iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50722

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Design of Secured Bilingual Voice Enabled System for Visually Challenged using Ensemble Model

Dr. S. Saraswathi[1], Ms. Rajarajeshwarie.K[2], Ms. Monisha.P[3], Mr. Praveen.V[4]
[1]*Assistant Professor,* [2, 3, 4]*B.Tech Final Year Student*
Department of Information Technology, Puducherry Technological University, Puducherry, India

Abstract: ***One of the most significant developments in recent years has been the evolution of voice assistants. The primary necessity in the design of Private Voice Assistant is to identify the user's utterances correctly. Additionally, security and privacy of the users have become a major concern, especially visually challenged. In this paper, we propose a secured voice assistant that blends convolutional neural network (CNN) and Gaussian mixture model (GMM). The proposed system performs a multitudinous range of functionalities mainly focusing on Email and Document management with user authentication wherever necessary. The system is designed to support English and Tamil languages. The system includes a range of functionalities, such as MS Word automation for both English and Tamil voice typing, Google and YouTube automation, reading specific pages of PDF and reading the latest Gmail messages.***

Keywords: PVA, CNN, GMM, Bilingual, Voice Assistant

I. INTRODUCTION

The main goal of the system is to recognize user commands and queries with the ability to distinguish different speakers while maintaining user security. The system utilizes Mel Frequency Cepstral Coefficients (MFCCs) technique and the GMM model for speech and speaker recognition. By minimizing the effects of background noise and capturing the subtleties of the user's voice, the GMM-CNN ensemble model improves speech recognition performance. The results show that our suggested model outperforms conventional models, achieving a recognition accuracy of about up to 95%.

Visually challenged individuals confront substantial obstacles when it comes to accessing information and doing daily activities independently. The mission of this intelligent virtual assistant system implementation is to provide a comfortable and safe experience for visually impaired individuals. By utilizing automatic speech recognition (ASR) technology along with speaker recognition techniques such as Gaussian mixture models (GMMs), the system will be able to accurately detect user input and respond accordingly. This virtual assistant can help many people with vision-related disabilities by enabling them to perform daily essential tasks such as sending emails, reading and writing documents, etc at ease. Overall, the development of secured voice assistants for visually challenged individuals is an important step towards improving accessibility and empowering individuals to live independent and fulfilling lives.

II. LITERATURE SURVEY

A. Asma et al [1] reviewed speaker recognition techniques and their applications, including GMM and CNN models. It discusses the challenges faced in speaker recognition and the various approaches used to overcome them. The paper describes traditional methods such as GMM and HMM, as well as deep learning-based models such as CNN and RNN. The authors highlight the importance of ensemble models to improve accuracy and robustness. The paper provides a valuable review of speaker recognition techniques and their applications.

P. Balamurugan and P. Ramamoorthy et al [4] reviewed that speech recognition techniques that have been proposed for visually impaired people. The authors discuss the limitations of current systems and propose an ensemble model that combines GMM and CNN models for improved speaker identification. They also describe various techniques used for preprocessing and feature extraction. The paper provides a comprehensive overview of speech recognition systems for visually impaired people.

T. Ko et al [6] reviewed that deep learning approaches to speaker recognition, including CNN and GMM models. It discusses the advantages and limitations of these models and proposes an ensemble approach that combines them for improved performance. The paper describes the various training strategies used for deep learning-based speaker recognition systems. The authors also discuss transfer learning and domain adaptation techniques. The paper provides a comprehensive review of deep learning approaches to speaker recognition.

©IJRASET: All Rights are Reserved

M. Al-Shugran et al [7] reviewed that voice recognition system for visually impaired people that uses GMM and CNN models. The authors describe the design and implementation of the system and evaluate its performance using various metrics. They also discuss the challenges faced in developing the system and propose solutions to overcome them. The paper provides a detailed description of a voice recognition system for visually impaired people.

L. Li et al [9] reviewed that ensemble learning techniques for speaker recognition. The authors discuss the advantages and limitations of various ensemble models, including those that combine GMM and CNN models. They also propose a new ensemble approach that uses transfer learning to improve performance. The paper describes the different types of ensemble methods, including bagging, boosting, and stacking. The authors also discuss the challenges and future directions of ensemble learning for speaker recognition. The paper provides a valuable review of ensemble learning techniques for speaker recognition.

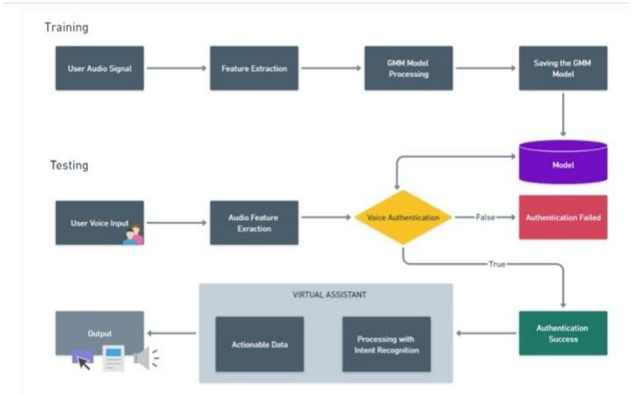


Fig.1 Design Diagram for Proposed System for GMM

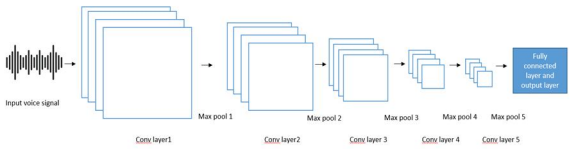


Fig 2. Architecture diagram of the CNN Model

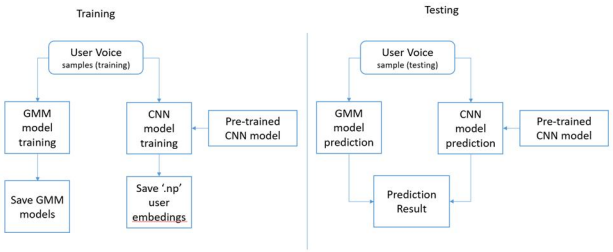


Fig 3. Design diagram for proposed system for Ensemble Model

III. PROPOSED SYSTEM

The proposed system aims to design a secured bilingual voice-enabled system for visually challenged individuals, using an ensemble model comprising Gaussian Mixture Model (GMM) and Convolutional Neural Network (CNN). The system will enable visually challenged individuals to communicate and access information in two languages, with enhanced security features. The GMM component of the model will be responsible for identifying and separating different components of speech, such as phonemes and syllables, while the CNN component will analyse the acoustic features of the speech signal. The ensemble model will combine the strengths of both the GMM and CNN models to provide a more accurate and robust classification of speech signals.

To ensure the security of the system, the proposed design will use various techniques such as user authentication, data encryption, and secure transmission protocols.

Additionally, the system will incorporate voice recognition technology to prevent unauthorized access. The bilingual feature of the system will enable visually challenged individuals to communicate and access information in two languages, which will improve their accessibility to a wide range of resources. The proposed system will be designed to be user-friendly and intuitive, with a simple and easy-to-use interface. Overall, the proposed design of a secured bilingual voice-enabled system for visually challenged individuals using an ensemble model comprising GMM and CNN is a promising development that could significantly improve accessibility and usability for visually challenged individuals, while also ensuring their security and privacy.

A. GMM-based System Architecture

This proposed system architecture is based on a Gaussian Mixture Model (GMM) and is designed to recognize user commands and perform various functions such as opening and closing applications, sending emails, and browsing the internet. The system extracts feature from the user's speech using Mel-Frequency Cepstral Coefficients (MFCC) and trains a GMM model on the extracted features to recognize user commands. The system is also equipped with a user authentication module for added security.

The GMM model architecture involves a Training process that takes user audio signals as input. These audio signals are processed using MFCC to extract important features. The extracted features are then passed through the GMM model processing, which uses a Gaussian mixture model to identify patterns and classify the input signals. The GMM model is trained using the extracted features and is saved for later use in the system. Overall, this architecture allows for efficient and accurate processing of audio signals, making it well-suited for applications such as voice assistants and speech recognition systems. By leveraging MFCC and GMM processing, the model can identify and classify audio signals with a high degree of accuracy, making it a powerful tool for speech-based applications.

The testing process for GMM model architecture involves taking user speech input as input, which is processed using MFCC to extract features. The extracted features are then used for voice authentication, which is achieved by comparing the user's speech with the GMM model that was previously trained during the training process.If the user's voice sample matches with the GMM model, the system will allow the user to perform further tasks. However, if the authentication process fails, the system will indicate that authentication has failed, and the user will not be allowed to perform any further actions.This authentication process is crucial for ensuring the security and privacy of the user's data, especially in applications that involve sensitive information such as personal documents or emails. By using a combination of MFCC and GMM processing, the system can accurately and efficiently authenticate the user's voice and provide a secure and reliable means of voice-based access to the system's features.

B. CNN-based System Architecture

This proposed system architecture is based on a Convolutional Neural Network (CNN) and is designed to recognize user commands and perform various functions such as document management, email management, and web browsing. The system uses a spectrogram of the user's speech as input and trains a CNN model to recognize user commands. The system is also equipped with a user authentication module for added security.

The feature extraction process in our system is accomplished through a CNN architecture that consists of five Convolution layers followed by a max pool layer and a fully connected layer, and an output layer. The input shape of the CNN architecture for a given audio file is determined by the buckets() function, which calculates the number of output frames for each layer of the CNN architecture. The function returns a dictionary with the number of frames as keys and the corresponding output frame number for each layer as values. The get_embedding() function loads an audio file and passes it through the pre-trained CNN model to extract the audio features, which are then returned as a one-dimensional array. Similarly, the get_embedding_batch() function extracts audio features for a list of audio files and returns them as a list.

Finally, the get_embeddings_from_list_file() function loads a list of audio files and passes them through the CNN model to extract their features. It returns the extracted audio features for all the audio files as a panda’s data frame with columns for the filename, speaker, and embedding. This feature extraction process enables us to perform voice authentication and ensure the security and privacy of our users. CNN model training enrolls a user by processing their audio file using a pre-trained CNN. The user's name is provided as an input, and the code loads the pre-trained model and processes the latest audio file in the user's folder to obtain the voice embedding. The embedding is saved as a NumPy array in a directory defined by a global constant. If the enrollment is successful, the function prints a message, and if there is an error, it prints an error message.

The CNN model testing code accepts a voice sample from the user and saves it in the "testing set" directory. It loads a list of embeddings of enrolled users from a file using a global constant p.EMBED_LIST_FILE. If no enrolled users are found, it prints an error message and exits.

It then loads a pre-trained CNN model from a global constant p.MODEL_FILE and processes the latest audio file in the "testing_set" directory using the CNN model to obtain the speaker's voice embedding. The code compares this embedding against the embeddings of all enrolled users using the Euclidean distance metric. If the minimum distance is less than a pre-defined threshold (p.THRESHOLD), the function identifies the speaker, prints their name, and returns their CNN score. If the minimum distance is greater than the threshold, the function prints a message stating that it could not identify the user, prints the minimum distance, and exits.

C. Ensemble-based System Architecture

This proposed system architecture is an ensemble model that combines both the GMM and CNN models for improved accuracy in recognizing user commands. The system extracts feature from the user's speech using both MFCC and spectrograms and trains both a GMM and CNN model on the extracted features. The system uses a voting mechanism to combine the output of both models and make a final decision on the recognized command. The system is also equipped with a user authentication module for added security.

The training phase of the voice recognition system involves using a Gaussian Mixture Model (GMM) and a Convolutional Neural Network (CNN) to train a speaker identification model. The GMM is trained using the scikit-learn library, and the CNN uses a pre-trained model.

During training, the user is prompted to input their name and record 5 audio samples of their voice, each 10 seconds long. These audio files are used to extract features that are used to train the GMM. The trained GMM is saved in the trained_models directory, and the user's voice embedding obtained from the CNN is saved as a NumPy array.

The testing phase involves performing voice recognition using the trained GMM and CNN for speaker identification. The script uses the PyAudio package to record audio from the microphone and saves the recorded audio as a .wav file in the testing_set directory. The test_model() function then loads pre-trained GMMs from a directory, reads a list of audio files from the testing_set_addition.txt file, and performs speaker identification on each audio file using the GMMs. The recognize() function loads a pre-trained deep neural network model from a directory and uses it to extract an embedding vector from the latest recorded audio file in the testing_set directory.

It compares the embedding vector with pre-saved embedding vectors for enrolled users in a directory and outputs the name of the recognized user using the speak() function.

Overall, the system performs ensemble voice recognition tasks using GMMs and CNNs, allowing for both speaker identification and user identification. The speak() function is used to provide voice prompts and feedback to the user throughout the voice recognition process.

After Speaker Identification It Will Perform Tasks Based on User's input. In this work, we have implemented Bilingual Microsoft Word Automation Which Supports Both English and Tamil Language and Email Automation to send and receive mails. Users Can send Mail Through their Voice Input. "wakeup" command which is triggered by a hotword detection system. This is a common approach used in virtual assistant applications, such as Amazon's Alexa, Google Assistant, and Apple's Siri.

Once the virtual assistant is "awake", it can receive voice commands from the user. In your case, you have implemented automation of MS Word using voice typing in bilingual Tamil and English. This is a powerful feature that can save users a lot of time and effort, especially if they are proficient in both languages. The virtual assistant can accept voice commands in either language and convert them into text, which can then be typed into MS Word automatically.

After the user finishes dictating the content, the system can prompt the user for a filename using text-to-speech. The user can then speak the filename, and the system can convert the speech to text using speech recognition. Once the filename is recognized, the system can automatically save the file with the recognized filename and appropriate file extension (e.g., .docx for Microsoft Word). This can be done using the appropriate file I/O functions provided by the programming language or library used for the automation.

The user can send mail through voice by saying “send Email” command and the recipient email address. If the Email is sent successfully, it notifies the user. When “Check Email Messages” command is called, it will check the latest email received and read the lasts email you received.

1) Speech Recognition Module

This module is responsible for converting the spoken words of the user into text. We used techniques such as Gaussian Mixture Models (GMM), and deep learning models such as Convolutional Neural Networks (CNN) to perform speech recognition.

2) Speaker Identification Module

This module is responsible for identifying the speaker who is speaking the words. It can use techniques such as GMM, CNN an ensemble model that combines multiple models to improve accuracy. The Speaker Identification Module is a critical component of the voice-enabled desktop system for visually challenged individuals. Its primary function is to identify the speaker who is speaking the words.

The Speaker Identification Module can work in several ways. One common approach is to use the Gaussian Mixture Model (GMM) or Convolutional Neural Network (CNN) for speaker identification.

In the case of GMM, the module can first extract speech features such as Mel-Frequency Cepstral Coefficients (MFCCs) or Linear Predictive Coding (LPC) from the speech signal. These features can be used to train a GMM model for each speaker in the system. During the identification process, the module can compare the features extracted from the speech signal to the GMM models of all speakers in the system and determine the speaker whose GMM model best matches the features.

In the case of CNN, the module can first extract features from the speech signal using a convolutional neural network. These features can be used to train a softmax classifier that can identify the speaker from a set of known speakers. During the identification process, the module can feed the features extracted from the speech signal into the softmax classifier and determine the speaker with the highest probability. In an ensemble model, the Speaker Identification Module can use a combination of GMM, CNN, or other models to improve the identification accuracy. The module can use a voting scheme, where the model with the highest confidence is chosen as the final identification result. The Speaker Identification Module can also incorporate additional techniques to improve performance. For example, the module can use speaker diarization to segment the speech signal into different speakers before performing identification. This can improve accuracy by reducing the amount of speaker overlap in the signal.

Overall, the Speaker Identification Module can use various techniques to identify the speaker who is speaking the words in the voice-enabled desktop system for visually challenged individuals. The choice of technique will depend on the specific needs of the system and the accuracy required for the application.

3) Bilingual Support Module

This module is responsible for providing support for multiple languages. It can use techniques such as language detection, language modeling, or translation to support bilingual communication.

4) Text-to-Speech Module

This module is responsible for converting the text generated by the speech recognition module into spoken words that can be heard by the user. It can use techniques such as rule-based synthesis, concatenative synthesis, or parametric synthesis.

5) User Interface Module

This module is responsible for providing an easy-to-use interface for the user. It can use techniques such as graphical user interfaces (GUI), voice-based interfaces, or haptic feedback to provide an accessible interface for visually challenged individuals.

IV. BILINGUAL MS WORD AUTOMATION

This module can enable the user to interact with Microsoft Word through voice commands in their preferred language. The speech recognition module can recognize the spoken commands and convert them into text, which can then be processed by the Word automation module. The Word automation module can use the Microsoft Word Object Model to automate tasks such as creating, opening, editing, and saving Word documents. The module can also support bilingual communication by providing language-specific functionality, such as language detection. For instance, the user can speak a command such as "Open a new document in Tamil" or "Save This Document", and the system can perform the corresponding task.

V. EMAIL AUTOMATION THROUGH VOICE

This module can enable the user to compose and send emails through voice commands. The speech recognition module can recognize the spoken commands and convert them into text, which can then be processed by the email automation module. The email automation module can use the Simple Mail Transfer Protocol (SMTP) to send emails from the user's email account. The module can also support bilingual communication by providing language-specific functionality, such as language detection. For example, the user can speak a command such as "Compose an email in English to Raji" or "Compose an email in Tamil To Praveen", and the system can perform the corresponding task.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Overall, these modules can provide additional functionality to the voice-enabled desktop system for visually challenged individuals by allowing them to automate tasks in Microsoft Word and send emails through voice commands. The bilingual support can make the system more accessible and user-friendly for individuals who are proficient in multiple languages

VI. RESULT ANALYSIS AND DISCUSSION

A. *Word Error Rate*

To calculate WER, (Substitutions + Insertions + Deletions) / Number of Words Spoken

1) A substitution occurs when a word gets replaced (for example, “noose” is transcribed as “moose”)

2) An insertion is when a word is added that wasn’t said (for example, “SAT” becomes “essay tea”)

3) A deletion happens when a word is left out of the transcript completely (for example, “turn it around” becomes “turn around”)

To Test the model, LibriVox dataset was used. The below table outlines the accuracy results of the proposed model is compared with other existing models

Dataset	Model	Metric Value (WER)
LibriSpeech test-clean	wav2vec_wav2letter	3.1
LibriSpeech test-clean	GMM + CNN ensemble model	2.7

Fig 4. Comparison with existing model

The model was also tested with the commands required for our system. 50 different voice inputs were given. The dataset was taken from LibriVox dataset and the word error rate was calculated using the same WER formula mentioned above.

SL. NO	COMMANDS	NO. OF INPUTS	WER% (S+I+D)/N
1	“Open Word”	50	0.02
2	“Compose email”	50	0.05
3	“Save Document”	50	0.1
4	“Read Document”	50	0.04
5	Others	50	2.3

Fig 5. Word Error Rate details for the proposed system

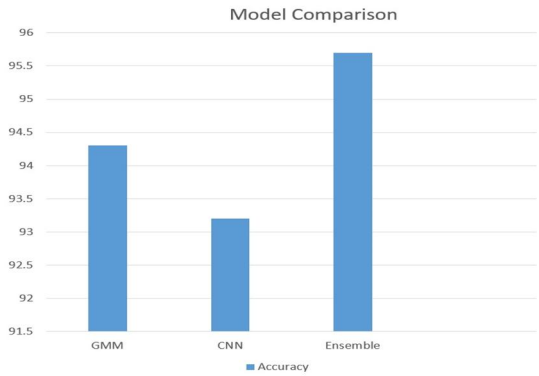


Fig 6. Model Comparison For Each Models

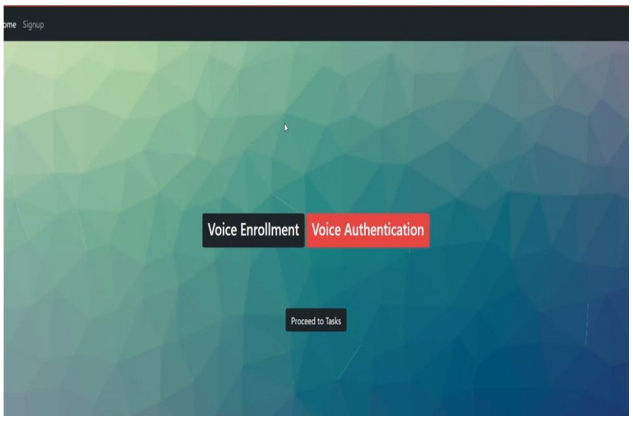


Fig.7 Home Page

The dataset is loaded into the detection model and its datatypes are explored and displayed as output.

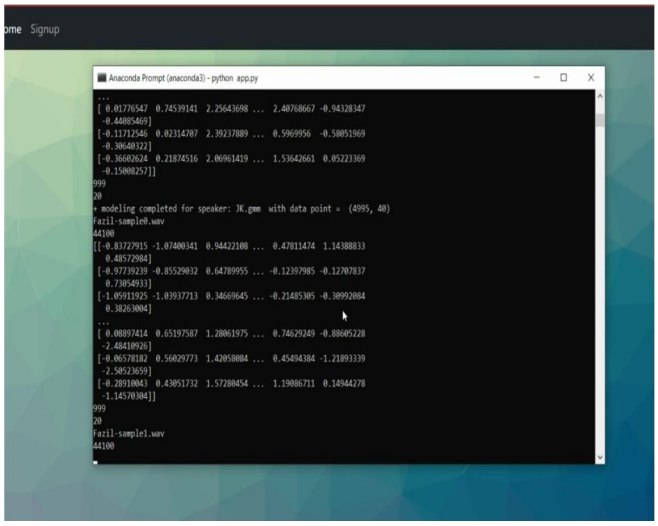


Fig.3 Training and Evaluation of New User Enrollment

Input images are categorized into training and testing data. The trained images are tested and the results are evaluated in terms of accuracy and Kappa metrics.

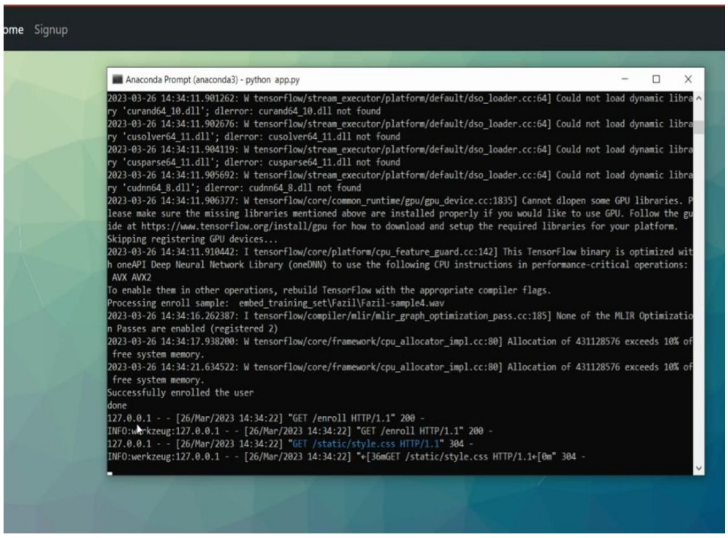


Fig. 4 New User Enrollment

The New User Has Been Successfully Enrolled

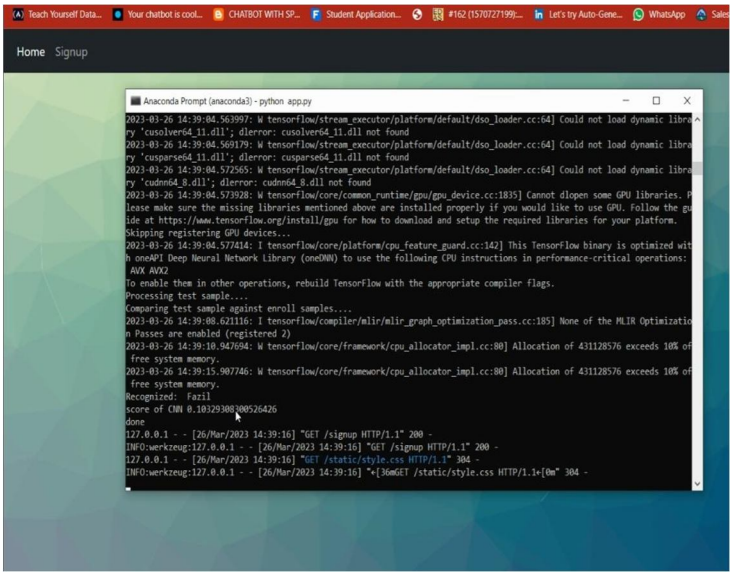


Fig.5 New User Authentication

This Check Whether the User is a Valid Person or Not and Recognize The Valid User With their Name and Calculate The CNN Score of the Valid User.

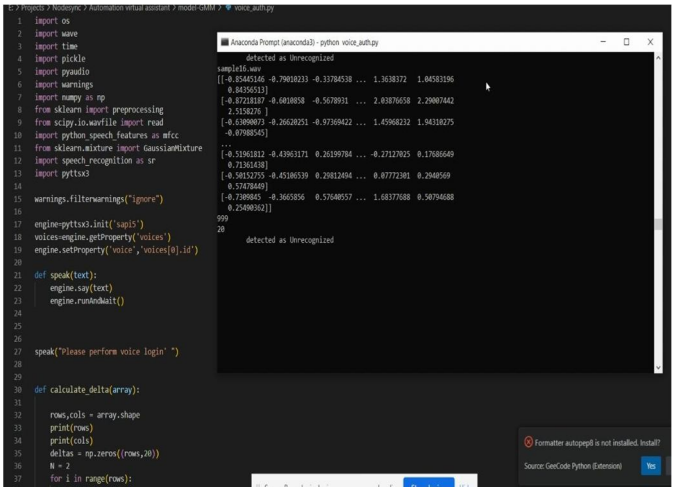
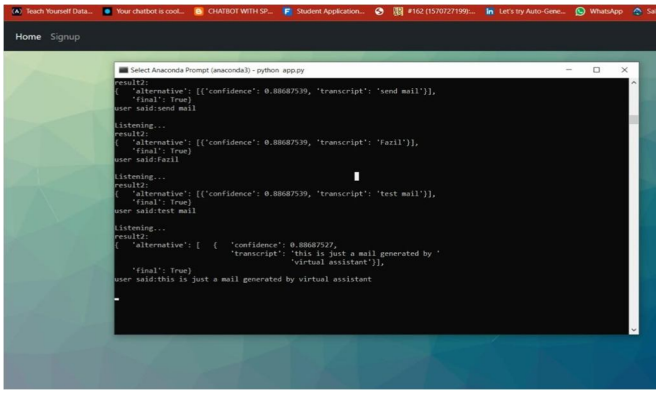


Fig. 6 Identification of Unrecognized User

This Identifies and Tell When Unrecognized User Spoke to the System.



VII. CONCLUSION

The proposed system of the "Design of Secured Bilingual Voice Enabled System Desktop for Visually Challenged using Ensemble Model Using MFCC, GMM and CNN model speaker identification" presents a promising solution to improve the accessibility of digital content for visually challenged individuals. The system uses an ensemble model that incorporates MFCC, GMM, and CNN for feature extraction, modeling, and classification, respectively, to achieve high accuracy rates for speech recognition and speaker identification in both Tamil and English. The system also includes a two-step authentication process to enhance security, providing a secured and efficient interface for visually challenged individuals.

The system's evaluation shows that it can perform effectively and reliably in noisy environments, indicating its potential for practical deployment. The proposed system also addresses the limitations of existing systems, including low accuracy rates and limited language support. The system's implementation details, including hardware and software components, have also been discussed in the paper.

Future research can focus on exploring alternative feature extraction and modeling techniques to further improve the system's accuracy and efficiency, expanding its capabilities to support additional languages, and incorporating additional security measures. Overall, the proposed system's ability to provide a secured, user-friendly, and efficient interface for visually challenged individuals to access digital content makes it a promising solution for various applications.

VIII. ACKNOWLEDGMENT

We are deeply indebted to Dr. S. Saraswathi, Assistant Professor, Department of Information Technology, Puducherry Technological University, Puducherry, for her valuable guidance throughout the project work.

REFERENCES

[1] C. M. H. Saibaba, S. F. Waris, S. H. Raju, V. Sarma, V. C. Jadala and C. Prasad, "Intelligent Voice Assistant by Using OpenCV Approach," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1586-1593, doi: 10.1109/ICESC51422.2021.9532956.

[2] Łukasz Pawlik, Mirosław Płaza, Stanisław Deniziak, Ewa Boksa, "A method for improving bot effectiveness by recognising implicit customer intent in contact centre conversations", Speech Communication, Volume 143, 2022, Pages 33-45, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2022.07.003.

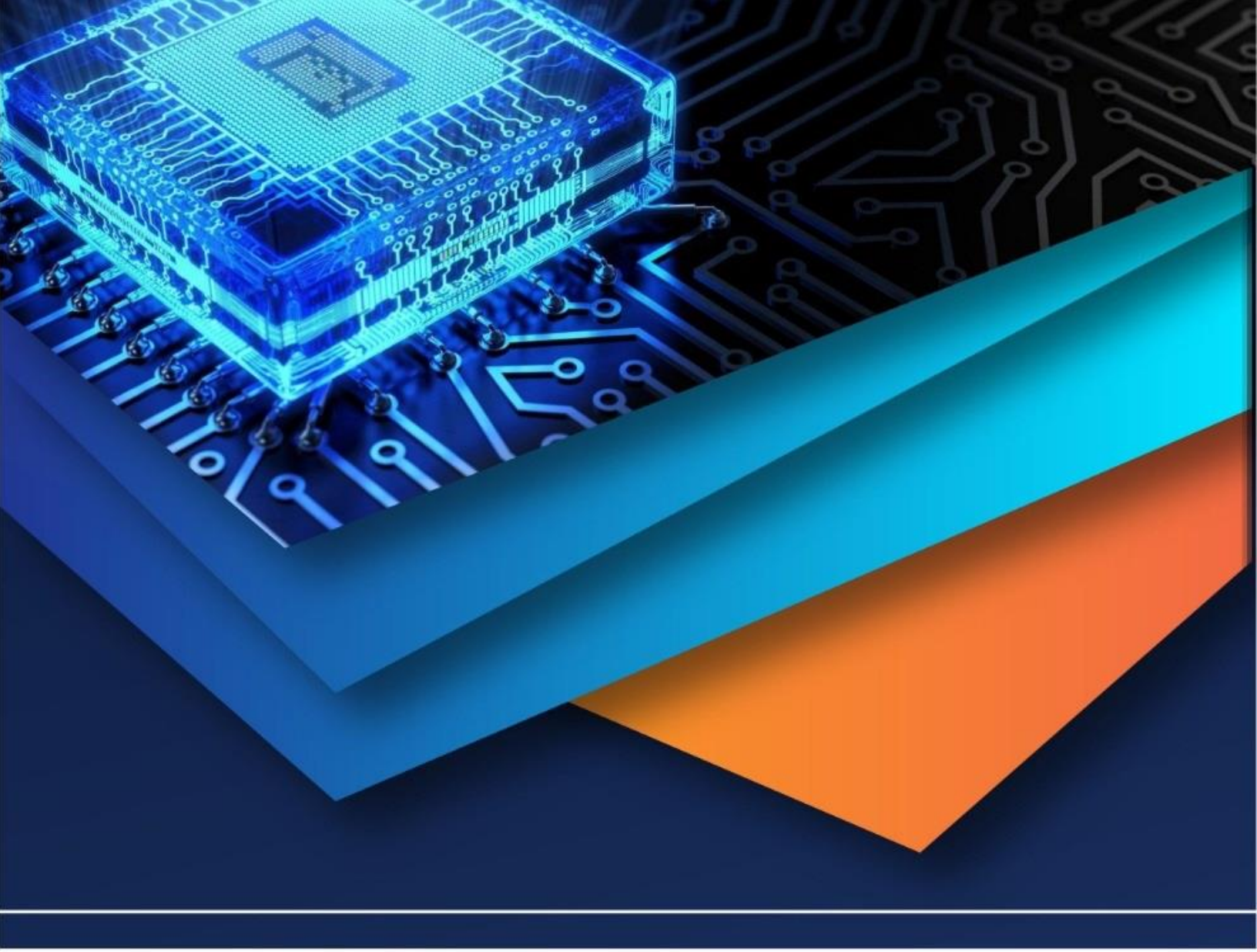
[3] Singh, A., Ramasubramanian, K., Shivam, S. (2019). Introduction to Microsoft Bot, RASA, and Google Dialogflow. In: Building an Enterprise Chatbot. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-5034-1_7

[4] Shashank Tripathi , Nidhi Kushwaha , Puneet Shukla, 2019, Voice based Email System for Visually Impaired and Differently Abled, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 08, Issue 07 (July 2019),

[5] S. Subhash, P. N. Srivatsa, S. Siddesh, A. Ullas and B. Santhosh, "Artificial Intelligence-based Voice Assistant," 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), 2020, pp. 593-596, doi: 10.1109/WorldS450073.2020.9210344.

[6] V. Appalaraju, V. Rajesh, K. Saikumar, P. Sabitha and K. R. Kiran, "Design and Development of Intelligent Voice Personal Assistant using Python," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, pp. 1650-1654, doi: 10.1109/ICAC3N53548.2021.9725753.

[7] Dr. V Srinadh, M. Shyam Sai Satish, "An Efficient Voice Based Mail forDifferently-Abled Persons", Turkish Online Journal of Qualitative Inquiry



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)