

VOICE LANGUAGE PREDICTION USING MACHINE LEARNING

BY

Eteka Sultana Tumpa

ID: 191-15-12121

AND

Shabrina Sharmin

ID: 191-15-12855

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Sheak Rashed Haider Noori

Professor & Associate Head

Department of CSE

Daffodil International University

Co-Supervised By

Mr. Saiful Islam

Assistant Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

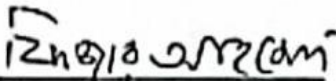
This Project/internship titled "Voice language prediction using machine learning", submitted by Eteka Sultana Tumpa, ID No: 191-15-12121, and Shabrina Sharmin, ID No: 191-15-12855 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 28 January 2023.

BOARD OF EXAMINERS



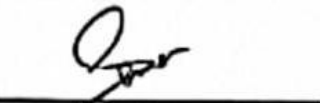
Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Taslima Ferdous Shuva
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Md Sazzadur Rahman
Associate Professor
Institute of Information Technology
Jahangirnagar University

External Examiner

DECLARATION

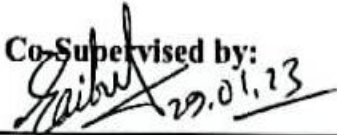
We hereby declare that this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, professor & Associate Head, the Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:


29.01.2023

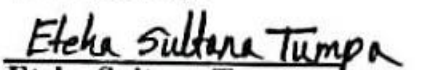
Dr. Sheak Rashed Haider Noori
Professor & Associate Head
Department of CSE
Daffodil International University

Co-Supervised by:


29.01.23

Mr. Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:


Eteka Sultana Tumpa
ID: 191-15-12121
Department of CSE
Daffodil International University


Shabrina Sharmin
ID: 191-15-12855
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing making us possible to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to **Dr. Sheak Rashed Haider Noori, professor & Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Field name*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan, Professor and Head**, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire course mate Daffodil International University, who participated in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

When a speaker cannot be positively recognized, speech-language identification may be used to determine the language they are speaking. Some basic machine-learning methods will also be covered. The ability to recognize human speech may be learned. First and foremost, we must establish search criteria. Differentiating languages by their spoken characteristics may help to derive a feature. After collecting all the necessary audio data from a variety of sources, we compress the resulting audio file. Methods of identifying languages have been used (LID). With the highest possible F1 score, Machine Learning is the most successful tactic. Therefore, we used machine learning classifiers in our studies. Our work focuses on predicting spoken language. To get the highest possible F1 score, we will employ a Machine Learning classifier in situations when identifying the other person's language is straightforward. Five algorithms were employed to get to this point. This work's Decision tree method has the highest F1 score and highest accuracy. Others have been more accurate, although they have been overfitting. Therefore, the decision tree approach proved effective for our project.

TABLE OF CONTENTS

CONTENTS	PAGE
Broad of examiners	i
Declaration	ii
Acknowledgment	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	2
1.3 The study's relationship	2
1.4 Asked in Questionnaires	3
1.5 Expected Outcomes	3
CHAPTER 2: BACKGROUND	4-10
2.1 Connected Works	4
2.2 Terminologies	7
2.3 Contrast this with a summary and compare	10
CHAPTER 3: RESEARCH METHODOLOGY	11-19
3.1 Methodology	11
3.2 Method for data collection	12
3.3 Statistics for Analysis	12
3.4 Proposed Approach	13
3.4.1 Pre-Processing	13
3.4.2 Feature extraction	14

3.4.2.1 (MFCC) Mel-frequency cepstral coefficients	14
3.4.2.2 The transformation of analog to digital	15
3.4.3 Machine Learning	15
3.4.3.1 Decision tree	15
3.4.3.2 Random Forest	16
3.4.3.3 KNN	16
3.4.3.4 Gaussian Naïve Bayes	17
3.4.3.5 SVM	17
3.5 Evaluation	18
3.5.1 Precision	18
3.5.2 Recall	19
3.5.3 F1 Score	19
3.6 Implementation prerequisites	19
CHAPTER 4: RESULT ANALYSIS	20-22
4.1 Discussion about the result	20
4.2 Discussion	22
CHAPTER 5: ASSESSMENT OF THE RESULTS AND IMPLICATIONS FOR FUTURE RESEARCH	23
5.1 Conclusion	23
5.2 Future study	23
APPURTENANCE	24
REFERENCES	25-26
PLAGIARISM REPORT	27

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1.1: Working method	11
Figure 3.4.1: Audio data plotting using librosa	13
Figure 3.4.2: Spectrogram plot	14
Figure:3.4.3.1.1: Diagram for Decision tree	16
Figure:3.4.3.5.1: Diagram for SVM	18

LIST OF TABLES

TABLE NAME	PAGE NO
Table 3.3.1 A chart for our data	12
Table 4.1.1 The result of the classification	20-21

CHAPTER 1

INTRODUCTION

1.1 Introduction

Due to recent technical advancements, the distances between people and countries have lessened, which presents a formidable obstacle. So, the globe is becoming more interconnected and hence smaller. Therefore, the most important thing right now is communication. However, we know that this is impossible without a shared vocabulary and that talking to one another is the first step toward building that bridge. Just about 6,500 spoken tongues may be found today. Given these constraints, it would be difficult to memorize all of the languages. There may be aspects of a language that we may pick up on even if we don't speak it. To organize this data, we need some form of the repository. The phrase "spoken language identification" describes a group of methods for automatically identifying the language of a spoken text. Multilingual speech recognition, voice transmission, and Speaker dualization systems all share the ability to identify spoken languages. By listening to and examining a recording of someone's speech, language identification software can determine their mother tongue. When it comes to identifying and classifying linguistic varieties, humans have performed well up to this point. When someone hears a language, they are fluent in, they may be able to recognize it immediately. However, difficulties emerge when they are unable to converse with one another in their native tongue. They won't be able to identify the spoken language. Machine learning is the best approach to use when this kind of problem emerges. Since it's impossible for one individual to be proficient in all of the world's languages. We will use AI to help us figure out a solution to this issue. who we were only able to educate in a rough approximation of all the languages in use today. This method is known as automatic language recognition. This method is an impressive technical achievement that has been put to use in applications such as voice-to-voice translation and the identification of audio recordings. Not only did some people try to tackle the problem in the same way, but others did, too. To be sure, the machine-learning process is unique. In the last several years, machine learning has made great strides. Nowadays, words alone are sufficient to command technological systems. However, these traditional methods have language constraints. For us to go forward, Daffodil International University, Level 1 tells our gadgets how to interpret our inputs. go to the next stage. Instead, we will reach previously unimaginable heights if we can

teach robots to detect several languages and advise individuals in their native tongue about the advantages of each. This success will be felt not just in this instance, but also in the security and intelligence sectors, where the ability to accurately identify speakers of different languages in the context of recorded communications and databases is crucial. For our first method of language recognition, we shall focus on spatial processing.

1.2 Motivation

Even while Google Translate performs a fantastic job at translating spoken languages, the difficulty of this task becomes clear when we consider that we must first decide which languages to translate. Even if one has Google Translator on their phone, it won't help ©Daffodil International University 2 them if they're in a position where they need to communicate with someone but they don't speak or understand the language they're speaking. Travelers spending a few months in a country where IELTS is not needed may struggle due to language barriers. One's ability to converse fluently in many languages would be severely tested under these conditions, making services like Google Translate indispensable. Google Translate, however, has not yet reached this functionality in cases when language is not an issue. It's possible that speech may be used in an automated selection process. So far, Google Translate hasn't continued the discussion in automatic mode after being given the order to do so.

1.3 The study's relationship

By 2022, Google Translate will provide limited support for 109 languages. As of April 2016, it had more than 500 million users worldwide and was processing 100 billion words per day in translation. By ensuring that the language is utilized correctly in a number of apps like Google Translate, we can considerably simplify the lives of many individuals and international students if we can reliably identify the language using voice recognition. In this project, we want to provide the best effective method for detecting languages.

1.4 Asked in Questionnaires

- Do you think this algorithm could identify the speaker's language?
- Methods for information retrieval are needed.
- In what ways will it be useful in practice?
- Where did we develop such a notion?
- Justification for using ML technology.

1.5 Expected Outcomes:

- On the aircraft, you won't have to worry about a thing when it comes to communicating.
- Our machine learning method will provide an F1 score indicating its precision.
- Any spoken language will be easily identifiable.
- Only those who rely heavily on their original language will benefit the most.

CHAPTER 2

BACKGROUND

2.1 Connected Works:

After years of study, we can confidently claim that several methods exist for honing language recognition skills. Improving output has always been a top priority. Differentiating across languages in conversation requires keeping various features and pieces of information from the original speech signal intact. Data that has previously been recorded is also used to establish the language.

Using language and gender as the two primary identifiers of the speech signal, a practical naming system was developed. The software was evaluated across 10 different languages. Using functional acoustic features and the best model training approach, Li, W., Kim, D.J., Kim, C.H., and Hong, K.S. [1] achieved a good accuracy identification rate. SLR back prediction using one.

Researchers Ferrer, Lei, McLaren, and Scheffer [2] examined and compared several deep neural networks (DNN) training methods and sensor-based deep neural network methodologies for interpreting spoken speech. We recommend using SLR, NIST, LRE, and RATS. Both jobs were performed by GMM/IV systems during the testing.

Language-independent features to be used in automated voice recognition To better characterize the acoustic properties of human speech, Siniscalchi SM, Reed J, Svendsen T, and Lee CH. [3] suggested a revolutionary universal technique (LRE).

H. Li, B. Ma and K. A. Lee [4], in their book "Spoken Language Recognition: From Concepts to Practice" (Spoken Language Recognition: From Concepts to Practice), set out to provide a foundation of theoretical principles and state-of-the-art approaches for spoken language recognition from phonological and computational aspects. According to Furui S and Juang BH[5], the first step towards true human-machine communication is automated speech interpretation and recognition. They made this claim using cepstral distance, forward-backward algorithm, Bayesian risk and linear prediction. An analysis of example-based language recognition systems by Wang MG, Song Y, Jiang B, Dai LR, and McLoughlin [6] focused on short speech segments. The cosine encoder is

shown to outperform a number of other, more complex encoding methods. SVM was used to improve already existing processes and regulate the launch of new functions.

They draw a conclusion based on the background data offered by Drugman T, Kane J, Raitio T, and Gobl C [7]. The voice was becoming squeaky. A more effective method of prediction has been shown to surpass the clumsy detection approach on which HMMs have been trained. Sometimes the quality of our conversations suffers. Deep machine listening (DESQL) was investigated by Ooster J. and Meyer B.T. [8] to quantify voice quality. Based on deep neural network-calculated phoneme posterior probabilities, DESQL generates predictions about how speech might be understood.

The authors of the paper on language recognition are Justin Pyron, Andrew Deveau, and Julien Boussard. Several ML approaches were analyzed in [9]. By combining a Gaussian Mixture Model with a Shifted Delta Cepstral (SDC) function, one may get the optimum outcomes. The use of long-short-term memory (LSTM) recurrent neural networks (RNNs) for automatic language recognition of brief utterances is examined in this literature analysis by Ruben Zazo*, Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez [10].

In order to identify languages, Rong Tong, Bin Ma¹, Donglai Zhu, Haizhou Li, and Eng Siong Chng [11] investigated the fusion of five traits at various levels of abstraction. On the 1996 NIST dataset, which included 30-second speech segments from 12 different languages, the 5-feature fusion technique produced an EER of 2.38%. One-level HTM organization surpasses the gold standard method in classifying four languages, as shown by research by Dan Robinson, Kevin Leung, and Xavier Falco [12]. The typical three-layer model has around 20,000 features. Given that our training error is often 0% at the speech level, this variance problem might be a major contributor to their blunders.

Bin MA and Haizhou LI confirmed the process for allocating language IDs to five Asian languages [13]. For sessions lasting 10 seconds, the average speech classification accuracy is 98.1%. With gratitude to Hongbin Suo, Yonghong Yan, Ming Li, Xiao Wu, and Ping Lu We employ several classifiers based on speaker groups to successfully map human speech into DLCSV space. the results of which are displayed below. Using information from the 2003 NIST Language Recognition Evaluation, we put the

proposed SVM system to the test. Chinese academics Yonghong Yan, Hongbin Suo, Xiao Wu, Ping Lu, and Ming Li [14] Using a group-based discriminant method, speakers' voices are effectively mapped into the DLCSV space. Classifiers are used in this investigation. We test the proposed SVM system using the NIST Language Recognition Evaluation data sets from 2003.

Yeshwant K. Muthusamy, Ronald A. Cole, and Beatrice T. Oshika[15] built the speech database to facilitate studies of automated language identification and multilingual voice recognition. We listened to recordings of the opinions of one hundred different individuals. Using phonological and lexical models, James L Hieronymus and Shubha Kadambe [16] created a five-language system. It was found that the system trained using phonetically hand-lapped data performed the best.

Every language has distinctive characteristics. For instance, the kinds, lengths, and phonemes of words might vary greatly from one language to the next. Altering the word frequency and word order are other possibilities. The researchers Berkling and Barnard [17] used a generalized phoneme to classify spoken languages. Both English and Japanese were in their vocabularies. They claimed a 90% success rate in differentiating between the two tongues.

Another method of distinguishing languages is to divide speech into various phonetic groups depending on the acoustic structure of the language [19]. In 1996, Zissman examined the four different methods of language identification to see which was most reliable [18].

Using the method described above, Biadisy and Hirschberg [19] tried to differentiate between four different Arabic dialects.

Zissman examined four different approaches to identifying languages in 1996 [18]. The acoustic model is an additional technique for language recognition. In this method, the speaker conveys the key idea verbally. Language is often identified using one of three data types: MF, L, or CP. In the field of automatic speech recognition, mel-frequency cepstral coefficients (MFCC) are now the gold standard. It turned out that MFCC was the best option. There are a plethora of methods presented by researchers who have

attempted to classify languages by accumulating evidence from spoken languages. Now, neural networks are being trained to better extract frequency from human speech.

2.2 Terminologies:

Years of study have led us to the conclusion that there are several methods for honing language recognition skills. Increasing output has always been prioritized. Differentiating across languages in conversation requires keeping various features and pieces of information from the original speech signal intact. Existing data is also used to establish the language.

Language and gender information were extracted from the speech stream to create a biometric verification system. The technology was tested with ten different languages. Accurate recognition rates were achieved using functional acoustic features and the best model training approach by Li, W., Kim, D.J., Kim, C.H., and Hong, K.S. [1]. To perform a spoken-language recognition (SLR) prediction using senone posterior.

Sensor-based deep neural network approaches for spoken voice interpretation were examined by Ferrer, Lei, McLaren, and Scheffer [2], who also compared several DNN training strategies. SLR, NIST, LRE, and RATS. Throughout the whole test, GMM/IV systems were used for both activities.

Automatic voice recognition relies on accurately characterizing the shared features of different spoken languages. An innovative universal method for characterizing the acoustic properties of human speech was suggested by Siniscalchi SM, Reed J, Svendsen T, and Lee CH. [3]. (LRE).

H. Li, B. Ma, and K. A. Lee [4] sought to offer a primer on the principles of theory and cutting-edge solutions from the phonological and computational viewpoints in their article "Spoken Language Recognition: From Principles to Practice." Furui S and Juang BH[5] saw automatic speech understanding and recognition as the first step toward true human-machine communication by integrating a forward-backward technique, cepstral distance, Bayesian risk, and linear prediction. For short speech segments, Wang MG, Song Y, Jiang B, Dai LR, and McLoughlin I [6] examined example-based language recognition algorithms. After looking into and comparing the efficacy of various coding

methods, it was discovered that the comparatively simple cosine coder yielded the best results. Standardized Variable Modeling (SVM) was utilized to refine existing processes and standardize newly introduced features.

Drugman T, Kane J, Raitio T, and Gobl C [7] attribute this to the influence of environmental variables. A squeaky voice was being developed. It is shown that the improved prediction technique performs better than the clumsy detection strategy on which HMMs have been trained. Some of our public speaking is less than stellar. Ooster J. and Meyer B.T. [8] investigated a technique for assessing speech quality called deep machine listening (DESQL). Using phoneme posterior probabilities obtained from a deep neural network, DESQL forecasts perceived speech quality.

The language recognition paper was written by Justin Pyron, Andrew Deveau, and Julien Boussard. (9) examined a wide range of ML techniques. The best outcomes are obtained when Gaussian Mixture Models and Shifted Delta Cepstral (SDC) functions are combined. The effectiveness of long-short-term memory (LSTM) recurrent neural networks (RNNs) for automatic language recognition for brief utterances was examined by Ruben Zazo*, Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez [10].

Five qualities were combined at various degrees of abstraction by Rong Tong, Bin Ma¹, Donglai Zhu, Haizhou Li, and Eng Siong Chng [11] to identify across languages. On the 1996 NIST dataset, the 5-feature fusion system scored 2.38 percent EER on 30-second speech segments in 12 different languages. In a job to classify four languages, Dan Robinson, Kevin Leung, and Xavier Falco [12] show that their one-level HTM setup outperformed the baseline system. Typically, their three-layer model would include more than 20,000 separate features. Given that our training error is usually 0% at the speech level, this variance issue may be a large contribution to their mistakes.

Bin MA and Haizhou LI [13] verified the procedure used to assign language IDs to five Asian languages. In 10-second speech sessions, 98.1 percent of the time, we are able to correctly categorize the speaker. The results of this study by these authors (Hongbin Suo, Xiao Wu, Ping Lu, Ming Li, and Yonghong Yan) are presented here. Spoken language is effectively mapped onto DLCSV space using various classifiers based on

speaker groups. NIST Language Recognition Evaluation datasets from 2003 are used to test the proposed SVM system. Authors: Hongbin Suo, Xiao Wu, Ping Lu, Ming Li, and Yonghong Yan [14]. To effectively map human speech into the DLCSV space, we use a group-based discriminant approach. This study makes use of classifiers. The suggested SVM system is tested, and the results are compared to those of the NIST Language Recognition Evaluation datasets from 2003.

The speech database was created by Yeshwant K. Muthusamy, Ronald A. Cole, and Beatrice T. Oshika[15] in order to aid research on automatic language identification and multilingual voice recognition. 100 people's comments were recorded and used. Using phonological and lexical models, James L. Hieronymus and Shubha Kadambe [16] created a five-language system. The best performance was achieved by the system that was taught using phonetically hand-lapped data.

It's true that every language has its own unique characteristics. Different languages, for instance, have a wide variety of word styles, lengths, and phonemes. Likewise, the frequency with which words appear in a sentence might shift. It was with a wide phoneme that Berkling and Barnard were able to classify languages [17]. They spoke well in two languages: English and Japanese. They claimed a 90% success rate in differentiating the two tongues.

Phonetic classifications based on the acoustic structure of the language allow for further differentiation across spoken languages [19]. To determine which of four language identification methods was most reliable, Zissman compared them all in 1996 [18].

Biadsy and Hirschberg used the method described above [19] to try to differentiate between four different Arabic dialects.

In 1996 [18], Zissman evaluated four distinct approaches to identifying languages to determine which was most reliable. An acoustic model is an additional tool for language recognition. The idea is conveyed orally in this method. More frequently than not, MF, L, or CP data types are employed to specify the language being spoken. Recent research has shown that mel-frequency circular efficiency (SR) coefficients are the most effective way for automated speech recognition. As expected, MF was the most

efficient. Numerous methods have been developed by researchers who are seeking to detect languages by accumulating data from spoken languages. The neural networks used to extract frequency from the human speech are still undergoing development.

2.3 Contrast this with a summary and compare:

There are several language identification systems with well-known names. We have also known how these methods function from the previous paragraph. The results of the study using these methods will now be discussed. Additionally, we will go deeper into the information formats they have employed.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Methodology:

Prior to the data being categorized, it must first be separated by language. In order to categorize the data, we will first separate it into its respective linguistic groups, then extract the characteristics from those groups. Before we could get started, we needed to gather datasets from several online resources. Once the data has been collected, we must decide how to organize and clean it before we can utilize it effectively. In order to isolate the feature, the phrase "MFCC" may have been removed from the audio recording, as we may find out by listening to a few videos on YouTube. [20]. Then, we work to understand MFCC better and how it operates. In order to do that, we try to understand the MFCC approach [21]. Python was the major language used to code this project. For machine learning, we used the Anaconda package. Decision trees, KNN, Random Forest, Support Vector Machine (SVM), and Gaussian Naive Bayes, are some of the approaches used.

How to Put Our Model to Work:

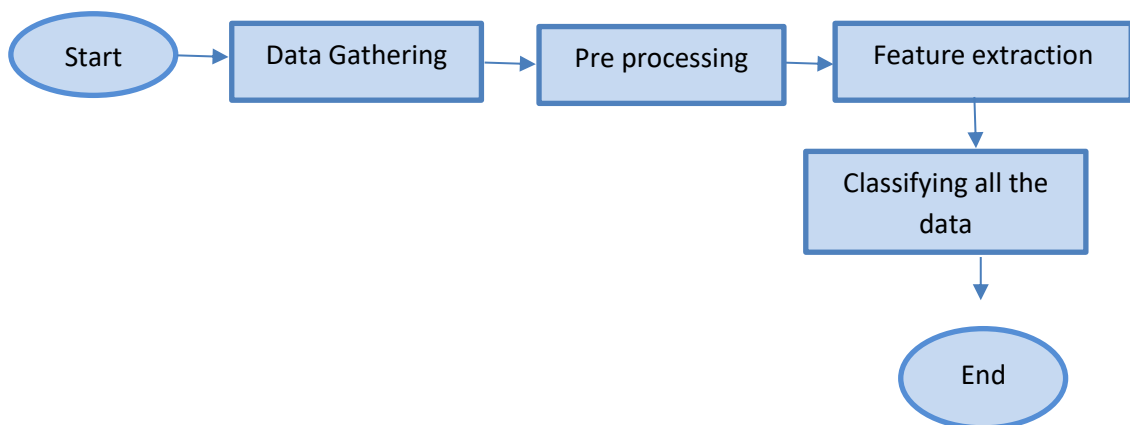


Figure 3.1.1: Flowchart for a working method

3.2 Method for data collection:

Data Collection Methods 3.2 About 800 unique entries make up our dataset. There are a hundred data points in all, distributed throughout the eight languages. Accessible online audio is offered in a variety of languages, including Korean, Chinese, Japanese, Bangla, Russian, German, Spanish, and English. To start, we use various techniques to get audio data from different sources for selected eight languages. It has been discovered that the size of the language files may provide a serious obstacle to reliable categorization. Therefore, we recommend creating brief snippets by cutting the audio in Audacity for each language individually. The results of our study show that male and female viewpoints are given equal consideration. Because we wanted to guarantee the reliability of our results, we restricted ourselves to a single data set while developing our classification algorithm.

3.3 Statistics for Analysis:

Table 3.3.1. A chart for our data

Selected Languages	Number of facts overall
Spanish	One Hundred
Russian	One Hundred
Korean	One Hundred
Japanese	One Hundred
Germany	One Hundred
China	One Hundred
Bangla	One Hundred
English	One Hundred

3.4 Proposed Approach:

Here, we'll try to go deeply into the system's approach to voice recognition. Our whole method consists of three distinct stages. Data preparation, feature extraction, and machine learning classification are a few examples of these subsystems. The practice of preparing raw data for further processing is known as "pre-processing." In lieu of self-training, the computer picks and chooses the characters it needs from the available set. The term "feature extraction" is used to describe the whole operation. For machine learning, the training phase consists of training and training. During the training process, the computer turns the data into knowledge. Following training, the device is put to the test. The computer does the tests by consulting its ever-growing knowledge store. Finally, a decision has been made.

Plotting the Librosa audio data:

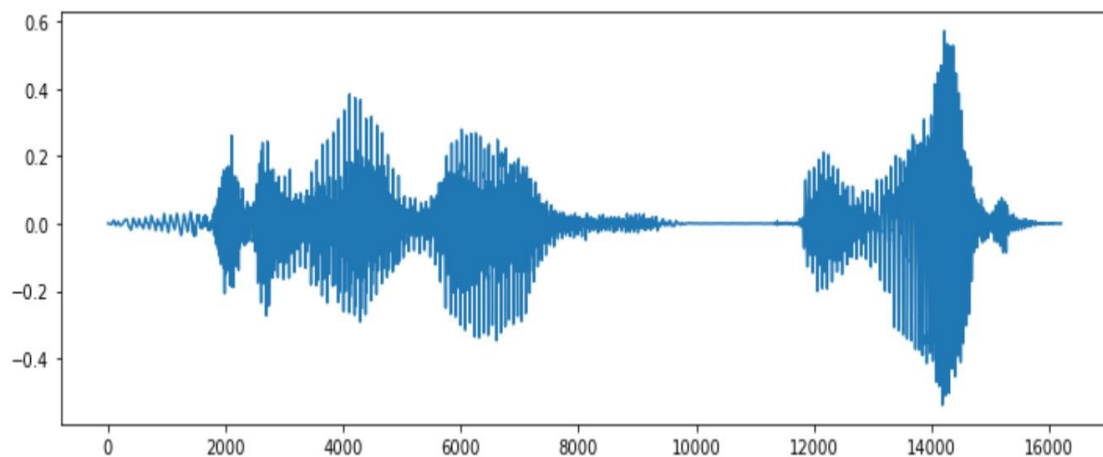


Figure 3.4.1: Audio data plotting using librosa

3.4.1 Pre-Processing:

Starting with pre-processing is the first stage in the procedure. Many methods are required at this time to bring the data into the same layout. A variety of male and female voices are used, and the resulting .wav files are compressed using the Audacity program. All of these features make this endeavor extraordinary.

3.4.2 Feature extraction:

Here, I used librosa's library to extract MFCC's features, which I labeled as 3.4.2 Feature Extraction. In this case, the development of a function called MFCC has taken place. The term "feature extraction" has a specific meaning here, and that is simply the characteristics that may be seen and potentially utilized to inform my prediction. Retrieved from the data are characteristics at many levels. Since we can only acquire pixels from images, our progress has only been significant with audio. However, chromatography is the source of the numerous characteristics that will be linked with my work. We may use a broad range of options. While MFCC is implemented in the case of language recognition, no additional action is performed. After extracting features, we split the data frame into features and classes. After that, I went and got the feature set. The characteristics have also been encoded.

3.4.2.1 (MFCC) Mel-frequency cepstral coefficients:

The first stage in MFCC extraction is to gather the necessary information for feature extraction. Many MFCCs are used in the process of voice recognition. More specifically, the frequency bands of the cepstral representation of the signal used by MFCCs are separated according to the mel-scale rather than being equally spaced. It is OK to use the word "spectrogram" if the MFCC has been located. Here we may see the spectrogram values evolving with time. The values of the MFCC were charted to create this diagram. In this illustration, we show the waveform structure of our data. Types of MFCC and their frequencies

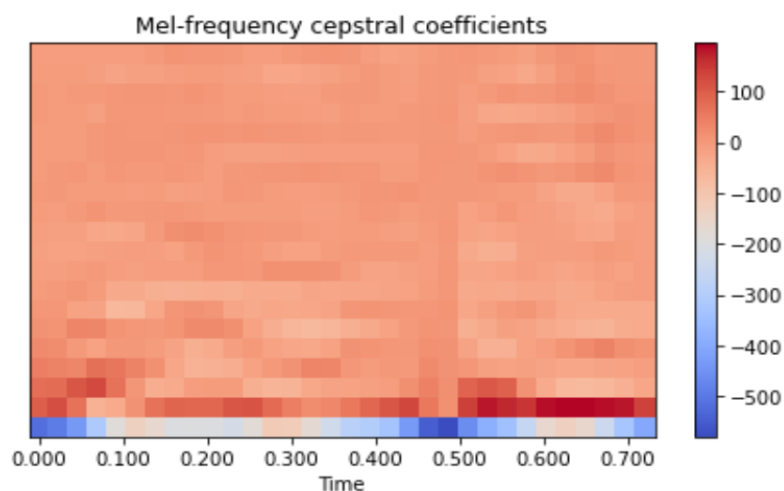


Figure 3.4.2.1.1: Spectrogram plot

3.4.2.2 The transformation of analog to digital:

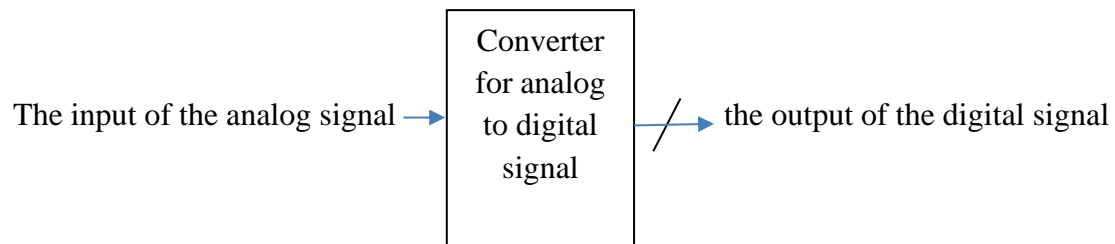


Figure 3.4.2.2.1: Analog to digital data transformation

3.4.3 Machine Learning:

Given that not all algorithms are created equal in terms of performance, we looked at quite a few of them. We used a combination of six machine-learning techniques for classification. among them Support Vector Machines, Random Forests, Naive Bayes, Decision Trees, and Kernel Neural Networks.

3.4.3.1 Decision tree:

If I have data and the data choose certain roots as the starting point for the decision tree, the roots will try to predict the classes above the condition. For the sake of argument, let's say I exclusively use the digits 0 and 1 in my categorization system. They're both either Class 0 or Class 1. Once this begins to happen, it begins to split apart conditionally. This way, it may be split into many distinct trees. to the lowest possible node in the root system from the leaf node. Finally, r goes to a node with no children or parents, where it selects a class value between 0 and 1. It is common practice to use a decision tree to do this. Still, if I have a continuous value, then changing any one of the values will affect the tree's overall structure. That judgment might be reversed under different conditions. The decision tree's worst flaw is exactly this. A decision tree has difficulty processing continuous data. It could be quite useful for fixing judgment problems. If there's a problem, it's good to think about every possible outcome. When compared to traditional approaches, more work is required to clean up the data. [23].

Decision-making graph:

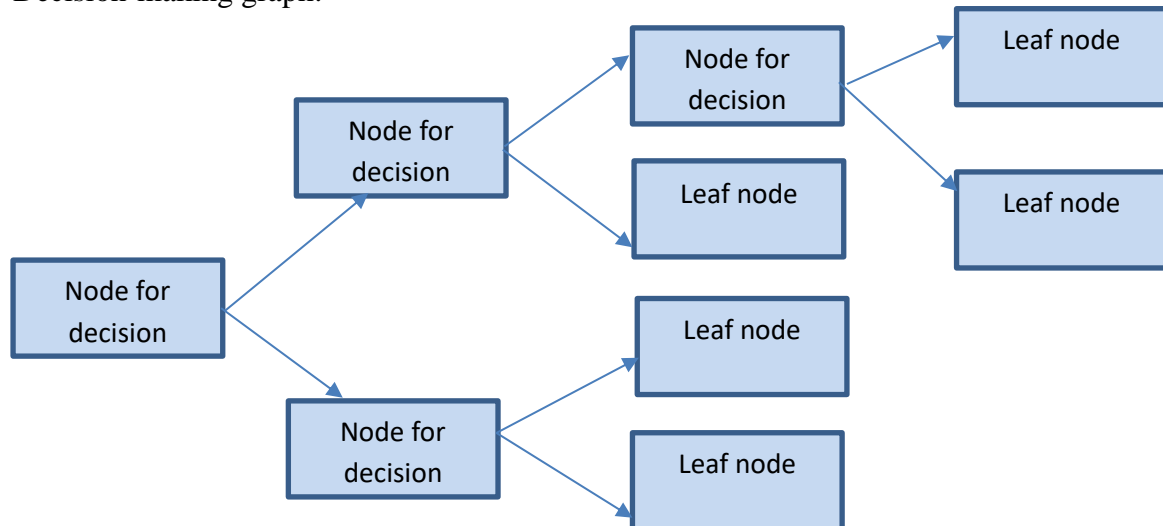


Figure:3.4.3.1.1 Diagram for decision tree

3.4.3.2 Random Forest:

The random forest provides a satisfactory response to continuous data. It's common for there to be discrepancies between the various groupings in random woods. This may imply that he will break the material down into manageable chunks. Rather of building a single large tree from all of my data, I'll use random sampling to choose smaller subsets of data and utilize them to build many trees. Finally, the outcome will be generated by linking together all of the trees. In other words, if a single tree out of many consistently yields the same results, then that tree will be labeled as a class. Using Random Forest has this benefit since it divides my whole dataset into many subsets in a completely random manner. Therefore, in all of these situations, Random Forest is beneficial [25].

3.4.3.3 KNN:

My data takes on a structure when I plot it for KNN, indicating groups of data that are probably connected. Due to the fact that I am working with 8 different languages, data from the equivalent languages will be shown together. Because it may theoretically be placed anywhere, we utilize KNN to make predictions about where it will likely end up. In order to get the KNN value, I will randomly arrange data points in a specific area and then choose that number of them. After picking a random starting location, we'll

calculate the average distance to each of the primary points. In a similar vein, the distance should be calculated using the Euclidean method. By computing the distance in this way, it may be shown whether two sets of data belong to the same class. This is how KNN works [23].

Distance in terms of Euclidean geometry:

$$\text{Distance analysis: } \sqrt{[(c_2 - c_1)^2 + (d_2 - d_1)^2]}$$

3.4.3.4 Gaussian Naive Bayes:

Simply said, the naive Bayes method is used for classifying data. The naive Bayes approach relies heavily on conditional probability when working with probability trees. What are the probabilities of one condition being fulfilled if another condition is also satisfied? This is, in essence, a Naive Bayes approach. In this instance, the results are continuous numbers. Bayes theorem:

$$p(B|C) = \frac{P(C|B)P(B)}{P(C)}$$

Taking on a continuous value, as opposed to a static one, stimulates the gaussian prediction's predictor values.

Theorem of Gaussian Naive Bayes:

$$P(bi|c) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(bi - \mu_y)^2}{2\sigma_y^2}\right)$$

According to the Gaussian Naive Bayes model, each parameter has the ability to predict the outcome variable on its own [22]

3.4.3.5 SVM:

It's hard to see myself with just two meals. Because of this, I'm going to draw a line through the middle of these people. Also, I am able to eliminate this stain regardless of the vantage point. This strategy allows me to establish connections between several

locations. In this case, though, I need to choose many lines and calculate the distance between each point and each selected line. Distances are determined between each line in order to exclude all except the best-fitting. It will be a closer and shorter distance. Like thus, it can be determined. I'll choose the closest line as my best-fit, and that's the one I'll use to classify everything [23] [27].

SVM diagram:

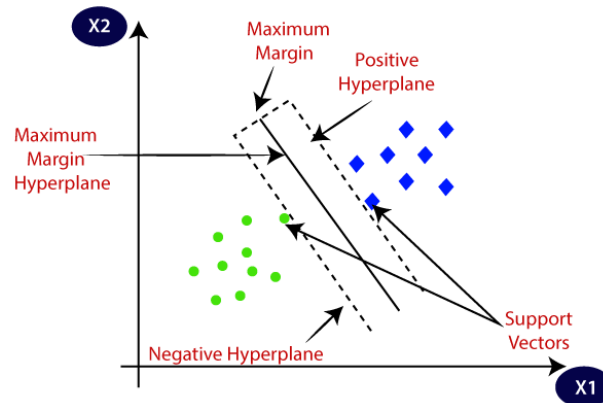


Figure:3.4.3.5.1 Diagram for SVM

3.5 Evaluation:

Simply training and testing our model won't provide us with an accurate picture of its performance. Our model was evaluated using a broader set of measures.

3.5.1 Precision:

It is the success rate of the most optimistic class predictions that matters. Here, we multiply the actual percentage of positive values by the total expected percentage of the values those are positive [26].

$$precision = \frac{TP}{TP + FP(\text{Total Positive Predicted values})}$$

Here, TP means True positive value and FP means False Positive value.

3.5.2 Recall:

The dataset's positive classes will have an impact on the anticipated total of all classes. The recall is the proportion of true positives that our algorithm accurately identified [26].

$$Recall = \frac{TP}{TP + FP(\text{overall real positive})}$$

3.5.3 F1 Score:

Mainly, F1 score takes accuracy and memory into consideration throughout the calculation. The accuracy of a model may be evaluated statistically with its help. To attempt parity in both Precision and Recall, an F1 Score is required. Since the F1 Score indicates an unequal class distribution, it may be more appropriate to utilize this statistic if we also need to prove equality in accuracy and recall [26].

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.6 Implementation prerequisites:

- The WAV format should be used for all audio conversions; a particular language should be utilized when creating the audio (English, Bangla, China, German, Japanese, Korean, Russian, Spanish).
- Here is a rundown of some of the items most often need for our daily activities
- The supplementary noise in the audio must be eliminated, and Anaconda must be installed and configured.
- Creating a working environment for our project on the computer.
- Librosa installation; MFCC audio extraction; research on machine learning methods for classifying languages

CHAPTER 4

RESULT ANALYSIS

4.1 Discussion about the result:

Therefore, we divided our information gathering in half. A test segment and a practice section are included. Two hundred audio speeches are used for evaluation, and another 600 are used for training out of a total of 800 data points. Sklearn Model Selection is used to help us put our data into useful categories. Now that this has been done, we can confidently say that Korean, Chinese, Japanese, Bangla, Russian, German, Spanish, and English are all separate tongues.

Table 4.1.1: The results of the classification

Name of algorithm	Selected language	Score of F1
Random forest	Spanish	1
	Bangla	0.99
	Japanese	1
	China	1
	Russian	1
	Korean	1
	English	0.99
	Germany	1
Gaussian Naïve Bayes	Spanish	1
	Bangla	1
	Japanese	1
	China	1
	Russian	1
	Korean	0.99

	English	0.99
	Germany	1
Decision tree	Spanish	0.88
	Bangla	0.83
	Japanese	0.84
	China	0.83
	Russian	0.95
	Korean	0.90
	English	0.84
	Germany	0.89
SVM	Spanish	1
	Bangla	1
	Japanese	1
	China	1
	Russian	1
	Korean	1
	English	1
	Germany	1
KNN	Spanish	1
	Bangla	1
	Japanese	1
	China	1
	Russian	1
	Korean	1
	English	1
	Germany	1

4.2 Discussion:

My model uses historical data, which results in a high f1 score and outstanding accuracy in the decision tree. My data has been overfit in this circumstance, even though other algorithms are showing better results. It's possible that the model has been overfitted to account for my outstanding results. However, the correct answer was discovered through a single branch of the decision tree. This is because, in contrast to the other approaches, the decision tree provided an objective review of the entire project.

CHAPTER 5

ASSESSMENT OF THE RESULTS AND IMPLICATIONS FOR FUTURE RESEARCH

5.1 Conclusion:

The purpose of this research was to identify the method of machine learning that exhibited the highest level of language recognition ability. The most important contribution that our study has made is the identification of many methods for the extraction of speech features and the implementation of the method that proved to be the most successful in the context of our project. As a result of this analysis, we now know that decision tree algorithms produce the best accurate F1 score for our project. The tree-based approach, such as the decision tree algorithm, produced satisfactory outcomes due to the absence of overfitting in the context of this scenario. Because of the constraints of our data, we also find that the performance of other algorithms is worse than that of random guessing under settings that are equivalent.

5.2 Future study:

It's possible that there are more audio clips available. Adding support for more languages is an important enhancement that should be implemented. For our current configuration, incremental machine learning might prove to be a game-changer. In addition, deep learning and more data will provide better results in future study. Since we will be able to make better informed decisions with the help of deep learning. Compared to visuals, audio's foundation is often far less weighty. Choosing a language based on its sounds might be a simpler alternative for the future, but only if it can be taught well.

APPURTENANCE

Summarization:

ML- Machine Learning

LER- Language Recognition

SVM- Support Vector Machine

SLR- Simple Linear Regression.

MFCC- Mel-frequency cepstral coefficients

LID- Language Identification

SAMME- Stage-wise additive multi-modeling with multi-class exponential loss function

SR- Speech Recognition

KNN- K-Nearest Neighbors

Reference

- [1]Li, Wei, et al. "Voice-based recognition system for non-semantics information by language and gender." *2010 third international symposium on electronic commerce and security*. IEEE, 2010.
- [2]Ferrer, Luciana, et al. "Study of senone-based deep neural network approaches for spoken language recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.1 (2015): 105-116.
- [3]Siniscalchi, Sabato Marco, et al. "Universal attribute characterization of spoken languages for automatic spoken language recognition." *Computer Speech & Language* 27.1 (2013): 209-227.
- [4]H. Li, B. Ma and K. A. Lee, "Spoken Language Recognition: From Fundamentals to Practice," in *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136-1159, May 2013, doi: 10.1109/JPROC.2012.2237151.
- [5]Juang, Bing-Hwang, and Sadaoki Furui. "Automatic recognition and understanding of spoken language-a first step toward natural human-machine communication." *Proceedings of the IEEE* 88.8 (2000): 1142-1165.
- [6]Wang, Meng-Ge, et al. "Exemplar based language recognition method for short-duration speech segments." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.
- [7]Drugman, Thomas, et al. "Prediction of creaky voice from contextual factors." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.
- [8]Ooster, Jasper, and Bernd T. Meyer. "Improving deep models of speech quality prediction through voice activity detection and entropy-based measures." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.
- [9]Boussard, Julien, Andrew Deveau, and Justin Pyron. "Methods for Spoken Language Identification." (2017).
- [10]Zazo, Ruben, et al. "Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks." *PloS one* 11.1 (2016): e0146917.
- [11]Tong, Rong, et al. "Integrating acoustic, prosodic and phonotactic features for spoken language identification." *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 1. IEEE, 2006.
- [12]Robinson, Dan, Kevin Leung, and Xavier Falco. "Spoken language identification with hierarchical temporal memories." (2009).
- [13]Ma, Bin, and Haizhou Li. "Spoken language identification using bag-of-sounds." *International Conference on Chinese Computing, Singapore*. 2005.
- [14]Li, Ming, et al. "Spoken language identification using score vector modeling and support vector machine." *Eighth Annual Conference of the International Speech Communication Association*. 2007.
- [15]Muthusamy, Yeshwant K., Ronald A. Cole, and Beatrice T. Oshika. "The OGI multi-language telephone speech corpus." *second international conference on spoken language processing*. 1992.
- [16]Hieronymus, James L., and Shubha Kadambe. "Spoken language identification using large vocabulary speech recognition." *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. Vol. 3. IEEE, 1996.

- [17]Berkling, Kay M., and Etienne Barnard. "Language identification of six languages based on a common set of broad phonemes." *ICSLP*. 1994.
- [18]Zissman, Marc A. "Comparison of four approaches to automatic language identification of telephone speech." *IEEE Transactions on speech and audio processing* 4.1 (1996): 31.
- [19]Biadisy, Fadi, and Julia Bell Hirschberg. "Using Prosody and Phonotactics in Arabic Dialect Identification." (2009).
- [20]YouTube, available at << <https://www.youtube.com/watch?v=dUmSHIduo3c>>> last accessed on 01-02-2020 at 9:49 PM.
- [21]Medium, available at <<https://medium.com/@jonathan_hui/speech-recognition-featureextraction-mfcc-plp-5455f5a69dd9>> last accessed on 01-02-2020 at 9:53 PM.
- [22]Vats, R.(n.d). Gaussian Naïve Bayes: What You Need to Know? Retrieved from upGrad.com: <https://www.upgrad.com/blog/gaussian-naive-bayes/#:~:text=features%20are%20independent,-,What%20is%20Gaussian%20Na%3%AFve%20Bayes%20algorithm%3F,theorem%20with%20strong%20independence%20assumptions>
- [23]Ilias G. Maglogiannis, K. K. (n.d.). Emerging Artificial Intelligence Applications in Computer Engineering. Retrieved from Books: https://books.google.com.bd/books?hl=en&lr=&id=vLiTXDHR_sYC&oi=fnd&pg=PA3&dq=SVM+classifier+in+machine+learning&ots=CZrtzy-Bhr&sig=GKfevaMJhcrjzuxbZRauGrKEwrk&redir_esc=y#v=onepage&q=SVM%20classifier%20in%20machine%20learning&f=false
- [24]Schapire, R.E., 2013. Explaining adaboost. In Empirical inference (pp. 37-52). Springer, Berlin, Heidelberg.
- [25]Ren Q, Cheng H, Han H. Research on machine learning framework based on random forest algorithm. In AIP conference proceedings 2017 Mar 13 (Vol. 1820, No. 1, p. 080020). AIP Publishing LLC.
- [26]Shung, K. P. (n.d.). Accuracy, Precision, Recall or F1? Retrieved from Towards Data Science: <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>
- [27]SupportVectorMachineAlgorithm.(n.d). Retrieved from javatpoint.com: <http://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

VOICE LANGUAGE PREDICTION USING MACHINE LEARNING

ORIGINALITY REPORT



PRIMARY SOURCES

1	Submitted to Jacksonville University Student Paper	17%
2	Submitted to Daffodil International University Student Paper	2%

Exclude quotes Off
Exclude bibliography On

Exclude matches < 1%