IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 3, May-June 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203073

Analyzing Gender Differences in Voice using Cross-Correlation

Karambir [1], Kirti Hooda [2]

P.G. Student, Dept. of ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India[1]

Assistant Professor, Dept. of ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India[2]

ABSTRACT: Engineers have long struggled with human speech analysis for a number of reasons, such as product evaluation, identifying emotional states, developing artificial intelligence, and much more. These days, determining gender is a significant problem in speech analytics. Using auditory information like pitch, median, and frequency to determine gender. For quick and effective computerized voice recognition systems, digital processing of speech signals and voice recognition procedures are essential. The voice is a transmission that contains a vast amount of information. There is too much data within the complex voice signal for simple synthesis and analysis. Consequently, the voice signal is described using methods for digital signal processing including feature extraction and feature matching. As soon as the signal has been pre-processed or filtered, the collection and matching process starts. Mel Frequency Cepstral Coefficients (MFCCs), a non-parametric technique for simulating the human auditory perception system, are utilized as extraction processes. The relationship between instructed and tested signals has been compared using the cross-correlation approach. The gender has been accurately predicted by the model 97% of the time.

KEYWORDS: Gender Prediction, Cross-correlation

I. INTRODUCTION

A voice signal can reveal a speaker's gender, age, accent, and emotional state, among other details. The most noticeable piece of information in the voice signal is the speaker's gender, which is utilized either directly or through inference in numerous applications. The study [1] suggests a novel method for identifying the speaker's gender by using hybrid features, which are produced by stacking various feature types. Besides spoken content, the human voice conveys a multitude of information, as gender, age, and emotional condition. Gender differentiation, or the capability to discriminate between male and female voices, is one of the main problems in speech signal processing. This procedure is based on the identification and examination of particular acoustic characteristics in speech signals that show discernible gender variations. Pitch, formants, energy distribution, and spectral properties are some of the speech aspects that are crucial in expressing the distinctive qualities of male and female voices. Because they are good at simulating the human auditory system, Mel Frequency Cepstral Coefficients (MFCC) and Mel Scaled Power Spectrogram (Mel Spectrogram) are two of these that are frequently employed in speech analysis [2].

MFCC: A linear cosine transmute of a log power spectrum on a nonlinear mel measure of frequency yields the MFCC, which stands for the short-term power spectrum of resonance. The timbral and phonetic characteristics of a speaker's voice, which vary depending on gender owing to physical variations in vocal tract length and tension, are especially well captured by MFCCs.

Mel Spectrogram: With the frequency axis scaled in accordance with the Mel scale, the Mel Spectrogram offers a time-frequency description of the signal. In addition to highlighting the voice's perceptual characteristics, this graphic representation shows how the energy concentration patterns of male and female voices evolve over time.

These characteristics provide the basis for the identification and measurement of gender-specific speech patterns using machine learning (ML) models and signal processing methods like cross-correlation. Researchers can create reliable systems that can accurately distinguish between male and female speech mechanically by utilizing MFCC and Mel Spectrogram. The approximate Mel spectrogram produced from a synthetic audio sample using common tools is displayed in Figure 1.

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203073

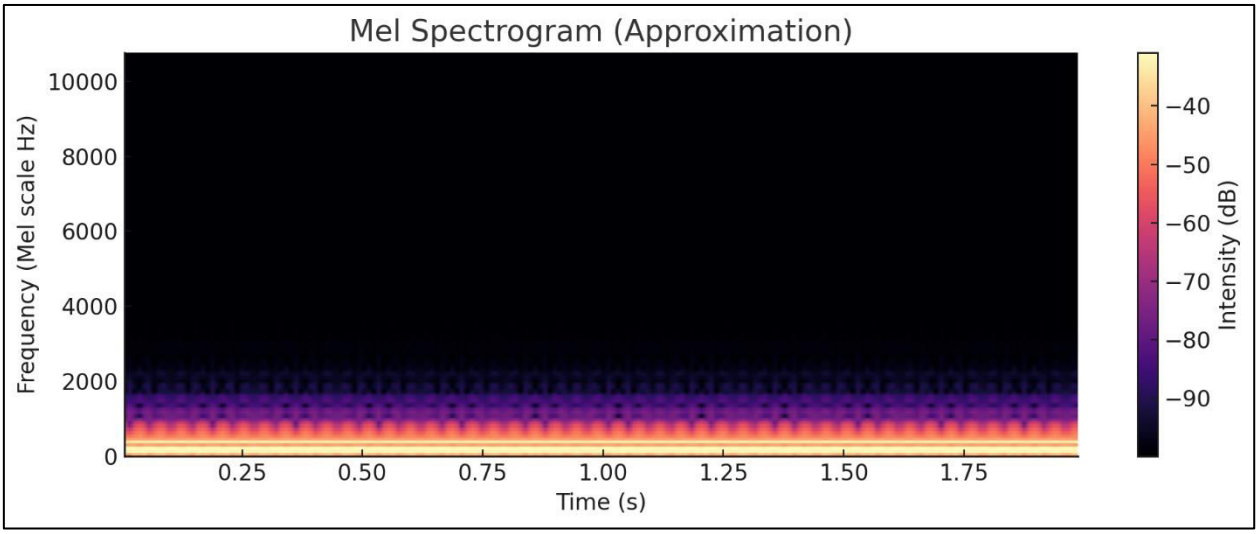

Figure 1: An approximate Mel Spectrogram

Role of MFCC in Speech Identification: The vocal tract that people create when they speak is represented mathematically by MFCCs. In order to document the key elements of human speech that are most perceptible to the human ear, the procedure consists of multiple phases.

Signal Analysis: Speech is a complicated signal with fluctuating amplitude and frequency. These signals are broken down into simpler elements with the use of MFCCs, which show how the properties and pace of sound waves vary over time.

Frequency Transformation: Frequencies are not linearly perceived by humans. Because the human auditory system is more susceptible to variations in lower frequencies than higher ones, the MFCCs employ a mel-scale that closely resembles this response.

Cepstral Illustration: The signal is first transformed to the mel scale and then returned to the cepstrum, a time-domain description. The gradual fluctuation (timbre), which contains the majority of the information necessary for speech recognition, is separated from the periodic variation (pitch) of the signal by the cepstrum.

Role of Mel-Scale in Speech Identification: The Mel-scale was created specially to replicate how people hear sound, especially how we can distinguish between different pitches. shifts in lower frequencies are more perceptible to human hearing than corresponding shifts in higher frequencies.

Linear Region: Our ears are capable of picking up on subtle variations in pitch at lower frequencies (below 1000 Hz). By linearly spacing these frequencies, the Mel-scale reflects this sensitivity, so that an alteration in frequency directly correlates to a similar shift on the scale.

Logarithmic Region: Our ears become less sensitive to frequency fluctuations above 1000 Hz. The Mel-scale becomes logarithmic at this point, clustering frequencies that would sound identical to the human ear closer together. Because of its logarithmic structure, greater adjustments are needed to produce a comparable apparent pitch disparity as frequencies rise.

This dual strategy aids in a number of applications, such as analyzing music and speech processing, where preserving the subtleties of human hearing can greatly improve the precision and efficacy of the technique.

II. RESEARCH BACKGROUND

The usage of computing devices that use more sophisticated techniques, including speech recognition in human–computer communication, has grown in popularity as technical developments have coincided with a rise in human–computer contact [3]. Artificial intelligence analyzes voice signals in systems that use speech recognition technology,

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203073

then uses the information gathered to decide what needs to be done. The method by which humans produce speech is intricate. In addition to linguistic data as vocabulary and grammar, the generated speech signal also includes paralinguistic data like the speaker's age, gender, accent, and emotional state [4]. Communication with computers has begun to use the paralinguistic cues that people frequently use to communicate with one another. One of the most noticeable paralinguistic clues is the speaker's gender, which is frequently used to tailor services. It is difficult to determine the speaker's gender from voice cues, though.

Developing successful promotion and publicizing techniques, improving human-computer interaction, catching convicts, and boosting customer happiness through tailored user services are just a few of the many applications that leverage the gender recognition job [5]. Furthermore, some speech-based recognition systems also employ data on gender to improve their accuracy and speed by creating gender-specific models or restricting the search area to speakers of a particular gender [6]. Voice excellence, variety, feature selection and extraction, classifier model development and evaluation, and other aspects all affect how accurate automated gender recognition systems are [7]. One of the most important processes that straight affects the recognition rate is selecting and retrieving appropriate features from the voice signal [8].

A number of the most often utilized speech features are tonal centroid features (Tonnetz), Chroma, spectral contrast, Mel Spectrogram, and MFCCs. The feature extraction procedure must be customized for every assignment because none of these features capture every facet of speech. Once the speech signals' attributes have been extracted, they are input into a classifier along with class labels that indicate the speaker's gender. The classifier is then trained. The model's efficacy on unseen data is next evaluated using a dataset that includes fresh voice data that wasn't used during the training phase. Numerous techniques have been created to categorize input data that is represented in various feature spaces. Because each strategy takes a different approach to tackling problems, they all have unique benefits and drawbacks [9]. In the past, gender detection was frequently accomplished using conventional machine learning (ML) techniques including (SVM), (KNN), (LDA), (HMM), and (GMM) [10,11]. Because of its success in a number of domains, including computer vision [12], natural language processing [13], and speech recognition [14], deep learning (DL) has emerged as one of the most prominent subfields of ML in recent years.

III. PROPOSED METHODOLOGY

The collected samples were undergone two phases. Training Phase and testing phase. The (MFCC) were used for voice feature abstraction (Figure 1).

a. Steps of MFCC

1. Pre-emphasis

Amplify high-frequency components by using a high-pass filter to the input voice signal.

$$y(n) = x(n) - \alpha . x(n-1) \quad (1)$$

2. Framing

Since Speech signals are non-stationary, but short segments (20-40ms) can be considered stationary , therefore the signal is split into overlapping frames

3. Windowing

In order to minimize spectral leakage during FFT, each frame is multiplied by a window function (e.g., Hamming window):

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (2)$$

Here, N is the number of samples in each frame.

4. Fast Fourier Transform (FFT)

The magnitude spectrum is obtained by applying the FFT to each windowed frame in order to transfer the time-domain signal to the frequency domains.

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203073

5. Mel Filter Bank

In order to simulate human ear perception (more sensitive to lower frequencies), applied a set of triangular filters spaced on the Mel scale to the power spectrum.

$$\text{Mel scale (m)} = \text{a log10}(1 + \frac{\text{f}}{700}) \quad (3)$$

6. Filter Bank Energies

To mimic human loudness perception (logarithmic), the log of each Mel-filtered energy has been carried out.

7. Discrete Cosine Transform (DCT)

To decorrelate the log filter bank energies, apply DCT to the log energies and get the MFCCs.

8. Feature Vector Formation

9. Gender Identification (Classification Stage)

Training:

- Extract MFCC features from labelled male and female samples.
- Train a classifier (e.g., SVM, CNN, k-NN, GMM).

Testing:

- Extract MFCC from an unknown sample.
- Classify using the trained model to predict gender.

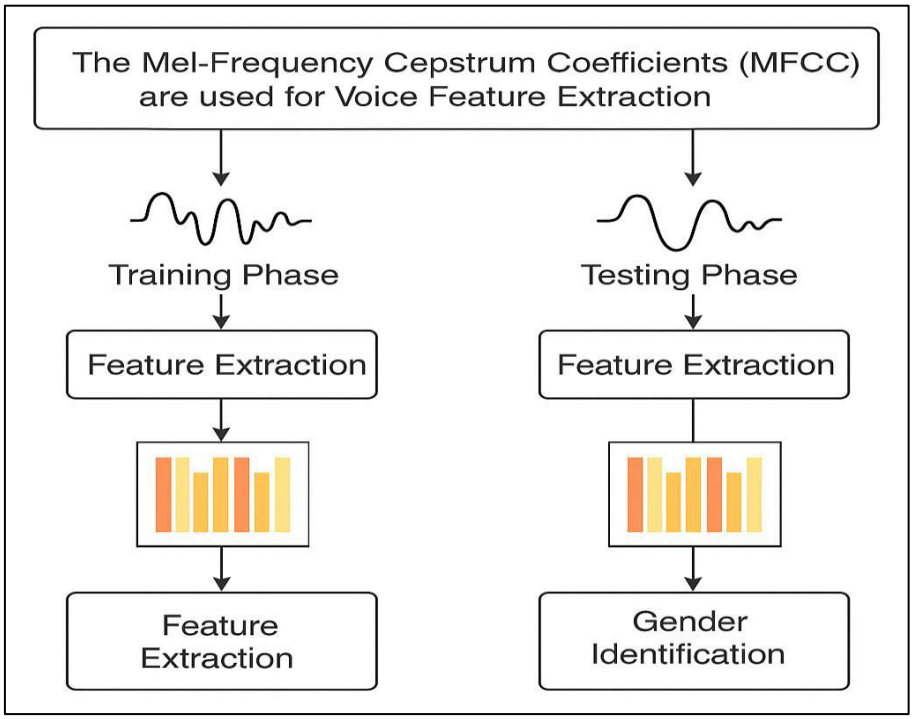

Figure 1: Process of Feature Extraction

b. Using cross-correlation (CC) to match signals

By comparing two signals, cross-correlation finds similarities between them. To do this, one signal is shifted over the other, and the correlation among the two signals is calculated at all locations. The result is a CC function that shows how comparable the two signals are to one another as a function of how far apart they are. In voice recognition, the input signal is the speech signal that has to be recognized, and the reference signal acts as a pattern for the word that is said. Since it reflects the physical character of the signal, the central frequency is one of the most important components

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203073

in voice recognition. Another name for it is the approximated frequency of a spoken speech signal that is quasi-periodic. To ascertain the fundamental frequency in our investigation, we used the CC methodology, a time-domain technique. It looks at the relationship between two speech signal levels and the evolution of their divergence.

IV. RESULTS AND DISCUSSIONS

We have collected the samples of a group of 50 peoples (both male and female). The samples has been used for training purpose. The linear transform of a sound's logarithmic spectrum is used to describe its short-term power spectrum. A cepstral visualization of audio clips utilizing MATLAB's.wav format serves as the foundation for the concept. In order to automatically identify spoken words from audio data, MFCCs are frequently use up in speech recognition systems. Uses for audio data recovery, such as gender classification and auditory resemblance metrics, use MFCCs. In order to lessen the influence of additive noise, its values can be normalized in speech recognition systems. The documented audio signal of a man voice is shown in Figure 2. Other audio signals in the collection are matched with this signal (Figure 3a-c).

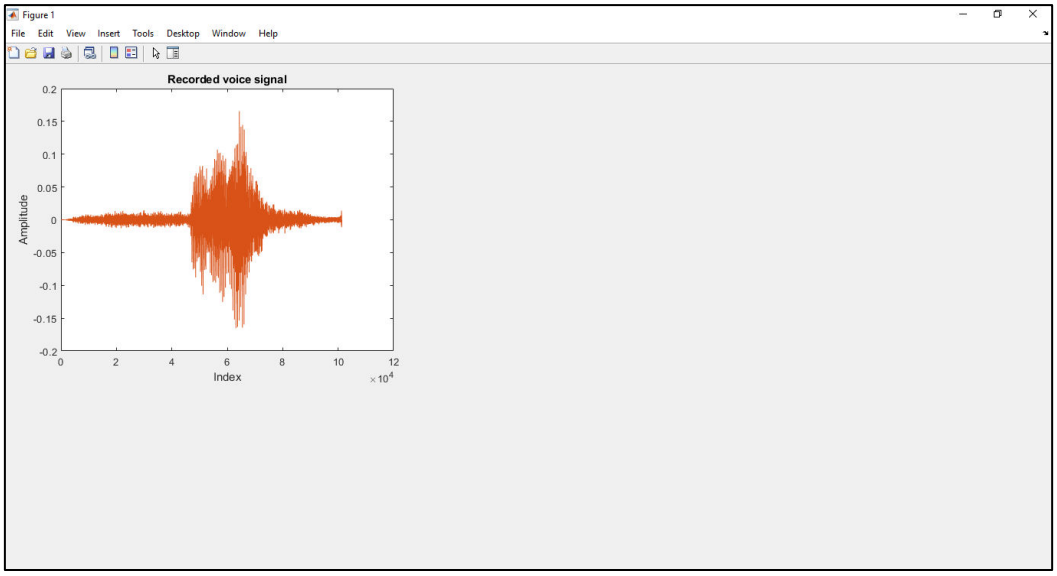

Figure 2: Documented male voice

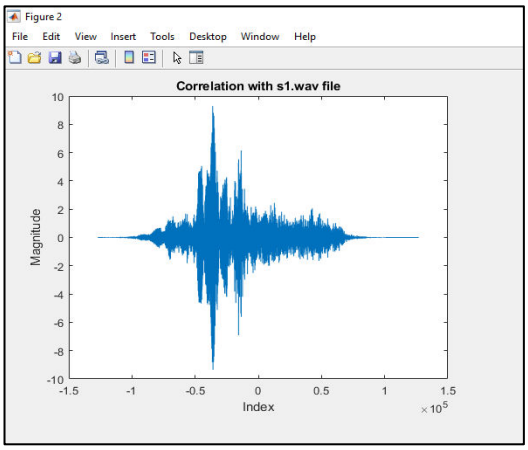

Figure 3a: Matching with first sample

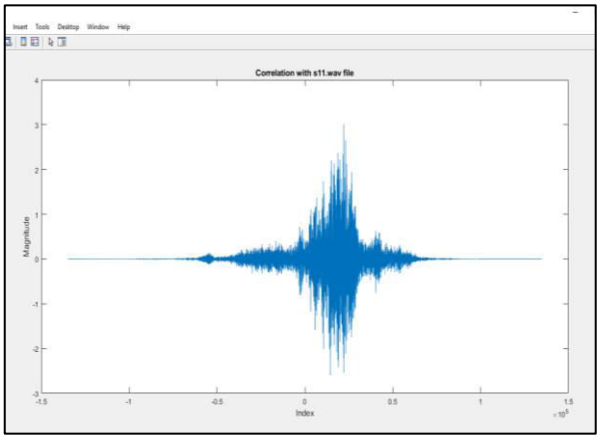

Figure 3b: Matching with second sample

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in | | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203073

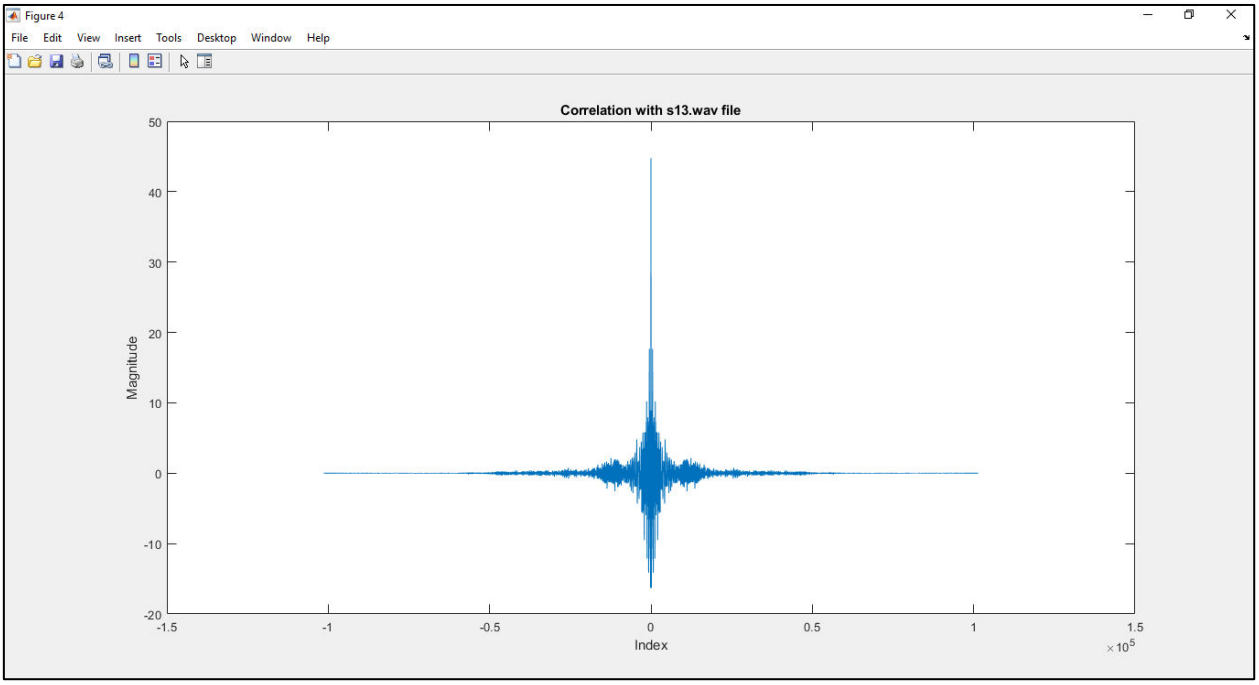

Figure 3c: Matching with third sample

The captured female voice is seen in Figure 4. Other audio signals in the collection have been linked with this signal. (see Figure 5 a-c).

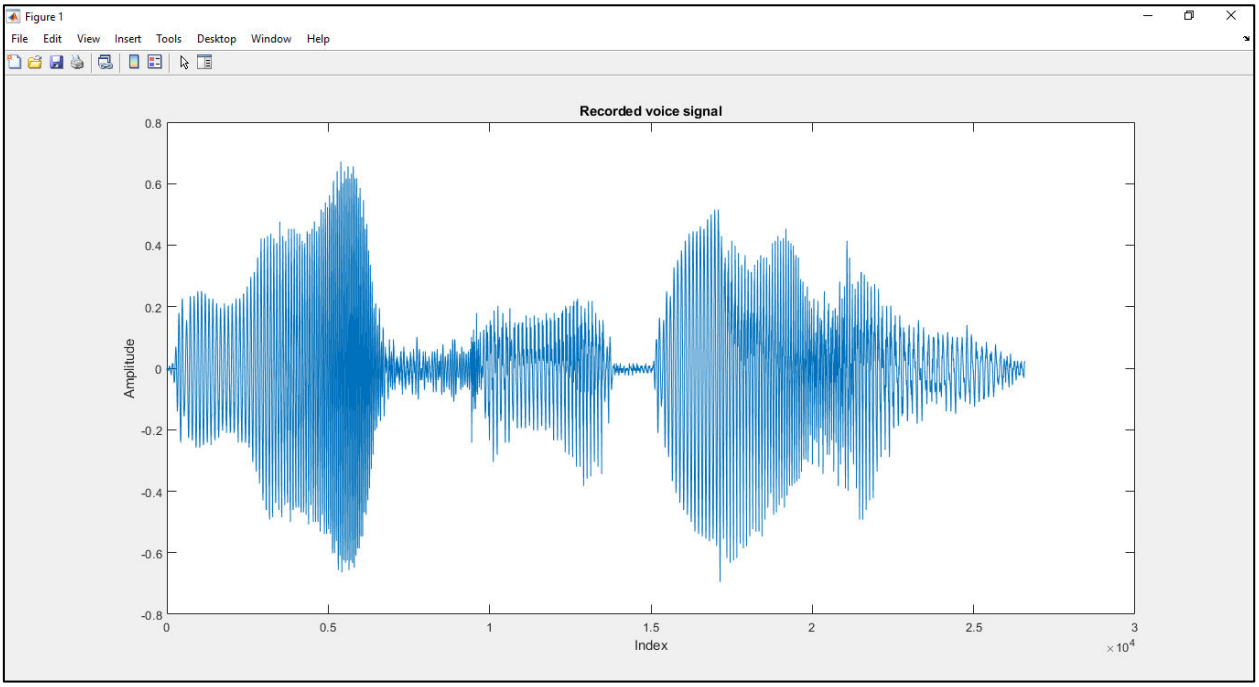

Figure 4: Documented female voice signal

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203073

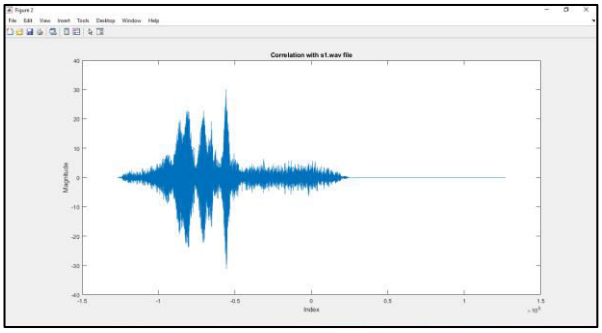

Figure 5a: Match with sample four

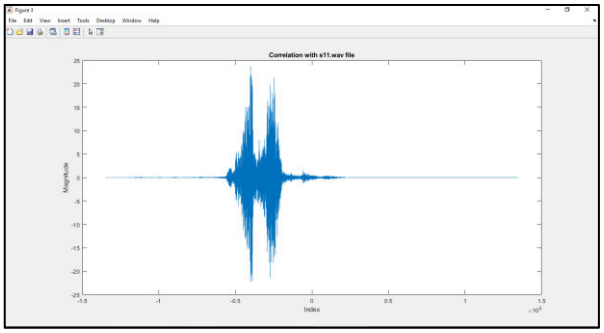

Figure 5b: Match with sample five

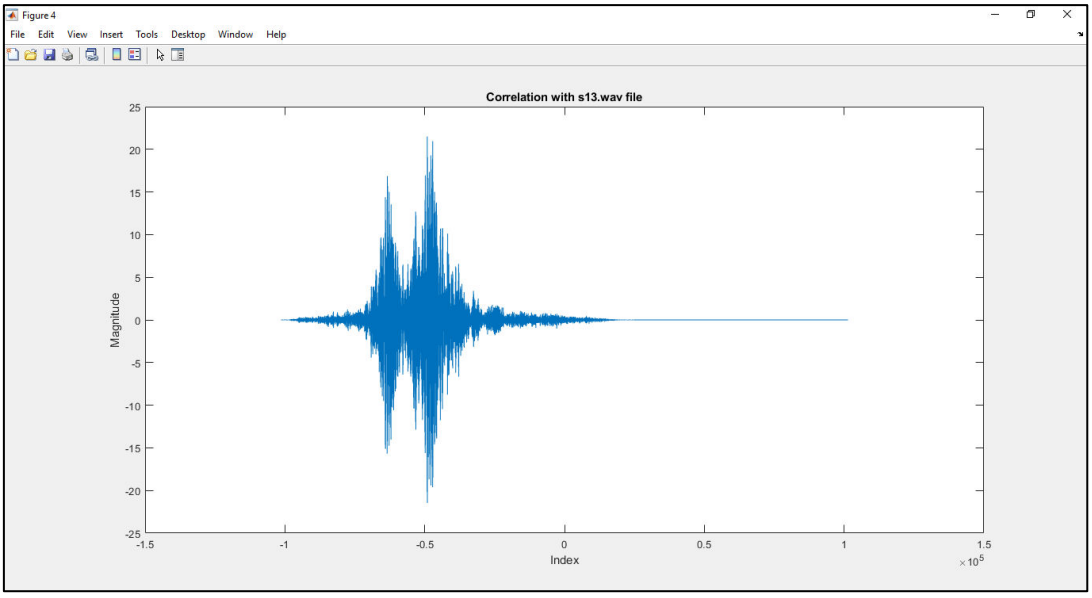

Figure 5c: Match with sample six

A female voice has been acknowledged. The two genders' different sounds were coded and separated using two main algorithms. In order to differentiate between the two users, the input user speech was sampled using the correlation and MFCC algorithms. The sound was then processed to create a single catalogue for each user based on the characteristics of their sampled voice records. There are two steps in the recognition process. In the first step, speech samples are documented using a system stereo and kept as.WAV files, while user input is kept in the appropriate directory of the testing folder. The voice signal is then further processed by MATLAB by accessing these files.

V. CONCLUSION

One important speech recognition technique is cross-correlation, which detects spoken words by comparing the input speech signal to a reference signal. It has several advantages, such as being easy to use, efficient, and noise-resistant. However, there are several disadvantages, such as sensitivity to variations in speaking speed and pitch, the need for a reference signal for every word, and challenges in recognizing speech in specific contexts. Generally speaking, cross-correlation is a successful speech recognition technique that should be combined with other techniques to increase precision and robustness. A voice-based gender recognition algorithm is presented in this work. By examining private details in the speaker's voice stream, the method correctly identified the speaker. Separating noise from information signals and using them to carry out practical tasks is the main objective of this effort. In order to distinguish between male and female voices, we employed the MFCC approach in our work.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203073

REFERENCES

1. Yücesoy, E. Gender Recognition Based on the Stacking of Different Acoustic Features. Appl. Sci. 2024, 14, 6564.
2. https://www.geeksforgeeks.org/mel-frequency-cepstral-coefficients-mfcc-for-speech-recognition/
3. Gondohanindijo, J.; Noersasongko, E. Multi-Features Audio Extraction for Speech Emotion Recognition Based on Deep Learning. Int. J. Adv. Comput. Sci. Appl. 2023, 14, 198–206.
4. Safavi, S.; Russell, M.; Jančovič, P. Automatic speaker, age-group and gender identification from children's speech. Comput. Speech Lang. 2018, 50, 141–156.
5. Alkhawaldeh, R.S. DGR: Gender recognition of human speech using one-dimensional conventional neural network. Sci. Program. 2019, 2019, 7213717.
6. Tursunov, A.; Khan, M.; Choeh, J.Y.; Kwon, S. Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. Sensors 2021, 21, 5892.
7. Rezapour Mashhadi, M.M.; Osei-Bonsu, K. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. PLoS ONE 2023, 18, e0291500.
8. Jiang, S.; Chen, Z. Application of dynamic time warping optimization algorithm in speech recognition of machine translation. Heliyon 2023, 9, e21625.
9. Reda, M.M.; Nassef, M.; Salah, A. Factors affecting classification algorithms recommendation: A survey. In Proceedings of the 8th International Conference on Soft Computing, Artificial Intelligence and Applications, Dubai, United Arab Emirates, 29–30 June 2019.
10. Tian, Q.; Arbel, T.; Clark, J.J. Deep LDA-runed nets for efficient facial gender classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 10–19.
11. Singhal, A.; Sharma, D.K. Estimation of Accuracy in Human Gender Identification and Recall Values Based on Voice Signals Using Different Classifiers. J. Eng. 2022, 2022, 9291099.
12. Chai, J.; Zeng, H.; Li, A.; Ngai, E.W. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Mach. Learn. Appl. 2021, 6, 100134.
13. Lauriola, I.; Lavelli, A.; Aiolli, F. An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing 2022, 470, 443–456.
14. Kwon, H.; Yoon, H.; Park, K.W. Acoustic-decoy: Detection of adversarial examples through audio modification on speech recognition system. Neurocomputing 2020, 417, 357–370.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com