Exploring Advancements in Speech Processing

Devansh Handa
Department of Computer Science
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham,
Bengaluru, India.
BL.EN.U4AIE21041@bl.students.amrita.edu

Kindi Krishna Nikhil
Department of Computer Science
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham,
Bengaluru, India.
BL.EN.U4AIE21068@bl.students.amrita.edu

Vaan Amuthu Elango
Department of Computer Science
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham,
Bengaluru, India.
e_vaanamuthu@blr.amrita.edu

***Abstract* – The purpose of this work is an inventory of the latest discoveries in speech technology represented in twelve bright works. They span from emotional voice conversion to speech analytics and language learning and so on and these studies represent the multidirectional nature of modern experiments. This review studies interconnections amongst domains of speech processing synthesizing key insights which leads to a multi-faceted view of the obstacles, inventions and future projections of speech processing.**

Keywords:* *Speech processing, emotional voice conversion, speech analytics, language learning, deep learning, machine learning, signal processing, voice conversion techniques, natural language processing, privacy protection, data security, human-machine interaction, computational intelligence.

I. INTRODUCTION

In speech processing field there are rapid technological development with the key contributions being evident in the intersection of signal processing, machine learning and linguistics. Behind the cutting-edge development, what is now visible is beyond the conception only, for the awareness of the intricate relationship between human language and logical computation now strikes hard upon us. Immerse in these dynamic shifting grounds, emotional voice conversion becomes an interesting research field [1]. By revealing of the subtleties that speech signals contain in the area of emotional expression, researchers seek to match machine cognition with human perception of emotion thus ensuring more caring human–machine interaction.

Also, the strategic integration of speech analytics into diverse domains has completely changed how we draw inferences from audio data streams [2]. From tracking information streams live to empowering seamless customer service, the introduction of automatic keywords detection systems has brought data driven decision making to a whole new level. In addition, due to the demand for efficient language learning approaches, scholars began to examine deep feature-based clustering algorithms [3]. These methods which employ the potential of machine learning provide personalized feedback and correction mechanisms tailored to suit the distinct learning needs and preference of the individual.

Simultaneously, the goal of ensuring the protection of the privacy and security of voice data has given rise to voice de-identification techniques [4]. With voice data growing ever omnipresent in our digital arena, the confidentiality and integrity of sensitive information turns out be one of the top items on the agenda. Researchers aim to find a middle ground through innovative voice conversion and anonymization methodologies utility and privacy thereby earning trust and an openness in voice-operated systems [5].

Recent studies have been directed into speech and prosodic error annotation and detection, for instance, see [6], [7]. This work may thus lead to more accurate language learning methods accompanied with speech enhancement methods. Likewise, recent advancements in voice conversion methods have brought about expressive voice conversion architectures [8], [9], providing speaker-dependent and high-fidelity synthesis opportunities. Although end-to-end frameworks of speech waveform re-construction and clean content extraction have greatly enhanced the quality and robustness of the speech conversions [10], [11]. Thus, the research of anonymization methods which is based on voice conversion has dealt with the privacy issues formatted in protected speaker characteristics [12].

II. RELATED WORKS

The characteristics of the human voice, which include its intricate warp of emotions, intentions and information, are irresistible ones. It is easy to understand why the domains of speech analytics and voice changing are booming right now, with scientists doing everything they can to discover the hidden treasures marinated in spoken language. This exponential exploration of human life spans into an even deeper exploration of the curated recent literature (in random order) and precise dissections of how it adds to our understanding of human experiences and the changes it can bring.

Beyond Business Metrics: Unleashing Societal Accelerator: Though the initiator of speech analytics' role in the business research, it has gone beyond merely the spreadsheets and profit margins. In their work, Farkhadov et al. [2] look at the use of AI in the information space monitoring systems to provide authorities with a better capacity to identify risks and regain situation awareness. For this application, there is tremendous societal significance, with the possibility of making the world and places of residence safer.

Emotions Take Center Stage: The Nonverbal Gap: The emote space is one of the features that act as a bridge during nonparallel emotional voice conversion that was discovered by Shah et al. [1]. This invention surpasses language barriers as it provides a viable platform whereby speakers can communicate their emotions, be they speakers and/or languages. Now visualize the impact on cross-cultural conversations or facilitative technologies that will enable people with communicative emotions challenge to authentically connect and listen to others.

Quality and Efficiency: The Two Wheels of Invention: Enhancing the speech quality brought about by the conversion is the top priority, and Sun et al. [3] fix this problem with their unique random cycle loss function. This development signifies more natural and exact voice confabulations which could bring revolutionary breakthroughs in audio books, documentation narration or even personalized education. Imagine that you'll be immersed in fascinating audio books voiced in your own language and receive personalized educational information based on your specific learning style through this great technological advancement.

Efficiency is another important aspect that researchers have been managing at the same level. Li et al. [4] presented STYLETTS-VC, which is a technique that can produce high quality transformations with a single training sample by applying knowledge transfer from text-to-speech models. Such options expands possibilities for unique, tailored experiences on-demand such as voice translations in real life which retains your pronunciation characteristics. Just imagine how it would be to have a conversation with an individual even somewhere across the world as if you were simply talking to yourself. Li et al. [9] improves FREEVC, by doing away with the text generation replacement step, and hinting at a future with voice conversion being comfortable and ubiquitous.

Control and Expression: Displaying the Need for control and Expressiveness: The ability of voice conversion to somewhat satisfy the need for control and expressiveness can be seen in the works of Hussain et al. [7] as well as Ning et al. [8]. The prior's ACE-VC tech that allows users to edit and enhance the converted language with self-taught features; the latter's EXPRESSIVE-VC model with attention fusion to achieve near real speech with high emotional content. These developments set the path for customizable voice avatars which spoils the uniqueness of your personality, personalized storytelling that casts a spell on you, and also that emotionally intelligent conversational AI which actually realizes and responds to your emotions and feelings.

Safeguarding Privacy in a Voice-Driven World: With emergence of voice technologies as the main component of our lives the privacy becomes of the greatest importance. This issue is tackled by Srivastava et al. [11] their work investigates security vunerabilities in voice conversion and countermeasures. Such a research is responsible for safe and conscious use of these technologies to guarantee data privacy and provide confidence in a voice-controlled society. Imagine using voice assistants or validation systems without worrying about where your voice data may appear.

In conclusion, this widened research opens up the colorful and multi-disciplinary research paradigm that the speech analytics and voice conversion are not just tools but means of reaching deeper perception and higher quality communication. The applications ranges from societal to personalized, and including emotional expression and subtle control. The potential gets even bigger and more varied as it continues to evolve. Instead, researchers can anticipate new technological breakthroughs that will be even more surprising in terms of natural languages spoken interaction and processing, creating a new world where the human voice becomes the key to digital universe. Perhaps one fine day, chats with AI buddies would be as seamless as communicating with a human friend by your side. Life could become easier as we could share our stories and emotions in the most convenient way- through the magic of our voice

III. Methodology

A. Lab 1

There are several operations done on the audio file referred to as "Lab_01.wav" using the librosa library in the Python programming language. At the beginning, audio file is loaded from the librosa's load function and the result is returned in y variable, which contains the audio data and in sr variable, which represents the sampling rate. The waveform of the audio signal is then plotted by the matplotlib, providing the visual representation of the amplitude over the time.

The audio shall be examined more closely by targeting the parts when words are spoken and the silence periods are marked using the remarks. These segments are now labeled on the waveform plot as colored spans between corresponding characters, e.g, "A", "I", "in", "Speech" and "Processing". In addition, the duration of the audio signal is in seconds is calculated by dividing the audio data array length by the sampling rate and the the value in decibels of the audio signal corresponds to the magnitude range and it is able to find the maximum value.

Furthermore, respective audio segments matching the indicated intervals are deconstructed from the original audio. Each part is plotted separately to show its waveform and therefore to grasp more easily the characteristics of different regions of the wave. Besides, the regions of silence in the audio signal are also trimmed using librosa's trim function with the threshold as 20 dB and the trimmed audio are plotted as a waveform in order to illustrate the effect of trimming.

Next, the audio signal is resampled to desired target sampling rates (e.g., 40,000 Hz, 8,000 Hz via librosa's resample function). Plotted "resampled" audio waveforms are provided for the purpose of visualizing the effect of the resampling process on the signal's characteristics. The spectrograms finally, are calculated using a Short-Time Fourier Transform (STFT) for the audio signal. Both the spectrogram and mel spectrogram are plotted for visualizing the frequency content of the audio signal as a function of time. This method gives detailed analyses of the audio data.

B. Lab 2

Operations are done using audio signal from the scipy, matplotlib, IPython, numpy, and librosa libraries in Python. At the beginning the program extracts the original speech signal from "Lab_01.wav" file that is in the audio format using the wavfile.read function from scipy. In case of stereo audio signals, it is converted to mono by taking the mean value of channels. The first derivative of the speech signal is obtained by the Finite Difference Method (FDM) and the resulting output is saved as another audio file named: "first_derivative.wav".

Then, matplotlib is deployed to plot the existing speech signal and its derivative. The original speech signal is plotted at the top subplot, but the first derivative is plotted below it as the bottom subplot. Both signals amplitude are given on plots which demonstrate variation over time. Therefore, this code provides an opportunity to hear the original utterance and its first derivative Audio IPython function.

Finally, the new first derivative is loaded again into memory using the librosa.load function. Finally, zeros are found in first derivative signal by librosa's zero_crossings function. The distances between consecutive zero crossings are being calculated and speech and silence intervals are being applied through a given threshold. The code interprets first derivative signal overlaid by speech and silence areas. It recognizes also zeros.

Besides, the program calculates and shows the mean lengths of the zero crossings between continuous speech and silence. This yields details on the periodic properties of spoken and silence segments in the audio signal.

Finally, the code assigns a score to two people ("Nikhil" and "Devansh") who say some specific words. The speech-word lengths are calculated by accessing the audio files of each word, determining duration of each sample, and holding on the durations in lists. The bar plots which are then used to represent and compare the speech lengths between the two individuals are being created.

Furthermore, the coding has a plot_audio_signals function to diagram the audio signals for individual words of individual speakers. The plotting function takes a list of audio filenames and plots their signals with their titles as x-axis and y-axis respectively. This type of these plots give visuals the audio signals amplitude and their reference speech are given by the two speakers "Nikhil" and "Devansh".

In a nutshell, the code carries out different analyses and comparisons of speech signals, ranging from derivatives computation, speech and silence region identification, word lengths comparison between individuals, to visual representations for individual words. These studies give interpretations regarding the temporal features as well as the pronunciation patterns of words in speech.

C. Lab 3

The methodology employed in this study involved three main procedures for silence detection and audio segmentation. First, the librosa.effects.trim() function was utilized to remove silent parts from the beginning and end of the recorded speech signal, ensuring the extraction of only meaningful audio segments. Subsequently, the librosa.effects.split() function was employed to segment the recorded speech signal based on detected silences, with varying thresholds (top_db values) to observe the effect on segmentation quality. Lastly, a custom silence detection algorithm inspired by the paper "Silence Detection and Removal Method Based on the Continuous Average Energy of Speech Signal." [13] was implemented, which relied on calculating the energy of the signal and comparing it against a predefined threshold to identify silent regions. These methodologies aimed to explore and compare different approaches for silence detection and segmentation in speech signals.

D. Lab 4

The use of various techniques was a crucial factor for the analysis of the spectral attributes and the temporal facts of the speech signal which was actually recorded. Firstly, the amplitude spectrum of the voice signal was observed using Fast Fourier Transform (FFT) by applying numpy.fft.fft() which demonstrated the distribution of amplitudes amongst different points. Next, the signal reconstruction process was assessed for its fidelity by making a comparison between the signal, derived from computing the inverse Fourier transform using the Fast Fourier Transform through numpy.fft.ifft(), and the original speech signal. The comparison of historical and reconstructed signals, as presented in the time domain, allowed to evaluate the reconstruction process adequacy. Besides that, certain portions of the speech signal were identified and featured for analysis; in this way we converted time-dependent signals into frequency domain to discern the corresponding frequencies. Moreover, Fourier spectrum was found by plotting it through a rectangular window function which has been applied to the speech signal. This emphasized localized spectral features in specific time windows only. In addition to this, Short-Time Fourier Transform (STFT) was computed using librosa.stft() to analyze the speech signal in time-frequency domain, As a result it provided an opportunity to understand the time-varying nature of its frequency content. Lastly, through scipy.signal.spectrogram(), a spectrogram of the speech signal was created which can be seen in the figure above. The spectrogram shows a time-frequency distribution of spectral intensity, hence enabling one to examine time-varying spectral characteristics within a speech signal. These detailed examinations helped me to acquire the knowledge about how the signal processing was performed and allowed me to learn about the features of the filtered speech signal.

IV. RESULTS

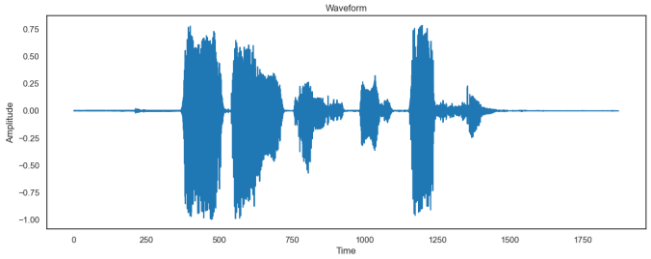

Fig. 1. The waveform “AI in Speech Processing”

As observed in Fig.1. the .wav file is loaded onto for the first time and visualized with the help of an amplitude vs time plot.

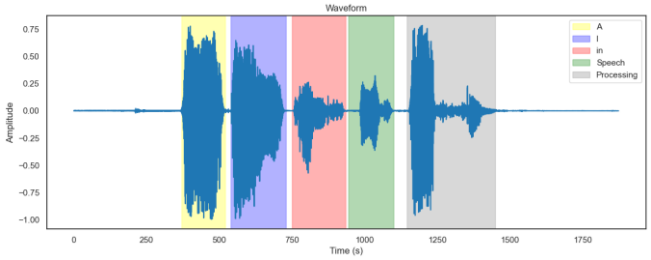

Fig. 2. Every word in the waveform being Highlighted

Every word is highlighted in the waveform as seen in Fig.2. The audio file, sections it into 5 different parts.

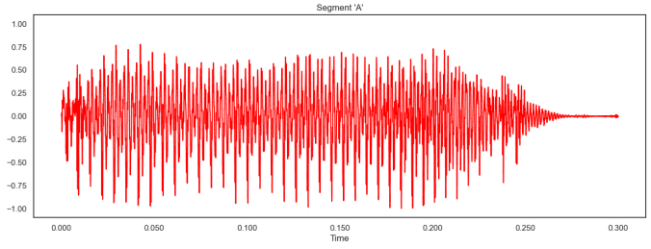

Fig. 3. A zoomed in segment of the audio for “A”

For better visualization and better understanding, the plot is zoomed in for every sectioned word, being “A”, ”I”, ”in”,

”Speech”, ”Processing”. The visualization of “A” can be observed in Fig.3.

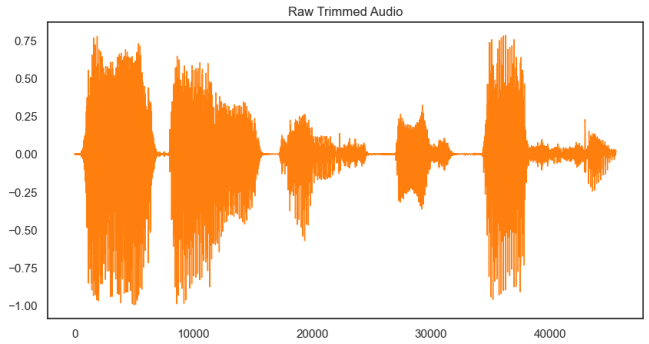

Fig. 4. Trimmed audio wave

Using inbuilt librosa functions, the silent zones are trimmed and the audio file is visualized in Fig.4.

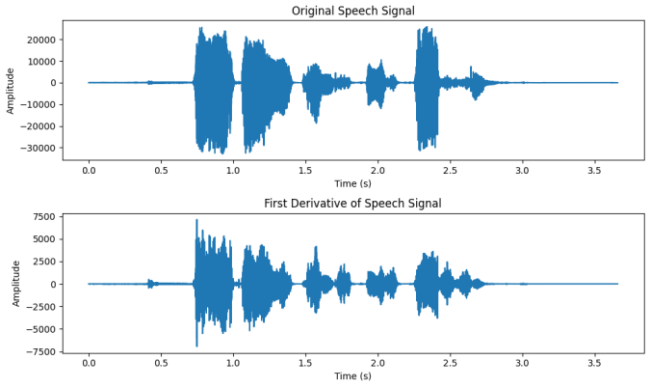

Fig. 5. The Original speech and its derivative

The comparative visualization between the Original Speech and the First Derivative can be observed in Fig.5.

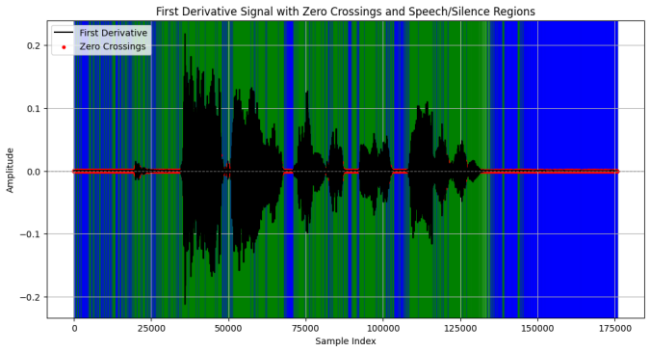

Fig. 4. Zero Crossing and Silent zones

For understanding the zero crossing and the speech and silent zones, Fig.4 can be observed, where the red dots represent the zero crossing, the blue strips represent the silent zones and the green strips represents the speech zones.
The zoomed in version for better visualization can be seen in Fig.5.

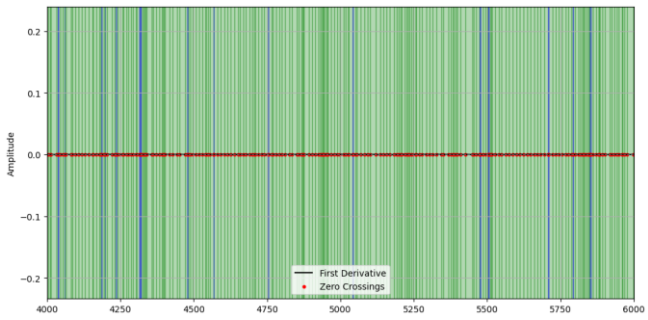

Fig. 5. Zoomed in Zero Crossing and Silent zones.

Fig.6. shows the different word lengths for specific words pronounced by “Devansh” and “Nikhil”.

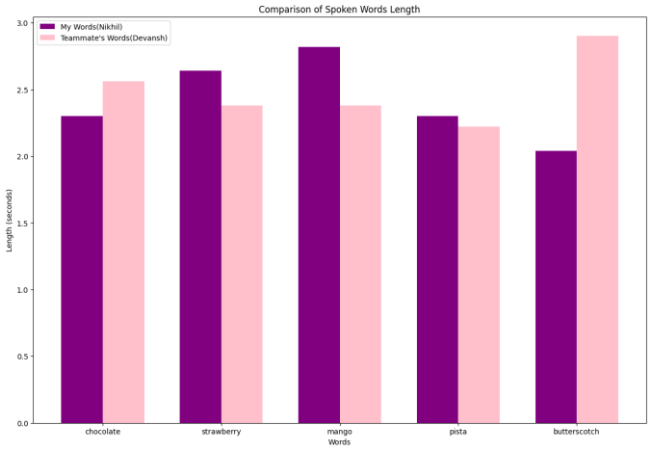

Fig. 6. Word length comaprisons

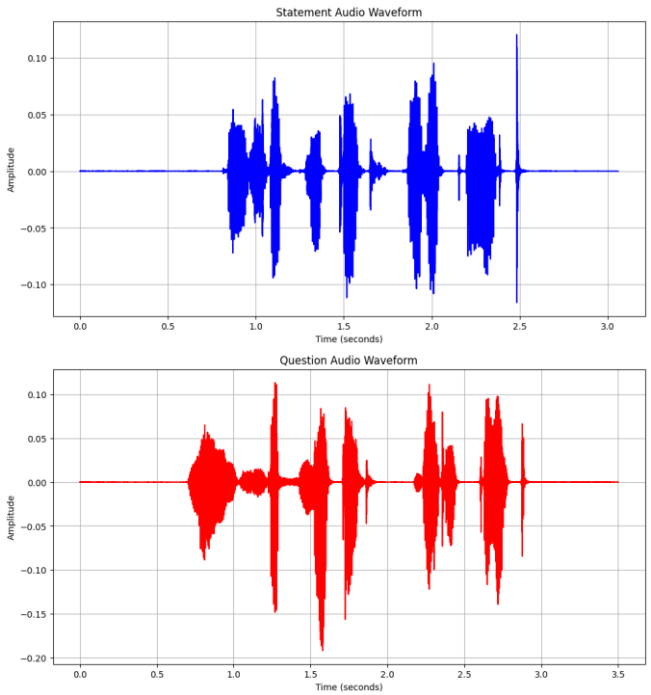

Fig.7. Comparing the Statement “You Submitted The report”

To make the analysis of the way in which a statement can be pronounced in order to change the stress levels and form a question of itself is done in Fig.7. where the statement “You Submitted The report” is said in a different tone representing

both the statement as well as a question. Their comparison results can be observed as the following

Duration of Statement Audio: 3.06 seconds
Duration of Question Audio: 3.5 seconds

Mean Amplitude of Statement Audio: 0.0044322964
Mean Amplitude of Question Audio: 0.0061697927

Peak Amplitude of Statement Audio: 0.12042236
Peak Amplitude of Statement Audio: 0.19189453

The above are the results of Lab1 and Lab2 based on the processing and experiments done with the audio signals.

As found in Fig. 8, the methods of silence detection and silence segmentation are visually presented by the graphs. The comparison of the subintervals created from the librosa.effects.split() function versus the silence motorl filter based on the method described in amid is demonstrated in. The observation is the lables of silent intervals both methods have the same but the partitioning is different. Therefore, the way that the methodology in employed has a bearing in the segmentation.

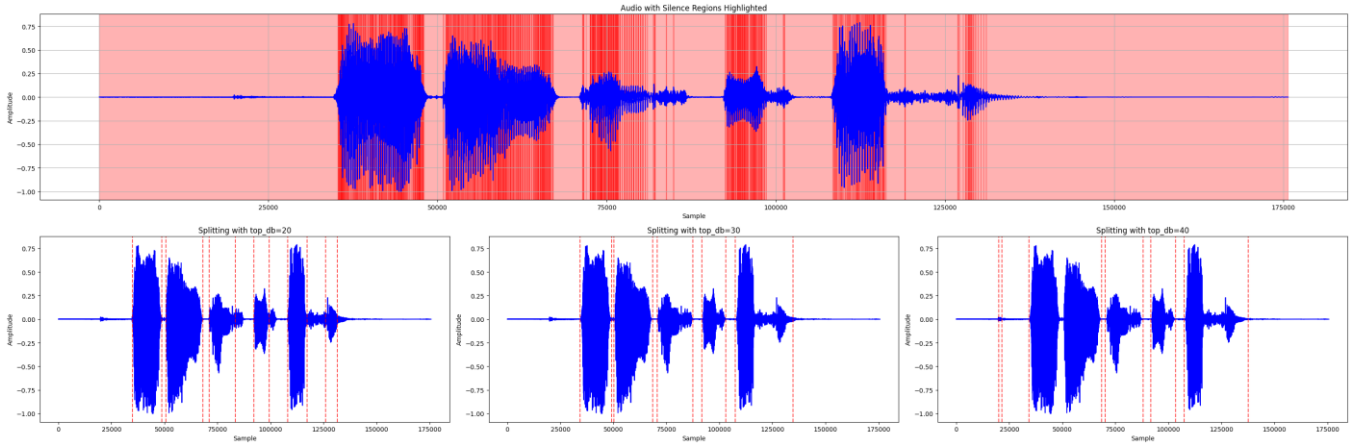

Fig. 8. Comparison of Silence Detection Between Calculating Signal Energy and pre defining a threshold value on a signal

Overall, this result provides insight into the effectiveness and applicability of different silence detection and segmentation techniques in speech signal processing.

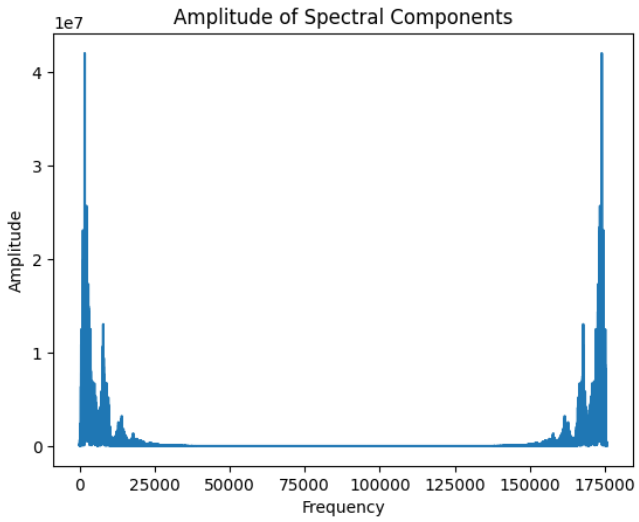

Fig. 9. Amplitude of Spectral Components

Fig. 9. Shows the amplitude spectrum of the speech signal of Lab1 (“AI in Speech Processing”), obtained through Fast Fourier Transform (FFT), displaying the distribution of amplitudes across different frequencies.

The signal is then put through inverse Fourier transform and compared with the original signal in the Fig. 10. (Red : Reconstructed signal, Greed: Original Signal).

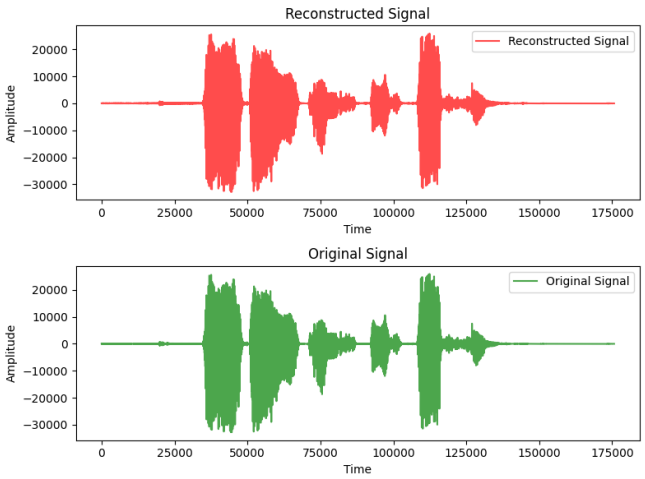

Fig. 10. Comparison of Reconstructed vs. Original Signal

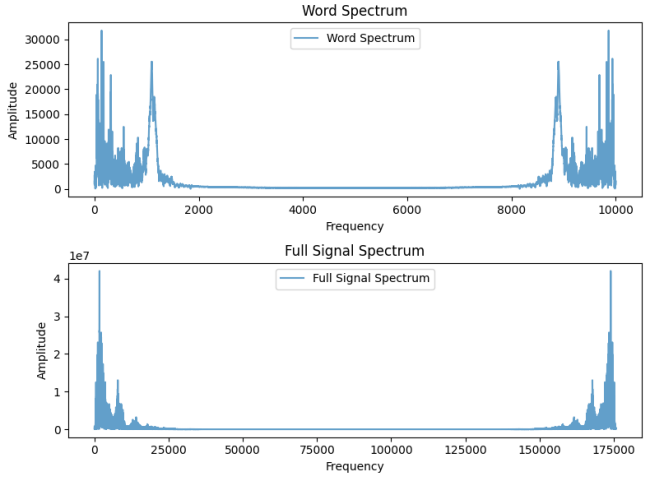

Fig. 11. Word Spectrum vs. Full Signal Spectrum

To observe any distinct frequency components associated with the word segment, a comparison plot of the frequency spectrum of a specific word segment from the speech signal with the spectrum of the entire speech signal can be seen in Fig. 11.

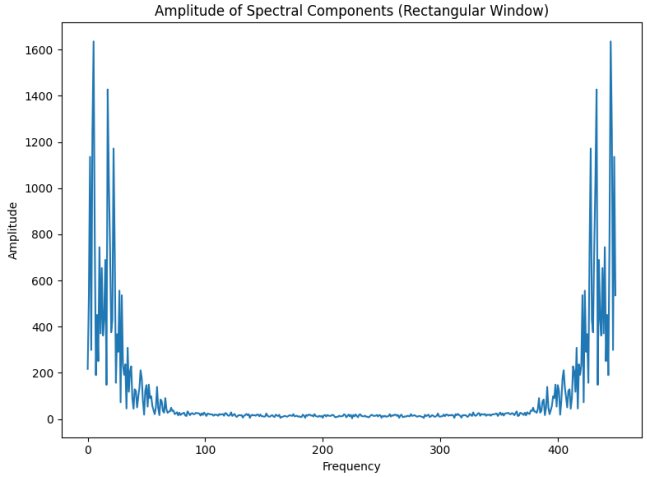

Fig. 12. Amplitude of Spectral Components with Rectangular Window

The spectrogram is obtained by measuring the magnitude within a rectangular window of the speech signal, Fig. 12. delivers the amplitude of spectral components of a specific time domain, thus the method may be used for the local power spectrum.

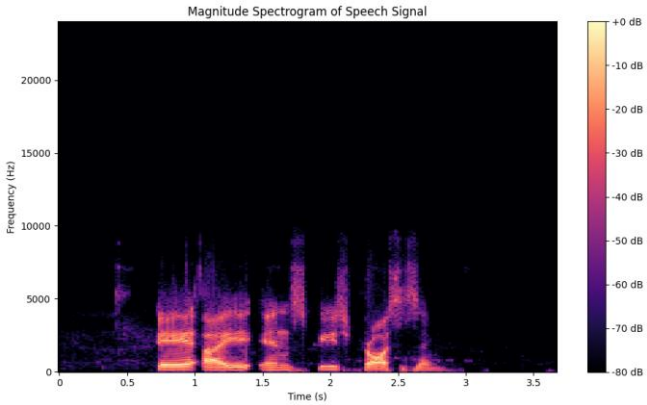

Fig. 13. Magnitude Spectrogram of Speech Signal

Fig. 13. depicts the magnitude spectrogram of the speech signal, computed using Short-Time Fourier Transform (STFT), providing a time-frequency representation of spectral intensity and facilitating the analysis of time-varying frequency content.

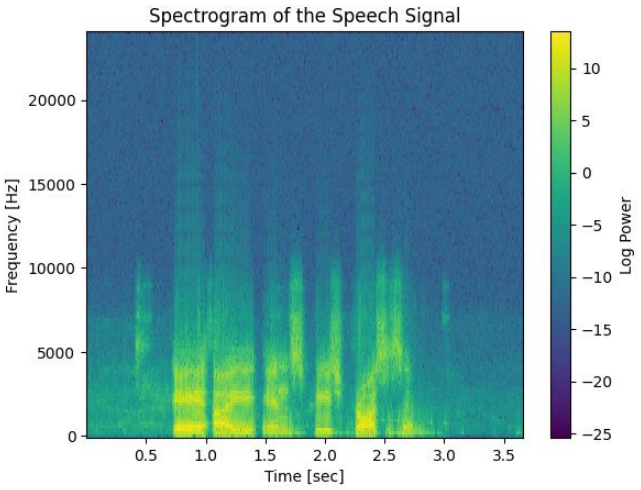

Fig. 14. Spectrogram of the Speech Signal

Fig. 14. Spectrogram is then displayed to show time frequency distribution of spectral intensity with color representation, allowing to focus on time varying spectral features embedded in the speech signal.

The above are the results of the processing done to the voice signal analysed up to Lab 4.

REFERENCES

[1] Shah, N., Singh, M., Takahashi, N. and Onoe, N., 2023, June. Nonparallel Emotional Voice Conversion For Unseen Speaker-Emotion Pairs Using Dual Domain Adversarial Network & Virtual Domain Pairing. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[2] Farkhadov, M., Smirnov, V. and Eliseev, A., 2017, November. Application of speech analytics in information space monitoring systems. In 2017 5th International Conference on Control, Instrumentation, and Automation (ICCIA) (pp. 92-97). IEEE.

[3] Sun, H., Wang, D., Li, L., Chen, C. and Zheng, T.F., 2023. Random Cycle Loss and Its Application to Voice Conversion. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[4] Li, Y.A., Han, C. and Mesgarani, N., 2023, January. Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models. In 2022 IEEE Spoken Language Technology Workshop (SLT) (pp. 920-927). IEEE.

[5] Scheidt, S. and Chung, Q.B., 2019. Making a case for speech analytics to improve customer service quality: Vision, implementation, and evaluation. International Journal of Information Management, 45, pp.223-232.

[6] Nazir, F., Majeed, M.N., Ghazanfar, M.A. and Maqsood, M., 2023. A computer-aided speech analytics approach for pronunciation feedback using deep feature clustering. Multimedia Systems, 29(3), pp.1699-1715.

[7] Hussain, S., Neekhara, P., Huang, J., Li, J. and Ginsburg, B., 2023, June. ACE-VC: Adaptive and Controllable Voice Conversion Using Explicitly Disentangled Self-Supervised Speech Representations. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[8] Ning, Z., Xie, Q., Zhu, P., Wang, Z., Xue, L., Yao, J., Xie, L. and Bi, M., 2023, June. Expressive-VC: Highly Expressive Voice Conversion with Attention Fusion of Bottleneck and Perturbation Features. In

ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE..

[9] Li, J., Tu, W. and Xiao, L., 2023, June. Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[10] Sisman, B., Yamagishi, J., King, S. and Li, H., 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp.132-157.

[11] Srivastava, B.M.L., Vauquier, N., Sahidullah, M., Bellet, A., Tommasi, M. and Vincent, E., 2020, May. Evaluating voice conversion-based privacy protection against informed attackers. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2802-2806). IEEE.

[12] Hildebrand, C., Efthymiou, F., Busquet, F., Hampton, W.H., Hoffman, D.L. and Novak, T.P., 2020. Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications. Journal of Business Research, 121, pp.364-374.

[13] Adjila, A., Ahfir, M. and Ziadi, D., 2021, December. Silence Detection and Removal Method Based on the Continuous Average Energy of Speech Signal. In 2021 International Conference on Information Systems and Advanced Technologies (ICISAT) (pp. 1-5). IEEE.