Real-Time end-To-End Target- Speaker ASR with Integrated Activity Detection in Multi-Speaker Environments

[1]N. Sai Sindhuri, [2]K. Chashmitha Lakshmi, [3]M. Venkata Karthika, [4]M. Keerthi, [5]A. Vyshnavi

[1]Assistant Professor , [2,3,4,5]UG Students, [1,2,3,4,5]Department of Computer Science & Engineering, Geethanjali Institute Of Science And Technology, Nellore, India

Abstract

Recognizing speech from a specific speaker amidst overlapping conversations presents a persistent challenge in automatic speech recognition (ASR). Conventional systems often rely on separate modules for target speech extraction and recognition, which increases latency and system complexity. This project introduces a streaming, end-to-end Target-Speaker ASR framework based on a neural transducer architecture, designed for real-time and edge-device deployment. The proposed system utilizes speaker embeddings to condition the recognition model on a predefined speaker's voice, allowing it to isolate and transcribe only the target speaker while suppressing interfering voices. Additionally, a built-in target-speaker activity detection (TSAD) module ensures the system remains silent during non- speaking intervals of the target speaker, reducing unnecessary transcriptions. This unified approach not only streamlines the ASR pipeline but also improves accuracy, efficiency, and responsiveness in multi-speaker, real-world scenarios.

Keywords:

Introduction

In modern speech processing systems, understanding and transcribing speech in real-time, especially in environments with multiple speakers, is a significant challenge. This project aims to develop a real-time, end-to-end automatic speech recognition (ASR) system that focuses on a target speaker— a specific individual among multiple voices—by integrating speaker activity detection mechanisms directly into the ASR pipeline. Isolate and recognize speech from a target speaker in environments with overlapping or background speech. Achieve real-time performance suitable for live applications like voice assistants, meetings, and surveillance. Improve the accuracy of speech recognition by detecting when the target speaker is active and ignoring irrelevant segments. This project combines advanced techniques in speech separation, speaker verification, and ASR, integrating them into a single, efficient model. It is particularly valuable for applications such as smart assistants, customer service bots, and voice-controlled systems in noisy or crowded environments.

The motivation for developing Real-time End-to-End Target-speaker ASR with Integrated Activity Detection in Multi-speaker Environments stems from several real-world challenges in speech recognition. Here's a breakdown of the motivation and a concise summary. Multi-speaker Challenges: In environments like meetings, call centers, or smart homes, multiple people may speak simultaneously or sequentially. Traditional ASR systems struggle to isolate and accurately transcribe only the target speaker. Need for Target-Speaker Focus: Many applications only require transcribing one specific person's speech (e.g., virtual assistants responding to their user, or diarization in meeting minutes). Thus, isolating and recognizing only the target speaker's voice improves performance and reduces unnecessary processing.

Real-Time Processing: To be practical (e.g., for voice assistants or live captioning), the system must operate in real-time with low latency. Integrated Activity Detection: Instead of using a separate voice activity detection (VAD) module, integrating activity detection directly into the ASR model simplifies the pipeline and enables better synchronization between detecting and transcribing speech.End-to-End Learning Benefits: An end-to-end model reduces the complexity of traditional modular ASR systems, potentially improving performance, training efficiency, and ease of deployment. The project proposes a real-time, end-to-end speech recognition system that focuses on transcribing only the target speaker in a noisy, multi-speaker environment. By integrating speaker activity detection directly into the ASR model, it simplifies the architecture and improves performance. This approach is particularly useful for applications like smart assistants, meeting transcription, and surveillance, where isolating and recognizing one speaker's voice in real time is critical.

Literature Review

2.1 YUYA FUJITA, YUSUKE FUJITA, SHINJI WATANABE, "Unified Target-Speaker Speech Recognition with Two-Pass Modeling " This work presents a unified offline TS-ASR system that combines TSAD (Target-Speaker Activity Detection) and TS-RNNT (Target-Speaker Recurrent Neural Network Transducer) in a two-pass framework. TSAD is used to detect the activity of the target speaker and to discard segments of audio where the target speaker is inactive. TS-RNNT is then used to transcribe only the relevant segments.Model training requires large, well-annotated datasets with speaker labels,real-time deployment not supported due to offline nature.

2.2 WANG, QUAN ET AL, "Voice Separation with Speaker Embedding Guidance" This paper proposes a deep learning model for extracting a known speaker's voice from a mixture using a speaker embedding as reference. The model guides the separation process with the embedding.Enables speaker-dependent separation and useful as pre-processing for ASR. Performance degrades in multi-speaker scenarios beyond two voices, assumes high-quality reference audio.

CHANG, XINGJIAN ET AL, "End-to-End Multi-Speaker Speech Recognition with Transformer" This paper introduces a transformer-based end-to-end model that transcribes speech from multiple overlapping speakers. It uses dual decoders to handle speaker separation and recognition jointly. The approach achieves high transcription accuracy in multi-speaker environments and provides robust sequence modeling. However, the model imposes a high computational load, and its complexity increases significantly as the number of speakers grows. Speech Separation Based on Speaker Extraction Using Time-Frequency Masking

2.4 HU HU, JIAN WU, JIE WANG, XIXIN WU, LEI XIE, "Speech Separation Based on Speaker Extraction Using Time-Frequency Masking" This paper introduces a time-frequency masking technique guided by speaker embeddings to isolatethe target speaker's voice from mixed audio. The approach is effective in enhancing the quality of speech input to ASR systems. Its main limitations include dependency on accurate speaker embeddings and potential delays introduced by the bedding separation step.

2.5 XIONG XIAO, SHAN LIU, LIANG LU, ENG SIONG CHNG, "Deep Learning for Target Speaker" The authors propose a deep learning approach for extracting the target speaker's voice using speaker embeddings. This method improves separation performance in mixed-speech environments and is applicable even in low-resource scenarios. However, it is sensitive to background noise and embedding quality, and may propagate errors into the ASR phase.

Proposed Model

To overcome the limitations of traditional cascade-based systems, the proposed system introduces an end-to-end Target-Speaker Automatic Speech Recognition (TS-ASR) framework using a Target- Speaker Recurrent Neural Network Transducer (TS-RNNT) along with Target-Speaker Activity Detection (TSAD). This unified architecture is designed to directly recognize the speech of a target speaker in real-time from audio streams that contain multiple speakers and background noise.

Key Components

Target-Speaker RNNT (TS-RNNT):

An advanced neural architecture that incorporates target-speaker embeddings directly into the RNNT encoder. This allows the model to condition the recognition process on who is speaking, not just what is being said. It effectively merges Target-Speaker Extraction (TSE) and ASR into a single model.

Target-Speaker Activity Detection (TSAD):

A lightweight module that detects whether the target speaker is active in the current audio segment. Helps suppress transcription of non-target speech, improving precision and efficiency. TSAD ensures that only relevant portions of the audio are processed by the TS-RNNT.

System Workflow

1. Reference Audio of the target speaker is provided.
2. The model listens to a multi-speaker audio stream.
3. TSAD detects when the target speaker is speaking.
4. TS-RNNT transcribes only the target speaker’s speech in real time. 5. Output is a clean, accurate transcription of the target speaker’s voice.

This study is carried out to evaluate the technical, operational, and economic impact that the proposed target-speaker speech recognition system will have. As real-world applications demand fast and accurate transcriptions from noisy, multi-speaker environments, traditional systems often fall short. Our project addresses this gap by introducing a unified and intelligent speech recognition pipeline capable of identifying and transcribing only the target speaker's speech with high precision.The proposed system integrates TS-RNNT and TSAD in a single framework, removing the need for separate voice separation and recognition modules. This simplification reduces latency and computational complexity, while improving transcription accuracy in real-time The technologies used are largely open-source and community-supported, reducing the economic burden. Cloud-based APIs and pre-trained models also support faster prototyping and deployment, making this system viable within reasonable infrastructure limits.

SYSTEM ARCHITECTURE

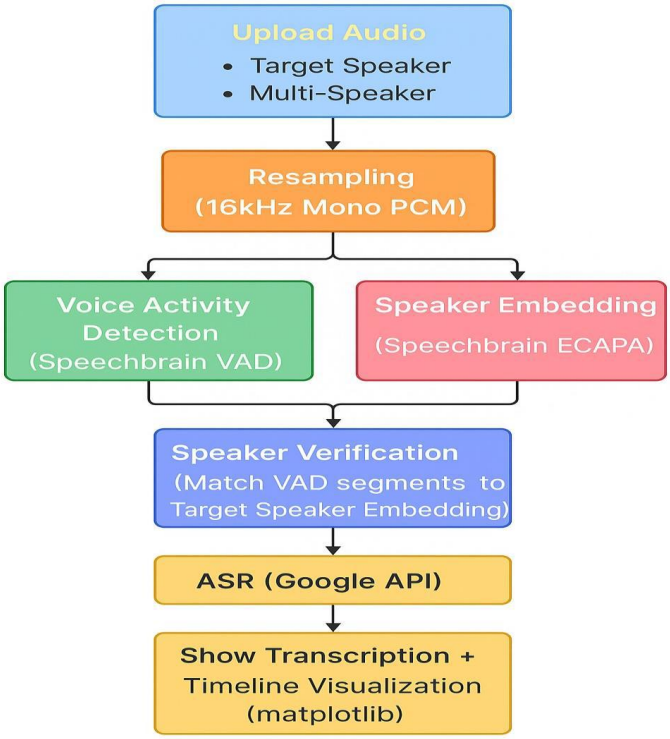


Fig 4.1: System Architecture

Audio Input

This is the first step of the system where two audio files are uploaded: Target Speaker Audio: A clean reference audio sample of the person we want to recognize in the noisy or mixed audio. This sample should contain only the voice of the target speaker. Multi-Speaker Audio: This is the test audio which may contain overlapping voices from multiple people, including the target speaker.The purpose of this module is to supply both the identity (via the clean sample) and the search area (via the mixed audio) for the system to operate on.

Re-Sampling

The uploaded audio files might differ in format or quality. This module ensures that all audio is standardized: Converts both audios to a sampling rate of 16 kHz, which is commonly used for speech processing. Converts stereo files to mono channel, making it easier to process using neural models. Ensures the format is PCM-encoded WAV, which is compatible with most speech-processing libraries and models. This step is crucial for consistent feature extraction and model compatibility.

Voice Activity Detection (VAD)

VAD is responsible for identifying which parts of the multi-speaker audio contain active speech (i.e., not silence or background noise). It works as follows: The system scans the audio and returns time intervals where speech is detected. Segments of interest are isolated and extracted from the original audio. By doing this, the system avoids wasting resources on silent regions and reduces the complexity of speaker verification in the next steps.

Speaker Embedding

In this module, the system extracts a speaker embedding (a fixed-length vector) from the clean target speaker audio. This is done using a pretrained speaker verification model (like ECAPA- TDNN). The

embedding is a digital signature of the target speaker's vocal characteristics — pitch, tone, accent, speaking style, etc. This embedding acts as a reference to identify the target speaker's voice in the mixed audio.

Speaker Verification Module

Each segment from the VAD output is analyzed to determine if it matches the target speaker: The same speaker embedding method is applied to each segment. Cosine similarity is calculated between each segment's embedding and the target speaker's embedding. Only segments with a similarity score above a threshold (e.g., 0.75) are retained. This module filters out all non-target speech, allowing the system to isolate only the voice of interest.

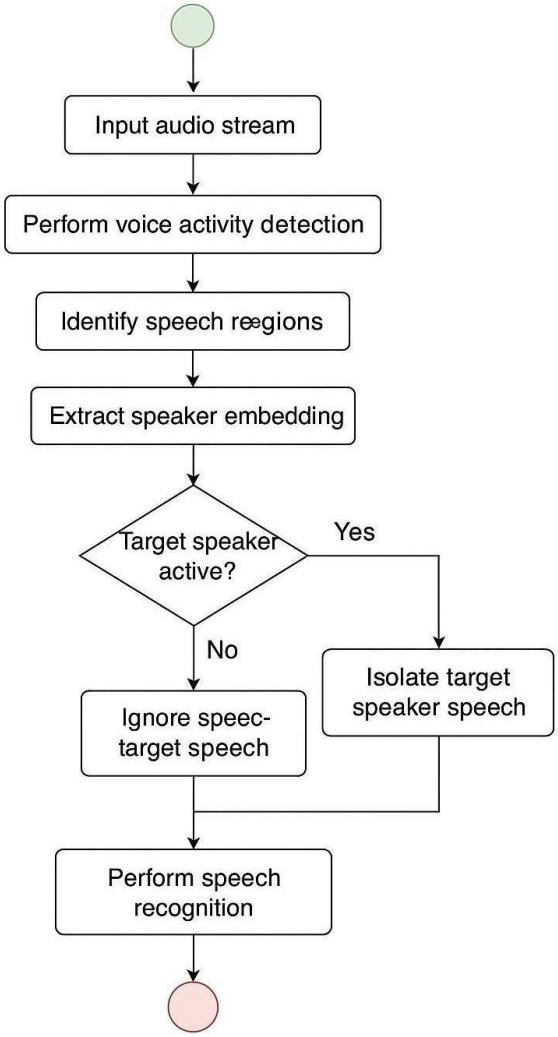
Segment Merging

After identifying and selecting the target speaker's segments: All verified segments are concatenated into a single audio stream. This final audio contains only the target speaker's speech, now free of background voices or noise. This step ensures that only the required audio is passed into the transcription module, improving accuracy and efficiency

Speech Recognition

After isolating and merging the segments containing only the target speaker's voice, the system sends this audio to the Google Speech-to-Text API for transcription. This cloud-based Automatic Speech Recognition (ASR) service by Google uses advanced deep learning models such as Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) to convert spoken language into written text. The API supports real-time and batch transcription, punctuation insertion, and speaker diarization (in multi-speaker mode), though in this system only one speaker is processed after filtering. It offers high accuracy across various languages and dialects, enabling multilingual transcription. The processed result is returned as a structured response containing the recognized text, word- level time stamps, and confidence scores.

An Activity diagram (also known as a workflow) provides a graphic overview of the business process. Using standardized symbols and shapes, the workflow shows step by step how your work is completed from start to finish. It also shows who is responsible for work at what point in the process. Designing a workflow involves first conducting a thorough workflow analysis, which can expose potential weaknesses. A workflow analysis can help you define, standardize and identify critical areas of your process. An event is created as an activity diagram encompassing a group of nodes associated with edges. To model the behavior of activities, they can be attached to any modelling



System Modules

Environment Setup:

The project requires the following libraries:

SpeechBrain for speaker recognition and VAD.

Torchaudio for audio processing.

Speech Recognition for ASR.

Streamlit for the web interface.

Audio Preprocessing:

The uploaded audio files are re-sampled to 16kHz, converted to mono, and normalized to float32 to ensure consistent processing

Speaker Recognition:

A speaker embedding is extracted from the target speaker's audio using a pretrained Speaker Recognition model. This embedding helps identify the target speaker in the multi- speaker environment.

Voice Activity Detection (VAD):

The system uses a VAD model to detect speech segments in the audio, filtering out silence and focusing only on speech portions

Automatic Speech Recognition (ASR):

The extracted target speaker's speech is transcribed into text using an ASR model. The transcription is displayed in the user interface

Results Display:

The system provides the following results to the user:

1. Transcription of the target speaker's speech.
2. Playback of the extracted speech.
3. .Download option for the transcription

Results & Analysis

The execution of the process will be explained clearly with the help of continuous screenshots.

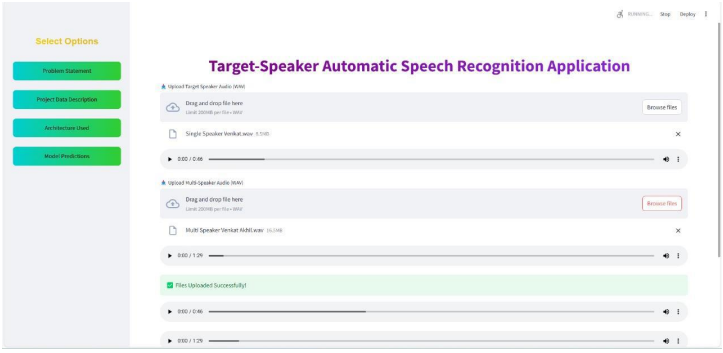


Fig 8.1: Audio inputs

Fig 8.2: output (VAD Visualization)

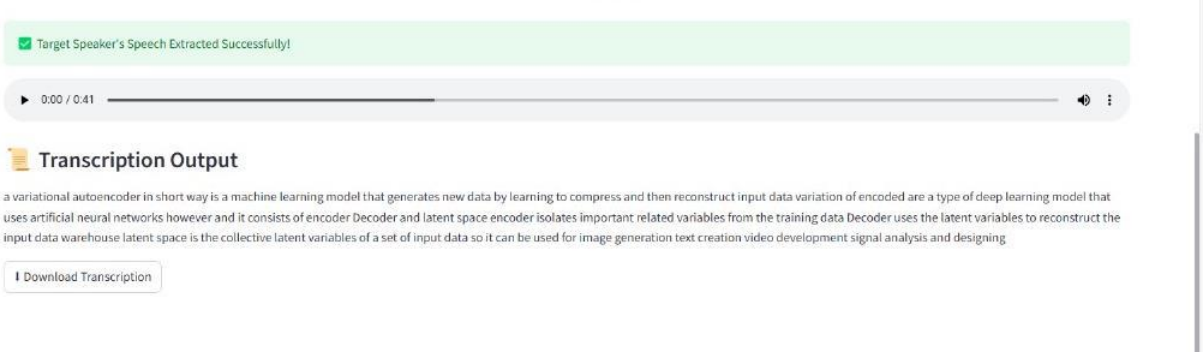


Fig 8.3: Transcription Output

Conclusion

The proposed system successfully implement Target-Speaker Automatic Speech Recognition pipeline that can identify, isolate, and transcribe speech from a specific speaker in a multi- speaker environment. By integrating advanced deep learning models such as CRDNN for Voice Activity Detection and ECAPA-TDNN for Speaker Verification, the system effectively filters out irrelevant speech segments. The final transcription is handled by a robust Automatic Speech Recognition (ASR) engine, ensuring accurate and efficient conversion of speech to text.This approach demonstrates practical applications in voice-driven interfaces, surveillance, personal assistants, and transcription services, especially where speaker-specific transcription is essential.

Future Scope

The system can be extended to support multi-language transcription by integrating multilingual ASR models such as OpenAI Whisper or Google's multilingual APIs. This would allow the application to transcribe speech from different languages, automatically detect spoken language, and handle code-switching (multiple languages in a single audio). Such enhancements will make the system globally adaptable and more inclusive for diverse users. To further enhance the capabilities of the system, language-specific tuning can be implemented to improve transcription accuracy for regional accents and dialects. This would ensure more precise recognition of diverse speech patterns. Additionally, by supporting low-resource languages, the system can promote greater linguistic diversity and inclusivity, enabling users from underrepresented communities to benefit equally. Furthermore, the system could be designed to adaptively learn individual user language preferences and usage patterns over time, leading to more personalized and efficient performance

References

1. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82–97.
2. Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. In Proceedings of Interspeech (pp. 5036–5040).
3. Sainath, T. N., He, Y., Narayanan, A., Botros, R., Pang, R., Rybach, D., Allauzen, C., Variani, E., Qin, J., Le-The, Q.-N., Chang, S.-Y., Li, B., Gulati, A., Yu, J., Chiu, C.- C., Caseiro, D., Li, W., Liang, Q., & Rondon, P. (2021). An efficient streaming non- recurrent on-device end-to-end model with improvements to rare-word modeling. In Proceedings of Interspeech (pp. 1777–1781).
4. Kurata, G., & Saon, G. (2020). Knowledge distillation from offline to streaming RNN transducer for end-to-end speech recognition. In Proceedings of Interspeech (pp. 2117 2121).
5. Li, B., Chang, S.-Y., Sainath, T. N., Pang, R., He, Y., Strohman, T., & Wu, Y. (2020). Towards fast and accurate streaming end-to-end ASR. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 6069–6073).
6. Chen, X., Wu, Y., Wang, Z., Liu, S., & Li, J. (2021). Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 5904–5908).

7. Moriya, T., Tanaka, T., Ashihara, T., Ochiai, T., Sato, H., Ando, A., Masumura, R., Delcroix, M., & Asami, T. (2021). Streaming end-to-end speech recognition for hybrid RNN T/attention architecture. In Proceedings of Interspeech (pp. 1787– 1791).
8. Moriya, T., Ashihara, T., Ando, A., Sato, H., Tanaka, T., Matsuura, K., Masumura, R., Delcroix, M., & Shinozaki, T. (2022). Hybrid RNN- T/attention- based streaming ASR with triggered chunkwise attention and dual internal language model integration. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 8282–8286).
9. Sato, H., Ochiai, T., Delcroix, M., Kinoshita, K., Kamo, N., & Moriya, T. (2022). Learning to enhance or not: Neural network-based switching of enhanced and observed signals for overlapping speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 6287–6291)
10. Delcroix, M., Watanabe, S., Ochiai, T., Kinoshita, K., Karita, S., Ogawa, A., & Nakatani, T. (2019). End-to-end SpeakerBeam for single channel target speech recognition. In Proceedings of Interspeech (pp. 451–455).