SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V12I4P124

Volume 12 Issue 4, 298-307, April 2025

Original Article

Advancing Audio Processing and Emotion Recognition through Deep Learning Techniques

N. Venkata Sailaja[1], CH V K N S N Moorthy[2]*, Chiranjiva Rao Atluri[1], Gaddam Shiva Kumar Reddy[1], Gorugantula V S J Karthik[1], C S N V Ram Sree Santhosh[1]

[1]*Department of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India.*
[2]*Department of Mechanical Engineering, Vasavi College of Engineering, Hyderabad, India.*

[*2]*Corresponding Author : krishna.turbo@gmail.com*

Received: 19 February 2025 Revised: 21 March 2025 Accepted: 20 April 2025 Published: 29 April 2025

Abstract *- Speaker diarization is essential in audio processing, distinguishing and attributing speech segments to individual speakers in applications like speech recognition, customer service analytics, social media monitoring, and broadcast transcription. Despite advancements, challenges such as overlapping speech, varying acoustic environments, and noise resilience persist. This work proposes a deep learning-based approach using a pre-trained transformer model to handle multiple tasks, including feature extraction, Voice Activity Detection (VAD), segmentation, and speaker change detection. The system employs the ECAPA-TDNN model for speaker embeddings, followed by Agglomerative Hierarchical Clustering (AHC) for speaker labeling. Additionally, it performs sentiment analysis by converting segmented speech into text and applying a lexicon-based model. The implementation supports real-time audio recording, speaker segmentation, sentiment analysis, and visualization. Tested on multi-speaker recordings, it achieved 93.4% transcription accuracy while reducing computational costs, making it more feasible for standard hardware compared to traditional resource-intensive methods. By integrating speaker diarization and sentiment analysis within a unified framework, this research contributes to real-time, speaker-aware applications in various domains.*

Keywords *- Speaker Diarization, Deep Learning, Audio Signal Processing, Sentiment Analysis, Speech-to-Text, Transformer Models, ECAPA-TDNN, Agglomerative Hierarchical Clustering, Real-Time Speech Processing.*

1. Introduction

Segmenting and identifying distinct speakers within an audio recording or conversation is termed diarization of the speaker. This task is fundamental in automatically determining "who said what" in spoken dialogues. The inherent challenges in speaker diarization arise from speech variability, background noise, and situations where multiple speakers may be talking simultaneously. It plays a pivotal role in numerous fields and applications, including transcription services, where it is used to create accurate transcripts of meetings and interviews, customer service and call centers for quality assurance, forensic investigations for identifying speakers in recorded conversations, and market research to help understand group and individual behaviours. The methodology for the speaker diarization task includes multiple key stages. At first, a self-attention-based transformer model [1], pre-trained on large and diverse audio datasets, is used to transcribe or translate the audio input into text. The encoder block of the transformer processes the log-Mel spectrogram to extract features from the audio data. The decoder block generates the text output based on the encoded representation from the encoder. The module now utilizes a speaker-embedding model, which is trained prior to generating numerical representations of each speech segment. The following step is a clustering algorithm - agglomerative hierarchical clustering, which groups the feature vectors into clusters, each representing a potential speaker. Once clusters are formed, speaker identification is performed by assigning clusters to specific individuals, often utilizing speaker recognition techniques. A refinement step can be employed to enhance speaker identification accuracy by considering speech energy, pauses, and speaker-specific characteristics. In the end, the diarization algorithm provides an output that includes speaker labels and timing information for further analysis or transcription by adapting to the complexity and quality of the audio data. Additionally, identify the speaker's emotions and represent them using an Interactive User Interface.

1.1. Related Work

Speaker diarization is the process of splitting an audio recording into different, uniform speech segments and

determining the exact moment each speaker begins and ends. Automatic recognition of speaker boundaries helps in transcription and analyzing conversations, meetings, or any spoken content whilst performing distinguishing among speakers. A generic framework for speaker diarization comprises a series of key stages designed to split audio recordings into homogeneous speech segments while accurately determining each speaker's speech boundaries.

The computational load is reduced, and more emphasis is built around speech segments by identifying speech and non-speech regions with each segment.

VAD algorithms are used to achieve this purpose [2]. The generic framework for speaker diarization is mentioned in Figure 1, it shows the series of essential stages that form a framework that serves as the foundation for various systems.

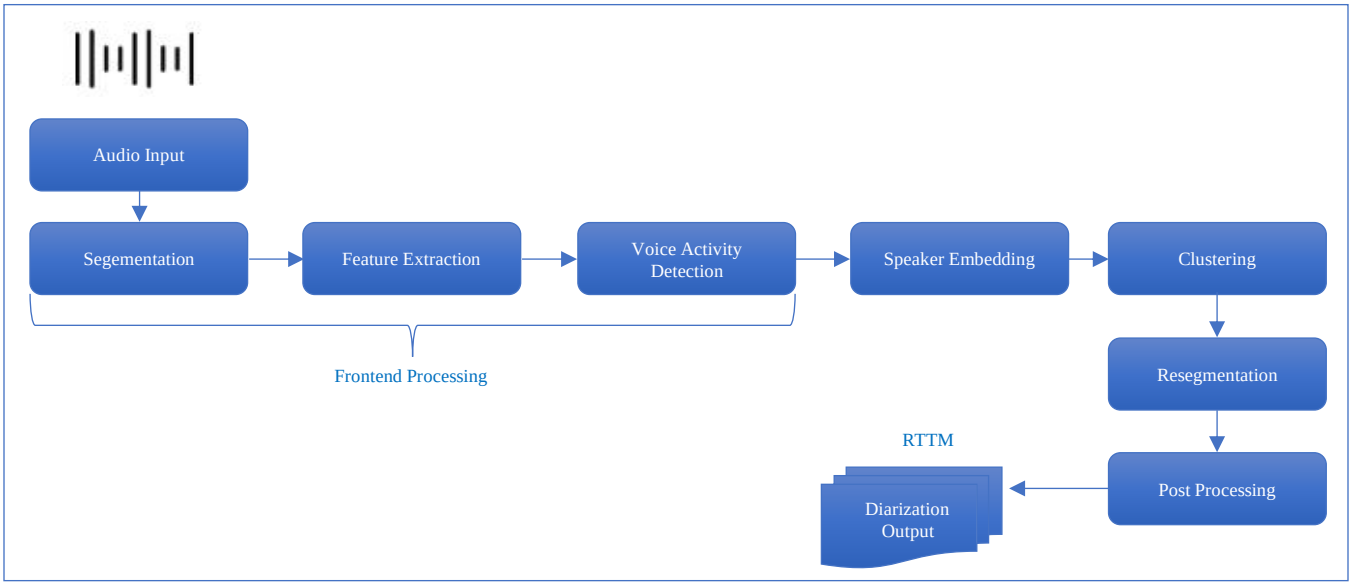

Fig. 1 Generic framework for speaker diarization

2. Traditional Techniques for Speaeker Diarization

The first part of the section is the study of Traditional methods and Deep-learning-based techniques. The next part constitutes the Hybrid Techniques along with improved versions of a few stages in the Speaker Diarization, like Speaker Segmentation.

2.1. Traditional Methods

A model proposed by Federico et al. employs speaker diarization through the Bayesian Hidden Markov Model (HMM) [3] aggregating of x-vector patterns. One kind of speaker embedding trained with a deep neural network is called an X-vector. A statistical technique for data clustering that considers data uncertainty is called Bayesian HMM clustering. This method outperforms conventional speaker diarization techniques in terms of resilience to noise and overlapping speech. It may, however, take longer than more conventional speaker diarization techniques. A model for speaker diarization using discriminative training of deep neural network language models was proposed by Chen et al. [4]. Using a machine learning technique called discriminative training, a model is trained to minimize error on a particular task. Comparing this method to conventional speaker diarization techniques, it is more accurate. On the other hand, training with it may cost more computationally.

2.2. Deep Learning-Based Methods

Garcia-Romero et al. [5] proposed a model that uses deep neural networks to extract embeddings from audio recordings 2018. These embeddings are then clustered to distinguish speakers in the audio. This approach is more robust to noise and overlapping speech. Huang et al. [6] developed a model for identifying speaker fragments in speech audio in 2019. The model uses a regional proposal network with a deep-learning neural network to classify the audio chunks. It produces a more accurate and robust diarization system than those used in the past. Fujita et al. [7] suggested training a deep neural network for speaker diarization with permutation-free objectives in 2020.

This model does not take all possible combinations of speakers, which makes it advantageous. This approach has improved over traditional systems in terms of scalability and performance. Singh and Ganapathy et al. [8] advanced a model using self-supervised deep learning-based hierarchical clustering in 2020. This method is robust to noise and overlapping speech and does not require labelled datasets. However, it is a computationally expensive approach in terms of training time. A combination of sophisticated Deep-Learning models was used to form a joint diarization system. This method enhances the system's performance by enabling the three disjoint tasks to learn from the data simultaneously.

This model needs a huge amount of data to be trained. [4, 9, 10]. A model for speaker identification in emotional and stressful settings that uses hybrid models based on DNN-based hybrid models [11]. This approach combines the advantages of deep learning and traditional speaker verification methods to achieve better performance in challenging environments. However, training and deploying are more complex than traditional speaker verification methods. A model that uses a convolutional neural network [13, 14] for speaker change detection in telephone speaker diarization systems. This approach is more robust to noise and overlapping speech than traditional speaker-change detection methods. However, a large amount of labelled data is required to train the model.

2.3. Hybrid Methods

Burget et al. [15] presented a system for speaker diarization that draws advantages from both deep learning and traditional speaker diarization. The team has used neural networks to extract the audio features and the Bayesian hidden Markov Model to cluster the speech segments based on their respective features. The hybrid system, which was developed by the team, produced state-of-art results on the DIHARD dataset [16]. The complexity of training, however, was computationally expensive compared to the traditional ones.

2.4. Speaker Embeddings Extraction

Speaker Embeddings are vectorial representations that are extracted from speech signals. These things help uniquely determine a speaker's speech identification, consequently aiding us in classifying different speakers involved in the audio recording. Convolutional and Recurrent Neural Networks (CNNs & RNNs) are commonly used to achieve this objective. Methods like x-vectors and d-vectors are commonly used to generate speaker embedding. The compact representations provided by these methods assist in clustering and speaker identification. [17]

2.5. Methods for Speaker Segmentation

Teimoori et al. [12] proposed an unsupervised learning-based approach for speaker segmentation in 2021. A pre-trained speaker embedding model is used to generate pseudo-labels for the input audio recording, which are subsequently used to train a speaker segmentation model. This method achieves good performance on speaker segmentation. The developed model is still expensive to train in terms of time. One advantage of using this system is the elimination of labelled datasets to train the model. A technique for LSTM neural networks, acoustic and language modelling, and speaker segmentation was used [18]. The audio properties and features are characterized using an LSTM model. LSTM maintains track of every audio segment to determine every speaker's unique identity. The approach is suitable for noisy environments. However, it is more complex to train and deploy than the unsupervised method-based help-training method described above.

2.6. Methods for Overlap Detection

Snyder et al. [5] devised a deep neural network-based method for overlapped speech detection in 2016. The system distinguishes between overlapped speech and non-overlapped speech by taking every other speech segment into account while training the neural network. The accuracy and performance were better than those of traditional systems. The requirement for a huge volume of data remains a setback for the system. Neeraj et al. [19] presented a technique for overlap detection in 2018. It uses LSTM to detect overlapped speech in the audio recording. In multi-speaker environments, there is always a probability of speech overlapping. It is a critical challenge in the speaker's diarization systems. LSTM models stand as good choices for handling overlapped speech detection by detecting the overlapping in each aural segment during training. This approach results in an accurate and noise-tolerant diarization system. The dataset must be large enough to train the model, which remains a drawback.

3. Traditional Techniques for Sentiment Analysis

Jbene et al. [20] proposed a novel sentiment analysis system, BE-Att-BiLSTM, designed exclusively to analyze sentiment in conversational systems in 2016. This model captures the nuances of how words are used within conversations by using the capabilities of pre-trained BERT contextual embeddings. The system analyzes the sequential nature of conversation text by deploying a Bidirectional Long Short-Term Memory (BiLSTM) layer. It also features an attention mechanism that focuses on important information to improve the sentiment classification further. Additionally, the authors explore text augmentation techniques to improve the model performance. However, the system still acknowledges several limitations. The impact of text augmentation techniques has not been extensively analyzed. This necessitates further investigation into the model's effectiveness. García-Ordás et al. [21] proposed an FCN-based approach for sentiment analysis in audio data in 2020. This approach is capable of processing variable-length audio files, which makes it suitable for the development of real-time applications. The usage of MFCC features outperforms Mel spectrograms in capturing emotional hints. The model achieves excellent accuracy on benchmark datasets. However, the model struggles to distinguish and differentiate between similar emotions. Future efforts could perhaps focus on enhanced differentiation of similar sentiments and emotions, exploring generalizability on broader datasets. Jain et al. [22] proposed a Support Vector Machine (SVM)-based system for speech emotion recognition in 2021. This system investigated the impact of feature extraction methods, classification strategies, and dataset quality. The system uses Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Prediction Cepstral Coefficients (LPCCs) for feature extraction. It achieves higher accuracy with MFCCs. The gender-dependent approach outperforms the One-against-all approach for classification, indicating the importance of

incorporating speaker gender information. Data quality is a major determinant of emotion recognition accuracy. The system's overall accuracy is 85.085%, performing much better with professional recordings than with student recordings. The study highlights the potential of using an SVM-based approach. However, future work could explore combining MFCCs with other features like Mel-Energy Cepstral Coefficients (MEDC) and investigate real-world applications such as emotion-based call routing in voicemail systems.

3.1. Research Gap

The majority of current diarization algorithms either just consider speaker segmentation without taking into account the speakers' emotional states, or they have high processing overheads that prevent them from being used in real-time, resource-constrained settings. Additionally, speaker diarization and sentiment analysis are not fully integrated into a single effective framework.

4. Proposed System

This study proposes a novel and innovative speaker diarization system that minimizes the issues of noisy environments and computational expense while also improving accuracy and efficiency, which were seen as some of the major limitations of traditional approaches. The proposed system uses sophisticated techniques for reliable outcomes across various audio sources. The presented solution includes vital components such as feature extraction approaches tailored for efficiently differentiating speakers, powerful clustering algorithms capable of handling altering acoustic conditions, and flexible models that adjust to speaker changes over time. Another distinguishing feature of the system is its adaptability to different audio content and recording conditions, which combines the best elements of supervised and unsupervised learning techniques to devise a system that can automatically adjust to new speakers and settings, eliminating the need for manual annotation and human intervention. We also introduce a mechanism for detecting the tone of the individual speakers through a simple lexicon-rule-based sentiment analysis module for a better understanding of the context. Through efficient implementation and optimization techniques, we ensure that our system can process large volumes of audio data with minimal computational overhead, making it practical for deployment in resource-constrained environments. The functionality of the modules and architecture of the proposed system is described in the following section.

4.1. Novelty of the Proposed Work

The suggested study presents a novel integrated system that conducts real-time sentiment analysis and speaker diarization. It greatly increases computational efficiency by utilizing a pre-trained transformer model for effective speaker change detection, Voice Activity Detection (VAD), and feature extraction. Agglomerative Hierarchical Clustering (AHC) in conjunction with ECAPA-TDNN embeddings guarantees reliable speaker labeling even in noisy environments, surpassing conventional GMM-HMM and x-vector-based methods. An emotional component frequently overlooked in previous diarization research is added by the system's integration of a lightweight, lexicon-based sentiment analysis module on recorded speech segments.

The system retains minimal computing costs, allowing real-time operation on conventional hardware, and achieves 93.4% transcription and diarization accuracy on multi-speaker, non-overlapping recordings. This study provides a unified, scalable, noise-resilient solution, in contrast to earlier approaches that either need intricate multi-stage pipelines or concentrate on a single task. It has better contextual richness and deployment readiness than previous systems [5-8, 13-15]. An important step toward intelligent, practical audio processing applications is the combination of speaker diarization with emotional characterization.

5. Methodology

Speaker diarization is the process of identifying who is speaking in an audio recording. Figure 2 illustrates the architecture of the proposed system. The steps involved in the system, as depicted in the diagram, are as follows:

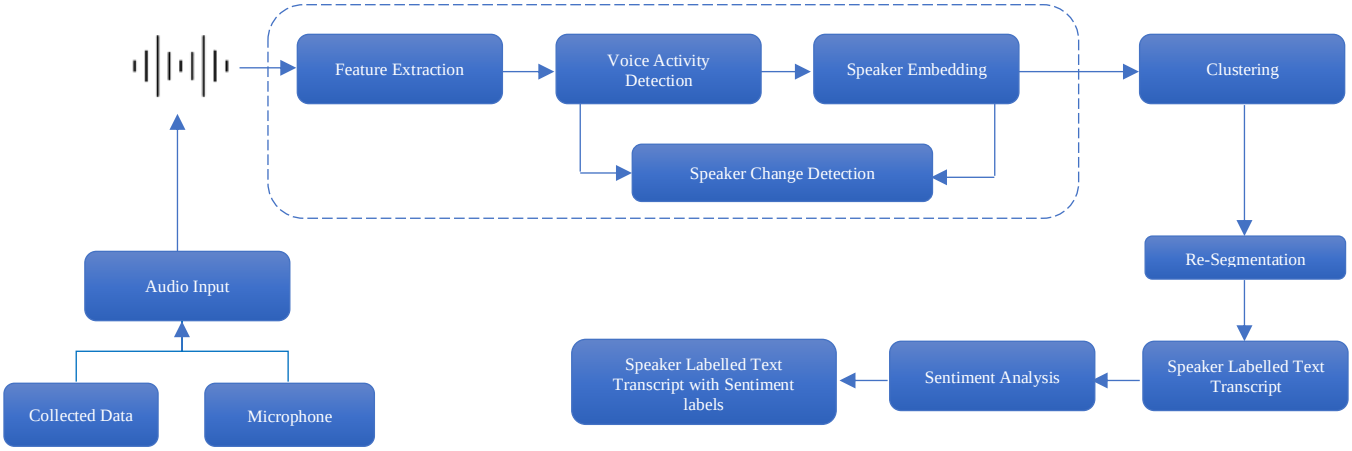

Fig. 2 Proposed system architecture

- Audio Input: This is the system's starting point. The audio is collected and given as input to the system. The collected audio data is then processed through several steps. Audio is thoroughly pre-processed to undergo further steps in the system. A noise reduction module is used to process the audio so that any background noise or distortion in the audio is eliminated. This makes the audio robust to disturbances that are common in real-time scenarios. The raw audio is converted into a spectrogram here. The system uses single-channel audio throughout instead of multi-channel audio. If multi-channel audio is used, we convert it into single-channel audio.

- Feature Extraction: The next step specific features are extracted from the audio data. These features could include things like the volume, pitch, and timbre of the speaker's voice. In this step, the model mathematically captures the speaker's vocal characteristics associated with a specific speech segment. The processed audio is divided into small segments, and within each segment, properties like Mel-Frequency Cepstral Coefficients (MFCCs) or pitch are calculated in this step. We have associated 192 features for each feature vector. The feature vector for each speech segment serves as a fingerprint-like profile that enables the model to differentiate between speakers.

  A 192-dimensional feature vector represents each voice segment during the feature extraction phase. This dimensionality was used to strike a compromise between preserving computational efficiency for real-time processing and capturing rich speaker-specific information. Several Low-Level Descriptors (LLDs) that describe basic speech characteristics, such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, spectrum flux, spectral entropy, formant frequencies, and their first- and second-order derivatives, are used to create the 192 features that were chosen.

  Incorporating dynamic information, such as derivatives, into the model aids in capturing speaking styles and temporal variations, both of which are essential for differentiating distinct speakers with comparable vocal traits. The ECAPA-TDNN model produces highly discriminative embeddings with 192 features, preventing overfitting and processing bottlenecks.

  While significantly higher dimensions (>300) would needlessly increase clustering complexity and memory needs without yielding significant performance advantages, choosing a lower dimensionality (e.g., <100) could result in the loss of crucial spectral and prosodic information essential for successful diarization. Consequently, a 192-dimensional representation offers a useful and efficient feature space for strong speaker distinction in a range of acoustic circumstances.

- Voice Activity Detection (VAD): This step attempts to identify the parts of the audio recording that contain speech and separate them from the parts that do not. This step acts as a doorkeeper. VAD analyzes the audio energy and spectral content to distinguish faint whispers and clear speech segments. This filtering step is very important because it ensures that the model is not overwhelmed by background noise and can effectively analyze the speaker's unique speech signal, even when presented in silent tones.

- Speaker Change Detection: This step attempts to identify the points in the audio recording where the speaker changes. This information can segment the audio recording into different speaker turns.

- Speaker Embedding: At this point, the system generates a mathematical representation, or embedding, for each speech segment. These representations are used for speaker diarization in the clustering step. We have used a Time Delay Neural Network (TDNN) model pre-trained on the VoxCeleb dataset to learn speaker-specific representations. The model takes the Mel Frequency Cepstral Coefficient (MFCC) as input, which captures the spectral envelope of the audio signal. The TDNN architecture helps effectively capture both short-term and long-term temporal dependencies of speech features, resulting in robust speaker embeddings. This step inter-cooperates with previous steps of feature extraction, speaker change detection and VAD. We have attempted to collectively handle the previous three steps using a separate model.

- Agglomerative Hierarchical Clustering (AHC) was chosen for the speaker diarization clustering phase because of its unique benefits within the framework of our suggested method. AHC is more suited for speaker embeddings that might create non-convex or uneven clusters because it does not rely on assumptions about spherical clusters, unlike K-means and other approaches that do. Furthermore, considering the possible variability and distortions found in real-world audio recordings, AHC's resilience to noise and outliers is crucial.

  AHC continuously maintains performance without requiring careful hyperparameter adjustment, in contrast to density-based techniques like DBSCAN, which may suffer when clusters have different densities (common in speech data due to fluctuating speaker loudness or background noise). Despite their strength, spectral clustering techniques are frequently computationally costly and impractical for real-time applications on conventional technology. Furthermore, when the number of speakers is not precisely known in advance, AHC's hierarchical structure enhances adaptability by allowing

flexibility in dynamically determining the number of clusters based on a distance threshold if necessary.

Overall, the application of AHC coincides with the goals of the proposed system: achieving high clustering accuracy, robustness to audio fluctuation, little computing cost, and real-time practicality - all required for scalable speaker diarization and sentiment analysis on varied.

- Clustering: This is one of the crucial steps of our system. We used the Agglomerative Hierarchical clustering step to perform the speaker diarization. In this step, we cluster and assign the speech segments to the respective speaker based on the similarity index. The algorithm is fed with the embedding generated in the previous step and the expected number of speakers in the meeting.

- Re-segmentation: This step addresses potential shortcomings in the primary segmentation by re-evaluating speaker boundaries.

- Sentiment Analysis: The speaker diarization model of our system generates a text transcript of the audio recording using a transformer-based speech-to-text transcription model, labelled with speaker labels determined in the clustering stage for each speech segment. Subsequently, we perform the sentiment analysis of the speaker's words in this step. We perform it on each speaker, and the overall conversation happens in the audio recording. We use a model that utilizes a lexicon-based approach combined with machine learning to perform the sentiment analysis. In this step, sentiment labels are generated and assigned to the speech segments.

6. Module-Level Functionality

This section intends to provide a brief overview of the core working modules in our system. Our system consists of the following modules that form the basis of the core functionality: the User Interface, the Speaker Diarization Module, and the Sentiment Analysis Module. We integrate these independent modules collectively to create an end-to-end speaker diarization and sentiment analysis application. The working modules are described as follows:

6.1. User Interface Module (UI)

The main functionality of this module is to take the user's audio input. The input can be taken by uploading a pre-recorded audio file or recording it live. Our system allows the user to upload audio recordings in various formats. Once the user gives the audio input to the system, the noise in the audio is reduced using the spectral grating method. The pre-processed audio is sent to the speaker diarization module, followed by a sentiment analysis module for speaker tone detection. The results of speaker diarization are displayed back to the user through the User Interface (UI). Our system generates a speech-to-text transcription file as output, which contains a transcript of all speech segments spoken by the speakers in the audio. Each speech segment is labeled with its start and finish timestamps, predicted speaker label (the label of the speaker who delivered this specific speech segment), and predicted sentiment label. The output also shows the sentiment label of all the speech delivered in the audio recording and the sentiment label for each speaker.

6.2. Speaker Diarization Module

The speaker diarization module receives the audio input from the user interface and performs speaker diarization. It leverages the advantages of several state-of-the-art deep learning models and machine learning algorithms to achieve high accuracy and robust performance. At first, a self-attention-based transformer model [1], pre-trained on large and diverse audio datasets, is used to transcribe or translate the audio input into text. The encoder block of the transformer processes the log-mel spectrogram to extract features from the audio data. The decoder block generates the text output based on the encoded representation from the encoder. The idea of using a powerful transformer model is to handle multiple vital tasks at once, including feature extraction, speech recognition, voice activity detection, and speaker change detection. This maximizes the efficiency and performance of the system with less computational overhead and processing time. It is often helpful for accurate transcription and speech segmentation because it can capture complex relationships between sounds by processing the complete data sequence at once.

The transformer model gives all the speech segments spoken by all the speakers sequentially (each speech segment can be thought of as a properly structured sentence). Subsequently, the module now utilizes a pre-trained speaker-embedding model to generate numerical representations of each speech segment. Each numerical representation is a vector of 192 features, including pitch, shrill, tone, rhythm, frequency, etc. The underlying model used for speaker-embeddings is ECAPA-TDNN (Emphasized Channel Attention in Time-Delay Neural Network) [23]. The ability to suppress irrelevant channels and focus only on informative ones using channel-wise self-attention mechanisms makes this model more efficient and more effective than other models. This self-attention mechanism also aids in creating robust and strongly discriminative embeddings, which in turn results in the effective acquisition of discriminative speaker-related information across various frequency bands. The model is tolerant to background noise, distortion, and variability. It effectively captures the temporal dependencies in speech signals, enabling the generation of more informative speaker embeddings. The model’s minimum computational overhead and efficiency make it suitable for real-time applications where accurate and reliable representation of speakers is important. Following the generation of speaker embeddings for every speech segment, a clustering algorithm is finally used to predict the speaker label for each speech segment. We have used Agglomerative Hierarchical Clustering (AHC) for

our system. We provide the speaker embeddings generated in the earlier step to the clustering model, and to improve the accuracy of clustering, the number of clusters in the audio recording is also given. The reason for choosing AHC is that it tends to be robust to noise and outliers and can handle large datasets well. This quality makes it suitable for real-time applications. The system is computationally less expensive and performs better because of AHC's simple implementation, which also has minimal computational overhead. This step marks the end of the speaker diarization module by determining who spoke what and assigning speaker labels to each speech segment.

6.3. Sentiment Analysis Module

After performing the speaker diarization, each transcribed speech segment is passed to a lexicon and rule-based sentiment analysis model [24] to determine the sentiment of the speech to better understand and interpret the tone/context. The model comes with a pre-built lexicon that eliminates the need for training data. Its rule-based approach enables it to capture complex sentiments even in informal text. This model is computationally efficient, making it suitable for real-time analysis. It can be adapted to diverse domains and provides a sentiment intensity score for a more nuanced understanding of the text. The sentiment assigned can be neutral, happy, fearful, surprised, or angry. Our system also computes the sentiment of the overall speech in the audio input and for individual speakers by aggregating all the speech they delivered in the audio input.

7. Implementation

The following algorithms are introduced to implement the proposed system architecture shown in Figure 2.

Algorithm for Speaker Diarization

Input: Path of the Audio file, number of speakers involved in the audio recording, path location of the output file.

Output: A text file containing labeled segments indicating speaker identities and corresponding transcriptions of speech segments along with their timestamp boundaries.

1. Check if the input audio file is in WAV format. If not, convert it into WAV format.
2. Load the audio data and extract its length.
3. Convert multi-dimensional audio to single-dimensional audio if necessary.
4. Initialize the model for speech-to-text transcription.
5. Transcribe the audio and get the raw segments.
6. Filter out segments that fall within the audio duration.
7. Read the WAV file to get the number of frames and frame rate to calculate the audio duration.
8. For each speech segment, do the following:
    a. Take a segment of audio data as input.
    b. Extract the waveform within the segment.
    c. Generate embedding using the embedding model and store them.
9. Perform clustering using Agglomerative Clustering on the embeddings generated in step 10 to determine speaker identities.
10. Label segments with speaker identities and filter out segments with low speech probability.
11. Write the labeled segments in the output text file.
12. Return the output text.

Algorithm for Sentiment Analysis

Input: Text associated with transcribed speech segments

Output: A label indicating the sentiment associated with the text segment.

1. Load the sentiment analysis model.
2. Break down the input text into individual tokens (words or phrases).
3. Look up each token in the lexicon containing a list of words along with their associated sentiment scores.
4. For each token found in the lexicon, retrieve its sentiment score.
5. Determine the sentiment intensity of the text by aggregating individual scores.
6. Calculate a compound score by summing up the normalized scores of all the words in the text.
7. Based on the compound score, categorize the sentiment into predefined categories using predefined thresholds and rules to map the compound score to sentiment categories.
8. Return the sentiment category along with the sentiment scores calculated for the input text.

8. Dataset Description

The speech-to-text transcription pre-trained model is pre-trained on 1 million hours of labelled audio data from web sources. Our system's speaker embedding pre-trained model is trained on the benchmark VoxCeleb1 and VoxCeleb2 training data [25, 26].

8.1. Evaluation Metrics

The accuracy is obtained by calculating the Word Error Rate (WER) and subtracting it from a whole. The formulae used are as follows:

$$WER = \frac{I+D+S}{N} * 100$$

$$Accuracy = 100 - WER$$

Where I = number of word insertions

D = number of word deletions

S = number of substitutions

N = Total number of actual words spoken

9. Results and Discussion

This section presents the results obtained from implementing a deep learning-based system for speaker diarization and sentiment analysis. The system utilizes a pre-trained Speech-To-Text (STT) model for speech recognition, followed by a deep learning model specifically designed for speaker diarization to identify and separate speakers.

Finally, a lexicon rule-based decision model performs sentiment analysis on the transcribed text for each speaker segment. The system provides a diarization output with around 93.4% accuracy in terms of speech transcription, appropriate assignment of speech segments to the corresponding user, and detection of the sentiment when tested against 10 audio files of non-overlapped speech with multiple speakers.

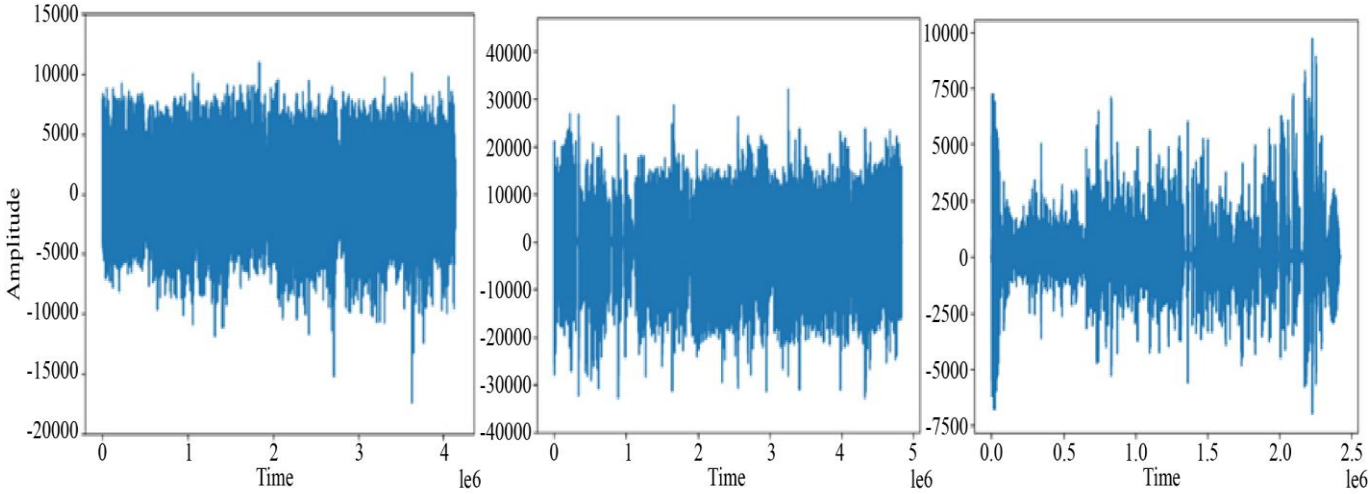

Fig. 3 Audio spectrograms of three audio files with 2 speakers, 4 speakers and 5 speakers respectively, each spectrogram displays the frequency content of the audio signals over time

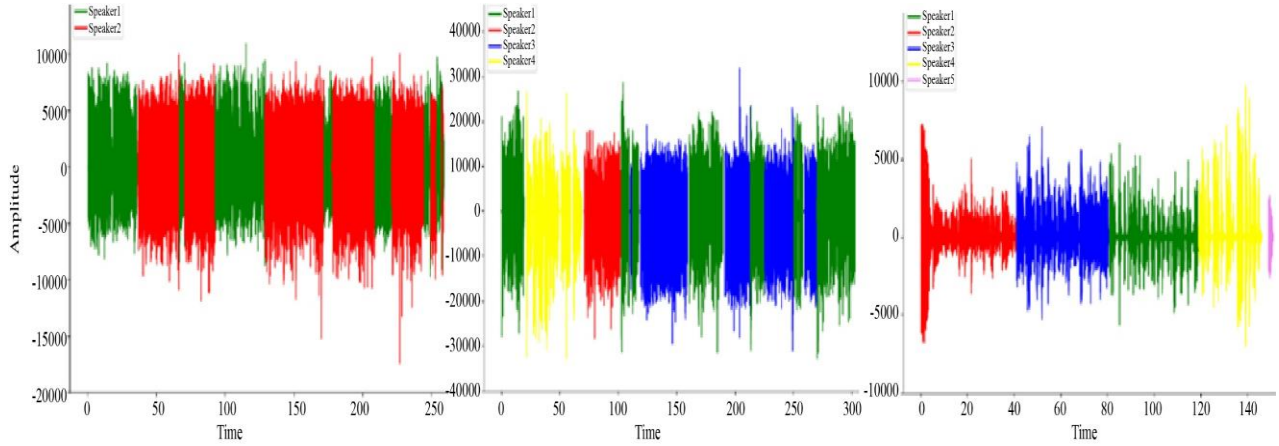

Fig. 4 Audio spectrograms of 3 audio files with 2 speakers, 4 speakers and 5 speakers respectively along with speaker segments. Each colour represents a unique speaker, A coloured portion in the spectrogram indicates that it is spoken and delivered by the speaker represented by its colour

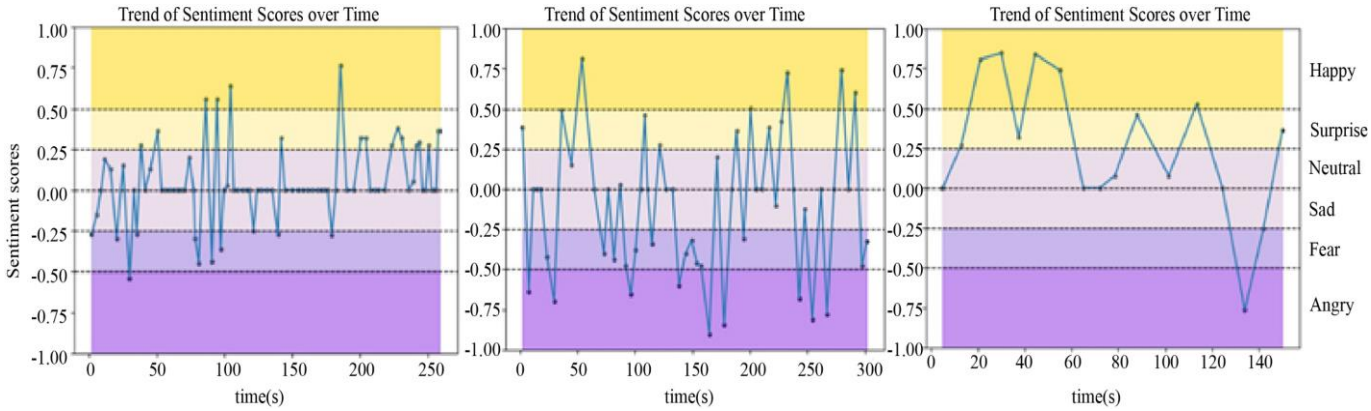

Fig. 5 Trend of sentiment scores over time for output files of audio with 2 speaker, 4 speakers and 5 speakers respectively, the colored portions in the graphs between the grid lines indicate various sentiment labels: happy, surprise, neutral, sad, and angry.

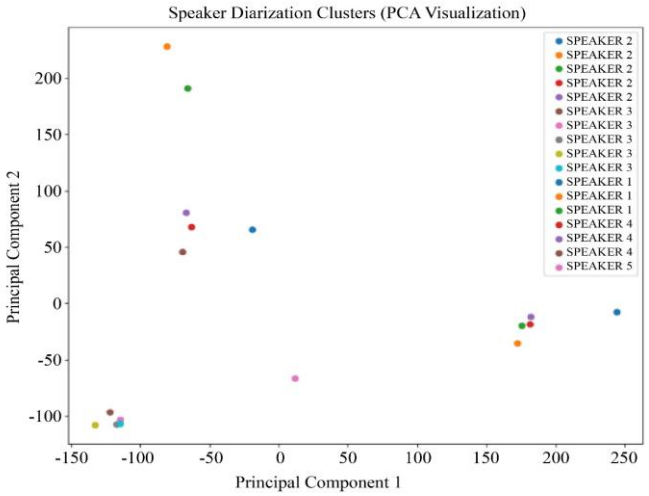

Fig. 6 PCA Visualization of Speech Segments assignment to individual speakers in a multi-speaker recording of 5 speakers

Fiure 6 shows a plot shows a scatterplot of speaker embeddings that are dimensionally reduced from 192 dimensions to 2 dimensions. We can see the correlation between the points that are close to each other. There are several clusters. Points in a particular cluster belong to a particular speaker. There are also fewer outliers, which indicates the robustness of speaker embeddings that distinguish each other. Our method achieves strong performance across large and diverse audio sources by utilizing cutting-edge deep learning, machine learning, signal processing, and speech processing approaches.

10. Conclusion

This work presented a novel deep learning-based system for speaker diarization, integrating sentiment analysis to capture emotional states within audio recordings. This approach addressed several key objectives: Accurate Speaker Segmentation and Noise Suppression – By leveraging pre-trained models like Whisper, the system effectively separates speaker turns within an audio recording and exhibits noise robustness, enhancing speaker diarization reliability in real-world environments. Unsupervised Learning – The system operates unsupervised, eliminating the need for large labelled datasets for training and making it more adaptable to diverse audio content. Sentiment Analysis Integration – Sentiment analysis, which is based on speaker transcripts created by the diarization module, offers insightful information about the conversation's emotional overtones. The suggested technique shows encouraging improvements in speaker diarization. Integrating sentiment analysis benefits from applications in various fields, such as product development, market research, healthcare, and customer experience management. To better comprehend emotional states and facilitate communication, the system can, for example, evaluate patient-doctor interactions in the healthcare industry. Similarly, customer experience management may determine the mood of customers during chatbot or contact centre interactions to improve the Caliber of service. Subsequent research endeavours will concentrate on enhancing the system's precision, specifically in managing concurrent speech and varied acoustic surroundings. Furthermore, investigating different speech-specific sentiment analysis methods may be able to extract even more detailed emotional information from audio recordings.

Acknowledgements

The authors thank the management of Vasavi College of Engineering for providing the necessary facilities and financial support.

References

[1] Alec Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," *Proceedings of the 40th International Conference on Machine, Learning*, Honolulu, Hawaii, USA, vol. 202, 2023. [Google Scholar] [Publisher Link]

[2] Ouassila Kenai et al., "A New Architecture based Vad for Speaker Diarization/Detection Systems," *Journal of Signal Processing Systems*, vol. 91, no. 11, pp. 827-840, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Federico Landini et al., "Bayesian Hmm Clustering of X-Vector Sequences (VBX) In Speaker Diarization: Theory, Implementation and Analysis on Standard Tasks," *Computer Speech and Language*, vol. 71, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Liang He et al., "Latent Class Model with Application to Speaker Diarization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Daniel Garcia-Romero et al., "Speaker Diarization Using Deep Neural Network Embeddings," *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[6] Zili Huang et al., "Speaker Diarization with Region Proposal Network," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yusuke Fujita et al., "End-To-End Neural Speaker Diarization with Permutation-Free Objectives," *arXiv*, 2019.[CrossRef] [Google Scholar] [Publisher Link]

[8] Prachi Singh, and Sriram Ganapathy, "Deep Self-Supervised Hierarchical Clustering for Speaker Diarization," *arXiv*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Amitrajit Sarkar et al., "Says Who? Deep Learning Models for Joint Speech Recognition, Segmentation and Diarization," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, pp. 5229-5233, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[10] Laurent El Shafey, Hagen Soltau, and Izhak Shafran, "Joint Speech Recognition and Speaker Diarization Via Sequence Transduction," *arXiv*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ismail Shahin et al., "Novel Hybrid Dnn Approaches for Speaker Verification in Emotional and Stressful Talking Environments," *Neural Computing and Applications*, vol. 33, no. 23, pp.16033-16055, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Farshad Teimoori, and Farbod Razzazi, "Unsupervised Help-Trained LS-SVR-Based Segmentation in Speaker Diarization System," *Multimedia Tools and Applications*, vol. 78, no. 9, pp. 11743-11777, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] Marie Kunesova et al., "Detection of Overlapping Speech for the Purposes of Speaker Diarization," *Speech and Computer 21st International Conference*, Istanbul, Turkey, pp. 247-257, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Marek Hrúz, and Zbyněk Zajíc, "Convolutional Neural Network for Speaker Change Detection in Telephone Speaker Diarization System," *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, pp. 4945-4949, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[15] Federico Landin et al., "But System Description for Dihard Speech Diarization Challenge 2019," *arXiv*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[16] Lei Sun et al., "Speaker Diarization with Enhancing Speech for the First Dihard Challeng," *Proceedings of Interspeech*, Hyderabad, India, pp. 2793-2797, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[17] David Snyder et al., "X-Vectors: Robust Dnn Embeddings for Speaker Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, pp. 5329-5333, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[18] India Massana et al., "LSTM Neural Network-Based Speaker Segmentation using Acoustic and Language Modelling," *Annual Conference of the International Speech Communication Association*, Stockholm, pp. 2834-2838, 2017. [Google Scholar] [Publisher Link]

[19] Neeraj Sajjan et al., "Leveraging LSTM Models for Overlap Detection in Multi-Party Meetings," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, pp. 5249-5253, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[20] Mourad Jbene et al., "User Sentiment Analysis in Conversational Systems Based on Augmentation and Attention-Based BiLSTM," *Procedia Computer Science*, vol. 207, pp. 4106-4112, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] María Teresa García-Ordás et al., "Sentiment Analysis in Non-Fixed Length Audios Using a Fully Convolutional Neural Network," *Biomedical Signal Processing and Control*, vol. 69, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Manas Jain et al., "Speech Emotion Recognition Using a Support Vector Machine," *arXiv*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," *Proceedings of Interspeech*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] C. Hutto, and Eric Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no.1, pp. 216-225, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[25] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A Large-Scale Speaker Identification Dataset," *arXiv*, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[26] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," *arXiv*, 2018. [CrossRef] [Google Scholar] [Publisher Link]