key projections capture acoustic style. A small linear projection adapts this 1024-D vector to the separator. The style token conditions the audio path twice: (i) via FiLM on the encoder features and (ii) as prefix tokens for the 3-layer transformer bottleneck. The decoder then reconstructs the separated speech.*

then describe how the textual description is converted into an acoustic style embedding 3.3. We next explain how the transformer bottleneck exploits that embedding during separation, motivate key design choices 3.4, and finally list training and fine-tuning protocols 3.5

3.1. Architecture Overview

The model architecture consists of four main components:

1. **Waveform Encoder:** A U-Net-based encoder processes the input audio mixture, extracting a sequence of latent features that compactly represent the acoustic content of the mixture.
2. **Text-to-Style Embedding Module:** The free-form textual description provided by the user is converted into a style embedding. This embedding captures key characteristics described in the text and acts as a conditioning signal throughout the separation process.
3. **Transformer Bottleneck:** The encoded audio features and the text-derived style embedding are jointly processed by a transformer bottleneck. This component enables rich cross-modal interactions, allowing the network to align acoustic features with the specified style, and effectively focus on the target speaker within the mixture.
4. **Waveform Decoder:** A decoder reconstructs the time-domain waveform of the separated speech from the transformed latent features, completing the separation process.

A schematic of the full architecture is shown in Figure 1. The following provide detailed descriptions of full pipeline.

3.2. Architecture Details

The input to the model is a 16 kHz mono audio mixture $a \in \mathbb{R}^T$. The waveform encoder consists of a five-level, one-dimensional U-Net that maps the mixture to a sequence of latent features. Specifically, the encoder transforms the input into a tensor of shape $\mathbf{A} \in \mathbb{R}^{C \times T_{\text{enc}}}$, where the channel dimension C = 768, and the temporal downsampling reduces the sequence length to T_{enc} = T / 4.

To condition the network on the desired speaker, a free-form textual description $\mathbf{z}$ is mapped to a 1024-dimensional style embedding, as detailed in 3.3. This style token serves a dual purpose. First, it is projected through a two layers of 1×1 convolution to $C = 768$ and is prepended to the encoded audio feature sequence $\mathbf{A}$, forming the input $\mathbf{H} = [\mathbf{z}^{\otimes 1}; \mathbf{A}] \in \mathbb{R}^{(L+T_{\text{enc}}) \times C}$ to the transformer bottleneck. The style token is projected to match the channel width of the audio features and pooled to form the conditioning vector for a FiLM generator. This generator applies feature-wise scaling and bias to the encoded audio, providing an additional conditioning pathway that allows for channel-selective gain control.

To condition the network on the desired speaker, a free-form textual description (e.g., A woman talking in a fast paced excited style with high pitch) is mapped to a 1024-dimensional style embedding, as detailed in 3.3. This style token serves a dual purpose. First, it is projected to match the channel width of the audio features through a two layers of 1×1 convolution to $C = 768$. The projection is prepended as a sequence of $\mathbf{L} = \mathbf{10}$, using a learnable positional encoding, as prefix to the encoded audio feature sequence, forming the input $\mathbf{H} = [\mathbf{z}^{\otimes 1}; \mathbf{A}] \in \mathbb{R}^{(L+T_{\text{enc}}) \times C}$ to the transformer bottleneck. Second, the style token is also pooled to form the conditioning vector for a FiLM generator. This generator applies feature-wise scaling and bias to the encoded audio, that applies channel-wise scaling/shifting to $\mathbf{A}$ to enhance the channels that match the target style and suppress others.

At the core of the network is a three-layer transformer encoder with eight attention heads (dim C) per layer, operating jointly on the sequence of audio and style tokens.

Given the concatenated sequence $\mathbf{H}$ each transformer layer performs

$$\text{CrossAttn}(\mathbf{Q} = \mathbf{H}, \mathbf{K} = \mathbf{H}, \mathbf{V} = \mathbf{H}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

$$\mathbf{H}\prime = \text{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \text{FFN}(\cdot)$$

This enables cross-modal interactions between the audio features and the text-derived conditioning. Audio queries learn to pull from the style keys that best match their timbre/prosody hypothesis, suppressing other speakers. While text queries can refine themselves by attending to the noisy audio, improving robustness when the prompt is imprecise and learn to ignore filler word.

The output of the transformer, denoted $\widehat{\mathbf{A}}$, is then processed by a mirrored U-Net decoder to reconstruct the estimated target waveform $\hat{a}_c$ at 16 kHz. A residual skip connection from the encoder to the decoder enables the preservation of fine acoustic details.

3.3. Text Description Encoding

TextSep employs a natural language description of the target speaker as its primary conditioning input. Given a caption $\mathcal{C}$, we extract the style token from Parler-TTS Mini v1.1. The description is first processed by a frozen Flan-T5 encoder to produce contextualized token embeddings $\mathbf{H}_{\text{T5}} \in \mathbb{R}^{L \times 1024}$. These embeddings are then projected to match the dimensionality of the Parler-TTS transformer decoders via a learned linear mapping. From Parler-TTS we extract the cross-attention key projection of the first transformer layer:

$$\mathbf{K} = W_k\, \mathbf{H}_{\text{dec}}$$

where W_k denotes the key projection matrix of the first transformer decoder layer. A simple mean pooling operation is then applied across the sequence to obtain a fixed-length 1024-dimensional style embedding $\mathbf{z} \in \mathbb{R}^{1024}$.

We observe that this embedding encodes key attributes such as gender, pitch, speaking rate, and recording environment, while remaining close to the pure language semantics of the prompt. Therefore the style token is well-suited to serve as a conditioning signal for the separation network, as it is both semantically informative and acoustically meaningful. This vector is only one linear layer away from pure language semantics, making it ideal for guiding a separator.

3.4. Design Decisions

The use of the projected key vectors from the first decoder layer of Parler-TTS was motivated by two factors: First, the separation model can operate with minimal discriminatory information about the target speaker [6]. Furthermore, it is the earliest point where language meets acoustics; higher-layer embeddings became increasingly entangled with speaker-independent prosody and codebook context, which are not directly useful for the separation task. We apply mean pooling over the token sequence which provides a fixed-length embedding that is computationally efficient and avoids the need for padding logic, with minimal impact on performance. The one-dimensional convolutional projection from 1024 to 768 dimensions acts as a per-timestep linear layer with negligible parameters. This allows the network to learn which dimensions of $\mathbf{z}$ matter, while matching the audio channel width. Finally, we combine prefix token injection with FiLM-based conditioning. The FiLM pathway allows channel-selective gain, biasing early convolutional features toward frequencies characteristic of the described voice. While the the token prefix allows for time-selective masking through the transformer cross-attention layers.

3.5. Training and Implementation Details

The training set consists of two-speaker mixtures synthesized from a subset of the MultiVSR corpus, with background noise from the DNS dataset added at random SNRs between 1–10 dB to simulate realistic conditions. Each instance pairs an audio segment with its corresponding description, using utterances from different speakers with at least two differing attributes. Mixtures are generated on-the-fly.

Training begins with pre-training the audio-only VoiceVector model on MultiVSR using standard speaker embeddings, following the original protocol [9]. The U-Net backbone is initially frozen while the projection and transformer modules are trained on mixtures with maximally distinct speaker attributes. This constraint is gradually relaxed to allow two differing attributes. In the final stage, the full separation model is jointly fine-tuned.

The loss function combines waveform ℓ_1 loss between predicted $\hat{a}_c$ and ground-truth a_c speech, along with proxy SDR loss [18] on short windows. Training uses Adam (=0.9, =0.95), a learning rate of 5×10, and cosine decay to 2×10^{-6}.

4. Experiments

This section describes the datasets, implementation details, experimental setup, and results used to evaluate TextSep for natural-language–guided speech separation. We conduct rigorous comparisons with relevant baselines and analyse the impact of various architectural and design choices.

4.1. Dataset

Training and evaluation are performed on synthetic two-speaker mixtures derived from a subset of the MultiVSR corpus, which contains 1,400 hours of transcribed audio-visual recordings spanning diverse accents, genders, topics, and speaking styles. Each mixture combines two randomly selected utterances that differ in at least two speaker attributes to ensure diversity and avoid trivial cases.

The dataset includes 1,000,000 (mixture, description) pairs, split into 950,000 (1300 hours) for training, and 25,000 each (50 hours) for validation and testing. Audio is resampled to 16 kHz, and each training unit is a 5.0-second (0.8–12 s) clip paired with a natural language description.

Following DataSpeech [3], we compute three acoustic features; C50 reverberation, pitch, and speaking rate discretized by global dataset statistics.

To enrich speaker descriptions, we use the DeSTA2-8B-beta [4] model to extract six additional attributes: age, accent, timbre, tone, speech topic, rhythm, and emotion. We also identify the most salient trait per speaker, yielding eleven descriptive categories when combined with gender and DataSpeech attributes. Generic terms like "natural" or "average" are excluded for specificity.

These keywords are input to Llama-3.1-8B-Instruct[1], which generates fluent, context-rich descriptions such as "A young woman with a lively tone and Australian accent…" This fully automated pipeline enables scalable, diverse, and natural data creation.

For evaluation, we construct the synthetic MultiVSR-Mix test set using unseen speakers and adding ambient noise. The 25,000 evaluation mixtures are built from held-out utterances, with minimal domain gap to real conversational speech, ensuring real-world relevance.

4.2. Evaluation Metrics

We assess model performance using standard metrics widely adopted in speech separation and perceptual quality evaluation. Signal-to-Distortion Ratio (SDR) is used to quantify separation quality, with higher values indicating more accurate recovery of the target speech. Short-Time Objective Intelligibility (STOI) is measured to estimate the intelligibility of the separated speech output, while Perceptual Evaluation of Speech Quality (PESQ) is used to reflect the perceived audio quality from a listener's perspective.

4.3. Results

To evaluate TextSep, we conduct a series of experiments comparing its performance to several baselines and ablations. The baselines include VoiceVector [9], which serves as a speaker-conditional reference and utilizes a 192-dimensional ECAPA-TDNN [19] speaker embedding derived from enrolment audio. We fine-tuned a variation of our model we name CLAP-Sep. It represents a text-conditional approach, employing a 512-dimensional LAION-CLAP [20] text embedding. Additionally, we compare against a previous state-of-the-art audio-visual (AV) system that requires video (lip) input, representing multimodal separation performance.

All baseline and ablation models are trained and evaluated using the same test set.

Table 1 summarizes the quantitative results in comparison

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Table 1: *Speech separation results on the synthetic test set. ✓ indicates modality used for conditioning: video (Vid), speaker embedding (Spk), and text description (Txt). ↑ means higher is better.*

Model	Vid	Spk	Txt	SDR ↑	STOI ↑	PESQ ↑
Noisy input	–	–	–	1.3	69.7	1.30
VoiceFormer [5]	✓	–	–	15.5	93.4	2.60
VF (Phonemes) [5]	–	–	✓	14.1	91.4	2.37
VoiceVector [9]	–	✓	–	14.4	91.1	2.52
CLAP-Sep	–	–	✓	11.8	89.4	2.20
TextSep (T5)	–	–	✓	12.9	89.6	2.24
TextSep (desc.)	–	–	✓	**13.8**	**91.1**	**2.36**

to models that use other modalities than text description such as lip-motion, speaker embeddings and transcription.

The noisy input baseline, representing the unprocessed audio mixture, shows an SDR of 1.3 db, 69.7 STOI, and 1.31 PESQ. This confirms the difficulty of the separation task. The previous audio-visual SOTA method, which leverages both audio and video (lip) information, achieves 15.5 dB SDR, highlighting the benefit of multimodal cues but requiring video input.

The VoiceVector (speaker embeddings), which is conditioned on a 192-dimensional ECAPA speaker embedding derived from enrolment audio, yields 14.4 dB SDR. This demonstrates the strength of speaker-conditional models when enrolment audio is available. The CLAP-Sep baseline conditions the separator on a 512-dimensional LAION-CLAP text embedding, which captures primarily semantic information. Despite this, CLAP-Sep attains 11.8 dB SDR, showing that generic text embeddings, while effective, fall short of acoustic conditioning.

For TextSep-transcription, where the model is conditioned on transcription tokens encoded by a T5 model [21], performance drops to 13.2 dB suggesting that phoneme-level information offers a stronger signal than transcription embeddings using T5. This is because there is a direct correlation between phonemes and the audio signals which the transformer bottleneck is able to attend to.

This also remains inferior to embeddings tuned for acoustic style. Our proposed TextSep-K0 model, which uses the layer-0 key vectors from Parler-TTS as style embeddings based on the provided natural language description, achieves 13.8 dB SDR. Notably, TextSep-K operates without any reference audio or visual information at test time. matching or exceeding the performance of enrolment-based speaker-conditional systems and outperforming all other text-based approaches by a substantial margin. This establishes TextSep as a practical and effective system for text-guided target speech extraction, closing the gap with audio-visual and speaker-conditioned systems that require explicit enrolment samples.

4.4. Ablation Study:

To better understand the factors contributing to the performance of TextSep, we conduct an ablation study focusing on several critical design choices: the richness of the textual prompt, the choice of embedding layer from Parler-TTS, the method of token aggregation, and the role of FiLM conditioning. Table 2 reports performance for each configuration on the MultiVSR test set.

Our results indicate that increasing the descriptive richness of the prompt, by adding more attributes, yields notable gains in separation quality. Moving from a prompt with 3 attributes to one with 4 increases the SDR from 12.9 dB to 13.7 dB, with a corresponding improvement in intelligibility and perceptual quality (STOI rising from 89.4 to 91.1 and PESQ from 2.28 to 2.36). Adding a fifth attribute results in marginal further improvement, suggesting diminishing returns as the prompt becomes saturated with relevant detail.

Table 2: *Ablation studies on TextSep architecture (SDR, STOI, PESQ on MultiVSR test set). ↑ higher is better.*

Variant	SDR (dB) ↑	STOI ↑	PESQ ↑
Prompt: 3 attributes	12.9	89.4	2.28
Prompt: 4 attributes	13.7	91.1	2.34
Prompt: 5 attributes	13.8	91.1	2.34
TextSep-K0: layer-0 keys	13.8	91.2	2.35
TextSep-K1: layer-1 keys	13.9	91.4	2.35
All tokens (no pooling)	13.9	91.3	2.36
FiLM removed	13.4	91.1	2.27

Examining the effect of the embedding source, we find that the default use of Parler-TTS layer-0 key vectors (TextSep-K0) provides strong results (13.8 dB SDR, 91.2 STOI, 2.35 PESQ), with layer-1 (TextSep-K1) keys and using all tokens without pooling yielding only very slight additional gains. This suggests that layer-0 already captures the critical cross-modal style information required for effective separation, and more complex or higher-level embeddings offer minimal advantage in this setting.

The ablation also highlights the importance of the FiLM pathway. Removing FiLM modulation results in a notable drop in performance, with SDR decreasing to 13.4 dB and PESQ to 2.27. This confirms that FiLM-based conditioning contributes non-trivially to TextSep's overall effectiveness, both in signal fidelity and perceived quality.

Collectively, these findings validate the choice of a succinct, attribute-rich textual prompt and the use of layer-0 Parler-TTS keys embeddings. They also underscore the crucial role of FiLM conditioning, while showing that additional complexity in the form of more attributes or alternative token aggregation schemes yields only incremental benefits.

5. Discussions

In this work, we introduced TextSep, a natural-language–guided speech separation framework that leverages advances in large-scale text-to-speech (TTS) models to enable extraction of a target speaker from noisy multi-speaker mixtures using only a natural textual description. Our approach inverts the TTS pipeline: instead of generating speech from text, we extract a style embedding from a natural language prompt using the earliest cross-modal layer of a pre-trained Parler-TTS model and use this embedding to guide waveform-level speech separation via a U-Net transformer architecture.

We demonstrated that TextSep achieves separation performance competitive with traditional speaker-conditional systems that rely on enrolment audio, as well as prior multimodal and phoneme-level text-conditioned baselines.

Acknowledgements.
This work is funded by the UK EPSRC AIMS CDT, the EPSRC Programme Grant VisualAI EP/T028572/1, and a Google-DeepMind Graduate Scholarship.

6. References

[1] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," 2024.

[2] K. R. Prajwal, S. Hegde, and A. Zisserman, "Scaling multilingual visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[3] Y. Lacombe, V. Srivastav, and S. Gandhi, "Data-speech," https://github.com/ylacombe/dataspeech, 2024.

[4] K.-H. Lu, Z. Chen, S.-W. Fu, H. Huang, B. Ginsburg, Y.-C. F. Wang, and H. yi Lee, "Desta: Enhancing speech language models through descriptive speech-text alignment," in *Interspeech 2024*, 2024, pp. 4159–4163.

[5] A. Rahimi, A. Triantafyllos, and A. Zisserman, "Reading to listen at the cocktail party: Multi-modal speech separation," in *Conference on Computer Vision and Pattern Recognition 2022*, 2022.

[6] T. Pan, J. Liu, B. Wang, J. Tang, and G. Wu, "Ravss: Robust audio-visual speech separation in multi-speaker scenarios with missing visual cues," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 4748–4756. [Online]. Available: https://doi.org/10.1145/3664647.3681261

[7] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Interspeech*, 2018.

[8] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, 2018.

[9] A. Rahimi, T. Afouras, and A. Zisserman, "Voicevector: Multimodal enrolment vectors for speaker separation," in *ICASSPW*, 2024, pp. 785–789.

[10] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Interspeech*, 2018.

[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," 2024. [Online]. Available: https://arxiv.org/abs/2310.07284

[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," 2025. [Online]. Available: https://arxiv.org/abs/2501.09169

[13] 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 *Interspeech 2022*, 2022, pp. 1801–1805.

[14] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 458–471, 2025.

[15] Y. Ohishi, M. Delcroix, T. Ochiai, S. Araki, D. Takeuchi, D. Niizumi, A. Kimura, N. Harada, and K. Kashino, "Conceptbeam: Concept driven target speech extraction," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 4252–4260. [Online]. Available: https://doi.org/10.1145/3503161.3548397

[16] G. Liu, Y. Zhang, Y. Lei, Y. Chen, R. Wang, Z. Li, and L. Xie. Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. [Online]. Available: http://arxiv.org/abs/2305.19522

[17] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, J. Wang, I. Cruz, B. Akula, A. Akinyemi, B. Ellis, R. Moritz, Y. Yungster, A. Rakotoarison, L. Tan, C. Summers, C. Wood, J. Lane, M. Williamson, and W.-N. Hsu, "Audiobox: Unified audio generation with natural language prompts," 2023. [Online]. Available: https://arxiv.org/abs/2312.15821

[18] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech 2020*, 2020.

[19] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech 2020*. ISCA, Oct. 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2650

[20] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[21] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," 2022. [Online]. Available: https://arxiv.org/abs/2210.11416