



TargetVoice: Single Channel Low-Latency Target Speaker Extraction

Arun Kumar Pallala, Nivedita Chennupati, Balaji Padmanaban, Rakesh Pogula
Umasubhashini Ravuri, Naveen Ellanki, Harish Rajamani, Naveen Ambati

¹Meeami Technologies, Hyderabad, India,

{arunkumar.pallala, nivedita.chennupati, balaji.padmanaban,
rakesh.pogula, umasubhashini.ravuri, naveen.ellanki, harish.rajamani,
naveen.ambati}@meeamitech.com

Abstract

We present TargetVoice, a lightweight, low-latency target speaker extraction (TSE) model optimized for edge devices. It isolates a target speaker’s voice from multi-speaker and noisy environments, making it ideal for use in call centers, conference calls, hands-free communication, and smart speakers. By streaming only the enrolled speaker’s voice, TargetVoice also improves speech recognition accuracy in real-world conditions. Unlike existing models that struggle with similar-gender speakers or varying acoustic environments, TargetVoice leverages a robust in-house data strategy and a specialized speaker embedding extraction system. The model uses a compact 10MB speaker encoder to generate a reliable embedding from a single 3 second enrollment. This embedding is fused with the input mixture in a 12MB extraction block with 6G MACs to isolate the target voice efficiently, enabling real-time performance on resource-constrained platforms.

Index Terms: speech recognition, human-computer interaction, target speaker extraction, speech enhancement

1. Introduction

In realistic environments, speech communication often occurs amidst multiple speakers and background noise, posing significant challenges in various scenarios such as online conferences, voice calls, and interactions with voice-driven applications. Target Speaker Extraction (TSE) is a critical speech processing task aimed at isolating and enhancing the voice of a specific speaker from a mixture of multiple voices and environmental sounds. Unlike traditional speech separation techniques, which attempt to separate all overlapping sources, TSE focuses solely on extracting the desired speaker using an enrollment audio reference.

One of the primary challenges in TSE is handling highly variable acoustic conditions, including reverberation, background noise, and dynamic interference from competing speakers. Traditional methods, such as Blind Source Separation (BSS) and beamforming, rely on spatial cues from multi-microphone arrays to isolate speakers. However, these approaches struggle in single-channel (mono) settings and fail when speakers share similar spatial characteristics. More recent deep-learning approaches have significantly improved TSE performance by leveraging speaker embeddings, attention mechanisms, and temporal-frequency modeling.

VoiceFilter and its lightweight variant, VoiceFilter-Lite [1, 2], were developed to improve ASR performance by suppressing competing speakers using speaker embeddings (d-vectors). While VoiceFilter achieved strong noise suppression, its computational demands limited real-time deployment. VoiceFilter-Lite addressed this with model compression and quantization,

enabling efficient on-device inference for smart devices. However, both were optimized for ASR enhancement rather than standalone target speaker extraction.

SpeakerBeam [3] proposed an adaptive neural network conditioned on speaker embeddings to extract the target voice. Though effective, it depends on clean enrollment audio, which is often unavailable in real-world settings.

Look Once to Hear [4] introduced a self-attentive cross-modal fusion method for target speaker extraction using binaural audio. Tailored for headphone-based use, it leverages spatial cues and speaker priors to isolate the target speaker in noisy environments. However, it requires binaural input, limiting its applicability in single-microphone scenarios.

Our proposed TargetVoice model draws inspiration from Look Once to Hear[4], but it is specifically designed to operate with a single microphone. This design choice allows for greater flexibility in deployment across a wide range of use cases. It has been observed that many open-source Target Speaker Extraction (TSE) models struggle when the interfering speaker is of the same gender as the target speaker, or when the speaker enrollment is recorded in a different acoustic environment. Our proposed system addresses these challenges through a proprietary data strategy and by training a model capable of extracting robust and environment-invariant speaker embeddings.

2. TargetVoice Model

Fig. 1 illustrates the block diagram of TargetVoice. The model utilizes an enrollment of the desired speaker to isolate his/her voice from a mixture of multiple speakers and background noise. It consists of two primary components:

1. Speaker Encoding Module
2. TargetVoice Enhancement Module

The speaker encoding module directly accepts the audio enrollment of the target speaker as input and generates a robust speaker embedding. A minimum of 3 seconds audio is used for speaker enrollment. An offline version of the TF-GridNet module [5] is employed for this purpose. The speaker encoder module is trained on the VoxCeleb dataset, which consists of 7,500 speakers. A combination of cross-entropy loss and hinge loss is used.

During inference, an intermediate layer of this network is used to derive a unique speaker representation, which is then L2-normalized to serve as the final speaker embedding. This ensures that the embeddings remain robust, speaker-discriminative, and invariant to acoustic conditions. Once generated, these speaker embeddings are used in the TargetVoice enhancement module to condition the network on the target speaker. Importantly, the speaker embedding model remains frozen during the training of the enhancement module.

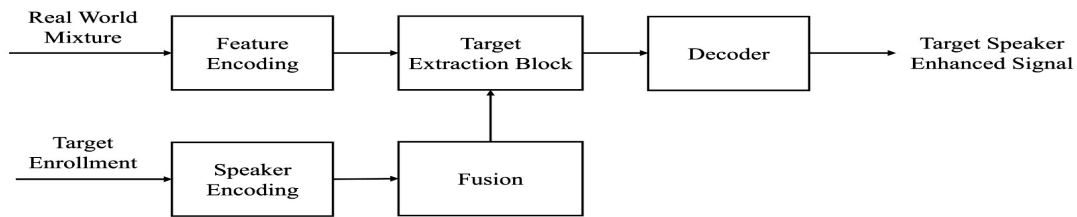


Figure 1: Block diagram of TargetVoice model

The TargetVoice enhancement module consists of four main components, as illustrated in Fig. 1. The Feature Encoder block processes the real-world audio mixture and transforms it into latent representations. The Fusion block integrates the speaker embedding with these latent representations. The Target Extraction block then takes the fused features and extracts enhanced latent representations corresponding specifically to the target speaker. Finally, the Decoder block reconstructs the target speaker’s audio signal from the enhanced latent features.

The input mixture signal is first processed by the Feature Encoder module using a 20 ms window and a 10 ms hop length. A Hanning window is applied, and the FFT size (N-FFT) is set equal to the window size, to extract both real and imaginary components of the signal. This time-frequency representation is then passed through a convolutional layer, which generates rich spectral features for subsequent processing.

The target extraction block consist of TF-GridNet blocks repeated for 3 times. Each TF-GridNet block consists of Intra-LSTM, Inter-LSTM, and a self-attention module. Several key modifications have been made to optimize real-time performance and computational efficiency, inspired by [4]. Global layer normalization has been removed to ensure streaming capability. De-convolution layers have been eliminated to reduce latency and computations. The cell state buffers are used in Inter-LSTM for real-time inference. The self-attention module is also applied to chunks of data using a buffer of 50 frames.

The L2-normalized speaker embedding produced by the Speaker Encoder module is first expanded and broadcasted to align with the time-frequency resolution of the input mixture. It is important to note that the speaker’s average embedding is derived from a single enrollment of the target speaker, augmented with noise and reverberation. This data augmentation strategy enhances the system’s robustness to real-world acoustic environments. The mixture features and the speaker embedding are then fused using a point-wise convolution layer, which effectively integrates speaker-specific information into the feature space.

The fused representation is progressively injected into target extraction block. A Transposed Convolution layer is applied to estimate the real and imaginary components of the enhanced signal. Finally, the output is processed through a feature decoder, which performs an inverse FFT operation to reconstruct the time-domain signal, delivering a high-quality, isolated speaker output.

The TargetVoice enhancement model was trained using the LibriSpeech and VoxCeleb datasets. The Pedalboard toolkit was used to generate reverberation in the training data, with Signal-to-Interference Ratio (SIR) ranging from -10 dB to 20 dB. The

dataset consisted of 60% reverberant and 40% non-reverberant speech, covering diverse real-world scenarios. In total, 200 hours of data is created, ensuring that speaker identities were disjoint across training and testing sets. The model is also finetuned on real data recorded with different microphones placed at varying distances in different room configurations. The Adam optimizer with an initial learning rate of 1×10^{-4} , along with learning rate schedulers and an early stopping mechanism, was used. Training was guided by SNR loss, while model performance was evaluated using the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR).

3. Results and Conclusions

The proposed model operates in real time with a processing latency of just 10 ms, making it well-suited for streaming applications. The speaker embedding module is compact at 10 MB. This module is invoked only once to extract the target speaker’s embedding.

Across multiple evaluation scenarios, the model consistently achieves an average improvement of 8–10 dB in SI-SDR, even when the signal-to-interference ratio ranges from -10 dB to 0 dB. These results underscore the system’s robustness in extracting the target speaker’s voice from complex, noisy, and overlapping speech environments.

Overall, TargetVoice offers a compelling combination of low computational complexity, real-time performance, and high speaker separation quality. Its scalability and efficiency make it ideal for deployment on edge devices in real-world applications such as smart assistants, voice bots, and hands-free communication systems even in highly dynamic, multi-speaker settings.

4. References

- [1] Q. Wang, W. Chan, C. Zhang, J.-C. Chou, and D. Tran, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," *Proceedings of INTERSPEECH*, 2019.
- [2] Q. Wang, Y. Wu, C. Zhang, Y. Jia, and Y. Cao, "VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition," *Proceedings of INTERSPEECH*, 2020.
- [3] M. Delcroix, K. Kinoshita, T. Nakatani, and A. Ogawa, "Speaker-Beam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures," *IEEE Journal of Selected Topics in Signal Processing*, 2018.
- [4] B. Veluri, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, "Look Once to Hear: Target Speech Hearing with Noisy Examples," *CHI Conference on Human Factors in Computing Systems*, 2024.
- [5] Z.-Q. Wang, S. Cornell, and D. Wang, "TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.