**Audio Spoof Detection Integrated with a Home Automation System Using Iot**

MAIN PROJECT REPORT

***Submitted by***

# GOURAV GOPAL 1905015

# NALIN SURIYA S 1905031

# VISHAL KARTHIK S 1905060

**YOKESH R S 1905062**

***in partial fulfilment for the award of the degree of***

**BACHELOR OF ENGINEERING**

**in**

**COMPUTER SCIENCE AND ENGINEERING**



**COIMBATORE INSTITUTE OF TECHNOLOGY**

***(Government Aided Autonomous Institution Affiliated to Anna University)***

**COIMBATORE – 641 014**

**ANNA UNIVERSITY – CHENNAI - 600 025**

**MARCH 2023**

COIMBATORE INSTITUTE OF TECHNOLOGY

***(A Govt. Aided Autonomous Institution Affiliated to Anna University)***

**COIMBATORE – 641 014**

**BONAFIDE CERTIFICATE**

Certified that this project **“AUDIO SPOOF DETECTION INTEGRATED WITH AN HOME AUTOMATION SYSTEM USING IoT”** is the bonafide work of **GOURAV GOPAL (1905015), NALIN SURIYA S (1905031), VISHAL KARTHIK S (1905060), YOKESH R S (1905062)** under my supervision during the academic year 2022-2023.

**SIGNATURE SIGNATURE**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dr. G. KOUSALYA, M.E., Ph.D., Dr.M. MOHANAPRIYA, M.E., Ph.D**

**HEAD OF THE DEPARTMENT ASSOCIATE PROFESSOR**

Department of CSE Department of CSE

Coimbatore Institute of Technology Coimbatore Institute of Technology

Coimbatore – 641 014 Coimbatore – 641 014

Certified that the candidates were examined by us in the **19CS65 Main Project** viva voce examination held on \_\_\_\_\_\_\_\_\_\_\_\_.

**Internal Examiner External Examiner**

Place :

Date :

|  |  |  |
| --- | --- | --- |
| CHAPTER NO. | TITLE | PAGE NO |
|  | ACKNOWLEDGEMENT |  |
|  | ABSTRACT |  |
| 1 | INTRODUCTION   * 1. INTRODUCTION   2. OBJECTIVE   3. PROBLEM STATEMENT   4. GOOGLE TREND ANALYSIS |  |
| 2 | LITERATURE SURVEY |  |
| 3 | SYSTEM SPECIFICATION  3.1 SOFTWARE SPECIFICATION  3.2 HARDWARE SPECIFICATION |  |

**ACKNOWLEDGEMENT**

**ACKNOWLEDGEMENT**

We whole-heartedly thank the management of Coimbatore Institute of Technology for providing us the necessary infrastructure which was very much helpful to complete our project.

We heart fully thank our Secretary **Dr.R.Prabhakar** **B.Tech.,(IIT, Madras), M.S.(OSU.,USA), Ph.D.(Purdue, USA),** Advisor **Dr. V. Selladurai**, **M.E., Ph.D.,** and our principal **Dr.A.Rajeswari, M.E., Ph.D.,** for providing us with the necessary facilities which was useful in the completion of our project.

We sincerely thank the efforts taken by **Dr.Kousalya G, M.E., Ph.D.,** Professor and Head, Department of Computer Science and Engineering, Coimbatore Institute of Technology, for her valuable ideas in completing our project efficiently.

We also express our profound gratitude to our supervisor **Dr.M.Mohanapriya,M.E.,Ph.D** Associate Professor, Department of Computer Science and Engineering, Coimbatore Institute of Technology, for guiding us to carry out our project successfully.

We also express our sincere thanks to our Senior Tutor, Tutors and Project Coordinators, Department of Computer Science and Engineering, Coimbatore Institute of Technology, for guiding us all through the successful completion of our project.

We also extend our thanks to all our teaching faculty and non-teaching staff of our department for their kind attitude towards us all through our project work.

**ABSTRACT**

**ABSTRACT**

Authentication has become very important aspect of our day to day lives starting from normal lock screen pin to human retina based authentication systems. One among those popular and complex authentication systems are the audio based authentication systems where in people use certain words to unlock devices and objects like mobiles, doors etc., Audio Authentication generally involves authentication based on words and voices. Issue in the existing system is that the system verifies and extracts the features of words and voices but it does not classify human voice and recorded human voices.

The proposed system overcomes the issue of audio spoof attack and recognizes the genuineness of the voice using model like RNN and LSTM for classification. This system can be further integrated to any Iot system or Home automation systems , adding more security to the accessibility of the device.

**INTRODUCTION**

**CHAPTER 1**

**INTRODUCTION**

* 1. **INTRODUCTION**

Development of VCD’s, have boosted the realization of smart homes, voice-controlled authentication systems etc.,.These VCDs are vulnerable to different spoofing attacks. Audio Authentication is becoming very essential part of our lives. Audio Authentication spoofing is becoming an issue. High-quality audio recorders enable bypassing this audio authentication system by just recording the human voice and reusing them for accessing the same system. Thus, there exists a need to develop a voice anti-spoofing framework capable of detecting multiple audio spoofing attacks.

* 1. **OBJECTIVE**

1. To develop a Deep-Learning Enabled Audio Spoof Detector which uses Recurrent Neural Networks(RNN) and Long Short Term Memory(LSTM) to classify human voice from recorded voices and integrating the system with an Iot system.

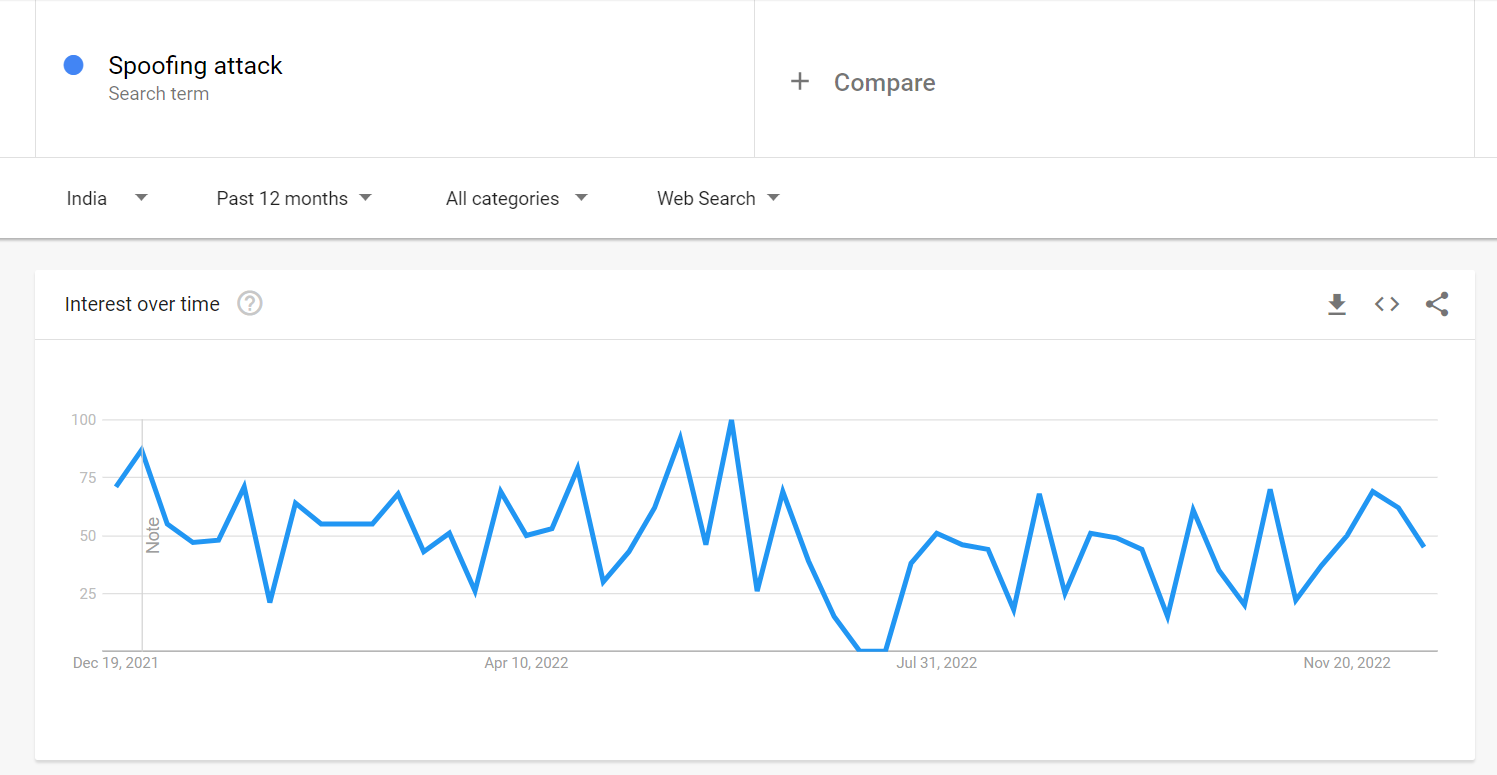
1. The proposed model will defend from the following attacks:
2. Replay attack
3. Voice conversion attack
   1. **PROBLEM STATEMENT**

The increase in the advancement of audio editing software’s , provide an easy way to the accessibility of voice controlled authentication systems which makes the Voice controlled devices(VCD) vulnerable for audio spoof attacks.

Audio spoof attack is the manipulation of genuine signal through recording or modifying to trick an audio verification system.

* 1. **GOOGLE TREND ANALYSIS**

The diagram shown below displays the Google trend analysis about the domain of the project.



*Figure 1.1 Google Trend analyses of spoof attacks over the time.*

**LITERATURE SURVEY**

**CHAPTER 2**

**LITERATURE SURVEY**

# 2.1 Fusion of Belts and GMM-UBM Systems for Audio Spoofing Detection (2019)

**Authors:** Ivan Rakhmanenko , Alexander Shelupanov , Evgeny Kostyunchenko

#### Description

#### In this study, Bidirectional Long Short Term Memory (BiLSTM) networks with constant Q cepstral coefficients (CQCC) are used to classify real audio from fake audio in anti-spoofing systems.

#### By fusing the BiLSTM and GMM-UBM systems, a fusion mechanism is used to increase the variability of the systems' decision-making processes and their accuracy.

#### Over the baseline systems, these proposed systems significantly improved performance.

#### Advantages

* Uses time domain dependency relation in audio and gives better results.
* Uses fusion mechanism to improve accuracy.

**Disadvantages**

* It doesn’t provide classification model for classification of various attacks.

# 2.2 IoT based speech recognition system(2022)

**Authors:** Kishor Kumar Sethy, L. M. Varalakshmi, Rajkumar E

#### Description:

#### In this paper, speech recognition based IoT system is carried out.For IoT system Raspberry pi board is used and for wifi communication ESP-32 module is used. For speech recognition Support vector machine model is used for training the speech recognition model.

#### Advantages:

* ESP-32 - Wifi mouse is efficient for data transmission.
* Raspberry pi - Iot system is powerful.

#### Disadvantages:

* Does not prove to be much effective in classification of audio spoofed data.

**2.3 Comparison of VQ and GMM for Text Independent Speaker Identification System for The Bengali Language**

**Authors:** Md Mahadi Hasan Nahid , Md Ashraful , Islam , Md Saiful Islam

**Description:**

#### Speaker identification (SI) is the system to identify the person by the signal pattern of their voices. With many speaker identification models have been proposed, but till now speaker identification technology do not reach their full potential. This paper presents a comprehensive comparative study of VQ and GMM to identify the speaker who speaks in Bengali accent. We consider the problem of text-independent speaker identification. We compare the performance/accuracy of VQ and GMM based Speaker Identification System (SIS). They’ve used Mel Frequency Cepstral Coefficients (MFCC) and Liner Predictive Coding Coefficients (LPCC) for feature extraction.

#### These extracted features are then sent to the Convolutional Neural Network as input and then are classified as either synthetic or replay attacks.

#### Advantages:

* The method GMM+LPCC and VQ+LPCC are very fast.
* The method (GMM+LPCC) gives tremendous improvement over the method (GMM+MFCC), and it can detect the correct speaker from much shorter speech samples.

**Disadvantages:**

* Method like VQ+MFCC is highly accurate but slow and it can be applied in security purpose where the number of users is limited.

**2.4 Detection of Various Speech Forgery Operations Based on Recurrent Neural Network (2020)**

**Author:** Diqun Yan and Tingting Wu.

**Description:**

In this paper, feature extraction methods like LFCC and MFCC are used to extract audio features and then these are sent as input to the Recurrent Neural Network (RNN) frame with two-layer LSTM to detect four common audio forgery operations.

These are experimented mainly on TIMIT and UME databases and various evaluations like intra-database evaluation as well as cross-database evaluation are done and the detection accuracies of each of the above are identified.

**Advantages:**

* Feature extraction techniques like MFCC and LFCC are used to extract audio features.
* In this work, RNN is used as it can capture the correlation between the frames in a speech recognition application.
* Hence it is considered better than CNN as it does not capture the sequential correlation well.

**Disadvantages:**

* The cross-database evaluation accuracy could be improved in this model.

# Fake Audio Speech Detection (2020)

# Author: Shilpa Lunagaria, Mr. Chandresh Parekh

# Description:

# In this paper, deep fake audio forgery is identified using Deep Learning algorithms. Audio files are taken as input and model is trained to uniquely identify features for voice creation and voice detection. The model could then classify between whether the audio is real or fake.

# The accuracy obtained for this model during training and validation phases are pretty high but the testing accuracy could be improved more by extracting more features and using different algorithms.

# Advantages:

# The real and fake voices can be identified.

# The accuracies obtained for training, validation are considerably high (99%, 95% respectively).

# Disadvantages:

* This work only focuses on deep fake audio forgery and it doesn’t detect or identify other audio forgery operations.
* The testing accuracy in this model could be improved with better algorithms (just 85% accuracy).

# Voice-Based Human Identification using Machine Learning(2022)

**Author:** Authors: Ivan Rakhmanenko , Alexander Shelupanov , Evgeny Kostyunchenko

**Description:**

In this study, Simple machine learning algorithms are used for voice Identification of 150 speakers. Dataset contains around 3000 samples. Mel Frequency cepstral coefficients(MFCC) is used for extracting features from audio samples. Machine learning algorithms like Support vector machine, Random forest algorithm are used for training the system.

**Advantages:**

* As simple machine learning algorithms are used, hence it is power efficient.

**Disadvantages:**

* Simple machine learning algorithms are used so accuracy is low.

# Voice spoofing countermeasure for voice replay attacks using deep learning. (2022)

# Author: Jincheng Zhou, Tao Hai1, Dayang N. A. Jawawi, Dan Wang, Ebuka Ibeke and Cresantus Biamba

# Description:

# The paper discusses about the immense usage of Automatic Speaker Verification (ASV) system which verifies users with their voices and it’s susceptibility to voice spoofing attacks - logical and physical access attacks.

# A secured voice spoofing countermeasure to detect voice replay attacks is proposed. This has enhanced the ASV system security by building a spoofing countermeasure dependent on the decomposed signals that consist of prominent information. It uses two main features— the Gammatone Cepstral Coefficients (GCC) and Mel-Frequency Cepstral Coefficients (MFCC) — for the audio representation. For the classification of the features, Bi-directional Long-Short Term Memory Network in the cloud, a deep learning classifier.

# Numerous audio features and respective feature’s capability to obtain the most vital details from the audio for it to be labelled genuine or a spoof speech is examined. Furthermore, it uses various machine learning algorithms to illustrate the superiority of the system compared to the traditional classifiers. The results of the experiments were classified according to the parameters of accuracy, precision rate, recall, F1-score, and Equal Error Rate (EER). The results were 97%, 100%, 90.19% and 94.84%, and 2.95%, respectively.

# Advantages:

# Avoids replay attacks in ASV.

# Voice biometrics.

# More accurate with the method of Speech Decomposition.

# Disadvantages:

* Does not discuss about many audio spoof attacks, focuses on Replay attacks only.

# LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems (2022)

# Author: Mohit Dua, Chhavi Jain, Sushil Kumar

# Description:

# In this paper , the system that is proposed tries to address the problem of classifying legitimate user and the malicious attacks using deep learning (DL) methods and ensemble of different neural networks. The first model that is discussed is a combination of time-distributed dense layers and long short-term memory (LSTM) layers.

# The other two deep neural networks (DNNs) are based on temporal convolution (TC) and spatial convolution (SC). Finally, an ensemble model comprising of these three DNNs has also been analysed. All these models are analysed with Mel frequency cepstral coefficients (MFCC), inverse Mel frequency cepstral coefficients (IMFCC) and constant Q cepstral coefficients (CQCC) at the frontend, where the proposed ensemble performs best with CQCC features.

# The proposed work uses ASVspoof 2015 and ASVspoof 2019 datasets for training and testing, with the evaluation set having speech synthesis (SS) and voice conversion (VC) attacked utterances. Performance of proposed system trained with ASVspoof 2015 dataset degrades with evaluation set of ASVspoof 2019 dataset, whereas performance of the same system improves when training is also done with the ASVspoof 2019 dataset.

# Advantages:

# LSTM models are used which increases performance.

# Ensemble method is used to get a consolidated decision from models.

# Disadvantages:

* Performance on combined dataset is low.
* Deep CNN can be used to improve performance.

**SYSTEM SPECIFICATION**

**CHAPTER 3**

**SYSTEM SPECIFICATION**

**3.1. SOFTWARE SPECIFICATION**

Operating system: Windows 10

IDE: Google Colaboratory / Jupyter Notebook

Coding Language: Python

**3.2. HARDWARE SPECIFICATION**

Processor: Intel core i5

Hard disk: 512GB

RAM: 8GB

**SYSTEM ANALYSIS**

**CHAPTER 4**

**SYSTEM ANALYSIS**

**4.1. INTRODUCTION**

Development of Voice Control Devices(VCDs), have boosted the realization of smart homes, voice-controlled authentication systems etc.,.These VCDs are vulnerable to different spoofing attacks. Audio Authentication is becoming very essential part of our lives. High-quality audio recorders enables bypassing this audio authentication system by just recording the human voice and reusing them for accessing the same system. Thus, there exists a need to develop a voice anti-spoofing framework capable of detecting multiple audio spoofing attacks.

**4.2. PROPOSED SYSTEM**

An Iot system which acts as an home automation system is proposed which uses a Bidirectional LSTM model to classify the audio samples and GMM model to identify the specific voice of the speaker. If the given audio input is found to be an spoofed audio or an invalid user , then the system denies the access to the system. Features fed into both the bidirectional LSTM model as well as GMM model are generated using MFCC methodologies.

The proposed system performs binary classification of audio data which are mapped to two classes

1. Authentic/Bonafide
2. Spoofed

And identifies the speaker voice using the GMM model and performs the necessary action in the Home automation system.

**4.3. DATASET USED.**

**1. Audio Spoof Detection Model**

The Dataset used in the model is Automatic Speaker Verification(ASV) 2019.

ASV Spoof dataset contains two types of Audio files

* Physical Access - Bonafide utterances are made in a real, physical space in which spoofing attacks are captured and then replayed within the same physical space using replay devices of varying quality.
* Logical Access - Bonafide and spoofed utterances generated using text-to-speech (TTS) and voice conversion (VC) algorithms are communicated across telephony and VoIP networks with various coding and transmission effects

The dataset includes genuine and spoofed speech from 20 speakers (8 male, 12 female).Each spoofed utterance is generated according to one of 2 voice conversion and 3 speech synthesis algorithms.

The voice conversion systems include those based on (i) neural-network-based and (ii)transfer-function-based methods.

The speech synthesis systems were implemented with (i) waveform concatenation, (ii) neural-network-based parametric speech synthesis using source-filter vocoders and (iii) neural-network-based parametric speech synthesis using Wavenet.

1. **Speaker Identification Model**

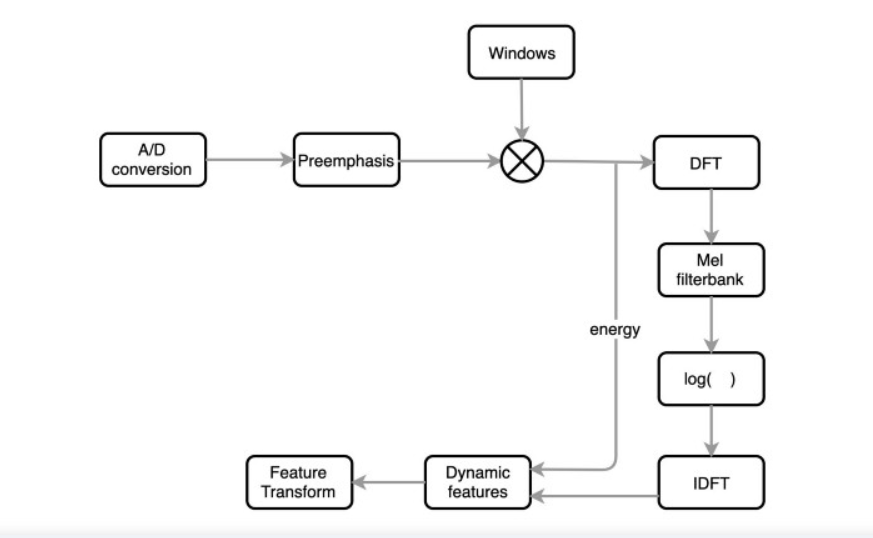
The model receives 5 audio samples from the user for training and creates an GMM model for that specific speaker. When an audio file is given as input it compares the scores of all the GMM model and finds the model that produces the highest score value. The speaker is then identified with the label for whom the it is made.

**4.4 METHODOLGIES USED.**

**Mel Frequency Cepstral Coefficients(MFCCs) :**

In general for any audio based task the raw audio signal cannot be given to the model as input because there will be a lot of noise in the audio signal. It is observed that extracting features from the audio signal and using it as input to the base model will produce much better performance than directly considering raw audio signal as input. MFCC is the widely used technique for extracting the features from the audio signal.

Mel-frequency cepstral coefficients (MFCC) which have 39 features. The feature count is small enough to force us to learn the information of the audio. 12 parameters are related to the amplitude of frequencies. It provides us enough frequency channels to analyze the audio.



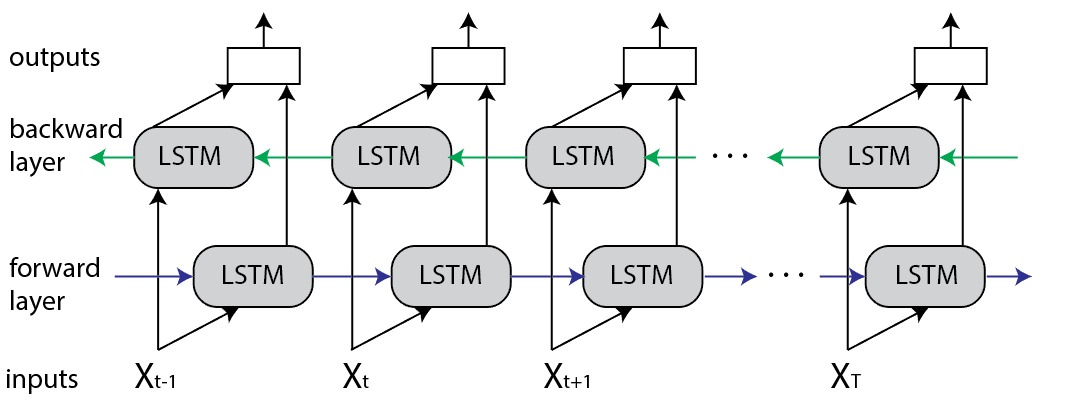
*Figure 4.1 MFCC feature extraction flow chart*

**Bi-directional Long Short Term Memory Network:**

Bidirectional LSTM (BiLSTM) is a recurrent neural network used primarily on natural language processing. Unlike standard LSTM, the input flows in both directions, and it’s capable of utilizing information from both sides. It’s also a powerful tool for modeling the sequential dependencies between words and phrases in both directions of the sequence.

BiLSTM adds one more LSTM layer, which reverses the direction of information flow. Briefly, it means that the input sequence flows backward in the additional LSTM layer. Then the outputs from both LSTM layers are combined in several ways, such as average, sum, multiplication, or concatenation.

To illustrate, the unrolled BiLSTM is presented in the figure below:



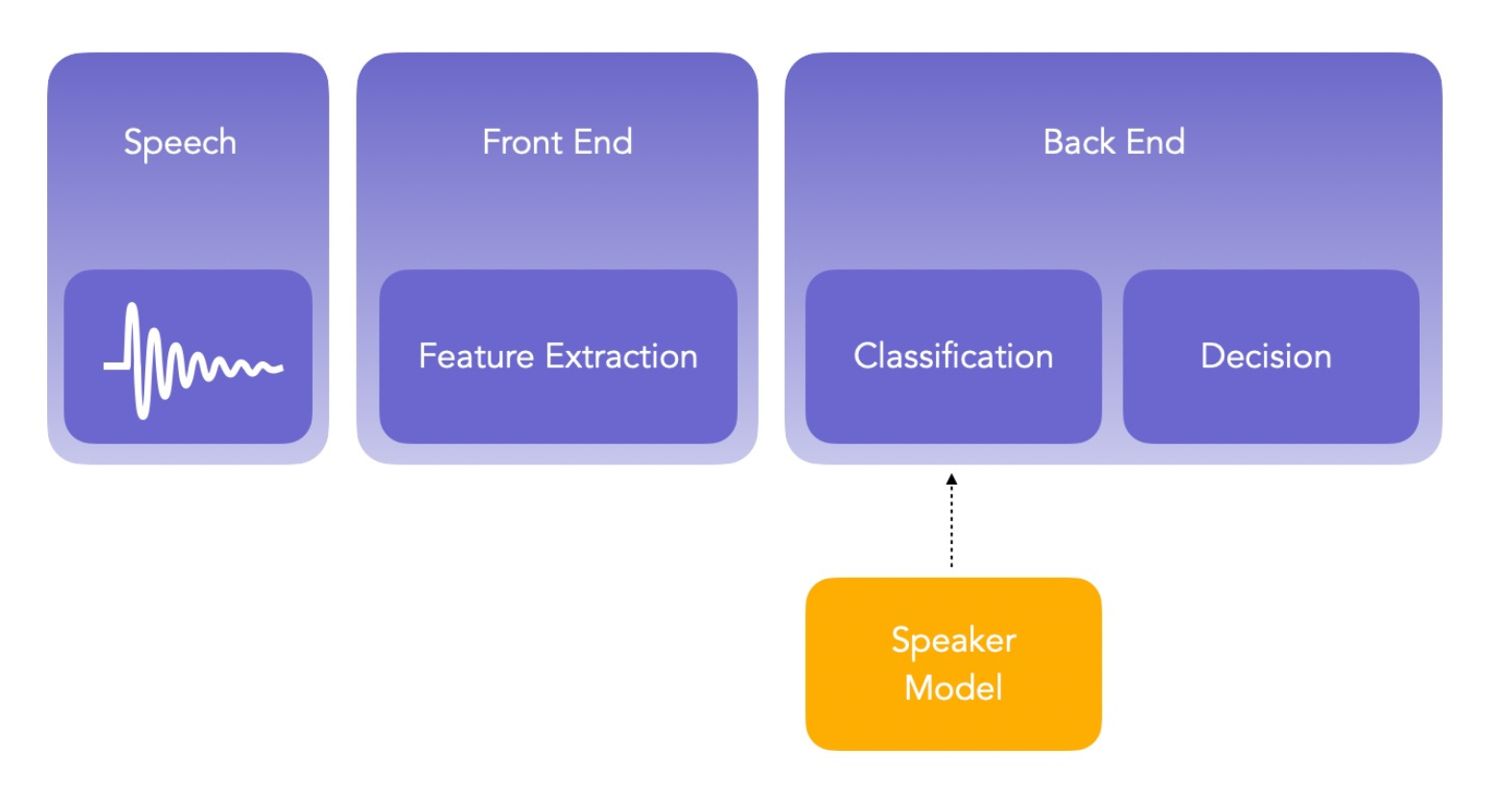
*Figure 4.2 Bi-directional LSTM architecture*

**Advantages of Bidirectional LSTM**

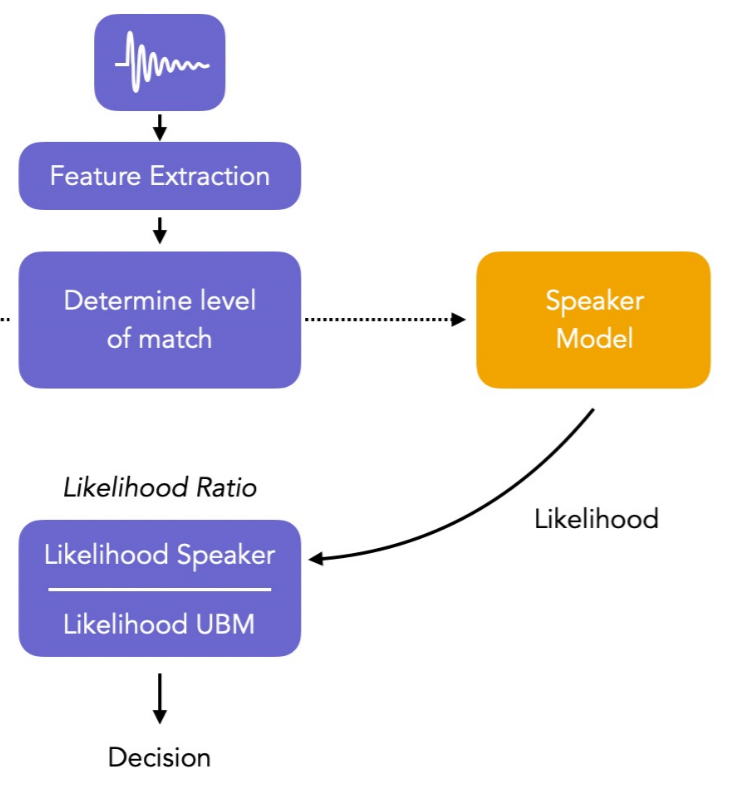
This type of architecture has many advantages in real-world problems, especially in NLP. The main reason is that every component of an input sequence has information from both the past and present. For this reason, BiLSTM can produce a more meaningful output, combining LSTM layers from both directions.

BiLSTM will have a different output for every component (word) of the sequence (sentence). As a result, the BiLSTM model is beneficial in some NLP tasks, such as sentence classification, translation, and entity recognition. In addition, it finds its applications in speech recognition, protein structure prediction, handwritten recognition, and similar fields.

**GMM:**

****

GMM is a generative speaker model because GMM model typically involves capturing data from targeted speaker. It is unique model which has de-facto reference method and generally used for robust SRS using short-speech statement. It has better advantage than other models because the training is quite fast and can be scaled and update the system to add new speakers with relative ease. A GMM model is composed of limited mixture of multivariate Gaussian components, it is a collection of several spectral features that are valid for deigning a speaker model for a targeted speaker. Suppose a speaker has 2 or 3 utterance, and from each utterance, we extract D-dimensional features .The MFCC features of each speaker are represented by Gaussian Mixture Model. MFCC coefficients are used for extracting features and minimum processing time in GMM is 10ms for speech utterance. The parameters for GMM model is mean vectors, densities (is a sum of M numbers component density), and covariance matrices.

****