# Localized Speech-to-Text for Inclusive Learning

Pranav H
Dept. of Computer Science
*Amrita School of Computing*
*Amrita Vishwa Vidhyapeetham,*
*Bengaluru, India*

Mayank Pandey
Dept. of Computer Science
*Amrita School of Computing*
*Amrita Vishwa Vidhyapeetham, Bengaluru, India*

## I. Introduction

The most natural way for communication as humans is speech, and there have been constant efforts to make speech a viable and effect way of communication with computing devices, this is achieved with the ever-evolving toolset known as speech to text. These tools constantly evolving as they are , have failed to account for variations in dialects [3] and pronunciation of words along with being available for regional and vernacular languages which varies the impact and effectiveness of these tools from region to region [4]. In this context, it is empirical that we strive to improve these tools to overcome this disparity, making them available to all irrespective of region or language.

In light of the pressing need for these technological tools to be not just widely accessible, but also effectively utilized across diverse regions and languages, this paper puts forth a novel proposition. It suggests the deployment of an on-device, self-supervised [5], speech-to-text module that has been localized specifically for the multitude of regional languages spoken across India. The primary objective of this module is not merely to transcribe speech to text, but to serve a greater purpose - to act as a facilitator in the learning process. By catering to the unique linguistic nuances of regional Indian languages[6,7], this module aims to bridge the gap between technology and effective learning, thereby making education more inclusive and comprehensive.

Here we strive to implement the following tools to facilitate our goal towards a smart and inclusive classroom :

- Discussion Logs : Leveraging Audio Fingerprinting and text independent speaker recognition systems [8] to keep an individualised, speaker separated log of the class room discussions, which will enable a quick and easy review of the classroom discussions.
- Summarization : Using the discussion logs made, prepare a short summary to assist in quick review.
- Individualised Remainders and Summary : Leveraging the discussion logs to make a personalised summary tailored to the individual comprising of individual assigned tasks and conversations.
- Live Captioning and Translation : Enable Seamless Real Time Captioning [9] and Translation of the classroom to the user's chosen language.
- Evaluation Aid : Identify the key words used in response to a question and cross referencing [10] it with the expected key words and helping the evaluator during a Viva.

The Scope for the use of this tool is vast and the speech to text module proposed in this paper can be leveraged to assist beyond these tools implemented herein.

## II. LITERATURE Review

This collection of research papers provides a comprehensive overview of advancements in the field of Speech-to-Text (S2T) technologies. The authors explore various aspects of S2T, including neural basis of speech production and comprehension, personalized federated learning, voice synthesis, text-to-speech conversion, and keyword extraction techniques. These studies highlight the ongoing efforts to improve the efficiency, accuracy, and personalization of S2T systems, while also addressing challenges such as communication overhead, performance degradation, aperiodic distortion, and the impact of punctuation on keyword extraction. The research underscores the potential of S2T technologies in facilitating seamless human-computer interaction and advancing the field of natural language processing.

Goldstein et al. [11] investigated how the brain processes realworld speech by utilizing a powerful speech-to-text model known as Whisper. By analyzing the model's predictions about brain activity in response to different aspects of speech, they discovered a widespread network across the brain's outer layer (cortex) dedicated to speech and language. This network involves sensory and motor areas responding to the sound features of speech, while higher-level language regions activate in response to grammatical structure and meaning. This research unveils a complex and distributed brain system responsible for understanding and producing spoken language in everyday situations.

Du et al. [12] proposed a new approach that tackles two problems commonly faced in training Speech-to-Text systems: high communication costs and inaccurate performance due to diverse data across participantsIn order to tackle the initial problem, FedLoRA, a lightweight module, is utilized by devices to adjust their models and communicate with a central server, reducing the need for extensive communication. Moreover, the FedMem technique tailors the central model by integrating a specific classifier that adapts to the distinct data patterns of each client. This dual method successfully customizes the model while decreasing communication costs for different S2T tasks.

Nuthakki et al. [13] discuss the challenges of traditional concatenation speech synthesis technologies and propose a new approach for voice synthesis. They suggest that their model’s minimal aperiodic distortion makes it an excellent candidate for a communication recognition model. The authors also highlight the need for more robust network foundations and optimization methods for their proposed algorithm to perform at its best. They believe that their approach is as close to human speech as possible, despite the fact that speech synthesis has a number of audible flaws.

Shastri et al. [14] explore the processes involved in translating English to Hindi, starting with written text and then converting it to spoken words. Two methods are utilized for recognizing characters from images: maximally stable extensible region (MSER) and grayscale conversion. The study also explores geometric filtering alongside stroke width transform (SWT). Once text sequences are identified and broken down into individual words, a spell check with 96% accuracy is conducted using naive Bayes and decision tree algorithms. The final steps involve optical character recognition (OCR) to digitize the text and a text-to-speech synthesizer (TTS) for converting it into Hindi.

Kontagora et al. [15] This study explored the effectiveness of common keyword extraction methods on speech-to-text data. They developed a new audio dataset and compared two popular methods, RAKE and TextRank, with different settings on both the original written text and its corresponding speech-to-text version. They measured performance using precision, recall, and F-score. TextRank with a specific list of stop words (FOX) performed best on both text and audio, achieving F-scores of 16.59% and 14.22%, respectively. While the audio F-score was lower, it was still considered suitable for applications involving spoken conversations. Notably, the lack of punctuation in the speech-to-text data negatively impacted all methods tested.

Adjila et al. [16] provides a novel method for locating and removing silent portions from voice signals. In applications like automatic voice segmentation and speech recognition, this strategy greatly improves system performance and accuracy by utilizing the continuous average energy of the signal. Notably, it maintains lower computing complexity and outperforms contemporary approaches based on spectral centroid and multi-scale product. The MATLAB evaluation of the study shows that the suggested strategy is resilient when dealing with speech signals in Arabic, French, and English.

## III. METHODOLOGY

The conducted experiment presents a meticulously structured methodology aimed at comprehensively analyzing the spectral characteristics of phonemes within recorded speech signals. Initially, the speech signal is loaded from a specified audio file using the librosa.load() function, ensuring compatibility with subsequent processing steps. Spectral analysis begins with a Fourier transform using numpy.fft.fft(), providing insights into the signal's frequency domain characteristics, followed by visualization of the amplitude components through matplotlib.pyplot plots.

The experiment expands its scope by applying conventional filters such as rectangular, bandpass, and highpass filters, crafted using scipy.signal.firwin(), to the time-domain signal reconstructed from the frequency domain. Additionally, less common filters like the Cosine filter and the Gaussian filter, also designed with scipy.signal.firwin(), are investigated. These filters, applied in the frequency domain, utilize cosine-shaped and Gaussian-shaped windows respectively, showcasing diverse effects on the filtered signals.

Furthermore, the experiment meticulously targets specific segments representing vowels, consonants, silence, and non-voiced portions within the speech signal. Each segment undergoes Fast Fourier Transform (FFT) analysis to derive its amplitude spectrum, revealing nuanced frequency profiles inherent to different phonetic contexts. This iterative process captures the distinct spectral signatures of vowels and consonants, considering factors such as airflow obstruction or turbulence. Additionally, slices of silence and non-voiced segments undergo rigorous identification and processing using FFT, contributing to a comprehensive understanding of the speech signal's overall composition.

The experiment culminates in the generation of a spectrogram from the entire speech signal, providing a visual representation of its frequency content evolving over time. This facilitates the identification of crucial change points corresponding to different speech segments. Overall, this methodical and comprehensive approach yields invaluable insights into the spectral representations of vowels, consonants, silence, and non-voiced portions, significantly enhancing the comprehension and analysis of speech signals across diverse applications and domains.

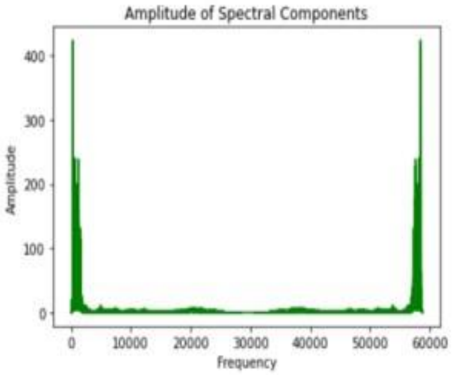


**Fig. 1.** The Amplitude part of the Spectral Components in this study.

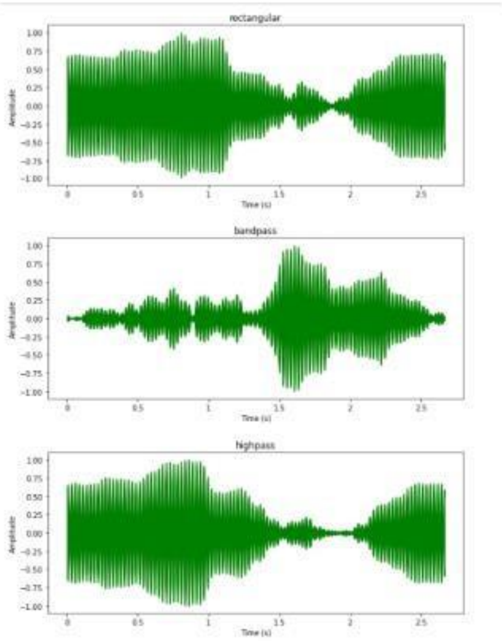


**Fig. 2.** Filtering applied using three conventional filter types - rectangular, bandpass, and highpass

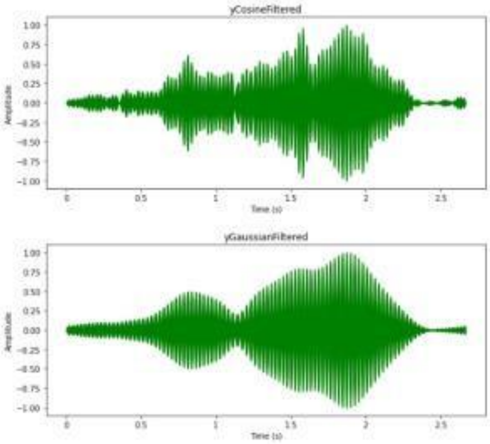


**Fig. 3.**Cosine and Gaussian filters applied in addition to the conventional filters.

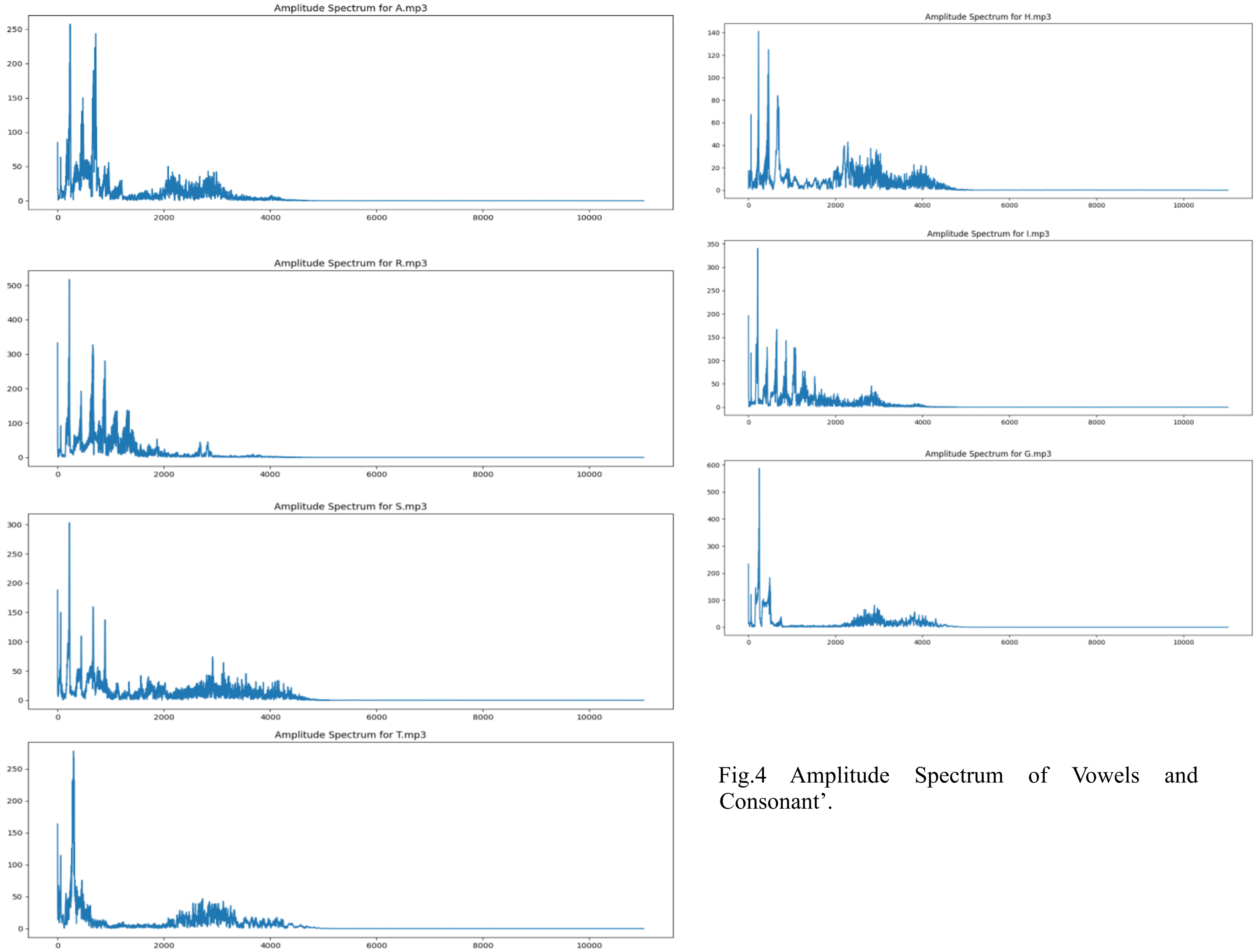
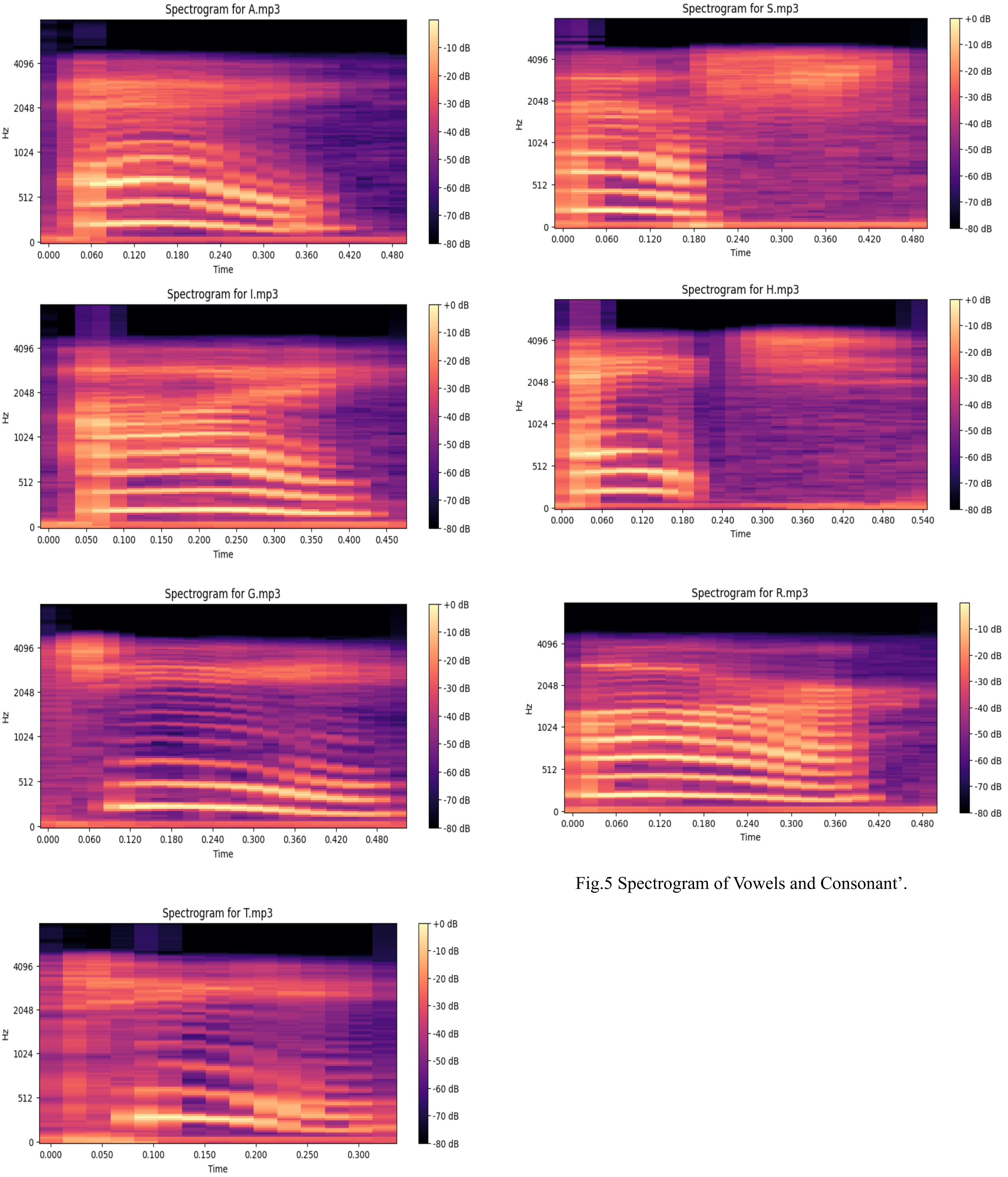


Fig.4 Amplitude Spectrum of Vowels and Consonant'.


Spectrogram for A.mp3

Spectrogram for S.mp3

Spectrogram for I.mp3

Spectrogram for H.mp3

Spectrogram for G.mp3

Spectrogram for R.mp3

Spectrogram for T.mp3

Fig.5 Spectrogram of Vowels and Consonant'.

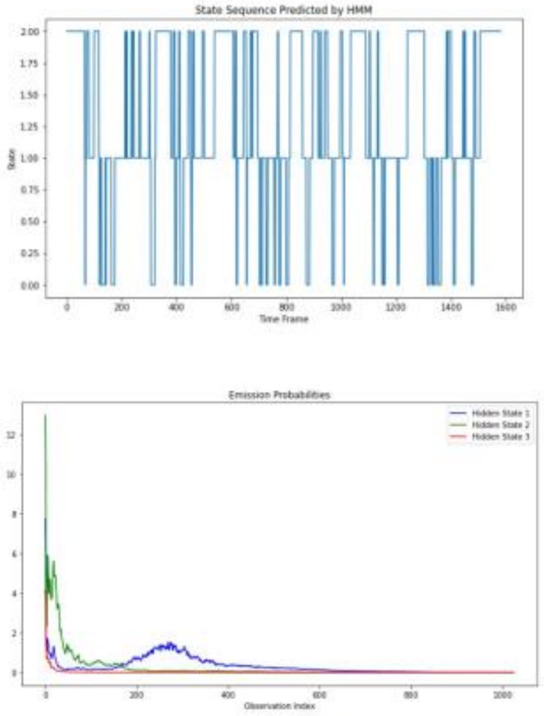


Fig. 6 State Sequence & Emission Probabilities

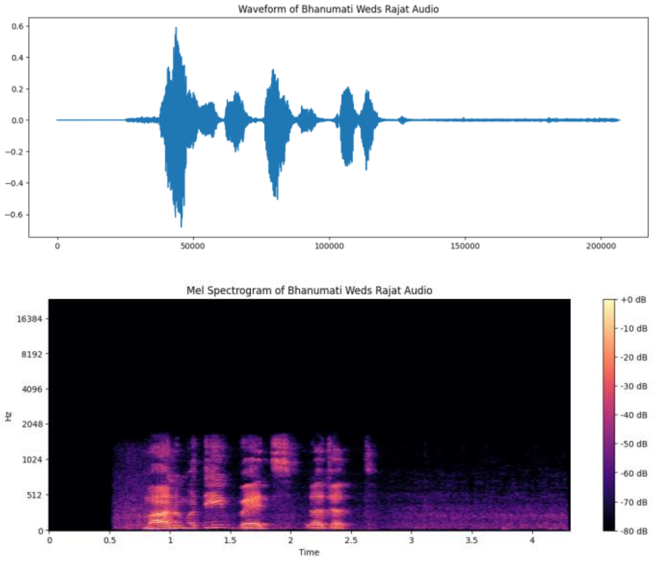


Fig. 7 Waveform & Mel Spectrogram of Bhanumati Weds Rajat Audio

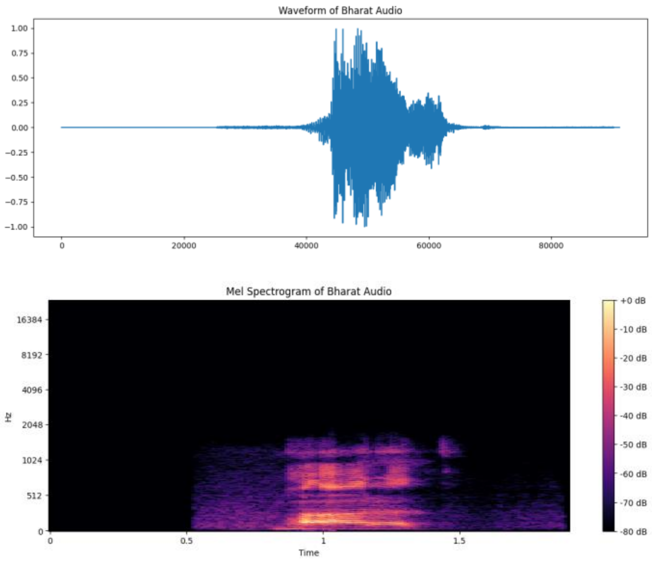


Fig. 8 Waveform & Mel Spectrogram of Bharat Generated Audio

The study in lab 7, the study outlines a methodology for speech classification using Hidden Markov Models (HMMs) and Short-Time Fourier Transform (STFT) features. Initially, the audio file containing the speech signal is loaded using the librosa library. Subsequently, the STFT features are computed from the signal to extract its time-varying spectral content, providing a representation that captures both temporal and spectral characteristics of the speech. These features are then flattened and standardized using StandardScaler from sklearn.preprocessing to ensure uniform contribution of each feature and facilitate model convergence during training. An HMM is instantiated using hmmlearn, configured with parameters such as the number of hidden states and covariance type, and trained on the standardized STFT features. Once trained, the HMM is employed to predict the most likely sequence of hidden states for the features, effectively classifying the speech signal. The resulting sequence of labels represents the classification of each frame of the speech signal based on its spectral characteristics. This methodology forms a foundational approach for speech classification tasks leveraging HMMs and STFT features, with potential applications across various domains including speech recognition and speaker identification.

The study in Lab 8 utilized LSTM and Bi-LSTM networks for speech recognition by extracting features from spectrograms produced through STFT or STCT. These features were then processed using methods like MFCC or LPC coefficients. Training the networks on labeled datasets helped in accurately identifying phonemes, allowing for precise segmentation of speech recordings like "Bhanumathi weds Rajat" into separate phonemes. After segmentation, the phonemes associated with the word "Bharat" were chosen and combined to create the speech signal for the target word. While simply putting words together may seem like a quick solution, it can cause problems such as awkward rhythm, gaps between sounds, and possibly saying words wrong because of missing context. Assessing the generated speech involves checking for problems like choppiness, mispronunciations, and overall sound quality, and fine-tuning the process to fix any issues by using better models and a wider range of data sources.

While accurate phoneme recognition has the potential to be successful, there are challenges when directly concatenating segmented phonemes for speech synthesis. The synthesized speech may not sound natural or coherent because contextual information is missing, and there are limitations in the concatenation process. Problems like gaps between phonemes, unnatural intonation, and mispronunciations can decrease the quality and understandability of the synthesized speech. To tackle these challenges, we need to continuously improve the synthesis process by using more advanced modelling techniques, being aware of context, and enhancing signal processing methods. This will help create more natural and quality synthesized speech.

## IV. Results

In our study, we delve into speech signal analysis using spectral transformations. Initially, we employ numpy.fft.fft() to transform the speech signal into the spectral domain, revealing amplitude distribution across frequencies. Subsequently, we explore three filter types: rectangular, bandpass, and high pass. Each filter selectively manipulates low-frequency, bandpass, and high-frequency components in the spectrum. The reconstructed time-domain signals are then plotted and saved as audio files (named ‘filteredrectangular.wav’, ‘filteredbandpass.wav’, and ‘filteredhighpass.wav’).

Expanding our investigation, we introduce two additional filter types: Cosine and Gaussian filters. These filters impart distinct characteristics to the original signal. The resulting waveforms, plotted in the time domain, showcase the versatility of signal processing techniques. We save the corresponding audio files as ‘filteredCosine.wav’ and ‘filteredGaussian.wav’ . Overall, this comprehensive exploration demonstrates how different filters impact the time-domain representation of speech signals, providing valuable insights for audio engineering applications.

On the Observations of Fig 4 and Fig 5

Vowel Sounds (A, I):
Amplitude Spectrum: Vowel sounds typically display distinct peaks at specific frequencies in their amplitude spectrum, representing characteristic formants. These peaks indicate resonant frequencies of the vocal tract during vowel articulation.
Spectrogram: Vowels are identifiable in the spectrogram by clearly defined bands of energy corresponding to their formants. These bands remain relatively stable over time, facilitating the recognition of vowel segments.

Consonant Sounds (G, H, R, S, T):
Amplitude Spectrum: Consonant sounds exhibit a broader energy distribution across frequencies compared to vowels. They lack distinct peaks in the amplitude spectrum and often show characteristics resembling noise.
Spectrogram: Consonants are typically observed as brief bursts of energy in the spectrogram, frequently occurring at higher frequencies. Their production may result in rapid changes in intensity and frequency content, discernible in the spectrogram.

Silence & Non-Voiced Portions:
Amplitude Spectrum: Silence and non-voiced segments of speech manifest as low energy levels across frequencies, resulting in a relatively flat amplitude spectrum with minimal peaks.
Spectrogram: Silence is represented by uniformly low energy levels across all frequencies and time intervals in the spectrogram. Non-voiced sounds may exhibit some energy, but lack the distinct patterns observed in voiced segments such as vowels and voiced consonants.

Identification from Spectrogram:
Vowels are characterized by bands of energy corresponding to formants, which appear as stable regions in the spectrogram.
Consonants are discerned as sudden changes or short bursts of energy, often occurring at higher frequencies. Their patterns in the spectrogram may be less consistent compared to vowels.

Transitions between consonants and vowels, or between different consonants, can be identified by abrupt changes in the spectrogram's energy distribution and frequency content.

## V. Conclusion

In our conducted experiments, we delved into signal processing techniques applied to speech signals, yielding valuable insights into their spectral characteristics and temporal dynamics. Initially, transforming the speech signal into the spectral domain provided a comprehensive view of amplitude distribution across frequencies. Subsequent experiments employing various filters—rectangular, bandpass, high-pass, Cosine, and Gaussian—demonstrated selective manipulation of frequency components, each imparting unique effects on the signal. The resulting waveforms and audio files vividly illustrated the distinct impact of each filter on the time-domain signal, highlighting the significance of signal processing in audio applications such as noise reduction and spectral shaping.

Furthermore, our analysis of speech signals through amplitude spectrum and spectrogram unveiled distinctive patterns characterizing vowels, consonants, and silent/non-voiced portions. Vowel sounds exhibited clear formants in both the amplitude spectrum and spectrogram, with stable bands of energy reflecting resonant frequencies of the vocal tract. In contrast, consonants showcased broader energy distributions, often resembling bursts of noise in the spectrogram, without the pronounced peaks observed in vowels. Silence and non-voiced segments were discernible by their low energy levels across frequencies, resulting in flat spectra and uniform energy distribution in the spectrogram.

These combined insights underscore the profound role of signal processing techniques in enhancing our understanding and manipulation of audio signals. By leveraging such techniques, we not only improve audio quality but also unlock new possibilities in audio engineering, contributing to advancements in various fields reliant on sound analysis and manipulation.

## References

[1] Y. Wei et al., "AdaStreamLite: Environment-adaptive Streaming Speech Recognition on Mobile Devices," in Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 7, no. 4, pp. 1-29, Dec. 2023, doi: 10.1145/3631460.

[2] J. Laures-Gore et al., "Dialect identification, intelligibility ratings, and acceptability ratings of dysarthric speech in two American English dialects," in Clinical Linguistics & Phonetics, pp. 1-12, doi: 10.1080/02699206.2023.2301337.

[3] S. Feng et al., "Towards inclusive automatic speech recognition," in Computer Speech & Language, vol. 84, p. 101567, 2024, doi: 10.1016/j.csl.2023.101567.

[4] V. Karthikeyan and S. Suja Priyadharsini, "Modified layer deep convolution neural network for text-independent speaker recognition," in Journal of Experimental & Theoretical Artificial Intelligence, vol. 36, no. 2, pp. 273-285, 2024, doi: 10.1080/0952813X.2022.2092560.

[5] P. Gambhir et al., "End-to-end Multi-modal Low-resourced Speech Keywords Recognition Using Sequential Conv2D Nets," in ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 23, no. 1, pp. 1-21, Jan. 2024, doi: 10.1145/3606019.

[6] M. Devare and M. Thakral, "Enhancing Automatic Speech Recognition System Performance for Punjabi Language through Feature Extraction and Model Optimization," in International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 8s, pp. 307–313, 2023.

[7] F. Wu et al., "Wav2Seq: Pre-Training Speech-to-Text EncoderDecoder Models Using Pseudo Languages," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096988.

[8] Y. Wei et al., "AdaStreamLite: Environment-adaptive Streaming Speech Recognition on Mobile Devices," in Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 7, no. 4, pp. not provided, Dec. 2023, doi: not provided.

[9] L. Liu, L. Liu, and H. Li, "Computation and Parameter Efficient MultiModal Fusion Transformer for Cued Speech Recognition," arXiv preprint arXiv:2401.17604, 2024, primary class: cs.CV.

[10] R. Shukla, "Keywords Extraction and Sentiment Analysis using Automatic Speech Recognition," arXiv preprint arXiv:2004.04099, 2020, primary class: eess.AS.

[11] A. Goldstein et al., "Deep speech-to-text models capture the neural basis of spontaneous speech in everyday conversations," bioRxiv, 2023, pp. 2023-06.

[12] Y. Du et al., "Communication-Efficient Personalized Federated Learning for Speech-to-Text Tasks," arXiv preprint arXiv:2401.10070, 2024.

[13] P. Nuthakki et al., "Deep Learning based Multilingual Speech Synthesis using Multi Feature Fusion Methods," in ACM Transactions on Asian and Low-Resource Language Information Processing, 2023.

[14] S. Shastri and S. Vishwakarma, "An Efficient Approach for Text-toSpeech Conversion Using Machine Learning and Image Processing Technique."

[15] B. Nuhu Kontagora et al., "Performance Evaluation of Keyword Extraction Techniques and Stop Word Lists on Speech-To-Text Cor," 2023.

[16] A. Adjila, M. Ahfir, and D. Ziadi, "Silence Detection and Removal Method Based on the Continuous Average Energy of Speech Signal," in