IDENTIFICATION OF SPOKEN LANGUAGE USING MACHINE LEARNING

MD. Asif Shahariar
ID: 162-15-7772

Iftekher Aziz
ID: 162-15-8182
AND

Shovan Banik
ID: 162-15-7790

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

Supervised By

Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Zahid Hasan
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH

JULY 2020

APPROVAL

This Project/internship titled **"Identification Of Spoken Language With Machine Learning"**, submitted by MD. Asif Shahariar, ID No: 162-15-7772, Iftekher Aziz, ID No: 162-15-8182 and Shovan Banik, ID No: 162-15-7790 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 08th July, 2020.

BOARD OF EXAMINERS

____

Dr. Syed Akhter Hossain **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Sadik
____

Md. Sadekur Rahman **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

____

Saiful Islam **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Motaharul Islam.
____

Dr. Md. Motaharul Islam **External Examiner**
Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mr. Abdus Sattar, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Zahid Hasan
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Asif.

Md. Asif Shahariar
ID: 162-15-7772
Department of CSE
Daffodil International University

Iftekher Aziz

Iftekher Aziz
ID: 162-15-8182
Department of CSE
Daffodil International University

Shovan

Shovan Banik
ID: 162-15-7790
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty ALLAH for His divine blessing makes us possible to complete the final year project successfully. Also, there are some others who gave us faith and right way to complete our task properly. From core of our heart, we want to thank them all.

We really grateful and wish our profound our indebtedness to **Mr. Abdus Sattar**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" 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 stage have made it possible to complete this project.

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

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

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

ABSTRACT

Identification of spoken language is the way to detect the specific language which is spoken by an anonymous speaker. We will also find out several techniques of machine learning for detecting spoken language. Our major task is to identify parameters and features from spoken language that can be used to separate languages. To extract feature from audio file we will use Mel Frequency Cepstral coefficient (MFCC). So far, many methods have been used for language identification (LID). Of all the techniques, the accuracy of machine learning is the best. That's why we also used machine learning in our project for lid. Our system will train with 30,000 data. This project aims to classify Spanish, German & English languages. Main goal of this project is to find out best algorithm for detecting specific language. We get the best accuracy from random forest algorithm.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Introduction:

The events of globalization have brought people together all over the world. There has always been a barrier to increase global communication and that is different people use different languages. This has basically made our communication weaker. To solve this problem, both parties should use a language of communication that is understandable to all of them. Language identification undoubtedly plays an important role here. The use of language as a medium of communication is very popular all over the world. But language among the means of communication is quite difficult compared to other mediums because it is based on complex rules and its meaning and acceptability may vary from place to place. Yet over the centuries, language has become the most widely accepted system for communication.

We know the process of language identification as spoken language identification (LID). In more detail, automatic language detection is the ability to tell which language a speaker is speaking by taking some samples from his speech and testing that sample. So far people have been doing this job of identifying language with best result. If a person hears a language that he is aware of, then he is able to tell in the blink of an eye what that language is spoken by the speaker. But the problem arises when they are not aware of the language. Then they can't guess what language they are listening to. Whenever we encounter this problem, we need the help of machine learning to solve it. Since it is not possible for one person to master all the languages in the world. So, we will take the help of Artificial Intelligence to solve this problem. Whom we are able to teach very easily

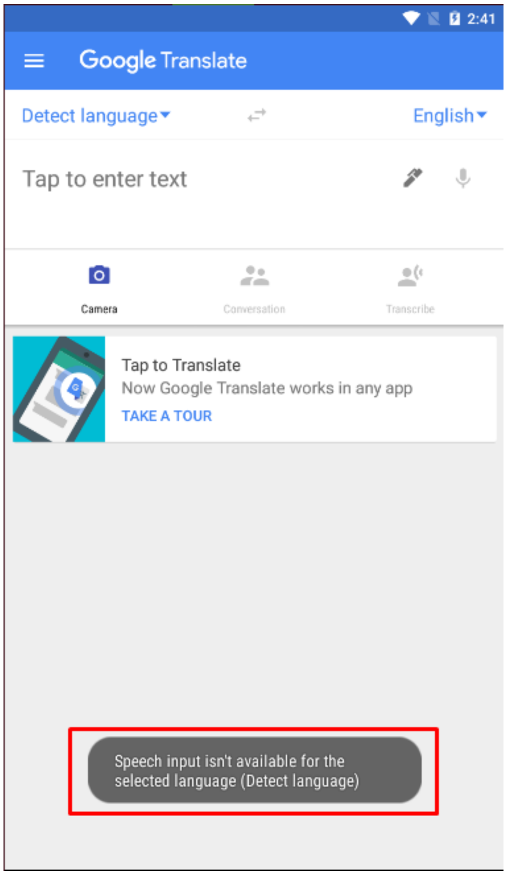
about all the languages we have in the world. This technology is known as automatic spoken language identification. This technology has been used in spoken language translation [1], spoken document retrieval [2] as a key technology. Many people tried to solve this problem in the same way. But the method of teaching machine, was different.
Machine learning has made great strides in recent years. Now a days we can control technology only through the use of voice. But these modern instruments also have language limitations. We have to teach our machines which specific language they will take as input. Then follow the next steps. On the other hand, if we can teach machines to determine all languages and give people the benefits of each in their own language, then the horizons of unprecedented success will be opened. Not only in this case but also in the field of intelligence and security success will be evident where identities of language are important in recorded messages and materials which are archived [3]. Space processing will be our first and foremost tool in language identification.

The pronunciation of words or sentences is an audio signal. Speech processing is the study of these signals and process methods those signals. Those signals processed in digital reorientation; speech processing is a part of digital signal processing. In order to identify a language, we need to separate its information from the audio file. Because we know that by breaking down different audio forms, we can collect different information from those audios. And through this we will try to find out different types of languages.

1.2 Motivation:

This problem comes to our notice when we see that the spoken language can be translated very well in Google Translator, but before translating, we have to select the languages. If someone encounters an event that requires him to continue the conversation with someone but he does not know or does not understand what language another person is talking to. In that case he will not be able to use Google Translator even if he has it. In this case, he could not use Google Translator, not only because he could not select the language. So far google translator has not been able to bring this service where language auto selection can be done from speech. When asked to continue the conversation in auto selection, google translator says they have not started it yet. The matter can be clearly seen through the following picture.

Figure 1.2.1: Language auto detection from speech isn't available in Google Translator [4]



Now Google Translate works in any app

Speech input isn't available for the selected language (Detect language)

1.3 Rationale of the Study:

More than 500 million people [4] are using google translator. If we can properly detect the language through speech recognition, we will be able to make the work of many people much easier by ensuring its proper use in many more apps including google translator. From this thought we will try to find the most accurate algorithm for determining English, Spanish, German language in our project.

1.4 Research Question:

- Create a system that can detect the language spoken by the speaker. We will first work on identifying English, Spanish and German languages.
- We will input audio file to train the system. Which the system will automatically extract and use as information for language identification.
- The system will also be taught with noise at different inputs so that users do not have to worry about applying this system in real life.

1.5 Expected Output:

- Creating a system that can automatically detect English, Spanish and German language.
- System will not dependent on specific vocabulary
- System is not affected by the gender of the speaker.
- The effectiveness of the system will not be reduced based on the type of speaker.

CHAPTER 2
BACKGROUND

2.1 Related Works:

Research on spoken language identification began in the 1970s. At the end of almost 5 decades of research, we have seen that language recognition has been practiced in different ways. Attempts have been made to achieve maximum efficiency. In the case of spoken language identification, various information and features have to be saved from breaking the speech signal. And that stored information is used to identify the language.

Phonotactic, prosodic or acoustic tactics are used to break down speech signals and extract different information from them. Of the 3 mediums mentioned above, the phonotactic method mainly deals with the syllable level or phoneme. The difference between the pronunciation of a word or a phrase is a phoneme. K.M. Berkling, T. Arai and E. Barnard worked in 1994 to identify the language using this difference in pronunciation. [5]

Hieronymous and Kadambe tries to identify language using Large Vocabulary Automatic Speech Recognition (LVASR) [6]. Each language has its own uniqueness. For example, words in different languages have different lengths, different phoneme and different types. Even word sequences and word frequencies are different.

Berkling and Barnard used A Broad Phoneme [7] to identify the language. They worked with English and Japanese. They claimed they could identify the two languages with 90% accuracy.

Language can also be identified by dividing speech into different phonetic categories based on the acoustic structure of language. [8]. In 1996 Zissman compares each of the four types of language detection processes [9] and tries to find out the best method between them. Those four models are respectively: Single-language phone recognition, Gaussian mixture model (GMM), n-gram language modelling (PRML) and language depended parallel phone recognition (PPR). Haizhou Li, Bin Ma and Chin-Hui Lee introduced A Vector Space modeling to identify language.

Prosodic method is another way through language identification can be done. Lin & Wang introduces us to this process of language identification [10]. Biadsy and Hirschberg works to identify the four dialects of Arabic in the above method [11]. The four languages are respectively: Iraqi, Levantine, Gulf & Egyptian. Boussard, Deveau and Pyron also worked with five different method for detecting language [12]. Those models are: Music-genre motivated approach, Feed-Forward Neural Network, Recurrent Neural Network, Convolutional Neural Network & Gaussian Mixture Model. They tested those models with English and Chinese languages. Long Short-Term Memory (LSTM) recurrent neural networks can be used to identify language. This method proved by Ruben Alicia & Javer [13].

Another method of language identification is called acoustic model. In this process cepstral is collected from speech. MFCC, LPC, PLPCC type cepstral data are most commonly used for language identification. Nowadays, Mei-frequency cepstral coefficients are most commonly used for automatic speech recognition (ASR). MFCC has proven to be the most effective.

We have learned the names of various methods through researchers have worked to identify languages by collecting information from speech. Currently, the use of neural network has been enriched to extract the frequency from the spoken word.

2.2 Terminologies:

There is some commonly used algorithm to identify the spoken language. One of them is Support Vector Machine (SVM) which is based on generalized linear discriminate sequence (GDLS) and it has been widely used in spoken language identification.

In a paper we got Ming Li and Yonghong Yan used SVM with GLDS kernel and Support Vector Modelling [22] to identify spoken language. Here is how the SVM works with GLDS kernel

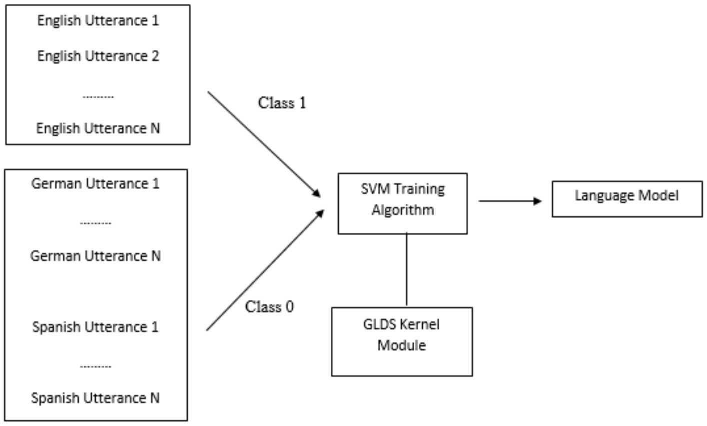


Figure 2.2.1: How SVM works with GLDS kernel [23]

They proposed & implemented a language identification (LID) system to make more efficient the SVM with shifted-delta-cepstral (SDC). The system works within the following steps:

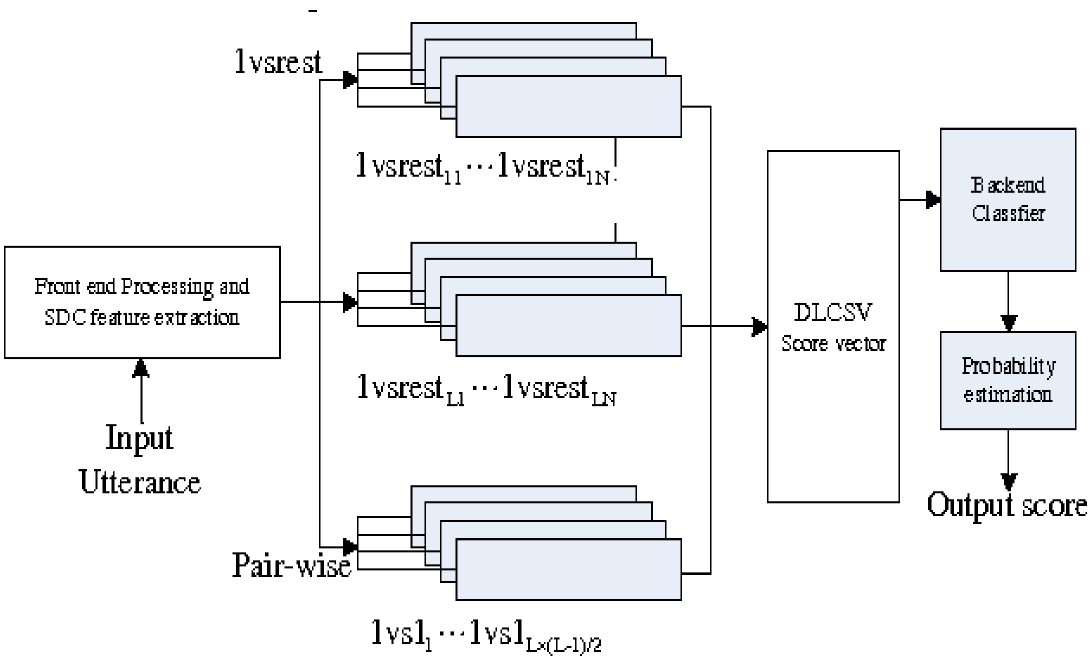

1vsrest
$1vsrest_{11} \cdots 1vsrest_{1N}$
Front end Processing and SDC feature extraction
Input Utterance
$1vsrest_{L1} \cdots 1vsrest_{LN}$
Pair-wise
$1vs1_{1} \cdots 1vs1_{L\times(L-1)/2}$
DLCSV Score vector
Backend Classfier
Probability estimation
Output score

Figure 2.2.2: Proposed LID System Overview [26]

The features of their system [23] ware 7 MFCC coefficients put an end with SDC 7-1-3-7 feature, where there were 56 dimensions in total coefficients per frame. While doing support vector model they converted every spoken utterance into a feature vector with its attributes so that a discriminative vector space classifier (DLCSV) can built the score vector space to make successful identification of targeted language. There a backend outspread premise work bit SVM classifier was done to separate objective dialects dependent on the likelihood dissemination in this DLCSV space. After test expressions

DLCSVs were created and the backend SVM classifier evaluated the back likelihood of each target language, which is utilized to align the final outputs.

Another spoken language identification method was proposed by Bin MA & Haizhou LI [22] to identify spoken language using a sound recognizer and a bag-of-sounds (BOS) classifier. In this paper they tried to build a language identifier for five Asian language. The BOS classifier system work like following diagram:

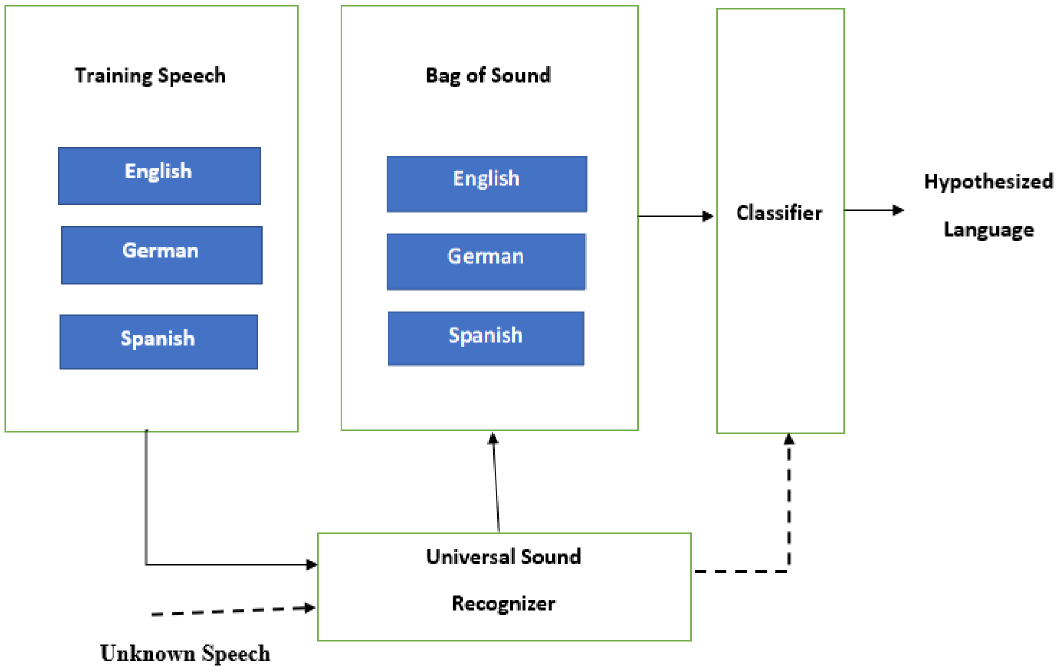


Figure 2.2.3 BOS Block Diagram [22]

The BOS system did new paradigm of feature extraction of specific language from different feature of languages features. Where sound recognizer symbol is like S0, S1, S2, ... Sn is a monophony symbol. Here high-dimensional vector classifier was used with SVM classifier on its original high dimension space. The sound recognizer was tested on an Abacus platform. The data was disjoint into two separate databases for training and testing the BOS classifier. On testing of this system database speech length were 5, 10 & 15 and a duration of speech utterances ranges were 2 to 5 seconds. By two-way end pointing the speech length was measured. It could make a successful classification accuracy 98.1% to identify a spoken language.

Spoken language identification is a unique concept so that many people try to solve this using machine learning program. In a paper we saw that Julien Boussard, Andrew Deveau and Justin Pyron used many methods and algorithms to solve this concept. [24] They used Music-genre motivated approach, Feed-Forward Neural Network, Recurrent Neural Network, Convolutional Neural Network and Gaussian Mixture Model in their project. Which one is give them best result and how these methods are work here we discuss about it.

First of all, they used Music-genre motivated approach. In this approach they used Mel Frequency Cepstral Coefficients (MFCCs) for capturing speech information. The MFCCs measure how the frequency distributes energy of a signal. This approach gave very poor result.

Then they apply Feed-Forward Neural Network approach. This method handle static MFCCs features only. Using this method, they get better result from music genre approach.

After these, they also apply recurrent neural network and convolutional neural network approach. In recurrent neural model, they used the Keras function named Long Short-Term Memory LSTM for constructed a model. Followed by a dense layer they trained a neural network constitute. But they get poor result again. Convolution neural network and feed-forward neural network are similar but output of a convolution layer is different. For binary classifier they used some Logistic regression like this-

$$\theta^{*} = arg\min_{\theta} \sum_{i}^{m} log\left(1 + e^{-y(i)\theta^{T}\varkappa^{i}}\right)$$

It gives the good result against big dataset but not exactly.

Finally, they used Gaussian Mixture Model (GMM) on their dataset. They used some terms in this model how these terms works are given below

$$p\left(X^{(i)}\middle|l\right) = \textstyle\sum_z P\left(x^i, z\middle|l\right)$$

$$= \sum_z p(x\hat{}(i)|z, l)P(z|l)_,$$

Where z is a phonetic expression and x^(i) is the utterance. For classified the data they use

$$\text{arg max} \qquad \textstyle\prod_{i=1} p\ (x\hat{}(i)|l)$$

Long short-term memory (LSTM) used in deep learning which is an artificial recurrent neural network (RNN) architecture. It can process single and entire sequences of data. LSTM is applicable for anomaly detection, speech recognition, unsegmented, hand writing recognition in network traffic. In a paper we got the uses of LSTM for language identification which is written by RubenZazo, AliciaLozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano and Joaquin Gonzalez-Rodriguez [13] in their paper they mainly work with i-vectors. How the i-vector help us in language identification we discuss about that here.

i-vector is a technology which is used in here as a front-end. Here they discussed about Baum-Welch statistics, Universal Background Model (UBM) and used some equations like this.

$$\boldsymbol{N_m} = \sum_{\boldsymbol{t}} \boldsymbol{P}(\boldsymbol{m}|\boldsymbol{O_t}, \boldsymbol{\lambda})$$

$$\boldsymbol{F_m} = \sum_{\boldsymbol{t}} \boldsymbol{p}(\boldsymbol{m}|\boldsymbol{o_t}, \boldsymbol{\lambda})(\boldsymbol{o_t} - \boldsymbol{\mu_m})$$

Where, $(\boldsymbol{m}|\boldsymbol{o_t}, \boldsymbol{\lambda})$ is the Gaussian occupation probability.

They also used the LSTM RNN model where DNN network topology is a model.

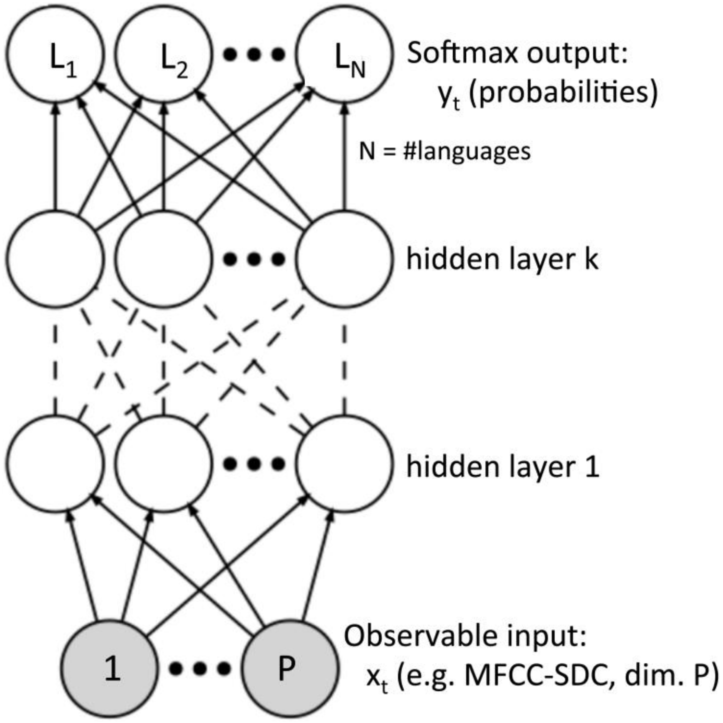


Figure 2.2.4: RNN model [26]

2.3 Comparative Analysis & Summary:

We already know the names of many methods that are used to identify languages. We also know about the working method of the methods as from the previous paragraph. We will now review the researcher's results using the methods. And we will discuss in detail what kind of data they have worked with.

Large vocabulary speech recognition:

Hieronymus and Shubha Kadambe use five languages to train their LID system [6]. They used different types of words amount and length for different languages.

Table 2.1.1: Lexicon sizes for each language

Language	Words	Ave. Length
English	2564	7.47
German	1844	8.34
Mandarin	1546	4.07
Spanish	2014	11.36
Japanese	1863	7.80

The result they have got:

Table 2.1.2. Result of five language identification in large vocabulary speech recognition

Length & Condition	English	German	Mandarin	Spanish	Japanese
50 sec normalized	98%	99%	98%	96%	98%
50 sec grammar	95%	95%	97%	94%	97%
50 sec no grammar	85%	90%	94%	76%	98%
Length & Condition	**English**	**German**	**Mandarin**	**Spanish**	**Japanese**
10 sec normalized	95%	96%	90%	90%	92%
10 sec grammar	85%	84%	85%	81%	87%
10 sec no grammar	81%	82%	81%	79%	83%

From their results, we can conclude that audio files of larger length give better results from smaller length in case of large vocabulary recognition system identification.

Integrating acoustic, prosodic & phonotactic feature:

They work with NIST 1996 & 2003 LRE dataset. In the same way they run same test on two datasets [14]. This is the equal rate error they get:

Table 2.1.3: EER% for individual language

Language	EER%	Utterances(tested)
English	1.56	478
Arabic	1.76	80
French	1.30	80
Farsi	3.15	80
Spanish	2.03	153
German	3.80	80
Mandarin	1.86	156
Hindi	7.92	76
Vietnamese	4.38	79
Japanese	1.20	79
Tamil	4.70	73
Korean	3.51	78

They get the least EER% in English, Japanese, Arabic and French language.

Using Machine Learning:

Patil, Akshay, Harsha & Pramod arranged their datasets randomly [15]. They collected their data from different web and online audio books. This is the result they got for different languages:

Table 2.1.4: Language identification accuracy

Language	Accuracy
English	98.5%
French	97%
Hindi	91.7%
Kannada	96.4%
Japanese	98.3%

Average accuracy they got in their system is 96.42%

Hierarchical Temporal Memories Identification method:

Dan, Kevin & Xavier worked with 8400 utterances including American English, British English, French, Russian & Japanese languages [16]. That's the accuracy they get:

Table 2.1.4: Language identification accuracy

Classification	**Accuracy**
English vs French	98.75%
American English vs British English	92.5%
Four language Classification	92%

Summary:

From the above information and data, we can see that it has been possible to find out the highest accuracy by using machine learning. That value is 96.42%. On the other hand, using large vocabulary speech recognition, the test result of 10 seconds data is 92.02%. And the accuracy of classification of four languages in the hierarchical temporal memories' identification method is 92%. Since the best accuracy is available in machine learning, we take help of machine learning to classify our selected languages.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation:

We will separate the features from our collected data by extraction and classify them using different algorithms with different features for each language. First, we had to take the help of various online websites for dataset collection. After the dataset collection we have to figure out how to break its feature and prepare it for full use. We can learn by watching some videos from YouTube [17] that MFCC is used to separate the feature from the audio file. Next, we try to find out about MFCC, how it works. We try to figure out methodology of MFCC [18]. Throughout the project we used python as the programming language. We have used anaconda software for machine learning. Used as algorithms linear regression, Decisiontree, Randomforest and GradientBoosting.

3.2 Data Collection Procedure:

We have about 30,000 data in our dataset. There are 10,000 data for each of the three languages. We have collected English, German and Spanish audio from online. Not all data was in the required format. We have not changed our format from online to our required format [19]. We keep all our data in flac format. All of our data is 10 seconds long. We use adobe audition to cut extra length audio and make them 10 seconds long. We didn't use just plane audio. To get great results and make real life usable, we have given different types of noise in the plane data. Notable among the types of noise we have given: a few cars passing sound, lots of car passing noise, inside airplane noise, inside train noise, few crowed noise, huge crowd noise, nature sound, nightmare sound, bird noise. By adding the above noise to each noise free audio, we collect different data for each noise. We have also controlled the speed and pitch of speech.[20]. Our data contains an equal number of male and female voices. From a raw data we have made 8 different data which have 8 different pitch and another 8 data also made which have 8 different type of speed. So, from a raw data we different types of more 28 data. To get a good accuracy in our project.

3.3 Statistical Analysis:

Table 3.3.1. Statistics of our data

Language	Data type	Male	Female	Total
English	Raw	172	172	344
	Noise (1-12)	2076	2076	4152
	Pitch (1-8)	1376	1376	2752
	Speed (1-8)	1376	1376	2752
Spanish	Raw	175	175	350
	Noise (1-12)	2025	2025	4050
	Pitch (1-8)	1400	1400	2800
	Speed (1-8)	1400	1400	2800

German	Raw	175	175	350
	Noise (1-12)	2100	2100	4200
	Pitch (1-8)	1361	1363	2725
	Speed (1-8)	1361	1363	2725
				30,000

3.4 Proposed Methodology:

In this section we will explain in detail how language detection works in our system. We have divided our whole process into three different parts. Those parts are: data pre-processing, feature extraction and classify with machine learning. The correct combination of these three parts creates a complete LID system.

Pre-processing includes all the tasks that need to be done to get each data as an input format. The machine extracts the feature from the input data as per its requirement. Because the machine will not be able to train itself with the data that is given. And we call this whole process feature extraction. Training and testing are two parts of the machine learning stage. In the training phase, the machine converts its extracted information into knowledge. At the end of the training process the machine is tested. The machine completes the test process using information from its accumulated knowledge. And it gives a result on it.

3.4.1 Pre-Processing:

The first stage of the whole process is pre-processing. At this stage we have to go through via many methods to bring the data in the same configuration. Notable works among them is to bring all the data in the same format, keep the length of each of them just right. And place the noises correctly.

3.4.2 Feature extraction:

Feature extraction is the process of breaking down the input data and taking information from it. It is important and complex in the whole process. We will discuss below how our system has accomplished this complex task. We have extracted 20 features from an audio through mfcc as well as extracted 6 more features. All those features are: chroma stft, rmse, spectral centroid, spectral bandwidth, rolloff, zero crossing rate. Each of these carries different characteristics in an audio.

3.4.2.1 Mel-frequency cepstral coefficients (MFCC)

The first thing we need to know about the extraction method of mfcc is the raw material required for feature extraction. An audio file with a length of at least 25ms is a major component for mfcc feature extraction. On the other hand, our each and every data is at least 10 second long. As an example of our data's. As an example, some of our data's wave form is given below:

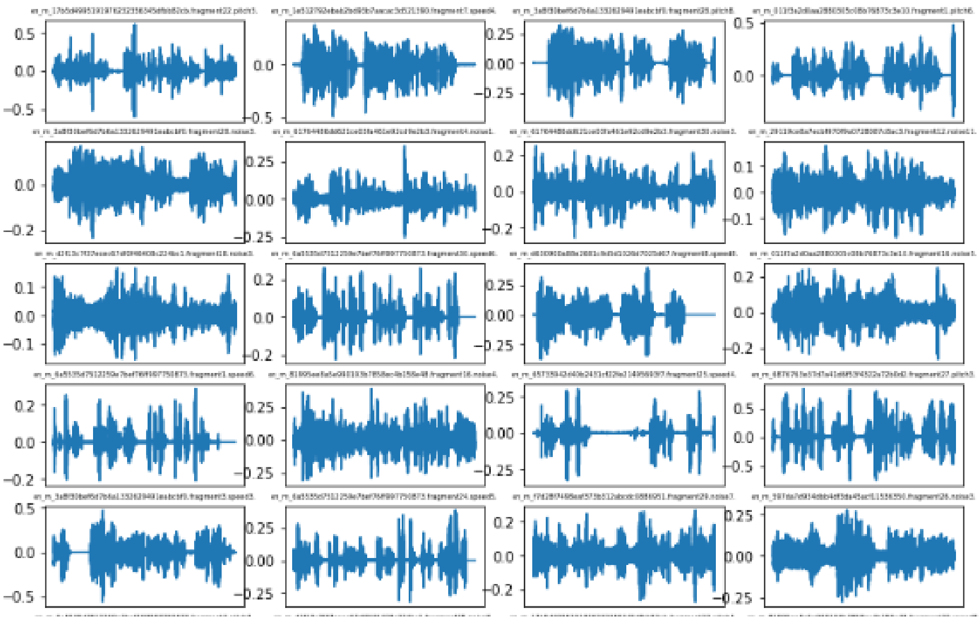


Figure 3.4.2.1: Wave form of our data

3.4.2.2 Analog to Digital conversation:

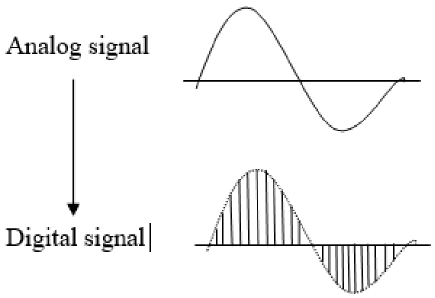


Figure 3.4.2.2: A/D conversion

In analog to digital conversation, the audio signal is converted to discrete space.

3.4.2.3 Pre-emphasis:

It basically increases the energy of high frequency. Alphabets like vowels have more energy in the lower frequency when it comes to pronunciation. This is called spectral tilt. More energy is available at that frequency through high frequency energy boosting. Phone detection accuracy will be higher by this method. Although the frequency of noise is also high. And when people are in the middle of high frequency, they also trouble to hearing. Pre-emphasis use filter to boost.

$$x'[t_d] = x[t_d] - \alpha x[t_d - 1] \qquad 0.95 < \alpha < 0.99$$

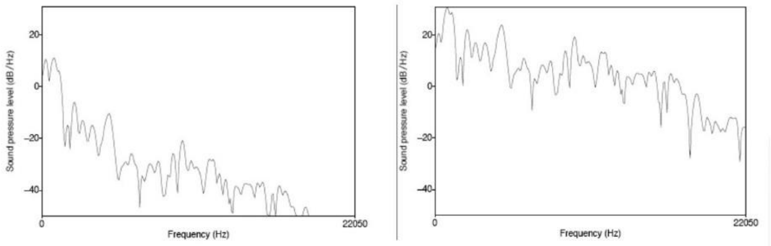


Figure 3.4.2.3: Boosting energy of high frequency

3.4.2.4 Windowing:

Windowing basically divides the waveform of an audio into tiny slides.

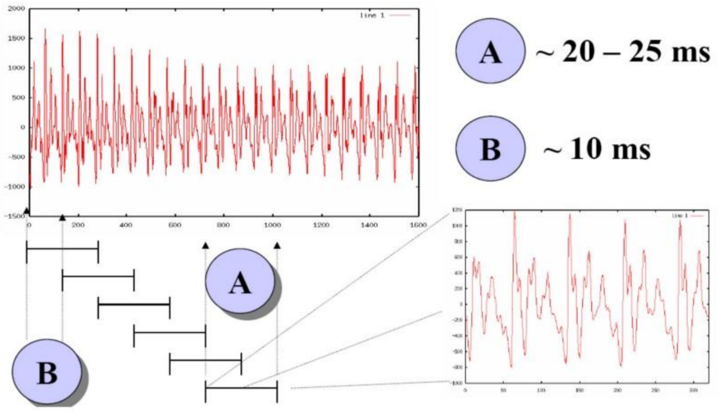


Figure 3.4.2.4: Example of Windowing

But I can't cut the audio as we want because it can cause a lot of noise in the audio. In this case we have to keep in mind, amplitude should drop near edge of frame.

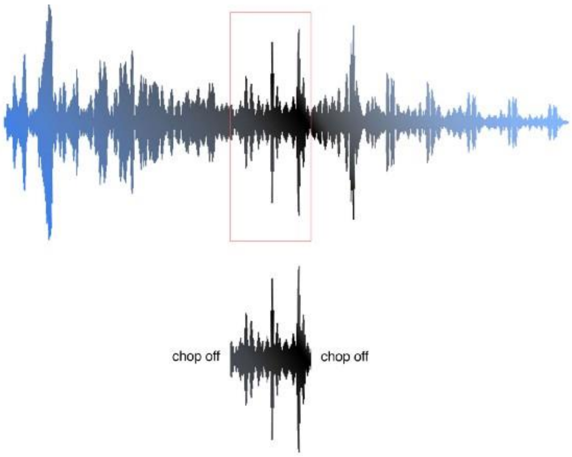


Figure 3.4.2.5: Slicing a audio

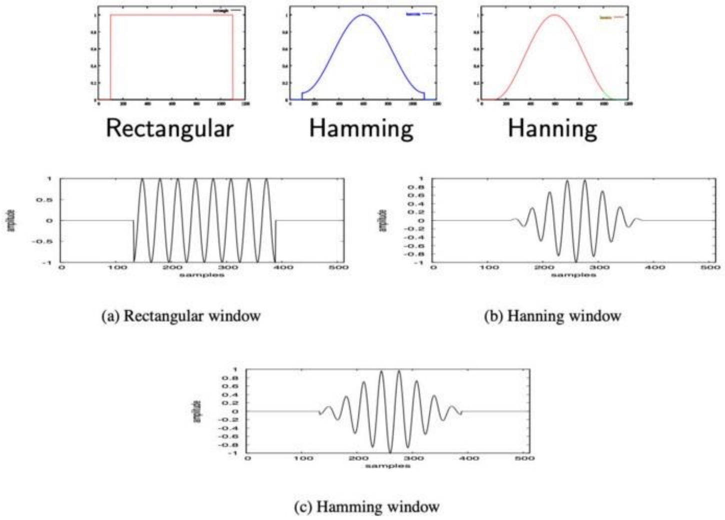
Suppose w is a part which is applied to the original audio clip as a time domain.

$$x[n] = w\ [n]\ s[n]$$

sliced audio

original audio clip

Hamming window and the hanning window are the alternatives of w window.



(a) Rectangular window

(b) Hanning window

(c) Hamming window

Figure 3.4.2.6: Hammig & Hanning Window

3.4.2.5 Discrete Fourier Transform (DFT):

After windowing DFT is applied in frequency domain

$$X[k] = \sum_{n=0}^{N-1} x[n] \exp\left(-j\frac{2\pi}{N}kn\right)$$

Figure 3.4.2.7: Discrete Fourier Transform

3.4.2.6 Mel filterbank:

The method of listening to an audio clip of the machine and the way human listen is totally different. In humans, it varies greatly in frequency. For higher frequencies human are less sensitive. In the diagram below we can understand how Mel scale work for measure frequency.

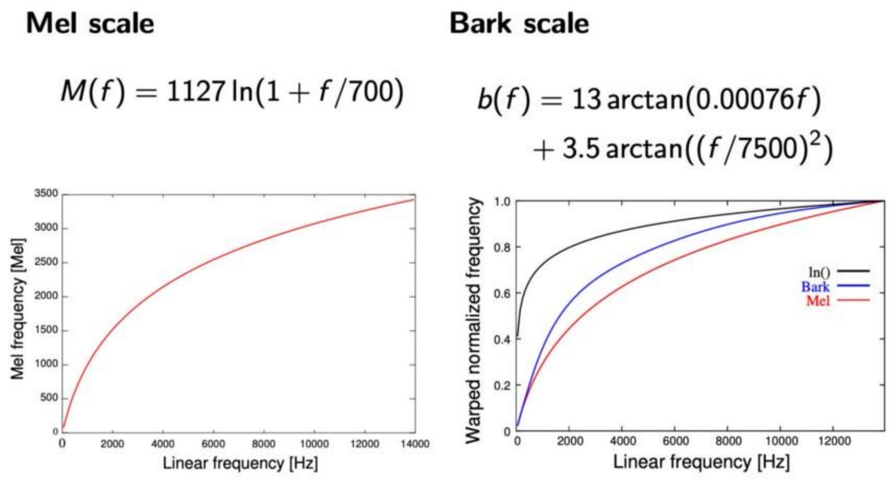


Figure 3.4.2.8: Frequency scale

3.4.2.7 Log:

Mel filterbank basically gives power spectrum as output. Our next task is to extract the log from the output of mel filterbank. This also reduces unnecessary pronunciation and frequency of words. At the end of this work, we need to do two more things. The first is to delete the f0 data or pitch and separate the extracted feature from the others.

3.4.2.8 Capstrum:

Our next task is to calculate the cepstral. Phone and pitch related information compose by log spectrum and peaks calculate formants of phones. But the most complex task is to separate them. Inverse Fourier Transformation is the key for separating task.

3.4.2.9 Dynamic features (delta):

Out of 39 features of mfcc, 12 features are known from the above-mentioned parts. The 13th feature is the energy of each frame.

3.4.2.10 Cepstral Mean and variance normalization:

After dynamic features part normalization of feature is important. Feature mean is dividing by its variance. And in this way normalization of feature value can be found. This help us to adjust values of different recording.

Each of the above sequential steps was to complete the mfcc perfectly. In addition to the features that came out of mfcc, we have extracted 6 more features from an audio. We will discuss them now

3.4.2.11 Root mean square error (rmse):

It is a feature of librosa. It's important to calculate root mean square for each frame.

3.4.2.12 Spectral centroid:

For characterize a spectrum, spectral centroid used in digital signal processing.

3.4.2.13 Spectral width:

Spectral width is the interval of wavelength.

3.4.2.14 Rolloff:

To find out harmonics of the waveform rolloff frequency used.

3.4.2.15 Zero crossing:

For changing mathematical sign (e.g. negative to positive or positive to negative) zero crossing is used.

3.4.2.16 Chroma stft:

For computing Chromagram from a wave chroma stft used normally. It's also a feature of librosa.

3.4.3 Machine Learning:

Of course, not all algorithms will perform the same, so we have worked with different algorithms. We have used 4 machine learning algorithms for classification. Those are:

3.4.3.1 Linear Regression:

Linear Regression is a supervised machine learning [21]. In linear regression there are two variables. One is explanatory variable and another is dependent variable. Where explanatory variable explains something and their value is independent. On the other hand, dependent variable is dependent on explanatory variable.

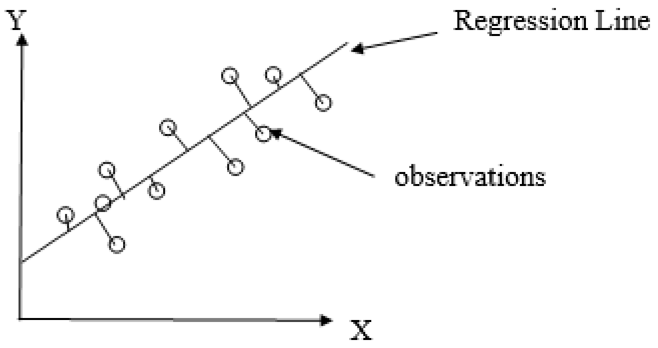


Figure 3.5.1: Linear Regression

In the above figure we can assume X for explanatory variable and Y for dependent variable. So, as we know there is a relationship between those two variables, the relationship can be positive or negative. If independent variable increases and dependent variable increases as well then, the relationship between them is called positive relationship. If Dependent variable decrease for increases of explanatory variable then the relationship is negative. Observations can be anywhere in the graph. It will depend on the data. Depending on the observations there will be a regression line. It will be based on least square method. After assuming the regression line, the difference between regression line and observation is called errors.

3.4.3.2 Decision tree:

The decision tree is usually based on the algorithm. He divides the data set based on different conditions. It is most commonly used for supervised learning. It is used for classification and regression. Decision tree is basically non parametric supervised learning.

3.4.3.3 Random Forest:

This algorithm also a supervised classification algorithm. And it is most popular and most powerful algorithm. It makes a forest with number of decision trees. If there is more tree then the result be more accurate in this algorithm. It also handles missing values. Accuracy will not affect for missing data. This algorithm will not overfit. And it works better with large data.

3.4.3.4 Gradient Boosting:

It's a machine learning to solve regression and classification. Gradient Boosting improves by combining shallow and weak by teaching them.

3.5 Implementation Requirements:

Below is a list of our most requires elements to complete our projects:

- Collect audio of specific languages (English, German & Spanish)
- Convert all audio to flac audio format
- Make sure all of them are 10 seconds long
- Adding different type of noise in audio file
- Install Anaconda
- Making an environment in our pc for our project
- Install librosa
- Using MFCC for extract audio
- Research on machine learning algorithm for language classification
- Find out best accuracy for our project

CHAPTER 4
EXPERIMENTAL RESULTS

4.1 Experimental results:

Our dataset was split into two part. One part is for train and another is for test. Our dataset has 30,000 data. We select 27,000 for train and we test with 3,000 audio speech. We use Sklearn Model selection for split our data. The classification result we got for English, German & Spanish languages are:

Table 4.1.1: Result of classification

Algorithm	Accuracy
Decision Tree	92%
Random Forest	98%
Linear Regression	21%
Gradient Boosting	84%

4.2 Discussion:

Only Linear Regression gave us less than 80% accuracy. Maximum accuracy we got from Random Forest algorithm that is 98%.

CHAPTER 5
CONCLUSION AND IMPLICATION FOR FUTURE RESEARCH

5.2 Conclusion:

The aim to our project was to identify best algorithm to classify language using machine learning. The main contribution of our paper is finding out different method to extract feature from speech and apply the best one in our project. Through this project we find out random forest gives us best accuracy between all other algorithm we use in our project. We also find out in this circumstances Liner Regression work poorly, and performed worse than random guessing.

5.2 Future Research:

LID system can be more accurate by increasing number of data for each language. More sample of speech can be added. If we can add more languages it will treat as an immediate improvement. Introducing incorporate incremental machine learning can be a great improvement for our system. Which classification system had done wrongly; system will learn accurately via user feedback mechanism.

ACRONYMS AND ABBREVIATION

LID – Language Identification

MFCC – Mel Frequency Cepstral Coefficient

LVASR – Large Vocabulary Automatic Speech Recognition

GMM – Gaussian Mixture Model

PRML – N-gram Language Modelling

PPR – Parallel Phone Recognition

LSTM – Long Short-Term Memory

ASR – Automatic Speech Recognition

SVM – Support Vector Machine

GDLS – Generalized linear discriminate sequence

SDC – Shifted-delta-cepstral

DLCSV – Discriminative Vector Space Classifier

BOS – Bag-of-sounds

RNN – Recurrent Neural Network

UBM – Universal Background Model

DFT – Discrete Fourier Transform

RMSE – Root Mean Square Error

REFERENCES

[1]. A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszcyina, "Multilinguality in speech and spoken language systems," *Proc. IEEE*, vol. 88, pp. 1181-1190Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszcyina, "Multilinguality in speech and spoken language systems," *Proc. IEEE*, vol. 88, pp. 1181-1190, Aug. 2000

[2]. P. Dai, U. Irugel, and G. Rigoll, "A novel feature combination approach for spoken document classification with support vector machines," *in Proc. Multimedia Information Retrieval Workshop*, pp 1-5, 2003

[3]. Haizhou Li, Bin Ma and Chin-Hui Lee, "A Vector Space Modeling Approach to Spoken Language Identification," *Proc. IEEE,* vol. 15, pp. 1-2, Jan. 2007

[4]. Google Play Store, available at <<https://play.google.com/store/apps/details?id=com.google.android.apps.translate>>, last accessed on 04-01-2020 at 3:12 PM.

[5]. K.M. Berkling, T. Arai and E. Barnard, "Analysis of phoneme-based features for language identification", *Proc. IEEE,* April 1994

[6]. J. Hieronymous and S. Kadambe, "Spoken Language Identification Using Large Vocabulary Speech Recognition", *proc. International Conference on Spoken Language Processing (ICSLP 96),* 1996

[7]. K. M. Berkling and E. Barnard, "Language Identification of Six Languages Based on a Common Set of Broad Phonemes" *Proc. 1994 International Conference on Spoken Language Processing,* September, 1994

[8]. Y. K. Muthusamy, "A Segmental Approach to Automatic Language Identification", Ph.D. thesis, Oregon Graduate Institute of Science & Technology, July1993

[9]. M. A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", *Proc. IEEE,* January, 1996

[10]. Chi-Yueh Lin, Hsiao-chuan Wang, "Language identification using pitch contour information", from Department of Electrical Engineering, National Tsing Hua University, Hisnchu, Taiwan

[11]. Fadi Biadsy, Julia Hirschberg, "Using prosody and Phonotactics in Arabic Dialect Identification", *Proc. 10th Annual Conference of the International Speech Communication Association,* Columbia University, New York, 2009

[12]. Julien Boussard, Andrew Deveau, Justin Pyron "Methods for Spoken Language Identification" December, 2017

[13]. Ruben Zazo, Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, Joazuin Gonzalez-Rodriguez "Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks" January, 2016

[14]. Rong Tong, Bin Ma, Donglai Zhu, Haizhou li and Eng Sking Chang "Integrating acoustic, prosodic and phonotactic features for spoken language identification" *Proc. IEEE*, pp. 207, May,2006

[15]. Adarsh D. Patil, Akshay Vishwas Johi, Harsha.K.C, Pramod.N "Spoken language identification using machine learning", *Visvesvaraya Technological University, Belgaum*, pp. 26, May,2012

[16]. Dan Robinson, Kevin Leung, Xavier Falco, "Spoken language identification with hierarchical temporal memories" pp. 2-3, 2009

[17]. YouTube, available at << https://www.youtube.com/watch?v=dUmSHIduo3c>> last accessed on 01-02-2020 at 9:49 PM.

[18]. Medium, available at << https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9>> last accessed on 01-02-2020 at 9:53 PM.

[19]. Audio Converter, available at << https://online-audio-converter.com>> last accessed on 15-12-2019 at 1.10 AM.

[20]. Natural readers, available at << https://www.naturalreaders.com/online>> last accessed on 13-12-2019 at 10.22 PM.

[21]. Geeks for Geeks, available at << https://www.geeksforgeeks.org/ml-linear-regression/>> last accessed on 20-02-2020 at 11:24 AM.

[22]. Bin MA & Haizhou LI, "Spoken Language Identification Using Bag-Of-Sounds"

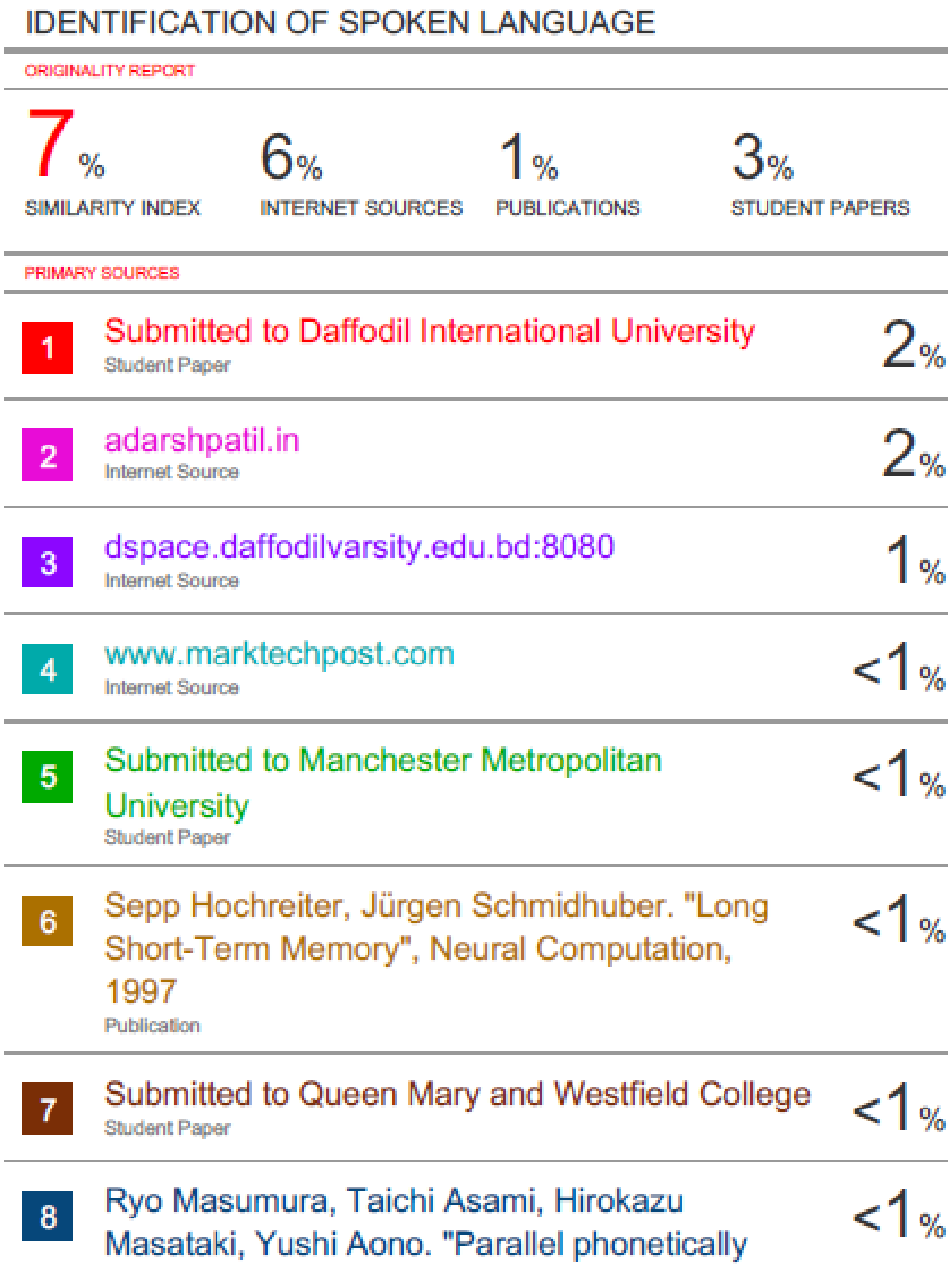
[23]. Ming Li, Hongbin Suo, Xiao Wu, Ping Lu, Yonghong Yan, "Spoken Language Identification Using Score Vector Modeling and Support" *proc. 8th annual conference of the international speech communication association*, 2007

[24]. R.A Cole and Y.K Muthusamy. "The OGI Multilanguage Telephone Speech Corpus". Proceedings International Conference on Spoken Language Identification, vol. 2 pp. 895899 Oct. 1992.

[25]. Ming Li, Hongbin Suo, Xiao Wu, Ping Lu, Yonghong Yan "Spoken Language Identification Using Score Vector Modeling and Support Vector Machine" *proc. 8th annual conference of the international speech communication association*, pp. 351, 2007

[26]. Ruben Zazo, Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, Joazuin Gonzalez-Rodriguez "Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks" pp. 5, January, 2016

PLAGIARISM REPORT



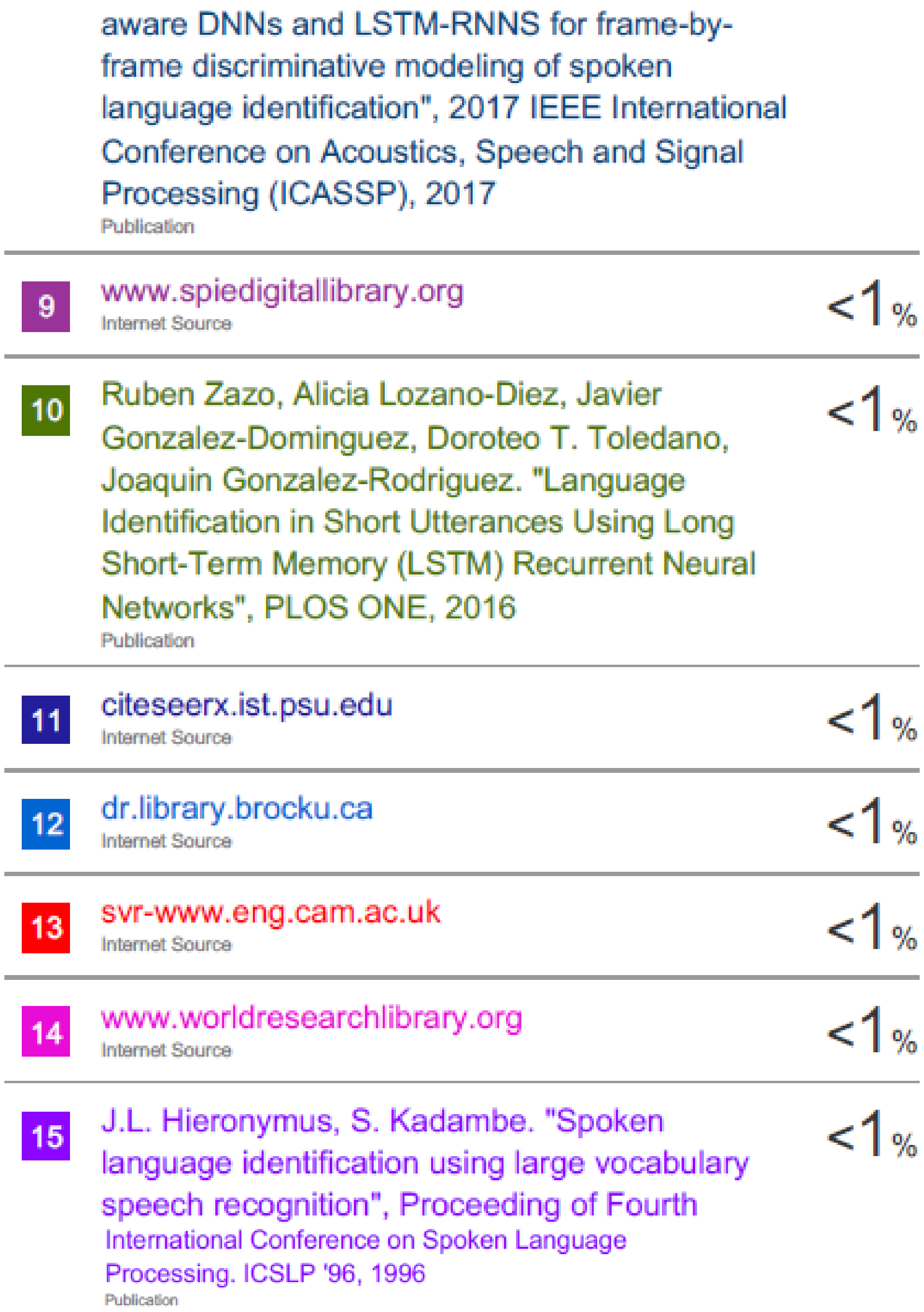
IDENTIFICATION OF SPOKEN LANGUAGE

ORIGINALITY REPORT

SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS
7%	6%	1%	3%

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	adarshpatil.in Internet Source	2%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	www.marktechpost.com Internet Source	<1%
5	Submitted to Manchester Metropolitan University Student Paper	<1%
6	Sepp Hochreiter, Jürgen Schmidhuber. "Long Short-Term Memory", Neural Computation, 1997 Publication	<1%
7	Submitted to Queen Mary and Westfield College Student Paper	<1%
8	Ryo Masumura, Taichi Asami, Hirokazu Masataki, Yushi Aono. "Parallel phonetically	<1%



	aware DNNs and LSTM-RNNS for frame-by-frame discriminative modeling of spoken language identification", 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017 Publication	
9	www.spiedigitallibrary.org Internet Source	<1%
10	Ruben Zazo, Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Doroteo T. Toledano, Joaquin Gonzalez-Rodriguez. "Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks", PLOS ONE, 2016 Publication	<1%
11	citeseerx.ist.psu.edu Internet Source	<1%
12	dr.library.brocku.ca Internet Source	<1%
13	svr-www.eng.cam.ac.uk Internet Source	<1%
14	www.worldresearchlibrary.org Internet Source	<1%
15	J.L. Hieronymus, S. Kadambe. "Spoken language identification using large vocabulary speech recognition", Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, 1996 Publication	<1%