Voice Assistant: User Authentication through Voice Biometrics

Kanaganandini Kanagaraja b, Keerthana Muraliac, Vaan Amuthu Elangoad *Department of Computer Science   
 Amrita School of Engineering,* Bengaluru, India a.  
[BL.EN.U4AIE21064@bl.students.amrita.edu](mailto:BL.EN.U4AIE21064@bl.students.amrita.edu) b, [BL.EN.U4AIE21067@bl.students.amrita.edu](mailto:BL.EN.U4AIE21067@bl.students.amrita.edu)c, [e\_vaanamuthu@blr.amrita.edu](mailto:e_vaanamuthu@blr.amrita.edu)d

***Abstract***: Voice biometrics plays an important role in revolutionizing user authentication and personalized interaction within various domains, including healthcare which we will be focusing on in this project. In this project, we propose voice biometrics to enable user authentication and a tailored medical assistant voice-based medical assistant experience. Our proposed system combines voice biometric authentication techniques with conversational interfaces to deliver personalized healthcare services such as scheduling, medication reminders and symptom analysis. The assistant ensures user confidentiality, enhances security, and delivers customized experiences based on individual user profiles and medical history by utilizing voice biometrics. By integrating with existing medical systems, and complying with relevant privacy regulations, we aim to implement a model that encompasses the design of a robust backend infrastructure for securely managing voice biometric templates.

1. *Introduction*

Speech has been playing a significant role in human communication, with approximately 7,000 languages spoken worldwide. To enable computers to understand and respond to human language, the field of speech processing has emerged. [1] The brain, cerebellum to be more specific, supports a variety of functions of speech processing like acceleration of orofacial gestures, timing and coordination of articulatory sequences, and control of brainstem reflexes. This field involves the analysis and utilization of human language for various applications, including medical support and voice assistance. [2] Immense research has been going on in developing speech assessment systems. We need speech assessment systems in order to evaluate speech quality, intelligibility and other parameters. We require these systems to assess the performance of speech processing algorithms and technologies. Studies are going on to meet the demands of modern communication technologies.

[3] A non-linear dynamical system can be used for speech assessment. This approach trains a non-linear predictor using speech samples to extract an effective dynamic system which then analyses the cast signal characteristics. A similar dynamical predictor may be used in time warping of speech, which acts on acoustic features. [4] A fundamental component of speech-based assessment systems, is related to automatic speech recognition technology. Advancements in ASR technology can directly benefit speech-based assessment systems because they improve their ability to accurately transcribe and analyse spoken language.

Several advancements are underway in the realm of speech processing and recognition. Augustine Gray and teams [1 5] have focused on distance measures in speech processing, utilizing cepstral measures for efficient estimation of spectral distances without FFT or DFT operations. Additionally, Tran Duc Chung's team [2 6] has developed an FPT.AI-based text

to-speech system using Django for Python, enabling users to convert Vietnamese text to speech with customizable options.

Vadaboyina Appalaraju and team [3 7] have introduced a Python-based intelligent voice personal assistant, integrating various Python libraries for diverse functionalities such as search, music playback, and messaging. It aimed to create a proficient voice assistant while referencing related research on voice assistants and AI technologies.

Furthermore, Eui sun Kim and team [4 8] have addressed stroke's impact on neurogenic bladder symptoms, proposing a platform for early detection and prevention by analyzing patients' voices. Their AI-based system analyzes voice recordings for unique features, triggering alerts for abnormalities via a mobile app. Testing demonstrated high accuracy in identifying neurogenic bladder risk. This goes to show how diverse the applications of speech processing range and how the advancements in the field pave milestones for development.

Additionally, Benjamin and team [5 9] investigated the effects of signal-to-noise ratio, speech level, and compression ratio on consonant recognition in noise. Their study revealed decreased recognition with higher compression ratios and presentation levels, with notable interaction effects observed. Higher compression ratios had less impact on speech intelligibility at higher presentation levels akin to hearing loss scenarios. [10]

We live in an era where technology contains the power to reshape the landscape of healthcare, voice biometrics emerges as a promising avenue for augmenting security and personalization within medical interactions. Traditional authentication methods, such as passwords or PINs, often provide inadequate safeguarding of sensitive medical information. This introduction sets the stage for exploring the integration of voice biometrics into a medical-based voice assistant, which we term "VoiceGuard Med." Through this project, we aim to demonstrate how voice biometrics can enhance the security and personalization of healthcare services, ultimately improving patient outcomes and satisfaction.

Voice Guard Med stands at the intersection of advanced speech processing technologies and the healthcare domain, offering a secure and intuitive platform for users to interact with medical services. By harnessing the power of voice biometrics, Voice Guard Med ensures that only authorized individuals can access sensitive medical information, thereby mitigating the risks associated with unauthorized access or data breaches.

Furthermore, we discuss the practical implications of deploying Voice Guard Med in real-world healthcare settings, addressing concerns related to security, privacy, and regulatory compliance. Through rigorous testing and user feedback, we strive to validate the effectiveness and usability of voice biometrics in enhancing healthcare security and personalization.  
Overall, Voice Guard Med exemplifies the potential of voice biometrics to revolutionize healthcare delivery, paving the way for a future where patients can access personalized medical assistance securely and seamlessly through voice interaction.

II. literature survey

[11] Shari Trewin and teams analysed the efficiency and user experience of three biometric authentication methods - voice, face, and gesture - alongside traditional password entry on mobile devices. Their study revealed that both face and voice recognition were quicker than password input. Speaking a PIN emerged as the swiftest biometric entry method, although face verification showed superior short-term memory retention. However, none of the authentication methods were deemed highly user-friendly. Combining two biometric methods reduced sample acquisition time but was unpopular and led to higher error rates in memory tasks. These results underscore the cognitive and motor distinctions among biometric modalities, aiding policymakers in selecting appropriate authentication approaches.

[12] R. Parkavi and teams work enhance authentication in biometric systems through multilevel authentication. Multimodal biometrics involves utilizing multiple biometric indicators for individual identification, providing higher authentication levels than unimodal biometrics. Our approach integrates fingerprint and iris data to automatically identify individuals at the matching-score level, employing techniques such as Minutiae matching and Edge detection. We evaluate the proposed method's performance, achieving increased accuracy by reducing the False Acceptance Rate (FAR) and False Rejection Rate (FRR).

[13] The authors have presented a human voice recognition system that integrates the Relative Spectra Algorithm and Decimated Wavelet (DW) with linear predictive coding. The feature vector with wavelet and linear prediction coefficients by extracting features from voice data. To identify the speech and speaker, a comparison is performed by computing the Euclidean distance between the feature vectors. The tested speech or speaker is considered to be a match with the training speech or speaker if the distance is almost zero. The suggested system performs better than LPC; it can get about 90% accuracy rate when tested with fifty preloaded voice signals from six different people

[14] The authors have worked on a research that develops a verification technique that can authenticate voice patterns using MATLAB(SIMULINK) function blocks. For each match in the voice sample, the system generates a logic '1' and for each mismatch a logic '0'. The system adapts it's security level to account for a person's voice changes during identification. It aims to provide medium security access control

[15] The study suggests employing the Relative Spectra Algorithm in conjunction with Linear Predictive coding and decimated wavelet (DW) for a human speech recognition system. The authors have used various methods to test and train speech signals in order to create feature vectors. To identify the speech and speaker, a comparison is performed by computing the Euclidean distance between the feature vectors. Their proposed technique performs better than the LPC scheme; in tests with fifty preloaded speech signals from six people, it achieved an accuracy rate of about 90%.

[16 ] The authors have explored the applications of AI and voice assistants in healthcare. The use of AI for disease diagnosis, treatment recommendations, medical document classification, and question answering. Voice assistants are utilized for providing critical information, scheduling appointments, and hands-free interaction with smartphones. Personalized medicine was highlighted for predicting diagnostic decisions based on individual patient data. The role of NLP and chatbots in understanding text and voice data in healthcare was emphasized.

[17] In this paper, the authors have discussed the readiness of voice assistants like Amazon Alexa, Google Assistant to help healthcare delivery during health emergency and pandemic, specifically focusing on the covid-19 outbreak. They have explored the use of intelligent conversational agents and virtual assistants to combine health service capacity, screen symptoms, deliver healthcare and virtual assistants to augment health service capacity, screen symptoms, deliver healthcare information, and reduce exposure. The authors examine the state of voice assistants as an emerging tool for remote healthcare delivery and assess the readiness of health systems and technology providers to use voice assistants for healthcare delivery during emergencies.

[18] The next five years' prospects for Voice-Controlled Intelligent Personal Assistants (VIPAs) in the healthcare industry are covered by the authors. In order to predict the growth and application of VIPAs in healthcare, the study employed the Delphi technique, which involves specialists in both information technology and healthcare. The following four theme areas were looked into: technology, consumer acceptability, possible applications, and privacy laws. Through desk research, expert panel evaluation, and consensus-building, projections had been developed. Advising healthcare institutions on the use and necessity of VIPAs is the goal of the study going forward.

III. Methodology

In this section, the authors performed multiple experiments using their own voice recordings, articulating the phrase "AI in speech processing." These experiments included adjusting the Signal-to-Noise Ratio, computing the first derivative of the signal, and identifying zero crossings. Key Python libraries, such as librosa for audio analysis and Matplotlib for visualization, were essential in these analyses.

The initial experiment involved modifying the SNR of the recorded speech signal. SNR, vital in signal processing, indicates the ratio of signal power to background noise. By manipulating SNR, researchers simulated different environmental conditions or transmission scenarios, evaluating speech processing algorithms' performance under varying noise levels. Librosa's versatility facilitated SNR manipulation and provided a robust framework for extracting audio features, crucial for understanding noise's impact on speech signals.

The subsequent experiment focused on computing the first derivative of the speech signal. The first derivative reflects the signal's rate of change over time, often used to analyse temporal dynamics in signal processing. By computing this derivative, researchers gained insights into speech's temporal characteristics, such as speech articulation rate or vocal pitch modulation. This analysis relied on mathematical concepts and numerical methods, with Python's NumPy library offering essential signal processing tools.

The final experiment aimed to detect zero crossings within the speech signal. Zero crossings mark points where the signal changes its sign [fig3], crucial in various applications like frequency measurement and signal segmentation. Identifying these points provided valuable information about the signal's waveform and temporal structure. This analysis underscored zero crossings' practical significance in signal processing tasks and emphasized accurate signal characterization's importance for speech processing algorithms.

Throughout these experiments, Matplotlib played a crucial role in visualizing results. This widely-used Python library offered a versatile platform for creating visual representations of data, aiding in understanding the effects of manipulating SNR values, computing derivatives, and identifying zero crossings in speech signals. These visualizations facilitated result interpretation, providing insights into speech signal dynamics under different experimental conditions.

Additionally, the authors performed silent detection on speech signal, silent detection helps in identifying period of silence within an audio signal. It is vital for segmenting audio, enhancing speech recognition accuracy, optimizing compression. Silence detection is done by using trim () and split() function, trim function is used to remove silence or to remove unwanted parts. In this project the signal energy is calculated and if the energy is less than threshold then it is considered as silence

*Fast Fourier Transform*

FFT is an algorithm used to computed DFT (discrete Fourier transform) of a signal. Fourier transform is used to convert time domain signal to frequency domain, DFT operates on discrete time signal, discrete time signals are easy to processes and analysis the information. This continuous time signal is converted into discrete time signal using sample method. Inverse Fourier Transform is the mathematical operation that allows us to reconstruct a signal from its frequency domain representation to time domain representation.

Spectrogram is used to represent frequency content of a signal as it varies with time. Its helps in analyse and visualize how the frequency components of signal change over time. A frequency heat map visually represents data using colours to indicate the frequency of occurrence across different categories or values.

*Filter*

Rectangle filter applied to speech signals for equalization or spectral analysis by isolating specific frequency bands with rectangular-shaped frequency response. Cosine filters in speech signal processing isolate specific frequency components, aiding in tasks like noise reduction and feature extraction. Gaussian filter smooths speech signals, reducing noise while preserving important features, enhancing clarity and intelligibility in communication.

LAB 06

Acoustic analysis reflected the phonological analysis process but was focused on phonation. FFT analysis was performed to understand the amplitude spectra of signal snippets representing different pitches, thus helping to identify their spectral features.The analysis was extended to silent and silent analysis steps that include silence and segmentation of non-voiced parts of the recorded speech signal In order to search for background noise or parts without speech content spectral properties, FFT was performed on these sections.

During the Spectrogram Generation and Analysis phase, a spectrogram of the entire signal was generated. The frequency of the signal with time was plotted visually in the spectrograms. By observing the change point in the spectrogram associated with speech features, insights from the previous study were able to be used to identify non-speech syllable and phonological features based on their spectral characteristics

Overall, this approach combined FFT analysis with spectrogram imaging to distinguish between vocal and non-speech vocal components in recorded speech signalsThis section explains the output of each section in the methodology.

***LAB07 HMM for classification of your speech signal using STFT features.***  
this section deals with the application of Hidden Markov Models (HMM) with Short-Time Fourier Transform (STFT) features for speech signal classification. We collect a diverse dataset, extract STFT features, and train HMMs to classify speech signals. Results show promising classification accuracy, demonstrating the potential for real-world applications such as speech recognition and speaker identification. Further experimentation and optimization are suggested to enhance model performance. Overall, this methodology offers a robust approach to analysing speech signals, contributing to advancements in speech processing and analysis.

LAB08

Lab discusses the application of LSTM and Bi-LSTM networks in speech recognition, along with feature extraction techniques like STFT, MFCC, and LPC coefficients. These networks excel at capturing temporal dependencies, making them suitable for processing sequential data like speech signals. Feature extraction transforms raw audio extracts important feature that is fed into neural network.

text-to-speech conversion, involves generating human-like speech from text. 2nd problem in the lab sheet tells to synthesizes the word "Bharat" in various Indian scripts by combining phonemes extracted from the speech recording of the sentence "Bhanumathi weds Rajat". Challenges include the accuracy of phoneme extraction, capturing prosody, and script/language support.

LSTM and Bi-LSTM networks, combined with feature extraction techniques, offer promising avenues for improving speech recognition accuracy. However, speech synthesis remains complex, with nuances in pronunciation, prosody, and script support posing challenges. Advancements in deep learning and TTS technologies drive progress in overcoming these challenges, aiming to produce more natural and accurate synthesized speech. Nonetheless, careful consideration and validation of Indian scripts are necessary for accurate speech synthesis applications.

IV. Results

This section explains the output of each section in the methodology. Fig1 represents the original audio it’s is plotted using amplitude vs time plot

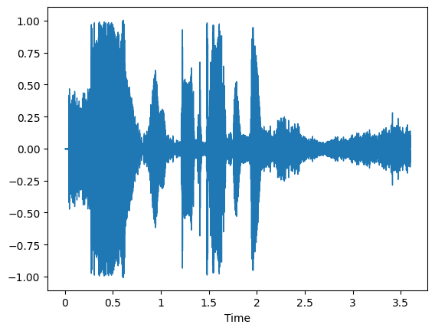


Fig 1. Original signal

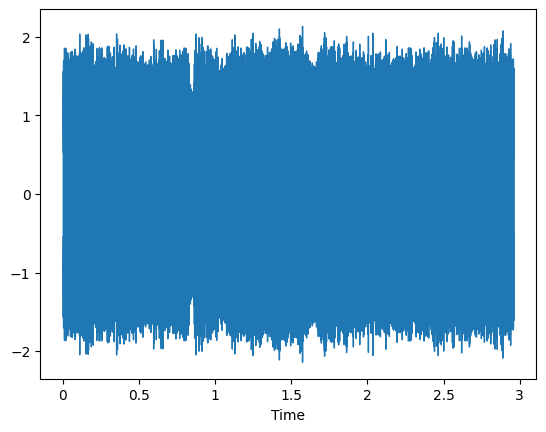


Fig 2. First derivative of a signal

Fig 1-2 is the representation of original signal vs first derivative of audio

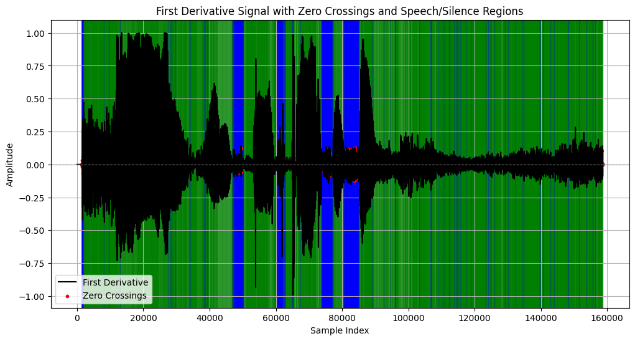


Fig 3. Zero crossing

In fig 3 the red dots represent the point where signal changes its sign

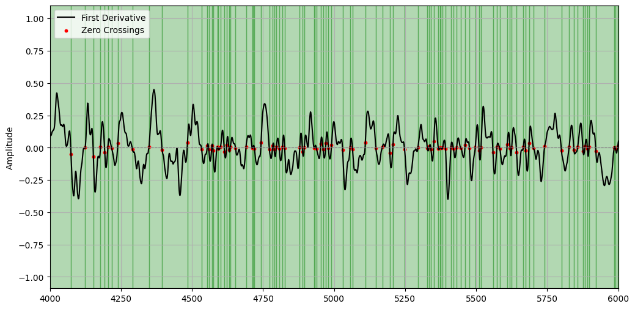


Fig5. Zero crossing and Frist derivative

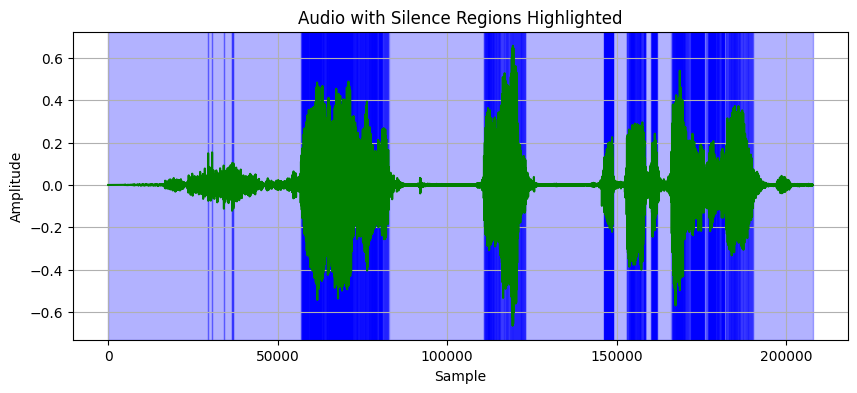


Fig 6 Represent the silence region

Fig6 shows the highlighted silence region with the help of this Fig we can identify the unwanted part and remove this will help to compress size, remove noise.

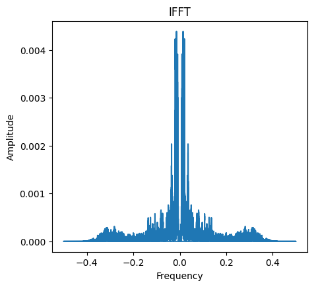
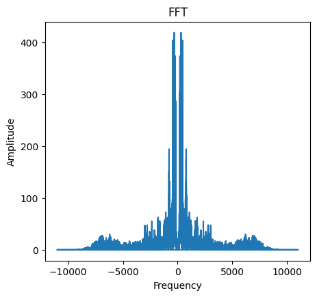


Fig 7. FFT and IFFT of speech signal

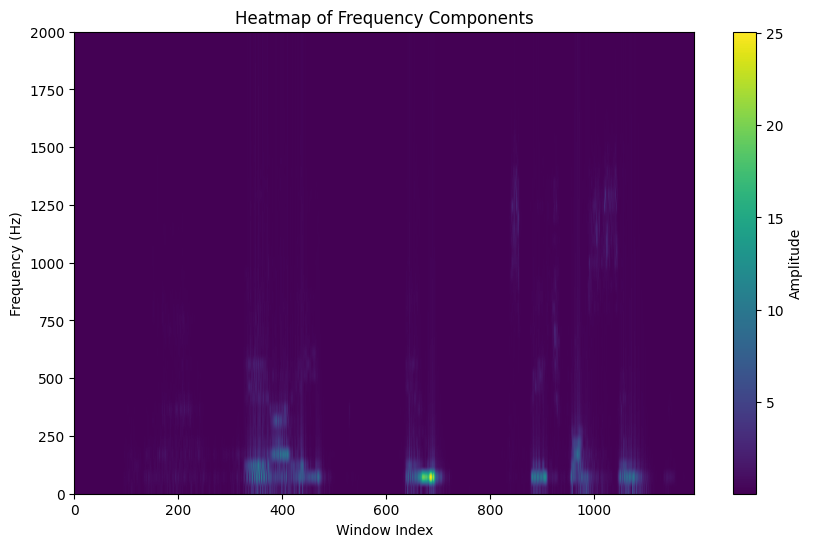


Fig8 Heatmap of Frequency components

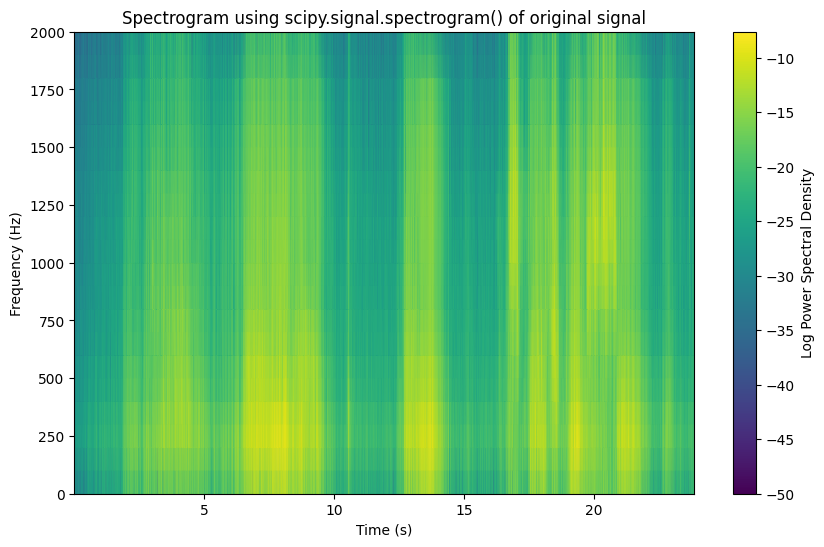


Fig 9 Spectrogram of original signal

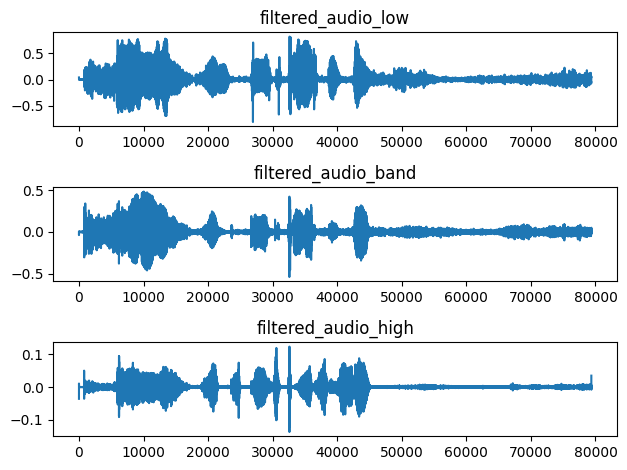


Fig 10 visual representation of speech signal after applying rectangle filter

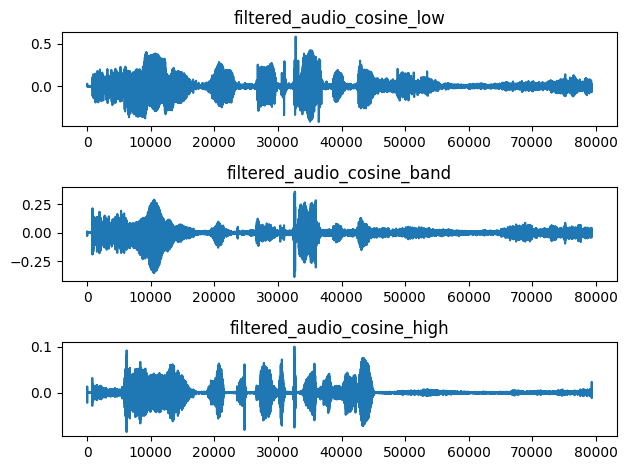


Fig 11 visual representation of speech signal after applying cosine filter

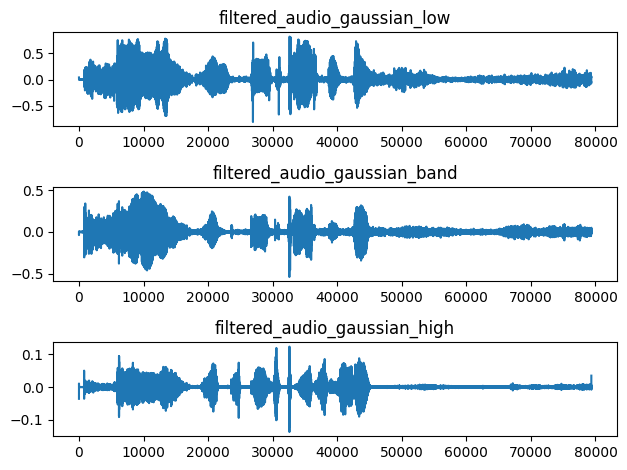


Fig 12 visual representation of speech signal after applying rectangle filter

In the above figure (10,11,12)The Rectangle Filter excels at isolating specific frequency bands, making it particularly useful for tasks like spectral analysis. The Cosine Filter is well-suited for frequency domain filtering, characterized by a flat passband and sharp transitions. The Gaussian Filter is highly effective in smoothing and reducing noise in images, with a notable ability to preserve edges compared to alternative filters. For this project the gaussian filter is more suitable.

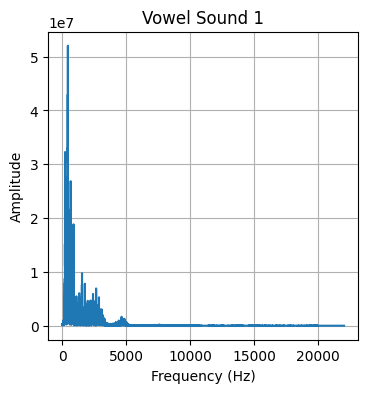
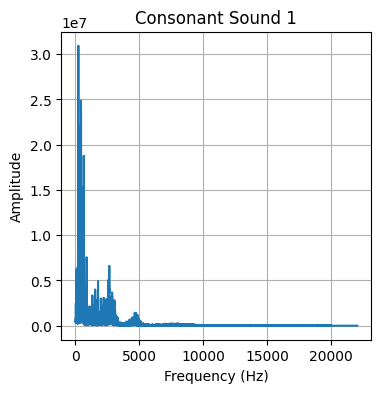
 

Fig13 representation of vowel and consonants

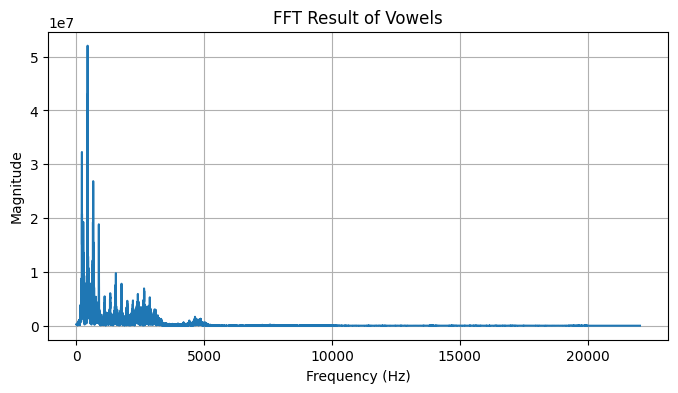


Fig 14 visual representation vowels after FFT applied

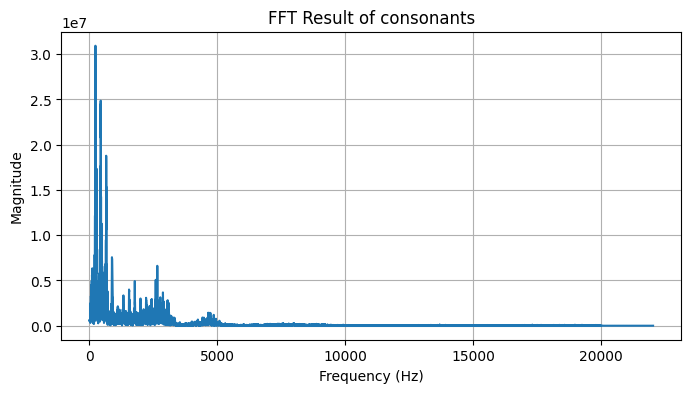


Fig 15 visual representation vowels after FFT applied

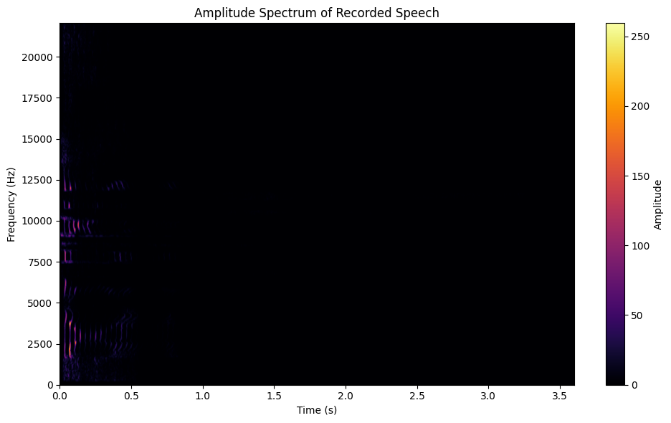
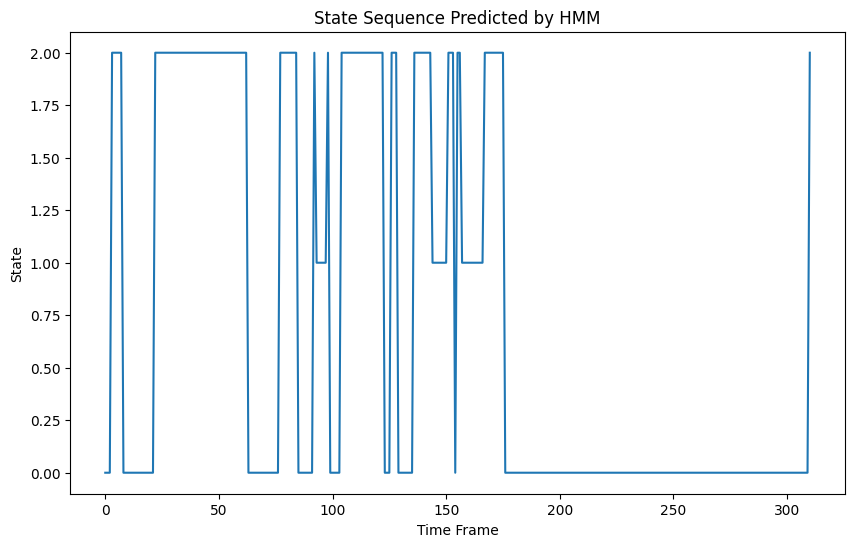


Fig 16 Amplitude spectrum of recorded speech

Fig 17

Using STFT features, the amplitude spectrum of the recorded speech signal is shown fig16, depicting temporal and spectral fluctuations. The HMM model, integrating Gaussian emissions and STFT features, undergoes training with 3 components over 100 cycles. Visualizing the state sequence (Fig 17) reveals segmentation of the speech signal into different states over time.

To sum up, these experiments showcased the varied applications of signal processing techniques in speech signal analysis. Leveraging Python libraries like librosa and Matplotlib, researchers manipulated SNR, computed derivatives, and identified zero crossings in recorded speech signals, silence detection and fast Fourier transform is applied, for advancing understanding of speech processing algorithms and their performance in diverse condition.

Lab 08 results

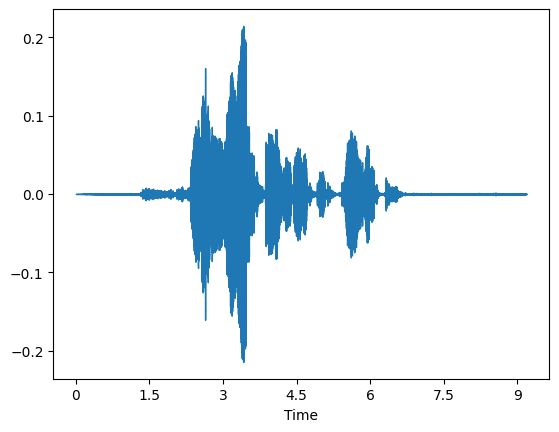


Fig 18 wave form of speech signal saying “Bhanumati weds rajat”

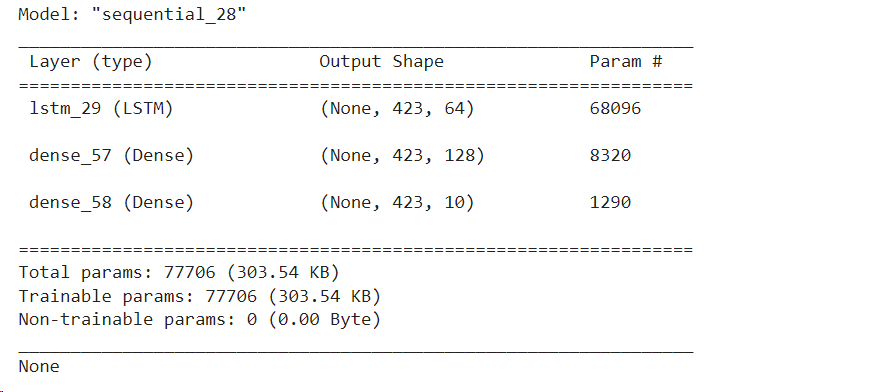


Fig speech recogonization using LSTM networks and STFT for feature extraction.

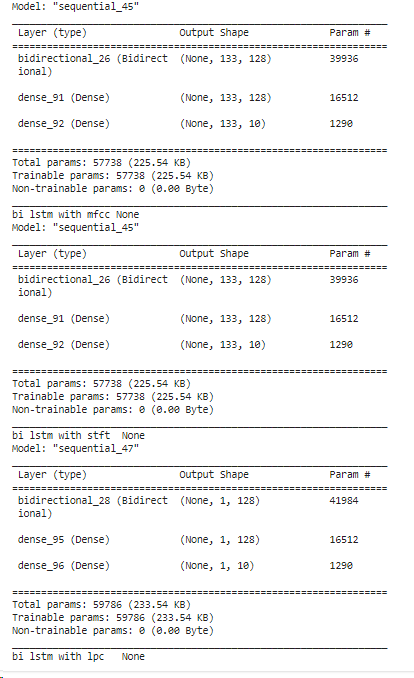


Fig 19 Speech recogonoze training using Bi lstm and STFT, MFCC, LPC for feature extraction

REFRENCES

[1] H Ackermann, I Hertrich, ”The contribution of the cerebellum to speech processing”, Journal of Neurolinguistics, Volume 13, Issues 2–3, 2000, Pages 95-116, ISSN 0911-6044

[2] B. H. Juang and Tsuhan Chen, "The past, present, and future of speech processing," in IEEE Signal Processing Magazine, vol. 15, no. 3, pp. 24-48, May 1998, doi: 10.1109/79.671130.

[3] N. Tishby, "A dynamical systems approach to speech processing," International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 1990, pp. 365-368 vol.1, doi: 10.1109/ICASSP.1990.115686.

[4] D. Yu, G. Hinton, N. Morgan, J. -T. Chien and S. Sagayama, "Introduction to the Special Section on Deep Learning for Speech and Language Processing," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 4-6, Jan. 2012, doi: 10.1109/TASL.2011.2173371.

[5] A. Gray and J. Markel, "Distance measures for speech processing," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 5, pp. 380-391, October 1976, doi: 10.1109/TASSP.1976.1162849

[6] T. D. Chung, M. Drieberg, M. F. Bin Hassan and A. Khalyasmaa, "End-to-end Conversion Speed Analysis of an FPT.AI-based Text-to-Speech Application," 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech), Kyoto, Japan, 2020, pp. 136-139, doi: 10.1109/LifeTech48969.2020.1570620448. keywords: {FPT.AI;TTS;performance;analysis;Vietnamese;voice},

[7] V. Appalaraju, V. Rajesh, K. Saikumar, P. Sabitha and K. R. Kiran, "Design and Development of Intelligent Voice Personal Assistant using Python," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 1650-1654, doi: 10.1109/ICAC3N53548.2021.9725753

[8] Kim ES, Shin DJ, Cho ST, Chung KJ. Artificial Intelligence-Based Speech Analysis System for Medical Support. Int Neurourol J. 2023 Jun;27(2):99-105. doi: 10.5213/inj.2346136.068. Epub 2023 Jun 30. PMID: 37401020; PMCID: PMC10325939.

[9] Benjamin W. Y. Hornsby, Todd A. Ricketts; The effects of compression ratio, signal-to-noise ratio, and level on speech recognition in normal-hearing listeners. J. Acoust. Soc. Am. 1 June 2001; 109 (6): 2964–2973. <https://doi.org/10.1121/1.1369105>

[10] O. Abdel-Hamid, A. -r. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, "Convolutional Neural Networks for Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, Oct. 2014, doi: 10.1109/TASLP.2014.2339736.

[11] Shari Trewin, Cal Swart, Larry Koved, Jacquelyn Martino, Kapil Singh, and Shay Ben-David. 2012. Biometric authentication on a mobile device: a study of user effort, error and task disruption. In Proceedings of the 28th Annual Computer Security Applications Conference (ACSAC '12). Association for Computing Machinery, New York, NY, USA, 159–168. https://doi.org/10.1145/2420950.2420976

[12] R. Parkavi, K. R. Chandeesh Babu and J. A. Kumar, "Multimodal Biometrics for user authentication," 2017 11th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, India, 2017, pp. 501-505, doi: 10.1109/ISCO.2017.7856044.

[13]. Kumar, Pala Mahesh. "A new human voice recognition system." *Asian Journal of Science and Applied Technology* 5.2 (2016): 23-30.

[14] R. A. Rashid, N. H. Mahalin, M. A. Sarijari and A. A. Abdul Aziz, "Security system using biometric technology: Design and implementation of Voice Recognition System (VRS)," 2008 International Conference on Computer and Communication Engineering, Kuala Lumpur, Malaysia, 2008, pp. 898-902, doi: 10.1109/ICCCE.2008.4580735.

[15] A. Kounoudes, V. Kekatos and S. Mavromoustakos, "Voice Biometric Authentication for Enhancing Internet Service Security," 2006 2nd International Conference on Information & Communication Technologies, Damascus, Syria, 2006, pp. 1020-1025, doi: 10.1109/ICTTA.2006.1684514.

[16] Ahanin, Elaheh, A. Bakar Sade, and Huam Hon Tat. "Applications of Artificial Intelligence and Voice Assistant in Healthcare." *Int. J. Acad. Res. Bus. Soc. Sci* 12 (2022): 2545-2554.

[17] Sezgin, E., Huang, Y., Ramtekkar, U. *et al.* Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. *npj Digit. Med.* **3**, 122 (2020). https://doi.org/10.1038/s41746-020-00332-0

[18] Ermolina, Alena, and Victor Tiberius. "Voice-controlled intelligent personal assistants in health care: International Delphi Study." *Journal of Medical Internet Research* 23.4 (2021): e25312.