National College of Ireland

Home automation framework through Voice recognition System for home security

MSc Research Project
M.Sc. in Cybersecurity

Vivek Singh
Student ID: 21120315

School of Computing
National College of Ireland

Supervisor: Michael Pantridge

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vivek Singh
Student ID:	21120315
Programme:	M.Sc. in Cybersecurity
Year:	2022
Module:	Research Project
Supervisor:	Michael Pantridge
Submission Due Date:	15/12/2022
Project Title:	Home automation framework through Voice recognition System for home security
Word Count:	4579
Page Count:	14

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	29th January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Home automation framework through Voice recognition System for home security

Vivek Singh
21120315

Abstract

Finding out who is speaking is what speaker recognition is all about. Speaker identification and speaker verification are two categories under which speaker recognition can be categorized. A text-dependent or text-independent method of speaker identification is available. There are different ways and algorithms available for the purpose of voice verification In this paper Mel-Frequency Cepstral Coefficients(MFCC) is used for the purpose of feature extraction and Gaussian Mixture Modeling (GMM) model is used for voice features modelling. In this procedure maximum Loglikelihood feature of voice is being compared for the purpose of decision making.In this study, a secure speaker recognition system is created to match the voice of one speaker with the sound of another speaker with system access permissions, rejecting the alien voice. The novelty which supports this project is when the user with real-time background noise registers its voice is able to successfully match the voice of the user with the registered voices in the system. The results show 70 percent accuracy with more than 34 db of different types of background noise while testing the project.

1 Introduction

Nowadays, the security of the home is becoming a serious concern, and the implementation of access control systems for residential doors that use conventional keys, smart cards, pins, or passwords is currently regarded as less trustworthy. The owner of the card cannot be determined by the card-based access system, which can only manage the identification of the plastic card. Anyone with a smart card can access the door, which cannot detect that the key is being used by the owner or not. The user must input a specified number to gain access in a similar manner to a PIN or password system. A door access system with enhanced security technology is required given this situation.

While talking about biometric technology, a few characteristics of humans cannot be changed or manipulated such as iris data (retina), fingerprints and voice. Therefore, these data can be used for the purpose of identifying the person for the purpose of authentication. Compared to other biometric authentication techniques, voice authentication is considered one of the fastest, most convenient and with very low false positive results therefore voice is considered for this project as an authenticator. We can add different features and machine learning techniques to suppress the background noise and hence a clear voice can be used for that purpose. Chauhan et al. (2019). Better security measures are required in light of the shortcomings of current ones. In this project the novelty is

focused on the noise authentication with real time background noise so that the project can be implemented easily, all the earlier projects have tested the project in a lab environment in which the background noise is less than 34 dB and because of that these cannot be implemented in the real world environment. Therefore to solve this issue research is being done and maximum accuracy is being achieved.

Voice recognition is another name for the voice biometric authentication method, which has two basic stages: classifying and extracting features. Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding are two methods that were formerly used to extract voice features. Through a process known as Linear Predictive Coding (LPC), the DFT amount of the temporal input signal is represented by a coefficient vector that shows the fine spectrum. Utilizing Linear Prediction coefficients alone for voice identification is ineffective since none of the presumptions about the pole transfer function of vocal cords is valid, even though this method is insufficient to discriminate excitation convolution from glottis and pole transfer function.MFCC improves LPC's capacity to describe sound by correcting this problem by replicating human hearing perception to frequency. In Patange and Alex (2017) it is mentioned that the sensitivity of the MFCC takes into account how sensitive human fetuses are to frequencies that are conducive to speaker identification. In order to gather significant sound features, this fashion uses Fourier analysis, placing short-range power gamuts at low frequency and logarithmically at high frequency. There are several classifiers that serve well with speech recognition systems, including the Gaussian Mixture Model(GMM), Vector Quantization(VQ), and Support Vector Machine(SVM). GMM is employed in this design to support the point categorization process for speech recognition operations. The study uses GMM as a classifier for speech features with MFCC birth in a perpetration design for an access control system employing identification software. In the report, different types of conditions are considered so that the accuracy of the matching algorithm can be understood.

This document is structured as follows for the remaining portions: The papers and research effort mentioned in Section 2 are briefly described there. The technique employed for this project's objectives is described in Section 3. The design criteria for this study are covered in Section 4. The design and execution of this study are described in Section 5. Section 6 provides information on the experiments and performance assessments of our system. Section 7 gives an honest evaluation of the research's findings. Section 8 offers this study's conclusions and perspectives. The primary sources that we used for this work are listed in the last section.

2 Related Work

The research papers listed below have a number of shortcomings, which can be concluded after a thorough analysis of them. After going through all the applicable paper and reading the Mel- frequency Cepstral Coefficient(MFCC) Method and statistical portions are employed to prize features of voice recognition and the Gaussian Mixture Model(GMM) for the pattern matching of voice samples are used. The methods which are being used are correct but the result obtained using this can be enhanced and can be applied to real-time environment.

The Recent paper Shofiyah et al. (2022) has provided the result of a 97 percent success rate but the environmental condition which is used for the purpose of the experiment is quite environment (34db noise) which cannot be considered for the application of real-

time application as we cannot expect every time a quite environment every time. Another condition which is suggested for a better result is within a distance of 10 cm the success rate is 93 percent.

The paper Yang et al. (2020) used different methods such as Extreme Learning Machine(ELM), Probabilistic Neutral Network(PNN), Support Vector Machine (SVM), and Back Propagation Neural Network (BPNN), in this paper different methods is being used for the purpose of synthesizing and recognising the processing time. In the research, it is being supported that MFCC and SVM methods for the purpose of speech recognition with low computational requirements when training and testing of the model are done. Therefore with this paper, we can conclude that the MFCC method for the extraction of data is considered the best and can be used for the purpose.

In Meng et al. (2019) research, In order to perform speech emotion identification for SER (Speech Emotion Recognition), the researchers have introduced a novel architecture called ADRNN (dilated CNN with residual block and BiLSTM based on the attention mechanism). They used our advised method to turn emotional dialogues into spectrograms in the experiment, extracting the values of the 3-D Log-Mel spectrums from raw data and feeding them in. The experiment with speakers who are dependent on each other produced a result of 74.96 percent unweighted accuracy and a result of 69.32 percent achieved accuracy. The outcomes show that the suggested networks are capable of learning high-level abstractions and differentiating characteristics of emotional input. In terms of average accuracy, ADRNN networks perform better than other popular feature representations and methods. In the future, the work will continue to focus on model diversity in order to create a flexible framework that can be applied to other speech corpora. They believe that by adopting a multimodal signal, the result will be improved on various emotional analysis tasks.

The research Etienne et al. (2018) to differentiate between emotions in audio analysis, researchers have built a neural network. When extracting high-level characteristics from unprocessed spectrograms, we use convolutional layers, and when aggregating long-term data, we use recurrent layers. Model performance is evaluated using a 10-fold cross-validation, which was better suitable for this dataset. To address scarcity and class imbalance, they used VTLP and minor class amplification to augment data. Using the IEMOCAP dataset, used a neural network to recognize emotions.

A common and serious medical condition called clinical depression, commonly known as Major Depressive Disorder (MDD), is explained in this Rejaibi et al. (2022). With an overall accuracy of 76.27 percent and a root mean square error of 0.4, the proposed method outperforms cutting-edge methods on the DAIC-WOZ database for evaluating depression. Calculations are made based on the Patient Health Questionnaire results and the binary classification of depression. It is suggested to use an MFCC-based Recurrent Neural Network to identify depression and assess its severity levels. It is being looked at how to transfer knowledge from one activity and supplement training data to make up for a lack of training data. In a further development, we'll include data classes for gender recognition and balancing. The MFCC features are extracted after the cordings have been preprocessed. Afterward, a deep recurrent neural network is fed with the MFCC coefficients.

In Park et al. (2019) SpecAugment is explained which is a straightforward data augmentation method for voice recognition. It is directly applied to the feature inputs of a neural network (i.e., filter bank coefficients). Warping the features, masking blocks of frequency channels, and masking time steps comprise the augmentation policy. We out-

perform all previous work by achieving state-of-the-art performance on the LibriSpeech 960 and Swichboard 300h challenges. We obtained 6.8 percent WER on test-other for Switchboard without using a language Afterwards

Rahmandani et al. (2018) explains the auscultation, Auscultation of heart sounds is an important tool for detecting heart disease symptoms. The Artificial Neural Network (ANN) approach is used to classify the outcomes of heart sound extraction. The Michigan Sound Heart Database was used to collect the data. In comparison to the prior study, accuracy has improved. In previous investigations, recognizing 13 varieties of apex heart sounds reached 92 per cent since 11 types of heart sounds were accurately Switchboard. However, two forms of aberrant cardiac sounds cannot be reliably diagnosed since they have nearly identical characteristics.

The study of Wahyuni (2017) addresses a difficult problem:identifying spoken Arabic letters, particularly three letters of the hijaiyah that, although they have similar pronunciations to Indonesian speakers, have different makhrajs in Arabic. The feature extraction method used in the study is Mel-Frequency Cepstral Coefficient (MFCC), while the classification method is Artificial Neural Network (ANN). With an average accuracy of 92.42 percentage, the MFCC-based feature extraction and ANN classification approach may produce a better recognize. The recognition of hijaiyah speech has become a difficult topic in learning algorithms. The extraction and classification methods' capability in the learning process are two critical factors in successful voice recognition.

In Tirumala et al. (2017) the method of identifying the speaker from a given utterance is known as "Speaker Identification," and it involves comparing the voice biometrics of the utterance with those utterance models that have been previously saved. One of the most crucial SI components, feature extraction has a big impact on how the system works and how well it performs. According to the findings, techniques that solely rely on Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction have been employed more frequently than any other methodology. The numerous characteristics extraction techniques and speaker recognition algorithms are presented in detail in this study. This analysis took into account about 190 publications that were published between 2011 and 2016. Regardless of the method used for the speaker classification step, the majority of feature recognition approaches use some variation of MFCC. The Kitchenham systematic review methodology was used to evaluate about 190 publications between 2011 and 2016 for this investigation. Future research will go into creating a reliable, all-encompassing speaker identification system that tackles the major issues of Speaker identification. This universal framework is propose in this paper. multiple languages can be modelled. Include the capacity to deal with noisy and all-channel data.

In FRANCIS and Alimi (2014) a security system for the building is proposed that uses voice activation to operate doors. Speech cannot be taken, copied, lost, forgotten, or accurately predicted, which is why it has been proposed. For feature extraction and pattern matching in the suggested system, respectively, the Mel Frequency Cepstrum and Gaussian Mixture Model are used. Using inexpensive and readily accessible parts, the door control system could be put together quickly. When compared to other access control systems currently in use, analysis of the results revealed an accuracy of more than 80 percent.

In Bagul and Shastri (2013) Recognizing and categorizing various speakers' speeches is the project's main objective. Mel Frequency Cepstral Coefficients (MFCCs), which are important features that may be extracted from those people's speech signals and used in this classification, are one of the main features that are generally used. Another suggestion

for enhancing speaker recognition effectiveness is to use the fractional Fourier transform to extract features. Results of the experiments on the created database show that Mel scale-based algorithms outperform FrFT-based algorithms in clean databases, but FrFT-based algorithms outperform them better in noisy environments. In order to successfully recognize speakers from distorted, unrestrained speech, gaussian mixture speaker models offer a reliable speaker representation. The models are simple to implement on a real-time platform and computationally cheap.

In Kinnunen et al. (2006) it is explained that When identifying unknown speakers, the major proposed computation comes from calculating the disbowelled likelihood between their feature vectors and the database's models. This study focuses on speaker recognition using vector quantization (VQ). The most effective variations are then translated to modelling based on Gaussian Mixture Model (GMM). A vector quantization-based speaker identification system (VQ) has been presented. The suggested method is quicker than the widely utilised nearest impostors method. It is proved that the techniques developed for VQ modelling apply to GMM modelling. It reduced the vectors using quiet detection and pre-quantization and reduced the speakers using speaker pruning. With little to no accuracy loss, clustering can be used to remove redundant information from the test sequence.

3 Methodology

The system architecture employs the MFCC approach for feature extraction and the GMM method for feature matching. Gathering and building a database of homeowners' voices, and extracting speech attributes, this design includes creating machine learning programme models for identifying homeowners' identities.

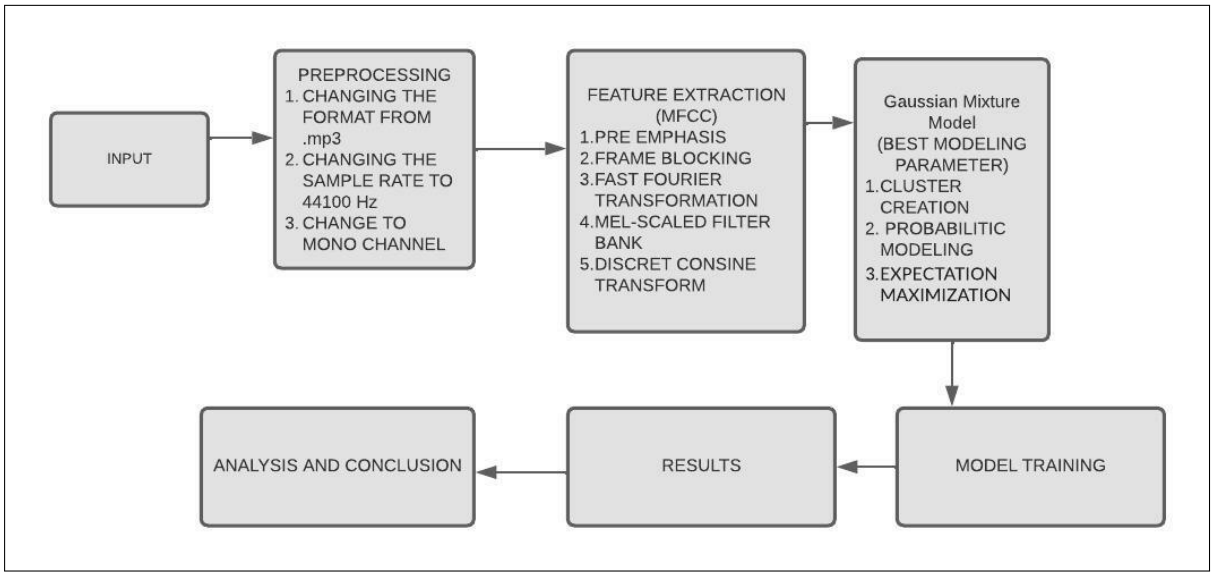

Figure 1: Research Methodology Overview

The system contains the input phase in which the audio sample of 10-20 people is collected and it is then converted to .wav format from .mp3 format, and then the data is transferred to the preprocessing stage for further feature extraction. The Mel Frequency Cepstral Coefficients (MFCC) approach will be used to search for characteristics or fea-

tures in the sound. The Gaussian mixture clustering approach will be used to match the MFCC extraction findings with the sound pattern.

1. **Pre-processing:** In the pre-processing stage we recorded the sound using the PRAAT tool. The voice is recorded at 44100 Hz in mono-channel form. A mono channel is used to streamline files in order to speed up computation in the following stage. The audio file is saved in form of .wav format and then extracted from the MFCC features using the PRAAT tool itself.

2. **Mel-Frequency Cepstral Coefficient:** Mel Frequency Cepstral Coefficients (MFCCs) are a component that is frequently utilized in speaker and automatic voice recognition Hatala (2019). A sound waveform model that accurately depicts the sound is what is intended here. While the properties of the sound signal are known to fluctuate over time, it has been shown that the sound has a reasonably stationary characteristic when reflecting noticeable variances over a shorter time frame (between 5 and 100 milliseconds). Because of this, the short-time spectrum analysis method known as the Mel Frequency Cepstral Coefficient (MFCC) method has gained popularity for extracting the features of the static sound.

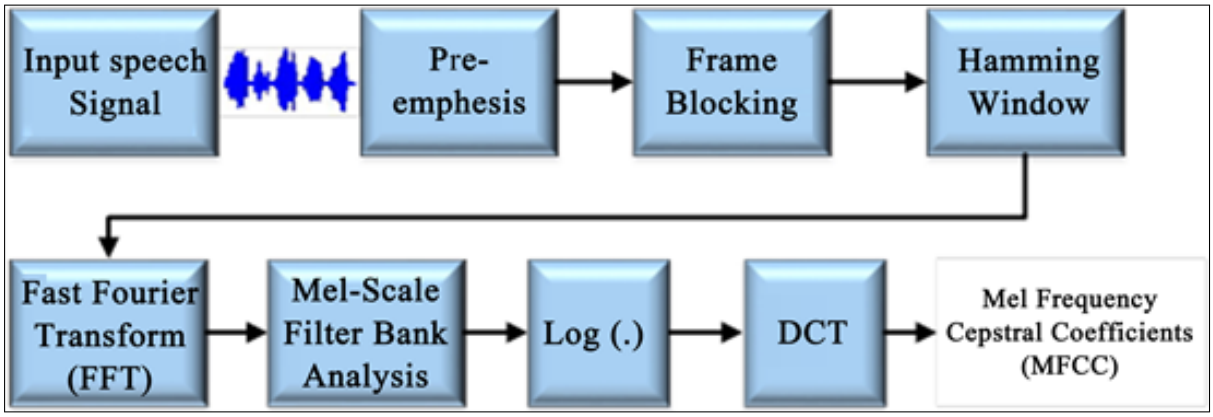

Figure 2: Block schematic for the MFCC

A 10ms overlap was added after the sound signal was windowed and frame-locked for 25ms. A total of 26 filter banks are used to process the frequency components of the framed signal. A discrete cosine transform is used to beautify the cepstral coefficients and compress the filtered output after a logarithmic function has compressed it (DCT).MFCC findings are considered the first 13 DCT block outputs.

3. **Gaussian Mixture Model:**An approach for machine learning called a Gaussian mixture model (GMM), They are employed to divide data into various groups in accordance with the probability distribution. To characterize the distribution of feature vectors in probability space, GMM combines a number of Gaussian probability density functions. The maximum likelihood estimation function is mostly used in speaker training by the GMM technique to determine the model's parameters Zhang and Yao (2020). Using the following formula and assuming that the training vector set is X x1, x2, x3,..., xi, one can determine the likelihood function of the Gaussian mixture model.:

$$p(X \mid \lambda) = \prod_{i=1}^{I} p\left(x_i \mid \lambda\right) \tag{9}$$

In order to increase the likelihood function of the GMM model, maximum likelihood estimation is used to identify the optimum model parameter. In the phase of speaker recognition, initially extract the associated I-dimensional feature parameters by using the following formula:

$$p\left(x_i\right) = \sum_{i=1}^{M} \alpha_i p_i\left(x_i\right) \tag{10}$$

Whenever p(x) and a(x) meet the criteria in the formula:

$$p_i\left(x_i\right) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} \exp\left\{ \frac{\left(x_i - \mu_i\right)^T \sum_i^{-1} \left(x_i - \mu_i\right)}{2} \right\} \tag{11}$$

$$\sum_{i=1}^{M} \alpha_i = 1 \tag{12}$$

The mean and covariance matrix can be used to describe each component of the Gaussian mixture model, as can be seen from the calculation above.

4 Design Specification

1. **Data collection:** The study's subjects were willingly invited to participate. Each of these individuals signed a consent form before the experimentation began, attesting to the voluntary nature of their involvement and their consent to the collection of all the information. They were asked to record the five voice samples, based on different voice levels so that I can train the model based on different types of voice amplitudes. Post to this voice is recorded using the Audacity tool in which the frequency of the sound is selected as 44100 and the mono audio option is chosen for the purpose.

2. **Background noise:** The voice signal interference caused by ambient noise was recorded. Based on different kinds of noise we have recorded the samples and trained them using the model.

   (a) Convolutional noise: The reverberation caused by enclosed spaces is referred to as convolutional noise or convolutional distortions. In an enclosed room, sound waves that are being produced by a speaker are reflected off of walls and other objects before reaching the microphone. A colourful, echoey, or reverberant sound will be produced when it is recorded at the microphone. More reverberant sounds will be recorded in larger enclosed spaces.

(b) Additive noise: It is a type of noise which contains the sound of a vacuum cleaner, air conditioner noise, crying of a baby and people talking in the background. As these noises are combined with the noise of the original speaker these types of noise are termed additive noise

(c) Nonlinear distortion: When the speaker and microphone are too close together or the volume of the device is too high, nonlinear distortion occurs.

5 Implementation

The final model used to accomplish the research goals is included in this section. For speaker recognition, two terms are relevant **speaker verification** and **speaker identification.** In order to achieve this, a machine-learning pipeline with a variety of features is created. For example, the MFCC feature is extracted, and the features are then compared using a GMM model. In the identification task, an unknown speaker X is compared to a database of known speakers, with the best matching speaker being given as the identification result. For the purpose of the verification task the stored value of the legitimate speakers was compared with the authorized speaker and then the decision for the acceptance and rejection is calculated. The figure below provides a visual representation of the same.

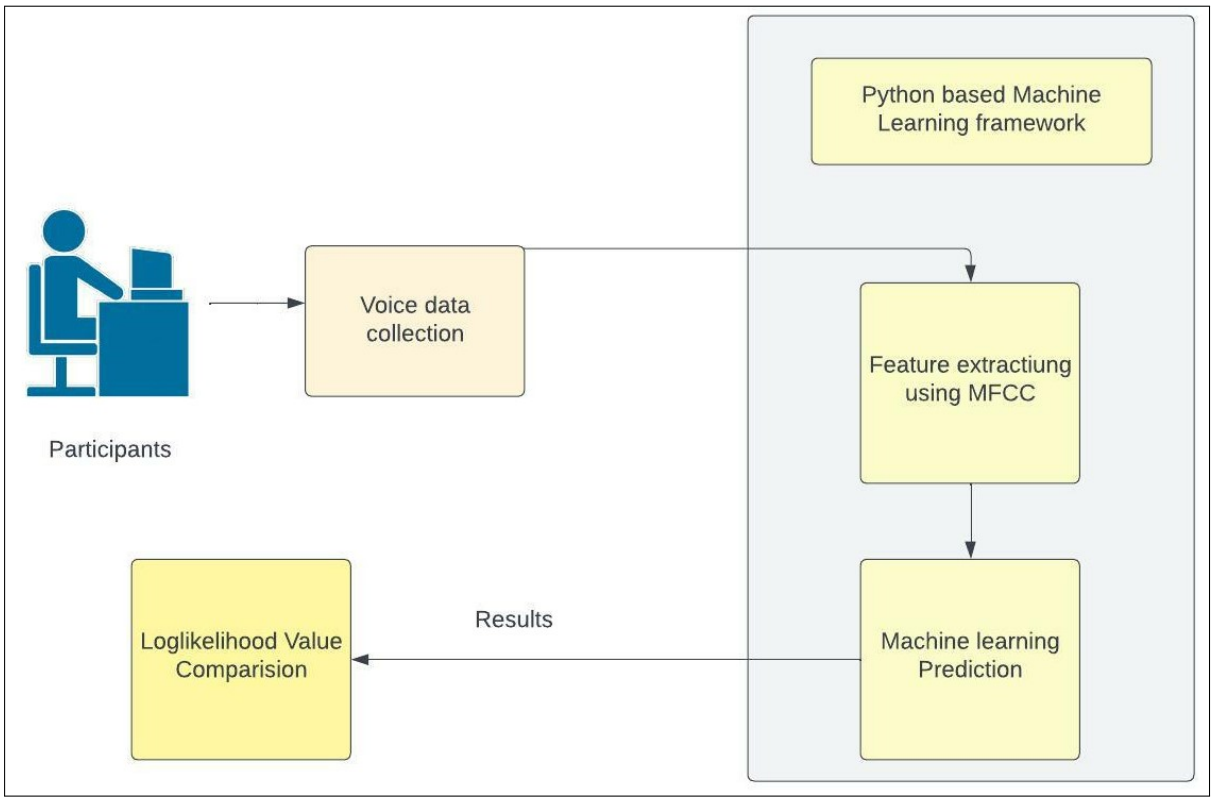

Figure 3: Implementation overview

1. **Creating a Voice Database:** The steps that are taken to establish a voice database are:

Voice recording: Sound captured with a microphone that both fits the test situation and the desired circumstances.As it does not differentiate between test conditions and training conditions. For instance, if the recording was done in a peaceful environment, but testing is done in a noisy environment, the testing accuracy will undoubtedly suffer.

The voice database's labelling: Then, subwords or labels are used to allocate names to each recorded file. Audacity and PRAAT are two programs that are used to label voice, record, and extract MFCC features.

2. **Extraction of feature:**Then, audio files are transformed into feature format. The Mel Frequency Cepstral Coefficient is the format that is extracted from the voice sample recorded. Following that, the Phase auto-correction approach is used to filter the 39 coefficient auto-correction that the input sound has been distorted by noise, this will boost the recognition accuracy.

3. **Training the model:**Six speakers total were used as training data, with each speaker uttering five words in a sentence that was practiced five times. In training, 30 data sets made up the practice of speech data. Two different models is being trained for the purpose of this project, one with audio with high frequency and the other with the audio with low frequency.

4. **Log-likelihood Comparison:**Regression model quality of fit is evaluated using the log-likelihood value. A model's ability to fit a data set is directly correlated with the log-likelihood value. For a certain model, the log-likelihood value can be either negative or positive infinity. In this model the negative value is being considered.

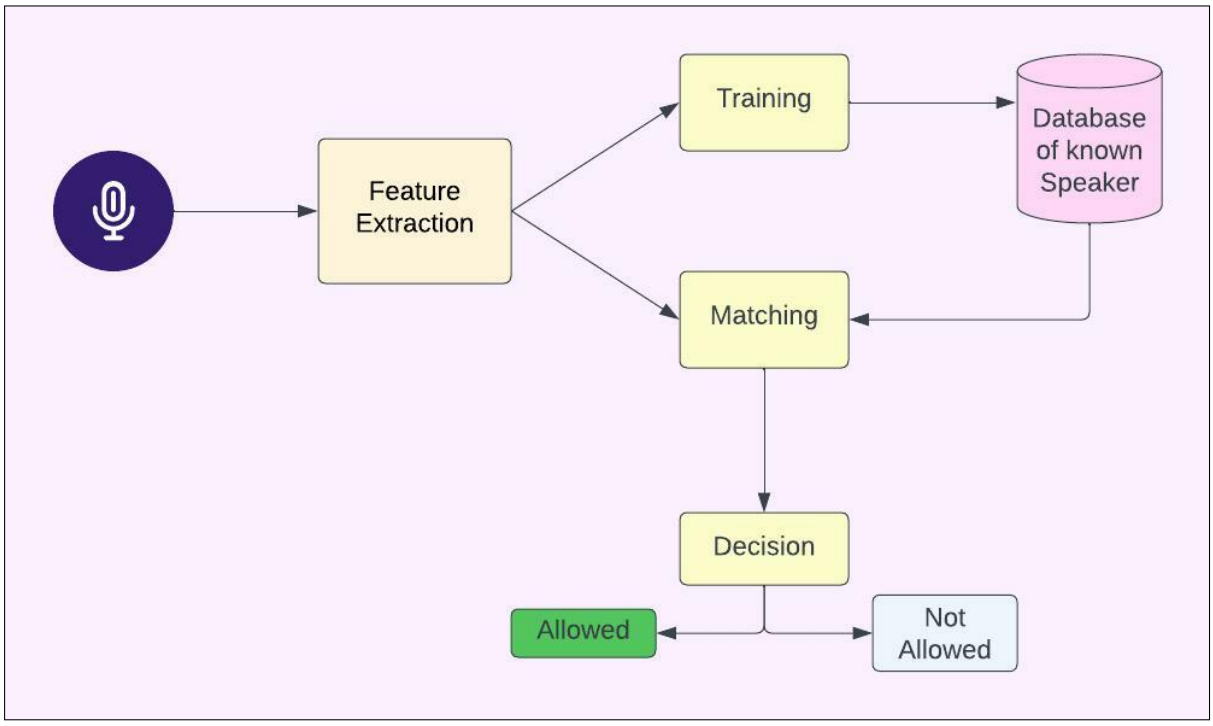

Figure 4: Decision flow diagram

From figure 4, we can see that once the voice is being recorded and the features are

being extracted the voice feature goes to the GMM matching algorithm and check for the voice feature present in the database and try to compare the log-likelihood value with the same, Upon comparison with the stored value, it will allow or not allow the user, whereas if the value does not match any of the stored value in the database then it will reject the voice and not allow that person.

6 Evaluation

The experiment is conducted with over 20 participants of which two participants were allowed to access the system whereas if the voice of other participants is being detected they were not allowed. At the start of the experiment it was done based upon different qualities of voice whereas now to enhance the model and check the accuracy of the model, the voice with different background noise is being also tested such as in quite an environment, with different noise in the background as well. Whereas the accuracy is achieved at 70 percent during different noises in the background as well.

The audio sample of three sound records is shown in the picture below.

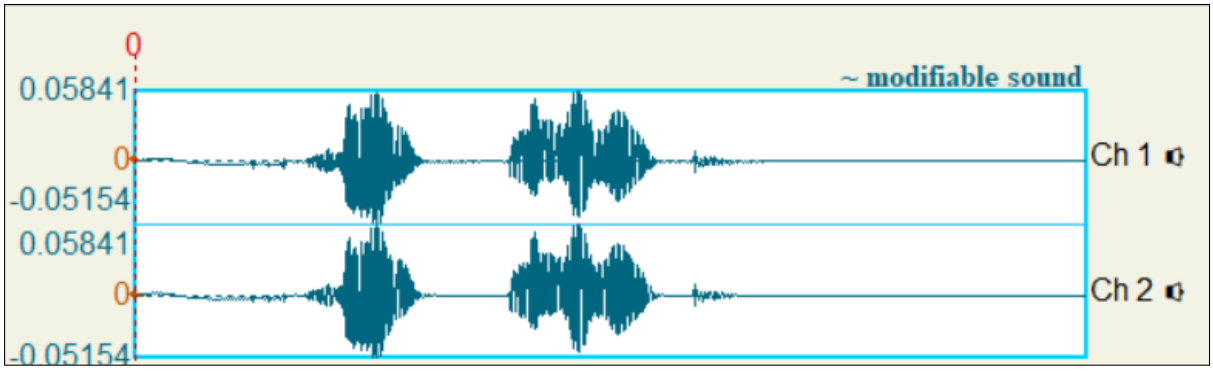

Figure 5: Digital speech signal of User 1

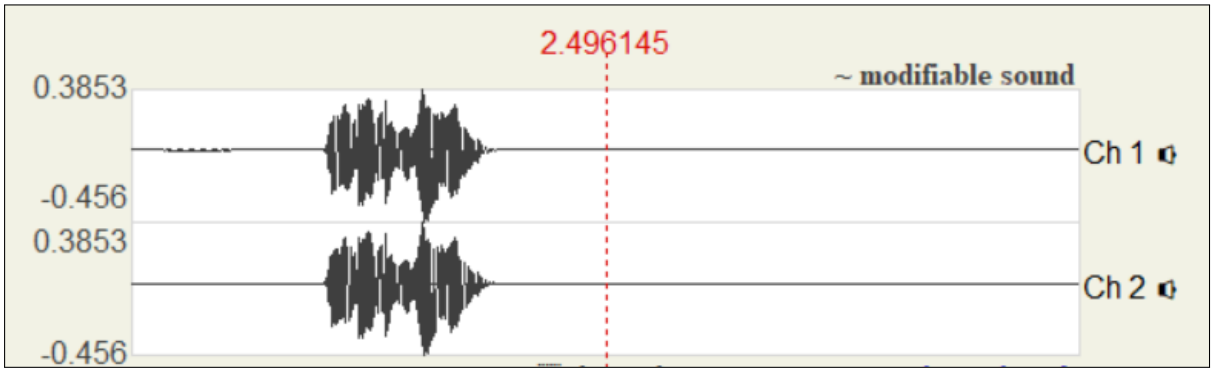

Figure 6: Digital speech signal of User 2

12 male and 8 female speakers between the ages of 20 and 28 supply the same phonetically balanced sentence utterances for the database used for system evaluation. The recordings for this database were made using the same microphone for all speakers across all sessions, in various places with various acoustic environments. An average sentence duration was about 3.5 seconds. The speaker-specific model is trained over a time period

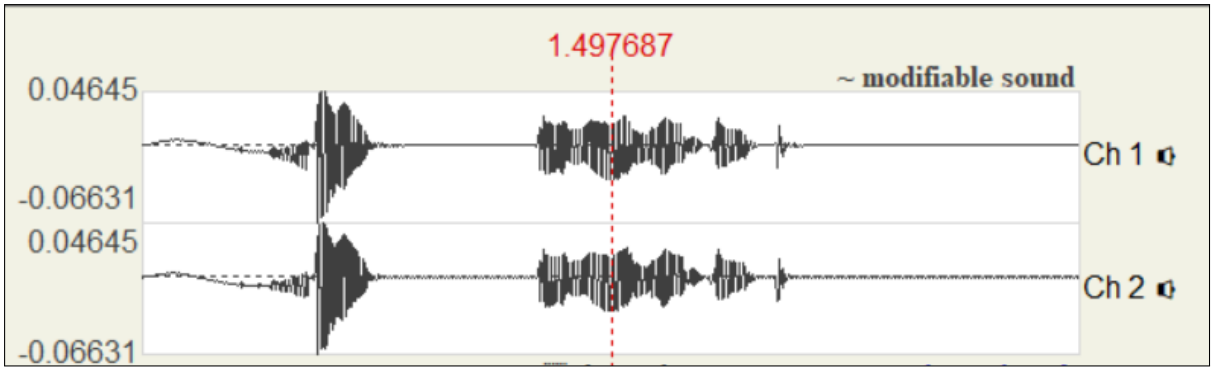

Figure 7: Digital speech signal of User 3

of 5 to 10 seconds and then tested over a period of 3 to 5 seconds for a selection of sentences. To assess the systems, a number of experiments were done. While conducting the experiment we have considered three different types of noise in the background so that system can be trained and tested. As we know not all methods work for every sort of noise interference, and some aspects of a loud setting are harder to fix. There are different types of noise available that usually frequently interferes with speech recognition technology includes.

6.1 Different Genders' Identification Rates

The experiment is performed on 12 males and 8 females so that the analysis of gender identity can also be done based on the background noise.

Table 1: Identification rate table

Type of speaker	**Success rate in Percentage**
Male	70
Female	75

It can be seen from the above table that the identification rate of the female speakers are higher than the male speaker,here out of 8 female speaker 6 were matched perfectly. Therefore we can say that the training model provides better results with the female voice.

6.2 Rate of speaker identification for various speaker counts

As we can see in the above table, as the number of speakers increases for the experiment, the rate of identification decreases. This is due to the fact as the number of speakers increases, the identification rate and log likelihood value of the system overlapped due to which the identification rate decreases.

6.3 Rate for the rejection of unauthorized users

The experiment is done on 20 participants, in which two users were authorized, where as rest of the users were not authorized. The rate for rejection for unauthorized users are the most crucial thing which will make the model more acceptable.

Table 2: Speaker count based identification percentage

Number of speakers	**Identification rate (in percentage)**
4	95
8	90
12	80
16	75
20	70

Table 3: Identification rate table

Speakers(Registered and Unregisterd)	**Identification Rate**
Authorized Speakers	75
Unauthorized Speakers	70
Known Speakers	83
Unknown Speakers	70

From the above results, we can see that the success rate for the unknown speaker is 70 percentage, which was not registered in the system. For this purpose of this 3 unknown speakers were selected to record their voices and tried ten times out of which seven times the system rejected their voices.

6.4 Experiment based upon different types of noise in the background

Table 4: Noise based identification percentage

Types of background noise	**Identification rate (in percentage)**
Additive noise	75
Convolutional noise	72
Nonlinear distortion	83

For case 1, the nonlinear distortion noise background is being considered, where the participants were asked to record their voice from a closer distance with the microphone and speak in a high pitch. six participants voice is being record out of which 5 voices were perfectly matched, which shows 83 percentage of accuracy. Similarly, 8 participants were asked to record the voice with the sound of fan in the background, out of which 6 voices were matched perfectly.

6.5 Discussion

An evaluation was conducted and took into account a few aspects to make sure the suggested system is implemented in accordance with the goal. The previous related work of this project was taken into consideration and tried to solve the real-world problem. The major purpose of this project is to solve the issue of voice authentication with background noise, which is being successfully established with this project. Three different

participants' voices are being recorded and considered for the purpose; for better training and results, some percentages with different background noise can also be considered for better results. While testing the model with the unknown voices which are not being used in the database we observed that the accuracy received for rejecting the unknown voices is 60 percent whereas the users which are registered as not allowed users were able to receive an accuracy of 80 percentage.

7 Conclusion and Future Work

Utilising the Mel-frequency Cepstral Coefficients and the Gaussian Mixture Model (GMM), a speaker recognition system has been developed in this study (MFCC), the objective of the research is to develop a model which can be used in a real-world environment with noise in the background as well, as all the models which are being developed earlier were developed based upon the environment with no noise in the background. The data from 20 people are being collected and the model developed is 70 percent successful. As the number of speakers rises from 4 to 20, the recognition rates drop by up to 12 percent. Compared to female speakers, male speakers in the system have a higher identification rate.

For future work, facial recognition can be added to the voice recognition system including the distress in a voice so that if at any point in time, the system observes that there is any kind of distress in the voice and body language of the user it can provide the access or deny the access for the user and inform the relevant authorities about the situation which is being recorded through the system.

References

Bagul, S. G. and Shastri, R. K. (2013). Text independent speaker recognition system using gmm, *2013 International Conference on Human Computer Interactions (ICHCI)*, pp. 1–5.

Chauhan, N., Isshiki, T. and Li, D. (2019). Speaker recognition using lpc, mfcc, zcr features with ann and svm classifier for large input database, *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, IEEE, pp. 130–133.

Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L. and Schmauch, B. (2018). Cnn+ lstm architecture for speech emotion recognition with data augmentation, *arXiv preprint arXiv:1802.05630* .

FRANCIS, A. K. and Alimi, I. A. (2014). Voice-based door access control system using the mel frequency cepstrum coefficients and gaussian mixture model, *Global Journals of Research in Engineering* **14**(F4): 39–42.

Hatala, Z. (2019). Speech recognition for indonesian language and its application to home automation, *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 323–328.

Kinnunen, T., Karpov, E. and Franti, P. (2006). Real-time speaker identification and verification, *IEEE Transactions on Audio, Speech, and Language Processing* **14**(1): 277–288.

Meng, H., Yan, T., Yuan, F. and Wei, H. (2019). Speech emotion recognition from 3d log-mel spectrograms with deep learning network, *IEEE access* **7**: 125868–125881.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D. and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition, *arXiv preprint arXiv:1904.08779* .

Patange, P. P. and Alex, J. S. R. (2017). Implementation of ann based speech recognition system on an embedded board, *2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2)*, IEEE, pp. 408–412.

Rahmandani, M., Nugroho, H. A. and Setiawan, N. A. (2018). Cardiac sound classification using mel-frequency cepstral coefficients (mfcc) and artificial neural network (ann), *2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE)*, IEEE, pp. 22–26.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S. and Othmani, A. (2022). Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech, *Biomedical Signal Processing and Control* **71**: 103107.

Shofiyah, Z., Mahmudah, H., Santoso, T. B., Puspitorini, O., Wijayanti, A. and Siswandari, N. A. (2022). Voice recognition system for home security keys with mel-frequency cepstral coefficient method and backpropagation artificial neural network, *2022 International Electronics Symposium (IES)*, pp. 497–501.

Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S. and Wang, R. (2017). Speaker identification features extraction methods: A systematic review, *Expert Systems with Applications* **90**: 250–271.

Wahyuni, E. S. (2017). Arabic speech recognition using mfcc feature extraction and ann classification, *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, IEEE, pp. 22–25.

Yang, N., Dey, N., Sherratt, R. S. and Shi, F. (2020). Recognize basic emotional statesin speech by machine learning techniques using mel-frequency cepstral coefficient features, *Journal of Intelligent & Fuzzy Systems* **39**(2): 1925–1936.

Zhang, N. and Yao, Y. (2020). Speaker recognition based on dynamic time warping and gaussian mixture model, *2020 39th Chinese Control Conference (CCC)*, pp. 1174–1177.