

# Novel Implementation of Speech to Multi-Language Text Conversion

**Dr. Imtiyaz Khan**

*Assistant Professor CS and AI dept MJCET OU Hyderabad*

**V. Anitha**

*Assistant Professor CS and AI dept MJCET OU Hyderabad*

**Dr. M. Upendra Kumar**

*Professor CSE and Associate Head CS and AI dept MJCET OU Hyderabad*

**Mohammed Yaser**

*BE CSE MJCET OU Hyderabad 1604-17-733-043*

**Mohammed Saeed Ali Khan**

*BE CSE MJCET OU Hyderabad 1604-17-733-056*

**Furqan Imtiaz Mohammed**

*BE CSE MJCET OU Hyderabad 1604-16-733-022*

---

**Abstract:** Right from the beginning of previous century, researchers have shown interest in areas like Automatic Speech Recognition, Image Processing and Natural Language Processing. The area of Automatic Speech Recognition (ASR) has received attention over the past five decades due to its application in both commercial and military. In the recent times this can be attributed to the advancements in Artificial Intelligence and Advanced Algorithms. ASR takes speech as input and converts it in to text. ASR is employed in electronic dictionaries, Customer Call Centers, Voice Dictation and Query based Information Systems, Speech Transcription, Avionics, Smart Houses and Access Systems and many more areas. ASR can also be used to interact with handicapped people. ASR enables human beings interact with computers using speech rather than using keyboards & mouse (Vimalaand Radha V., 2012). ASR aims to provide natural machine interface where in speech acts input to the machine.

Generally, ASR is based on two tasks viz. Identification of Phoneme and Whole-Word Decoding. A relationship between speech signal and speech segment that has dissimilar physical or perceptual features usually termed as phones is established in two steps. The first step deals with dimensionality reduction and second step deals with the estimation of likelihood of each phoneme. In the dimensionality reduction phase, the volume of the speech signal is decreased by extracting the relevant information using task-specific knowledge. In the next phase, the system recognizes the word sequence using a discriminative program. Traditionally ASR systems preferred the Mel frequency Cepstral coefficients (MFCC) for the first phase and Discriminative techniques for the second phase. Over the years ASR systems have evolved from being an integration of multiple trained components to “end-to-end” Deep neural architectures that link speech to text directly. The proposed work implements an MLP with AdaBoost Classifier. The MLP will be used to extract discriminative features from the speech data. Later AdaBoost classifier will map these features to the relevant set of words.

---

## 1. Introduction

Right from the beginning of previous century, researchers have shown interest in areas like Automatic Speech Recognition, Image Processing and Natural Language Processing. The area of Automatic Speech Recognition (ASR) has received attention over the past five decades due to its application in both commercial and military. In the recent times this can be attributed to the advancements in Artificial Intelligence and Advanced Algorithms. ASR takes speech as input and converts it in to text. ASR is employed in electronic dictionaries, Customer Call Centers, Voice Dictation and Query based Information Systems, Speech Transcription, Avionics, Smart Houses and Access Systems and many more areas. ASR can also be used to interact with handicapped people. ASR enables human beings interact with computers using speech rather than using keyboards & mouse (Vimalaand Radha V., 2012). ASR aims to provide natural machine interface where

in speech acts input to the machine.

Generally, ASR is based on two tasks viz. Identification of Phoneme and Whole-Word Decoding. A relationship between speech signal and speech segment that has dissimilar physical or perceptual features usually termed as phones is established in two steps. The first step deals with dimensionality reduction and second step deals with the estimation of likelihood of each phoneme. In the dimensionality reduction phase, the volume of the speech signal is decreased by extracting the relevant information using task-specific knowledge. In the next phase, the system recognizes the word sequence using a discriminative program. Traditionally ASR systems preferred the Mel frequency Cepstral coefficients (MFCC) for the first phase and Discriminative techniques for the second phase. Over the years ASR systems have evolved from being an integration of multiple trained components to “end-to-end” Deep neural architectures that link speech to text directly.

The proposed work implements an MLP with AdaBoost Classifier. The MLP will be used to extract discriminative features from the speech data. Later AdaBoost classifier will map these features to the relevant set of words.

### **1.1 Background**

Over the years, the scale and complexity of investigations and developments in ASR for a wide range of applications has increased gradually. Around 1970s most ASR systems were developed using isolated speech recognition methodology (Baker. J 1975). They used Hidden Markov Models (HMMs) and steadily improved the recognition accuracy (Baker. J 2009).

Around late 1980s the focus shifted to continuous speech recognition methodology. Due to this the vocabulary size of recognition increased by many folds. For example, vocabulary size of in the Resource Management (RM) Task had about 900 words which increased to 20,000 words in the Wall Street Journal (WSJ) task (P. Douglas and J. M. Baker 1992). Current ASR Systems like voice-based search engines require virtually unlimited vocabulary size (Bacchiani. M et al. 2008). The difference between real time operating conditions and standard laboratory conditions also impacted the recognition accuracy of the ASR systems. Before 1995 testing and analysis of ASR systems were performed using speech recorded in noiseless environment (C.H.Lee 1996). These systems lost their accuracy as soon as they were applied to real time applications because in real time ASR Systems had to deal with noisy data. From this point on the focus of research in ASR Systems shifted to recognition of spontaneous and conversational speech in noisy environments. Currently End-to-End Systems that perform Acoustic Frame to Phone mapping in one go are being developed. In these types of systems all the modules are trained simultaneously using Advanced Machine Learning algorithms. These models take raw speech as input and generate phoneme class conditional probabilities as output. Applications like Broadcast News Transcription, Telephone Conversation Analytics gained from this new approach. Recent Commercial ASR Systems have to tackle with challenging tasks like voice search, short message dictation, YouTube transcription, Multilingual Conversion and so on. These systems work on data recorded from natural conversations which do not follow any predefined data collection procedures. Moreover, they have to operate in diverse acoustic environments catering the needs of millions of users. The ASRs have to adapt to these challenges and provide the user accurate transcriptions. This adaptation can be performed by feature-based methods or model-based methods. In feature-based methods the feature vectors of test data are normalized whereas in model-based methods acoustic model attributes are modified to better model the real conditions.

### **1.2 Rationale of the Study**

#### **What is the research issue?**

Automatic Speech Recognition Systems have to deal with challenges like acoustic diversity of speech data, Environmental noise. Voices spoken by people of different backgrounds, accents and differing styles contribute to Acoustic Diversity. Channel Characteristic differences add up to the variations in speech signals. Speech services for global users require the ASR systems to have the ability to adapt to both acoustic and channel diversity. These non-speech diversities usually confuse the ASR system and significantly degrade its recognition accuracy. To solve these issues Adaptation techniques that can separate speech diversity from non-speech diversities and then absorb these non-speech diversities have to be developed.

### **1.3 Research Gap**

Although there are many studies pertaining to Automatic speech Recognition, there is little work present in the literature that accurately recognizes the speech and converts it into text of Multiple Languages by employing hierarchical training for MLP and AdaBoost for testing.

There is still a need to develop models that are robust to reverberation and variable Acoustics. This thesis will focus on the model-based approach to address the issues of reverberation and variable acoustics.

#### 1.4 Purpose of Study

The purpose of the work presented in this thesis is to contribute to the improvement of speech recognition. It aims to deal with issues like acoustic diversity and environment diversity. The objective is to build an ASR that reads speech as input and gives text as output. The text generated can be later translated in to other languages of the user's choice. This study will concentrate on model based acoustic adaptation technique for speech recognition. The research question that this thesis answers is

1. How to optimize the training time of MLP on speech data?
2. Can AdaBoost improve the accuracy of an ASR system?

#### 1.5 Report Outline

This thesis is divided in to six chapters. First chapter introduces the topic, explains the background of the problem and outlines the research issues. It also discusses the purpose and objectives of this study. The second chapter conducts a detailed review of available literature related to Automatic Speech Recognition Systems. It also highlights the achievements and pitfalls of the existing works in this area. The third chapter will throw light on the software required to answer the research questions raised. The programming language, packages and necessary libraries that are required to conduct this research is discussed broadly. Model Design, Experimental Setup and Test strategies implemented for achieving accurate results in emotion classification is explained in the fourth chapter of this thesis. The fifth chapter gives a detailed analysis of results obtained. It performs a comparative analysis of the proposed technique with existing techniques present in the literature. The final chapter draws meaningful conclusions from the result analysis and suggests how this study can be further improved.

## 2. Literature Review

### 2.1 Investigation

Kuldip K. Paliwal and et al in the year 2004 had discussed that without being affected by their popularity for front end parameter in speech recognition, the cepstral coefficients which had been obtained from linear prediction analysis is sensitive to noise. Here, the use of spectral sub band centroids had been discussed by them for robust speech recognition. They discussed that performance of recognition can be achieved if the centroids are selected properly as in comparison with MFCC. to construct a dynamic centroid feature vector a procedure had been proposed which essentially includes the information of transitional spectral information (Jingdong Chen, 2004).

Esfandier Zavarehei and et al in the year 2005, studied that a time-frequency estimator for enhancement of noisy speech signal in DFT domain is introduced. It is based on low order auto regressive process which is used for modeling. The time-varying trajectory of DFT component in speech which has been formed in Kalman filter state equation. For restarting Kalman filter, a method has been formed to make alteration on the onsets of speech. The performance of this method was compared with parametric spectral subtraction and MMSE estimator for the increment of noisy speech. The resultant of the proposed method is that residual noise is reduced and quality of speech in improved using Kalman filters (Hakan Erdogan, 2005).

Ibrahim Patel and et al in the year 2010, had discussed that frequency spectral information with mel frequency is used to present as an approach in the recognition of speech for improvement of speech, based on recognition approach which is represented in HMM. This approach of Mel frequency utilizes the frequency observation in speech within a given resolution resulting in the overlapping of resolution feature which results in the limit of recognition. In speech recognition system which is based on HMM, resolution decomposition is used with a mapping approach in a separating frequency. The result of the study is that there is an improvement in quality metrics of speech recognition with respect to the computational time and learning accuracy in speech recognition system (Ibrahim Patel, 2010).

Kavita Sharma and Prateek Hakar in the year 2012 has represented recognition of speech in a broader solution. It refers to the technology that will recognize the speech without being targeted at single speaker. Variability in speech pattern, in speech recognition is the main problem. Speaker characteristics which include accent, noise and co-articulation are the most challenging sources in the variation of speech. In speech recognition system, the function of basilar membrane is copied in the front-end of the filter bank. To obtain better recognition results it is believed that the band subdivision is closer to the human perception. In speech recognition system the filter which is constructed for speech recognition is estimated of noise and clean speech (Kavita Sharma, 2012).

### 2.2 Proposed Idea

There is a gap is found after analyzing the previous works in speech recognition. The speech recognition can be done in two steps. In this project the same is proposed. In this project first the audio is converted in to

text then in the text the keywords are searched. The audio files can be easily converted into to text using python speech recognition library function. Once the audio file is converted in to text then it is easy to search the keywords in the files.

### 3. System Analysis & Feasibility Study

This chapter includes, existing system, proposed system, methodologies (or) algorithms, software development life cycle, feasibility study.

#### 3.1 Existing System

- Existing System did not used any Machine Learning Techniques
- Previous Systems are not efficient
- The Systems are developed for One-to-One Language Conversion
- The behavior of the System is not user friendly

#### 3.2 Disadvantages of Existing System

- Lack of Accuracy and Misinterpretation
- Time Costs and Productivity
- Accents and Speech Recognition
- Background Noise interference

#### 3.3 Proposed System

- There is a gap is found after analyzing the previous works in speech recognition. The speech recognition can be done efficiently by using machine-learning technique. The machine learning techniques increase the efficiency and accuracy of the model.
- In this project, we are proposing a novel technique of speech recognition based on Machine learning.
- In this project the Audio data set is collected from the kaggle repository and the audio file are used for training the model. Adaboost algorithm is used in this project. Then the audio file, which we want, is fed to the model then the audio file will be translated to English text. Then after using the google translator, the text will be converted in to any other recognized language.
- By removing the noise and for smoothening of the speech while converting the Speech to Text Conversion we use AdaBoost Algorithm

##### 3.3.1 Advantages of Proposed System

- Accurate Speech to Text Conversion
- User friendly Interface
- Speed in Conversion

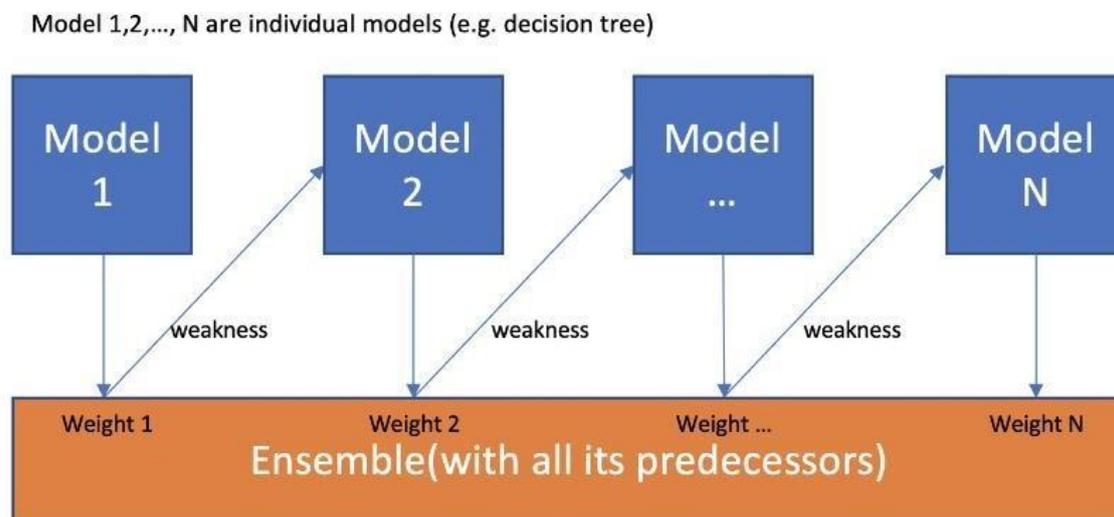
## 4. Methodologies (Or) Algorithms

### 4.1 Adaboost Algorithm

**AdaBoost algorithm**, short for Adaptive Boosting, is a Boosting technique that is used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights to incorrectly classified instances. Boosting is used to reduce bias as well as the variance for supervised learning. It works on the principle where learners are grown sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. Adaboost algorithm also works on the same principle as boosting, but there is a slight difference in working. Let's discuss the difference in detail.

## 5. Working

First, let us discuss the working of boosting. It makes  $n$  number of decision trees during the training period of data. As the first decision tree/model is made, the record which is incorrectly classified during the first model is given more priority. Only these records are sent as input for the second model. The process will go on until we specify a number of base learners we want to create. Remember, the repetition of records is allowed with all boosting techniques.



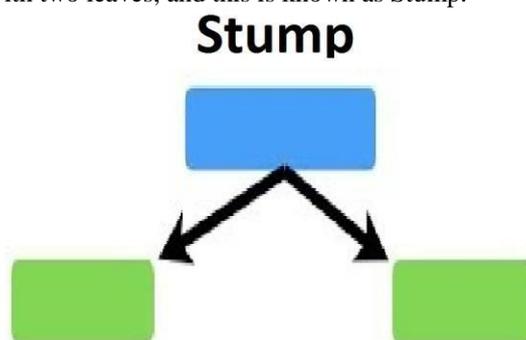
Source: Google

**Figure 4.1 Working of Adaboost**

### 6. Algorithm

This figure shows that when the first model is made and the errors from the first model are noted by the algorithm, the record which is incorrectly classified is given as the input for the next model. This process is repeated until the specified condition is met. As you can see in the figure, there are  $n$  number of models made by taking the errors from the previous model. This is how boosting works. The models 1,2, 3, N are individual models that can be known as decision trees. All types of boosting models work on the same principle.

Since we know the boosting principle, it will be easy to understand the AdaBoost algorithm. Let's deep dive into the working of Adaboost. When the random forest is used, the algorithm makes  $n$  number of trees. It makes proper trees that consist of a start node with several leaves nodes. Some trees might be bigger than others, but there is no fixed depth in a random forest. But with Adaboost, that's not the case. In AdaBoost, the algorithm only makes a node with two leaves, and this is known as Stump.



Source: Google

The figure here represents the stump. It can be seen clearly that it has only one node with only two leaves. These stumps are weak learners, and boosting techniques prefer this. The order of stumps is very important in AdaBoost. The error of the first stump influences how the other stump is made. Let's understand this with an example.

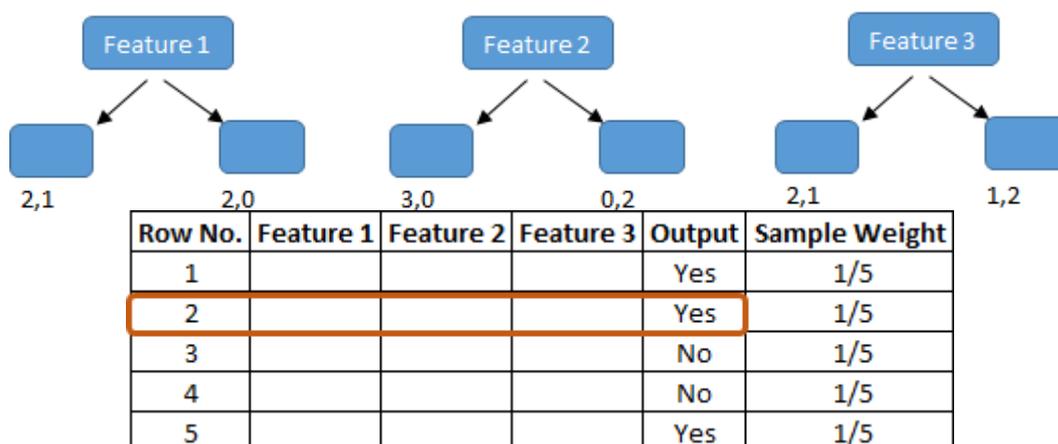
Row No.	Feature 1	Feature 2	Feature 3	Output	Sample Weight
1				Yes	1/5
2				Yes	1/5
3				No	1/5
4				No	1/5
5				Yes	1/5

Here I have created a sample dataset that consists of only three features, and the output is in categorical form. The image shows the actual representation of the dataset. As the output is in binary/categorical form, it becomes a classification problem. In real life, the dataset can have any number of records and features in it. Let us consider 5 datasets for explanation purposes. The output is in categorical form and, here it's *Yes* or *No*. All these records will get a sample weight. To assign some sample weight, the formula used is,  $W=1/N$  where N is the number of records. In this dataset, there are only 5 records, so the sample weight becomes  $1/5$  initially. Every record gets the same weight. In this case, it's  $1/5$ .

### Step 1: Creating First Base Learner

Now it's time to create the first base learner. The algorithm takes the first feature, i.e., feature 1, and creates the first stump f1. It will create the same number of stumps as the number of features. Here, it will create 3 stumps as there are only 3 features in this dataset. From all these stumps it will create three decision trees and can be called stumps base learner model. Out of these 3 models, the algorithm selects only one. For selecting a base learner, there are two properties, those are, Gini and Entropy. We must calculate Gini or Entropy the same way it is calculated for decision trees. The stump that has the least value will be the first base learner. In the below figure, all the 3 stumps can be made with 3 features. The number below the leaves represents the correctly and incorrectly classified records. By using these records, the Gini or entropy index is calculated. The stump that has the least entropy or Gini will be selected for the base learner. Let's assume that the entropy index is the least for stump 1. So, let's take stump 1, i.e., feature 1 as our first base learner.

## Base Learners/Stumps



Here, feature (f1) has classified 2 records correctly and 1 incorrectly. The row in the figure that is marked red is incorrectly classified. For this, we will be calculating the total error.

### Step 2: Calculating the Total Error (TE)

The total error is the sum of all the errors in the classified record for sample weights. In our case, there is only 1 error, so **Total Error (TE) = 1/5**.

### Step 3: Calculating Performance of Stump

Formula for calculating Performance of Stump is:

$$\text{Performance of Stump} = 1/2 \ln \left[ \frac{1-TE}{TE} \right]$$

where, **ln** is natural log and **TE** is Total Error.

In our case, TE is 1/5. By putting the value of total error in the above formula and after solving, we get the value for **the performance of Stump as 0.693**. You must be wondering why it's necessary to calculate the TE and performance of stump? The answer is, we must update the sample weight before proceeding for the next model or stage because if the same weight is applied, we receive the output from the first model. In boosting, only the wrong records/incorrectly classified records got more preference than the correctly classified records. Thus, only the wrong records from the decision tree/stump are passed on to another stump. While in AdaBoost, both records were allowed to pass, the wrong records are repeated more than the correct ones. We must increase the weight for the wrongly classified records and decrease the weight for the correctly classified records. In the next step, we will be updating the weights based on the performance of the stump.

**Step 4: Updating Weights**

For incorrectly classified records the formula is:

**New Sample Weight = Sample Weight \* e<sup>^(Performance)</sup>**

In our case Sample weight = 1/5 so, **1/5 \* e<sup>^(0.693)</sup> = 0.399**

And for correctly classified records, we use the same formula with a negative sign with performance, so that the weight for correctly classified records will reduce compared to the incorrect classified ones. The formula is:

**New Sample Weight = Sample Weight \* e<sup>- (Performance)</sup>**

Putting the values, **1/5 \* e<sup>-(0.693)</sup> = 0.100**

Row No.	Feature 1	Feature 2	Feature 3	Output	Sample Weight	Updated Weight
1				Yes	1/5 ↓	0.1
2				Yes	1/5 ↑	0.399
3				No	1/5 ↓	0.1
4				No	1/5 ↓	0.1
5				Yes	1/5 ↓	0.1
-	-	-	-	-	1	0.799

Row No.	Feature 1	Feature 2	Feature 3	Output	Sample Weight	Updated Weight	Normalized Weight
1				Yes	1/5 ↓	0.1	0.13
2				Yes	1/5 ↑	0.399	0.50
3				No	1/5 ↓	0.1	0.13
4				No	1/5 ↓	0.1	0.13
5				Yes	1/5 ↓	0.1	0.13
-	-	-	-	-	1	0.799	1

The updated weight for all the records can be seen in the figure. As is known, the total sum of all the weights should be 1. But in this case, one can see that the total updated weight of all the records is not 1, it's 0.799. To make the total sum 1, one must divide every updated weight by the total sum of updated weight. For example, if our updated weight is 0.399 and we divide this by 0.799, i.e., **0.399/0.799=0.50**.

**0.50** can be known as the normalized weight. In the below figure, we can see all the normalized weight and their sum is approximately 1.

**Step 5: Creating New Dataset**

Now, it's time to create a new dataset from our previous one. In the new dataset, the frequency of incorrectly classified records will be more than the correct ones. While considering these normalized weights, we have to create a new dataset and that dataset is based on normalized weights. It will probably select the

wrong records for training purposes. That will be the second decision tree/stump. To make a new dataset based on normalized weight, the algorithm will divide it into buckets.

Normalized Weight	Buckets
0.13	0 - 0.13
0.50	0.13 - 0.63
0.13	0.63 - 0.76
0.13	0.76 - 0.89
0.13	0.89 - 1.02

So, our first bucket is from  $0 - 0.13$ , second will be from  $0.13 - 0.63(0.13+0.50)$ , third will be from  $0.63 - 0.76(0.63+0.13)$ , and so on. After this the algorithm will run 5 iterations to select different-different records from the older dataset. Suppose, in 1st iteration, the algorithm will take a random value  $0.46$ , then it will go and see in which bucket that value falls and selects those records in the new dataset, then again it will select a random value and see in which bucket it is and select that record for the new dataset and the same process is repeated for 5 times.

There is a high probability for the wrong records to get selected several times. This will be the new dataset. It can be seen in the below image that row number 2 has been selected multiple times from the older dataset as that row is incorrectly classified in the previous dataset.

Row No.	Feature 1	Feature 2	Feature 3	Output
2				Yes
3				No
2				Yes
5				Yes
2				Yes

Based on this new dataset, the algorithm will again create a new decision tree/stump and it will repeat the same process from step 1 till it sequentially passes through all stumps and finds that there is less error when compared with normalized weight that we had in the initial stage.

## 7. Deciding Output

Suppose with the above dataset, the algorithm constructed 3 decision trees or stumps, the test dataset will pass through all the stumps which have been constructed by the algorithm. While passing through the 1st stump, it gives the output as 1, passing through 2nd stump it again gives the output as 1, and while passing through 3rd stump it gives the output as 0. So, in AdaBoost algorithm also, the majority of votes take place between the stumps, the same as in random trees. And in this case, the final output will be 1. This is how the output with test data is decided.

### 7.1 Coding AdaBoost in Python

In Python, it is easy with only 3-4 lines of code for AdaBoost algorithm. We must import the AdaBoost classifier from the sci-kit learn library. Before applying AdaBoost to any dataset, one should split the data into train and test. After splitting the data into train and test, the training data is ready to train the AdaBoost model. This data has both the input as well as output. After training the train data, our algorithm will try to predict the result on the test data. Test data only consists of the inputs. The output of test data is not known by the model. So, test data is given to the model. One can check the accuracy by comparing the actual output of the test data and the predicted output by the model. This can help us conclude how our model is performing. How much accuracy can be considered depends on the problem statement. If it's a medical problem, then accuracy should be above 90%. Usually, 70% accuracy is considered good. Accuracy also depends on more factors apart from the type of model. The below figure shows the code to implement AdaBoost.

```
[ ] from sklearn.ensemble import AdaBoostClassifier
    ad=AdaBoostClassifier()

[ ] pred=ad.fit(xtrain, ytrain).predict(xtest)

[ ] accuracy_score(ytest, pred)

0.7533333333333333
```

Figure 3.3.1.3 Code for AdaBoost in Python

At last, I would like to conclude that Adaptive Boosting is a good ensemble technique and can be used for both Classification and Regression problems. But in most cases, it is used for classification problems. It is better than any other model as it improves model accuracy, one can check this by going in sequence. First try decision trees and then go for the random forest, next apply to boost and finally go for AdaBoost. We can see that the accuracy keeps increasing as we follow the above sequence. The weight assigning technique after every iteration makes AdaBoost algorithm different from all other boosting algorithms. And that's the best thing about the AdaBoost algorithm.

## 7.2 System Requirements

### Software Requirements:

- Operating System : Windows 7+
- Server-side Script : Python 3.6+
- Platforms : Jupiter Notebook
- Libraries used : Pandas, Scikit-learn, Matplotlib, Google trans and Librosa
- Dataset : AN4 dataset from Kaggle

## 8. Design

### 8.1 System Design

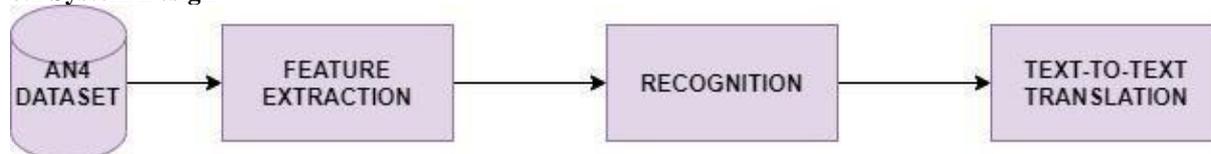


Figure 5.1 Speech to Multilingual Text System

### Architecture

The design of the project is described as, the user has the predefined data set which contains all the information related to audio files. Later, Python libraries are imported for the data set. The libraries include NumPy, Panda, Sklearn, Matplotlib.

Pandas is an open-source, Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. An introduction to Matplotlib is also provided. All this is explained with the help of examples for better understanding.

## 9. Implementation

### 9.1 Introduction

This chapter explains the implementation of the proposed system. Setting up the environment, selection of dataset, preprocessing of dataset, modeling and conclusions are included with appropriate code snippets and output screen shots. Besides this chapter focus on explanation of functions which are used in various stages of

implementation.

This chapter reviews the hierarchy components required to design an Automatic Speech to Multilingual Text Converter. The goal of this model is to convert speech into text automatically.

This project uses a workflow model to develop the proposed model. This model offers a robust solution for the problem of building software models from initiation to completion. The work flow model is a pipeline of various phases that are to be executed for developing a software project.

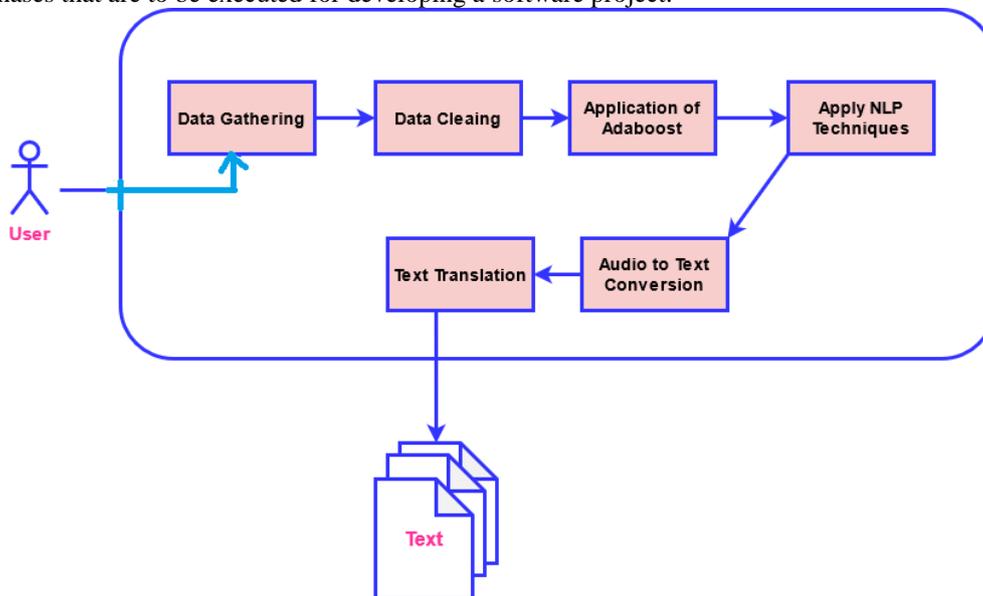


Figure 6.1 Proposed System Architecture

This model design contains five phases namely, data Collection, data preprocessing, data analysis, feature selection, speech recognition, speech to text conversion and text-to-text conversion. The workflow model is depicted in figure 6.1.

#### Installation required:

- **Python Speech Recognition module:**

pip install speechrecognition

#### 9.2 Speech Input Using a Microphone and Translation of Speech to Text

- **Allow Adjusting for Ambient Noise:** Since the surrounding noise varies, we must allow the program a second or too to adjust the energy threshold of recording so it is adjusted according to the external noise level.
- **Speech to text translation:** This is done with the help of Google Speech Recognition. This requires an active internet connection to work. However, there are certain offline Recognition systems such as PocketSphinx, but have a very rigorous installation process that requires several dependencies. Google Speech Recognition is one of the easiest to use.

#### 9.3 Data Collection and Preprocessing

Data is very important in any project. Data plays a major role in drawing the correct results. The collection and data preprocessing are discussed in the chapter 3. The audio file is read using a function rate, data = wavfile.read(audio file path).

Generally .wav format audio files are accepted in the instruction. If the audio file is not in .wav format is should be converted then its path is given as parameter for the function. The noise can be removed from the audio files by using filters. The filters can be designed to eliminate the noise. Noise is removed using function removenoise () which is supported by librosa.

#### 9.4 Data Analysis

The data analysis can be used to study the data type and patterns. After analyzing, the data if required changing of data types is applied for the parameters, which are not suitable for processing. If the data type is objects or character then the data is converted to suitable formats.

The amplitude, frequency and pitch of the audio data is analyzed using 'wave' library functions. To analyze the amplitude of the audio the function 'np.fromstring(soundInfo, np.int16)' is used.

### 9.5 Speech Recognition

Speech recognition is one more important phase in this project. This is done with the help of Speech Recognition library. The speech recognition can be done using Recognizer function.

```
import speech_recognition as sr
from guessing_game.py import recognize_speech_from_mic r = sr.Recognizer()
```

### 9.6 Audio to Text Conversion

Audio to text conversion is last and crux of the thesis. In this phase the audio file is given as input. That audio file is read using a function AudioFile() which is a predefined function present in the Speech Recognition library. After Reading the file that audio is recorded using function record(). The audio file is converted to text data using a function recognize\_google(audio\_data). This function returns the text for the given audio file. The code snippet for conversion of speech to text is given in the figure 6.6.

```
with sr.AudioFile(filename) as source:
    audio_data = r.record(source)
text = r.recognize_google(audio_data) print(text)
```

Figure 6.6 Speech to Text Conversion Code

### 9.7 Text to Text Conversion

In Text-to-text conversion phase the text translation in to other languages is done. This is language translation phase. The language translation can be done using 'googletrans' library. In googletrans library supports a function to translate the text from one language to other language. The language form which the text to convert is known as source language. The language to which that text has to be translated is known as destination language. The function translate('text', src='la1', dest='la2') is used to convert the text from one language to other language.

## 10.Results

### 10.1 Introduction

In this chapter the results projected using appropriate visualization diagrams. The results are compared with the systems which are existed.

The main aim of this thesis is to convert the speech to text and language translation. The speech to text converting is done using speech recognition supporting library functions. Text translation is done using google translator library functions.

### 10.2 Results Discussion

The Audio analysis results are given in the figure 7.2.1 and 7.2.2

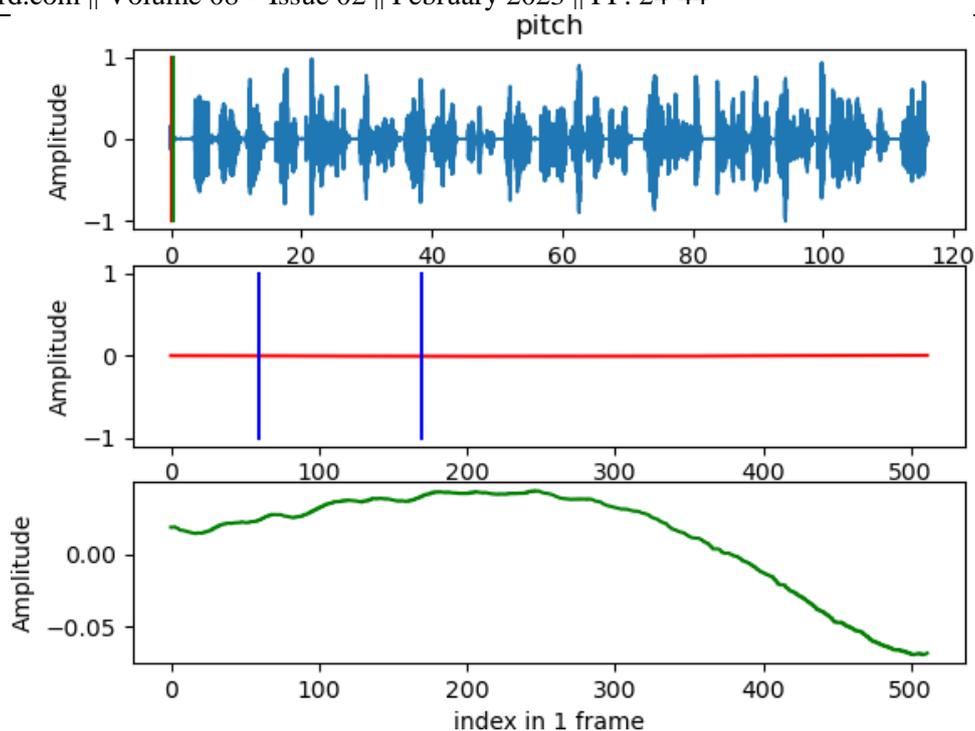


Figure 7.2.1 Pitch of the given audio

The Figure 7.2.1 shows the pitch of the audio file, which we want to convert in to text in amplitude domain. By observing the figure we can say that pitch of the audio file not constant.

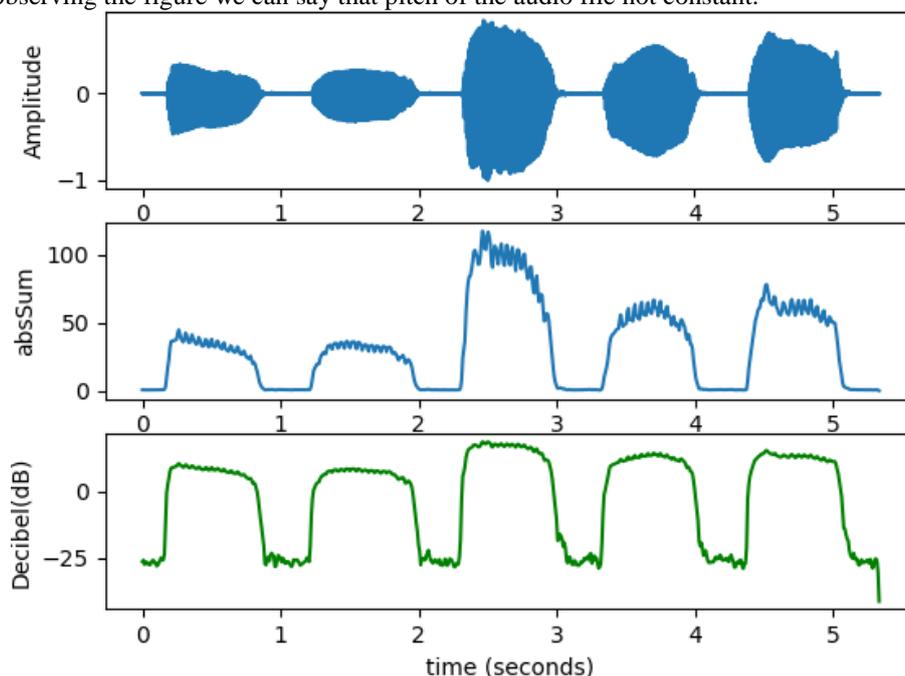


Figure 7.2.2 Amplitude of the given audio

The Figure 7.2.2 shows the volume of the audio file, which we want to convert in to text in amplitude domain and in terms of Decibels also. By observing the figure we can say that volume of the audio file not constant.

**Speech to Text Conversion Result:**

For speech conversion recognizer function is used, the function recognizer is available in speech recognition library.

The given input audio file: 4.wav The output text:



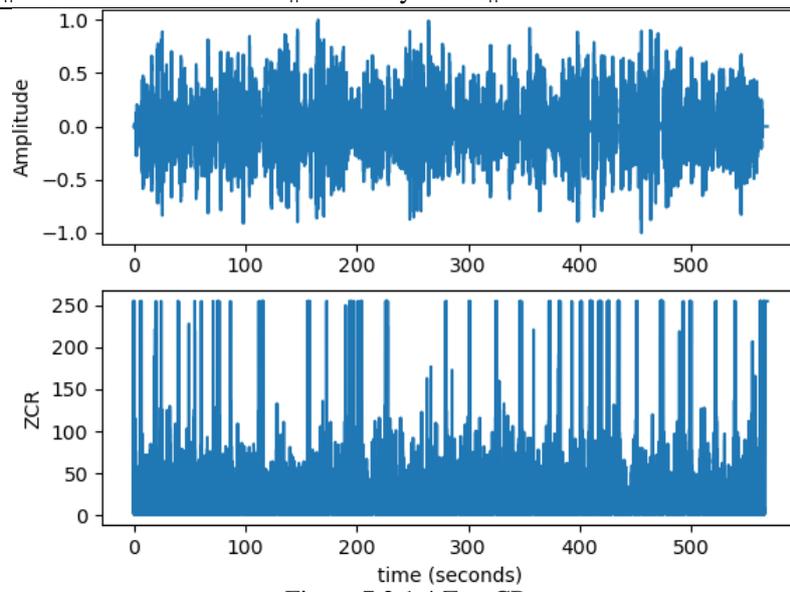


Figure 7.2.1.4 ZeroCR

x Hindi ▼

hi

new English Khwab community is important when learning anything expression Express live learning English it's important to have a community I have a community club other institution after coronavirus this is why people continue to go to school can you go to school even though they know the score now the schools message the terrible people want keep 0 want a community want to join with API join with other people they treat motivation motivation and inspiration director min direction Unity provides why our website is called the effortless English club verb lessons is also community of very positive and interviews enthusiastic learners in fact we are very we are very careful about membership in our community in our community best English honours you are very positive and enthusiastic Dastak can we have zero tolerance protein negative negative insulting or childish behavior our Puspa here are usually seen in innate immunity immunity define the massive amount of insults in and arguing with a loud that at such that such members are quickly and decisively eliminated eliminated from the club and never allowed to read how to read join this is a tough policy it is necessary International learning club online running club online and I am not interested in Excel not interested in accepting and tolerating everyone my go is to create and enter an international English learning Club of only ab only the very best learners I want the more I want the most enthusiastic the most supportive the most what is the most family the most energetic members in the members in the world and that in fact is exactly is exactly what we have the members of the effortless English club English club are absolutely amazing nasm and friendliness instrument new member new members are always very happy to discover such a fun and Pan and supportive learning club many super members who answer your questions questions give you warning advice encourage you can you feel tired when you feel tired and inspire you with their success we will be making the community community even stronger when we launch our new when will launch our new Masta member video site virya website is designed as a club for the best of the best membership site with the top 1% 18 about members we meet go get weekly new videos from me use for me I can understand everything but more importantly the videos buffer videos for kids on 4 powerful topics Jism Larsen abs exercise finance of success and daily life life in North America remember your side will focus not only on English English learning and success in general ingenero 1% and Lauren together powerful club in community of the best of the best the best close the launching the beta test version of the site this site were planning to start testing next week next week all star members 68 and view of remember Vivo offer membership to our email subscribers only sscribers only master member video site will only be available to you are email subscribers saaho in your email inbox for updates about on newest most exclusive English learning Club and Resort and as always enjoy your English learning

Translated(src=en, dest=hi, text=नया अंग्रेजी ख़ाब समुदाय कुछ भी सीखते समय महत्वपूर्ण है अभिव्यक्ति लाइव सीखना अंग्रेजी एक समुदाय होना महत्वपूर्ण है मेरे पास एक सामुदायिक क्लब है जो कोरोनावायरस के बाद अन्य संस्थान है यही कारण है कि लोग स्कूल जाना जारी रखते हैं क्या आप स्कूल जा सकते हैं, भले ही वे स्कोर जानते हों अब स्कूल संदेश भयानक लोग रखना चाहते हैं ओ चाहते हैं कि एक समुदाय एपीआई के साथ जुड़ना चाहता है अन्य लोगों के साथ जुड़ना वे प्रेरणा प्रेरणा और प्रेरणा निर्देशक न्यूनतम दिशा प्रदान करते हैं एकता प्रदान करती है कि हमारी वेबसाइट को सहज अंग्रेजी क्लब क्रिया पाठ भी बहुत सकारात्मक का समुदाय क्यों कहा जाता है और साक्षात्कार उत्साही शिक्षार्थी वास्तव में हम बहुत हैं हम अपने समुदाय में सदस्यता के बारे में बहुत सावधान हैं हमारे समुदाय में सबसे अच्छे अंग्रेजी सम्मान आप बहुत सकारात्मक और उत्साही हैं दस्तक क्या हमारे पास शून्य सहनशीलता प्रोटीन नकारात्मक नकारात्मक अपमानजनक या बचकाना व्यवहार है यहां पुस्त्या आमतौर पर जन्मजात में देखा जाता है इम्युनिटी इम्युनिटी अपमान की भारी मात्रा को परिभाषित करती है और इसके साथ बहस करती है जोर से कि ऐसे सदस्यों को क्लब से जल्दी और निर्णायक रूप से हटा दिया जाता है और कभी भी पढ़ने की अनुमति नहीं दी जाती है कि कैसे शामिल हों यह एक कठिन नीति है, यह आवश्यक है इंटरनेशनल लॉनिंग क्लब ऑनलाइन रनिंग क्लब ऑनलाइन और मुझे एक्सेल में कोई दिलचस्पी नहीं है सभी को स्वीकार करना और सहन करना मेरी कोशिश है कि मैं केवल सबसे अच्छे शिक्षार्थियों का एक अंतरराष्ट्रीय अंग्रेजी सीखने का क्लब बनाना और उसमें प्रवेश करना चाहता हूँ मैं जितना अधिक चाहता हूँ मैं सबसे उत्साही सबसे अधिक सहायक सबसे अधिक परिवार सबसे ऊर्जावान सदस्यों में से सबसे अधिक ऊर्जावान सदस्य चाहता हूँ। दुनिया में और वास्तव में ठीक वैसे ही है जैसा कि हमारे पास सहज इंग्लिश क्लब इंग्लिश क्लब के सदस्य बिल्कुल अद्भुत नस्म और मित्रता साधन हैं नए सदस्य नए सदस्य हमेशा इस तरह के एक मजेदार और पान और सहायक शिक्षण क्लब की खोज करके बहुत खुश होते हैं सुपर सदस्य जो आपके सवालों के जवाब देते हैं, आपको चेतावनी सलाह देते हैं प्रोत्साहित करते हैं कि जब आप थका हुआ महसूस करते हैं तो क्या आप थका हुआ महसूस कर सकते हैं और आपको प्रेरित कर सकते हैं? h उनकी सफलता हम समुदाय समुदाय को और भी मजबूत बना रहे होंगे जब हम अपना नया लॉन्च करेंगे जब हमारी नई मस्ता सदस्य वीडियो साइट लॉन्च होगी virya वेबसाइट को सर्वश्रेष्ठ सदस्यता साइट के लिए एक क्लब के रूप में डिज़ाइन किया गया है जिसमें शीर्ष 1% 18 सदस्यों के बारे में है। मीट गो मेरे लिए साप्ताहिक नए वीडियो प्राप्त करें मेरे लिए उपयोग करें मैं सब कुछ समझ सकता हूँ लेकिन इससे भी महत्वपूर्ण बात यह है कि 4 शक्तिशाली विषयों पर बच्चों के लिए वीडियो बफर वीडियो जिस्म लार्सन एब्स एक्सरसाइज फ़ाइनस ऑफ़ सर्वसेस एंड डेली लाइफ़ लाइफ़ इन नॉर्थ अमेरिका याद रखें कि आपका पक्ष न केवल पर ध्यान केंद्रित करेगा अंग्रेजी अंग्रेजी सीखने और सामान्य ज्ञान में सफलता 1% और लॉरेन एक साथ सबसे अच्छे के समुदाय में शक्तिशाली क्लब साइट के बीटा परीक्षण संस्करण को लॉन्च करने के करीब यह साइट अगले सप्ताह सभी स्टार सदस्यों का परीक्षण शुरू करने की योजना बना रही थी 68 और याद रखने का दृश्य वीडियो हमारे ईमेल ग्राहकों को सदस्यता प्रदान करता है केवल स्क्राइबर केवल मास्टर सदस्य वीडियो साइट आपके लिए उपलब्ध होगी ईमेल ग्राहक साहो आपके ईमेल इनबॉक्स में नवीनतम सबसे विशिष्ट अंग्रेजी शिक्षण क्लब और रिज़ॉर्ट के बारे में अपडेट करें और हमेशा की तरह अपने अंग्रेजी सीखने का आनंद लें, pronunciation=[[]], extra\_data="{ 'translat..."

Figure 7.2.1.5 English to Hindi

x Chinese(Simplified) ▼

zh-cn

zh-cn

new English Khwab community is important when learning anything expression Express live learning English it's important to have a community I have a community club other institution after coronavirus this is why people continue to go to school can you go to school even though they know the score now the schools message the terrible people want keep 0 want a community want to join with API join with other people they treat motivation motivation and inspiration director min direction Unity provides why our website is called the effortless English club verb lessons is also community of very positive and interviews enthusiastic learners in fact we are very we are very careful about membership in our community in our community best English honours you are very positive and enthusiastic Dastak can we have zero tolerance protein negative negative insulting or childish behavior our Puspa here are usually seen in innate immunity immunity define the massive amount of insults in and arguing with a loud that at such that such members are quickly and decisively eliminated eliminated from the club and never allowed to read how to read join this is a tough policy it is necessary International learning club online running club online and I am not interested in Excel not interested in accepting and tolerating everyone my go is to create and enter an international English learning Club of only ab only the very best learners I want the more I want the most enthusiastic the most supportive the most what is the most family the most energetic members in the members in the world and that in fact is exactly is exactly what we have the members of the effortless English club English club are absolutely amazing nasm and friendliness instrument new member new members are always very happy to discover such a fun and Pan and supportive learning club many super members who answer your questions questions give you warning advice encourage you can you feel tired when you feel tired and inspire you with their success we will be making the community community even stronger when we launch our new when will launch our new Masta member video site virya website is designed as a club for the best of the best membership site with the top 1% 18 about members we meet go get weekly new videos from me use for me I can understand everything but more importantly the videos buffer videos for kids on 4 powerful topics Jism Larsen abs exercise finance of success and daily life life in North America remember your side will focus not only on English English learning and success in general ingenero 1% and Lauren together powerful club in community of the best of the best the best close the launching the beta test version of the site this site were planning to start testing next week next week all star members 68 and view of remember Vivo offer membership to our email subscribers only sscribers only member video site will only be available to you are email subscribers saaho in your email inbox for updates about on newest most exclusive English learning Club and Resort and as always enjoy your English learning

Translated(src=en, dest=zh-cn, text=新的英语 Khwab 社区在学习任何表达时都很重要 快速现场学习英语 有一个社区很重要 我有一个社区俱乐部 冠状病毒后其他机构 这就是为什么人们继续上学的原因 即使他们知道分数, 你也能上学现在学校消息可怕的人想要保持 0 想要一个社区想要加入 API 加入他们对待其他人的动机和灵感总最小方向 Unity 提供了为什么我们的网站被称为毫不费力的英语俱乐部动词课程也是非常积极的社区和采访热情的学习事实上我们非常我们非常谨慎地加入我们的社区在我们的社区中最好的英语荣誉你非常积极和热情Dastak我们可以零容忍蛋白质消极的负面侮辱或幼稚的行为这里的Puspa通常是与生俱来的免疫免疫定义了大量的侮辱和争论大声说这样的成员被迅速果断地从俱乐部淘汰出局, 永远不允许阅读如何阅读加入这是一项严厉的政策是必要的国际学习俱乐部在线跑步俱乐部在线我对Excel不感兴趣接受和包容每个人我的目标是创建和进入一个只有最好的学习者的国际英语学习俱乐部我想要的越多我想要的越热情最支持什么是最家庭成员中最有活力的成员在世界上, 事实上正是我们所拥有的 轻松英语俱乐部 英语俱乐部的成员绝对令人惊叹 nasm 和友好的乐器 新成员 新成员总是很高兴发现这样一个有趣、泛和支持的学习俱乐部 很多回答你问题的超级会员给你警告建议鼓励你在你感到疲倦时会感到疲倦并激发你的智慧他们的成功我们将在我们推出新会员视频网站时使社区社区变得更加强大 virya 网站旨在成为最好的会员网站中最好的俱乐部, 其中前 1% 的会员是我们的会员见面去每周从我那里获取新视频 为我使用我可以理解一切, 但更重要的是视频缓冲视频为孩子们提供了 4 个强有力的主题 Jism Larsen abs 锻炼成功和北美日常生活的财务 记住你的身边将不仅英语 英语学习和一般成功 ingenero 1% 和 Lauren 一起强大的俱乐部在最好的最好的最好的社区 关闭 推出该网站的 beta 测试版 本网站计划下周开始测试 下周所有明星会员 68和视图记住 Vivo 为我们的电子邮件订阅者提供会员资格 只有抄写员只有主会员视频网站将只对您是电子邮件订阅者的电子邮件收件箱中的 saaho 可用r 最新最独特的英语学习俱乐部和度假村的更新, 一如既往地享受您的英语学习, pronunciation=[[[]], extra\_data="{ 'translat...")

Figure 7.2.1.6 English to Chinese

The given input audio file: 5.wav The output text:

```
[ [10.94713983 9.99398292 12.18790791 12.8504111 13.62553786 12.29264385
 10.92248824 11.55679057 12.25664498 11.51486177 8.38093661 8.81039205
 8.46320476 9.27358183 8.84892385 8.84642652 9.43254083 8.7842305
 8.91336212 6.71144106 6.21862968 6.32807392 6.42951027 8.40391954
 8.34886804 8.45447038]
 [11.12858872 11.2736033 13.86072626 14.51233543 15.49314012 12.93437362
 10.75359312 13.08680835 14.08713444 13.01643775 9.467426 8.9429215
 9.78233427 10.60088878 11.54396398 9.99562533 10.919647 11.20341505
 11.79934337 7.21605933 4.05816941 4.21901235 6.81349061 10.92982219
 10.12483045 10.10499301]]
```

Figure 7.2.1.7 MFCC

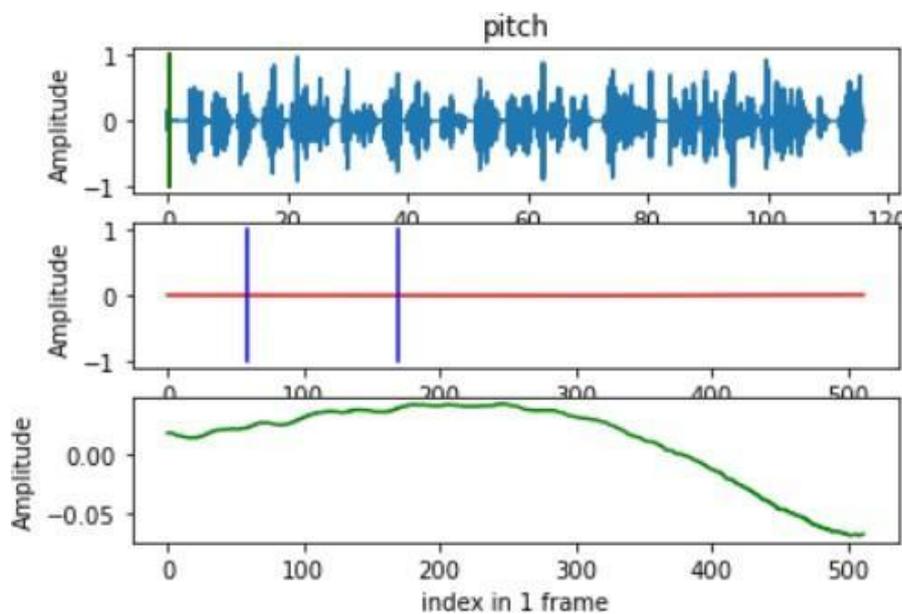


Figure 7.2.1.8 Pitch

Wave and spectrogram of 5.wav

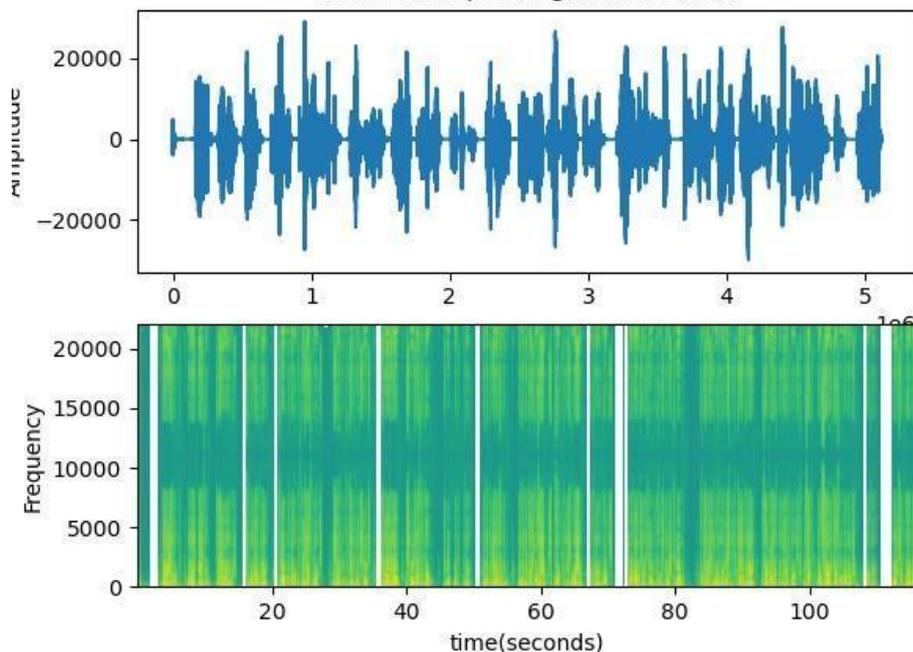


Figure 7.2.1.9 Spectrogram

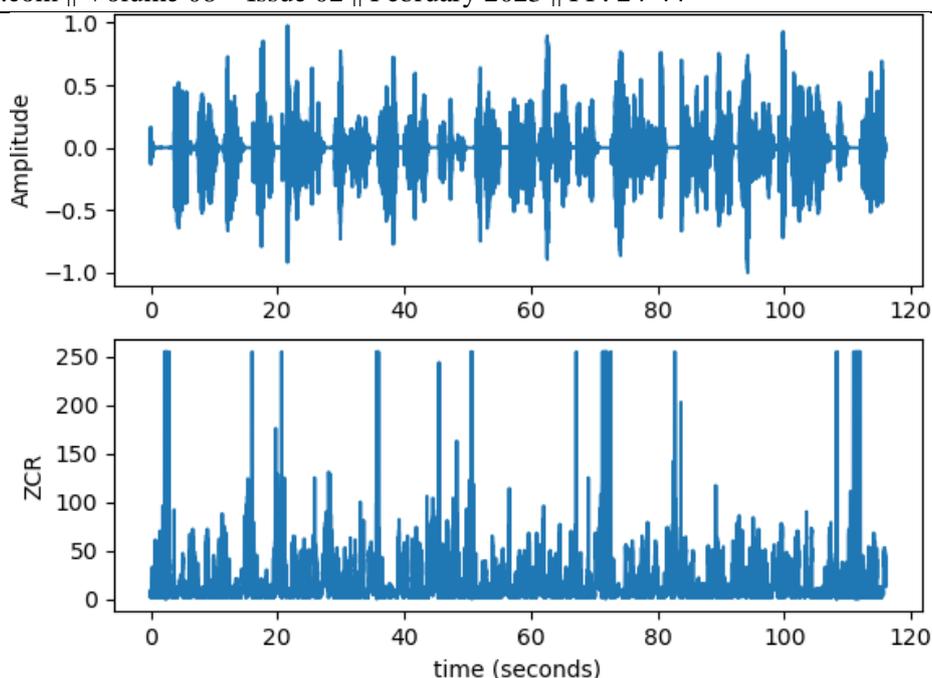


Figure 7.2.1.10 ZeroCR

x Chinese(Simplified) ▼

zh-cn

zh-cn

community is important when learning any when running anything especially when learning English it's important to have a community Club of other ab other into the after coronavirus this is why people this is why people continue to go to school even though even though they know the schools message the terrible terrible people want a community I want to join with other people I ndian trees motivation support in inspiration an inspiration that our community provides which is why our website is called the effortless English effortless English club percentage is more than GRE lessons is also a community of a very positive and enthusiastic learners in fact in fact we are very careful about membership about membership in our community only accept the very best

Translated(src=en, dest=zh-cn, text=社区在学习任何东西时都很重要, 尤其是在学习英语时, 重要的是让其他 ab 的社区俱乐部进入冠状病毒之后这就是人们这就是为什么人们即使知道学校信息也继续上学的原因可怕的人想要一个社区 我想和其他人一起加入 印度树木 激励支持 我们社区提供的灵感 这就是为什么我们的网站被称为轻松英语 轻松英语俱乐部的百分比超过 GRE 课程也是一个社区一个非常积极和热情的学习者 事实上, 我们对会员资格非常谨慎 关于我们社区的会员资格只接受最好的, pronunciation=[[ ]], extra\_data={'translat...'})

Figure 7.2.1.11 English to Chinese



hi  
community is important when learning any when running anything especially when learning English it's important to have a community Club of other ab other into the after coronavirus this is why people this is why people continue to go to school even though even though they know the schools message the terrible terrible people want a community I want to join with other people I ndian trees motivation support in inspiration an inspiration that our community provides which is why our website is called the effortless English effortless English club percentage is more than GRE lessons is also a community of a very positive and enthusiastic learners in fact in fact we are very careful about membership about membership in our community only accept the very best

Translated(src=en, dest=hi, text=समुदाय महत्वपूर्ण है जब कुछ भी सीखते समय विशेष रूप से अंग्रेजी सीखते समय अन्य लोगों का एक समुदाय क्लब होना महत्वपूर्ण है, यही कारण है कि लोग स्कूल जाना जारी रखते हैं, भले ही वे स्कूल संदेश जानते हों भयानक भयानक लोग एक समुदाय चाहते हैं मैं अन्य लोगों के साथ जुड़ना चाहता हूं भारतीय पेट प्रेरणा प्रेरणा में समर्थन एक प्रेरणा जो हमारा समुदाय प्रदान करता है, यही कारण है कि हमारी वेबसाइट को सहज अंग्रेजी कहा जाता है, अंग्रेजी क्लब प्रतिशत जीआरई पाठों से अधिक है। एक बहुत ही सकारात्मक और उत्साही शिक्षार्थी वास्तव में हम सदस्यता के बारे में बहुत सावधान हैं हमारे समुदाय में सदस्यता के बारे में केवल सबसे अच्छा स्वीकार करते हैं, pronunciation=[[]], extra\_data="{ 'translat...")

Figure 7.2.1.12 English to Hindi

**Text to Text Conversion:**

In text-to-text conversion google translator is used. The function translator consists of three parameters one those are

- 1) text: the text which is to be translated is specified
- 2) src: this is the source language
- 3) dst: this is destination language The text is converted English to Telugu:

```
tr=translator.translate(st, src='en', dest='te')
print(tr)
```

community is important when learning any when running anything especially when learning English it's important to have a community Club of other ab weather in today after coronavirus this is why it this is why people continue to go to school even though even though they know the schools message the terrible terrible people want a community I want to join with other people Indian trees motivation support in inspiration an inspiration that our community provides which is why our website is called the effortless English effortless English club percentage is more than grey lessons is also a community of a very positive and enthusiastic learners in fact in fact we are very careful about membership about membership in our community only accept the very best

Translated(src=en, dest=te, text=వీధి నా నేర్చుకునేటప్పుడు వీధి నా నేర్చుకునేటప్పుడు కమ్మనిటి ముఖం, ముఖం ఇంద్రియ నేర్చుకునేటప్పుడు కరోనావైరస్ వ్యాధి ఈ రోజు ఇతర కమ్మనిటి వాతావరణం కఠిన ఉండటం చాలా ముఖం, అందువల్ల వాతావరణం సందేశం తెలిసినప్పటికీ ప్రజలు వాతావరణం వ్యాధి కొనసాగుతుంది. భయంకరమైన భయంకరమైన వ్యక్తులు నేను ఇతర వ్యక్తులతో చేరాలని కోరుకునే సమాజం కోరుకుంటున్నాను భారతీయ చెట్ల ప్రి రణ మధ్య న్యూరీతో మా సభం అందించే ప్రి రణ, అందుకే మా వాచ్ సైట్ ను అప్రయత్నంగా ఇంద్రియ అప్రయత్నంగా ఇంద్రియ క్లబ్ శాశం అని పిలుస్తారు బుడిద వాలా కంటే కూడా ఒక సభం చాలా సానుకూల మరియు ఉత్సాహభరితమైన అభ్యాసకులు వాస్తవం మా సమాజంలో సభ్యత్వం గురించి సభ్యత్వం గురించి చాలా జాగ్రత్తగా ఉన్నాము, pronunciation=[[]], extra\_data="{ 'translat...")

Figure 7.2.2.1 English to Telugu

The text is converted English to Irish:

community is important when learning any when running anything especially when learning English it's important to have a community Club of other ab weather in today after coronavirus this is why it this is why people continue to go to school even though even though they know the schools message the terrible terrible people want a community I want to join with other people Indian trees motivation support in inspiration an inspiration that our community provides which is why our website is called the effortless English effortless English club percentage is more than grey lessons is also a community of a very positive and enthusiastic learners in fact in fact we are very careful about membership about membership in our community only accept the very best

Translated(src=en, dest=ga, text=tá an pobal tábhachtach nuair atá siad ag foghlaim rud ar bith agus iad ag rith aon rud go háirithe agus iad ag foghlaim Béarla tá sé tábhachtach go mbeadh Club pobail ab aimsire eile ann inniu tar éis coronavirus agus sin an fáth go leanann daoine orthu ag dul ar scoil cé go bhfuil teachtaireacht na scoile ar eolas acu tá na daoine uafásacha uafásacha ag iarraidh pobail ba mhaith liom a bheith páirteach le daoine eile Tacaíocht spreagtha crainn Indiach mar inspioráid inspioráid a sholáthraíonn ár bpobal agus sin an fáth ar a dtugtar ár suíomh Gréasáin mar chéatadán an chlub gan iarracht Béarla Béarla gan iarracht ná gur ceachtanna liath é pobal de foghlaimeoirí an-dearfacha agus díograiseacha i ndáiríre i ndáiríre táimid an-chúramach faoi bhallraíocht faoi bhallraíocht inár bpobal ach glacadh le is an gcuid is fearr, pronunciation=[[ ]], extra data="{ 'translat..."}")

Figure 7.2.2.2 English to Irish

## 11. Conclusion and Future Work

### 11.1 Conclusion

In this project, we have implemented Speech to text conversion. In this project the Google API used to convert speech to text using the microphone in English audio file. This can be useful in natural language processing projects for handling audio files and transcripts as well.

The purpose of the project is achieved with expected results and the speech to text conversion is also done in efficient manner.

### 11.2 Future Work

This project is mainly concentrated in converting the English audio files to other recognized languages. This project can be extended to convert any audio file to any text. This project can be implemented using other techniques like deep learning and Multilayer perceptron's etc.

One more important extension for this project is instead of audio to text conversion, this can be extended for converting video file to text file.

## 12. References

- [1]. Vimala, C. and Radha, V. (2012). A Review on Speech Recognition Challenges and Approaches. World of Computer Science and Information Technology Journal (WCSIT), 2 (1), PP. (1-7). 2221-0741.
- [2]. J. Baker. The DRAGON system—An overview. IEEE Transactions on Acoustics, Speech, and Signal Processing, 23(1):24–29, 1975.
- [3]. J. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy. Developments and directions in speech recognition and understanding. IEEE Signal Processing Magazine, 26(3):75–80, 2009
- [4]. P. Douglas and J. M. Baker. The design for the Wall Street Journal-based CSR corpus. In Proceedings of the Workshop on Speech and Natural Language, pages 357–362, 1992.
- [5]. M. Bacchiani, F. Beaufays, J. Schalkwyk, M. Schuster, and B. Strope. Deploying goog411: early lessons in data, measurement, and testing. In Proceedings of ICASSP, pages 5260–5263, 2008
- [6]. C.-H. Lee, F. K. Soong, and K. K. Paliwal. Automatic speech and speaker recognition: advanced topics. Springer, 1996
- [7]. N. Sharma and S. Sardana, "A real time speech to text conversion system using bidirectional Kalman

- filter in Matlab," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 2353-2357, doi: 10.1109/ICACCI.2016.7732406.
- [8]. <https://www.analyticsvidhya.com/blog/2019/07/learn-build-first-speech-to-text-model-python/>
- [9]. Vlado Delić et al. "Speech Technology Progress Based on New Machine Learning Paradigm" Volume 2019 Computational Intelligence and Neuroscience
- [10]. Ahad, A. & Fayyaz, A. & Mehmood, T.. (2002). Speech recognition using multilayer perceptron. 103 - 109 vol.1. 10.1109/ISCON.2002.1215948.
- [11]. <https://heartbeat.fritz.ai/a-2019-guide-for-automatic-speech-recognition-f1e1129a141c>
- [12]. M.A.Anusuya and S.K.Katti "Speech Recognition by Machine: A Review" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009.
- [13]. R. Sultana and R. Palit, "A survey on Bengali speech-to-text recognition techniques," 2014 9th International Forum on Strategic Technology (IFOST), Cox's Bazar, 2014, pp. 26-29, doi: 10.1109/IFOST.2014.6991064.
- [14]. K. Nguyen, T. Ng and L. Nguyen, "Adaptive boosting features for automatic speech recognition," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 4733-4736, doi: 10.1109/ICASSP.2012.6288976.