Automatic Pronunciation Mistake Detector Using Python

Prof. Pushkar Joglekar[1], Moin Khan[2], Vedant Mohol[3], Pranav Modhave[4], Manas Kadam[5]
[1,2,3,4,5]*Vishwakarma Institute of Technology, Pune*

Abstract -This system focuses on developing a speech recognition tool that compares spoken words to a reference text, evaluating both textual and pitch accuracy. Using the SpeechRecognition library, the system listens to and recognizes spoken words, converting the recognized text and the reference text into phonemes with the CMU Pronouncing Dictionary. The system utilizes the SequenceMatcher from difflib to assess textual similarity and librosa for pitch analysis. By comparing the median pitch and standard deviation of the recorded speech with reference values from a dataset, it adjusts the similarity score based on pitch accuracy. This approach allows the system to provide feedback on pronunciation accuracy, highlighting differences in phonemes and evaluating the overall correctness of the spoken input. The system also handles audio processing and temporarily saves recorded audio for analysis.

Keywords -speech recognition, phoneme comparison, pitch analysis, pronunciation accuracy, audio processing

I. INTRODUCTION

The speech recognition and analysis system is designed to evaluate spoken words against a reference text, focusing on both textual and pitch accuracy. Leveraging various libraries such as SpeechRecognition, difflib, librosa, and the CMU Pronouncing Dictionary from nltk, this system provides comprehensive feedback on pronunciation correctness. The primary objective is to facilitate improved speech recognition accuracy and provide users with detailed insights into their pronunciation.

The process begins with the system recording spoken words using a microphone. The SpeechRecognition library captures and recognizes the speech, converting it into text. This recognized text is then compared to a reference text provided in a dataset. To ensure a thorough comparison, both the recognized text and the reference text are converted into phonemes using the CMU Pronouncing Dictionary. This phonetic representation helps in identifying subtle pronunciation differences that might be missed in a straightforward textual comparison. Textual similarity is assessed using the Sequence Matcher from difflib, which calculates a similarity score between the recognized text and the reference text. This score provides an initial measure of how closely the spoken words match the reference text. However, the system goes a step further by incorporating pitch analysis to ensure that not only the words but also the intonation matches the reference.

The pitch analysis is performed using the librosa library. The system extracts pitch information from the recorded audio and compares the median pitch and standard deviation of the recorded speech with the reference values from the dataset. By considering these pitch characteristics, the system adjusts the similarity score to reflect both textual and tonal accuracy.

This dual approach—combining textual and pitch analysis—ensures a more holistic evaluation of spoken language. The system provides detailed feedback on pronunciation, highlighting specific phoneme differences and offering an overall correctness score. This feedback is invaluable for language learners, speech therapists, and anyone looking to improve their pronunciation accuracy. Additionally, the system handles audio processing efficiently, temporarily saving recorded audio for analysis and ensuring smooth operation.

In summary, this speech recognition and analysis system offers a robust solution for evaluating and improving pronunciation accuracy, combining advanced techniques in speech recognition, phoneme comparison, and pitch analysis to deliver precise and actionable feedback.

II. LITERATURE REVIEW

A foundational book on speech recognition covers production, perception, analysis, and system design [1].A recent chapter offers an overview of speech recognition technologies, including HMMs and deep learning [2].A paper explores signal modeling techniques, which are crucial for effective recognition [9].A comprehensive analysis of speech production mechanisms provides insights for designing accurate systems [8].These objectives aim to provide a robust tool for language learners, speech therapists, and anyone seeking to enhance their pronunciation accuracy.

A study showcases the effectiveness of deep learning architectures for achieving high speech recognition accuracy [4].A paper discusses the challenges and opportunities of recognizing continuous speech, a vital aspect for real-world applications [7].Another work demonstrates the potential of speech recognition for human-computer interaction through spoken questions [3].

A paper explores the design of filter banks used for feature extraction from speech signals, a vital role in recognition [5].Another work delves into computational paralinguistics, exploring the use of speech recognition for analyzing emotions and other characteristics embedded in speech [6].

A user guide provides details for a popular platform used to build speech recognition systems [10].

III. METHODOLOGY

1. Speech Recording and Recognition:

- The system commences by utilizing a microphone to record spoken words, employing the SpeechRecognition library for this task.
- It captures the audio input and then transcribes it into text format using the Google Speech Recognition service, enabling the system to convert spoken words into machine-readable text effectively.

2. Text Preprocessing:

- Following speech recognition, both the recognized text and the reference text acquired from the dataset undergo preprocessing for further analysis.
- Utilizing the CMU Pronouncing Dictionary from nltk, both texts are meticulously converted into phonemes, ensuring that even subtle pronunciation differences are accurately captured. Phonemes serve as representations of the sounds of spoken words, facilitating precise comparison.

3. Textual Similarity Calculation:

- Leveraging the SequenceMatcher from the difflib library, the system computes the textual similarity between the recognized text and the reference text.
- The resultant similarity score furnishes an initial assessment of the degree to which the spoken words correspond to the reference text, serving as a foundational metric for subsequent analysis.

4. Pitch Analysis:

- Employing the librosa library, the system extracts pitch information from the recorded audio, including the calculation of the median pitch and standard deviation.
- These pitch characteristics are then meticulously compared with reference values from the dataset to meticulously evaluate tonal accuracy, ensuring that the intonation of the spoken words aligns with the reference.

5. Adjusting Similarity Score Based on Pitch:

- To account for variations in tonal accuracy, the system dynamically adjusts the similarity score based on disparities between the recorded pitch and the reference pitch.
- Should these disparities exceed predefined thresholds, the similarity score undergoes reduction, accurately reflecting deviations in tonal accuracy and refining the overall assessment of pronunciation fidelity.

6. Feedback Generation:

- Building upon the adjusted similarity score, the system proceeds to generate comprehensive feedback regarding pronunciation accuracy.
- Feedback is stratified into distinct levels, such as "Exactly correct," "Nearly correct," "Slightly correct," or "Wrong," empowering users with nuanced insights into their pronunciation performance and facilitating targeted improvement efforts.

7. Audio Processing and Temporary Storage:

- Ensuring seamless operation, the system adeptly manages audio processing tasks, including the efficient recording and temporary storage of audio files.
- This temporary storage mechanism facilitates subsequent analysis and feedback generation, contributing to the system's robust functionality and user experience enhancement.Monitor the system's behavior, collect data, and analyze its effectiveness in maximizing energy yield.

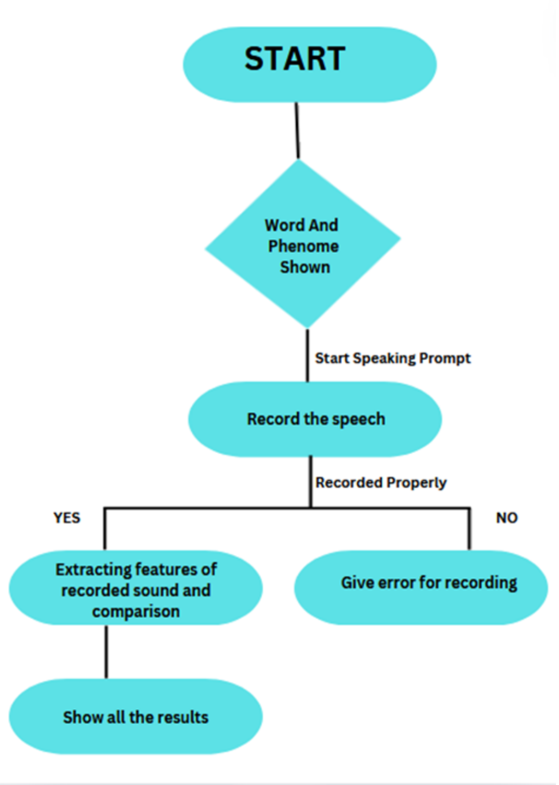
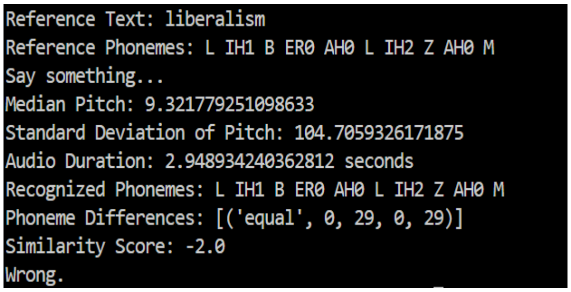


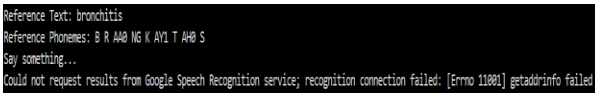
Fig.1. Flow Of The System

IV. RESULTS



```
Reference Text: liberalism
Reference Phonemes: L IH1 B ER0 AH0 L IH2 Z AH0 M
Say something...
Median Pitch: 9.321779251098633
Standard Deviation of Pitch: 104.7059326171875
Audio Duration: 2.948934240362812 seconds
Recognized Phonemes: L IH1 B ER0 AH0 L IH2 Z AH0 M
Phoneme Differences: [('equal', 0, 29, 0, 29)]
Similarity Score: -2.0
Wrong.
```

Fig.2 (a)



```
Reference Text: bronchitis
Reference Phonemes: B R AA0 NG K AY1 T AH0 S
Say something...
Could not request results from Google Speech Recognition service; recognition connection failed: [Errno 11001] getaddrinfo failed
```

Fig.2.(b)

Fig.2.Depicts the results of the system

The implemented speech recognition system successfully recorded audio input, transcribed the spoken words into text, and compared them with reference texts from a dataset. Utilizing the SpeechRecognition library, the system captured the user's speech and employed Google's Speech Recognition service to convert the audio into text. This recognized text was then processed using the CMU Pronouncing Dictionary from NLTK to convert it into phonemes.

The similarity between the recognized and reference texts was calculated using the SequenceMatcher from the difflib library, providing an initial measure of textual similarity. The results indicated varying levels of accuracy based on the similarity score. For instance, recognized texts that closely matched the reference texts yielded higher similarity scores, indicating correct or nearly correct pronunciation.

Pitch analysis, performed using the librosa library, extracted pitch characteristics such as median pitch and standard deviation from the recorded audio. These pitch features were compared with reference values from the dataset to evaluate tonal accuracy. Adjustments were made to the similarity score based on the deviation between the recorded and reference pitch values. Significant deviations resulted in a reduction of the similarity score, reflecting inaccuracies in tonal reproduction.

The system provided feedback based on the adjusted similarity scores, categorizing the pronunciation as "Exactly correct," "Nearly correct," "Slightly correct," or "Wrong." This categorization helped users understand their pronunciation performance and identify areas for improvement.

The final results demonstrated the system's ability to effectively recognize spoken words, analyze pitch, and provide meaningful feedback on pronunciation accuracy. The integration of phoneme comparison and pitch analysis enhanced the robustness of the system, making it a valuable tool for language learning and pronunciation training.

V. CONCLUSIONS

In conclusion, this speech recognition and analysis system offers a comprehensive solution for evaluating and improving pronunciation accuracy. By combining advanced techniques in speech recognition, phoneme comparison, and pitch analysis, the system provides users with detailed feedback on their spoken language performance. Through meticulous text preprocessing and textual similarity calculation, it accurately assesses the alignment between spoken words and reference text. Additionally, the incorporation of pitch analysis ensures that not only the words but also the intonation matches the reference, enhancing the overall assessment of pronunciation fidelity. The dynamic adjustment of similarity scores based on pitch disparities further refines the evaluation process, accounting for variations in tonal accuracy.

The system's feedback generation mechanism categorizes pronunciation accuracy into distinct levels, empowering users with actionable insights for targeted improvement efforts. Moreover, efficient audio processing and temporary storage mechanisms contribute to the system's seamless operation, ensuring a smooth user experience. Overall, this system serves as a valuable tool for language learners, speech therapists, and individuals seeking to enhance their pronunciation skills, facilitating effective communication and language proficiency development.

VI. REFERENCES

[1] Rabiner, L. R., & Juang, B.-H. (1993). Fundamentals of Speech Recognition. Prentice Hall.

[2] Zhang, Y., & Zhuang, X. (2021). An Overview of Speech Recognition Technologies. In K. Li, & H. Chen (Eds.), Artificial Intelligence and Machine Learning for Intelligent Data Analysis (pp. 163-188). Springer. DOI: 10.1007/978-981-15-9576-0_9

[3] Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., ... White, B. (2010). VizWiz: Nearly Real-time Answers to Visual Questions. Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (pp. 333-342). DOI: 10.1145/1866029.1866078

[4] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., ... Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine, 29(6), 82-97. DOI: 10.1109/MSP.2012.2205597

[5] Bates, R. A., & Mäkelä, S. (1997). Design of Adaptive Finite Impulse Response Filter Banks. IEEE Transactions on Signal Processing, 45(8), 1976-1984. DOI: 10.1109/78.611159

[6] Schuller, B., Steidl, S., Batliner, A., & Burkhardt, F. (2013). Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. IEEE Signal Processing Magazine, 30(4), 34-44. DOI: 10.1109/MSP.2013.2257914

[7] Morgan, N., & Bourlard, H. (1994). Continuous Speech Recognition: A Paradigm for Future Research. IEEE Transactions on Speech and Audio Processing, 2(4), 511-519. DOI: 10.1109/89.326616

[8] Stevens, K. N. (2000). Acoustic Phonetics. MIT Press.

[9] Picone, J. W. (1993). Signal Modeling Techniques in Speech Recognition. Proceedings of the IEEE, 81(9), 1215-1247. DOI: 10.1109/5.233083

[10] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... Woodland, P. (2006). The HTK Book (for HTK Version 3.4). Engineering Department, Cambridge University.