Speaker Recognition Model

Vipin Sharma
Department of Computer Science and Engineering
Indian Institute of Information Technology Surat
Surat, India
ui21cs65@iiitsurat.ac.in

***Abstract*—This project focuses on filtering the voice of a person from mixture of voices using speaker recognition technique using machine learning based on recurrent neural network (RNN). The objective is develop a system which can remove anybody's voice from any audio sample. The system operates by accepting audio input from the user and proocessing it through the trained RNN model. If the person's voice is present in that audio file then it will be removed and the final filtered audio will be given to the user.**

***Index Terms*—Speaker Recognition Machine Learning Recurrent Neural Networks (RNNs) Supervised Learning Audio Processing Identity Authentication Security Protocols**

I. Introduction

Speaker recognition is a special skill of computer to compute who is talking based on the sound of their voice. Researchers been continuously working for a long time in this fields to extract more and more features from the audio. Many cool gadgets and smart devices are available now. Imagine talking to your smart device and it will automatically recognise it's you just by hearing your voice - that's speaker recognition.

Nowadays, when lots of people talk to smart devices, it's become important for the device to not just understand what is being said but also it should know who is talking. This helps in improving of providing the right information to the right person. This is where speaker recognition become important and useful to make things easier.

Speaker recognition is also useful in security where knowing password can be easy to guess. So security experts are looking for biometric to secure the data. Biometric include fingerprint, face recognition, speaker recognition using ASR or other unique feature to confirm that it's really you. So improving speaker recognition is a key to make sure our information stays safe and secure.

Speaker recognition is also used in filtering the other people voice from an audio to just focus onto the main speaker which makes the speech even more clear and audible for the others.

II. Existing Solutions

In the field of Automatic Speech Recognition (ASR), there are two components: speech recognition and speaker recognition. Initially, reasearchers focused on speech recognition to understand the audio saying. After speech recognition they moved to speaker recognition where they noticed the peaks, frequency, vocal tract's shape and many other features to identify who is speaking. The logical progression led to significant advancement in speech recognition, starting wth Davis et. al's system in 1952 at Bell Laboratories, which could recognize spoken digits based on formant frequenies in vowels.

On the other hand, research of speaker recognition began in 1960, when Pruzansky initiated studies at Bell Laboratories focusing on corelating digital spectrograms for speaker recognition.

The evolution of speaker recognition further divided into text-dependent and text-independent approcahes. In text-dependent approach, the speaker is allowed to speak a given set of words which makes trainig easier but it is less conveninent for the user but on the other hand, the text-independent approach doesn't requires any specific set of words, means user can speak in any way as per the convenience. The drawback is that it takes longer time durign training phase and more complex verification and validation.

OOverall, the journey from early to modern speaker recognition techniques reflects the higher accuracy and effeciency in ASR system. Driving continuous innovation and exploration in freature extraction and classification mehtods.

III. Proposed Model

A. RNN

Recurrent Neural Network(RNN) is a type of neural network whose previous output act as an input for the next step. In traditional neural network, the inputs and outputs are independent from each ohter. For instance, to predict the next word for a sentence, we need the previous word which means we have to remember the previous output. The main and most important feature of this neural network is its Hidden state, which remembers some useful information from a sequence.

B. Multiclass Classification

Multiclass classification is a machine learning algorthm focused on categorizing the data into more than two labels/classes. In short, the goal is to train a model that can effectively sort the instances into various predefined categories, providing a subtle solution for scenarios where items can belong to more than two labels. This method is commonly used in email categorization, handwriting recognition, and image classification.

In the given model, there is a speakers-folder array containing the list of speakers act as labels for the model architecture. Out of these labels, the target label index is used to attenuate the voice from the input voice data.

C. LSTM

LSTM network feature memory cells and gates. These gates includes the forget and input gates and modulate the content of the memory cell. If both gates are closed, the memory cell remains unchanged. This gate structure retains information across multiple time-steps.

An LSTM network comprises of four gates:

1. Forget Gate (f): Combines the input and previous output to determine how much previous state to preserve. An output of 1 means remeber everything and 0 means forget everything.

2. Input Gate (i): Decides which new information enters into the LSTM state.

3. Input Modulation Gate (g): Modulates information written into the Internal State Cell.

4. Output Gate (o): Combines input and previous state to generate a scaling fraction and then combined with the output of the tanh block to produce the current state. The output and state are feedback into the LSTM block.

D. RelU

ReLU is used in hidden layers of neural networks. It is mathematically defined as f(x) = max(0, x), which means that negative inputs are zero and the inputs for positive inputs are as it is. It introduces non-linearity to the model, facilitating in learning complex patterns.

E. Softmax

Softmax is an activation function used in the output layer of a neural network for multi-class classification problems. The output of the softmax function represents each element as the probability of the corresponding class. Final output is the class having the highest probability. In the given model, the last layer contains softmax activation to obtain probabilities over the different classes (speakers) to obtain the final result.

F. Loss Function

During model compilation, the chosen loss function is 'sparse_categorical_crossentropy'. Different types of loss functions serve different purposes. Here are some commonly used loss functions:

1) Mean Squared Error (MSE): Used in regression based problems. Restricts the model for large errors.
2) Binary Crossentropy: Used for binary classification based problems. Suitable when the target variable is binary (0 or 1).
3) Categorical Crossentropy: Used for multi-labeled classification based problems. Appropriate when the target variable is categorical and each field belongs to one and only one class.
4) Sparse Categorical Crossentropy: Similar to categorical crossentropy, but more suitable when the target variable is represented as integers. It avoids the need to convert the target variable to one-hot encoding values. In the provided code, the problem is a multi-class classification task with a categorical target variable representing the different speakers. Therefore, 'sparse_categorical_crossentropy' is a suitable choice for the loss function.

IV. Flowchart and Procedure

A. Flowchart

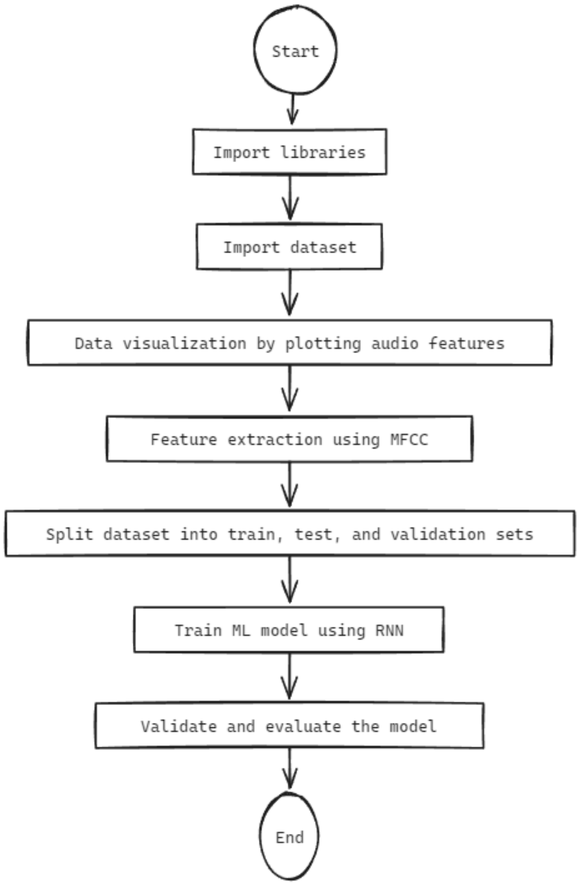
Start
Import libraries
Import dataset
Data visualization by plotting audio features
Feature extraction using MFCC
Split dataset into train, test, and validation sets
Train ML model using RNN
Validate and evaluate the model
End

Fig. 1. Flowchart for Proposed Model

B. Feature Extraction

Feature extraction is a crucial step in the preparation of data for machine learning tasks, and some of its importances are:

1) Dimensionality Reduction: Raw data, especially in audio, image, or text processing, can be high-dimensional. Extracting relevant features can reduce the dimensionality of the data, making it more manageable and efficient.
2) Relevant Information Capture: Not all raw data is equally informative. Feature extraction allows to capture only the essential feature from the data. It helps identify and retain the most relevant information from the data.
3) Noise Reduction: By focusing on specific features, we can filter out noise or irrelevant details present in the raw data.
4) Improved Model Performance: Feature extraction contributes to more improved model performance. Models performs better when trained on a set of well-defined features.

5) Facilitation of Learning: Extracted features can highlight patterns, relationships, or structures in the data that makes learning more easy. This can lead to faster and more effective training of ML models.
6) Human Interpretability: In some cases, Extracted data is more understandable to the human as compared to the raw data.
7) Domain-Specific Adaptation: Feature extraction facilitates the tailoring of data representation to the specific requirements and characteristics.
8) Handling of Multimodal Data: In cases, where data comes from multiple sources, feature extraction facilitaets the integration of information from different domains into a common representation.

C. MFCC

The array contains extracted MFCCs (Mel-frequency cepstral coefficients) features for each frame of audio. Each row corresponds to a frame, and each column corresponds ti a specific MFCC feature.

The output consists of an array of numbers representing the MFCCs for the first frame of the first audio file.

Understanding these numbers can be difficult and depends on your audio data specific context and the speakers' characteristics. Generally, MFCCs coefficients contains information about the vocal tract's shape. For instance, the first MFCC may shows the overall energy, while higher-order coefficients indicate the finer spectral details of speech signal.

In a typical set of MFCC coefficients, each coefficient serves a distinct purpose:

- First Coefficient (MFCC 0): Often termed the "constant" term, it denotes the signal's overall energy.
- Second Coefficient (MFCC 1): Reflects the overall spectral slope and correlates with perceived pitch.
- Third Coefficient (MFCC 2): Captures the spectral features related to the vocal tract shape,linked to speech formants(frequency peaks in the spectrum which have higher degree of energy).
- Fourth Coefficient (MFCC 3): Indicates changes in the spectral envelope, potentially related to quality of the voice produced by nasal resonator.
- Higher-order Coefficients (MFCC 4 and above): It offer insights about finer spectral characteristics, revealing more complex audio structures.

Interpretation of these coeeficients may vary based on characteristics of audio signal. Understanding MFCCs is often more straightforward within speech processing and analysis contexts.

V. Results

During training of the model, audio samples of five people were taken for model evaluation process. The testing sample is a set of shuffled audio of these five people. The obtained Accuracy Score is 0.9644970414201184 and Weighted F1 Score is 0.9645248921343329

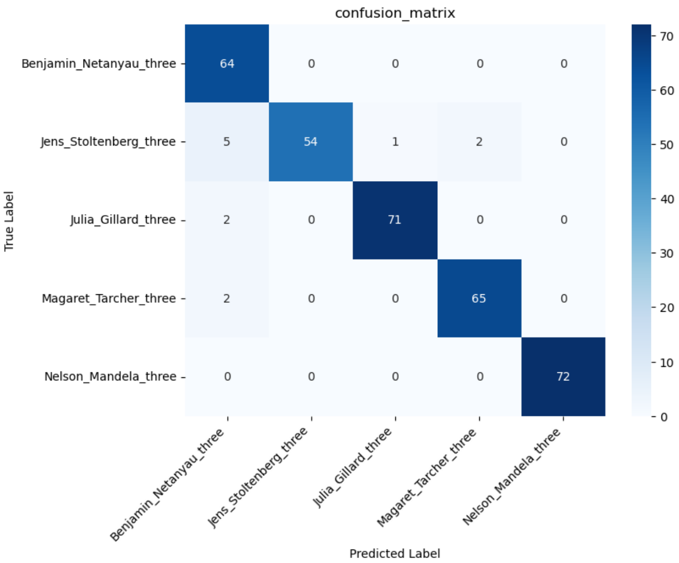

Fig. 2. Confusion Matrix

VI. Conclusion

This study is divided into three parts :- audio preprocessing, feature extraction and multilabel classification. The crucial part is splitting the training in three sec sample and testing data in nine second sample as the testing audio samples should be greater than the training sample data. During training and testing MFCCs coefficient are extracted to find the information about the vocal tract's shape and overall energy of the signal. These coefficients are used to train and test the data. There are some of the challeneges faced during the process like differentiating between male and female voice , language difference creates a large difference between the pitch, noise reduction, normalisation of the audio and many more. Lastly choosing an appropriate machine learning algorthm is imprtant too to determine how the model process the data.

References

[1] Multiclass Classification using Scikit-learn. (n.d.). Retrieved from https://www.geeksforgeeks.org/multiclass-classification-using-scikit-learn/ [2] Introduction to Recurrent Neural Network. (n.d.). Retrieved from https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/ [3] Simple Audio Recognition: Recognizing Keywords. (n.d.). Retrieved from https://www.kaggle.com/code/janesser777/simple-audio-recognition-recognizing-keywords/notebook
[4] Speaker Recognition. (n.d.). Retrieved from https://www.kaggle.com/code/alkanerturan/speakerrecognition/notebook [5] Alam, T. (n.d.). MFCCs Made Easy. Retrieved from https://medium.com/@tanveer9812/mfccs-made-easy-7ef383006040 [6] Alam, T. (n.d.). Understanding MFCCs: A Practical Guide. Retrieved from https://medium.com/@tanveer9812/mfccs-made-easy-7ef383006040