Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 10, October 2024

A Novel Approach to Speech Diarization using Embedded Clustering and Voice Activity Detection

Ravisham Santha
VLSI Design and Embedded Systems
BMS College of enginnering
Bengaluru, Karnataka.

Dr . K P Lakshmi
VLSI Design and Embedded Systems
BMS College of enginnering
Bengaluru, Karnataka.

***Abstract*—In various applications such as voice recognition, speech-to-text, and other sound-related functionalities, isolating the required voice is crucial. Many available datasets for these applications consist of singular voices with included noise, suitable for cleaning. However, real-world scenarios often involve multiple voices, making the cleaning process challenging. In such cases, voice diarization becomes essential to differentiate and split different voice notes within the audio. Speaker diarization, the process of partitioning an audio stream into homogeneous segments, addresses this need.**

This paper presents a two-step process for speaker diarization. Firstly, the audio is segmented into different voices using Gaussian Diversion, Bayesian Information Criterion, Hierarchical Agglomerative Clustering, and resegmentation. Secondly, voice activity detection is employed to distinguish between individual speakers and background noise. The achieved accuracy is 82 percent, acknowledging a decrease caused by the breakdown of voice notes during Hierarchical Agglomerative Clustering. Future improvements could focus on refining the separation of combined speakers, potentially leveraging a voice recognition module on already separated voice notes for enhanced precision.

***Keywords*— Hierarchical Agglomerative Clustering , Voice activity Detection (VAD) , Bayesian Information Criterion, MFCC.**

I. INTRODUCTION

An audio signal encapsulates all the essential information required for sound reproduction, serving as a comprehensive representation of sound. In practical applications, the transformation between digital and analog waveforms is frequent and indispensable. Speaker recognition is the process of identifying an individual based on voice characteristics, specifically answering the question, "Who is speaking?" The broader term "voice recognition" encompasses both speaker recognition and speech recognition.

Speaker verification, distinct from speaker diarization and identification, verifies a claimed identity. Speaker diarization, also known as diarization, is crucial for segmenting an input audio stream into homogeneous sections based on speaker identity. This segmentation significantly enhances the comprehensibility of automatic speech transcriptions by organizing the audio stream into speaker turns. When integrated with speaker recognition systems, it aids in accurately identifying the speaker. Addressing the question, "Who spoke when?" speaker diarization combines speaker segmentation and clustering. Speaker segmentation identifies transition points between speakers, while clustering groups speech segments based on speaker characteristics.

These evaluations involve distinguishing between speech and nonspeech segments and marking speaker changes within detected speech. Nonspeech segments include music, silence, noise, and other similar elements.

Speaker diarization offers several advantages: it allows for speaker-specific audio searches, enhances transcript readability, and aids in speaker adaptation for speech recognition systems.

Focus on advancing speaker classification and speaker diarization techniques. The primary objective is to proficiently identify and classify speech segments, determining the appropriate number of speakers present in the audio data. The development of the speaker diarization system encompasses various stages, including preprocessing the data to address domain-specific considerations like noise reduction. Additionally, Voice Activity Detection (VAD) tools are applied to delineate speech segments, and essential features are extracted to facilitate accurate classification. The project adopts a systematic approach, initializing cluster numbers based on the chosen system type, and executing speaker segmentation and clustering tasks sequentially or in parallel.

Unsupervised and supervised diarization systems are increasingly pivotal in enhancing the accuracy and usability of automatic speech recognition (ASR) and related technologies. Unsupervised diarization, in particular, significantly improves the clarity and organization of speech-to-text transcriptions by segmenting audio recordings based on speaker turns[2]. This approach is particularly beneficial in the processing of extended audio recordings involving multiple speakers, as it enables the division of lengthy audio streams into manageable segments prior to speech recognition [3].Additionally, unsupervised diarization finds utility in automatic translation systems, where it partitions audio data into distinct speech segments, facilitating more efficient and accurate translations by enabling segment-wise processing[4][5].

Supervised diarization, on the other hand, is becoming increasingly critical for applications requiring not only transcription but also speaker identification. This is especially relevant in domains such as online meetings and video conferencing, where there is a growing demand for systems that can automatically transcribe conversations and identify individual speakers within the dialogue. Such functionality allows for advanced features such as voice search and indexing of audio content, making recorded conversations more searchable and structured [6].

A notable use case of supervised diarization is in the annotation of telephone conversations, where the system segments the dialogue between a caller and an operator, sending these segments to speech recognition systems for transcription. Another prominent application lies within healthcare, where supervised diarization can isolate and focus on a doctor's speech. This capability enhances the accuracy and reliability of transcriptions related to diagnoses, prescriptions, and medical procedures[7], thereby supporting more specialized use cases in medical and clinical environments.

To contribute to the existing body of knowledge in speaker diarization, my work involves experimenting with different algorithms, optimizing parameter settings, and exploring innovative approaches to enhance the accuracy and efficiency of speaker classification. The aim is to provide valuable insights into improving speaker diarization systems, offering applications in various domains such as transcription services, voice-controlled interfaces, and audio indexing. The outcomes of this work are expected to contribute to the broader field of speech processing and automated audio analysis.

A. Spectrogram and MFCC

Spectrogram visually represents the spectrum of frequencies within an audio signal, capturing its temporal and frequency characteristics. It is created by applying the Short-Time Fourier Transform (STFT) to the signal, which involves performing the Fast Fourier Transform (FFT) on small time segments. Since human perception of sound is logarithmic, with greater sensitivity to lower frequencies, the mel scale—a logarithmic scale where equal distances represent perceptually equivalent differences—was introduced. A mel-spectrogram is a spectrogram with frequencies transformed into the mel scale.

Extracting Mel Frequency Cepstral Coefficients (MFCC) features is a common practice in text-independent speaker recognition. The process involves several steps:

1. Framing and Windowing: The signal is divided into 20ms windows, as speaker signals are stationary within these short durations.
2. Computation of the DFT: The Discrete Fourier Transform (DFT) is calculated to extract spectral information from each window, determining the energy available in each frequency band.
3. Computation of the Mel Filter Banks: Mel Filter Banks, triangular filters non-linearly placed across the bandwidth and Mel scale, estimate energy in different frequency regions, converting the signal from the frequency domain to the Mel scale.
4. Computation of the Logarithm: After computing the Mel filter banks, the logarithm of each filter bank is calculated, enabling the use of cepstral mean subtraction for channel normalization.
5. Computation of the DCT: The Discrete Cosine Transform (DCT) of the filtered signal is computed, resulting in Mel Frequency Cepstral Coefficients. Typically, only the first 12 coefficients are utilized, as these represent information about the vocal tract filter, distinct from other information.

The cepstrum, representing the rate of change in spectral bands, is a crucial aspect of MFCC. It is essentially a spectrum of the log of the spectrum of the time signal, residing neither in the frequency nor time domain. These coefficients, known as Mel-Frequency Cepstral Coefficients (MFCCs), are critical in analyzing and characterizing audio signals for various applications, including speaker recognition [9][10].

B. Design of Neural networks for Speaker Recognition

To address the binary classification problem a Backpropagation (BP) neural network model with a consistent structure is used. These BP network models consist of three layers: an input layer with 143 nodes, a hidden layer with 120 nodes, and an output layer with two nodes to provide the expected probability for the binary problem. Cross-entropy is used as the loss function, with a learning rate of 0.01. Training stops upon reaching the maximum number of iterations, and the final model is selected based on the highest validation accuracy. When adding new speakers, only the binary classifiers for the new speakers are retrained, leaving other classifiers unchanged.

Both the Voice Activity Detection (VAD) and Coherent Point Drift (CPD) models function as frame-level binary classifiers based on neural networks. The VAD model employs a Deep Neural Network (DNN) structure with seven fully connected layers and Rectified Linear Unit (ReLU) activation functions. This DNN uses a large input window covering 55 consecutive frames (27 on each side) to ensure sufficient information for accurate speech and non-speech classification.

Conversely, the CPD model uses a ReLU Recurrent Neural Network (RNN) to encode past and future inputs (spanning 50 frames on each side) into two vectors. These vectors are fused using the Hadamard product, followed by a SoftMax fully connected layer to classify frames as a speaker change point or not. The RNN's output vectors, representing speaker characteristics for past and future audio segments, enable decisions on speaker identity change by comparing representations before and after the current time [1].

To enhance the RNN's ability to encode effective speaker representations, frame-level d-vectors from a Time-Delay Neural Network (TDNN) model, trained by classifying speakers in the training set for each frame, are used as input. The CPD model, encompassing the TDNN, RNN, and output layer, undergoes joint training to effectively execute speaker change/non-change classification.

C. Current advancements in diarization

Recent advancements in single-speaker identification and diarization rely heavily on embedding models, with two prominent types being probabilistic subspace models like i-vector and deep neural network-based models such as x-vector and d-vector. These models are trained to optimize either a cross-entropy loss over a predefined set of training classes or an embedding loss (e.g., triplet, generalized end-to-end) in a one-shot learning scenario [8].

For speaker identification and diarization in single-speaker contexts, embedding models are the predominant methodology. During training, the speakers at test time are generally unknown. In speaker identification, the model receives one enrollment example per speaker in isolation, aiming to identify the individual speaking in any given input audio. In speaker diarization, without explicit enrollment examples, the process involves computing distances for clustering, similar to comparing test examples with enrollment examples in identification.

Compositional embeddings, recently explored in computer vision for object recognition, involve models with distinct embedding and set union functions. These embeddings translate high-dimensional vectors into a lower-dimensional space, facilitating machine learning on large inputs like sparse word vectors. The goal is to capture semantic similarities by positioning semantically similar inputs closely in the embedding space, promoting efficiency and consistency across multiple models [13].

In multi-person speaker diarization, the traditional approach involves using an overlapping speech detector to estimate the set of speakers during segments of overlap, often selecting the top k closest speakers in the embedding space. However, compositional embedding models offer a novel approach, allowing for the joint prediction of both the number of speakers and their identities. This innovative methodology represents a shift from traditional strategies, providing a more comprehensive and unified solution for multi-person speaker diarization.

II. PROPOSSED ALGORITHM

The operational facets for the aforementioned methodology are delineated as follows:

A. Pre - processing

- Acquire a Diverse Multi-Speaker Audio Dataset: Gather a dataset containing various multi-speaker audio clips in wav (Waveform Audio File) and mp3 (MPEG-1 Audio Layer 3) formats to form the foundation of the approach.
- Data Pre-Processing: Conduct essential pre-processing steps prior to feature extraction, starting with pre-emphasis using a first-order Finite Impulse Response (FIR) filter, followed by frame blocking to partition the speech signal into frames, reducing acoustic artifacts at the beginning and end of the signal.
- Utilize Python Audio Libraries: Employ robust Python audio libraries such as Librosa and PyAudio for audio processing. Librosa is selected for its user-friendly features, while built-in Python modules handle fundamental audio functionalities, enabling audio file cleansing and sampling, and facilitating waveform extraction for visualization.
- Generate and Analyze Spectrograms: Apply the Short-Term Fourier Transform (STFT) to analyze the amplitude of different frequencies at specific time intervals within an audio signal. Create spectrograms to visually represent the frequency distribution over time, providing comprehensive insights into the audio signal's dynamic nature and aiding in the interpretation of spectral characteristics for further analysis.

B. Audio feature extraction

Feature extraction serves as the process of transforming the speech waveform into a parametric representation that maintains valuable information at a reduced data rate, facilitating subsequent processing and analysis. The effectiveness of classification is intricately linked to the quality and precision of the extracted features input. Below tables TABLE I and TABLE II outlines the different techniques.

After seeing both the tables we can confer that Mel frequency cepstral coefficient (MFCC) is the best feature extraction technique for the purpose of Speech Diarization because of the high speed of computation, high reliability and good noise resistance.

TABLE I. SPEECH FEATURE EXTRACTION TECHNIQUES - I

Speech feature extraction Techniques	**Type of filter**	**Computation Speed**	**Types of coefficients**
Mel Frequency Cepstral Coefficient	Mel	High	Cepstral
Linear Prediction Cepstral Coefficient	Linear prediction	Medium	Low and medium
Line spectral Frequency	Linear prediction	Medium	Spectral
Discrete Wavelet Transform	Low and High pass	High	Wavelets
Perpetual Linear prediction	Bark	Medium	Autocorrection

TABLE II. SPEECH FEATURE EXTRACTION TECHNIQUES - II

Speech feature extraction Techniques	**Noise resistance**	**Reliability**	**Frequency Captured**
Mel Frequency Cepstral Coefficient	Medium	High	Low
Linear Prediction Cepstral Coefficient	High	Medium	Low and medium
Line spectral Frequency	High	Medium	Low and medium
Discrete Wavelet Transform	Medium	Medium	Low and medium
Perpetual Linear prediction	Medium	Medium	Low and medium

1) Sample Rate Selection

- Strategic Sampling Frequency: A sampling frequency of 44.1kHz is chosen to ensure optimal performance and minimize distortions in the audio signal.
- Transition Band Requirements: This frequency selection is driven by the need for a transition band from 20kHz (pass-band) to 22.05kHz (stop-band).
- Nyquist Zone Consideration: Ensuring that at least half of the transition band resides within the first Nyquist zone is essential to prevent potential aliasing from the second Nyquist zone.

2) Audio Framing Strategy

- Handling Non-Stationary Audio: Audio is divided into concise frames, assuming short-term stationarity to address potential distortions from the Fast Fourier Transform (FFT).
- Incorporating Frame Overlap: Overlapping frames are used to foster correlation and compensate for information loss at the edges after applying a window function.
- Applying the Hanning Window: A Hanning window is applied to each frame to maintain continuity and prepare the signal for seamless analysis.

3) Transformation to Frequency Domain

- Frequency Domain Conversion: The Fast Fourier Transform (FFT) is performed on each windowed frame, utilizing the Hanning window to transition the audio signal to the frequency domain.
- Exploring Spectral Characteristics: This frequency domain representation facilitates the exploration and analysis of the signal's spectral characteristics for advanced processing and feature extraction.

4) Calculation of Filter Banks

- Constructing Filter Points: Filter points are constructed to delineate the starting and stopping positions of the filters, converting the filter bank edges into the Mel frequency space.
- Generating and Normalizing Filter Banks: A linearly spaced array is generated within the two Mel frequencies, transformed back into the frequency space, and normalized with respect to the FFT size to provide a structured set of filter banks.

5) Filtering with Mel-Spaced Filter Banks

- Passing Audio Through Filters: The framed audio is passed through Mel-spaced filter banks to capture power distribution across various frequency bands.
- Adaptation for Frequency Bands: Mel-spaced filter banks are designed with an exponential growth pattern, allowing adaptation for any desired frequency band.

6) Building Filter Banks

- Area Normalization: Triangular Mel weights are divided by the width of the Mel band to ensure proper calibration and prevent noise increase with frequency.
- Maintaining Consistency and Accuracy: This normalization refines the filter banks to maintain consistency and accuracy across the frequency spectrum, enhancing reliability for subsequent analyses.

7) Derivation of Cepstral Coefficients

- Applying the Discrete Cosine Transform (DCT): The final phase involves applying DCT to extract high-frequency and low-frequency variations within the signal.
- Capturing Essential Features: The DCT analysis results in cepstral coefficients that provide a comprehensive representation of the spectral content and dynamics, refining and extracting critical information for further processing.

C. Voice Activity Detection

Voice Activity Detection (VAD), or speech activity detection, identifies the presence or absence of human speech in an audio signal. It is primarily used in speech processing applications like speech coding and recognition, optimizing processes during non-speech segments. For example, in Voice over Internet Protocol (VoIP) applications, VAD prevents the unnecessary coding or transmission of silent packets, saving computational resources and bandwidth.

The VAD process begins with acquiring Mel Frequency Cepstral Coefficients (MFCC), a crucial component in detecting voice activity. A notable benchmark for VAD performance is the bRTC implementation, which uses Gaussian Mixture Models (GMM) to model speech and non-speech sounds. This model relies on logarithmic energies from six frequency bands ranging from 80Hz to 4000Hz. VAD then classifies audio segments as voiced or unvoiced using fixed-point operations, which is particularly useful for telephony and speech recognition applications.

Google's VAD for the bRTC project is renowned for its speed, modernity, and cost-free availability, making it one of the best options for tasks requiring fast and accurate voice activity detection. Its effectiveness makes it a valuable tool in various speech processing applications.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 10, October 2024

D. Speaker Change Detection

In speaker diarization systems, accurately identifying boundaries between speech turns of different speakers is crucial, especially when multiple speakers are active in a single voice activity.

- Post-MFCC feature extraction: After Audio feature extraction further segmentation is done by clustering smaller groups to refine differentiation between simultaneous speakers.
- Gaussian Divergence Model: Utilizes the Gauss Divergence Theorem to divide a device into smaller data instances effectively, aiding in speaker change detection.
- Bayesian Information Criterion (BIC): Serves as a model selection criterion that addresses overfitting by introducing a penalty for the number of parameters. It is used in various domains, including time series and linear regression.
- Hierarchical Agglomerative Clustering (HAC): Employs a "bottom-up" approach where each observation starts in its own cluster, and clusters are merged progressively, resulting in a Dendrogram that illustrates the clustering process.
- Optimal Clustering Configuration: Analyzing the Dendrogram helps identify significant sections without intersections to determine the best clustering configuration for speaker change detection.

E. Re-Segmentation

The original audio is segmented into different sections according to the clustering of the most probable sound notes. Voice Activity Detection (VAD) further classifies the original audio into noise and voice activity segments. By combining the results of both clustering and VAD, we can distinguish the audio signal into separate segments for each speaker and noise. This integrated approach ensures a comprehensive analysis of the audio, effectively differentiating between multiple speakers and background noise, enhancing the accuracy and reliability of the speaker diarization system.

III. EXPERIMENTAL RESULTS AND ANALYSIS

The below images shows the Dendrogram and the Audio segmentation of a call with 2 speakers.

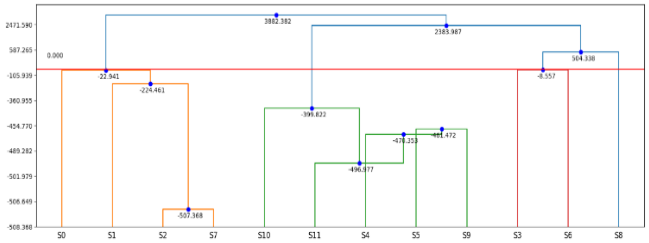

Fig. 1. Dendrogram of 2 speakers.

In Fig. 1, the x-axis represents individual speech segments, while the y-axis denotes the dissimilarity or distance between these segments. This dendrogram effectively illustrates the clustering of acoustically similar speech segments, with segments at lower y-axis levels indicating greater similarity. It is useful for visualizing speaker grouping based on acoustic properties.

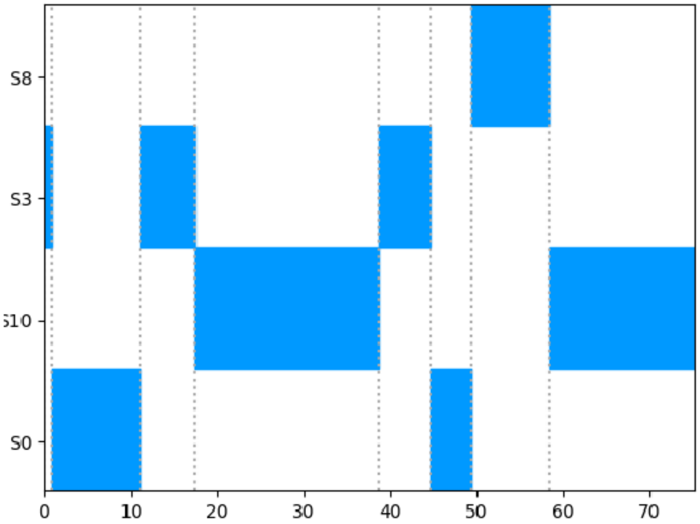

Fig. 2. Segmented Audio file

Fig. 2 presents the segmented audio files, where the Y-axis represents the clustered voice segments, and the X-axis denotes the time duration of the audio. This figure illustrates the final partitioning of the complete audio file, as determined by diarization, with distinct segments corresponding to different speakers.

Using Based on Voice Activity Detection (VAD), we observe the following results:

- Sound note S0 and Sound note S10: These segments represent two different speakers.
- Sound note S3: This segment depicts noise or silence, indicating no voice activity.
- Sound note S8: This segment shows an overlap between the two speakers, which the algorithm identifies as a separate voice note.

The results indicate an accuracy of 82.7 percent, which could potentially be improved to approximately 97 percent with clearer audio and no overlapping segments. However, since this is not feasible in real-world applications, the expected efficiency of this method remains at 82.7 percent.

IV. CONCLUSION

In conclusion, this manuscript introduces a novel two-step process to address the complexities of real-world speaker diarization scenarios involving multiple voices. The first phase involves audio segmentation using Gaussian Divergence, Bayesian Information Criterion, and Hierarchical Agglomerative Clustering, followed by resegmentation. The second phase employs voice activity detection to differentiate between individual speakers and background noise. With an accuracy of 72 percent, the proposed approach demonstrates effectiveness, although there is a decline in accuracy during the Hierarchical Agglomerative Clustering phase due to voice note breakdown. The results highlight the methodology's value in establishing a robust foundation for speaker diarization while identifying areas for improvement, particularly in separating combined speakers. Future iterations could integrate a voice recognition module on already separated voice notes to enhance individual speaker identification accuracy and overall system performance.

The future of speaker diarization holds promising opportunities for advancements in audio processing and recognition technologies. As computational capabilities evolve, integrating sophisticated algorithms and machine learning models into diarization systems is feasible. Research can focus on leveraging deep learning techniques, such as neural networks, to improve voice segmentation and identification accuracy. Additionally, incorporating contextual information, sentiment analysis, and language understanding into diarization systems can lead to a more comprehensive understanding of audio content. Collaborations with linguistics and psychology experts could result in systems capable of discerning emotional nuances and social dynamics in conversations.

Moreover, applying speaker diarization in customer service analytics, forensics, and social media monitoring offers opportunities for tailored solutions. Exploring real-time processing capabilities and edge computing could enable deployment in resource-constrained environments. In summary, the future of speaker diarization involves advancements in algorithmic sophistication, contextual information integration, and expanding applications across diverse domains, driven by evolving computational technologies and interdisciplinary collaborations.

ACKNOWLEDGMENT

I extend my gratitude to Dr. K P Lakshmi for her expertise and assistance in every aspect of our study, as well as her help in writing the manuscript.

REFERENCES

[1] G. Sun, D. Liu, C. Zhang, and P. C. Woodland, "Content-Aware Speaker Embeddings for Speaker Diarisation," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 7168-7172.

[2] Bai, Z.; Zhang, X.-L. Speaker recognition based on deep learning: An overview. Neural Netw. 2021, 140, 65–99. ISSN 0893-6080. [CrossRef] [PubMed].

[3] Mao, H.H.; Li, S.; McAuley, J.; Cottrell, G.W. Speech Recognition and Multi-Speaker Diarization of Long Conversations. arXiv 2020, arXiv:2005.08072.

[4] Inaguma, H.; Yan, B.; Dalmia, S.S.; Gu, P.; Jiatong Shi, J.; Duh, K.; Watanabe, S. ESPnet-ST IWSLT 2021 Offline Speech Translation System. arXiv 2021, arXiv:2107.00636.

[5] Ueda, Y.; Maiti, S.;Watanabe, S.; Zhang, C.; Yu, M.; Zhang, S.X.; Xu, Y. EEND-SS: Joint End-to-End Neural Speaker Diarization and Speech Separation for Flexible Number of Speakers. 2022.

[6] Zajíc, Z.; Kunešová, M.; Müller, L. Applying EEND Diarization to Telephone Recordings from a Call Center. In Speech and Computer. SPECOM 2021; Lecture Notes in Computer Science; Karpov, A., Potapova, R., Eds.; Springer: Cham, Switzerland, 2021; Volume 12997.

[7] Fürer, L.; Schenk, N.; Roth, V.; Steppan, M.; Schmeck, K.; Zimmermann, R. Supervised Speaker Diarization Using Random Forests: A Tool for Psychotherapy Process Research. Front. Psychol. 2020, 11, 1726.

[8] Z. Li and J. Whitehill, "Compositional Embedding Models for Speaker Identification and Diarization with Simultaneous Speech From 2+ Speakers," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 7163-7167.

[9] A. Anand, R. Donida Labati, M. Hanmandlu, V. Piuri and F. Scotti, "Text- independent speaker recognition for Ambient Intelligence applications by using Information Set Features," 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2017, pp. 30-35.

[10] Q. Sun, L. Luo, H. Peng, and C. An, "A Method of Speaker Recognition for Small-scale Speakers Based on One-versus-rest and Neural Network," 2019 14th International Conference on Computer Science & Education (ICCSE), 2019, pp. 771-774.

[11] S. R. Hasibuan, R. Hidayat and A. Bejo, "Speaker Recognition Using Mel Frequency Cepstral Coefficient and Self-Organising Fuzzy Logic," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020, pp. 52-55.

[12] N. M and A. S. Ponraj, "Speech Recognition with Gender Identification and Speaker Diarization," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-4.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329-5333.

[14] Tomi Kinnunen, Haizhou Li," An overview of text-independent speaker recognition: From features to supervectors, "Speech Communication, Volume 52, Issue 1,2010, Pages 12-40, ISSN 0167-6393.

[15] S. Nisar, I. Shahzad, M. A. Khan and M. Tariq, "Pashto spoken digits recognition using spectral and prosodic based feature extraction," 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI), 2017, pp. 74-78, doi: 10.1109/ICACI.2017.7974488