International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

An AI-Based Multi-Speaker Audio Transcription and Speaker Diarization System Using Deep Learning Models

D.Prashanth
Department of Computer Science and Engineering ,Rajiv Gandhi University of Knowledge & Technologies-Basar, India

T.Manohar
Department of Computer Science and Engineering , Rajiv Gandhi University of Knowledge & Technologies-Basar, India

P.Rajinikanth
DDepartment of Computer Science and Engineering ,Rajiv Gandhi University of Knowledge & Technologies-Basar, India

Dr.B. Venkat Raman
Department of Computer Science and Engineering ,Rajiv Gandhi University of Knowledge & Technologies-Basar, India

ABSTRACT

This paper presents an end-to-end deep learning-based system for multi-speaker audio transcription and speaker diarization. The proposed system addresses the challenge of identifying "who spoke when" in real-world multi-speaker conversations, where traditional transcription systems fail to distinguish between speakers. The system integrates advanced AI models including OpenAI's Whisper for automatic speech recognition and ECAPA-TDNN for extracting speaker embeddings. The extracted embeddings are clustered using agglomerative clustering to group similar voice segments, followed by cosine similarity-based matching for speaker identification.Unlike traditional approaches based on MFCC and Gaussian Mixture Models, the proposed method leverages modern deep learning techniques to improve accuracy and robustness in noisy and overlapping speech conditions. The system also provides visual representations such as speaker timelines and embedding cluster plots for better interpretability. Experimental results demonstrate that the proposed system effectively performs speaker diarization and transcription, making it suitable for applications such as meetings, interviews, and conversational analysis.
Multi-Speaker Transcription, Speaker Diarization, Whisper, ECAPA-TDNN, Speaker Embeddings, Agglomerative Clustering, Deep Learning, Speech Processing. [KEYWORDS]

1. INTRODUCTION

Recent advancements in artificial intelligence and speech processing have significantly enhanced the ability to analyze, interpret, and extract meaningful information from audio data. Speech recognition systems are widely used in applications such as virtual assistants, automated transcription services, call center analytics, and multimedia processing. However, most conventional systems focus only on converting speech into text and fail to distinguish between multiple speakers present in a conversation. In real-world scenarios such as meetings, interviews, and panel discussions, conversations typically involve multiple speakers, making it essential not only to transcribe speech but also to identify "who spoke when." This requirement leads to the concept of speaker diarization, which involves segmenting an audio stream into speaker-specific regions and assigning appropriate labels to each segment.Despite its importance, speaker diarization remains a challenging problem due to variations in speaker characteristics such as pitch, tone, accent, and speaking speed, along with external factors like background noise, overlapping speech, and rapid speaker transitions. Traditional approaches rely on handcrafted feature extraction techniques such as Mel Frequency Cepstral Coefficients (MFCC) combined with statistical models like Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). Although these methods have been widely used, they often struggle to capture complex speech patterns and fail to perform effectively in real-world environments.With the emergence of deep learning, significant improvements have been achieved in both speech recognition and speaker representation tasks. Modern models are capable of learning complex features directly from raw audio data, eliminating the need for manual feature engineering. In this work, we propose an end-to-end deep learning-based system for multi-speaker audio transcription and speaker diarization. The system integrates OpenAI's Whisper model for accurate speech-to-text conversion and ECAPA-TDNN for extracting robust speaker embeddings that capture unique voice characteristics. These embeddings are further processed using agglomerative clustering to group similar speaker segments, while cosine similarity is employed for accurate speaker identification. The proposed approach improves transcription accuracy, enhances speaker separation, and performs effectively in noisy and overlapping speech conditions. Additionally, the system provides structured transcripts along with visual representations such as speaker timelines and embedding clusters, making it suitable for real-world applications.

2. BACKGROUND AND METHODS

Speaker diarization and multi-speaker transcription systems involve multiple stages, including speech recognition, speaker representation, clustering, and speaker identification. Traditional approaches rely on handcrafted features and statistical models, which often struggle to handle real-world challenges such as background noise, overlapping speech, and variations in speaker characteristics.With the advancement of deep learning, modern systems have shifted towards data-driven approaches that learn complex representations directly from audio signals. These methods provide improved robustness, scalability, and accuracy compared to conventional techniques.

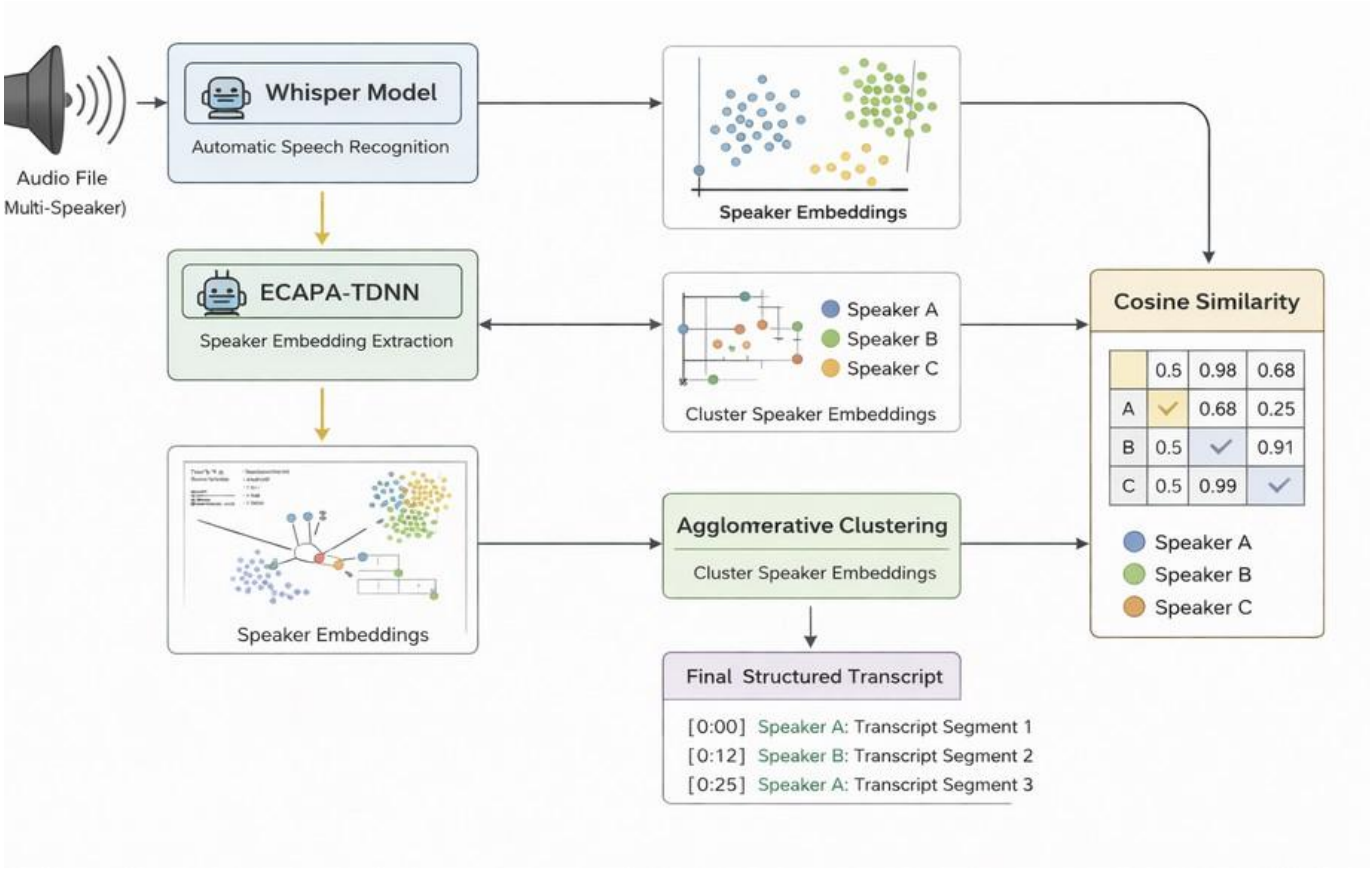

Fig. 1. Deep learning transcription and diarization flowchart

In this work, we adopt a deep learning-based framework that integrates advanced models for each stage of the diarization pipeline. The system combines automatic speech recognition using Whisper, speaker embedding extraction using ECAPA-TDNN, clustering using agglomerative hierarchical methods, and similarity-based speaker identification. The following subsections describe each component of the proposed system in detail.

2.1 Automatic Speech Recognition using Whisper

Automatic Speech Recognition (ASR) is a fundamental component of any speech processing system, responsible for converting spoken language into textual form. Traditional ASR systems are typically based on acoustic and language models that rely on handcrafted features such as MFCC. These systems often face limitations when dealing with real-world audio due to background noise, speaker variability, accents, and overlapping speech.In the proposed system, Whisper, a transformer-based deep learning model, is used for transcription. Whisper is trained on large-scale multilingual and multi-domain datasets, allowing it to generalize well across different types of audio inputs. Unlike traditional systems, it does not rely on manually engineered features but instead learns complex speech patterns directly from data.

One of the key advantages of Whisper is its robustness to noisy environments and its ability to handle long audio sequences. It generates accurate transcriptions along with time-aligned segments, which are essential for mapping spoken content to corresponding speakers. This time-stamped output forms the foundation for the subsequent speaker diarization process. Additionally, Whisper supports multiple languages and varying speech conditions, making it suitable for real-world applications such as meetings, interviews, and conversational recordings. Its ability to maintain high transcription accuracy significantly improves the overall performance of the system.

2.2 Speaker Embedding using ECAPA-TDNN

Speaker embedding plays a crucial role in distinguishing between different speakers in an audio stream. It involves converting segments of speech into fixed-length numerical vectors that capture the unique characteristics of a speaker's voice. Traditional methods rely on MFCC-based features combined with statistical models, which are often insufficient for capturing complex speaker traits.

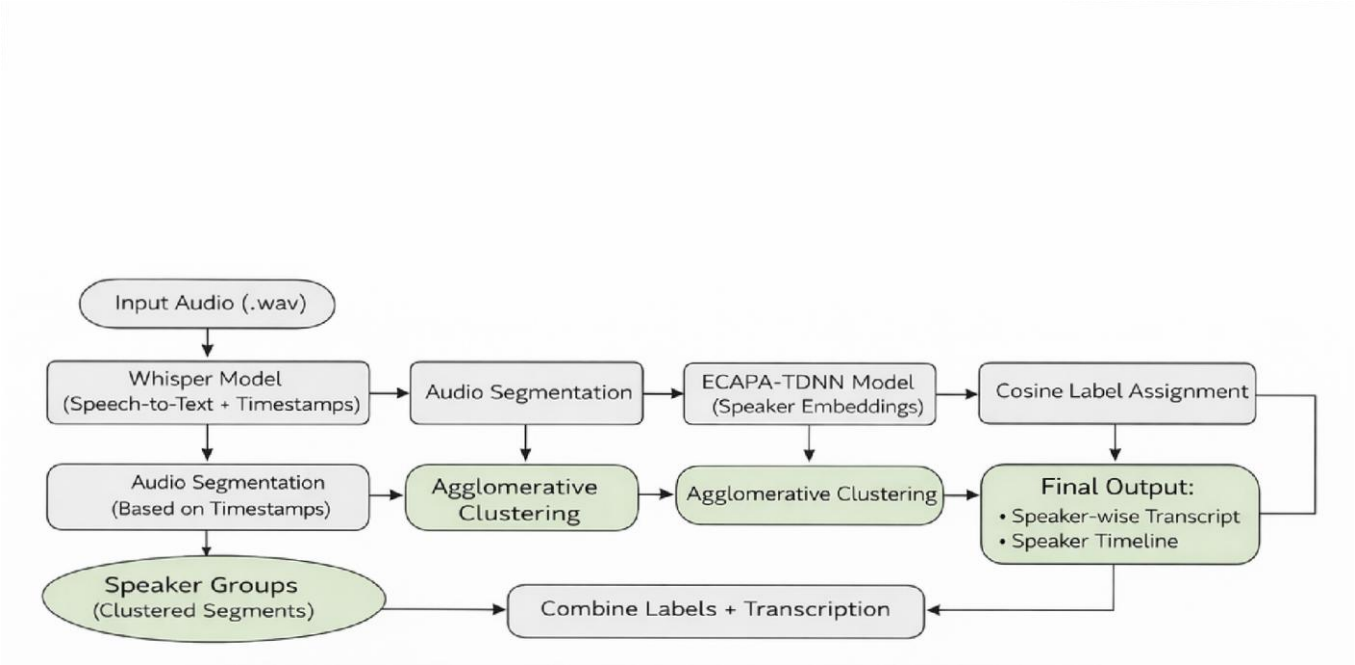

Fig 2. Flowchart of speaker diarization and transcription process

To address these challenges, the proposed system utilizes ECAPA-TDNN for speaker embedding extraction. ECAPA-TDNN is an advanced deep learning architecture designed specifically for speaker recognition tasks. It incorporates channel attention mechanisms, feature aggregation, and temporal modeling to capture both short-term and long-term speech characteristics. Each audio segment is processed by the model to generate a high-dimensional embedding vector. These embeddings are highly discriminative, meaning that segments from the same speaker produce similar vectors, while segments from different speakers are clearly separated in the embedding space.

Another important advantage of ECAPA-TDNN is its robustness to variations such as background noise, recording quality, and speaking style. This ensures reliable speaker representation even in challenging conditions. The quality of these embeddings directly impacts the effectiveness of clustering and speaker identification stages.

2.3 Speaker Clustering using Agglomerative Clustering

After extracting speaker embeddings, the next step is to group similar speech segments belonging to the same speaker. This is achieved through clustering, which is a key component of the speaker diarization process.In the proposed system, agglomerative hierarchical clustering is used due to its effectiveness in handling unknown numbers of speakers. This method follows a bottom-up approach, where each embedding initially forms its own cluster. The algorithm then iteratively merges clusters based on similarity until a stopping condition is reached.

One of the major advantages of agglomerative clustering is that it does not require prior knowledge of the number of speakers, making it highly suitable for real-world scenarios. It can dynamically adapt to varying audio inputs and speaker distributions.

The clustering process relies on similarity measures to determine which clusters should be merged. By grouping embeddings that are close in the feature space, the system effectively segments the audio into speaker-specific regions. This step is crucial for ensuring accurate diarization results.

2.4 Speaker Identification using Cosine Similarity

After clustering the speaker embeddings, the system needs to assign consistent speaker labels to each segment. This process is known as speaker identification, and it ensures that all segments belonging to the same speaker are labeled correctly throughout the audio.In the proposed system, cosine similarity is used as the primary metric for comparing speaker embeddings. Since embeddings are high-dimensional vectors that represent voice characteristics, cosine similarity measures how similar two vectors are based on the angle between them rather than their magnitude. This makes it particularly effective for speaker comparison, as it focuses on the direction (pattern) of the embeddings rather than their absolute values.

$$CosineSimilarity = (ei \cdot ej)/(||ei||||ej||)$$

A cosine similarity value close to 1 indicates that the two embeddings are highly similar and likely belong to the same speaker, while values closer to 0 indicate dissimilar speakers.

In practical implementation, a similarity threshold is defined (typically between 0.7 and 0.85). If the similarity between two embeddings exceeds this threshold, they are considered to belong to the same speaker. Otherwise, they are treated as different speakers.This step is essential for maintaining consistency across the entire audio, especially when a speaker appears multiple times at different intervals. It also helps in correcting minor clustering errors by refining speaker labels based on similarity scores.

Additionally, cosine similarity is computationally efficient and scalable, making it suitable for processing large audio datasets. Its simplicity and effectiveness make it a standard choice in embedding-based speaker recognition systems.

2.5 Integrated Diarization Pipeline

The proposed system follows an end-to-end pipeline that integrates multiple components to perform multi-speaker transcription and diarization efficiently. Each stage of the pipeline contributes to transforming raw audio into a structured and meaningful output. The process begins with the input audio, which may contain multiple speakers, background noise, and varying speech patterns. This audio is first processed by the Whisper model, which performs automatic speech recognition and generates a transcription along with precise timestamps. These timestamps play a critical role in segmenting the audio into smaller time intervals. Next, each segmented audio portion is passed through the ECAPA-TDNN model to extract speaker embeddings. These embeddings capture the unique vocal characteristics of each speaker and serve as the basis for distinguishing between different speakers. Once embeddings are generated, agglomerative clustering is applied to group similar segments together. This step organizes the audio into clusters, where each cluster ideally represents a single speaker. Since the number of speakers is not known beforehand, the clustering process dynamically determines the grouping based on similarity. After clustering, cosine similarity is used to refine the results by comparing embeddings within and across clusters. This ensures that speaker labels are consistent and reduces errors caused by incorrect grouping. Finally, the system combines transcription, timestamps, and speaker labels to produce the final output. The output is structured in a readable format, indicating which speaker spoke at a particular time along with the corresponding text. In addition to textual output, the system can also generate visual representations such as speaker timelines and clustering plots. These visualizations help in understanding speaker distribution and verifying the correctness of diarization. Overall, the integration of Whisper, ECAPA-TDNN, clustering, and similarity measurement creates a robust and scalable pipeline capable of handling real-world multi-speaker audio. Compared to traditional methods, this approach offers improved accuracy, better handling of noisy conditions, and enhanced interpretability.

3. SYSTEM ARCHITECTURE

The proposed system follows a unified and modular architecture designed to process multi-speaker audio and generate structured transcripts with accurate speaker identification. The system accepts a WAV audio file as input, which may contain multiple speakers, background noise, and varying speech patterns. Initially, the input audio is processed using the Whisper model, a transformer-based automatic speech recognition system. Whisper converts the audio into text and simultaneously provides timestamps for each spoken segment. These timestamps play a crucial role in segmenting the audio into smaller intervals corresponding to speech activity. Each segmented audio portion is then passed through the ECAPA-TDNN model to extract speaker embeddings. These embeddings are fixed-length vectors of size 192 that capture the unique vocal characteristics of individual speakers. The embedding space is structured such that segments from the same speaker are closely grouped, while those from different speakers are well separated.

The extracted embeddings are then processed using agglomerative hierarchical clustering. This clustering method groups similar embeddings into clusters, where each cluster ideally represents a single speaker. Since the number of speakers is not predefined, the clustering process dynamically determines the grouping based on similarity between embeddings. To ensure consistent speaker labeling, cosine similarity is applied across embedding vectors. This step refines the clustering results and assigns accurate speaker labels to each segment. Finally, the system integrates speaker labels, timestamps, and transcribed text to produce a structured output. The architecture is designed to be scalable, robust, and adaptable to real-world scenarios, making it suitable for applications such as meeting transcription, interviews, and conversational analysis.

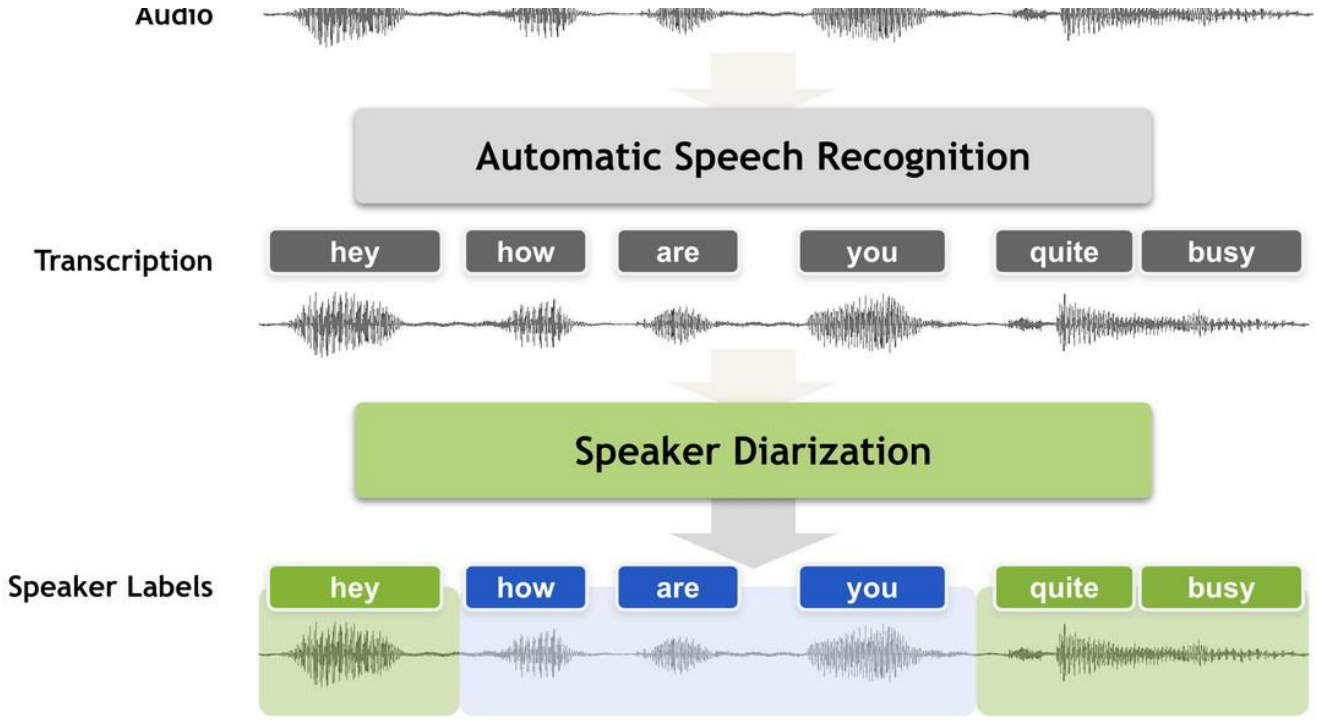

Fig. 3. Illustration of automatic speech recognition and speaker diarization process.

	Speaker	Start Time	End Time	Text
0	manohar	0:00:00	0:00:02	Hi, how are you today?
1	prashanth	0:00:02	0:00:04	I'm good, thank you. How about you?
2	manohar	0:00:04	0:00:06	I'm doing well. What are you working on these ...
3	prashanth	0:00:06	0:00:10	I'm currently learning Salesforce development ...
4	manohar	0:00:10	0:00:12	That sounds interesting. What kind of project ...
5	prashanth	0:00:12	0:00:16	It is a CPU system where we manage codes, disc...
6	manohar	0:00:16	0:00:19	Oh nice. That must involve Apex and Lightning ...
7	prashanth	0:00:19	0:00:23	Yes, exactly. I'm using Apex triggers and Ligh...
8	manohar	0:00:23	0:00:26	Great. I hope your project goes well.
9	prashanth	0:00:26	0:00:28	Thank you. I'm learning a lot from it.

Fig. 4. Generated speaker-wise transcript with timestamps from the proposed system.

4. SYSTEM WORKFLOW

The system workflow describes the sequential process through which the proposed system converts raw multi-speaker audio into a structured and labeled transcript. It outlines the interaction between different components of the system, including transcription, segmentation, embedding extraction, clustering, and speaker identification. The workflow is designed to ensure efficient processing and accurate diarization by integrating deep learning models with clustering techniques. Each stage of the workflow contributes to refining the output, ultimately producing a clear representation of speaker-wise dialogue along with corresponding timestamps. The following subsection provides a detailed step-by-step explanation of the workflow.

4.1 Workflow Description

The workflow of the proposed system represents a structured pipeline that transforms raw multi-speaker audio into an organized and speaker-labeled transcript. The process is designed to handle real-world audio conditions, including background noise, overlapping speech, and variations in speaker characteristics, while maintaining high accuracy and consistency. The workflow begins with a WAV audio file as input, which may contain multiple speakers interacting in a conversational setting. This audio is first processed by the Whisper model, which performs automatic speech recognition and generates transcribed text along with precise timestamps. These timestamps are essential for identifying speech boundaries and segmenting the audio into meaningful intervals.

Using the timestamp information, the audio is divided into smaller segments, where each segment corresponds to a specific portion of speech. This segmentation ensures that the system can process speech at a finer granularity, improving the accuracy of speaker identification. Each segmented audio portion is then passed through the ECAPA-TDNN model to extract speaker embeddings. These embeddings are fixed-length vectors of size 192 that capture the unique vocal characteristics of each speaker. The embedding space is designed such that segments from the same speaker are positioned closer together, while those from different speakers are well separated. Once embeddings are generated, agglomerative clustering is applied to group similar embeddings into clusters. This step plays a crucial role in identifying distinct speakers in the audio. The clustering process follows a hierarchical approach, where segments are iteratively merged based on similarity, without requiring prior knowledge of the number of speakers.After clustering, cosine similarity is used to refine the speaker labeling process. By comparing embeddings within and across clusters, the system ensures that segments belonging to the same speaker are consistently labeled. This step helps in correcting minor clustering errors and improves the overall accuracy of diarization. Finally, the system integrates speaker labels, timestamps, and transcribed text to generate the output. The output is presented in a structured format that clearly indicates which speaker spoke at a particular time, along with the corresponding text. Additionally, a speaker timeline visualization is generated to provide a clear view of speaker transitions throughout the audio. This workflow enables the system to perform end-to-end multi-speaker transcription and diarization efficiently, making it suitable for practical applications such as meeting analysis, interviews, and conversational intelligence systems.

Fig. 5. Speaker embedding clustering visualization showing separation of different speakers.

5. RESULTS AND ANALYSIS

conversational scenarios. The dataset included audio with varying speaker characteristics, background noise, and different speaking styles, allowing comprehensive testing of the system's robustness and adaptability. The Whisper model demonstrated strong performance in generating accurate transcriptions, even in the presence of noise and variations in speech patterns. The inclusion of timestamp information enabled precise segmentation of audio into meaningful speech intervals, which significantly improved the downstream diarization process. The ECAPA-TDNN model effectively extracted speaker embeddings that captured distinctive vocal features. These embeddings showed clear separability in the embedding space, enabling efficient differentiation between speakers. This representation played a critical role in ensuring accurate clustering and speaker identification. Agglomerative clustering successfully grouped similar speech segments into clusters corresponding to individual speakers.

The method dynamically adapted to different numbers of speakers without requiring prior knowledge, making it suitable for real-world applications. Cosine similarity further enhanced the clustering results by refining speaker labels and ensuring consistency across multiple segments belonging to the same speaker. The system generated structured outputs that included speaker-wise transcripts along with timestamps. These outputs clearly indicated speaker transitions and provided a coherent representation of the conversation. Additionally, the speaker timeline visualization offered valuable insights into speaker distribution, interaction patterns, and conversational flow. Overall, the results demonstrate that the proposed system achieves reliable performance in multi-speaker environments. The integration of deep learning models significantly improves transcription accuracy and speaker differentiation compared to traditional methods. However, certain limitations were observed in scenarios involving heavy speaker overlap or very short speech segments, where embedding quality and clustering accuracy may be affected.

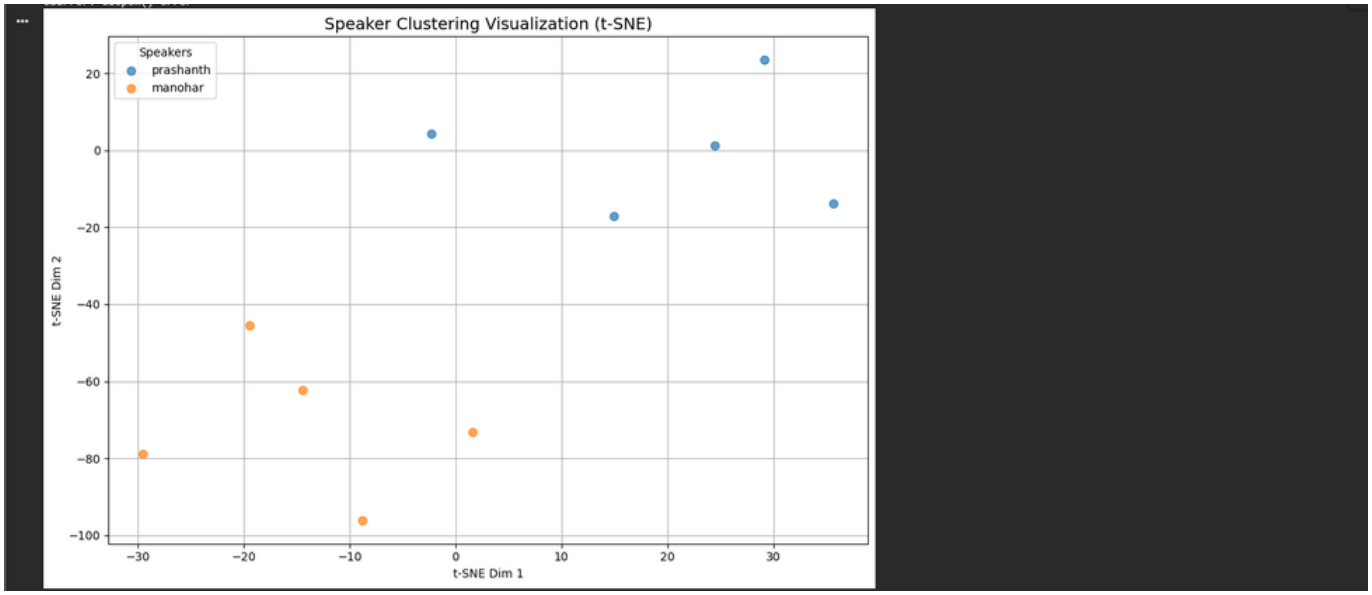

Fig. 6 Speaker Cluster Visualization

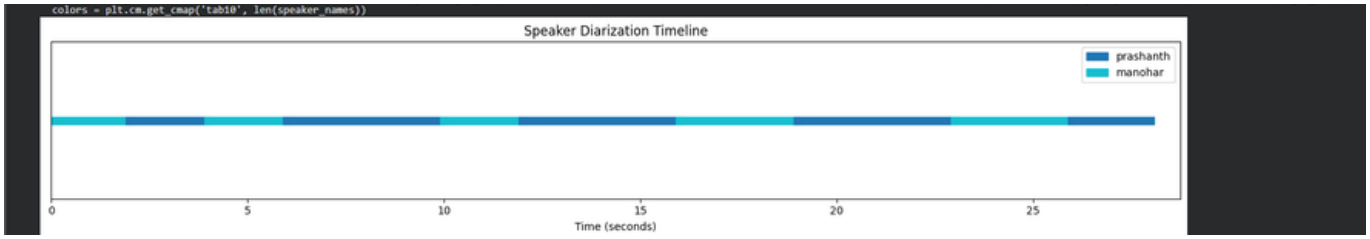

Fig 7. Speaker diarization timeline representing speaker transitions over time.

6. CONCLUSION AND FUTURE SCOPE

This paper presented a comprehensive deep learning-based system for multi-speaker audio transcription and speaker diarization. The proposed system integrates Whisper for accurate speech recognition, ECAPA-TDNN for robust speaker embedding extraction, agglomerative clustering for grouping speaker segments, and cosine similarity for consistent speaker labeling. By combining these components into a unified pipeline, the system effectively addresses the challenge of identifying "who spoke when" in complex multi-speaker audio. The results demonstrate that the system performs reliably across diverse audio conditions, including variations in speaker characteristics and background noise. The use of transformer-based models significantly enhances transcription quality, while embedding-based speaker representation improves clustering accuracy. Compared to traditional approaches, the proposed system offers better scalability, adaptability, and overall performance. Despite these advantages, certain challenges remain. The system may experience reduced accuracy in cases of overlapping speech, rapid speaker switching, or extremely short audio segments.

Additionally, the performance of clustering may depend on the quality of embeddings and the chosen similarity threshold. Future work can focus on several key improvements. One important direction is the implementation of real-time speaker diarization to support live applications such as meetings and streaming platforms. Another area of enhancement is the integration of speech separation techniques to handle overlapping speech more effectively. Incorporating more advanced transformer-based speaker models and self-supervised learning techniques can further improve embedding quality. Furthermore, extending the system to support multilingual environments and domain-specific adaptations can increase its practical applicability. The development of an interactive user interface and visualization tools can enhance usability for non-technical users. Optimizing the system for large-scale deployment and improving computational efficiency are also important areas for future research. Overall, the proposed system provides a strong foundation for intelligent multi-speaker audio analysis and opens new possibilities for advanced speech processing applications.

```
manohar 0:00:00
Hi, how are you today?
prashanth 0:00:02
I'm good, thank you. How about you?
manohar 0:00:04
I'm doing well. What are you working on these days?
prashanth 0:00:06
I'm currently learning Salesforce development and working on a project.
manohar 0:00:10
That sounds interesting. What kind of project is it?
prashanth 0:00:12
It is a CPU system where we manage codes, discounts and approvals.
manohar 0:00:16
Oh nice. That must involve Apex and Lightning components, right?
prashanth 0:00:19
Yes, exactly. I'm using Apex triggers and Lightning Web components.
manohar 0:00:23
Great. I hope your project goes well.
prashanth 0:00:26
Thank you. I'm learning a lot from it.
```

Fig. 8. Final structured multi-speaker transcript generated by the system.

7. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering for providing the necessary resources and support to carry out this research work. The authors also thank the project guide for their valuable guidance, continuous support, and encouragement throughout the development of this project.

8. REFERENCES

[1] A. Vaswani et al., "Attention is All You Need," Advances in Neural Information Processing Systems, 2017.

[2] OpenAI, "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," 2022.

[3] B. Desplanques et al., "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," Interspeech, 2020.

[4] D. Snyder et al., "X-vectors: Robust DNN Embeddings for Speaker Recognition," ICASSP, 2018.

[5] N. Dehak et al., "Front-End Factor Analysis for Speaker Verification," IEEE Transactions on Audio, Speech, and Language Processing, 2011.

[6] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.

[7] Scikit-learn Documentation, "Agglomerative Clustering," Available: https://scikit-learn.org

[8] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.

[9] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition," IEEE Transactions, 2010.

[10] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Transactions, 1981.

[11] J. S. Garofolo et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, 1993.

[12] H. Bredin et al., "pyannote.audio: Neural Building Blocks for Speaker Diarization," ICASSP, 2020.

[13] Y. Fujita et al., "End-to-End Neural Speaker Diarization," Interspeech, 2019.

[14] E. Vincent et al., "Speech Separation and Enhancement," Academic Press, 2018.

[15] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," International Conference on Learning Representations (ICLR), 2014.

[16] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

[17] T. N. Sainath et al., "Deep Convolutional Neural Networks for LVCSR," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.

[19] J. Chorowski et al., "Attention-Based Models for Speech Recognition," Advances in Neural Information Processing Systems (NeurIPS), 2015.

[20] V. Panayotov et al., "Librispeech: An ASR Corpus Based on Public Domain Audio Books," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[21] D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," International Conference on Machine Learning (ICML), 2016.

[22] Y. Liu et al., "Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Advances in Neural Information Processing Systems (NeurIPS), 2020.

[23] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019.

[24] A. Baevski et al., "wav2vec: Unsupervised Pre-training for Speech Recognition," Interspeech, 2019.

[25] J. Chung et al., "VoxCeleb2: Deep Speaker Recognition," Interspeech, 2018.

[26] A. Nagrani et al., "VoxCeleb: Large-Scale Speaker Identification Dataset," Interspeech, 2017.

[27] D. Povey et al., "Kaldi Speech Recognition Toolkit," IEEE Workshop, 2011.

[28] Y. Bengio et al., "Deep Learning," MIT Press, 2016.