AI-Powered Speech Analysis Tool for Speech Therapists

Akshaya Jayaraj, Sreedevi R Krishnan, Greena Maria Rajan*, Christo Martin,
Elswin P Eldho, Gouri Anil
Department of Computer Science, Adi Shankara Institute of Engineering and Technology, India

* Corresponding author: greenamaria123456@gmail.com
doi: https://doi.org/10.21467/proceedings.7.5.4

ABSTRACT

Speech disorders present a considerable challenge within the realms of clinical diagnosis and therapeutic intervention, thereby necessitating the development of practical tools for the analysis and monitoring of speech patterns. This mini project, entitled ″AI-Powered Speech Analysis Tool for Speech Therapists,″ was developed to integrate advanced technology into the diagnostic process and provide an innovative solution for therapists. The system enabled therapists to record patients' speech, analyze essential acoustic features such as pitch, speech rate, volume, and articulation patterns, and visualize the results through intuitive graphical representations, including spectrograms and pitch contours. It further facilitated comparative analyses against normative speech patterns, identified anomalies, and suggested potential speech disorders based on the derived analyses. Developed using the Python programming language and leveraging libraries such as librosa for audio processing, matplotlib for data visualization, and SQLite for data storage, the tool ensured broad accessibility and expandability. Diagnostic suggestions, derived from either rule-based systems or machine learning models, augmented its utility, while strict adherence to data privacy regulations safeguarded patient confidentiality. The application holds the potential to be extended into a web-based interface employing frameworks like Flask or Django, thereby enhancing its accessibility for therapists. This project demonstrates technical proficiency in speech processing and data visualization and contributes to the field of speech-language pathology by improving the accuracy and efficiency of speech disorder diagnosis. By facilitating the monitoring of patient progress and enhancing therapeutic outcomes, this tool presents opportunities for further advancements in AI-powered healthcare applications.

Keywords: Speech analysis, Speech disorders, AI healthcare tool.

1.Introduction

Language disorders significantly affect substantial segments of the population and present complex clinical diagnostic and treatment challenges due to their inherent variability [1-3]. Traditional methodologies are frequently characterized by time-consuming processes, subjectivity, and reliance on manual analysis alongside the expertise of therapists. These limitations underscore the necessity for automated tools that offer objective, precise, and efficient analyses of linguistic patterns, ultimately enhancing diagnostic and treatment procedures [3]. The proposed system, entitled "AI-Powered Speech Analysis Tool for Speech Therapists," endeavors to address these challenges by integrating advanced language processing techniques with user-friendly visualization tools. This application permits therapists to record and analyze patients' language, focusing on acoustic properties such as pitch, speech rate, volume, and intelligibility. By facilitating comparative analyses that encompass normative language patterns, identification of abnormalities, and recommendations for potential language disorders, the system enhances evidence- based decision- making for therapists. Recent advancements in artificial intelligence (AI) and language processing provide a robust foundation for this work [2] [4]. Libraries such as Librosa for audio functionality extraction, Matplotlib for graphical visualization, and frameworks like Flask and Django facilitate the development of powerful, scalable, and user-friendly tools tailored for clinical applications. Additionally, prior research in this domain has demonstrated the efficacy of AI models in diagnosing and monitoring language disorders, thereby emphasizing the

Proceedings DOI: 10.21467/proceedings.7.5; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-7-1

significance and timeliness of this investigation [2] [4]. The objective of this study is to design a system capable of concurrently recording, analyzing, and presenting language data for diagnostic purposes while ensuring adherence to data protection regulations. Through its contributions to the field of speech-language pathology, the tool aims to enhance the precision and efficiency of clinical diagnoses, improve patient outcomes, and lay the groundwork for further advancements in AI-powered healthcare solutions.

2. Methodology

2.1 Hardware

The system requires a computer or device equipped with either an inbuilt or external microphone to facilitate audio recording, which is essential for analysing speech samples.

2.2 Software Libraries

Several Python libraries were utilized to implement the core functionalities of the tool. pydub was employed for converting audio files to mono and preprocessing them for uniformity. The librosa library played a key role in extracting acoustic features such as pitch, volume, and speech rate, which are critical for speech analysis [5]. For visual representation of the extracted features, matplotlib was used to generate spectrograms and pitch contour plots, enabling easier interpretation of the data.

2.3 Web Framework and Database

To provide a user-friendly web-based interface, the project was developed using the Django web framework, adhering to industry-standard design patterns and best practices [6]. Data storage and management were handled using SQLite, which offers lightweight and secure handling of audio recordings and their corresponding analyses.

2.4 Environment

The entire system was built and tested using Python version 3.12 on a Windows operating system.

3. Analysis

3.1. Audio Recording

The Audio Record Module facilitates the recording of languages directly through a web browser, utilizing the MediaCorder API integrated within web applications. The recorded audio is stored in .WAV format to ensure compatibility with analysis tools.

3.2 Audio Preprocessing

Audio preprocessing involves two key steps. First, audio files are converted to mono format using the Pydub library. This conversion simplifies the analysis and reduces inconsistencies often associated with stereo sound. Second, the recorded audio undergoes normalization to ensure a uniform signal level, effectively minimizing the impact of background noise on subsequent analysis [7].

3.3 Feature Extraction

The sound properties essential for linguistic analysis are employed utilizing the Librosa library [5].

Voice Rate: This is computed by evaluating the duration of voice and silent segments, along with quantifying the syllables produced per second [5].

Volume: It is quantified as the root mean square (RMS) energy of the audio signal [5].

Spectrum features: MEL frequency cepstral coefficients (MFCC) are extracted to facilitate the analysis of speech clarification patterns [2] [8].

3.4 Visualization

The tool provides real-time visualization of the extracted features. Spectrograms, generated using Matplotlib, are employed to illustrate the time-frequency characteristics of the audio [9]. These visualizations offer insights into the analyzed speech, including predictions regarding the presence or absence of speech disorders.

Proceedings DOI: 10.21467/proceedings.7.5
ISBN: 978-81-989164-7-1

Series: AIJR Proceedings
ISSN: 2582-3922

3.5 Data Storage

All recorded audio files are stored in the upload/directory, while the analysis results and visualization outputs are systematically organized within the static/spectrogram/directory. SQLite is utilized to maintain a structured database, thereby ensuring efficient data retrieval and protection of patient information.

3.6 Deployment

The web-based interface has been developed utilizing the Django framework. This tool functions locally on a development server and enables seamless interaction between therapists and the analytical tool, built upon robust web development techniques [6].

4. Theory and Calculation

The system employs Librosa to extract acoustic features, including pitch, Mel-frequency cepstral coefficients (MFCCs), and speech rate. Pitch is computed using the YIN algorithm [10], while MFCCs facilitate the analysis of articulation patterns [8]. Speech rate is assessed by examining the number of syllables articulated within a specified time frame. Subsequently, these features are compared against established speech patterns to identify anomalies.

4.1Mathematical Expressions and Symbols

Pitch Calculation:

$$Pitch = librosa.yin(y, fmin = 50, fmax = 300)$$

MFCC Calculation:

$$MFCC = librosa.features.mfcc(y = y, sr = sr, nmfcc = 13)$$

5. Results and Discussion

The system was evaluated utilizing a diverse range of vocal samples, with the results subsequently visualized through the application of Matplotlib. This tool effectively generates spectrograms and pitch contours, thus offering therapists a lucid visual depiction of vocal patterns [9]. The diagnostic module encompassed precise recommendations grounded in the analysis of characteristics, such as atypical pitch and joint disorders.

The visual outputs, such as the spectrograms and pitch contours generated by the system, are foundational to its utility [9]. These visualizations provide therapists with an objective and readily interpretable representation of a patient's vocal patterns. For instance, irregular patterns in the spectrogram, or significant deviations in the pitch contour from normative ranges, can immediately highlight areas of concern related to articulation, vocal fold vibration, or breath control. This visual clarity supports the therapist's perceptual assessment by providing quantifiable evidence of speech characteristics.

5.1 Figures and Tables

Table 1 provides a comprehensive summary of the testcases executed to validate the core functionalities of the speech analysis tool, ranging from audio file loading to visualization of acoustic features. The successful execution of these tests confirms the system’s operational integrity and its ability to perform intended analyses

Table 1: *Summary of Test Cases*

Test Case	Description	Expected Result
1	Audio File Loading	System successfully loads the audio file.
2	Pitch Detection	The system accurately calculates and prints the mean pitch.
3	Tempo Detection	System accurately calculates and prints the tempo in BPM.
4	Volume Calculation	The system accurately calculates and prints the mean volume in dB.
5	Visualization of Pitch	The system displays a plot showing pitch variation over time.

5.2 Data Pipeline

5.2.1 Data Distribution

The balanced distribution of samples (approximately 2300 for both non-dysarthric (0) and dysarthric (1) labels) depicted in Figure 1 is critical for training a robust machine learning model [11]. This balanced dataset helps prevent bias towards the majority class and ensures that the model learns equally from both types of speech, contributing to its generalizability and accuracy in clinical settings. This dataset was sourced from Kaggle [12] [13].

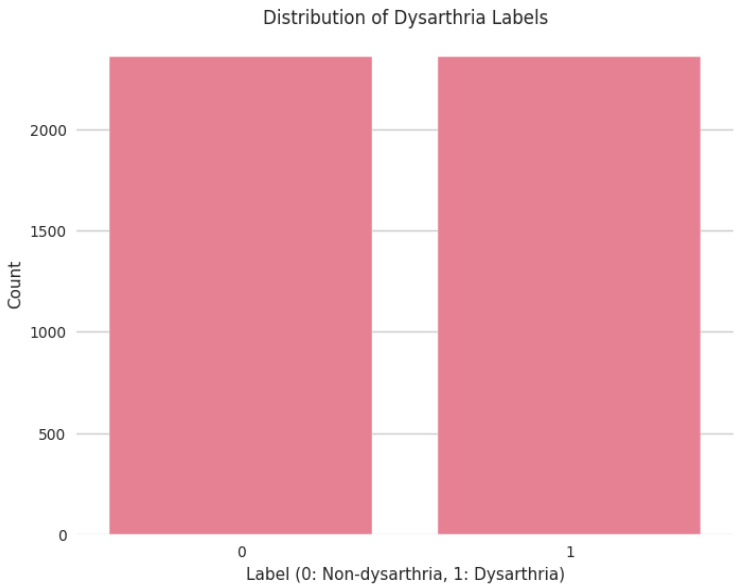

Figure 1: Distribution of Dysarthria Labels showing the balance between non- dysarthric (0) and dysarthric (1) samples

5.2.2 Feature Engineering

The feature importance analysis presented in Figure 2 provides valuable scientific insight into which acoustic properties are most salient for detecting dysarthria using our Random Forest model [14]. The prominence of MFCCs 32, 33 and 38 suggests that specific high-frequency spectral components, or nuanced variations within them, are highly correlated with the presence of dysarthria. This finding

Proceedings DOI: 10.21467/proceedings.7.5
ISBN: 978-81-989164-7-1

Series: AIJR Proceedings
ISSN: 2582-3922

aligns with known speech pathology, where dysarthria can affect precise articulation and vocal tract control, leading to distinct spectral characteristics. Technically, identifying these key features can potentially optimize future models by focusing on more discriminative inputs.

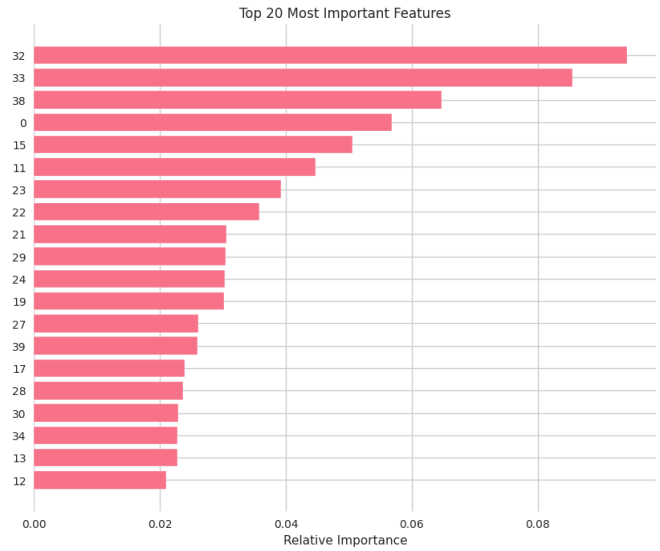

Figure 2: Top 20 Most Important MFCC features, illustrating their relative importance in model predictions

5.2.3 Confusion Matrix

The confusion matrix presented in Figure 3 is a crucial component of the model's performance assessment. It indicated that out of the total test samples, 467 samples were correctly identified as non-dysarthric (True Negatives) and 466 samples were correctly identified as dysarthric (True Positives). There were only 5 instances where non-dysarthric speech was misclassified as dysarthric (False Positives), and 6 instances where dysarthric speech was misclassified as non-dysarthric (False Negatives). This translates to a high overall accuracy, approximately 98%, which aligns with the performance requirements of the tool.

Scientific Significance: The low number of false positives and false negatives is scientifically significant. A low false positive rate (incorrectly diagnosing dysarthria) is crucial in a clinical context to avoid unnecessary interventions and patient distress. Similarly, a low false negative rate (missing a diagnosis of dysarthria) is vital to ensure that patients receive timely and appropriate therapy. This demonstrates the model's ability to discriminate effectively between the two classes.

Technical Significance: From a technical perspective, these results indicate that the Random Forest model, with its ensemble learning approach, is well-suited for this binary classification task on acoustic features [2] [14]. The model's robustness and high accuracy provide strong evidence for its potential utility as an assistive tool for speech therapists, reducing diagnostic subjectivity and enhancing efficiency. The performance suggests that the extracted MFCC features, combined with the Random Forest algorithm, capture sufficient information to reliably identify dysarthric speech patterns.

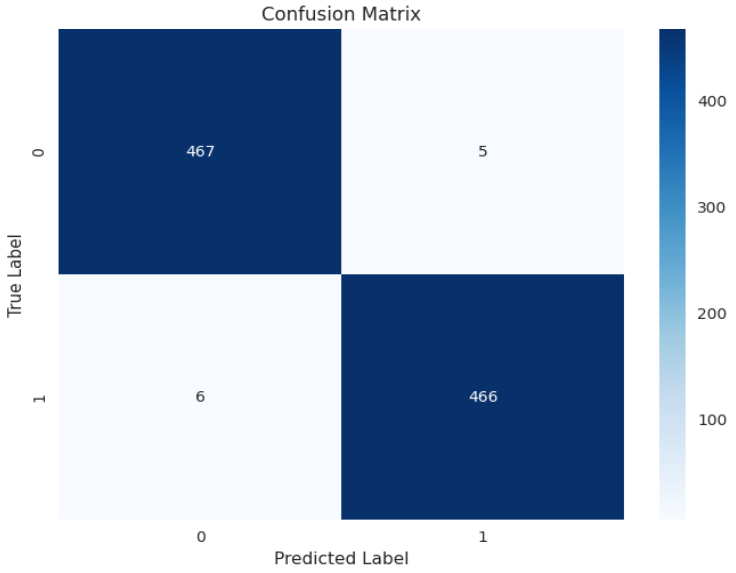

Figure 3: Confusion Matrix for Model Performance Assessment. This matrix displays the true positive, true negative, false positive, and false negative classifications of Random Forest model on the test dataset.

6 . Conclusions

AI-driven speech analysis tools offer a reliable and efficient solution for speech therapists in diagnosing and analyzing speech disorders. By automating the analysis of acoustic features and providing intuitive visual representations, these tools enhance the accuracy and efficiency of diagnosis. The system's adaptability to web-based interfaces ensures accessibility and flexibility, making it a valuable asset for therapists worldwide. Future developments are expected to include the integration of machine learning models to further enhance diagnostic capabilities and expand the tools' functionalities to support multilingual speech analysis.

7 . Declarations

7.1 Acknowledgements

We wish to express our sincere gratitude to the Department of Computer Science at the Adi Shankara Institute of Engineering and Technology for their unwavering support and guidance throughout the duration of this project. The resources and expertise provided by the esteemed faculty have significantly contributed to our efforts, and we sincerely appreciate their commitment to cultivating a conducive learning environment.

7.2 Study Limitations

One limitation of this study was the limited availability of diverse and comprehensive datasets representing different types of speech disorders [15]. Only a few publicly accessible datasets were available at the time of development, which constrained the range of disorders that could be analyzed and tested within the tool. Specifically, our model was trained on the Dysarthria and Non-Dysarthria Speech Dataset from Kaggle [12]. This limitation may affect generalizability of the diagnostic capabilities to other speech disorders or more varied speech patterns, and future work will docus on integrating more diverse datasets to improve system accuracy and robustness.

7.3 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

How to Cite

Jayaraj *et al.*, “AI-Powered Speech Analysis Tool for Speech Therapists”, *AIJR Proc.*, vol. 7, no. 5, pp. 23-29, Sep. 2025. doi: https://doi.org/10.21467/proceedings.7.5.4

References

[1] P. F. Jr., "Emergence and prevalence of persistent and residual speech errors," *Seminars in Speech and Language,* vol. 36, no. 4, pp. 217-223, November 2015.

[2] D. A. M. e. al., "Classifying Speech Disorders Using Voice Signals and Machine Learning," in *2025 22nd International Learning and Technology Conference (L&T)*, 2025.

[3] B. B. Lubker, "Epidemiology: An essential science for speech-language pathology and audiology," *J. Commun. Disord.,* vol. 30, no. 4, pp. 251-267, 1997.

[4] G. Ying, "Research on English Speech Data Analysis Algorithm Based on Deep Learning," in *2024 International Conference on Language Technology and Digital Humanities (LTDH)*, Bhubaneswar, 2024.

[5] A. R. K. e. al., "Voice Sentiment Analysis System Using Librosa," in *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, 2024.

[6] A. Ravindran, Django Design Patterns and Best Practices: Industry-standard web development techniques and solutions using Python, Packt Publishing Ltd., 2018.

[7] J. C. Y. H. a. I. C. J. Benesty, Noise Reduction in Speech Processing, vol. 2, Springer Science & Business Media, 2009.

[8] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *Int. J. Advance Res. Eng. Technol.,* vol. 1, no. 6, pp. 1-4, 2013.

[9] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," *IEEE Spectrum,* vol. 7, no. 8, pp. 57-62, August 1970.

[10] A. d. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America,* vol. 111, no. 4, pp. 1917-1930, April 2002.

[11] G. E. A. P. A. Batista, R. C. Prati and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD,* vol. 6, no. 1, pp. 20-29, June 2004.

[12] P. Gupta, "Dysarthria and Non-Dysarthria Speech Dataset," Kaggle, 2023.

[13] F. R. L. K. M. S. S. P. A. N. J. Y. P. M. v. L. D. J. O. J. R. Tobin, "The TORGO Database of Dysarthric Speech," University of Toronto, 2010.

[14] S. V. Thambi, K. T. Sreekumar, C. S. Kumar and P. R. Raj, "Random forest algorithm for improving the performance of speech/non-speech detection," in *2014 First International Conference on Computational Systems and Communications (ICCSC)*, 2014.

[15] K. Kimiafar, M. Sarbaz, D. Sobhani-Rad, A. S. Mousavi, M. R. Mehneh, F. D. Kemmak, S. F. M. Baigi and M. Esmaeili, "A comparative study of minimum data set of speech therapy: A systematic literature review," *Frontiers in Health Informatics,* vol. 12, 2023.