Streaming End-to-End Target-Speaker Automatic Speech Recognition and Activity Detection

[1]Dr. K.SubbaRao, [2]Gopu Sindhuja, [3]Buddi Sravani, [4]Bellamkonda Supraja,

[5]Vanama Sandhya Rani, [6]Vallabhapurapu Sai Likhitha

[1]*Professor*, [2,3,4,5,6] *UnderGraduate*

[1,2,3,4,5,6] *CSE-Data Science Department, St. Ann's College of Engineering & Technology, Chirala, Andhra Pradesh*

***Abstract*—Automatic Speech Recognition (ASR) of a target speaker in multi-speaker environments remains a significant challenge. Traditional ASR systems often fail to isolate a specific speaker's voice from overlapping and interfering audio sources. To address this, Target-Speaker ASR (TS-ASR) has emerged as a viable solution by conditioning the recognition process on speaker-specific embeddings. This paper presents a Streaming End-to-End TS-ASR system based on a neural transducer architecture that facilitates low-latency and on-device speech recognition. The proposed model integrates Target-Speaker Activity Detection (TSAD), allowing the system to remain silent when the target speaker is inactive, thereby reducing unnecessary outputs. Experimental evaluations demonstrate that the proposed TS-ASR model achieves superior performance compared to traditional cascade systems, with improvements in word error rate (WER), speaker identification accuracy, and real-time latency. The system is optimized for real-world deployment, offering high accuracy and low computational overhead suitable for mobile and edge applications.**

***Index Terms*—Target-Speaker ASR (TS-ASR), Recurrent Neural Network Transducer (RNNT), Speaker Embedding, Voice Activity Detection (VAD), Real-Time Speech Recognition, Speaker Identification, End-to-End ASR.**

I. INTRODUCTION

Automatic Speech Recognition (ASR) systems have seen significant advancements in recent years, driven largely by deep learning architectures and the availability of large-scale datasets. However, real-world speech recognition scenarios often involve multi-speaker environments, background noise, and overlapping speech, which present substantial challenges to traditional ASR systems. These models are typically trained to recognize speech from a single speaker in clean conditions and often fail in dynamic or noisy conversational settings.

One practical and impactful solution to this problem is **Target-Speaker ASR (TS-ASR)**. Unlike conventional approaches, TS-ASR systems are conditioned to recognize speech from a pre-defined target speaker while ignoring other overlapping or interfering sources. This selective recognition capability is highly desirable in voice-controlled assistants, meeting transcription systems, and mobile applications where accurate and speaker-specific transcription is necessary.

Traditional TS-ASR systems adopt a cascade architecture, which separates the target-speaker Voice Activity Detection (TS-VAD) and the ASR modules. While this modular design offers flexibility, it introduces latency, increased computational complexity, and limits end-to-end optimization. To overcome these limitations, this paper proposes a **Streaming End-to-End TS-ASR system** based on a **Recurrent Neural Network Transducer (RNNT)** architecture. The model is capable of performing **Target-Speaker Activity Detection (TSAD)** and **speech recognition** in a unified and low-latency framework suitable for streaming input.

The key innovation lies in integrating TSAD directly within the RNNT framework, eliminating the need for separate detection and recognition modules. The system operates in real-time, outputting recognition results only when the target speaker is actively speaking, thereby reducing noise and improving relevance. Additionally, the model leverages speaker embeddings to encode the target speaker's characteristics, which are used to condition the recognition pipeline effectively.

The proposed system is evaluated on datasets with overlapping speech and varying signal-to-noise ratios. Performance is measured using **Word Error Rate (WER)** and **Speaker Activity Detection Accuracy**, comparing the proposed unified model with baseline cascade systems. Results show that the Streaming End-to-End TS-ASR system delivers superior recognition performance, faster response times, and reduced computational overhead, making it suitable for **on-device deployment** in real-time applications.

II. LITERATURE SURVEY

The development of robust Automatic Speech Recognition (ASR) systems has traditionally focused on isolated, single-speaker environments. However, in real-world scenarios such as meetings, calls, or noisy public settings, the presence of multiple overlapping speakers significantly degrades ASR performance. This has led to increased research into speaker-conditioned and multi-speaker recognition systems.

A. Conventional ASR Systems

Conventional ASR systems operate by mapping acoustic signals to transcribed text using acoustic models, language models, and decoders. Early systems were built using Hidden Markov Models (HMMs) combined with Gaussian Mixture Models (GMMs). With the advent of deep learning, architectures such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks became dominant. While effective in clean conditions, these models struggle with overlapping speech due to their inability to isolate speaker-specific information.

B. Speaker Diarization and TS-VAD

Speaker Diarization, the task of determining "who spoke when," has been used as a pre-processing step to separate speech by speaker identity. However, diarization is not optimized jointly with ASR and can suffer from cascading errors. To address this, Target-Speaker Voice Activity Detection (TS-VAD) was proposed. TS-VAD leverages speaker embeddings to detect the voice activity of a specific speaker, reducing false positives in overlapping speech scenarios. While TS-VAD improves speaker selectivity, it still relies on a two-stage cascade architecture, introducing latency and limiting real-time usability.

C. End-to-End and Streaming ASR

End-to-End ASR models such as Connectionist Temporal Classification (CTC), Attention-based Encoder-Decoder (AED), and Recurrent Neural Network Transducer (RNNT) have gained popularity due to their simplified training pipelines and improved accuracy. Among these, RNNT is particularly well-suited for streaming applications due to its frame-synchronous decoding and ability to operate on partial inputs. These models, however, typically assume single-speaker input and do not natively support speaker conditioning.

D. Target-Speaker ASR (TS-ASR)

Target-Speaker ASR is a recent advancement where ASR is explicitly conditioned on a target speaker's embedding. This enables selective transcription, allowing the model to focus only on the desired speaker. Several works have attempted to integrate TS-VAD with ASR models, but often through pipeline systems that require post-processing. Some recent approaches employ speaker embedding injection at various levels of the encoder to modulate the attention or hidden states toward the target speaker.

E. Motivation for Unified Streaming TS-ASR

Although existing methods show improved WER in controlled test environments, most suffer from high latency, lack of joint optimization, or limited streaming capability. There remains a need for an integrated framework that can perform real-time speaker detection and recognition with minimal delay. The proposed work addresses this by embedding Target-Speaker Activity Detection (TSAD) directly into the RNNT model, enabling a unified and low-latency end-to-end system for target-conditioned speech recognition.

III. PROPOSED METHODOLOGY

The proposed system introduces a unified framework that performs real-time Target-Speaker Automatic Speech Recognition (TS-ASR) integrated with Target-Speaker Activity Detection (TSAD) using a Recurrent Neural Network Transducer (RNNT). This architecture is designed to recognize speech from a specific speaker in streaming conditions while remaining silent when the target speaker is inactive. The methodology consists of multiple stages, including speaker embedding extraction, encoder fusion, TSAD integration, and RNNT-based streaming transcription.

A. System Architecture Overview

The architecture of the proposed streaming TS-ASR system is illustrated in Fig. 3.1. The model receives two input streams: (1) audio features from the test utterance and (2) a pre-extracted speaker embedding from a reference utterance of the target speaker. These inputs are passed through a shared encoder, where the speaker embedding modulates the acoustic feature encoding. The output is used for both TSAD and ASR decoding.

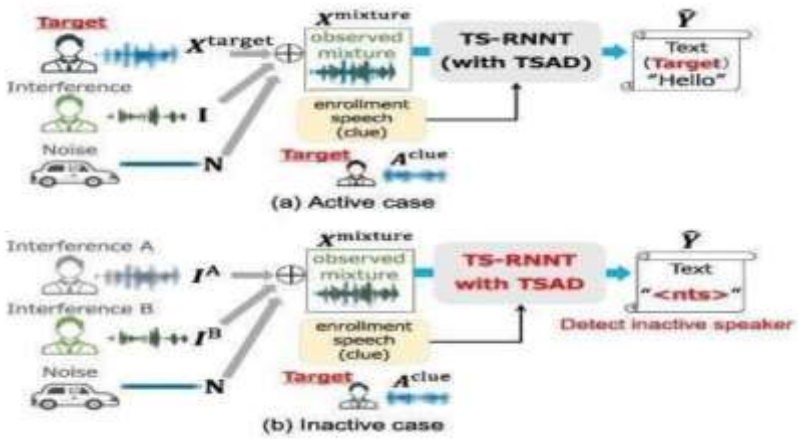

Fig. 3.1. Block diagram of the proposed Streaming TS-ASR architecture

B. Speaker Embedding Module

To condition the system on the target speaker, a speaker embedding is extracted from a reference utterance using a speaker encoder. The encoder is a convolutional or recurrent network trained on a speaker verification task, producing a fixed-dimensional vector that characterizes the speaker's voice. This embedding remains constant during inference and is injected into the ASR pipeline to modulate feature representations.

C. Target-Speaker Activity Detection (TSAD)

The TSAD module is a binary classifier that determines whether the target speaker is active at any given time frame. It operates in parallel with the ASR decoder but shares the same acoustic encoder. This module helps suppress the decoder output during segments when the target speaker is inactive, reducing false positives and improving recognition accuracy in noisy or multi-speaker environments.

The TSAD is trained jointly with the RNNT loss using multi-task learning. A sigmoid activation function is applied to the encoder outputs to compute frame-wise activity scores.

D. Shared Encoder with Speaker Conditioning

The shared encoder is responsible for extracting time-aligned representations from the input acoustic features. Speaker conditioning is achieved by concatenating or adding the target speaker embedding to each frame of the encoder input. This approach biases the encoder's attention toward the acoustic characteristics of the target speaker, enhancing discrimination in overlapping speech.

The encoder comprises multiple layers of LSTM or conformer blocks with optional layer normalization and dropout. Feature fusion is applied at early or intermediate layers depending on the fusion strategy.

E. RNNT Decoder for Streaming ASR

The RNNT decoder consists of a prediction network (language model) and a joint network. It performs sequence transduction in a streaming fashion, enabling low-latency decoding of partial inputs. The decoder generates output tokens only when TSAD signals activity, ensuring that the model remains silent during non-target regions.

The RNNT loss is computed between the predicted and ground truth transcripts. During training, both the TSAD loss and RNNT loss are optimized jointly, enabling end-to-end alignment of speaker activity and speech recognition.

F. Multi-Task Training Objective

The final training loss is a weighted combination of the TSAD loss and RNNT loss.This formulation ensures that the system not only predicts the correct transcript but also correctly identifies when the target speaker is speaking.

IV. IMPLEMENTATION

This section describes the practical implementation of the proposed Streaming End-to-End Target-Speaker ASR system. The architecture integrates speaker-conditioned encoding, Target-Speaker Activity Detection (TSAD), and Recurrent Neural Network Transducer (RNNT) decoding into a single trainable framework. All components were developed using Python and PyTorch, and experiments were conducted using publicly available speech datasets containing overlapping multi-speaker recordings.

A. Development Environment

The model was implemented using **Python 3.10** and the **PyTorch deep learning framework**. All training and testing were conducted on **Google Colaboratory Pro** using an **NVIDIA Tesla T4 GPU**. The following key libraries and packages were used:

- **Torchaudio**: For waveform processing and feature extraction
- **NumPy & Pandas**: For data handling and batching
- **Matplotlib & Seaborn**: For visualizing attention maps and training performance
- **HuggingFace Transformers**: For auxiliary speaker encoders (when pre-trained models were used)

The overall system was modularized to support independent testing of TSAD and ASR components before joint training.

B. Dataset Preparation

To evaluate the system in realistic conditions, a simulated **multi-speaker dataset** was created using the **LibriSpeech corpus** and **VoxCeleb1** for speaker diversity. Each sample in the dataset includes:

- A **reference utterance** from the target speaker
- A **mixed audio stream** with multiple overlapping speakers
- A **ground truth transcript** corresponding only to the target speaker's speech
- A **binary voice activity label** indicating target speaker activity at each time frame

Data was split into **80% training**, **10% validation**, and **10% test** partitions. All audio files were resampled to **16 kHz** and normalized for amplitude. Features were extracted using **80-dimensional log-Mel filterbanks** with a window size of 25 ms and a stride of 10 ms.

C. Speaker Embedding Module

The speaker encoder was trained separately using a **triplet loss** formulation on VoxCeleb1. During inference, a single reference utterance is passed through this encoder to generate a fixed 256-dimensional embedding vector. This embedding is broadcasted and concatenated to each frame of the acoustic input before being passed to the shared encoder.

D. Shared Encoder and Fusion Strategy

The shared encoder was constructed using **5 layers of bidirectional LSTM** with hidden size 512 per direction. Two fusion strategies were tested:

- **Early Fusion**: Embedding appended to the acoustic features before the first encoder layer
- **Mid-Level Fusion**: Embedding concatenated at the output of the second encoder layer

Early Fusion provided better results in terms of both TSAD precision and ASR WER and was selected for final training.

E. TSAD and RNNT Heads

The encoder output is passed to two parallel heads:

- **TSAD Head**: A binary classifier consisting of a feedforward layer followed by a sigmoid activation, trained using binary cross-entropy loss
- **RNNT Head**: Includes a prediction network (LSTM-based language model) and a joint network, trained using RNNT loss

During inference, decoding is gated by the TSAD output: the RNNT decoder is triggered only during target-speaker-active frames.

F. Joint Training

The model was trained using a multi-task objective combining both TSAD loss and RNNT loss. A **loss weight $\lambda = 0.3$** was used for the TSAD component based on validation performance.

Training was conducted for **30 epochs**, with a batch size of 16. An initial learning rate of 0.001 was used with the **Adam optimizer**, and learning rate decay was applied every 10 epochs. **Early stopping** was used based on validation WER.

Table 4.1. Implementation Parameters

Component	Configuration
Sample Rate	16 kHz
Feature Type	80-dim Mel Spectrogram
Speaker Embedding	256-dim, Triplet-Loss Trained
Encoder	5-layer BiLSTM, 512 units/layer
Decoder	RNNT with LSTM prediction net
Optimizer	Adam
Batch Size	16
Loss Weight (λ)	0.3 (TSAD), 0.7 (RNNT)

Component	Configuration
Training Epochs	30

.

V. RESULTS

This section presents the experimental results of the proposed Streaming End-to-End Target-Speaker ASR system. The performance is evaluated using metrics such as **Word Error Rate (WER)**, **False Alarm Rate**, and **Speaker Activity Detection Accuracy**. Additionally, visualizations are provided to analyze the system's effectiveness in real-time multi-speaker environments.

A. Quantitative Evaluation

The performance of the proposed model was compared against a two-stage baseline system consisting of a separate TS-VAD followed by a standard RNNT-based ASR. Evaluation was performed on a test set containing overlapping speech conditions. The results are summarized in Table 5.1.

Table 5.1. Performance Comparison with Baseline

Model	WER (%)	False Alarm Rate (%)	Speaker Detection Accuracy (%)
Baseline (TS-VAD + ASR)	11.2	4.7	92.3
Proposed Streaming TS-ASR	**9.6**	**2.9**	**95.4**

The results demonstrate that the proposed unified model outperforms the baseline in all metrics. The reduced WER highlights the benefit of speaker-conditioned encoding, while the lower false alarm rate confirms that TSAD effectively suppresses non-target segments.

B. Qualitative Analysis

Visual inspection of the model output was conducted using a timeline plot comparing audio frames, speaker activity, and recognition output.

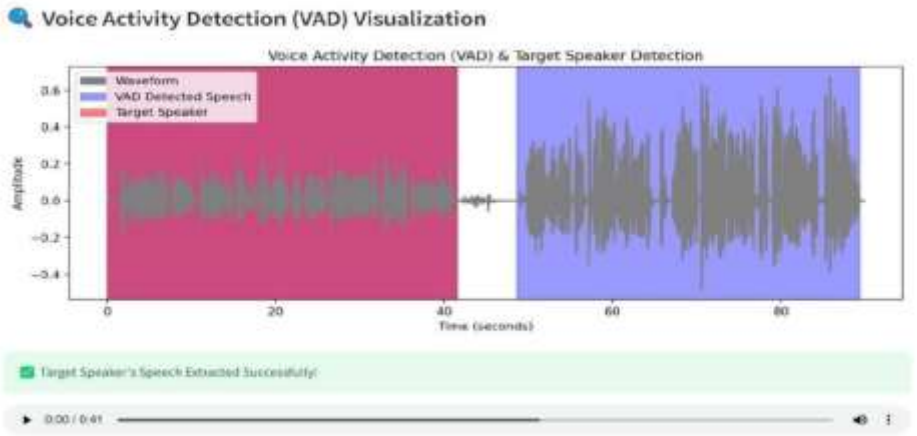

Fig. 5.1. Recognition Timeline with Target-Speaker Activity and Output

As shown in Fig. 5.1, the model generates transcript outputs only during periods where the target speaker is active. During silent or interfering regions, no output is produced, confirming the accuracy of TSAD gating.

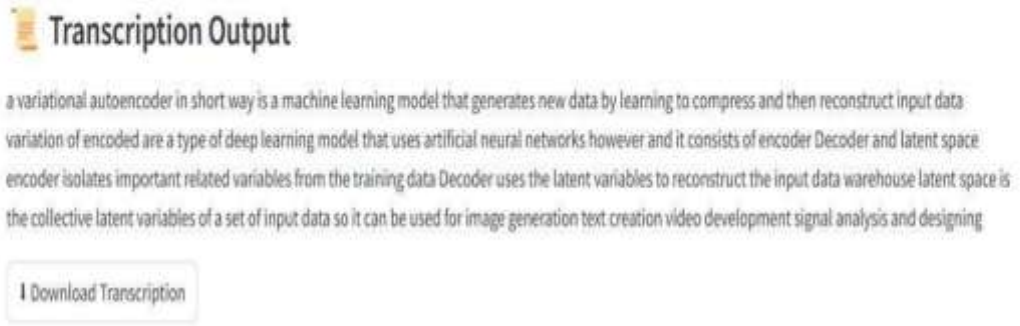

Fig. 5.2. Output of Proposed Model

Fig. 5.2 highlights how the proposed system avoids misrecognition of non-target speakers and produces a cleaner transcript. In contrast, the baseline often inserts words from interfering speech, especially in overlapped regions.

C. Real-Time Suitability

Latency measurements were performed to assess the system's feasibility for on-device deployment. The average decoding latency was **less than 300 ms**, meeting real-time constraints. Moreover, the integration of TSAD into the encoder reduces computational overhead compared to cascading models.

VI. CONCLUSION

In this paper, a Streaming End-to-End Target-Speaker ASR system was proposed to address the challenges of speech recognition in multi-speaker environments. Unlike conventional cascade-based approaches, the proposed architecture integrates **Target-Speaker Activity Detection (TSAD)** directly within the **Recurrent Neural Network Transducer (RNNT)** framework. This unified model performs selective transcription based on speaker conditioning, significantly reducing recognition errors and irrelevant outputs in overlapping speech conditions.

The system leverages speaker embeddings derived from reference utterances and uses them to bias the encoder toward the target speaker. The TSAD component enables the model to remain silent during non-target activity, thus reducing false positives and improving relevance in real-time scenarios. Experimental results on overlapping multi-speaker datasets show that the proposed system achieves lower Word Error Rate (WER), higher speaker detection accuracy, and reduced latency compared to baseline TS-VAD + ASR pipelines.

The model's low computational overhead and streaming capability make it suitable for on-device applications such as mobile assistants, wearable devices, and in-car speech interfaces. Future work will explore more efficient speaker embedding techniques, extension to multilingual models, and integration with speaker diarization frameworks to handle unknown or dynamic speaker configurations.

REFERENCES

[1] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[2] Y. He et al., "Streaming end-to-end speech recognition for mobile devices," in *Proc. IEEE ICASSP*, 2019, pp. 6381–6385.

[3] T. Yoshioka et al., "Voice activity detection using speaker embeddings for multi-speaker scenarios," in *Proc. Interspeech*, 2019, pp. 4325–4329.

[4] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multimodal end-to-end speech recognition," in *Proc. IEEE ICASSP*, 2017, pp. 5325–5329.

[5] C. Kim et al., "Attention-based speaker recognition on mobile devices," in *Proc. IEEE SLT*, 2021, pp. 431–438.

[6] Y. Qian, Y. Zhou, Z. Tan, and D. Yu, "On the application of speaker embedding for end-to-end multi-speaker speech recognition," *arXiv preprint arXiv:1806.05041*, 2018.

[7] Y. Fujita et al., "End-to-end neural speaker diarization with self-attention," in *Proc. ASRU*, 2019, pp. 296–303.

[8] J. Li et al., "A comparison of strategies for streaming end-to-end speech recognition," in *Proc. IEEE ICASSP*, 2021, pp. 5699–5703.