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 human voice, with its intricate tapestry of emotions, intentions, and information, holds an undeniable power. It's no wonder, then, that the fields of speech analytics and voice conversion are experiencing a meteoric rise, fueled by researchers on a quest to unveil the secrets locked within spoken language. This extended exploration delves even deeper into a curated selection of recent publications (presented in random order), meticulously dissecting their contributions and the transformative potential they hold.

Beyond Business Metrics: Unveiling Societal Value: While Hildebrand et al. [12] lay the foundation for speech analytics' role in business research, its impact stretches far beyond spreadsheets and profit margins. Farkhadov et al. [2] envision harnessing its power in information space monitoring systems, empowering authorities with an enhanced ability to detect threats and maintain situational awareness. This application holds immense societal value, potentially contributing to safer communities and a more secure world.

Emotions Take Center Stage: Bridging the Nonverbal Divide: Stepping into the realm of emotional expression, Shah et al. [1] present a groundbreaking method for nonparallel emotional voice conversion. This innovation transcends language barriers, enabling seamless communication of emotions even when speakers and languages differ. Imagine the impact on cross-cultural interactions or assistive technologies that empower individuals with emotional communication challenges to truly connect and be heard.

Quality and Efficiency: The Twin Engines of Innovation: Enhancing the quality of converted speech is paramount, and Sun et al. [3] tackle this challenge head-on with their novel random cycle loss function. This advancement promises more natural-sounding and accurate voice conversions, potentially revolutionizing applications like audiobooks, narrated documentaries, or even personalized educational tools. Imagine listening to captivating audiobooks narrated in your own voice or receiving educational content tailored to your unique learning style, all made possible by this technological leap.

Efficiency is another crucial aspect, and researchers are making impressive strides. Li et al. [4] introduce STYLETTS-VC, a method that achieves high-quality conversions with just one training sample by leveraging knowledge transfer from text-to-speech models. This opens doors for personalized experiences on-demand, like real-time language translation that retains your individual voice characteristics. Imagine effortlessly conversing with someone across the globe, yet still sounding like yourself. Li et al. [9] take it a step further with FREEVC, a text-free one-shot method, hinting at a future where seamless and ubiquitous voice conversion becomes a reality.

Control and Expression: Shaping the Voice We Hear: The desire for control and expressiveness in voice conversion is evident in the works of Hussain et al. [7] and Ning et al. [8]. The former's ACE-VC method empowers users to manipulate and refine converted speech with self-supervised representations, while the latter's EXPRESSIVE-VC utilizes attention fusion to achieve highly nuanced and expressive conversions. These advancements pave the way for customizable voice avatars that reflect your unique personality, personalized storytelling experiences that immerse you in the narrative, and even emotionally intelligent conversational AI that can truly understand and respond to your feelings.

Safeguarding Privacy in a Voice-Driven World: As voice technologies become increasingly integrated into our lives, the issue of privacy becomes paramount. Srivastava et al. [11] address this crucial concern by investigating security vulnerabilities in voice conversion and proposing countermeasures. This research ensures responsible development and deployment of these technologies, protecting user privacy and fostering trust in a voice-driven world. Imagine interacting with voice assistants or using voice-based authentication systems without fearing the misuse of your voice data.

In conclusion, this expanded exploration unveils a vibrant and diverse research landscape where speech analytics and voice conversion are not just tools, but gateways to deeper understanding and richer communication. From societal applications to personalized experiences, and from emotional expression to nuanced control, the potential is vast and continues to evolve. As researchers delve deeper, we can expect even more transformative advancements in the ways we interact with and analyze spoken language, unlocking the full potential of the human voice in our digital world. And who knows, the future might hold conversations with AI companions that feel as natural as talking to a friend, or the ability to effortlessly share your stories and emotions with the world, all through the power of your unique 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.

IV. Results

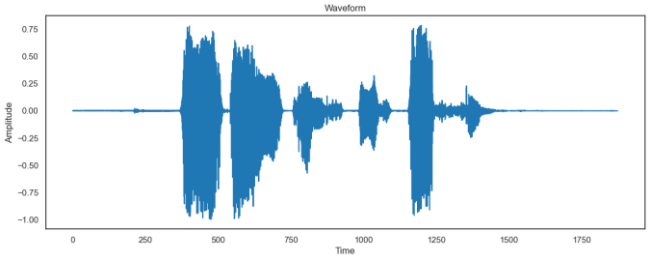

Fig. 1. The waveform “AI in Speech Processing”

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.

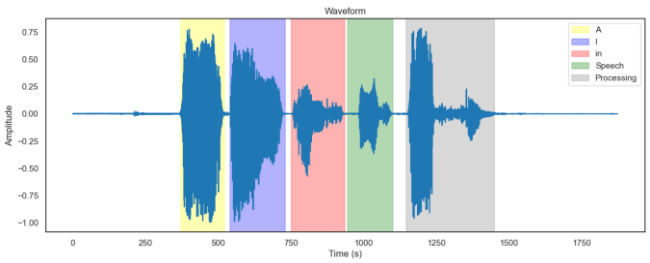

Fig. 2. Every word in the waveform being Highlighted

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

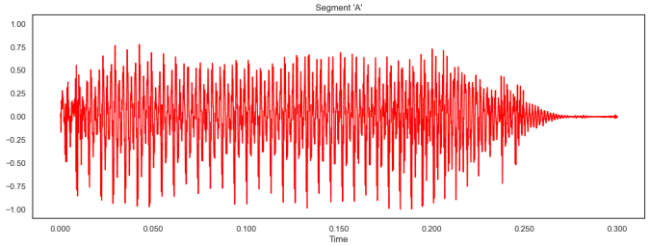

Fig. 3. A zoomed in segment of the audio for “A”

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.

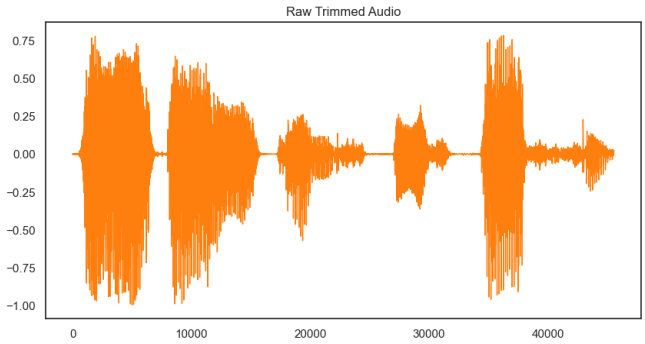

Fig. 4. Trimmed audio wave

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

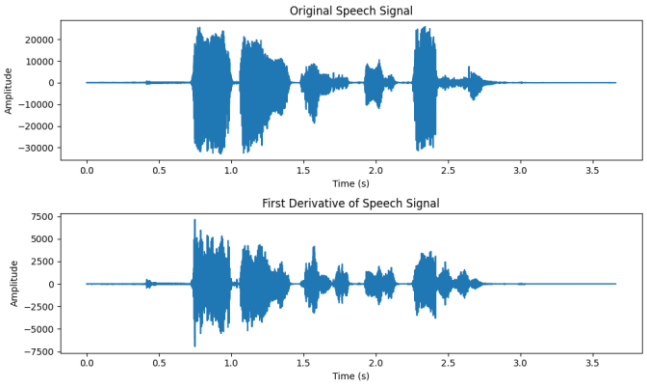

Fig. 5. The Original speech and its derivative

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

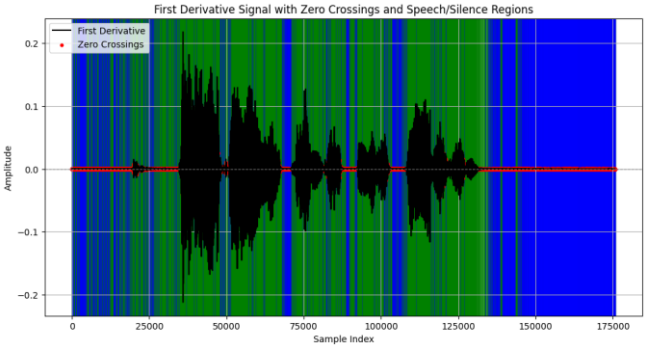

Fig. 4. Zero Crossing and Silent zones

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.

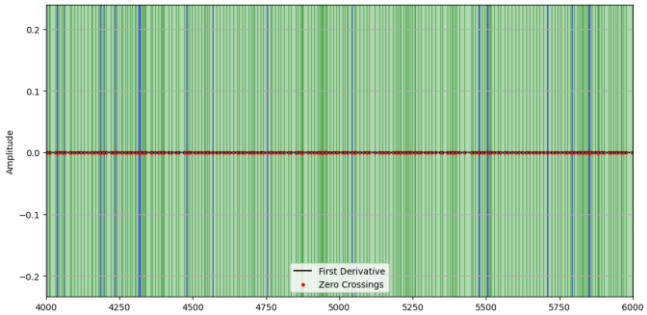

Fig. 5. Zoomed in Zero Crossing and Silent zones.

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

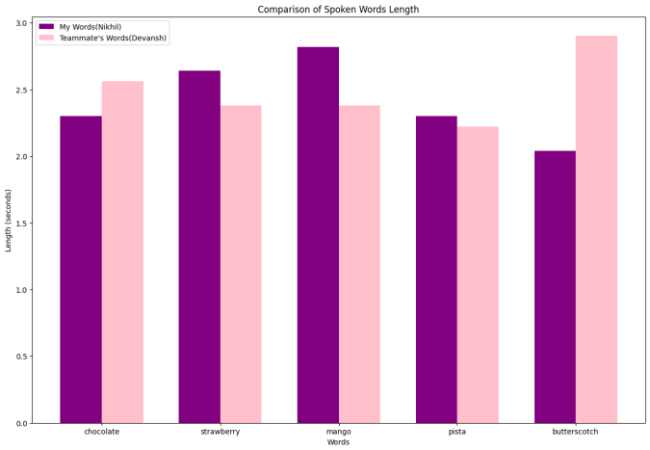

Fig. 6. Word length comaprisons

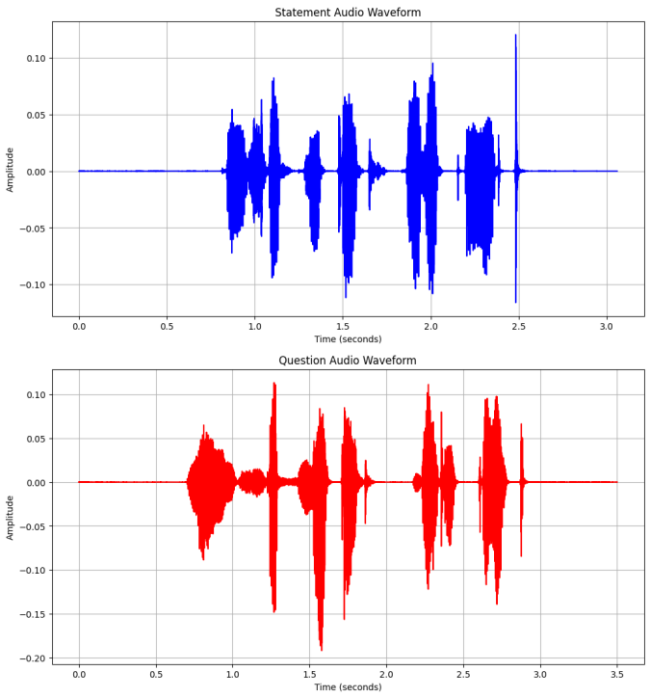

Fig.7. Comparing the Statement “You Submitted The report”

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.

REFERENCES

[1] Shah, N., Singh, M., Takahashi, N., & Onoe, N. (2023). Nonparallel Emotional Voice Conversion for Unseen Speaker-Emotion Pairs Using Dual Domain Adversarial Network & Virtual Domain Pairing.

[2] Farkhadov, M., Smirnov, V., & Eliseev, A. (2017). Application of Speech Analytics in Information Space Monitoring Systems.

[3] Sun, H., Wang, D., Li, L., Chen, C., & Zheng, T. F. (2023). Random Cycle Loss and Its Application to Voice Conversion.

[4] Li, Y. A., Han, C., & Mesgarani, N. (2023). STYLETTS-VC: One-Shot Voice Conversion by Knowledge Transfer from Style-Based TTS Models.

[5] Scheidt, S., & Chung, Q. B. (2019). Making a Case for Speech Analytics to Improve Customer Service Quality: Vision, Implementation, and Evaluation.

[6] Nazir, F., Majeed, M. N., Ghazanfar, M. A., & Maqsood, M. (2023). A Computer-Aided Speech Analytics Approach for Pronunciation Feedback Using Deep Feature Clustering.

[7] Hussain, S., Neekhara, P., Huang, J., Li, J., & Ginsburg, B. (2023). ACE-VC: Adaptive and Controllable Voice Conversion Using Explicitly Disentangled Self-Supervised Speech Representations.

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

[9] Li, J., Tu, W., & Xiao, L. (2023). FREEVC: Towards High-Quality Text-Free One-Shot Voice Conversion.

[10] Sisman, B., Yamagishi, J., King, S., & Li, H. (2021). An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning.

[11] Srivastava, B. M. L., Vauquier, N., Sahidullah, M., Bellet, A., Tommasi, M., & Vincent, E. (2020). Evaluating Voice Conversion-Based Privacy Protection Against Informed Attackers.

[12] Hildebrand, C., Efthymiou, F., Busquet, F., Hampton, W. H., Hoffman, D. L., & Novak, T. P. (2020). Voice Analytics in Business Research: Conceptual Foundations, Acoustic Feature Extraction, and Applications.