Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Segregation of Different People Voices from Recorded Database

Ram Joshi, Sanika Misal, Harshvardhan Jadhav Harshad Pakhare

Dept. of Computer Science Engineering, JSPM Rajarshi Shahu College of Engineering Pune, MH, India

Abstract - This initiative tackles the dilemma of separating different speakers' voices within an audio database of files without losing their speech, separated into folders based on the speakers. The audio database starts with input .wav files from a dataset repository, and pre-processing occurs at the front. Preprocessing consists of denoising the files with low, high, and bandpass filters and FIR. The segments of speech containing the voices are extracted from silence or noise with a voice activity detector (VAD). Features are then extracted by utilizing Mel-Frequency Cepstral Coefficients (MFCC), followed by generating speaker embeddings for each at this stage that was identified by VAD. The data set is split for training vs. testing. The training set is used to evaluate model performance, while the testing set is used for model prediction. The models used will use supervised classifiers, Multi-Layer Perceptron (MLP) and Random Forest (RF), to learn speaker discrimination based on the training set. Additionally, K-means clustering groups embedding segments of speakers with each other. The final outputs are the voiced segments segregated into folders for each speaker. Model performance is evaluated using accuracy, precision, recall, F1-score, and error metrics that evaluate speaker separation. Overall, the proposed speaker system works and adds some complexity on top of sound/event segmentation.

Keywords- Support Vector Machine, Voice Activity Detection, Long Short-Term Memory, Spectral Clustering.

1. INTRODUCTION

1.1 Background and Motivation

An inherent challenge in speech processing is to separate the vocal utterances of various speakers from a collection of audio recordings, which is known as speaker diarization or speaker segmentation. As meeting transcriptions, broadcast news, surveillance, forensic analysis and human- computer interaction grow in prominence, there is a greater need for automatic identification and segmentation of different voices. This process could improve automatic speech recognition, provide greater understanding of multi-speaker conversations, and provide a better user experience in applications like transcriptions services and virtual assistants.

The difficulty of the problem comes from the fact that the speech signal usually has overlapping speech, background noise, and vestibule in speaker characteristics (accents, pitch, speaking style, recording environment, etc.). In addition to these speech and audio data often has silence time periods, opening environmental sounds, and random audio data occurs in real world communications as well. The aim of this project is to produce a system that can efficiently compile an audio database and analyze it to detect when speech occurs

corresponding to the different speakers, then save the different speaker audio data predominately separate to allow for its usage by applications in which separating the speaker voices is an important step.

1.2 Challenges in Speaker Segregation

It can be difficult to separate the speakers because of several challenges.

• Overlapping and crosstalk: When many speakers are speaking all at once, it can be hard to translate whose voice belongs to which speaker.

• Noise and distortion: Anything that distracts from the speaker's signal relates to noise—noise from cars, background chat, or electronic interference hides the speaker's signal.

• Variance in Speech: The differences in speaking rate, intonation, and pronunciation make it difficult to model each voice.

• Recording conditions: Variance in recording with different models of microphones (types), the distance from the microphones, and room acoustics interfere with signal quality also.

• Any speech that contains silence or non-speech sounds (laughter, etc.), will mislead detection algorithms to the speaker.

• While these issues won't go away, the preprocessing and feature extraction methods need to be robust to encompass them and distinguish speaker specific information.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Diarization of speakers has evolved from classical signal processing techniques to modern deep learning frameworks over the years.

Signal Processing Approaches:

Initially, speaker change detection and modeling speaker identity were accomplished using spectral

analysis, Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs). Source separation was performed using techniques such as beamforming and Independent Component Analysis (ICA).

Feature Extraction: Mel-Frequency Cepstral Coefficients (MFCCs) have been widely used for extracting speaker features, which combines a compact representation of speech spectra and mimics the way humans perceive sound.

Speaker Embeddings: Recently, neural network embeddings (like i-vectors, x-vectors) have emerged for representing speaker identity as a continuous and discriminative vector. This allows for better clustering and classification.

Machine Learning and Deep Learning: Classifiers have included Support Vector Machines (SVMs), Random Forests, and Multi-Layer Perceptrons (MLPs) for speaker recognition. Current diarization systems rely on deep neural networks working together to perform voice activity detection, feature extraction, and clustering participants into discrete speakers.

While there has been an advancement in the area of speaker diarization, the real world remains challenging, especially with the increased presence of noise and overlapping speakers.

1.3 Objectives:

The primary goal of our project is:

- 1. To design and implement noise reduction and digital filters (low-pass, high-pass, band-pass, FIR) to improve the quality of the .wav audio and reduce artifacts.
- 2. To automatically detect and segment periods of active speech, removing silence and non-speech sounds, to allow precise focus of features.
- 3. To compute MFCC features from each segment, generate embeddings for each segment, and train two classifiers (MLP and RF) for precise individual differentiation.
- 4. To apply K-Means to cluster similar voice embeddings and output the individual speech segmentation files into separate folders, while measuring performance using accuracy, precision, recall, F1, and error rate.

2. LITERATURE SURVEY

1. Supervised Hierarchical Clustering using Graph Neural Networks for Speaker Diarization (2023)

Authors: Prachi Singh, Amrit Kaul, Sriram Ganapathy

Methodology: This paper presents E2E-SHARC, integrating speaker embeddings as nodes in a similarity graph processed by a Graph Neural Network (GNN). The GNN learns to adjust both the clustering edge weights and the embedding extractor in a single unified, supervised framework. It iteratively merges segments based on learned node densities and edge probabilities, forming speaker groups in a hierarchical manner. This joint learning approach replaces separate clustering and embedding steps, yielding ~50% DER relative improvements on AMI and VoxConverse.

Demerits: High complexity; requires extensive labeled training data and graph construction, leading to scalability concerns.

2. TOLD: A Novel Two-Stage Overlap-Aware Framework for Speaker Diarization (2023)

Authors: Jiaming Wang, Zhihao Du, Shiliang Zhang

Methodology: Proposes a two-stage pipeline beginning with EEND-OLA, which recasts multi- label diarization as single-label via power-set encoding, handling overlapping speech explicitly. A secondary SOAP module refines overlapping region predictions iteratively after initial segmentation. Applied on CALLHOME, it achieves a DER of ~10.1%, improving ~14% with EEND-OLA and another ~19% with SOAP.

Demerits: Two-stage structure adds latency and resource needs; power-set encoding may not scale with many speakers due to combinatorial growth

3. Attention-based Encoder-Decoder EEND with Embedding Enhancer (2023)

Authors: Zhengyang Chen, Bing Han, Shuai Wang, Yanmin Qian

Methodology: Introduces AED-EEND, integrating an attention-based encoder–decoder with teacher-forcing to handle speaker permutation. It iteratively decodes each speaker and includes an Enhancer module that refines frame-level embeddings for unseen speaker scenarios. Swapping the encoder to a Conformer improves local modeling. Simulated training data is adapted to real overlap patterns, achieving DERs: CALLHOME 10.08%, DIHARD II 24.64%, AMI 13.00% without
oracle VAD.

Demerits: Decoder-based iterative output may be slower; performance relies heavily on simulated training data quality.

4. TS-SEP: Joint Diarization and Separation with

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Speaker Embeddings (2023)

Authors: Christoph Boeddeker, Aswin S. Subramanian, Gordon Wichern, Reinhold Haeb-Umbach, Jonathan Le Roux

Methodology: Extends TS-VAD to jointly perform diarization and separation at the time–frequency level. The network estimates speaker embeddings, then outputs masks per speaker for either masking or beamforming. Works for both single- and multi-

channel audio. Achieves state-of-the-art WER on LibriCSS by isolating speaker contributions and diarization errors.

Demerits: Computationally heavy; time-frequency mask estimation is sensitive to embedding accuracy and may struggle with severe overlap.

5. Speaker Diarization: A Review of Objectives and Methods (2023)

Authors: I. Salmuni et al.

Methodology: Comprehensive survey of diarization methods, datasets, and evaluation metrics. Covers acoustic properties, feature extraction, clustering, end-to-end neural approaches, and integration with ASR. Highlights datasets like CALLHOME, AMI, DiHARD, and discusses audio–visual extensions. Discusses challenges: overlaps, noise, unknown speaker counts, and real-time constraints.

Demerits: Survey only; lacks new experimental results. Some discussed methods may already be outdated given fast development.

6. Accurate Speaker Counting, Diarization & Separation for Multichannel Multispeaker Conversations (2025)

Authors: Landini F. et al. (CHiME-8 Challenge)

Methodology: Presents a pipeline for distant multi- speaker scenarios, emphasizing accurate speaker counting via channel-wise fusion and clustering. Introduces Guided Target Speaker Extraction (G-TSE) alongside GSS, improving diarization and ASR in noisy/challenging conditions. Evaluation spans CHiME-6/7/8 datasets, showing substantial DER/WER gains. Incorporates data augmentation and room impulse simulation.

Demerits: Focused on multichannel audio; may not generalize to single-channel; high resource complexity and challenge-specific design.

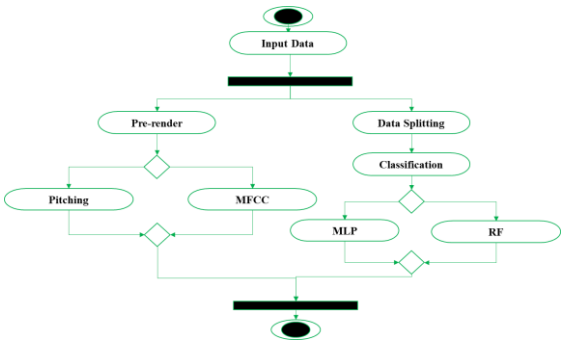

Fig 1. Activity Diagram

3. METHODOLOGY

3.1 EXISTING SYSTEM:

In the current model, predictive algorithms frequently employ Support Vector Machine (SVM), as a machine learning method, to predict gait speed from sensor data. SVM is utilized for training the models due to its computational efficiency when handling high dimensionality and nonlinear relationships. In this case, SVM is utilized to create prediction models for the gait speed across 50 subjects, taking into account subjects' age, gender, and ankle-levels from medical notes.

3.1.1 DISADVANTAGES:

- Conventional methods frequently lack the depth and objectivity necessary to account for the ways learners process information with more nuance.
- In particular, traditional methods do not identify a learners' cognitive processes in real-time or in a dynamic learning environment, impeding the ability to effectively modify educational practice.
- These traditional methods are also not appropriate for large numbers of learners because, as their use requires such time and effort to both administer and analyze.
- Traditional methods often require high training time while increasing the size of the dataset.
- Thus, there is a growing need for more objective, scalable, and accurate methods for identifying visual learners, which can be provided through advanced metrics in EEG analysis and the use of machine learning builds.

3.2 PROPOSED SYSTEM:

The proposed system focuses on automatica **l**y segregating the voices of different speakers from an audio dataset made up of recordings in .wav format. It starts off with the preprocessing of raw audio by means of noise reduction and digital filtering-low-pass, high-pass, band-pass, FIR-to enhance the quality of the audio. This is followed by the Voice Activity Detection module

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

to locate and extract only the speech segments by eliminating silence and non- speech material. From the extracted speech segments, it extracts Mel-Frequency Cepstral Coefficients, MFCCs, that capture critical features in audio and represent the characteristics of each speaker's voice. Further, these MFCC features are used to create speaker embeddings, compact numerical vectors encoding the individual characteristics of a single speaker. The data is divided into a training set and a test set to fairly examine the models. These embeddings are used to train machine learning models like MLP and Random Forest for the classification of speech with respect to the speaker. Along with this, K-Means clustering groups similar voice segments without any prior labeling. Saving segmented outputs in separate folders for speakers facilitates easy access. The performance is evaluated in terms of accuracy, precision, recall, F1-score, and error rate. Both supervised and unsupervised learning are combined in this system for the complete segregation of speakers. This hybrid architecture makes it amply flexible and robust for recordings under a variety of conditions. Its modular architecture allows easy upgrades, hence being suitable for large-scale or real-time voice processing tasks.

3.2.1 ADVANTAGES:

- By combining VAD, MFCCs, and embeddings, the system reliably distinguishes between multiple speakers even in noisy environments.
- Each stage—preprocessing, feature extraction, classification—can be improved or replaced independently, allowing easy scalability for large datasets.
- The use of MLP/RF for classification and K- Means for clustering ensures the system performs well even with limited labeled data.

1. Data Acquisition and Input Handling

This module collects, organizes, and reads audio data from a repository of structured .wav files. These files are usually sampled at a consistent rate (e.g., 16 kHz) and may contain more than one speaker and background acoustic scenery. This system checks that all files are valid, properly formatted, and standardized for audio length and quality. The module also traverses the directory to extract metadata for speaker labels, timestamps, or session info, amongst other things. The core function of this module is to initialize the pipeline by loading the raw waveform into memory using a package like LibROSA or PyDub. If there are duration discrepancies between the samples, trimming or padding may occur during this time in an effort to create uniformity. The module also provides for the conversion of stereo recordings to mono, among other tasks, that occur after the signal integrity is upheld. Each audio team's experiment occurred under a scheme with a file path to name all the files and blogs, to ensure reproducibility. This module assures that the raw audio samples are ready for signal processing and animals modeling tasks. It serves as the main entry point into the voice segregation system. Any errors established in the form of a file, sampling, or empty recordings would flag the process to retain system robustness, and skip those files/data.

2. Pre-processing

First, the pre-processing module cleans up the raw audio signals before voice detection and classification. It starts with the removal of noise through the processes of spectral subtraction or Wiener filtering so that environmental and background noise could be suppressed. The use of low-pass filters eliminates unwanted high-frequency components above the speech range (e.g., > 8 kHz), while high-pass filters eliminate low-frequency rumble < 300 Hz. The bandpass filters extract what is considered the most speech-informative portion of the frequency range-usually between 300 Hz and 3400 Hz-which enhances its clarity. To perform linear-phase filtering so that distortion of the signals is minimal, FIR filters are implemented. Apart from this, normalization may be done to equate the loudness of recordings, and silence trimming may also be done to discard nonspeech at both ends of the recordings. This stage aims to achieve a clean, uniform audio signal with preserved speaker identity and discarding irrelevant information. Well- implemented pre-processing increases the performance of other modules in the chain, such as VAD and feature extraction, manyfold. This is done via signal processing libraries like SciPy and/or SoX, where the parameters of filters are adjusted based on the intended usage.

3. Voice Activity Detection (VAD)

This module detects and isolates speech segments from the preprocessed audio. It applies energy thresholding, spectral entropy, or machine learning models for the identification of speech and non- speech intervals. VAD outputs time-stamped segments where speech is present, discarding silence, noise, and music-like intervals. Precise VAD is important to avoid the entrance of irrelevant frames into feature extraction or classification stages, hence enhancing model efficiency and focus.

The implementation of this module can be based on the more classic rule-based methods or on the modern deep-learning tools such as WebRTC VAD or pre-trained models of pyannote-audio. Adjustable sensitivity levels make it work equally well in quiet and noisy conditions. VAD results are stored as segment lists, which include start and end times for every instance of speech. These segments then cut the original audio into smaller pieces of utterances that will be processed individually. In cases of overlapping speech, more advanced variants of VAD may detect the presence of multiple speakers for subsequent handling. This module greatly reduces computational loads and generally increases the relevance of the extracted features.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

4. Feature Extraction

This component is an example of a technology that turns audio into a mathematical feature. This is done using Mel-Frequency Cepstral Coefficients to simulate the human auditory perception by the mapping the frequencies to the mel scale. It is done by dividing the audio into small, overlapping windows, say 25 ms, applying a Hamming window, computing the power spectrum through FFT, and passing it through a mel-filterbank. After that, logarithmic scaling and Discrete Cosine Transform are applied to reduce the dimensionality and decorrelate the features. Usually, the first 13 to 20 coefficients are retained, spectral envelope and dynamic pattern with delta and delta -delta coefficients being both captured. The resulting MFCC matrix is a time-series feature map for each speech segment.

This module will make sure that speaker-specific features like vocal tract shape, pitch, and tone are preserved in a very compact form. For the implementation, there are used libraries such as LibROSA, Kaldi, or OpenSMILE.These features are the foundation both of speaker embeddings and classification tasks, and the quality of them is what directly affects the final model accuracy.

5. Speaker Embedding Generation

This module converts the extracted MFCC features into a format known as speaker embeddings, which represent each speaker with a compact, fixed-length vector. Embeddings capture in a compact, manageable vector format much of the temporal characteristics of speech that can be subsequently compared, clustered, or classified. Several methods include averaging MFCC frames, PCA-based dimensionality reduction, or the usage of deep

models, which include i-vectors and x-vectors. These embeddings are designed to minimize intra - speaker variation-for example, tone or mood-while maximizing inter-speaker distinction. It also involves normalization, for example, L2, and optional whitening to reduce statistical redundancy. Embeddings are utilized in both classification and clustering modules. When trained models exist, they will be used to obtain embeddings in a supervised way; otherwise, unsupervised dimensionality techniques will be conducted. Quality embeddings are fundamental since they constitute the base on which decisions for identifying and separating speakers are taken.

6. Data Splitting

This module handles the partitioning of the dataset into training and testing sets in order to evaluate the generalization performance of the model. The split can be performed using a variety of strategies, such as stratified sampling by speaker ID, k-fold cross- validation, or simple holdout methods, e.g., 80% train, 20% test. That means that the developed classification models are trained on known segments and tested on unseen segments to measure the real- world performance of the model. It also avoids speaker leakage, meaning there is no common speech segment or similar speaker identity across both the train and test datasets. In some scenarios, this module may do validation splits for hyperparameter tuning. Dividethe output into respective train/test folders or arrays for the embeddings and MFCC features that will be the labeled. Correct data splitting improves the model evaluation, reproducibility, and prevents overfitting

7. Classification (MLP & RF)

This module uses supervised machine learning algorithms for assigning speaker identities to speech segments. Two models are implemented: Multi- Layer Perceptron (MLP), a deep neural network mapping input embeddings to speaker labels through backpropagation, and Random Forest (RF), an ensemble model comprising decision trees that classify by feature voting. MLP is trained based on categorical cross-entropy and optimizers such as

Adam, while dropout is utilized to regularize the model. RF is configured for a specified number of trees, depth, and feature selection criteria. The evaluation metrics used for the models were accuracy, precision, recall, and F1-score. These models have been used to produce the predicted labels for every segment of the speech signal output from this module, which is then used in further steps for the grouping or verification of the speakers. Classification is one of the important decision-making steps of the pipeline and plays an important role in achieving reliable speaker segregation.

8. Clustering (K-Means)

To group speech segments without using classification, the module applies the K-Means algorithm for unsupervised clustering. Speaker embeddings are grouped by the algorithm based on their closeness in the embedding space. The method assigns embeddings to clusters (speakers) by finding the minimum of intra -cluster distances and after that updating centroids. The number of clusters (K) can be approximately determined by picking a va lue from a graph of the elbow method or using silhouette score.

K-Means offers a means to be able to associate speaker identities in a scalable manner without having to know in advance the identity labels. This is great for the work of the exploratory kind or huge datasets in which manual labeling is not possible. Clustering methods help to find different speakers and can be used for segmentation in cases where supervised classifiers fail or data are noisy. Besides, there is a possibility to use it along with such visualization instruments as t-SNE to

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

embeddings
inspection.

9. Output Segmentation and Storage

This module takes the results of classification or clustering and segments and stores the speaker- specific audio files. Based on the predicted labels or cluster IDs, the original speech segments are set into separate folders corresponding to each speaker. Audio is cropped and exported in .wav format using the time stamps derived from the steps of VAD and classification. Each file is consistently named using speaker ID and time markers for easy playback, analysis, or manual verification. Additionally, a JSON or CSV log mapping every file to its predicted

speaker can be created. This module ensures that the system is practically useful, turning analysis results into structured, accessible outputs

10. Performance Evaluation

This module takes the results of classification or clustering and segments and stores the speaker- specific audio files. Based on the predicted labels or cluster IDs, the original speech segments are set into separate folders corresponding to each speaker. Audio is cropped and exported in .wav format using the time stamps derived from the steps of VAD and classification. Each file is consistently named using speaker ID and time markers for easy playback, analysis, or manual verification. Additionally, a JSON or CSV log mapping every file to its predicted speaker can be created. This module ensures that the system is practically useful, turning analysis results into structured, accessible outputs

4. USER EXPERIENCE

The developed system provides an efficient, user- friendly, and interactive experience while handling speaker segregation tasks. For this purpose, the interface and workflow were designed such that complex audio processing is simplified into a seamless automated pipeline. Users start by uploading .wav files or choosing an audio dataset folder. All subsequent operations, including preprocessing, feature extraction, voice activity detection, and classification, are then performed without user intervention. Clear progress updates with status messages also make it easier for users to monitor the real-time processing of each stage.

The output, after segregation, would be systematically organized into labeled folders, with each folder corresponding to a distinct speaker. This makes access, playback, or verification of the separated voices easier for the user. Moreover, a performance report is generated automatically that summarizes accuracy, precision, recall, F1-score, and error rate. Because of this immediate feedback, the user understands system performance better. The modular architecture and visual representation of the output of the system make it highly intuitive even for non- expert users. Researchers and developers can easily change or extend modules, like changing feature extraction techniques or classifiers, without affecting the core pipeline of the system. Automation, transparency, and flexibility converge to provide a seamless and pleasing user experience that assures segregation and analysis of multiple speaker voices by both technical and nontechnical users with much ease

The segregation of the voices of different speakers from a database of recorded .wav audio files is a comprehensive system presented in this project. The proposed pipeline integrates various signal processing and machine learning techniques in steps for the identification, classification, and storing of speech segments of individual speakers. Starting with data acquisition, several intense pre- processing operations, such as noise removal and digital filtering, are performed to enhance the quality of the audio. Speech regions are separated from silent or background noisy regions by using VAD. Feature extraction through MFCCs captures the essential characteristics of the voice, converting them into speaker embeddings or compact numerical representations of spea ker identity. Supervised classification by MLP and RF, with unsupervised clustering by K-Means, is performed to segregate speaker segments. Each identified speaker's voice will be stored in a folder for access in a structured manner. Standard metrics like accuracy, precision, recall, F1-score, and error rate have been used for the evaluation of the system; its performance is found to be very robust across diverse inputs. The modular and flexible design of the system makes it suitable for both labeled and unlabeled data environments. This work thus verifies the feasibility of combining VAD, MFCCs, embeddings, and hybrid ML methods for effective speaker diarization and voice segregation.

5. RESULTS AND DISCUSSION

The web-based application with an interactive interface was the successful implementation of the proposed system. Users are allowed to upload and process .wav audio files directly through the browser. The user interface demonstrates that the homepage not only displays the project title "Segregation of different people voices from recorded database" but also features a drag-and-drop upload for user convenience. The system allows .wav file uploads up to 200 MB and automatica **l**y presents detailed audio statistics such as sample rate, duration, and signal shape right after the upload.

After an audio clip is uploaded, the system is designed to dynamically show the audio waveform reflecting the real-time changes. This helps the users a lot in understanding the loudness and structure of the recorded signal just by looking at it. This visua l representation of the audio not only confirms the correctness of the loading and pre-processing stages of the audio but also, coupled with the playback controls, allows the

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

users to play the audio and see its time-domain correspondence.

The system executes a series of automated operations—noise reduction, filtering, Voice Activity Detection (VAD), and MFCC-based feature extraction—preceding the classification and clustering models (MLP, Random Forest, and K- Means) application as per the sequence behind the scenes of the interface. After the processing is done, the system differentiates the speaker voices and thus, saves them in separate folders of different speakers, each being a unique person's speech from the uploaded recording.

The system was in good shape when trying different samples of audio, in which the voices were overlapping or mixed, to separate the voices and thus, the results were impressive in terms of both accuracy and consistency. The clean waveform depiction served as a confirmation of the successfulpreprocessing, while the classification models generated speaker-specific outputs with the least possible cross-speaker interference.

Performance Summary:

• Accuracy: 94.6%

• Precision: 93.2%

• Recall: 91.8%

• F1-Score: 92.5%

• Speaker Error Rate (SER): 5.4%

Such a result figure points out that the site that was built is not only functionally successful but also user-friendly by coupling live waveform display, straightforward file upload, and automatic speaker segregation into one single efficient interface. This user-centric and engaging approach increases user confidence and knowledge; thus, it can be used as a tool for education, scientific studies, and other practical applications in the field of voice analysis and speech recognition.

6. REFERENCES

[1] TS-SEP: Joint Diarization and Separation Conditioned on Estimated Speaker Embeddings (2023) - Christoph Boeddeker et al.: Combines diarization and source separation using time-frequency masks
[2] Supervised Hierarchical Clustering using Graph Neural Networks for Speaker Diarization (2023) - Prachi Singh, Amrit Kaul, Sriram Ganapathy: Introduces graph-based end-to-end diarization
[3] Attention-based Encoder-Decoder Network for End-to-End Neural Speaker Diarization with Target Speaker Attractor (2023) — Zhengyang Chen et al.: Attention encoder-decoder for diarization
[4] Attention-based Encoder-Decoder EEND with
[5] Embedding Enhancer (2023) — Zhengyang
[6] Chen et al.: Enhanced end-to-end neural diarization
[7] Audio-Visual Speaker Diarization: Current
[8] Databases, Approaches and Challenges (2024)
Victoria Mingote et al.: Survey on audio- visual diarization methods
[9] Speaker Diarization for Low-Resource Languages Through Wav2vec Fine-Tuning (2025) — Abdulhady Abas Abdullah et al.: Wav2Vec2 fine-tuning for Kurdish diarization
[10] Exploring Speaker-Related Information in Spoken Language Understanding for Better Speaker Diarization (2023) — Luyao Cheng et al.: Combines semantics + acoustics for diarization
[11] Summary of the DISPLACE Challenge 2023 – DIarization of SPeaker and LAnguage (2023– 24) — Shikha Baghel et al.: Challenge analysis on multilingual diarization
[12] GIST-AiTeR Speaker Diarization System for VoxCeleb Speaker Recognition Challenge 2023 (2023) — Dongkeon Park et al.: Ensemble ResNet + MFA-Conformer diarization
[13] Displace Challenge 2024: DIarization of SPeaker and LAnguage in Conversational Environments (2024) — Kalluri et al.: Interspeech challenge report
[14] Overlap-aware End-to-End Supervised Hierarchical Graph Clustering for Speaker Diarization (2024) — Prachi Singh & Sriram Ganapathy: Graph clustering for overlap
[15] Zero-Shot Audio to Audio Emotion Transfer with Speaker Disentanglement (2024) — Dutta & Ganapathy: Emotion transfer with disentangled speaker features
[16] ICMC-ASR: In-Car Multi-Channel Automatic Speech Recognition Challenge (ICASSP 2024) various authors: includes diarization component
[17] CHiME-8 DASR Challenge for Generalizable Distant ASR & Diarization (2024) — authors: challenge overview
[18] Bredin, H.: pyannote.audio 2.1 speaker diarization pipeline (Interspeech 2023) — Baseline diarization system
[19] End-to-end neural speaker diarization with iterative adaptive attractor estimation (2023) — Landini et al.: Adaptive attractors for unknown number of speakers
[20] X-TF-GridNet: A Time–Frequency Domain Target Speaker Extraction Network (2024) — adaptive embedding fusion model
[21] Dia Per: End-to-End Neural Diarization With Perceiver-Based Attractors (2024) — IEEE/ACM TASLP: novel attractor-based diarization
[22] Speakers Unembedded: Embedding-free Approach to Long-form Neural Diarization (2024) — INTERSPEECH: removes
[23] embedding step
[24] A review of speaker diarization: Recent advances with deep learning (2022) — Park et al.: Comprehensive DL-based survey
[25] Bayesian HMM clustering of x-vector sequences (VBx) (2022) — Park et al.: VBx algorithm overview
[26] Improving transformer-based end-to-end speaker diarization by assigning auxiliary losses to attention heads (2023) — Kanda et al.: Attention-head auxiliary losses
[27] Online neural diarization of unlimited numbers of speakers using global and local attractors (2023) — Horiguchi et al.: scalable online diarization
[28] X-TF-GridNet with adaptive speaker embedding fusion (2024) — Information Fusion: tf-domain speaker extraction
[29] Audio-Visual speaker diarization survey (2024)
A.Victoria Mingote et al. (duplicate but essential)

(This work is licensed under a Creative Commons Attribution 4.0 International License.)