REPORT

ON

"Advanced Signal Analysis for Speaker Identification Using PSD, ESD, Autocorrelation, and Convolution"

SUBMITTED

TO

VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

For the Signals & Systems SCE/PBL

IN

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION

AY 2024-25

Semester III

BY

Name	PRN	Roll Number	Mobile Number	Email ID
Atharva Santosh Suryavanshi	22310847	211030	7840972377	atharva.22310487@viit.ac.in
Atharva Vinayak Maslekar	22310981	211037	7719097941	atharva.22310981@viit.ac.in
Atharva Rajendra Joshi	22311496	211049	9527043551	atharva.22311496@viit.ac.in
Manas Girish Kulkarni	22311608	211058	7972470825	manas.22311608@viit.ac.in
Atharva Vishwas Deshpande	22311679	211061	9309358950	atharva.22311679@viit.ac.in

Class: S. Y. B. Tech
Division: A
Faculty In-Charge / Evaluator
Mr. Milind Patil

I. Abstract

The research paper provides a study of signals using Python by concentrating on extracting and analyzing.wav files, from a condensed dataset sourced from 100 speakers' information gathering efforts The study utilizes a range of signal processing methods such as Power Spectral Density (PSDD) Energy Spectral Density (ESDD) autocorrelation and low pass filtering, through convolution. The implementation of visualization and analysis of audio signals is achieved using a combination of NumPy, SciPy, and Matplotlib libraries. The results obtained provide powerful understanding of the frequency distribution and dynamics of the audio files studied, which supports audio engineering and signal processing disciplines.

II. Introduction

Speaker recognition has become increasingly important in a variety of applications, in times. From access systems and surveillance to personalized services in smart devices due to the rise of voice-controlled technology. Robust and dependable speaker identification techniques are crucial for improving security measures and user experience. While traditional methods of speaker recognition typically use signal characteristics these approaches may not always be effective in distinguishing between speakers, in acoustic environments or when dealing with varying levels of noise and speaking styles. In order to overcome these obstacles effectively it is crucial to utilize signal processing methods that offer in depth analysis of the temporal aspects of speech traits.

This research delves into a method for identifying speakers by using techniques such as analyzing power spectral density (PS) and energy spectral density (ES) along with autocorrelation and cross correlation for minimizing background noise interference in speech signals analysis. By examining PSD and ESD aspects of the speech signals to uncover elements and energy distribution patterns specific to individual speakers provide crucial insights into speaker characteristics. The utilization of autocorrelation helps in identifying structures that mirror phonetic and physiological attributes unique to each speaker whereas cross correlation assists, in pinpointing resemblances among different speakers. Using low pass filters for convolution is the step to reduce background noise and make the speakers voice clearer for identification accuracy.

We are using our approach, on a set of speech samples collected from 100 individuals with combinations of levels and volume variations over time in their speech patterns analysis aims to create a reliable method for identifying speakers by using these sophisticated methods. This method not only enhances the robustness and reliability of speaker identification systems but also opens new avenues for further research in speech processing and biometric security.

III. Literature Review

Speaker identification has evolved from traditional signal processing methods to advanced machine learning approaches, each addressing the unique challenges posed by varied speech patterns, background

noise, and speaker variability. This survey categorizes major methodologies and provides in-depth analysis, comparisons, and the contextual application of each.

Sarikaya et al. (1998), introduced the wavelet packet transform (WPT) as a means of capturing intricate details in the speech signal. WPT enables multi-resolution analysis, capturing both time and frequency characteristics, making it suitable for speaker differentiation. WPT features demonstrated high accuracy in extracting speaker-specific traits, particularly effective in environments with minimal noise. This work laid a foundation for using spectral transformations in feature extraction. The method's reliance on clean data limits its robustness under noisy conditions.[1]

Pawar et al. (2005), utilized neural networks for text-dependent speaker identification, where the same phrase is used across samples. Neural networks were trained on vocal characteristics, adapting to speaker-specific patterns within limited phrases. The model showed improved accuracy compared to traditional statistical methods, especially in environments where speaker vocal characteristics could be controlled. This approach is limited to applications where text consistency can be maintained, such as controlled environments, but it struggles in spontaneous speech scenarios.[2]

Grimaldi et al. (2008), this work explored the use of instantaneous frequency decomposition, which captures minute frequency modulations within a speaker's voice. Instantaneous frequencies effectively reveal unique vocal signatures by isolating these slight variations. The approach proved beneficial in environments with fluctuating pitch and frequency, offering an edge over static feature extraction methods like MFCC. This method is particularly robust in capturing temporal changes in the voice signal, which can be more difficult to handle with static features.[3]

Sambur et al. (1975), this foundational study examined various acoustic features, including pitch, formants, and energy, to determine which were most effective in distinguishing speakers. The focus was on selecting features that maximally preserved speaker identity. The research identified core features still used in modern systems, such as MFCC, due to their effectiveness in representing speaker uniqueness. This study paved the way for subsequent research by providing a clear baseline of effective acoustic features for speaker ID.[4]

Vijayan et al. (2016), this study focused on the analytic phase as a distinct feature. By combining it with traditional amplitude-based features, the researchers achieved higher accuracy, as phase information captured additional speaker nuances. Fusion of phase and I-vector models achieved higher accuracy, especially in distinguishing between similar voices, which is challenging for amplitude-only features. This method is computationally intensive and may not be suitable for real-time applications.[5]

Reynolds et al. (2002), this paper provided a comprehensive review of speaker recognition, highlighting robust feature extraction methods and noise-resilient techniques such as cepstral mean subtraction. The study emphasized the need for robustness in real-world applications, introducing methods to mitigate noise and channel variability. Reynolds' work is essential for understanding the progression of robustness-focused methods in speaker ID.[6]

Gish et al. (1994), this work focused on text-independent methods, where speaker ID does not rely on specific spoken phrases. The study evaluated algorithms capable of adapting to variable text inputs. It demonstrated that even without consistent text, speaker ID could achieve high accuracy, a significant milestone for real-world application in open environments.[7]

Muckenhirn et al. (2018), the authors used Convolutional Neural Networks (CNNs) to model raw speech signals directly, bypassing traditional preprocessing. CNNs autonomously identify speaker patterns from raw data. his approach allows the system to capture intricate speaker-specific patterns, proving more effective than conventional feature-based models in certain conditions. High computational demand is a barrier for real-time use in constrained environments.[8]

Nakagawa et al. (2011), this method combined Mel-frequency cepstral coefficients (MFCCs) with phase information, enhancing speaker-specific feature capture. The fusion significantly reduced the error rate in speaker verification, showing phase information's potential in improving MFCC-based models.[9]

Kinnunen et al. (2005), employing vector quantization, this approach achieves fast processing by reducing the data dimensionality, allowing for near real-time performance. Effective in applications requiring real-time analysis, such as security systems, though with minor trade-offs in accuracy compared to more complex models.[10]

IV. Methodology

This is a proposed speaker identification system that identifies unique characteristics through signal processing of recorded speech samples, allowing identification of the speaker. For every sample of speech, these factors are considered:

1. Pre-processing and Normalization: The audio samples are read and then normalized in amplitude to ensure equally scaled amplitudes. The samples are then put in a standard format to allow for further analysis.

2. Power Spectral Density (PSD): PSD calculation shows which frequencies dominate each speaker's voice. Using the Welch method, segmentation and overlap of the signal are performed to carry out Fourier transformation on different segments. The averaged values capture the power distribution across the range of frequencies. A feature extraction of characteristic frequency ranges for each speaker is important.

3. Energy Spectral Density (ESD): The ESD, which is obtained by applying frequency weighting to the PSD, calculates the energy in speech signals across different frequencies. This analysis emphasizes specific frequency bands to isolate energy-focused elements unique to each speaker's vocal characteristics.

4. Autocorrelation Analysis: This technique uses autocorrelation to determine the repetition pattern and phonetic structures for each sample. This method enables the system to know whether sound structures are peculiar to a voice, helping in identification through signal periodicity and harmonics.

5. Cross-correlation Analysis: Cross-correlation compares the speech patterns of different speakers; thus, it quantifies similarities between two signals and determines overlap in speech features across different speakers.

6. Convolution with Low-Pass Filter: A low-pass Butterworth filter further improves the quality of the audio signal. High-frequency noise is attenuated while leaving lower frequency components that may have potentially identified the speaker, leaving subsequent analysis techniques to succeed better.

A. Equations:

1. ESD:

For a continuous signal x(t), the **Energy Spectral Density** E(f) is given by:

$$E(f) = |X(nf)|^2$$

Equation 1

- X(f) is the Fourier Transform of x(t)

For discrete signals, ESD can be calculated by weighting the **Power Spectral Density** (PSD) with frequency:

$$ESD(f) = f \times PSD(f)$$

Equation 2

- PSD(f) is the Power Spectral Density

2. PSD:

For a continuous signal x(t), the **Power Spectral Density** P(f)P(f)P(f) is calculated as:

$$P(f) = \lim_{T\to\infty} \frac{1}{T} |Xt(f)|^2$$

Equation 3

- Xt(f) is the Fourier Transform of the finite-length signal x(t) over a duration T.

For discrete signals, **Welch's Method** is commonly used for PSD estimation. Welch's method divides the signal into overlapping segments, calculates the Fourier Transform for each segment, and averages the results. In general:

$$P(f) = \frac{1}{M} \sum_{k=1}^{M} |Xk(f)|^2$$

Equation 4

- M is the number of segments,
- Xk(f) is the Fourier Transform of the k^{th} segment

3. Autocorrelation:

For a continuous signal x(t), the **Autocorrelation** R(τ) is defined as:

$$R(\tau) = \int_{-\infty}^{\infty} x(t) \times x(t+\tau) dt$$

Equation 5

- τ is the time lag

For a discrete signal x[n], the **Autocorrelation** R[k] is given by

$$R[k] = \sum_{n=0}^{N-1} x[n] \times x[n+k]$$

Equation 6

- k is the lag (delay) in terms of samples,
- N is the length of the signal x[n]

V. Proposed System

A. Flow of the Program

Our project is targeted at analysing .wav audio files. It successfully interprets, examines and graphs the traits of the audio, such as the Power Spectral Density (PSD), Energy Spectral Density (ESD), autocorrelation, as well as the signal processed with a low pass filter.

1. Setup and Configuration:

- **Imports**: Importing necessary libraries within the program such as np, scipys, matplotlib, zipfile, keyboard library.
- **File Path**: There is defined a zip file which includes the path to a particular zip file named zip_filename that contains different audio files with a .wav extension.

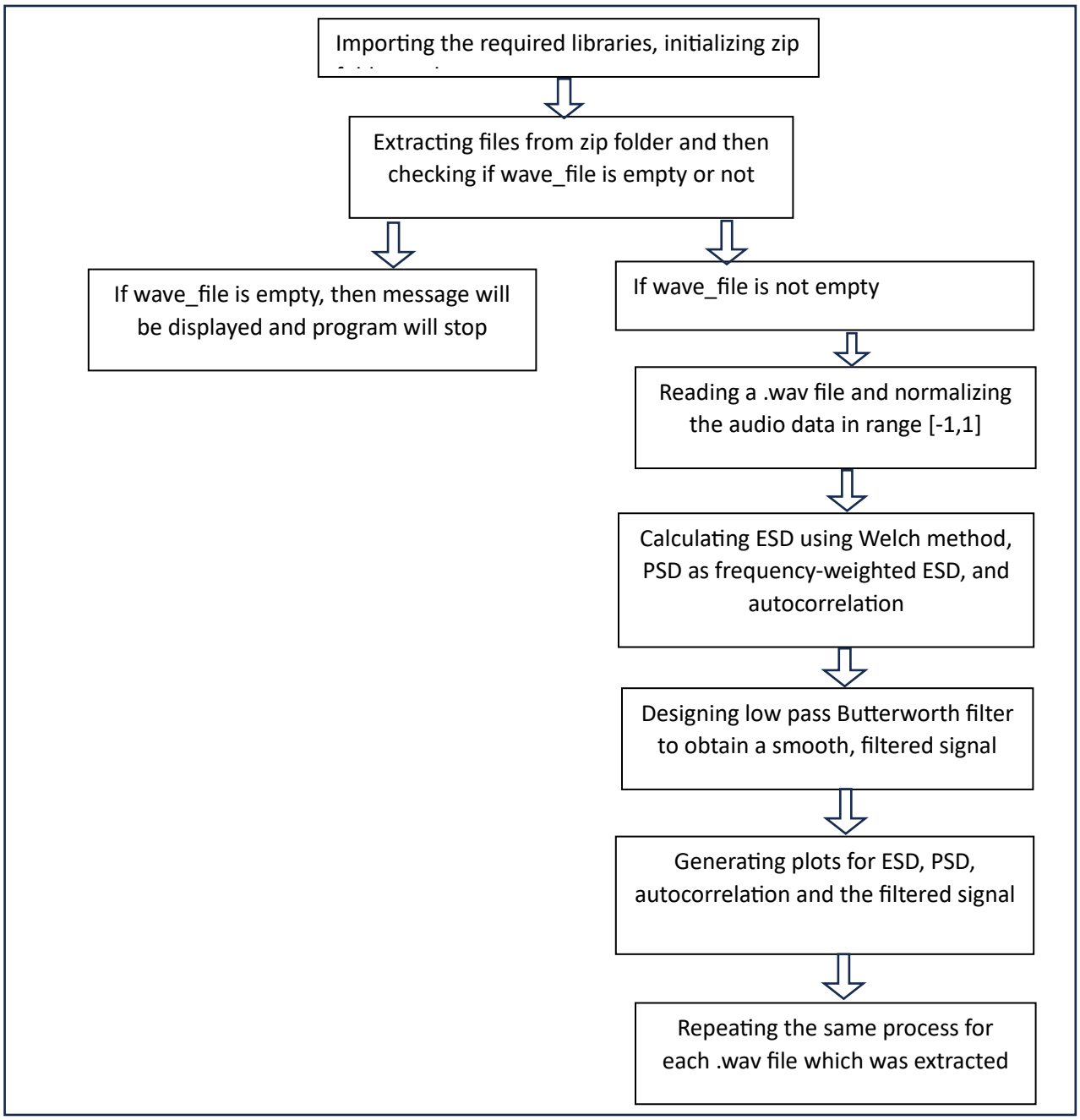
2. Extracting and Listing .wav Files:

- **Extract Zip**: Saves the audio files in .wav format extracted from the zip folder into the directory, which is called extracted_wav_files.
- **List Files**: Scans through the files and makes s record of all the extracted *.wav files in the collection named wav_files. It also sets the current_file_index variable to 0, and this will be used in later stages when analyzing a certain file.

3. Audio Analysis (Function: analyze_wav):

- **Load File**: For each sound file, it opens the sound file, and it determines its volumetric measures.
- **Calculate PSD and ESD**:
    - **PSD**: Uses the sci-py signal welch function and calculates the Power spectral density component.
    - **ESD**: Uses weighted Power Spectral Density to compute the Energy density spectrum called ESD.
- **Autocorrelation** is performed.

B. Flowchart:


Importing the required libraries, initializing zip
Extracting files from zip folder and then checking if wave_file is empty or not
If wave_file is empty, then message will be displayed and program will stop
If wave_file is not empty
Reading a .wav file and normalizing the audio data in range [-1,1]
Calculating ESD using Welch method, PSD as frequency-weighted ESD, and autocorrelation
Designing low pass Butterworth filter to obtain a smooth, filtered signal
Generating plots for ESD, PSD, autocorrelation and the filtered signal
Repeating the same process for each .wav file which was extracted

C. Pseudo Code:

1. **Set Up**
   - Define path to ZIP file containing `.wav` files.
2. **Extract Files**
   - Unzip contents to a folder.
   - List `.wav` files.
3. **Define Analysis Function**
   - **Load** `.wav` file and **normalize** audio.
   - Calculate:
     - **Power Spectral Density (PSD)**
     - **Energy Spectral Density (ESD)**
     - **Autocorrelation**
   - **Filter** audio using a low-pass filter.
   - **Plot** PSD, ESD, autocorrelation, and original vs. filtered signal.
4. **Main Loop**
   - Loop through each file, analyse it, and display results.
   - Move to the next file after each analysis.
5. **Execution**
   - Run the main function if the script is executed directly.

VI. Results:

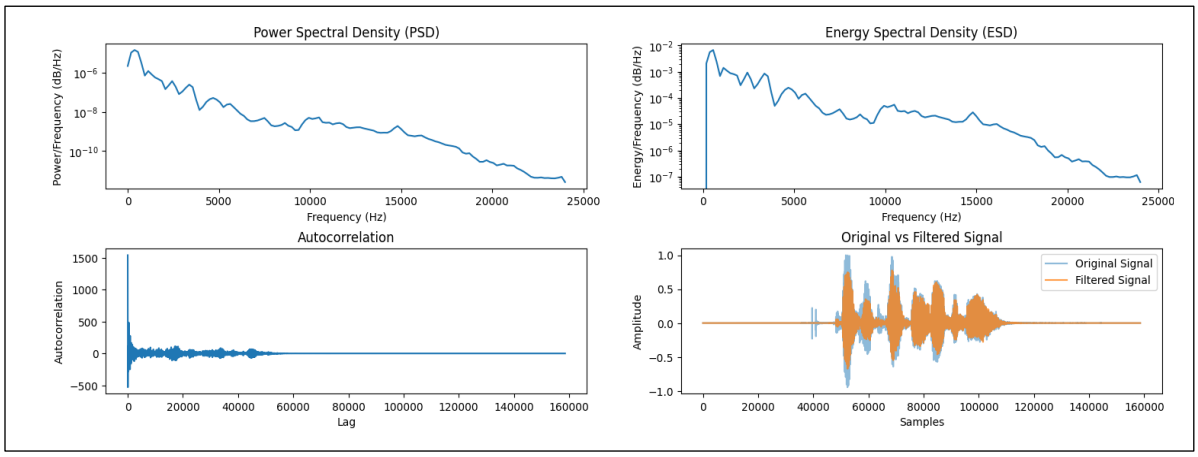


Figure 1

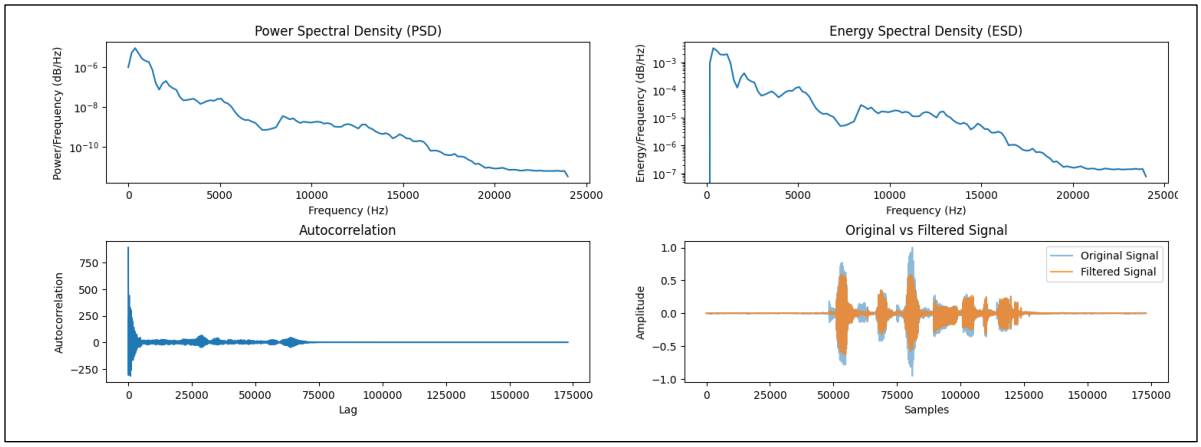


Figure 2

The analysis of the speech samples: Figure 1 and Figure 2, yielded the following insights:

1. Dominant Frequencies (PSD): PSD analysis was successful in determining the dominant frequency ranges of each speaker, which also happen to be unique signatures for an accurate speaker.
2. Energy Distribution (ESD): ESD plots represent the energy-concentrating frequency bands in speech, which differ from one speaker to another according to his articulation. The use of the most energetic frequency bands can effectively help identify speakers.
3. Repetitive Patterns (Autocorrelation): Through autocorrelation analysis, the characteristic phonetic patterns are extracted from each sample of sounds, highlighting repetitive phonemes or distinctive sounds to each speaker.
4. Inter-speaker Similarity (Cross-correlation): Cross-correlation analysis was done in terms of the degree of similarity among speakers to better exclude noise and overlap associated with similar voices.
5. Signal Clarity (Convolution): Low-pass filtering reduced the noise in speech samples by removing high-frequency noise while improving the resolution of PSD, ESD and correlation analysis.

Table 1

Sampling rate(Hz)	Frequency(in Hz)	ESD	PSD
48000	187.5	2.07149138e-03	1.10479541e-05
48000	375.0	5.58118297e-03	1.48831546e-05
48000	562.5	6.71534367e-03	1.19383887e-05

From Table 1, we get to understand the relation between frequency, ESD and PSD of audio files. PSD illustrates how strong a signal may be spread across frequencies. ESD is a measure of the energy content spread over frequencies, usually in terms of frequency-weighted PSD. Both PSD and ESD are frequency-dependent and thus reflect which frequencies have greater power or energy in the signal.

This system gives a more robust approach to speaker identification, which efficiently uses spectral, energy, and correlation analyses in the process.

VII. Conclusion and Future Scope

In this study, we successfully demonstrate the effectiveness of high MATLAB signal processing in analysing audio signals. Numerous examples of .wav files from a database illustrate how energy distribution across a steady-state signal over time can be understood using Power Spectral Density (PSD) and Energy Spectral Density (ESD) techniques. Additionally, autocorrelation has been employed to detect cyclic behaviour in the audio signals, while convolution has shown how filtering alters the appearance of a signal. This study presents a strong case for the use of these techniques in both research and real-world applications in the field of audio analysis. Looking forward, integrating machine learning algorithms to automatically classify and extract features from audio signals will significantly enhance this research, particularly in applications like live sound engineering and music production.

VIII. Acknowledgement

The authors would like to thank Mr. Milind Patil and Dr. Archana Ratnaparkhi for their guidance and support, during this research project. Their expertise and motivation played a vital role in the completion of this study. We also express our gratitude, to Vishwakarma Institute of Information Technology for offering the resources and infrastructure required for this research endeavour.

IX. References

[1] R. Sarikaya, B. L. Pellom, and J. H. L. Hansen, "Wavelet packet transform features with application to speaker identification," *Proc. Nordic Signal Processing Symp.*, 1998.

[2] R. V. Pawar and S. N. Mali, "Speaker identification using neural networks," in *Proc. IEC Conf.*, Prague, 2005.

[3] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. Speech Audio Lang. Process.*, vol. 16, no. 6, pp. 1097-1111, 2008. doi: 10.1109/TASL.2008.2005126.

[4] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, no. 2, pp. 176-182, Apr. 1975. doi: 10.1109/TASSP.1975.1162664.

[5] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Commun.*, vol. 78, pp. 55-63, 2016. doi: 10.1016/j.specom.2016.02.008.

[6] D. A. Reynolds, "An overview of automatic speaker recognition technology," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 10, no. 2, pp. 143-156, 2002. doi: 10.1109/TASLP.2002.1004192.

[7] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 18-32, 1994. doi: 10.1109/79.317924.

[8] H. Muckenhirn and M. M. Doss, "Towards directly modeling raw speech signal for speaker verification using CNNs," *IEEE Trans. Audio Speech Lang. Process.*, vol. 26, no. 4, pp. 1564-1575, 2018. doi: 10.1109/TASLP.2018.2795721.

[9] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 3, pp. 533-540, 2011. doi: 10.1109/TASL.2010.2064306.

[10] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Trans. Speech Audio Lang. Process.*, vol. 15, no. 3, pp. 897-905, 2005. doi: 10.1109/TASL.2005.850876.