Speaker Identification System

Sneha Jakati Shreya S Nadgir Charvi Kolloori Jayapriya A N
School of Electronics and
Communication Engineering
KLE Technological University
Hubli, Karnataka

Prof. Akash Kulkarni
School of Electronics and
Communication Engineering
KLE Technological University
Hubli, Karnataka

***Abstract*—This report presents a text-dependent speaker identification system for access security and fraud prevention applications. The system employs signal-processing techniques for signal capture, preprocessing, and feature extraction. Utilizing cepstral analysis, the system aims for accurate speaker identification. While acknowledging challenges, including accuracy limitations, the framework contributes to enhancing security in sectors like banking and law enforcement, emphasizing its role in preventing unauthorized access and safeguarding sensitive information.**

***Index Terms*—Fast Fourier transforms (FFT), Mel-Spectrograms, Speech signal processing.**

I. Introduction

Speaker recognition, a subset of biometric authentication, has emerged as a pivotal generation in making sure of stable entry to sensitive facts and preventing fraudulent activities. Unlike conventional authentication methods reliant on passwords or tokens, speaker recognition systems provide a continuing and herbal means of identity verification by analyzing precise characteristics present in a character's voice.

The evolution of digital sign processing (DSP) strategies has drastically advanced the field of speaker reputation, permitting the extraction of tricky features from speech signals for accurate identification of speakers. By leveraging DSP methodologies, researchers and practitioners have evolved state-of-the-art algorithms and frameworks capable of discerning subtle nuances in speech styles, intonations, and vocal traits that distinguish one speaker from any other.

The importance of speaker recognition of the usage of DSP strategies is underscored with the aid of its extensive-ranging programs across numerous domain names. In sectors including banking, telecommunications, law enforcement, and access control, the potential to reliably authenticate individuals based totally on their voice signatures holds tremendous value for enhancing safety protocols, streamlining consumer authentication strategies, and combating identification fraud.

This paper explores the basics of speaker reputation and the usage of DSP strategies, delving into the underlying concepts of sign processing, characteristic extraction, and sample popularity that underpin modern-day speaker reputation systems. By examining key methodologies which include cepstral analysis, dynamic time warping, and machine learning algorithms, we intend to explain the problematic interplay between DSP strategies and speaker reputation accuracy.

Furthermore, this paper discusses the realistic challenges and issues inherent in designing and enforcing DSP-based speaker reputation structures, which include problems related to facts variability, environmental noise, and machine scalability. Through a complete evaluation of existing literature and research findings, we seek to offer insights into modern-day developments, improvements, and destiny directions inside the subject of speaker reputation using DSP strategies.

In summary, speaker reputation using DSP strategies represents a critical frontier inside the realm of biometric authentication, offering an amazing arsenal of equipment and methodologies for boosting access security, stopping fraud, and safeguarding sensitive information in an increasingly digital and interconnected international.

II. Literature Survey

The paper explores the efficacy of diverse Wavelet Transforms, including Daubechies, Symlets, and Coiflets, inside the context of automatic speaker reputation as compared to conventional signal processing parameters. Applying signal pre-processing techniques which includes preemphasis and normalization for robustness, the take a look at conducts special experiments, assessing recognition accuracy primarily based on factors like schooling statistics length, checking out information duration, and exclusive speaker counts in corpora. Additionally, the paper evaluates the overall performance of two classifiers, Gaussian Mixture Model (GMM) and Multi-Layer Perceptron (MLP), losing mild on their comparative effectiveness. Acknowledging scope barriers, which includes specializing in a closed set of speakers, and attaining high accuracy in perfect conditions; the research identifies gaps and proposes destiny work. This includes exploring models for spotting unknown audio system in real-world situations and carrying out research on large corpora with diverse noise conditions. Furthermore, the mixing of audio facts with other modalities is suggested for more desirable get right of entry to structures. Overall, the have a look at contributes to the expertise of Wavelet Transforms in speaker popularity, addressing limitations and imparting guidelines for future research [1].

The literature outlines the methodology for speaker popularity, emphasizing the adoption of Mel Frequency Cepstral Coefficients (MFCC) for effective feature extraction. It emphasizes the need for schooling a speaker database with speech samples for next reputation of unknown audio system. The key steps in speaker popularity include framing, windowing, Fast Fourier Transform (FFT), Mel-frequency scaling, and cepstrum computation. However, the literature lacks precise algorithmic info for MFCC and Vector Quantization (VQ) and does no longer offer records on the accuracy or challenges related to the methods. It highlights the significance of speaker reputation in various packages and recognizes the advanced accuracy of machines, particularly with a big range of audio system and short utterances. The textual content also mentions the utility of MFCC, an established approach in speech processing. On the drawback, the absence of specific algorithmic details and performance metrics, as well as the shortage of debate on challenges, limits the complete expertise of the speaker reputation strategies. The literature evaluates fails to explore latest advancements in deep learning techniques and neural networks in speaker recognition. Additionally, there is a gap in supplying a comprehensive review of the contemporary country of the art and actual-world programs or case research within the field [2].

The research makes a speciality of textual content-impartial speaker identification by way of using Learning Vector Quantization (LVQ) because the number one set of rules to evaluate the suitability of numerous spectral capabilities extracted from speech signals. It demonstrates the effectiveness of LVQ for refining codebook vectors, thereby enhancing speaker reputation precision. This deserves encompass successful spectral feature assessment and practical implications for safety and authentication structures. However, boundaries include a dataset typically focused on male speakers, doubtlessly overlooking gender-associated voice variations, and a lack of complete exploration of actual-global applications. The absence of in-intensity dialogue at the effect of varying test sentence lengths on speaker reputation performance increases questions on device adaptability. Gaps in the studies encompass constrained exploration of practical packages, a gender bias in speaker consciousness, and inadequate investigation into the effect of varying check sentence lengths. Future guidelines contain diversifying the dataset, exploring realistic packages more comprehensively, and investigating the impact of test sentence lengths for a extra adaptable speaker identification machine. In end, the look at contributes valuable insights into LVQ-primarily based speaker identification however suggests areas for in addition exploration and improvement [3].

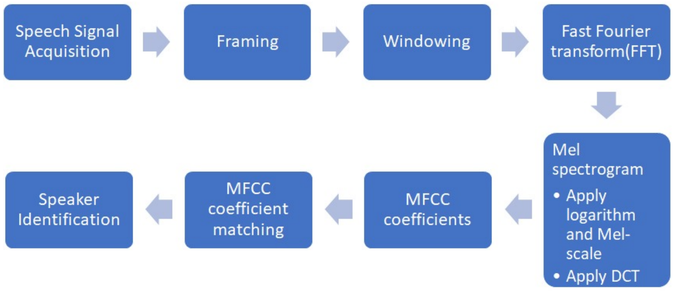
The proposed speaker identity gadget utilizes a piezoelectric transducer attached to a collar for shooting vocal twine vibrations, deviating from conventional microphone-based totally strategies. It focuses on efficient actual-time identification the usage of short utterances and demonstrates resistance to prosody variations. The machine achieves an excessive accuracy of 91%, outperforming other time-frequency techniques, whilst keeping simplicity in its method. Identified gaps consist of the need for accuracy development, the goal to construct a larger database, and exploring the industrial viability of the proposed collar. Future works have to address these gaps, behaviour robustness testing, and check scalability in real-world packages [4].

The described speaker identity device employs a multi-step method, encompassing silence removal, pre-emphasis, and segmentation of speech samples. Feature extraction is achieved through a completely unique mixture of the Mel Frequency Cepstral Coefficients (MFCC) technique with the Discrete Fractional Fourier Transform (DFRFT), introducing a singular method to enhance characteristic illustration. Speaker modelling includes the application of Vector Quantization (VQ) to create a codebook from feature vectors, supplying a established way for speaker representation. Matching rankings, important for decision-making, are calculated using Mean Squared Error (MSE). The technique's deserves lie in its progressive fusion of traditional and advanced strategies, supplying a capability improvement in characteristic extraction. However, demanding situations consist of the sensitivity of cepstral representation to environmental noise and the potential complexity hindering actual-time implementation. Notably, the text lacks a complete analysis of the proposed method in contrast to current speaker identification techniques, and the absence of information approximately the dataset and benchmarking limits the study's scope. A thorough literature survey is usually recommended to discover related research, in particular the ones addressing comparable challenges and innovations in speaker identity [5].

The proposed methodology for speaker popularity integrates each low-degree and excessive-degree auditory sign processing models, emphasizing temporal dynamics at the cochlea level and modulation additives evaluation within the inferior colliculus. This technique pursuits to beautify speaker-unique styles within the auditory sign. The dimension discount techniques, involving averaging and centroid calculation, contribute to the extraction of significant functions from temporal envelopes, presenting a greater green speaker popularity machine. The method demonstrates merits which includes a holistic representation of speech signals thru comprehensive auditory functions, the fusion of auditory and mel-cepstral capabilities for complementary information, and a found development in overall performance. However, capability demerits encompass elevated computational complexity and variable impact on overall performance advantage across exclusive speaker reputation responsibilities.

The study acknowledges barriers, along with the shortage of real-world checking out, comprehensive exploration of demanding situations, and a discussion on ethical concerns. Future directions contain accomplishing experiments in various settings for generalizability, addressing computational complexity and useful resource necessities, and exploring ethical implications related to privacy worries in speaker reputation systems using auditory functions [6].

III. Methodology



Speech Signal Acquisition
Framing
Windowing
Fast Fourier transform(FFT)
Mel spectrogram
• Apply logarithm and Mel-scale
• Apply DCT
MFCC coefficients
MFCC coefficient matching
Speaker Identification

Fig. 1. Proposed pipeline

- Speech signal acquisition: This is the first step in the process, where the sound of your voice is captured by a microphone. The microphone converts the sound waves into an electrical signal.
- Framing: The continuous speech signal is broken down into smaller chunks of time called frames. Each frame typically consists of a few milliseconds of speech.
- Windowing: A window function is applied to each frame to reduce the discontinuity at the edges of the frame. This helps to prevent artifacts in the spectrum of the signal.
- Fast Fourier Transform (FFT): The FFT is used to convert each frame from the time domain to the frequency domain. This reveals the different frequencies that make up the speech signal.
- Mel spectrogram: The output of the mel-filter is a mel spectrogram, which is a visual representation of the speech signal in the frequency domain. The mel spectrogram shows the intensity of different frequencies over time.
- MFCC extraction: Mel-frequency cepstral coefficients (MFCCs) are extracted from the mel spectrogram. MFCCs are a set of features that are derived from the mel spectrogram and that are used to represent the speech signal.
- MFCC feature vector formation: The MFCCs from each frame are combined to form an MFCC feature vector. The MFCC feature vector is a representation of the speech in a single frame.
- Speaker identification (optional): The MFCC feature vectors can be used to identify the speaker. This is done by comparing the MFCC feature vectors of the unknown speaker to those of known speakers.
- Speech recognition: The MFCC feature vectors are used to recognize the speech. This is done by training a speech recognition model on a large corpus of labeled speech data. The model learns to associate certain patterns of MFCCs with certain words or phrases.

IV. Dataset

For the data acquisition phase, we are gathering speech samples, each lasting 1 second. All speakers are required to utter identical phrases within the same time frame and under consistent environmental conditions.

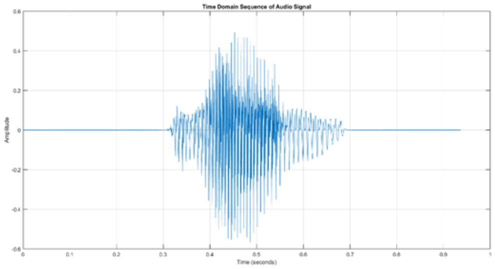


Fig. 2. Input signal

- Signal length:1 second
- Amplitude range:+/-0.6
- Sampling frequency:16000Hz
- Frequency range:100Hz to 8000Hz

V. Frequency Analysis

A. Discrete Fourier Transform(DFT)

Discrete Fourier Transform is a mathematical technique used in signal processing and digital image processing to analyze the frequency content of a discrete-time signal or a sequence of values.

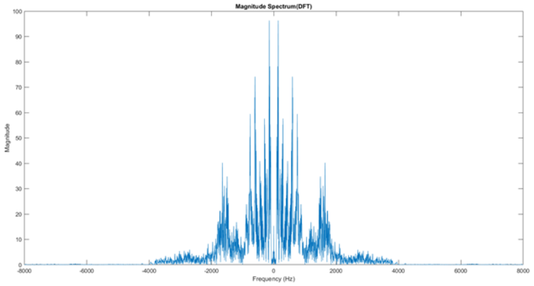


Fig. 3. DFT of the input signal

- Number of samples:14976
- Computation time: 35.999539 seconds
- Execution time:161.490108 seconds

B. *Fast Fourier Transform(FFT)*

Fast Fourier Transform is an algorithm used to compute the Discrete Fourier Transform (DFT) of a sequence or an array of data points. The FFT algorithm efficiently computes the DFT and is widely used in various applications where the analysis of frequency components of a signal is required.

The FFT algorithm was developed to overcome the computational inefficiencies of directly computing the DFT, especially for large data sets. It significantly reduces the number of arithmetic operations needed to compute the DFT, making it feasible to analyze signals in real-time applications and large-scale data processing.

The two primary approaches within the realm of FFT algorithms are indeed DIF (Decimation-In-Frequency) FFT and DIT (Decimation-In-Time) FFT.

1) Decimation-In-Time Fast Fourier Transform(DIT-FFT): DIT FFT algorithms begin by decomposing the input sequence in the time domain.

- Number of samples:14976
- Computation time: 0.579495 seconds
- Execution time:1.097778 seconds

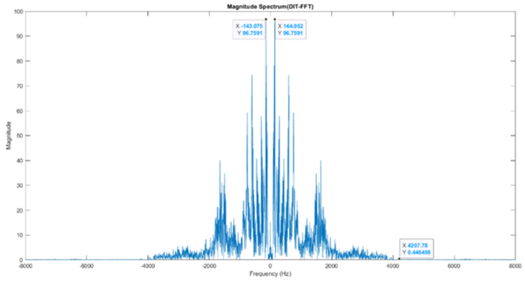


Fig. 4. DIT FFT of the input signal

2) Decimation-In-Frequency Fast Fourier Transform(DIF-FFT): DIF FFT algorithms initially decompose the input sequence in the frequency domain.

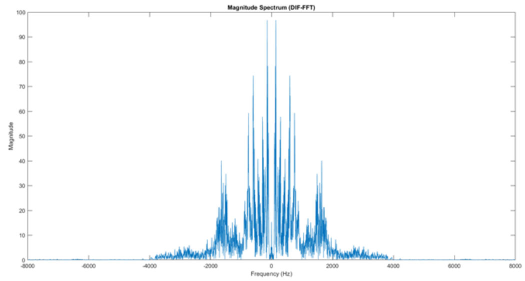


Fig. 5. DIF FFT of the input signal

- Number of samples:14976
- Computation time: 0.579495 seconds
- Execution time:1.097778 seconds

VI. Digital Filter Design

From the FFT of the audio signal it is clear that we only need frequencies till 500Hz as that will be in the human range and cover the maximum frequencies that we need. A filter with an order of 10 is chosen to strike a balance between achieving the specified pass band and stop band characteristics while maintaining computational efficiency.

- Passband edge frequency:1Hz
- Stop band edge frequency:500Hz
- Pass band ripple: 3dB
- Stop band ripple: 40dB
- Order of the filter:10
- Sampling frequency:16000Hz

A. *Digital IIR Butterworth Filter Design*

A digital Butterworth Infinite Impulse Response (IIR) filter is a fundamental tool in digital signal processing, renowned for its maximally flat frequency response in the pass band and gradual roll-off characteristics. Derived from its analog counterpart, the Butterworth filter exhibits a uniform distribution of poles around the z-plane, a key feature influencing its frequency response and filtering behavior.

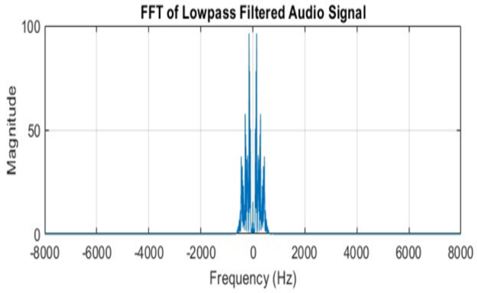


Fig. 6. FFT signal of the Digital IIR Butterworth Filter

B. *Digital FIR Hamming window Filter Design*

A digital Finite Impulse Response (FIR) filter designed with a Hamming window is a versatile tool extensively used in digital signal processing applications. Leveraging the finite-duration impulse response characteristic, FIR filters provide stable and linear phase responses essential for preserving signal integrity. The Hamming window, a commonly employed windowing function, plays a pivotal role in shaping the frequency response of the FIR filter. By minimizing side lobes in the frequency domain, the Hamming window enhances the filter's performance, attenuating unwanted frequency components effectively.

A low-pass filter is a type of electronic filter that allows signals with frequencies lower than a certain cutoff frequency to pass through while attenuating signals with frequencies higher than the cutoff frequency. The characteristics of a low-pass filter are defined by its cutoff frequency, which is the frequency at which the filter begins to attenuate the signal, and its roll-off rate, which describes how quickly the filter attenuates frequencies beyond the cutoff.They find applications in various fields such as audio equalization, telecommunications, and control systems, where the emphasis is on preserving the fundamental aspects of a signal while suppressing higher-frequency interference.

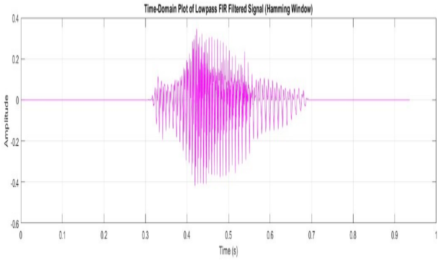


Fig. 7. Time-Domain of Lowpass FIR Filtered signal (Hamming Window)

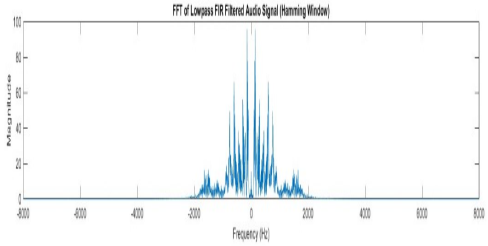


Fig. 8. FFT of Lowpass FIR Filtered input signal(Hamming Window)

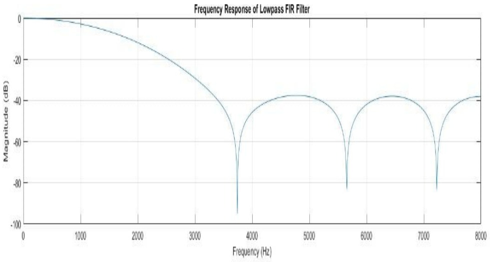


Fig. 9. Frequency Response of Lowpass FIR Filter(Hamming Window)

VII. Results and Discussion

Applying only filters to the audio will not serve our purpose. Hence we have opted for the Mel Frequency cepstral coefficient

A. *Mel-Frequency Cepstral Coefficients (MFCCs)*

Mel-Frequency Cepstral Coefficients (MFCCs) represent a critical feature extraction technique extensively used in speech and audio processing domains. Built upon the Mel-frequency scale, which models the non-linear human auditory response to frequencies, MFCCs offer a perceptually meaningful representation of audio signals.

Beginning with the Fourier transform of a signal to obtain its power spectrum, the process involves logarithmic scaling and transformation using the discrete cosine transform (DCT). The resulting coefficients, known as MFCCs, encapsulate the essential spectral characteristics of the signal.

TABLE I
MFCC Coefficients

MFCC	Speaker 1	Speaker 2
1	-380.523	-71.814
2	0.137	0.043
3	14.604	288.788
4	0.090	-0.037
5	2.233	-19.135

- MFCC 1: Often represents the overall energy or loudness of the speech signal within a specific frequency range.
- MFCC 2: Captures spectral variations related to the fundamental frequency and formant structure of the speech signal.
- MFCC 3-5: These coefficients usually encode information about spectral peaks and valleys, highlighting resonant frequencies and distinguishing features of the speech signal.

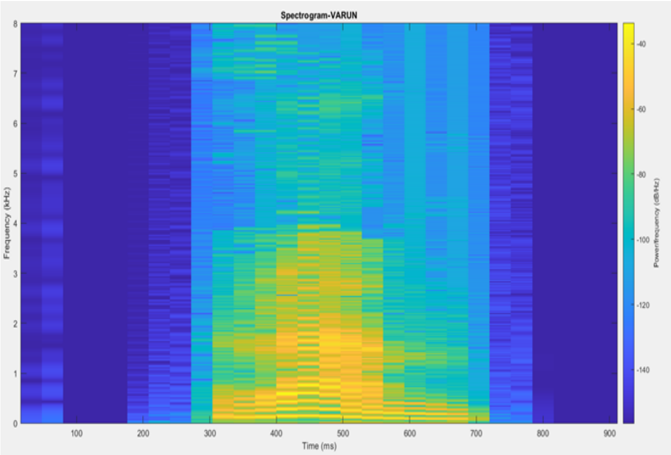


Fig. 10. Spectrogram of Speaker 1

In MFCC analysis:

- The x-axis commonly denotes time or frame number, illustrating the temporal progression of the audio signal. This enables the observation of spectral changes over time, crucial for understanding dynamic audio features.
- The y-axis in MFCC plots corresponds to the MFCC coefficients, which capture the spectral details within each frame of the audio signal. These coefficients represent the distribution of signal energy across different frequency bands, providing insights into the underlying spectral characteristics.

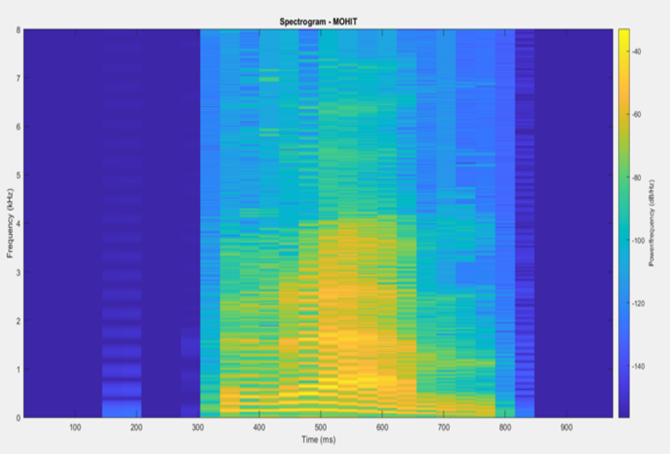


Fig. 11. Spectrogram of Speaker 2

- Intensity, which signifies the loudness or amplitude of the signal, indirectly influences spectral characteristics. Variations in intensity impact the perceived prominence of spectral features, influencing the computation and interpretation of MFCC coefficients. However, intensity considerations are typically addressed separately from MFCC analysis, often through preprocessing techniques to ensure consistent feature representation.

B. Euclidean Distance

The Euclidean distance calculation is a fundamental method for measuring the straight-line distance between two points in Euclidean space.

In the context of comparing two feature vectors such as MFCC coefficients, the Euclidean distance can determine how dissimilar or similar the feature vectors are. Smaller Euclidean distances imply more similarity, while larger distances indicate greater dissimilarity.

$$A = [a_1, a_2, \ldots, a_n];$$
$$B = [b_1, b_2, \ldots, b_n];$$

$$\text{Euclidean\_distance} = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2};$$

VIII. Conclusion

The report presents an in-depth exploration of speaker identification systems, emphasizing the critical role of digital signal processing (DSP) techniques, particularly Mel-Frequency Cepstral Coefficients (MFCCs), in achieving accurate speaker recognition. By leveraging DSP methodologies, the system aims to extract intricate features from speech signals, enabling reliable authentication based on unique vocal characteristics. Through a thorough literature survey, the paper highlights advancements in various techniques, such as Wavelet Transforms and Gaussian Mixture Models, while acknowledging challenges like data variability and environmental noise. The methodology outlined in the paper encompasses signal acquisition, framing, FFT computation, MFCC extraction, and speaker identification, culminating in experimental analysis and results evaluation.It helps improve speaker recognition technology by tackling problems and suggesting areas for future research. It aims to make access security stronger and protect important information in our digital world.

References

[1] P. Král, ”Discrete Wavelet Transform for automatic speaker recognition,” 2010 3rd International Congress on Image and Signal Processing, Yantai, China, 2010, pp. 3514-3518, doi: 10.1109/CISP.2010.5646691.

[2] Seema P, 2020, Speech Signal Analysis and Speaker Recognition by Signal Processing, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH and TECHNOLOGY (IJERT) IETE – 2020 (Volume 8 – Issue 11),

[3] Li Liu, Jialong He, and G. Palm, “Signal modeling for speaker identification,” 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 1996, pp. 665-668 vol. 2, doi: 10.1109/ICASSP.1996.543208.

[4] D. Ishac, A. Abche, E. Karam, G. Nassar, and D. Callens, ”A text-dependent speaker-recognition system,” 2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Turin, Italy, 2017, pp. 1-6, doi: 10.1109/I2MTC.2017.7969677.

[5] M. S. Walia, ”Discrete Fractional Fourier Transform and Vector Quantization Based Speaker Identification System,” 2014 Fourth International Conference on Advanced Computing and Communication Technologies, Rohtak, India, 2014, pp. 459-463, doi: 10.1109/ACCT.2014.41.

[6] T. F. Quatieri, N. Malyska and D. E. Sturim, ”Auditory signal processing as a basis for speaker recognition,” 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684), New Paltz, NY, USA, 2003, pp. 111-114, doi: 10.1109/ASPAA.2003.1285832.