Pronunciation Perfect: A Smart Pronunciation Evaluation System

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

Ch. Partha Sai Saketh
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 Project *strives to redefine language learning by offering personalized feedback on pronunciation accuracy. Through cutting-edge speech recognition technology, the system analyses user speech against correct pronunciations, providing tailored exercises for improvement. With its intuitive interface and adaptive learning features, Pronunciation Perfect aims to enhance language proficiency effectively and engagingly.***

Keyword: ***Speech-Based, Emotion-Enhanced Learning and Assessment System, Pronunciation Perfect, language learning, pronunciation accuracy, speech recognition, adaptive learning, real-time feedback, personalized exercises***

I. INTRODUCTION

Pronunciation Perfect, an integral part of the broader "Speech-Based Emotion-Enhanced Learning and Assessment System," zeroes in on refining pronunciation—a cornerstone of language mastery. The project's core objective is to furnish users with a comprehensive toolset for real-time pronunciation enhancement.

In today's interconnected world, effective communication stands as a linchpin skill. Pronunciation Perfect rises to this challenge by delivering instantaneous feedback on pronunciation, empowering users to refine their language skills across diverse scenarios, spanning from language acquisition to public speaking endeavours.

At the heart of Pronunciation Perfect lies state-of-the-art speech recognition algorithms. These algorithms scrutinize user speech, juxtaposing it against correct pronunciations to pinpoint areas necessitating improvement. Leveraging adaptive learning methodologies, the system tailors exercises based on individual performance metrics.

The system commences by capturing user speech input, followed by processing via sophisticated speech recognition algorithms. Phonetic representations are identified, then cross-referenced with correct pronunciations to furnish personalized feedback. Acoustic modelling, language modelling, and machine learning algorithms bolster accuracy and adaptability.

Acoustic modelling empowers the system to discern between diverse speech sounds, while language modelling ensures context-sensitive evaluation. Machine learning algorithms dynamically adjust feedback based on user performance, facilitating continuous skill enhancement.

Pronunciation Perfect heralds a paradigm shift in language learning and assessment, empowering users to attain proficiency in spoken language skills. With its user-friendly interface and adaptive learning capabilities, the system presents a transformative learning journey, tailored to meet the diverse needs of language learners across the globe.

II. LITERATURE SURVEY

In this paper [1], the authors introduce an innovative approach to speech recognition by blending data augmentation and ensemble methods, leading to enhanced accuracy. They employ feature perturbation for data augmentation, augmenting the training dataset, and employ ensemble techniques to integrate multiple models. This novel system proves effective in real-world French ASR tasks, showcasing its potential for improving speech recognition performance in practical applications.

The paper [2] introduces a language learning app for low-cost tablets in rural India, addressing the lack of teachers and technology. It uses speech recognition to give instant feedback on pronunciation, even without internet access. Learners can progress at their own pace and fix mistakes on the go, making learning engaging and motivating. Features like recognizing longer sentences and providing visual feedback enhance the learning experience. Future plans include improving support for Indian languages like Hindi and Malayalam.

This research [3] is based on innovation in speech recognition which has experienced through the experiment on nano physics specific quantum convolutional neural network (QCNN) that we have made. The paradigm is a new one that get the best out of Quantum as well as classical protocol

treating inter alia the issues of scalability and accuracy of voice recognition apps. The ability of the quant filters to reduce the input strength as such is confirmation that quantum computers have what it takes upstage the speech recognition as well as the text converting, and their potential is vast. Furthermore, the second part of the process which is the said layers added to the QCNN model, the tool demonstrated that it is a great candidate for chatbots which are among the very known virtual assistants like Alexa and Siri.

In this research, [4] the suggested multimodal intermediate-level fusion workspace integrates the speech recognition system output and visual movement information. This infrastructure alleviates the deficiencies of the one-way audio speech recognition systems, which are unable to accommodate the hearing-impaired, and they are also unable to handle the disruptive noise and differing pronunciations. It comes up with the analysis for a model based on transformers and turns out that the former has lower WER coefficients in terms of noisy audio than baseline systems. Benchmark datasets LRS2 and Grid are used for the evaluation of this multimodal approach because it decreases the WER significantly and provide better insights than the unimodal methods on some datasets. Finally, the future studies would use unlabelled data and knowledge distillation methodology alongside others to also improve the performance of the model.

This work [5] aims at a speech emotion recognition system, utilising different machine learning algorithms which are coupled to deep learning classifiers to improve the accuracy and tolerance of the system. To improve the recognition system and make it more generalised and stronger, 3 databases are used namely Berlin, SAVEE and TESS. It is seen that the model which involves CNN and LSTM accounts for 94% accuracy among all other classifiers suggesting its potency of classifying emotional speech using data from various databases. The research discloses the way SER can be designed to utilize data fusion and deep learning models which will ultimately improve the accuracy of emotion recognition by AI.

The paper [6] focuses on a spell correction framework that is based on an Automated Speech Recognition (ASR) system developed from a deep recurrent neural network (RNN). This framework is supported by a model that is built using Bidirectional Encoder Representations from Transformers (BERT) in the aim of improving transcription accuracy. Ongoing performance deficiencies of ASR systems, where the job of proper transcription remains a challenging one, push decision makers towards advanced spell-checking solutions. It is methodology of experimenting that evaluates the WER, CER sentence, and BLEU score using three different accent corpora to examine the effect of spell correction with BERT pre-trained model on ASR. The outcomes of the experiments reveal that spelling error detection and correction model is capable to notice and solve spelling mistakes successfully. It attained a considerably low WER in several datasets generated by voxforge, NPTEL, and librispeech corpora.

The paper [7] describes about the modern uses of hidden markov models (HMMs) are typically described in this piece by analysing the current capabilities of the speech recognition system. Voice Quest is an important step forward in digital communication technology. This essay aims to appraise Voice Quest and explain its significance. Simplification and ease of searching are what Voice Quest seeks to achieve. Essentially, we should establish a database that can host some questions, which are typically asked for, as well as their corresponding replies. Every time there is a new query from any user, our database is checked first to see if there exists an exact match before converting the input question into text form. After finding such answer, it is saved inside the database so that when user asks same question voice response could be played back accordingly with it being written like text in the system. All this makes Voice Quest important because it allows for no typing or pressing buttons. Every word must countif you want to maximize your writing's impact.

This research [8] presents a speech augmentation method for automated speech recognition (ASR) systems using a dense convolutional recurrent network (DCRN). It suggests using voice enhancement for two purposes: augmenting data and acting as an ASR preprocessing frontend. The paper uses a three-step training approach for preprocessing and a KL divergence-based consistency loss for data augmentation. Results on an English video dataset from social media show notable gains in ASR performance: an average relative improvement of 11.2% was achieved with data augmentation, 8.3% with preprocessing, and 13.4% when both strategies were combined. The results demonstrate the effectiveness of speech enhancement approaches in enhancing the resilience and accuracy of ASR, especially when used as preprocessing and data augmentation techniques.

The research of the paper [9] is concentrated on identifying speakers in harsh acoustic conditions by putting emphasis on the positives of DNN embeddings and i-vectors. The methodology is concerned with the evaluation of both the single-channel (monaural) and multi-channel speech enhancement approaches. It employs the masking based MVDR beamformers for multi-channel enhancement as well as the convolutional recurrent networks (CRNs) for monaural speech enhancement. Results of experiment explicitly show that the complex summarization of the spectrum employs convolutional recurrent networks jointly with gammatone frequency cepstral coefficients (GFCCs) leads to drastic decrease in the speaker verification errors for the systems of both i-vector and x-vector. Continually, spear in smooth weight \(x\) performance by multi-channel speech augmentation; this error diminishing is more apparent in a rank-1 estimation of the MVDR beamformer, hampered only by accurate steering vector estimation. The research paper highlights the importance of speech hostile augmentation methods in effective speaker recognition under noisy conditions.

In this paper [10], you will find a comprehensive review of the deep learning methods applying to audio-visual speech

enhancement and separation. It uses various aspects, such as training goals, evaluation methods, techniques of fusion, sound and visual features, and deep learning techniques. This illustrates the application of multimodal sources as well as the effectiveness of data-driven approaches. The authors also uncover strategies for audio-visual sound source separation and speech restoration from silent videos, which can be used for audio-visual speech separation and enhancement. Moreover, the diverse audio-visual speech datasets and evaluation mechanisms including often-employed ones will be paid attention from various angles to compare and evaluate the system performance. In sum, the study possesses a solid understanding of current state of this field highlighting both the progress and the gap as well.

III. METHODOLOGY

Data:

Have used sp2_anvith.wav for all the experiments

Method:

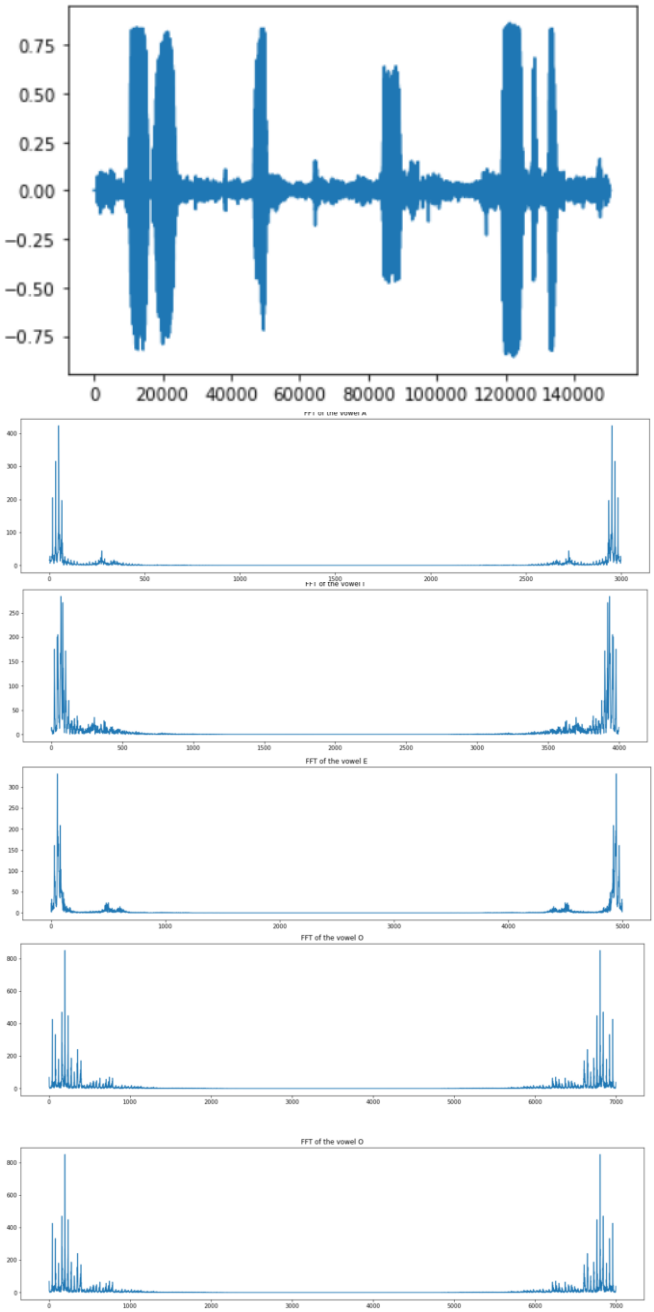
In A1, we aim to analyze the spectral characteristics of vowel sounds by taking a portion of the recorded signal representing a vowel sound. Firstly, we load the audio file using the **librosa** library and extract snippets corresponding to different vowel sounds. Then, we perform Fast Fourier Transform (FFT) on each snippet to obtain its frequency domain representation. By plotting the amplitude spectrum of the FFT results, we visualize the frequency content of each vowel sound.

For A2, we follow a similar approach as in A1, but this time focusing on consonant sounds. We extract portions of the recorded signal representing different consonant sounds, perform FFT on each snippet, and plot the resulting amplitude spectrum. This allows us to observe the spectral characteristics specific to consonant sounds.

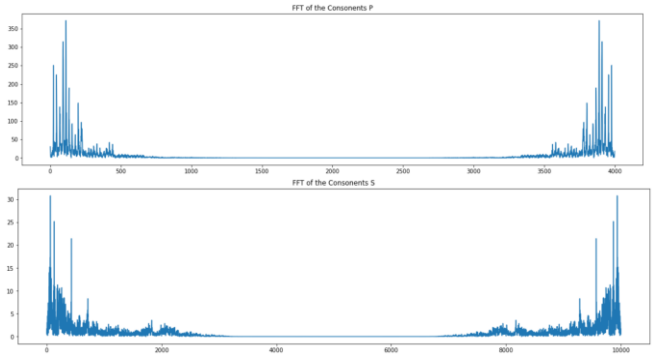
In A3, our objective is to analyze slices of silence and non-voiced portions of the recorded speech signal. We extract these portions from the audio file, perform FFT on each snippet, and visualize their amplitude spectra. This helps us understand the spectral characteristics associated with silence and non-voiced segments in speech signals.

Finally, in A4, we generate a spectrogram of the entire speech signal to observe the change points associated with different speech segments, including vowels and consonants. We utilize the **specgram** function to generate the spectrogram, which provides a comprehensive visualization of the frequency content over time in the speech signal.

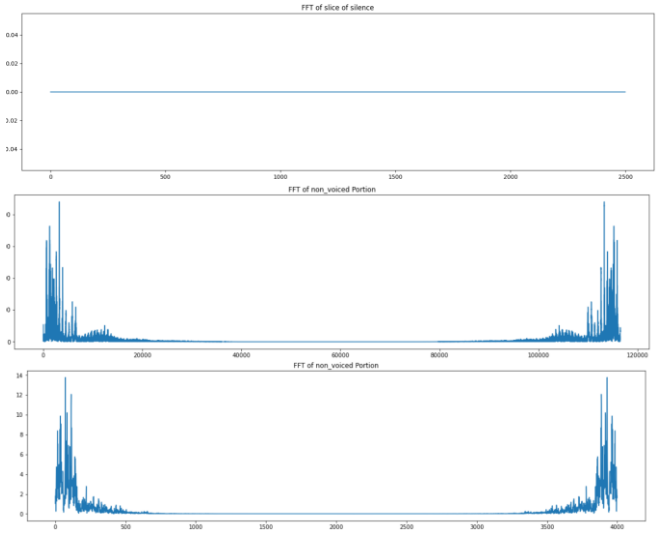
IV. RESULTS



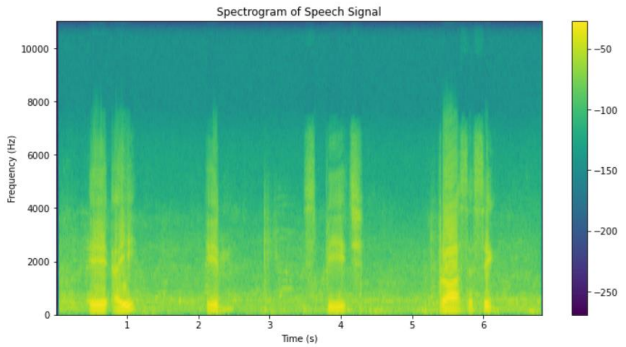
(i) Figure of FFT of Vowels in the audio clip



(ii) Figure of the FFT of the Consonants in the audio clip



(ii) Figure of FFt's of Slice Silence and Non-Voiced Portions



(iv) Figure of Spectrogram with change point of signals

V. CONLCUSION

In conclusion, the analysis conducted in this experiment provides valuable insights into the spectral characteristics of different phonemes in speech signals. By performing FFT on portions of the recorded signal representing vowels and consonants, we observed distinct patterns in their amplitude spectra. Additionally, analyzing slices of silence and non-voiced portions helped in understanding their spectral properties. Furthermore, generating a spectrogram of the entire speech signal enabled us to visualize the distribution of phonemes over time. Overall, this experiment enhances our understanding of speech processing and lays the groundwork for further research in this field.

REFERENCES

[1] Ilyes Rebai, Yessine BenAyed, Walid Mahdi, Jean-Pierre Lorré,"Improving speech recognition using data augmentation and acoustic model fusion",2017,Pages 316-322.

[2] A. Jayakumar, Raghunath, M., S, S. M., S, A., Sadanandan, A., and Prof. Prema Nedungadi, "Enhancing speech recognition in developing language learning systems for low cost Androids", in 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 2016

[3] Thejha B., Yogeswari S., Vishalli A, Jeyalakshmi J., "Speech Recognition Using Quantum Convolutional Neural Network" , Proceedings of 8th IEEE International Conference on Science, Technology, Engineering and Mathematics, ICONSTEM 2023(SCOPUS)

[4] A. Kumar, D. K. Renuka, S. L. Rose and M. C. Shunmugapriya, "Attention based Multi Modal Learning for Audio Visual Speech Recognition," 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Delhi, India, 2022, pp. 1-4,

[5] Sasidharan Rajeswari, S., Gopakumar, G., Nair, M. (2021). Speech Emotion Recognition Using Machine Learning Techniques. In: Sharma, H., Saraswat, M., Yadav, A., Kim, J.H., Bansal, J.C. (eds) Congress on Intelligent Systems. CIS 2020. Advances in Intelligent Systems and Computing, vol 1335, pp 169–178, 2020.

[6] Shunmuga Priya, M.C., Karthika Renuka, D., and Ashok Kumar, L. 'Towards Improving Speech Recognition Model with Post-processing Spell Correction Using BERT'. 1 Jan. 2022 : 4873 – 4882, IoS Press

[7] T. S. Sarika, Sreekumar, S., and A. G. Hari Narayanan, "Enhancement of speech recognition (Voice quest)", International Journal of Applied Engineering Research, vol. 10, pp. 708-711, 2015.

[8] A. Pandey, C. Liu, Y. Wang and Y. Saraf, "Dual Application of Speech Enhancement for Automatic Speech Recognition," 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 2021, pp. 223-228

[9] H. Taherian, Z. -Q. Wang, J. Chang and D. Wang, "Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp.

[10] Michelsanti, Daniel, et al. "An overview of deep-learning-based audio-visual speech enhancement and separation." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021):