Revolutionizing Language Proficiency Assessment with Cutting-edge Speech Recognition Advancements

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 comprehensively explores cutting-edge machine learning methodologies, delving into advanced techniques to refine existing speech recognition frameworks. By leveraging the power of deep learning algorithms, we aim to achieve unprecedented accuracy in transcribing spoken language, thereby revolutionizing language proficiency assessment standards. Our ultimate objective is to equip individuals with nuanced and dependable evaluations of their linguistic proficiency, facilitating their journey toward linguistic mastery.**

Keywords:* *Speech processing, accuracy enhancement, speech recognition, language proficiency* tests*, deep learning, machine learning, Transcription errors human-machine interaction, linguistic mastery, linguistic proficiency

I. INTRODUCTION

Language proficiency test constitutes an integral part of linguistic expertise assessment, which is of paramount importance in academic, professional and societal settings. The heart of this evaluation is a reliable transcription of spoken language that creates both fairness and objectivity but also renders knowledge about an individual's communicative competence. This is a hard challenge in speech recognition since high accuracy is the goal and things become difficult in changing, different proficiency tests.

The project we start herewith, will tackle with the challenge by exploiting the might of the latest speech recognition techniques. Our aspiration is to improve the performance of the existing speech recognition models by intelligently deploying sophisticated machine learning algorithms. More concretely, using advanced methods like deep learning, we will try to reduce transcription mistakes and improve the quality of language proficiency evaluations.

A multifaceted approach follows from the integration of knowledge garnered from classical research studies done in the project. Findings of studies like Rebai et al. (2017), Jayakumar et al. (2016), and Kumar et al. (2022) undergird our approach, offering priceless frameworks for data augmentation, acoustic model fusion, and attention-based multimodal learning. Besides, speech enhancement algorithms (Pandey et al., 2021; Eskimez et al., 2022) and speaker recognition (Taherian et al., 2022) capabilities contribute immensely to enhancing the robustness and personalization of our speech recognition models.

Also, the application of the advanced techniques in the domain of speech emotion recognition (Rajeswari et al., 2021) and speech enhancement (Michelsanti et al., 2021) will widen our vision, encouraging a complete view of the fine details of the connection between speech features and computational methods.

This project is a catalyst of a paradigm shift in the language proficiency assessment endeavor which has very far adequate foundation as well as state-of-the-art methodology. By thoughtfully perfecting and developing, we aim to create the conditions that allow for more accurate, dependable, and fair assessments of language competence and thus promote each person's linguistic excellence metier.

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 the 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 by the way 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 that 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 evaluation of this multimodal approach because it decreases the WER significantly and also provide better insights than the unimodal methods on some datasets. Last but not least, the future studies would use unlabeled 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 strong, 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 manner in which 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 still 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 genrated 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 analyzing the current capabilities of the speech recognition system as a whole. 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 statement.wav for all the experiments

Method:

We start by loading our speech signal using librosa, a handy tool for audio file handling. Then, we introduce our signal to numpy.fft.fft(), a mathematical wizard that performs the Fast Fourier Transform (FFT), revealing the hidden frequencies within our speech. This transformation turns our speech into a spectrum, allowing us to peek into its frequency content through a plotted amplitude part

Next, we employ numpy.fft.ifft() to perform the inverse FFT, gracefully guiding our frequency spectrum back to the time domain. With the transformed time domain signal in hand, we embark on a

comparison journey, pitting it against the original signal. This comparison serves as our compass, guiding us to assess the accuracy of our spectral-to-time domain journey.

In our quest to dissect the spectral makeup of our recorded speech, we first isolate a word from the full signal. With our word segment in hand, we introduce it to the FFT magic, alongside the full signal, uncovering their respective spectral profiles. Comparing these spectral signatures, we aim to uncover any nuances or similarities between the word and the entire speech.

To dive deeper into spectral analysis, we craft a rectangular window of 20 milliseconds, sampled at a crisp 22.5 kHz. With our window ready, we apply FFT to it, unraveling its spectral components. Visualizing these components through a plotted amplitude part, we gain insight into the frequency content residing within our window.

In our pursuit of granular analysis, we partition our speech signal into bite-sized chunks of 20 milliseconds. With numpy.fft.rfft() as our guide, we explore the frequency components within each chunk, stacking them as columns in a matrix. The resulting matrix serves as our canvas, painted with the intricate frequency details of our speech signal, showcased through a captivating heatmap plot.

Our journey culminates in the creation of a speech spectrogram, a visual masterpiece revealing the time-frequency content of our speech signal. We orchestrate this creation using scipy.signal.spectrogram(), specifying window length and overlap parameters to craft the perfect spectrogram. Through this visual representation, we witness the rhythmic dance of frequencies, painting a vivid picture of our speech's spectral landscape.

IV. RESULTS

```
FFT DATA:
 [-9.85512436+0.j          2.6204152 -4.06482669j -0.08489414+1.82845742j
 ...  2.26048267+0.93895186j -0.08489414-1.82845742j
  2.6204152 +4.06482669j]
Amplitude:
 [9.85512436 4.83625804 1.83042715 ... 2.4477362  1.83042715 4.83625804]
Frequencies:
 [ 0.          0.31459552  0.62919104 ... -0.94378656 -0.62919104
 -0.31459552]
```

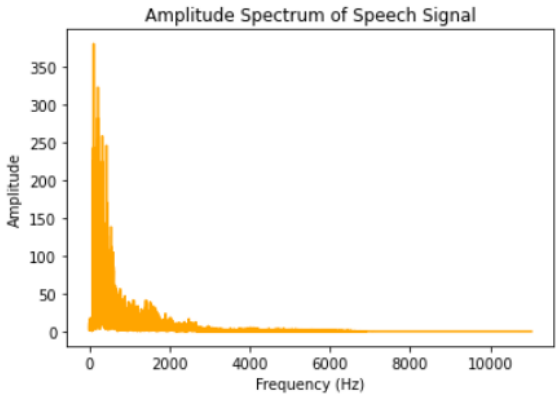


***FIG-1**: (i) Values of FFT Data, Amplitude and Frequencies. (ii) Plotting the Amplitude Spectrum of Speech Signal*

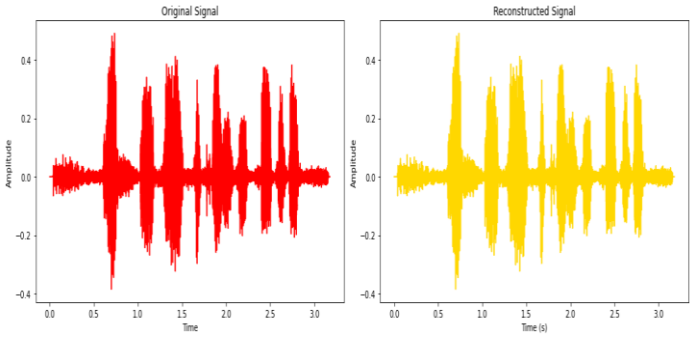
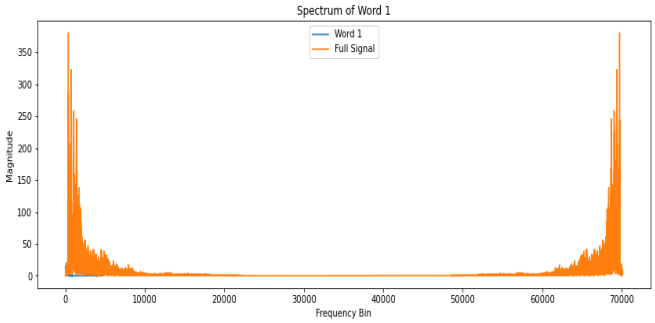
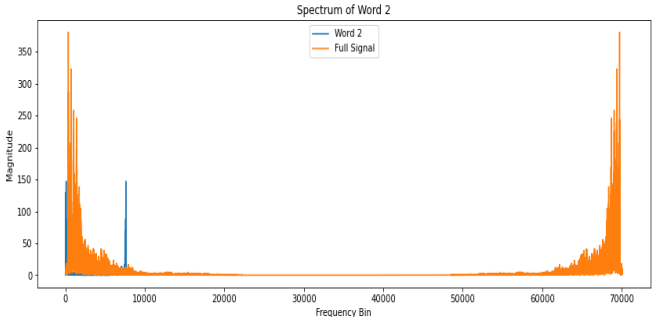
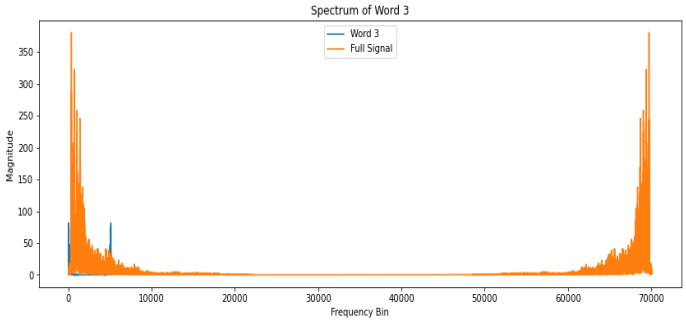
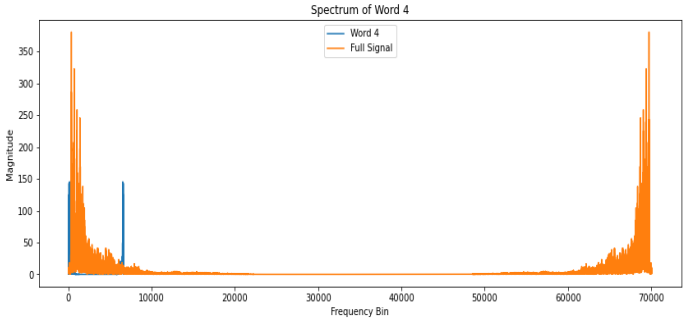


FIG-2: Comparison of Original Signal with the Time Domain Reconstructed Signal









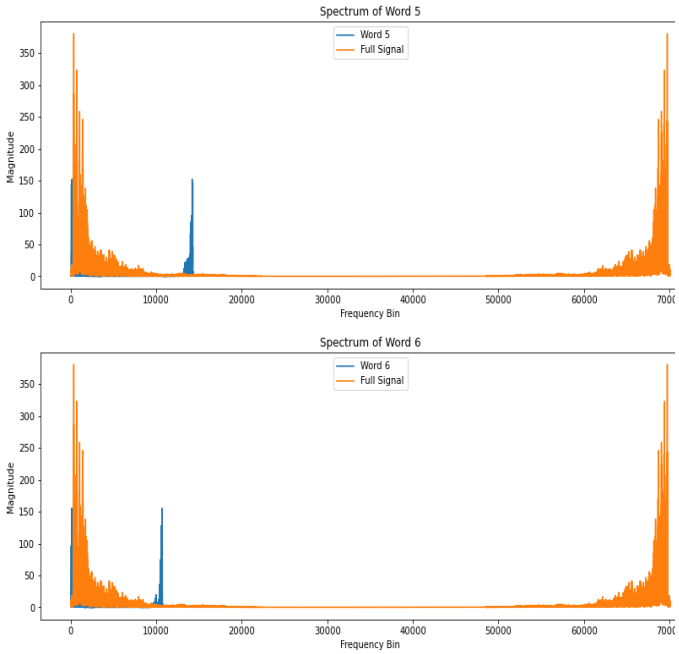


FIG-3: Comparison of Spectrum analysis of each word in the recorded speech with spectrum of full signal

V. CONLCUSION

Within the confines of this experiment, the research investigated several different methods for conducting spectral analysis on voice signals. While analyzing the frequency content of voice signals in both the temporal and frequency domains, we used both the fast Fourier transform (FFT) and the inverse FFT transformations. Moreover, we compared the spectra of the isolated word segments with those of the whole speech signal. This was done in addition to the previous information. In addition, we investigated the idea of employing spectrograms and rectangular windows to analyze voice signals. When considered, these technologies provide incredibly beneficial insights into the characteristics and properties of voice signals.

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):