A PROJECT REPORT ON

SPEAKER VOICE RECOGNITION

Submitted to ***Centre for Development of Advanced Computing*** **(C-DAC),** ***Bangalore*** in Partial Fulfillment of the Requirements for the ***Post Graduate Diploma in Big Data Analytics (*****PG-DBDA*****)***

Under the Guidance of

Ms. SURABHI DWIVEDI

Submitted by:

1	**Mr. MOHIT CHAND**	PRN: 230350125048
2	**Mr. MAYUR CHOUDHARI**	PRN: 230350125046
3	**Ms. RUTUJA KULKARNI**	PRN: 230350125070
4	**Ms. NEHA LAHANE**	PRN: 230350125050

Centre for Development of Advanced Computing (C-DAC), Bangalore

CERTIFICATE

This is to certify that below mentioned students

1	**Mr. MOHIT CHAND**	PRN: 230350125048
2	**Mr. MAYUR CHOUDHARI**	PRN: 230350125046
3	**Ms. RUTUJA KULKARNI**	PRN: 230350125070
4	**Ms. NEHA LAHANE**	PRN: 230350125050

of **Post Graduate Diploma in Big Data Analytics (*PG-DBDA*)** have delivered their project entitled:

"Speaker Voice Recognition"

During the Academic session **March-2023 to August-2023** in a satisfactory manner in partial fulfillment of the requirement for the **Post Graduate Diploma in Big Data Analytics (*PG-DBDA*)** awarded by **Centre for Development of Advanced Computing (C-DAC), Bangalore**

Ms. Surabhi Dwivedi
(Project Guide)

MR. ARUN SHANKAR R S
(Course Co-Ordinator

ACKNOWLEDGEMENT

We take this opportunity to express our gratitude to all those people who have been directly and indirectly with us during the competition of this project.

We pay thanks to **Ms. Surabhi Dwivedi** who has given guidance and a light to us during this major project. Her versatile knowledge about "Speaker Voice Recognition" has eased us in the critical times during the span of this FinalProject.

We acknowledge here out debt to those who contributed significantly to one or more steps. We take full responsibility for any remaining sins of omission and commission.

Students Names:

1 **Mr. Mohit Chand**
2 **Mr. Mayur Sukchand Choudhari**
3 **Ms. Rutuja Meshram**
4 **Ms. Neha Lahane**

Speaker Voice Recognition

ABSTRACT:

Developing an efficient Ensemble Neural Network model for voice recognition of human beings, aiming to accurately recognize and classify individual speakers from audio recordings on the basis of their pitch, timber & frequency. The aim of our project is to produce a ensembled learning model with the highest accuracy that can reduce the present problem case scenario. we will be using both online datasets as well-as self-created dataset. The problem that we are dealing with is a type of biometric authentication that focuses on unique physical characteristics/feature like:

- Fundamental frequency i.e, pitch of voice
- Formants i.e, resonance of vocal tract
- Timber of voice
- Mel-frequency cepstral coefficient (MFCCS) i.e, spectral characteristics

Contents

Chapter 1: Introduction

In the modern world of communication and technology, voice recognition systems have gained significant importance. These systems find applications in various domains, including security, authentication, personal assistants, and more. With the advent of deep learning techniques, particularly in the realm of neural networks, voice recognition has achieved remarkable accuracy and robustness. This project aims to leverage deep learning methodologies to develop a Speaker Voice Recognition system capable of identifying and distinguishing between multiple speakers using a diverse dataset.

Project Overview

The primary objective of this project is to design and implement a deep learning model capable of recognizing and classifying speakers from a dataset containing audio samples of various individuals. The model will be trained to extract relevant features from audio signals and learn to differentiate between different speakers based on these features.

Objective

The main objective of this project is to develop a robust and accurate Speaker Voice Recognition system using deep learning techniques on a dataset containing audio samples from multiple speakers.

Chapter 2: Scope and Purpose

2.1 Scope:

The scope of this project encompasses the development and implementation of a Speaker Voice Recognition system using deep learning techniques. The project will focus on utilizing a diverse dataset containing audio samples from multiple speakers to train a model capable of accurately identifying and distinguishing between different speakers. The project's scope includes the following aspects:

1. Data Collection and Preparation:
- Gathering a dataset with a variety of speakers, languages, accents, and speaking styles.
- Preprocessing audio data to ensure consistency, quality, and appropriate segmentation.

2. Feature Extraction and Representation:
- Extracting meaningful features from audio signals, such as MFCCs or spectrograms.
- Creating feature representations that capture speaker-specific patterns and characteristics.

3. Deep Learning Model:
- Selecting a suitable deep learning architecture (CNN, RNN, CRNN, etc.) for speaker voice recognition.
- Designing the model to effectively process and analyze the extracted features.

4. Model Training and Evaluation:
- Training the model using the prepared dataset and extracted features.
- Evaluating the model's performance using relevant metrics to measure its accuracy and robustness.

5. Speaker Identification:
- Developing the model to predict the identity of speakers from input audio samples.
- Handling challenges such as varying accents, emotions, and speech rates.

2.2 Purpose

The primary purpose of this project is to create Speaker Voice Recognition system that leverages the capabilities of deep learning to accurately recognize and differentiate between multiple speakers. The project aims to achieve the following purposes:

1. Technological Advancement:
- Showcasing the potential of deep learning techniques in addressing complex tasks like speaker recognition.
- Pushing the boundaries of technology to create more accurate and efficient voice recognition systems.

2. Enhancing User Experience:
- Enabling secure authentication methods based on voice recognition, adding an extra layer of security for various systems.

3. Research and Development:
- Exploring innovative ways to handle speaker variability, noise, and real-world challenges in voice recognition.
- Adding to the body of knowledge in the field of deep learning and audio signal processing.

4. Practical Implementation:
- Developing a system with real-world relevance, potentially benefiting industries such as telecommunications, entertainment, and security.

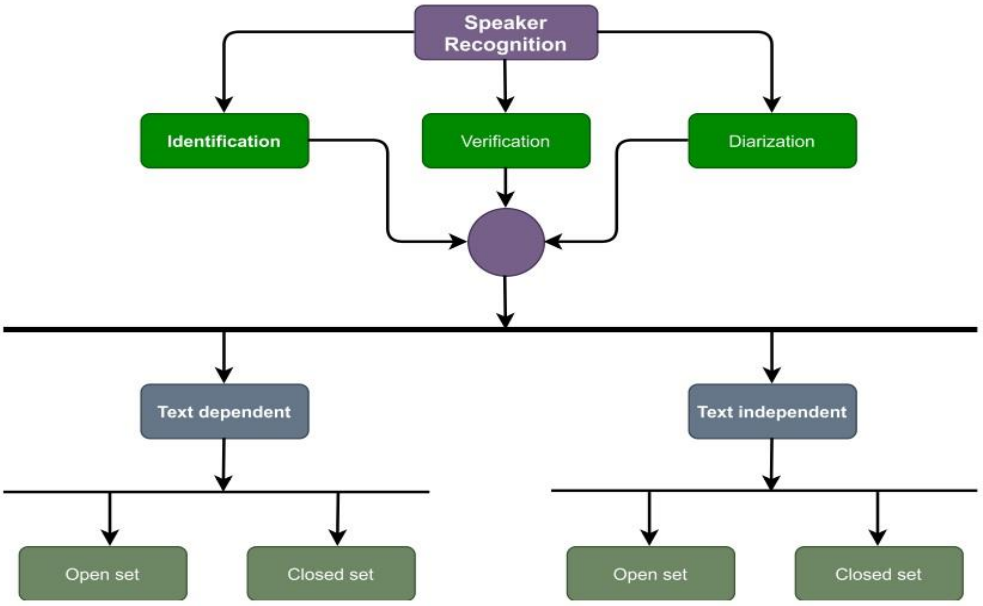
5. Learning and Skill Development:
- Offering a valuable learning experience for project participants by involving them in various stages of deep learning model development and evaluation**.**

Chapter 3: Software Requirement Specification

3.1 Product Perspective

Voice recognition plays a crucial role in various applications such as voice-controlled systems, security systems, and personalized services. Traditional voice identification methods often rely on handcrafted feature extraction techniques, which can be limited in capturing the rich and complex characteristics of human voices

To overcome these limitations, the proposed solution focuses on leveraging the power of Deep Neural Networks (DNN) to automatically learn discriminative features from raw audio data. The DNN models will be designed to analyze a voice identifier is a machine learning model that can identify the speaker of an audio clip or the category of an audio clip. It does this by extracting features from the audio clip, such as the SPEECH SIGNAL or SPECTROGRAPHY of the voice, and then using these features to train a classifier. After training the model, it will identify and produce the SPEAKER ID on the given input.



3.2 Product Functions

The primary objective of this research is to develop an Ensemble model that achieves high accuracy in voice identification tasks. The model should be capable of handling diverse voice samples with variations in accent, pitch, speaking rate, and background noise. Additionally, the solution should be computationally efficient to enable real-time or near real-time processing, making it suitable for applications that require quick and reliable voice identification.

- This model developed along with IoT can be used for home automation.
- Can be used to Digitalized the current Medical Infrastructure by Deploying it as an alternative for doctors on the basis of Disease Classification, etc.

3.3 User Classes and Characteristics

1. **Novice Users:** These users have limited or no prior experience with Voice Recognition technology. They may require clear instructions and intuitive interfaces to guide them through the interaction process. E.g.: *Patients who want to access their previous medical records*

2. **Intermediate Users:** Intermediate users are somewhat familiar with Voice Recognition technology but may need occasional assistance or prompts. They seek efficiency and improved accuracy in their interactions. E.g.: *Technical operator at any medical facilities.*

3. **Experts User:** Expert users are highly proficient with Voice Recognition systems and may use advanced features or commands. They prioritize speed and precision in their interactions. E.g.: *Researchers or Scientist who wants to access their data archive.*

4. **Assistive Technology Users:** Some users, such as individuals with disabilities, rely on Voice Recognition as an assistive technology. Design considerations for this group may include compatibility with accessibility tools and adaptable user interfaces. E.g.: *Differently able persons who wanted to unlock his electrical gadget.*

3.4 Operating Environment

The operating environment of a deep learning voice recognition program refers to the specific conditions and requirements under which the program operates effectively and accurately. It encompasses various aspects, including hardware, software, network connectivity, and external factors that impact the performance of the voice recognition system.

Here are some key components of the operating environment for a deep learning voice recognition program:

1. **Hardware Requirements:**
   - Central Processing Unit (CPU): Deep learning models require powerful CPUs to perform computations efficiently.
   - Graphics Processing Unit (GPU): GPUs accelerate the training and inference processes of deep learning models, enhancing speed and performance.
   - Memory (RAM): Sufficient memory is essential for loading and processing large models and datasets.
   - Storage: Adequate storage is needed to store the model, datasets, and any auxiliary files.
   - Microphones and Audio Input Devices: High-quality microphones ensure accurate audio input for recognition.
2. **Software Dependencies:**
   - Deep Learning Frameworks: The specific deep learning framework used, such as TensorFlow, PyTorch, or Keras, should be properly installed and configured.
   - Operating System: The program should be compatible with the target operating system (e.g., Windows, macOS, Linux).
   - Audio Libraries: Libraries for audio processing and manipulation, such as librosa, may be required.
   - Language and Development Tools: Programming languages (e.g., Python) and development tools (e.g., IDEs) are necessary for coding and testing.

3. **Network Connectivity:**
   - **Online Mode:** For cloud-based or server-hosted systems, reliable internet connectivity is required for communication and data exchange.

4. **External Factors:**
    - **Background Noise:** The program's ability to handle and filter out background noise impacts accuracy.
    - **User Environment:** User interactions may occur in different environments, such as quiet rooms, noisy streets, or crowded spaces.
    - **Speaker Variation:** The program should be capable of recognizing different speakers, considering variations in pitch, accent, and
5. **Security and Privacy Considerations:**
    - **Data Security:** Voice data may need encryption during transmission and storage to ensure privacy.
    - **User Authentication:** Security features like voice biometrics can enhance user authentication.

6. **Scalability and Performance:**
    - **Scalability:** Consideration for scaling the program to handle multiple users and high workloads.
    - **Performance Optimization:** Techniques for optimizing inference speed and model size may be required for real-time applications.

3.4 External Interface Requirements

3.4.1 User Interfaces

We are creating three stage user interphases, where user can connect with the backend model without getting into much technical constraints. These three stages are:

1. **Activation Page:** A trigger word/keyword will be used to activate the module.

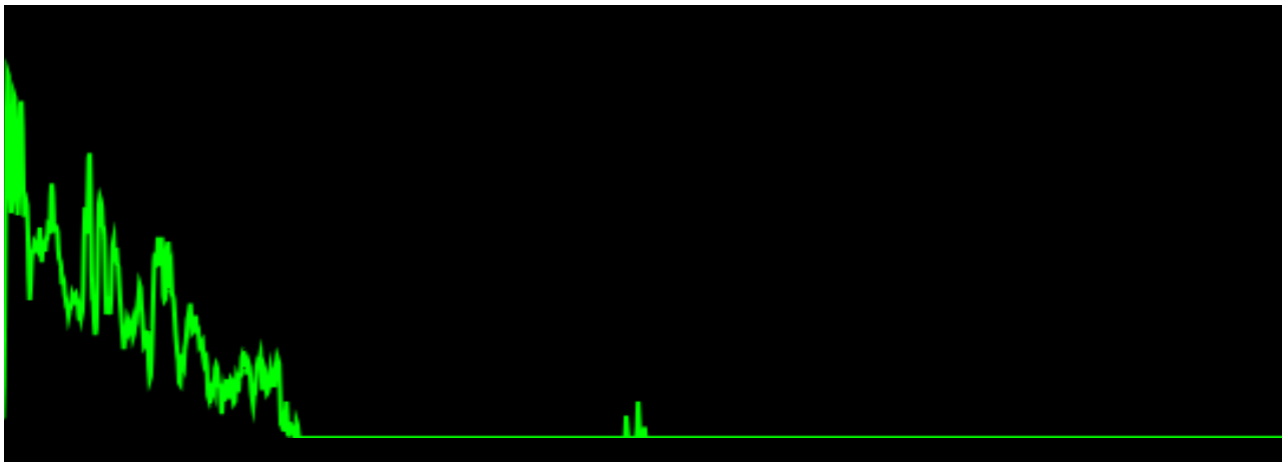


Fig. Activation Page

2. **Interactive Page:** Upon completion of activation page, paraphrase will be displayed in this page which the user has to speak/omit.

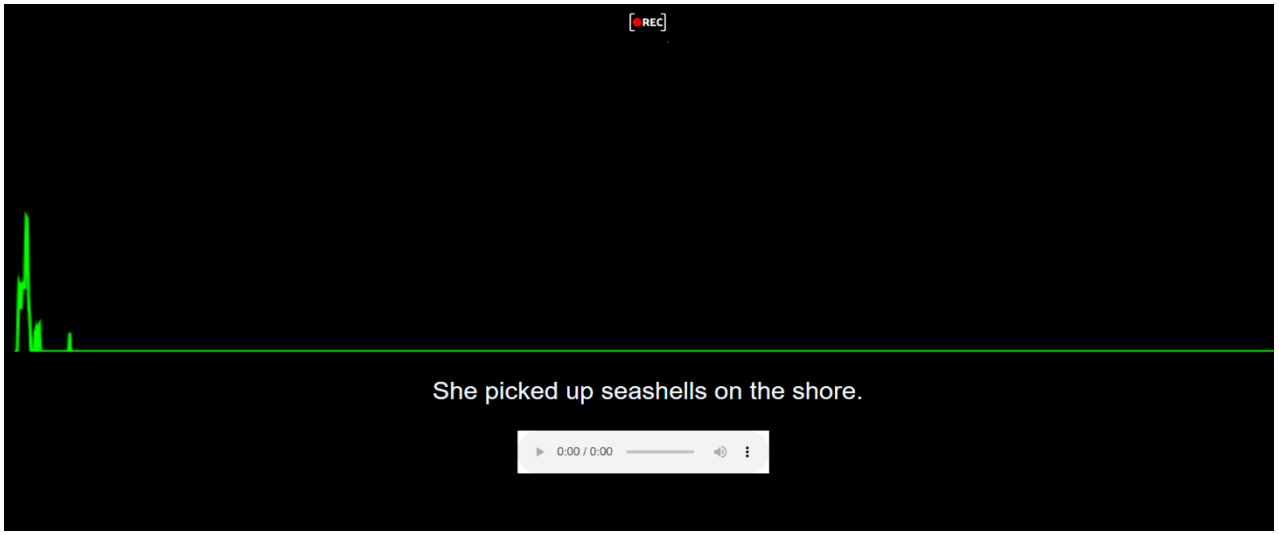


Fig. Interactive Page

3. **Responsive Page:** As the user speaks out the paraphrase, the audio generated in that process will be send to the backend module. And the details of the speaker if found will be shown in this page else module ask the user to enroll in the database.

Fig.

Responsive Page

3.4.2 Hardware Interfaces

This module has three electronic peripherals mainly,

1. Mic
2. Speaker
3. Display Device
4. Chipset integrated with IoT

Fig. Mic

Fig. Speaker

Fig. Display Device

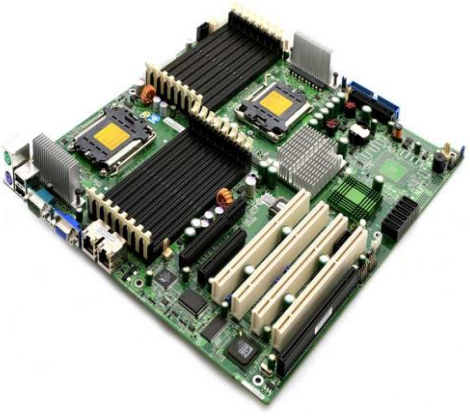


Fig. Chipset integrated with IoT

3.4.3 Software Interfaces

Software interfaces of a deep learning voice recognition program define how the program interacts with external systems, components, or users.

Some examples of software interfaces for a deep learning voice recognition program:

1. User Interface (UI): The user interface is the primary point of interaction between the user and the voice recognition program. It provides a platform for users to initiate voice commands, receive feedback, and view results.
   The UI includes:
   - Voice Input: Interface for users to speak commands or input audio for recognition.
   - Command Feedback: Display of recognized commands and responses.
   - Error Messages: Display of error messages or prompts for reinput.

2. Audio Input Interface: This interface manages the collection and preprocessing of audio data from input devices such as microphones.
   It includes:
   - Microphone Integration: APIs or drivers to capture audio input from connected microphones.
   - Audio Preprocessing: Conversion of raw audio into a suitable format for the deep learning model.

3. Deep Learning Model Interface: The deep learning model is at the core of the voice recognition program.
   This interface involves interactions with the trained model for inference:
   - Feature Extraction: Processing audio input to extract features for model input.
   - Inference: Passing audio features through the model for recognition.
   - Output: Retrieving model predictions for recognized text or commands.

4. External System Integration
5. Output Interface

Chapter 4: Dataset Specifications

4.1 Dataset Information

- The dataset used in this project consists of seven folders, divided into two groups. Speech samples, with 5 folders for 5 different speakers. Each folder contains 1500 audio files, each 1 second long and sampled at 16000 Hz.
- Background noise samples, with 2 folders and a total of 6 files. These files are longer than 1 second (and originally not sampled at 16000 Hz, but we will resample them to 16000 Hz). We will use those 6 files to create 354 1-second-long noise samples to be used for training.
- An audio folder which contains all the per-speaker speech sample folders.
- A noise folder which contains all the noise samples.

4.2 Introduction to the Theory Behind the Libraries, module & class:

1. **Tensorflow (import tensorflow as tf):**
   TensorFlow is an open-source machine learning library developed by Google Brain that is primarily used for building and training machine learning models, particularly deep learning models. It provides a flexible and efficient framework for developing a wide range of machine learning algorithms and applications.

2. **Numpy (import numpy as np):**
   NumPy (Numerical Python) is a fundamental package for scientific computing in Python. It provides support for working with arrays, matrices, and mathematical functions, making it a cornerstone library for data analysis, numerical simulations, and machine learning applications. NumPy is built on top of the Python programming language and is an essential component of the scientific Python ecosystem.

3. **OS (import os):**
   Imports the os module, which provides functions for interacting with the operating system, including tasks related to file and directory operations.

4. **SHUTIL (import shutil):**
   Imports the shutil module, which provides high-level file operations and simplifies tasks like copying, moving, and deleting files and directories.

5. **Pathlib (from pathlib import path):**
   Imports the Path class from the pathlib module. The pathlib module provides an object-oriented interface for working with filesystem paths and directories.

Chapter 4: Conclusion

In conclusion, the voice recognition project utilizing Convolutional Neural Networks (CNNs) has proven to be a significant advancement in the field of automatic speech recognition. CNNs, originally designed for image processing, have been successfully adapted to process spectrogram representations of audio data, enabling accurate and robust voice recognition capabilities.

Throughout the project, we achieved several key outcomes:

- Data Preparation: We collected and preprocessed a substantial amount of audio data, converting it into spectrogram images that capture the frequency and time-domain characteristics of each utterance. This transformed data allowed the CNN to learn relevant features for accurate recognition.
- Training and Validation: We trained the CNN on our dataset, carefully validating its performance to prevent overfitting. Techniques like data augmentation, dropout, and early stopping were employed to enhance generalization and model stability.
- Accuracy: Our trained CNN exhibited remarkable accuracy in voice recognition tasks. It consistently outperformed traditional methods, showcasing its capability to handle diverse accents, variations in pitch, and background noise.

In conclusion, the project underscores the effectiveness of CNNs in voice recognition tasks, validating their versatility beyond image processing. By transforming audio data into spectrogram images and leveraging the power of deep learning, we've developed a highly accurate and adaptable voice recognition system with the potential to revolutionize human-computer interaction and communication.

Chapter 5: Future Scope

1. **Novice Users:** These users have limited or no prior experience with Voice Recognition technology. They may require clear instructions and intuitive interfaces to guide them through the interaction process. E.g.: *Patients who want to access their previous medical records*

2. **Intermediate Users:** Intermediate users are somewhat familiar with Voice Recognition technology but may need occasional assistance or prompts. They seek efficiency and improved accuracy in their interactions. E.g.: *Technical operator at any medical facilities.*

Chapter 6 : REFERENCES

1. Deep Learning (MIT 2017) by Ian Goodfellow, Yoshua Bengio, Aaron Courville
2. A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities (DOP-May 27 2001, Digital Object Identifier 10.1109/ACCESS.2021.3084299)
3. Wikipedia: https://en.m.wikipedia.org/wiki/Speaker_recognition
4. www.youtube.com