Deep Learning-Based Text-Independent Speaker Identification for Voice Authentication

Mahabuba Sultana and Jannatul Ferdous

A Thesis in the Partial Fulfillment of the Requirements

for the Award of Bachelor of Computer Science and Engineering (BCSE)

Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Summer 2022

Deep Learning-Based Text-Independent Speaker Identification for Voice Authentication

Mahabuba Sultana and Jannatul Ferdous

A Thesis in the Partial Fulfillment of the Requirements for the Award of Bachelor of Computer Science and Engineering (BCSE)

The thesis has been examined and approved,

____

Prof. Dr. Utpal Kanti Das
Chairman and Professor

____

Dr. Hasibur Rashid Chayon
Coordinator and Associate Professor

____

Md. Saidur Rahman
Supervisor and Assistant Professor

Department of Computer Science and Engineering

College of Engineering and Technology

IUBAT – International University of Business Agriculture and Technology

Summer 2022

Letter of Transmittal

3 September 2022
The Chairman
Thesis Defense Committee
Department of Computer Science and Engineering
IUBAT–International University of Business Agriculture and Technology
4 Embankment Drive Road, Sector 10, Uttara Model Town
Dhaka 1230, Bangladesh

Subject: Letter of Transmittal.

Dear Sir,

It gives us enormous please to submit the paper on "**Deep Learning-Based Text-Independent Speaker Identification for Voice Authentication**" as per instruction. We expect this paper to be informative as well as comprehensive.

While conducting the thesis paper, we have gathered lots of knowledge about machine learning. We have tried our level best to collect the relative information as comprehensive as possible in preparing the thesis. During preparation of the report we have experienced practically a lot that will help us a great in my career. We will be able to explain anything for more clarification if necessary.

We would like to thank you, for giving us the opportunity to do a report on the above mentioned topic

Yours sincerely,

Mahabuba Sultana	Jannatul Ferdous
18303045	18303002

Student’s Declaration

We hereby declare that this thesis is based on results obtained from our own work. All the materials that were used for the purpose of completing this thesis are duly acknowledged and mentioned in reference. This thesis, neither in whole nor in part, has been previously submitted to any other University or Institute for the award of any degree or diploma. We carried our research under the supervision of Md. Saidur Rahman.

______________	______________
Mahabuba Sultana	Jannatul Ferdous
18303045	18303002

Supervisor’s Certification

This is to certify that Thesis report on “**Deep Learning-Based Text-Independent Speaker Identification for Voice Authentication**” has been carried out by Mahabuba Sultana bearing ID# 18303045 and Jannatul Ferdous bearing ID# 18303002, of IUBAT – International University of Business Agriculture and Technology as a partial fulfillment of the requirement of thesis defense course. The report has been prepared under my guidance and is a record of the accomplished work carried out successfully. To the best of my knowledge and as per their declaration, no parts of this report has been submitted anywhere for any degree, diploma or certification.

Now they are permitted to submit the report. I wish them success in all their future endeavors.

Md. Saidur Rahman
Assistant Professor

Department of Computer Science and Engineering

IUBAT–International University of Business Agriculture and Technology

Abstract

Speech authentication, security and surveillance, electronic voice eavesdropping, and identity verification all make extensive use of speaker identification techniques. This study looks at how speech recognition can be used to make door systems more accessible to persons with disabilities and to increase the safety of people's lives and property. The goal of this research project is to develop a voice recognition algorithm-based door access control system that will enable users to quickly unlock their doors while maintaining their security and safety. MFCC traits have been shown in numerous studies to be highly effective in precisely identifying speakers. Most speaker identification experiments have used shorttime characteristics like PLP coefficients and MFCC because of their ability to capture the recurring nature of signals. This work presents a novel architecture that uses the mel frequency cepstral coefficient (MFCC) for feature extraction and a convolutional neural network (CNN) for classification in order to identify the speaker. There is no text involved in the use of this architecture. The key to any text-independent speaker identification procedure is the system's ability to learn elements that are useful for categorization. A dataset of 700 audio clips, three known classes of three individuals each with 150–200 audio clips, and one unknown class of thirty individuals with five voice samples per speaker were all constructed. The accuracy percentage of our suggested approach was 98%. With more optimization, this method could be used to a number of research projects pertaining to speaker identification.

Keywords: MFCC, CNN, SoftMax

Acknowledgment

We would like to offer our sincere gratitude towards my thesis supervisor, Mr. Saidur Rahman for his patient guidance and constant support for our work. In addition to that, we would also like to thank sir for his critical analysis and regular feedback of our work's progress, which guided us to remain in the aligned path to our goal. Furthermore, we would like to extend our gratitude to all the other faculty members of IUBAT, Department of Computer Science and Engineering who played a vital part in the development process through the whole program. We are also grateful to all the resources found on the internet which helped us to grasp a better understanding of machine learning. "Super Data Science", a website which contains machine learning tutorials, proved out to be really helpful along with the sites related to group forum and discussions.

Table of Contents

List of Figures

List of Tables

Chapter I: Introduction

In this era of information, a plethora of high-tech gadgets are gradually infiltrating our everyday lives, drastically altering the way we live and behave. Biometric identification technology, which provides us with easier and more useful ways to identify people, has progressively surpassed some traditional authentication techniques, which use pin numbers or passwords to authenticate users [1]. However, passwords and pin numbers are no longer considered to offer a high level of protection because they can be mistyped or forgotten. The iPhone's Siri speech assistant and the face recognition technology used in airport terminals1 are two instances of biometric identification systems. Speaking has, on the one hand, always been the easiest and most natural way for humans to communicate, convey our ideas, and engage with one another. Therefore, identifying people based on their dialogue voice and content and then providing the relevant services would be a better method to genuinely simplify our daily lives. Up till now, speech recognition technology has advanced significantly and been incorporated into our everyday life [2]. Automatic Speaker Recognition is the ability to identify the right person just by listening to his voice. This method enables the use of speech waves containing the speaker's information to confirm the user's identity and manage access to a variety of services, such as voice dialing, phone banking, database access services, online shopping, voice mail, information services, remote computer access, transaction security for bank trading, and remote payment, among others [3]. Speaker identification is the process of identifying human speech through the application of artificial intelligence methods. The method of utilizing a machine to identify a speaker from a collection of recognized speech signals is known as automatic speaker identification, or ASI. Voice signals are effective communication tools that consistently transmit rich and valuable data, including a speaker's gender, emotion, accent, and other distinctive qualities.

Despite the speakers’ lack of physical presence, these distinctive qualities allow researchers to discern between them during phone talks. These features enable machines to learn the speech of speakers and become as familiar with them as people are. Using test utterances, speakers are recognized once their utterances have been educated by machine learning algorithms on the gathered dataset [4]. There are two methods for identifying speakers: text-dependent and textindependent. The text that is uttered during testing and system training must be same for the text-dependent speaker identification system to function. In contrast, text-independent speaker identification system does not rely on the speaker’s spoken words in order to identify the speaker. Since there are no restrictions on the text that can be used during the test or train phases and the speaker needs to be identified regardless of what is being said, text-independent speaker recognition is more difficult for the system to manage but more versatile for the users [5]. Moreover, speaker identification and speaker verification are the two processes that make up speaker recognition. The process of identifying a speaker utterance from a collection of trained speaker utterances is known as speaker identification. The speaker who has the highest likelihood of making an exam utterance is then designated as the speaker. As an alternative, speaker verification entails using binary classification to ascertain whether a speaker of a test utterance is a member of a group of speakers. The text-independent speaker identification challenge is taken into consideration in this study because of its applications in the current speech advancement of technology. Due to its numerous applications, speaker recognition has attracted a great deal of attention from researchers. These include forensic voice verification used by law enforcement to identify suspects [6], access control to various services, including computer access control [7], voice dialing, mobile banking, and mobile shopping [8]. Speaker identification and recognition are difficult challenges with crucial applications in automation, security, and authentication. Deep learning techniques such as AM-SincNet and SincNet performed remarkably well on these tests [8]. Finding discriminative elements in voice signals that can help classification algorithms perform better is the main difficulty in speaker identification. Numerous research have suggested various feature-engineering methods in this area, including spectrum features, time-domain features, power-normalized cepstral coefficient, MFCC, and linear prediction cepstral coefficient (LPCC) [4]. The recognition of voice is the task of supervised learning. The speech recognition problem will take an audio signal as input, and we have to infer the speaker's identity from the audio signal. We are unable to feed our model the raw audio signal since there would be too much noise in it. It has been discovered that the basic model performs noticeably better when its characteristics are used as input rather than the audio stream in its raw form. Mel-Frequency Cepstrum Coefficients (MFCC) are the most widely used technique for eliminating specific characteristics from an audio source. The MFCC approach was employed in this work to extract speech signal features [9]. The method of feature extraction involves locating a value or vector that can be used to identify a person or an object. The most widely used technique in the field of voice processing is MFCC since it is believed to be fairly effective at conveying signal. Two often used efficient models for speaker verification are the i-vector and the Gaussian Mixture Model-Universal Background Model (GMM-UBM). The primary disadvantage of these models is their unsupervised character, which results from the fact that they were not objectively trained for speaker verification setup [10]. In this paper, we offer a CNN architecture that can detect speakers automatically without textual input. The basic objective is to identify one speaker from a big group of speakers using a short speech snippet. We created a CNN architecture to address classification issues pertaining to speech and voice. One of the main concerns for any individual or business is security, and as technology has advanced over time, numerous strategies have been used to protect people and property through door access control systems. Traditional door opening techniques involve turning the doorknob or using a physical key. Physical door locks that require keys are easily copied and could be misplaced by someone. In addition to other failures with standard biometric technologies and other technologies, a person's finger can be severed to do a fingerprint scan, a pin can be compromised by a variety of techniques or combinations, and a person's picture can be utilized for facial recognition. It is also more difficult for people who are physically limited. For example, opening a door system on one's own is difficult for a wheelchair user without extra support or aid from another person. Consequently, it is inevitable that a voice recognition access control system that works with both able-bodied and disabled people will be needed. The purpose of this project is to construct and install a speech recognition-enabled door access control system, which will facilitate the granting of door system access while safeguarding individuals' safety and belongings. The particular goals of this investigation are to:

1. User verification: In the digital age, identity creation involves more than just traditional, physical authentication methods. Speech recognition is often a useful tool for user authentication. Our model can significantly increase the accuracy of user verification on individuals who have not yet been trained the system. The model will have a high degree of accuracy in identifying the speaker's audio sample.

2. Authentication with biometrics: Speaker recognition offers several advantages over traditional biometrics, such as low cost, high adoption, and non-invasive speech acquisition. It is possible to construct a speaker identification system without the

requirement for costly hardware or active speaker involvement. Speaker recognition may replace many other online services, including the use of credit cards and debit cards, as well as any security precautions like remembering bank account passwords.

3. Security: When combined with other verification methods like facial recognition, speaker recognition could be utilized as an identification mechanism in credit card transactions. Speaker recognition technology has several applications, including computer access control, monitoring, and long-distance speech authentication. Speech biometrics, another name for automatic speaker recognition, is a reliable way to verify a person's unique identification for secure device access management. Instead of focusing on the content of speech, automatic speaker identification technology recognizes the identity of the speaker. Speaker recognition is also utilized in conversational interfaces, messaging apps, and Internet of Things (IoT) devices like smart speakers and connected automobiles to provide seamless security and personalization.

1.1 Background and context

Speech is the primary means of interpersonal communication in the world tradition. The technique of recognizing speakers by their voice's acoustic properties is known as speaker identification. Speaker recognition has drawn a lot of attention from researchers because of its many applications, which include forensic speech verification for government law enforcement agencies to identify suspects. Feature extraction is an essential step in the speaker recognition process because of its massive influence on the performance of a speaker identification classification model. Novel components recently proposed by speaker recognition researchers have proven useful in classifying human voices. The training phase and the testing (identification or matching) process are typically the two basic procedures that make up the approaches for speaker identification tasks. The four steps of the training process are as follows: Entering the speech signal, Pre-speech preparation, the process of normalization and Extraction of features. As was previously said, the most important step in speaker identification is feature extraction. Because of this, the majority of studies have attempted to investigate the various methods for obtaining features from speech data with the goal of identifying each speaker. We can divide the techniques for feature extraction of the speaker identification task into MFCC-based approach and non-MFCC-based approach since MFCC-based feature extraction approaches have been provisionally employed more than any other approach. The filter banks of the MFCC-based technique are made to function similarly to how humans perceive frequencies through their aural senses [11]. Numerous MFCC-based feature fusions have been the subject of ongoing research [[12] - [13]]. The performance of the MFCC-GMM speaker recognition systems can be greatly enhanced by integrating two distinct sets of features from MFCCs and Perceptual Linear Predictive Coefficients (PLPC) utilizing ensemble classifiers in conjunction with principal component transformation [12]. Additionally, MFCC features in conjunction with Residual Phase Cepstrum Coefficients (RPCC) were examined by Wang and Johnson [14]. It provided a notable enhancement in terms of overall resilience and accuracy for speaker recognition tasks. Ma, Yu, Tan, and Guo [15] recently employed a text-independent speaker recognition challenge using MFCC features integrated with the histogram transform feature. In a comparative analysis, various time-domain statistical variables and machine learning classifiers are used to characterize speakers. It was discovered that the multilayer perceptron classifier had the highest accuracy, at roughly 94%. Even though the experimental results of the study obtained good classification accuracy, the results cannot be transferred on a larger scale because the experimenters used just 16 speaker voices from the PDA speech dataset. A small number of speaker utterances were also used in the study's training and testing sets [16]. A speaker identification model was presented that can identify ten distinct speakers using the dataset for the Urdu language. The study combined MFCC with deep learning-based features for speaker categorization using the support vector machine (SVM) method. The categorization accuracy of the experimental results is 92%. As such, the results are promising. On the other hand, there are some issues with the dataset used in the testing. First off, there were only ten speaker utterances in the studies. Second, every statement contained a single word. Therefore, for complex human voices, the authors' proposed fusion-based features might be ineffective and inefficient [15]. The ELDSR dataset's 22 speakers were classified through the use of an ANN classifier and clustering-based MFCC features. The experimental results of the investigation showed a 93% categorization accuracy [17]. For each of the 50 speaker utterances in 5 the MEPCO voice dataset, a collection of discriminative features was suggested. To categorize speaker utterances, these authors extracted RASTA-MFCC characteristics. The gathered features were fed into a classifier constructed using a GMM-universal background model in order to learn the classification rules. The results showed a 97% accuracy rate in classification. Despite the reasonable categorization accuracy of the data, the study only included six utterances, with one speech lasting a mere three seconds, therefore the results cannot be generalized. RASTA-MFCC features might therefore not be helpful for speaker utterances longer than three seconds [18]. To classify 50 and 138 speakers from the CMU and YOHO datasets, respectively, MFFC features and a GMM classifier were demonstrated. The experiment's results indicated that, when employing the recommended feature extraction strategies, classification accuracy was 86% and 88% [19]. Short speech portions are a challenge for today's interactive gadgets, such as smart speakers and phone assistants. Nevertheless, current speaker recognition programs struggle with brief utterances and need rather lengthy speech to function successfully. Deep neural network designs built on recurrent neural networks (RNN) and convolutional neural networks (CNN) were established with the goal of resolving this issue and improving speaker recognition performance for short utterance speaker recognition applications. Using the traditional i-vector and the Probabilistic Linear Discriminant Analysis (PLDA) technique, the suggested method is assessed. When significant and brief utterance durations are used, the experimental results demonstrate that the model might surpass the performance of the i-vector -PLDA baseline system and improve the speaker recognition capabilities [5]. Traditional Speaker Identification techniques have been surpassed by Deep Neural Learning approaches, particularly Convolutional Deep Neural Networks (CDN), which have become a potent tool in the field of speech processing. With a particular focus on speaker recognition from spoken Hindi language, a novel method for speaker identification and audio classification utilizing 1-Dimensional Convolutional Residual Blocks was developed. With an astounding accuracy rate of 86.02%, the suggested Residual architecture greatly improves speaker recognition even in situations with low Signal Noise Ratio. For the same set of speakers, this performs better than the conventional Gaussian Mixture Model (GMM) and Feed Forward Back-propagation Network (FFBN) model [20]. Using deep neural networks to extract speech features for individual speakers is an interesting avenue to pursue. One simple way to get started is by using a metric learning loss function to train the discriminative feature extraction network. A single loss function does, however, frequently have significant drawbacks. Consequently, three distinct losses—triplet loss, n-pair loss, and angular loss—were introduced for this problem and deep multi-metric learning was suggested as a solution. To train a feature extraction network with residual connections and squeeze-and-excitation attention, the three loss functions collaborate. Through extensive testing on the massive VoxCeleb2 dataset, which comprises more than a million utterances from more than 6000 speakers, the suggested deep neural network achieves an equivalent error rate of 3.48%—a very competitive outcome [21]. For automatic speaker identification from brief speech samples, a CNN architecture is suggested. In contrast to the well-known CNN architectures, the architectural design 6 seeks to capture the temporal character of the voice signal in an optimal convolutional neural network with few parameters. On the large and clean dataset, the proposed CNN-based technique performs better, while all DL approaches are outperformed by the classical method on the other dataset with less data. The suggested model achieves 99.5% top-1 accuracy on 1-second voice samples from the LibriSpeech dataset [22].

1.2 Problem Statement

Security is one of the top concerns for any person or company, and as technology has developed through time, many methods have been employed to safeguard people and property through door access control systems. The conventional method of opening a door is by turning the door knob or using a physical key to open the door. Physical keys that are used to unlock doors are prone to duplication and might be lost by people. A person's finger can be cut off to do a fingerprint scan, a pin can be hacked using various methods or permutations, and a person's photo can be used for face recognition, among other failures, with common biometric technologies and other technologies. Furthermore, it is harder for those who have physical limitations. For instance, it is challenging for a person in a wheelchair to open a door system without additional assistance or help from someone else. Therefore, the requirement for a speech recognition access control system that can accommodate both those with disabilities and those who are able-bodied is unavoidable.

1.3 Research Aim and Objectives

The goal of this project is to build and implement a door access control system with

voice recognition functionality, in order to make it easier to grant access to door systems and to provide protection for people's lives and property. The specific objectives of this research study are to:

User verification: To create identities in the digital age, user verification goes beyond conventional and physical techniques of authentication. User verification can frequently be carried out using speech recognition. On users for whom the system has not yet been taught, our model's accuracy in user verification can be greatly improved. Model will be able to identify the speaker's audio sample with high accuracy.

Biometric authentication: In comparison to traditional biometrics, speaker recognition has a variety of benefits, including low cost, high acceptance, and non-invasive voice acquisition. A speaker identification system can be built without the need of expensive equipment or direct speaker participation. The use of a debit card, credit card, keeping track of a bank account password or any other security measures, as well as many other online services, might be replaced by speaker recognition.

Security: Speaker recognition could be used as an authentication mechanism in credit card transactions in conjunction with other verification techniques like facial recognition. Among other things, speaker recognition technology can be used for long-distance speech authentication, monitoring, and computer access control. Automatic speaker recognition, also referred to as speech biometrics, is a trustworthy method for confirming an individual's distinct identity for safe device access control. Automatic speaker recognition technology is used to identify who is speaking rather than what is being spoken. In order to offer frictionless security and personalization, speaker recognition is also being used in conversational interfaces, messaging apps, and IoT devices such as smart speakers and linked cars.

1.4 Significance of the study

- When compared to other biometric systems like iris and fingerprint scan, can speech recognition technology offer a better usability and performance?
- Would this study take into account a person's speech tone in relation to their moods? For instance, a person's vocal tone differs depending on whether they are just waking up or feeling angry.

1.5 Research Question

- When compared to other biometric systems like iris and fingerprint scan, can speech recognition technology offer a better usability and performance?
- Would this study take into account a person's speech tone in relation to their moods? For instance, a person's vocal tone differs depending on whether they are just waking up or feeling angry.

Chapter II: Literature Review

In the global tradition, people communicate with one another by speech. Speaker identification is the process of identifying speakers using the acoustic characteristics of the human voice. Due to the variety of uses for speaker recognition, including forensic voice verification for suspect identification by government law enforcement organizations, speaker recognition has become a focus of considerable research. Due to the enormous impact it has on a speaker identification classification model's performance, feature extraction is crucial to the speaker recognition process. Researchers in the field of speaker recognition have recently put forth innovative elements that have been effective in categorizing human voices.

A comparison study to categorize speakers using different time-domain statistical variables and machine learning classifiers; they found that the multilayer perceptron classifier had the highest accuracy of about 94%. Because the experimenters used just 16 speaker voices from the PDA speech dataset, even if the experimental results of the study achieved good classification accuracy, the results cannot be applied to a larger scale. Additionally, in both the training and testing sets of the study, a scant number of speaker utterances were utilized [24]. Recently put up a speaker identification model that uses the dataset for the Urdu language to recognize 10 different speakers. The support vector machine (SVM) approach was utilized in the study to combine MFCC and deep learning-based characteristics for speaker classification. The experimental findings have a classification accuracy of 92%. Therefore, the outcomes are encouraging. The dataset employed in the tests, however, has a number of flaws. First, the studies only included 10 speaker utterances. Second, there was only one word in each utterance. As a result, the authors' suggested fusion-based features may be inefficient and ineffectual for complex human voices [25]. It was examined at the use of an ANN classifier and clustering-based MFCC features to categorize 22 speakers from the ELDSR dataset. The study's experimental findings had a classification accuracy of 93% [26]. A set of discriminative features for the MEPCO voice dataset's 50 speaker utterances. For the purpose of classifying speaker utterances, these writers retrieved RASTA-MFCC traits. To learn the classification rules, the collected characteristics were fed into a classifier built using a GMM-universal background model. 97% classification accuracy was attained by the findings. Although the results showed reasonable categorization accuracy, they cannot be generalized because the study only used six utterances, with one statement lasting barely three seconds. Because of this, RASTA-MFCC features may not be useful for speaker utterances that are longer than 3 s [27]. MFFC features and a GMM classifier were examined to categorize 50 and 138 speakers from the CMU and YOHO datasets. The experiment's outcomes showed 86% and 88% classification accuracy using the suggested feature extraction techniques [28].

Chapter III: Research Methodology

In this study, we propose a CNN architecture for automatically identifying speakers without the need for text. Using a brief speech fragment, the main goal is to distinguish one speaker from a large group of others. Deep CNNs, which were initially created for computer vision problems, are the subject of most current research. Additionally, the majority of the present speaker identification techniques require query phase audio samples longer than 3 seconds to achieve high accuracy. For voice and speech-related classification problems, we developed a CNN architecture which is displayed in Figure 1

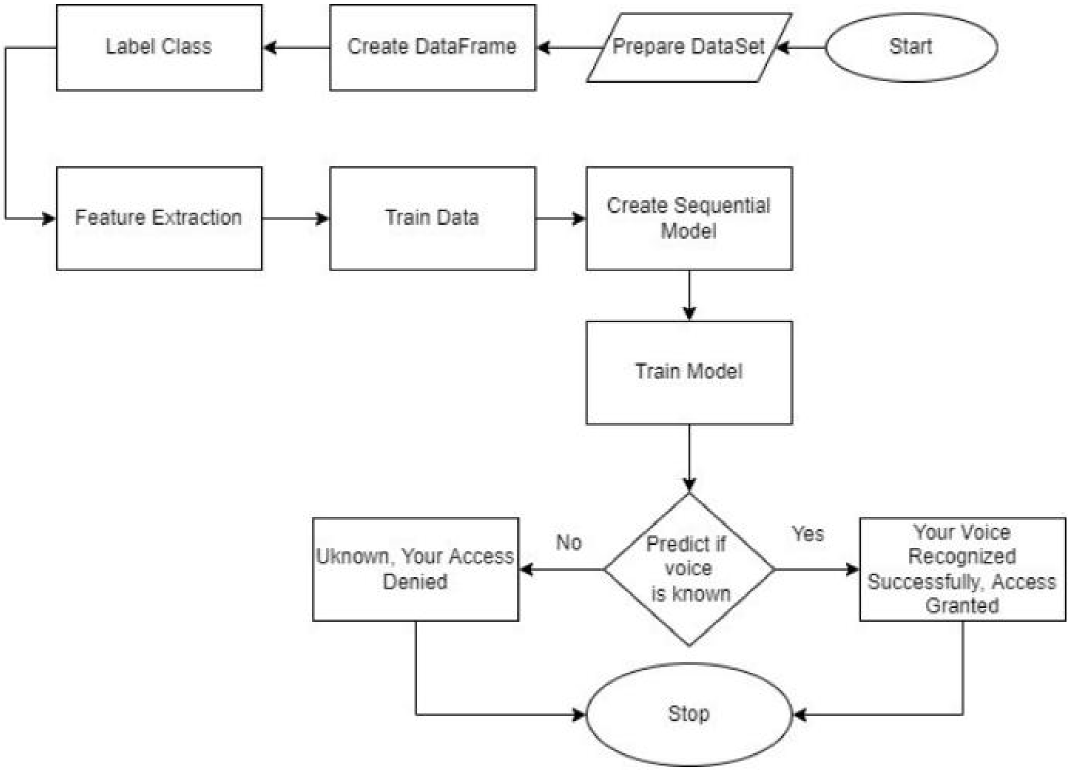


Figure 1 Workflow of the proposed model

3.1 Dataset

Three known classes of three speakers, each with 150–200 audio clips, and one unknown class of thirty speakers, each with five voice samples, comprised the 700 audio clips in the dataset we constructed. Though the majority of the audio files in the collection are of male and female speakers speaking English, there are recordings of speakers with a range of accents and languages. Every audio clip in the dataset has a duration of 10 seconds. In addition, only 20% of the data was used for testing and 80% was used for training. In order to identify the speaker using CNN and MFCC architecture, the classes were placed into a Data Frame and labeled.

3.2 Mel Frequency Cepstral Coefficients (MFCC)

Preprocessing and normalizing the voice or acoustic signal as needed is the first step in the process. After that, a temporal analysis employing windowing and framing techniques comes after the FFT procedure. Next, the logarithmic scale is encircled by the Mel-scale filter bank. Next, the Discrete Cosine Transform (DCT) is used to compute the Cepstral Mean Subtraction (CMS). The characteristics acquired by applying the MFCCs technique are represented by these CMS coefficients. The MFCCs-based feature extraction approach carried out frequency analysis based on International Journal of Machine Learning and Computing in order to advance the voice frame via the filter bank. Figure 2 illustrates the steps of the feature extraction method based on MFCCs. The retrieved features are represented by the MFCCs coefficients, which are the method's output. These MFCCs coefficients can therefore be applied to additional analysis or categorization for any purpose. Figure 3 illustrates the Mel Frequency Cepstral Coefficients categorization.

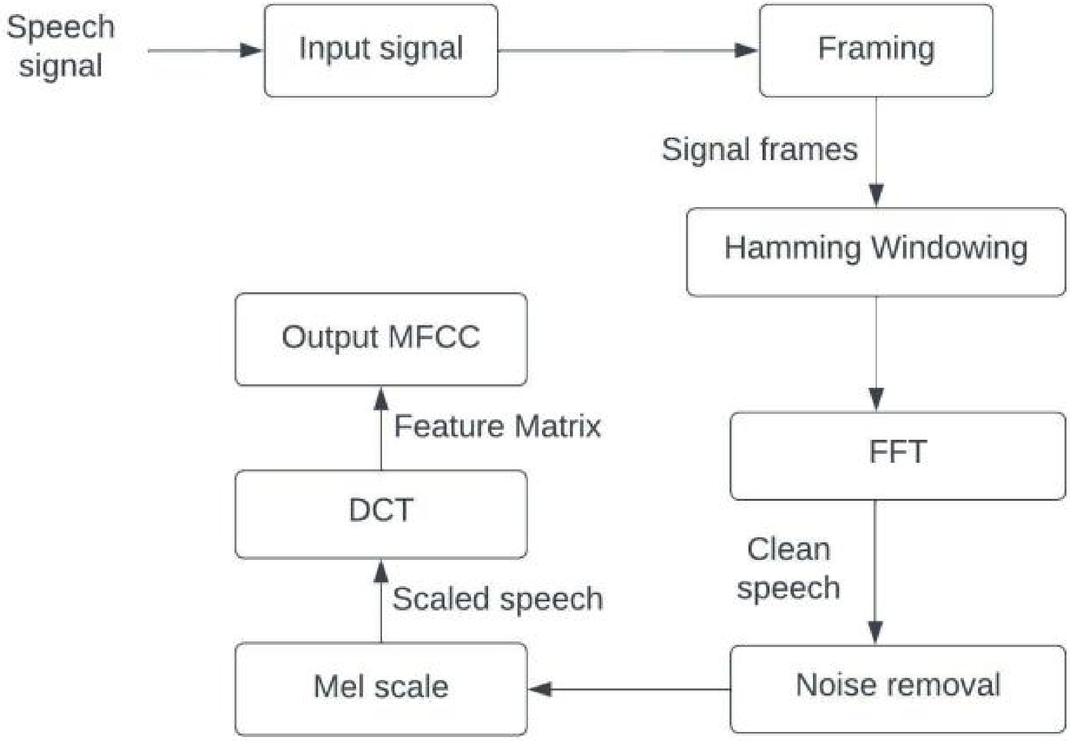


Figure 2 Workflow of Mel Frequency Cepstral Coefficients

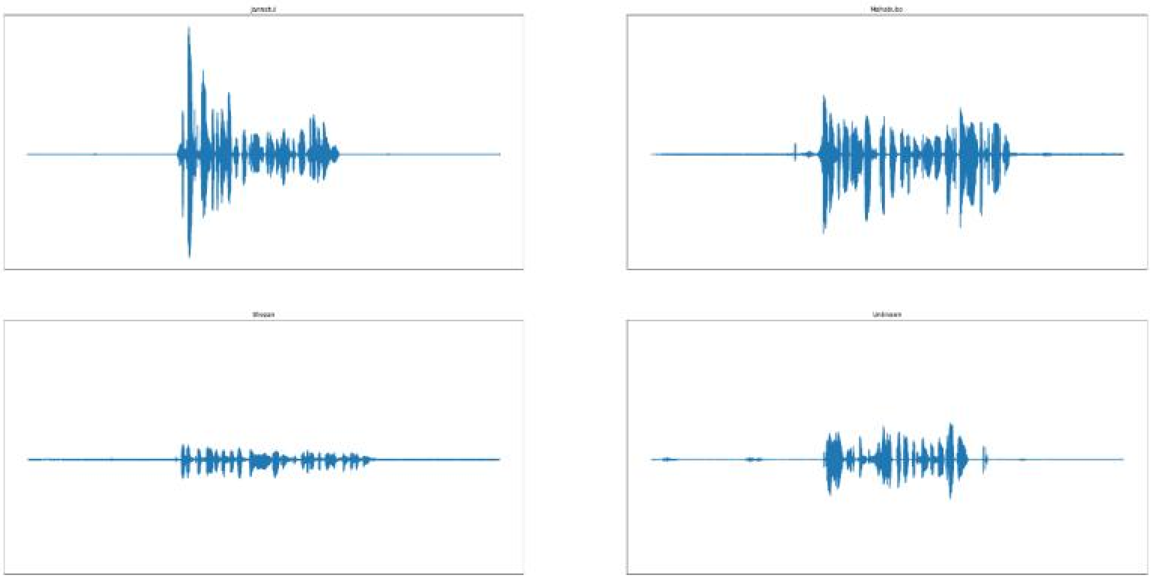


Figure 3 Mel Frequency Cepstral Coefficients

3.2.1 Noise Reduction

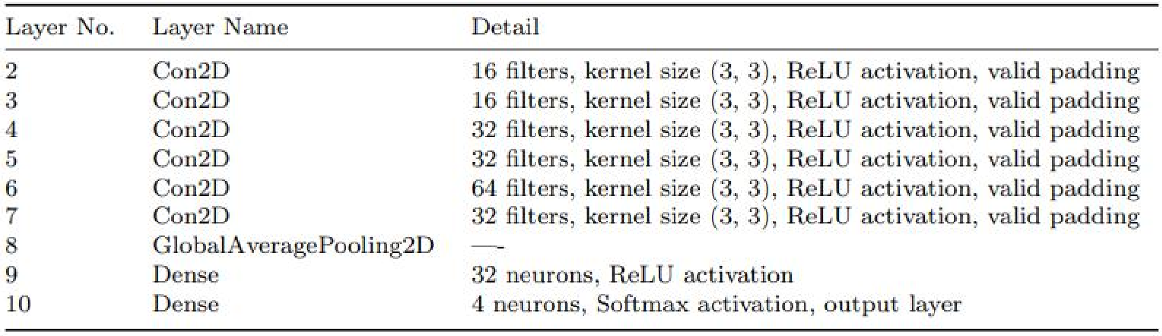
Voice, bioacoustics, and physiological signals are examples of time-domain signals that can now be made quieter with the use of the Python noise reduction method. It makes use of a Noise Gate technique called ”spectral gating.” After calculating the spectrogram of a signal and, if needed, a noise signal, it estimates a noise threshold, also known as a gate, for each frequency band including both the signal and the noise. A mask is calculated using the frequency-varying threshold to gate noise below it.

3.3 Convolutional Neural Network

Convolutional Neural Network (ConvNet/CNN) is a Deep Learning method that takes an image as input and distinguishes between various aspects/objects in the image while also assigning significance (Weights and biases that can be learned) to them. Money is very important to ConvNet. Compared to previous classification algorithms, this technique requires fewer pre-processing steps. While manual engineering is necessary for basic approaches, ConvNets may educate themselves to understand certain filters and attributes. The journal Neurons in the Human Brain published a description of the structure of a ConvNet, an analog connection network that is impacted by the organization of the visual cortex. Individual neurons can only respond to stimuli in the Receptive Field, which is a tiny fraction of the visual field. Tables 1. describes the structure of the proposed CNN model.

3.3.1 Convolutional Neural Network (CNN)

The two parts of a CNN architecture are as follows and displayed in Figure 4.



Layer No.	Layer Name	Detail
2	Con2D	16 filters, kernel size (3, 3), ReLU activation, valid padding
3	Con2D	16 filters, kernel size (3, 3), ReLU activation, valid padding
4	Con2D	32 filters, kernel size (3, 3), ReLU activation, valid padding
5	Con2D	32 filters, kernel size (3, 3), ReLU activation, valid padding
6	Con2D	64 filters, kernel size (3, 3), ReLU activation, valid padding
7	Con2D	32 filters, kernel size (3, 3), ReLU activation, valid padding
8	GlobalAveragePooling2D	—-
9	Dense	32 neurons, ReLU activation
10	Dense	4 neurons, Softmax activation, output layer

1. Feature extraction is the process of identifying and separating the unique qualities of a data set for analysis using a convolution tool.

2. The feature extraction network consists of multiple pairs of convolutional or pooling layers

3. A fully connected layer that uses the output of the convolutional process to identify the data's class based on previously extracted features. 4. Reducing the number of features in a dataset is the goal of this CNN feature extraction model. It creates new features that combine the preexisting features of an initial collection into a single new feature.

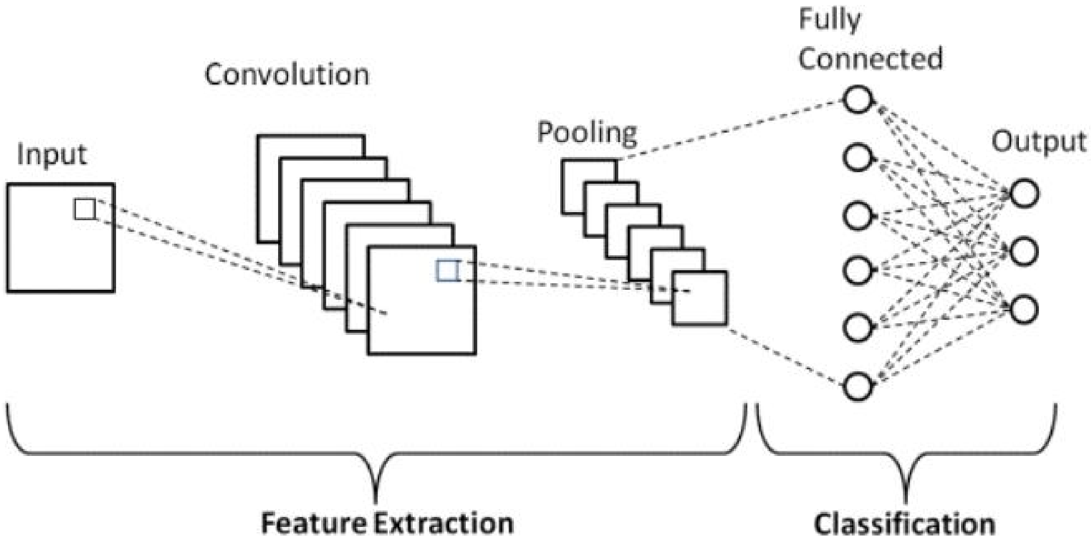


Figure 4 CNN Architecture

The convolutional layer analyzes the weights scalar product position of each input volume to calculate the yield of neurons associated with surrounding input regions. Convolutional neural networks apply the activation from the previous layer to the adjusted straight unit points in a manner similar to a sigmoid. Convolution is a particular type of simple operation used all through the extraction procedure. The input is 10 tensors, which are collections of numbers. A chunk is a compact collection of related, low-dimensional, small-sized numbers. As a result, the kernels were dispersed equally throughout the input density. Every channel in a convolutional layer is convolved over the input's spatial dimensions as the data enters it. Kernel convolution is currently used in many other computer vision algorithms, not simply CNN. By using this method, we can project a small network of numbers onto our image and change it according to its values. The elements of the pattern are found using the following equation, where f is the input image and h is our bit. The line and column lists of the consequence network are validated using m and n, respectively. This is illustrated by the formula that follows: $G[m, n] = (f h)[m, n] = h[j, k]f [m j, n – k]$ After positioning our network over a specific pixel, we use bit pairs of values extracted from the images to repeat each evaluation. We finally added everything up and inserted the result, with the total, in the proper spot within the result.

3.3.2 Pooling Layer

Down-sampling is the best way to characterize pooling, which facilitates layer promotion. In the field of image processing, it's comparable to reducing the determination. Combining the intermediate progressive Conv layers with a Pooling layer in a ConvNet design is not very common. It should be possible to control over-fitting by reducing the entity's borders and computations. The Pooling Layer acts independently on each profundity cut, resizing each spatially sliced depth cut of the data using the maximum operation. The feature map is used in the majority of Max Pooling calculations. The components in a predefined-size portion of a picture are averaged using average pooling. Sum pooling is used to get the total sum of all the items in a particular segment. The Pooling Layer is often used to link the Convolutional and Fully Connected layers.

3.3.3 Dropout

Dropout Typically, when every feature in the training dataset is linked to the FC layer, overfitting may occur. Overfitting is the term used to describe a situation in which a model performs so well on training data that it negatively impacts the model's performance on new data. A dropout layer, which reduces the size of the model by removing a small number of neurons from the neural network during training, is used to address this problem. Dropout improves the performance of a machine learning model by reducing overfitting by making the network simpler. Neurons are eliminated from neural networks during training.

3.3.4 Activation Function

One of the most important components of the CNN model is the activation function. They are employed in the identification and approximation of complicated and continuous links between network variables. Put simply, at the network's end, it decides which model information should advance and which should not. This gives rise to nonlinearity in the network. Activation functions that are often utilized include the Sigmoid, tanH, ReLU, and Softmax functions.

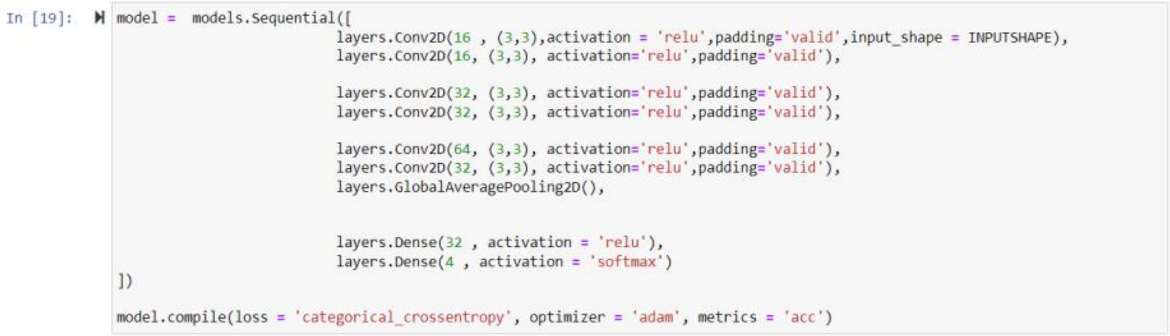
The ReLU, or Rectified Linear Unit: ReLUs are piecewise linear functions that take an input sequence and provide a straight output. There are multiple sizes available (0, 0). In the era of deep learning, the most widely used ReLU is an activation function, which is briefly summarized in the formula below: $y = \max(0, x)$ (2.4) ReLU's primary benefit is that it addresses the problem of fading gradients; TanH, on the other hand, has propagation error of 0% and one-sided activation (50%) but sparse activation. The derivatives of this function are also monotonic. One disadvantage is that this function is not zero-centered and cannot be zerodifferentiable. Another downside is the dying ReLU problem (returned as 0), which arises when half of the results for non-zero-cantered action are idle.

Softmax: A vector of K real values can be converted into a vector of 1 by using the Softmax function. The input data is converted by the Softmax into probabilisticsounding numbers between 0 and 1. Regardless of how little, negative, or large one of the inputs is, it will always fall between 0 and 1, big or positive inputs are translated into big probability by the Softmax, while tiny or negative inputs are translated into small probabilities. Other names for the Softmax function are Softmax and multi-class logistic regression. This is because there is a striking similarity between the equations of the Softmax, a logistic regression extension for multi-class classification, and the sigmoid employed in logistic regression. Only in classifiers where the classes are mutually exclusive can the Softmax function be applied.

3.3.5 Final Sequential Model

Finally, our model satisfies all of the conditions of the CNN (Convolutional Neural Network).

As can be seen below from Figure 5:

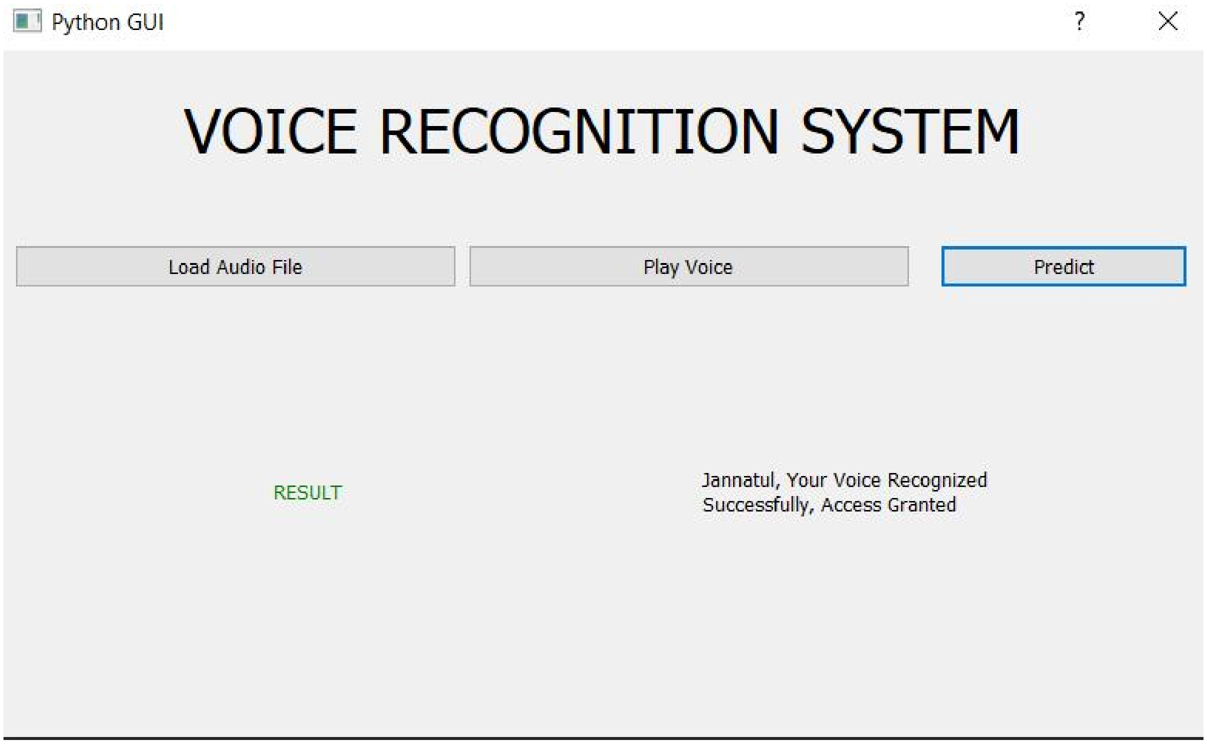


```
In [19]:  model =  models.Sequential([
                          layers.Conv2D(16 , (3,3),activation = 'relu',padding='valid',input_shape = INPUTSHAPE),
                          layers.Conv2D(16, (3,3), activation='relu',padding='valid'),

                          layers.Conv2D(32, (3,3), activation='relu',padding='valid'),
                          layers.Conv2D(32, (3,3), activation='relu',padding='valid'),

                          layers.Conv2D(64, (3,3), activation='relu',padding='valid'),
                          layers.Conv2D(32, (3,3), activation='relu',padding='valid'),
                          layers.GlobalAveragePooling2D(),


                          layers.Dense(32 , activation = 'relu'),
                          layers.Dense(4 , activation = 'softmax')
          ])

          model.compile(loss = 'categorical_crossentropy', optimizer = 'adam', metrics = 'acc')
```

Figure 5 Sequential Mode

3.4 Python Graphical User Interface

The GUI for this system was made using PyQt4, a complete set of Python bindings for Digia's Qt cross-platform GUI toolkit. The GUI contains a predict button and shows the speaker’s name if the system successfully predicts the name. The GUI contains a predict button and if the system successfully predicts the speaker, the speaker's name will display on the screen along with the message "accept granted." otherwise, it will display "accept denied."

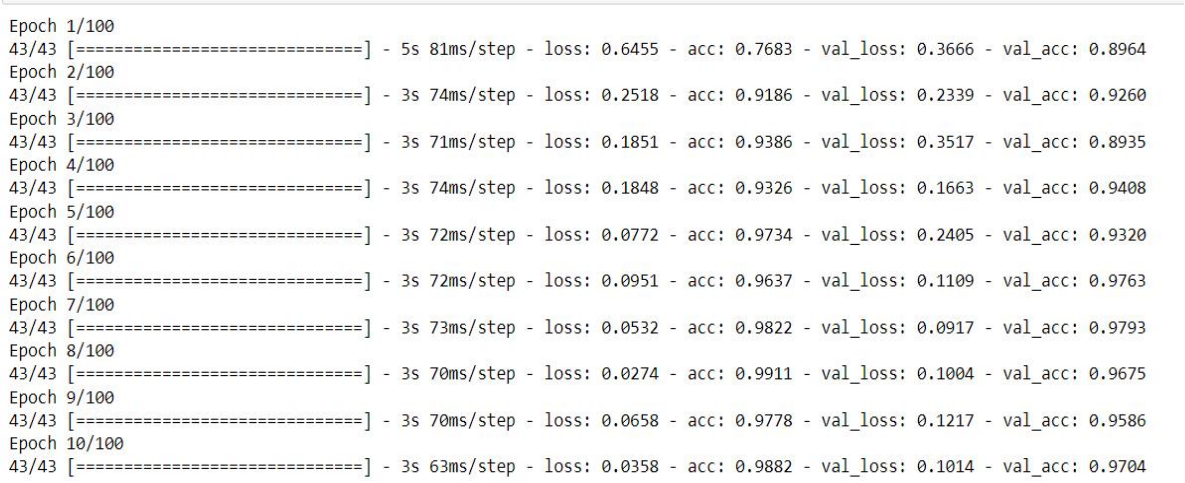


Jannatul, Your Voice Recognized
Successfully, Access Granted

Figure 6 Python Graphical User Interface

Chapter V: Result and Discussion

Three factors must be taken into account while assessing a novel's performance: test, accuracy and loss function, and period. When evaluating accuracy per epoch, it displays the accuracy rate of the specified model; when determining the test loss, it displays the discrepancy between our created and accuracy findings. In order to guarantee that the model was appropriately trained on the provided data to produce the lowest feasible loss function in our experiment, we selected the maximum epoch while utilizing a sufficiently large dataset. We developed a model that has the maximum accuracy possible. Epochs and precision displayed below in Figure 8. Therefore, our accuracy rate of over 98% is far higher than that of other recent studies on this subject. Regarding the detection outcome, our model was able to identify the voice that the machine was testing. Since the text independent speech recognition system that underpins our study topic allows users to talk anything they choose, they are not limited to speaking a certain phrase when they speak. Since we've used a large dataset with about 700 audio clips. There is separate audio content in the audio snippets. In this case, the audio sample is primarily in English, and the dataset includes



```
Epoch 1/100
43/43 [==============================] - 5s 81ms/step - loss: 0.6455 - acc: 0.7683 - val_loss: 0.3666 - val_acc: 0.8964
Epoch 2/100
43/43 [==============================] - 3s 74ms/step - loss: 0.2518 - acc: 0.9186 - val_loss: 0.2339 - val_acc: 0.9260
Epoch 3/100
43/43 [==============================] - 3s 71ms/step - loss: 0.1851 - acc: 0.9386 - val_loss: 0.3517 - val_acc: 0.8935
Epoch 4/100
43/43 [==============================] - 3s 74ms/step - loss: 0.1848 - acc: 0.9326 - val_loss: 0.1663 - val_acc: 0.9408
Epoch 5/100
43/43 [==============================] - 3s 72ms/step - loss: 0.0772 - acc: 0.9734 - val_loss: 0.2405 - val_acc: 0.9320
Epoch 6/100
43/43 [==============================] - 3s 72ms/step - loss: 0.0951 - acc: 0.9637 - val_loss: 0.1109 - val_acc: 0.9763
Epoch 7/100
43/43 [==============================] - 3s 73ms/step - loss: 0.0532 - acc: 0.9822 - val_loss: 0.0917 - val_acc: 0.9793
Epoch 8/100
43/43 [==============================] - 3s 70ms/step - loss: 0.0274 - acc: 0.9911 - val_loss: 0.1004 - val_acc: 0.9675
Epoch 9/100
43/43 [==============================] - 3s 70ms/step - loss: 0.0658 - acc: 0.9778 - val_loss: 0.1217 - val_acc: 0.9586
Epoch 10/100
43/43 [==============================] - 3s 63ms/step - loss: 0.0358 - acc: 0.9882 - val_loss: 0.1014 - val_acc: 0.9704
```

Figure 7 Test Accuracy

both male and female voices to help the model learn to anticipate. Additionally, our model was able to identify the speakers' voices from the audio recordings and sent a confirmation message indicating that the voice recognition was successful. Along with that, it displays the user's authorization to access the door. If the machine recognizes the user's voice, it indicates that access has been allowed; otherwise, it denies entry to the user. Here, we use a bar diagram to demonstrate the accuracy of our model in Figure 7.

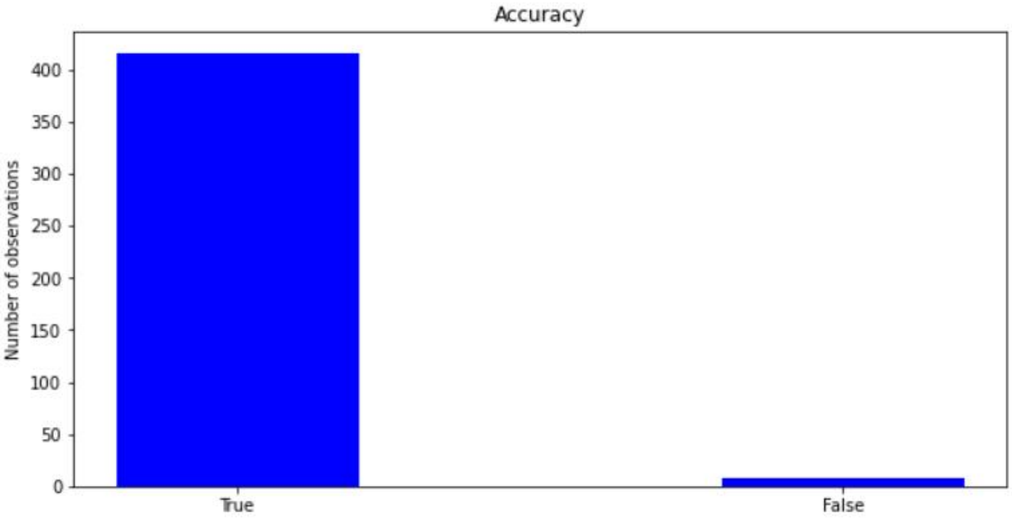
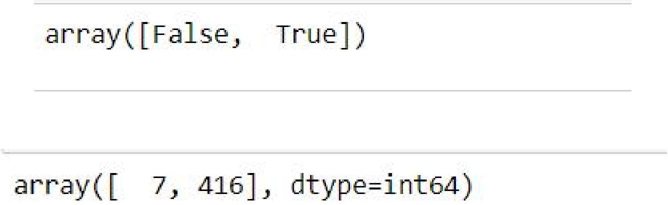


Figure 8 Accuracy in Bar Diagram

Here from the bar diagram, we can clearly see our accuracy. From the dataset it correctly (True) predicts the voice up to 416 records and mispredict (False) only 7 records.



```
array([False,  True])
```

```
array([  7, 416], dtype=int64)
```

Figure 9 Prediction Count

Chapter VI: Conclusion

A thorough analysis of a text-independent voice recognition system for voice-activated door opening is provided by the suggested technique. Our system has the ability to identify user voices. A dataset of over 700 audio clips from four courses has been generated by us. When a user wishes to utilize the door, they submit their request into the local computer through the CNN architecture as soon as the computer recognizes their voice. SoftMax will provide a four-class probability. The architecture assigns the class with the highest probability among the four classes as the output, or predicts that the voice in question belongs to that class. The experimental findings demonstrated that the suggested MFCC functions performed with an overall accuracy of roughly 97%–98%. Additionally, it was discovered that DNN could effectively detect speakers using the

Table 1: Accuracy Comparison of different studies

Reference	Research Name	Accuracy
[4]	Feature Fusion and Deep Neural Network	83.5%–93%
[15]	Hybrid features from a deep belief network	92%
[16]	Hybrid SFX time-series preprocessing and ensemble feature selection	94%
[18]	RASTA–MFCC feature with quadrilateral filter bank structure	97%
[19]	Fast binary features in embedded systems	86%
[23]	MFCC	94%

suggested MFCC functions. In comparison to the other identification model, experimental results show that the suggested speech recognition system is reliable, accurate, and efficient. Positive outcomes demonstrate that a variety of application domains, such as security and access management, can leverage the suggested speech recognition system. Our challenge is figuring out how to add the unknown audio snippets to the algorithm so that the outcome can be improved. In order to enable the model to recognize user-independent voice audio clips and provide accurate predictions, we attempted to build a model based on independent voice input. The model was successful in doing so. Despite our best efforts, we were unable to achieve a 99%–100% accuracy rate with the suggested solution due to time constraints and data limitations, which prevented our model from accurately predicting the known voice in real time. Our goal is to use deep learning with deeper architectures in the future to reduce categorical errors/loss and enhance categorical accuracy. With a larger dataset, we think we can achieve real-time speech recognition with significantly higher accuracy and a better user interface than we now have.

References

[1] Drozdowski, P., Rathgeb, C., Busch, C.: Computational workload in biometric identification systems: an overview. IET Biometrics 8(6), 351–368 (2019)

[2] Jansen, W.: Authenticating mobile device users through image selection. WIT transactions on information and communication technologies 30 (2004)

[3] Chakroun, R., Frikha, M.: A deep learning approach for text-independent speaker recognition with short utterances. Multimedia Tools and Applications 82(21), 33111–33133 (2023)

[4] Jahangir, R., Teh, Y.W., Memon, N.A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M.Z., Ali, I.: Text-independent speaker identification through feature fusion and deep neural network. IEEE Access 8, 32187–32202 (2020)

[5] Chakroun, R., Frikha, M.: Efficient text-independent speaker recognition with short utterances in both clean and uncontrolled environments. Multimedia Tools and Applications 79(29), 21279–21298 (2020)

[6] Morrison, G.S., Sahito, F.H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., Dorny, C.G.: Interpol survey of the use of speaker identification by law enforcement agencies. Forensic science international 263, 92–100 (2016)

[7] Hunt, A.K., Schalk, T.B.: Simultaneous voice recognition and verification to allow access to telephone network services. Acoustical Society of America Journal 100(6), 3488 (1996)

[8] Nunes, J.A.C., Macˆedo, D., Zanchettin, C.: Am-mobilenet1d: A portable model for speaker recognition. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2020). IEEE

[9] Islam, M.A., Jassim, W.A., Cheok, N.S., Zilany, M.S.A.: A robust speaker identification system using the responses from a model of the auditory periphery. PloS one 11(7), 0158520 (2016)

[10] Al-Kaltakchi, M.T., Woo, W.L., Dlay, S.S., Chambers, J.A.: Comparison of i-vector and gmm-ubm approaches to speaker identification with timit and nist 2008 databases in challenging environments. In: 2017 25th European Signal Processing Conference (EUSIPCO), pp. 533–537 (2017). IEEE

[11] Ma, Z., Leijon, A.: Super-dirichlet mixture models using differential line spectral frequencies for text-independent speaker identification. In: INTERSPEECH, pp. 2349–2352 (2011)

[12] Bose, S., Pal, A., Mukherjee, A., Das, D.: Robust speaker identification using fusion of features and classifiers. International Journal of Machine Learning and Computing 7(5), 133–138 (2017)

[13] Ma, Z., Yu, H., Tan, Z.-H., Guo, J.: Text-independent speaker identification using the histogram transform model. Ieee Access 4, 9733–9739 (2016)

[14] Wang, J., Johnson, M.T.: Physiologically-motivated feature extraction for speaker identification. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1690–1694 (2014). IEEE

[15] Ali, H., Tran, S.N., Benetos, E., Garcez, A.S.: Speaker recognition with hybrid features from a deep belief network. Neural Computing and Applications 29, 13–19 (2018)

[16] Fong, S., Lan, K., Wong, R., et al.: Classifying human voices by using hybrid sfx time series preprocessing and ensemble feature selection. BioMed research international 2013 (2013)

[17] Soleymanpour, M., Marvi, H.: Text-independent speaker identification based on selection of the most similar feature vectors. International Journal of Speech Technology 20, 99–108 (2017)

[18] Selva Nidhyananthan, S., Shantha Selva Kumari, R., Senthur Selvi, T.: Noise robust speaker identification using rasta–mfcc feature with quadrilateral filter bank structure. Wireless Personal Communications 91, 1321–1333 (2016)

[19] Laptik, R., Sledeviˇc, T.: Fast binary features for speaker recognition in embedded systems. In: 2017 Open Conference of Electrical, Electronic and Information Sciences (eStream), pp. 1–4 (2017). IEEE

[20] Chetouani, Mohamed, Marcos Faundez-Zanuy, Bruno Gas, and Jean-Luc Zarader. "Investigation on LP-residual representations for speaker identification." Pattern Recognition 42, no. 3 (2009): 487-494.

[21] Xu, J., Wang, X., Feng, B., Liu, W.: Deep multi-metric learning for text-independent speaker verification. Neurocomputing 410, 394–400 (2020)

[22] Fasounaki, M., Y uce, E.B., Onc ul, S., G okhan, I.: A comparative assessment of text-independent automatic speaker identification methods using limited data. Avrupa Bilim ve Teknoloji Dergisi (26), 217–222 (2021)

[23] Liu, J.-C., Leu, F.-Y., Lin, G.-L., Susanto, H.: An mfcc-based text-independentspeaker identification system for access control. Concurrency and Computation: Practice and Experience 30(2), 4255 (2018)

[24] Fong, S., Lan, K. and Wong, R.: Classifying Human Voices by Using Hybrid SFX Time-Series Preprocessing and Ensemble Feature Selection. BioMed research international, 2013(1), p.720834. (2013)

[25] Bhattu, S. Nagesh, Satya Krishna Nunna, Durvasula VLN Somayajulu, and Binay Pradhan. "Improving code-mixed POS tagging using code-mixed embeddings." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 19, no. 4 (2020): 1-31.

[26] Li, Ming, Kyu J. Han, and Shrikanth Narayanan. "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion." Computer Speech & Language 27, no. 1 (2013): 151-167.

[27] Singh, L. and Chetty, G.: A comparative study of recognition of speech using improved MFCC algorithms and Rasta filters. In International Conference on Information Systems,

Technology and Management (pp. 304-314). Berlin, Heidelberg: Springer Berlin Heidelberg (2012)

[28] Jahangir, Rashid, et al. "Text-independent speaker identification through feature fusion and deep neural network." IEEe Access 8 (2020): 32187-32202.