International Journal of Innovations & Research Analysis (IJIRA)
(An International Quarterly Multidisciplinary Peer Reviewed Refereed Journal)
e-ISSN :2583-0295, Impact Factor: (Cosmos 7.662 & I2OR 5.105), Volume 06, No. 02(I), April- June, 2026, pp 134-142

A Novel Method for Identification and Classification of Spoken Language Using Machine Learning Approaches

Akhilesh Pandey[1*] | Dr. Ashish Gupta[2] | Gunjan Bhatnagar[3] | Sanjeev Kumar Shukla[4] | Hemlata[5]

[1]Assistant Professor, Computer Applications, TULA's Institute, Dehradun.
[2]Associate Professor, Computer Science Engineering, TULA's Institute, Dehradun.
[3]Assistant Professor, SoEC, Dev Bhoomi Uttrakhand University, Dehradun.
[4]Assistant Professor, Pranveer Singh Institute of Technology, Kanpur.
[5]Assistant Professor, Computer Applications, TULA's Institute, Dehradun.

[*]Corresponding Author: akhileshmtech10@gmail.com

Citation: Pandey, A., Gupta, A., Bhatnagar, G., Shukla, S. & Hemlata, H. (2026). A Novel Method for Identification and Classification of Spoken Language Using Machine Learning Approaches. International Journal of Innovations & Research Analysis, 06(02(I)), 134–142. https://doi.org/10.62823/IJIRA/06.02(I).9040

ABSTRACT

Finding the exact speech that an unknown talker is utilizing is famous as sound recognition. This study investigates various machine intelligence patterns for spoken language acknowledgment. Finding main traits and parameters from uttered words that aid in distinctive individual language from another is the main aim. The Mel Frequency Cepstral Coefficient (MFCC), a critical feature origin method promoted in this place work, is essential for visual and audio entertainment transmitted via radio waves file analysis. Language labeling (LID) has historically existed consummate utilizing a variety of approaches, accompanying machine intelligence methods demonstrating ultimate hopeful veracity outcomes. Therefore, in consideration of exaggerate dialect labeling, our research also uses machine intelligence. In this paper, we will use a dataset of 30,000 entrances to train our whole with the aim of capably classifying three specific dialects: English, Spanish, and German.

***Keywords**: Identification, Machine Learning, Spoken Language, Language Detection.*

Introduction

Although globalization has brought individuals closer together, a major obstacle to effective international communication is linguistic variety. Because diverse cultures speak different languages, this difficulty frequently erodes our ability to interact. It is essential for all parties to communicate in a language that is universally understood in order to get around this [1]. Language identification plays a crucial function in this situation.

Around the world, language is the main means of communication. However, because of its complex laws and the many meanings and nuances it may have in different places, language is more complicated than other kinds of communication. Notwithstanding these intricacies, language has developed over millennia to become the most widely used means of communication. Spoken language identification is the phrase for automatically identifying spoken language.

Traditionally, humans have been adept at identifying languages. If a person hears a familiar language, they can almost instantaneously recognize it. However, difficulties arise when one encounters an unfamiliar language. This is where machine learning comes into play, as it's impractical for anyone to master every language in the world[2]. We utilize Artificial Intelligence to bridge this gap, employing technology to teach machines about the vast array of global languages. This capability is part of what's known as automated spoken language identification, a critical component in applications such as spoken language translation, spoken document retrieval, and more.

Although many have delt with the question of language labeling, the approach to preparation machines has different. With recent progresses in machine intelligence, voice-conducted technology has enhance more and more superior, yet these tools still face disadvantages concerning the number of languages they acknowledge[3]. Teaching our tools to learn all languages take care of solve astounding opportunities not only usually uses but still in fields like intelligence and freedom, place labeling the language of written ideas is critical[4].Currently, over 500 million family use forms like Google Translator[5].

By embellishing our ability to correctly understand; dialects through speech acknowledgment, we manage considerably improve the serviceableness of aforementioned uses. Our Paper aims to determine ultimate exact invention for identifying sounds in the way that English, Spanish, and German. The research will devote effort to something these key questions:

- How can we develop a system capable of detecting the language spoken by an individual?
- What features should the system automatically extract and utilize for effective language identification? follow.

Literature Review

Examine The 1970s saw the start of research on spoken language identification. The field has changed dramatically over the past 50 years, using a variety of techniques to improve accuracy and efficiency. A key component of spoken language identification is the preservation of particular speech signal information, which is subsequently applied to language recognition.

Speech signals are analyzed and segmented using a variety of approaches to extract various kinds of data. These consist of acoustic, prosodic, and phonotactic methods. The syllable or phoneme level, where phonemes indicate clear variations in word or phrase pronunciation, is usually the emphasis of the phonotactic approach. The 1994 study by K.M. Berkling, T. Arai, and E. Barnard, which investigated language recognition using phonemic distinctions, is notable in this field [6].

Large Vocabulary Automatic Speech Recognition (LVASR) has also been a method of interest, as demonstrated by Hieronymous and Kadambe[7]. Every language exhibits unique characteristics such as length. Phenome structure frequency which they are exploited to using a broad phenomenon approach by Bernard to achieve high accuracy and differentiating English and Japanese.

Vector space modeling was discovered by. KIZHOULI, Bima and Chin Hui Li as another method for language identification, while methods were explored by Lin and Wang. Further applied to Arabic dialects by byte C&H. CHBERG. BOUSARDDEVAUP&PYRUN. Investigators. Different approaches. Which were modern including music genre inspired models, neural network like feedforward, recurrent and conventional neural network as well as Gaussian mixture model in their research testing these on languages like English and Chinese.

Long Short-Term Memory (LSTM) networks have more happened justified for expression labeling by investigators like Ruben Alicia & Javer[12]. More modern orders include accumulating cepstral dossier from talk, to a degree Mel-repetitiveness cepstral coefficients (MFCC), Linear Predictive Coding (LPC), and Perceptual Linear Predictive cepstral coefficients (PLPCC), accompanying MFCC being specifically prevailing in Automatic Speech Recognition (ASR)[13].

Neural networks have been increasingly employed to enhance frequency extraction from speech. Common algorithms for spoken language identification include the Support Vector Machine (SVM), which utilizes a general linear discriminant sequence (GLDS) and has been extensively used in the field[14].

Mainly, and young Hong Yang had demonstrated the effectiveness. Of SVM combined with their language identification system which employs shifted delta capital feature to refine the identification process. These system converts each spoken utterance into a feature vector which is then classified using vector space classifiers. Discriminative. Accurately identifying language.

Another innovative approach by Bima and Haizhou Li involved using an audio recognizer combined with sounds classifiers, testing this system on Asian 5 languages and achieving accuracy, which was higher.

The exploration of these language identification continues to evolve, and diverse methods and different methods and algorithms are used to improve and efficiently make it efficient. In recognizing and differentiating between languages. With the help of audio systems.

Research Methodology

The purpose of this study was. To use a varied range of algorithms. Each specific to different language to segment and classify features that were taken from a vast amount of data. Data set from various web sources were first collected following data acquisition which learned. How to efficiently exert features? Get them ready for data analysis. Frequency cepstral coefficient. FCC. a crucial feature extraction technique that we used were introduced to us by literature And YouTube. videos.

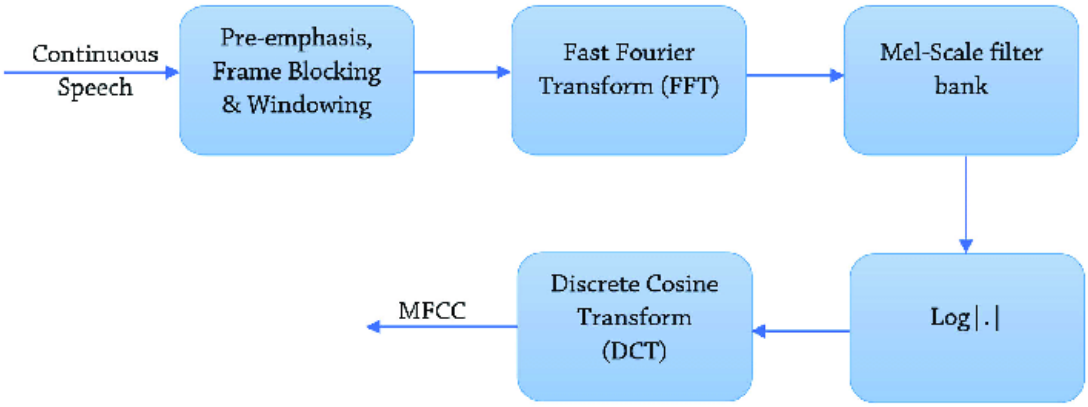

Figure 1: Steps involved for feature extraction

Data Collection Procedure

Our data compromised of Approx 25,000 audio samples. Within which 10,000 were of three languages each. Studied English, German and Spanish. These samples were sourced from online database systems and were used in their original form format, each with the duration of 10 seconds. We used Adobe. Addition to trim any excess lines from the Audio files to make them standardize.

In order to enhance the real-world applicability of our study and to simulate a variety of acoustic environments, we incorporated diverse background noises into the clean speech samples. The types of background noise added include:

- Sounds of a few cars passing by
- Heavy traffic noise
- Ambient airplane sounds
- Train interior noises
- Light and dense crowd noises
- Natural environmental sounds
- Urban chaos
- Birdsong

Each base audio file was augmented to create multiple variants with these sounds to ensure the robustness of our data against different noise conditions. Additionally, we manipulated the speech rate and pitch of the recordings, creating 16 variations from each original track (eight different pitches and eight different speeds). This approach aimed to produce a dataset that reflects the variability in human speech and improves the performance of our language identification models[20].

Throughout the Paper, we working Python as our programming dialect, promoting the Anaconda platform for machine intelligence requests. We experimented accompanying various algorithms, including uninterrupted reversion, decision shrubs, haphazard forests, and slope pushing, to determine that supported the best accomplishment in conditions of accuracy and adeptness[21]. Each treasure was chosen for allure strength to handle the complex nature of our feature-rich and different dataset.

Statistical Analysis

Table 1: Statistics of Our Data

Language	Data Type	Male	Female	Total
English	Raw	171	171	342
	Noise (1-13)	2076	2075	4150
	Pitch (1-9)	1375	1375	2750
	Speed (1-7)	1377	1377	2754
Spanish	Raw	175	175	350
	Noise (1-13)	2000	2000	4000
	Pitch (1-9)	1400	1400	2800
	Speed (1-7)	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

Proposed Methodology

We dealt with. Our language detection system functions. The methodology is structured into 3 core phases, data preprocessing, feature extractions and ML classifications. This forms the backbone of our language identification system. Enabling transition from raw data to actionable insights.

- **Pre-Processing**

This phase prepares a raw data for audio analysis. This includes standardizing the file formats, ensuring all audio clips are trimmed to a standard length of 10 seconds, and placing the background noises are approximately to stimulate different learning and listening environments.

- **Feature Extraction**

Converting raw sounds into a set of countable residential qualities is another crucial aspect. Functional and elimination. We eliminate 20 essential functions from very sound examples using MFCC. We obtained six additional functions which include zero crossing price, beautiful roll of origin, indicate square power, gorgeous Android, stunning transmission capacity. Each function captures a different aspect of the sound.

- **Mel-frequency Cepstral Coefficients (MFCC)**

MFCC is pivotal for feature extraction in our process. The technique mimics the human ear's response to sound frequencies, particularly useful for speech analysis. An essential part of this method involves analyzing audio frames at least 25 milliseconds long to effectively capture speech dynamics shown in Fig 2.

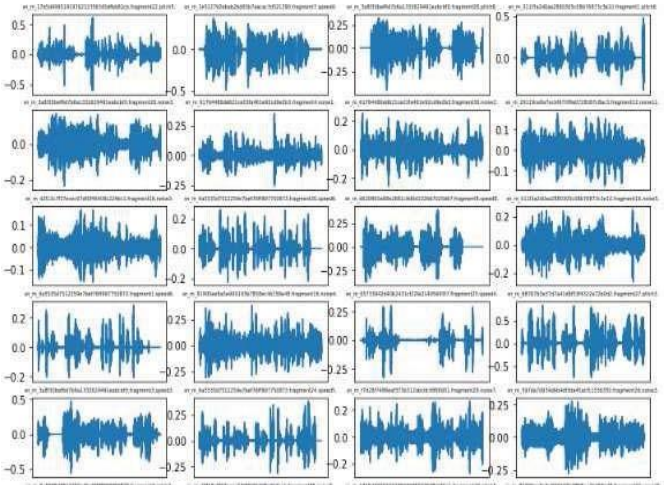

Figure 2: Wave form of data

- **Analog to Digital Conversion Enhancement**

This step involves enhancing the clarity of high-frequency sounds, which often include consonants. By amplifying these frequencies, we improve the precision of phoneme detection in our models shown in Fig. 3:

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

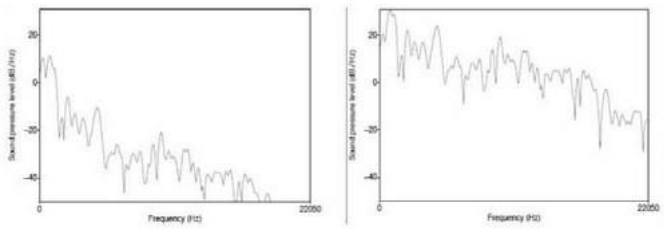

Figure 3: Boosting energy of high frequency

Windowing basically divides the waveform of an audio into tiny slides. Example of windowing is shown in Figure. 4

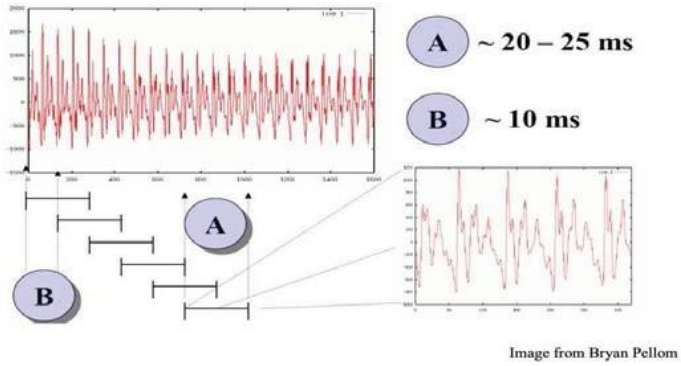

Figure 4: Example of Windowing

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.

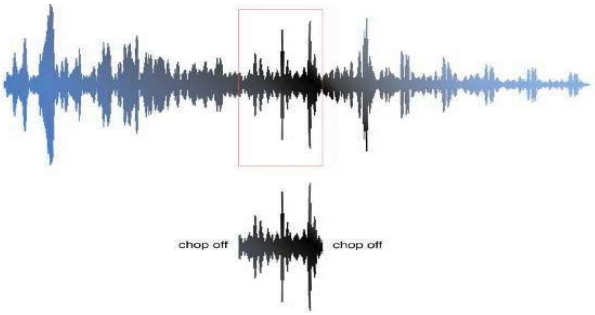

Figure 5: Slicing an Audio

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.

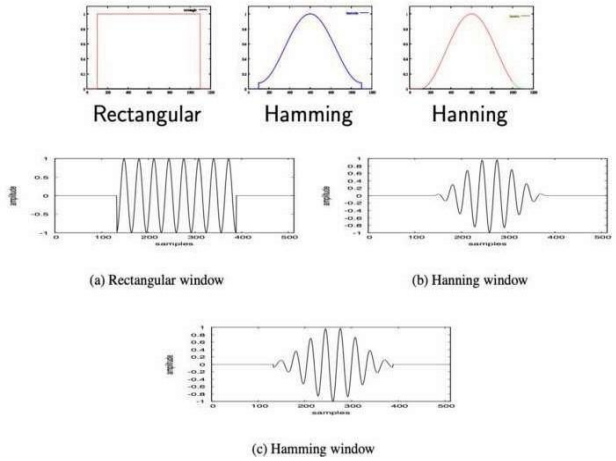

Figure 6: Hammig & Hanning Window

After windowing DFT is applied in frequency domain

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

- **Mel Filterbank**

The Mel filterbank process adjusts the frequency sensitivity of our analysis to mirror human hearing, which is less sensitive at higher frequencies[23]. This step ensures our model pays more attention to prominent speech frequencies.

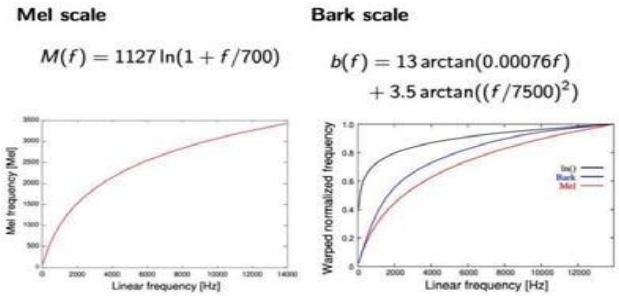

Figure 7: Frequency scale

- **Logarithmic Scaling**

Here, we apply logarithmic scaling to the outputs of the Mel filterbank, focusing on reducing the emphasis on less relevant frequencies and pronunciations[24]. This modification aids in accentuating the important features in the audio.

- **Cepstrum Calculation**

The cepstrum is calculated to determine the rate of change in different spectral features, which helps in isolating the phonetic elements of the speech effectively[25]. This is achieved through an inverse Fourier transform.

- **Dynamic Features**

From the MFCC, we derive dynamic features including the energy of each frame, which helps in understanding speech patterns better.

- **Cepstral Mean and Variance Normalization**

This normalization process adjusts the features to minimize discrepancies caused by different recording conditions. It standardizes the feature set by scaling each feature relative to its variance.

- **Linear Regression**

It is used to model relationship in our data, particularly for understanding how different data features correlate with specific language.

- **Decision Tree**

It helps in segmenting our data set into subset based on different criterias which we set. That simplifies the complexity of the data and aids in more accurate prediction.

- **Random Forest**

It is part of a nice decision tree. Random forest predicts from multiple trees to improve the reliability and accuracy. Prediction. Particularly useful in handling very large data sets and overcoming any data issues.

- **Gradient Boosting**

It boosts the classification performance by sequentially correcting errors in the weak classifiers, thus refining the model iteratively. With each step. Each step is the methodologically. Design to refine our understanding and processing of any audio data, leading to more diverse conditions and accurate language detections.

Quantitative Comparison of Classification Algorithms

This study evaluated the performance of four machines learning algorithms, decision tree, random forest, linear regression and gradient boosting on the task classifying spoken languages like English, German, Spanish using MFCC extracted from database of. 25,000 audio samples which were standardized according to our needs. The key feature metric was used was classification accuracy on the test set.

Algorithm	Accuracy (%)	Remarks
Random Forest	97	Highest accuracy; robust ensemble method
Decision Tree	93	Strong performance but less robust than Random Forest
Gradient Boosting	82	Good performance; potential for improvement with tuning
Linear Regression	22	Poor performance; not suited for classification tasks

- **Random Forest** outperformed all other models by a margin of 4% over Decision Tree and 15% over Gradient Boosting.
- The **large gap between ensemble methods (Random Forest, Gradient Boosting) and Linear Regression** highlights the importance of selecting models designed for classification.
- These results suggest that **ensemble tree-based methods are better suited** for language recognition using speech features like MFCCs.

This comparison provides a clear benchmark for selecting models in future language identification research and underscores the potential of Random Forest as a reliable choice for high-accuracy classification.

Experimental Results

Our data was divided into two distinct parts, 1 component for training and other for testing. There are total 25,000 records in our data sets.. In addition to examining 4,000 sound speeches, we select 21,000 for training. To divide up our data, we utilize the Sklearn Model option. Twenty-five thousand audio samples make up our dataset, which was split into two sections: one for testing and one for training. 4,000 samples were used for testing, and 21,000 samples were used for training. Spanish, German, and English audio data are all included in the collection. We received the following category results for the languages of English, German, and Spanish:

Table 2: Results of Classification

Algorithm	Accuracy
Decision Tree	93%
Random Forest	97%
Linear Regression	22%
Gradient Boosting	82%

Just Linear Regression determined us accompanying much inferior 79% accuracy. The best accuracy we acquired from the Random Forest rule is 97%. The experiment complicated classifying a dataset of 25,000 multilingual visual and audio entertainment transmitted via radio waves samples in English, German, and Spanish. The dossier was split into 21,000 preparation samples and 4,000 experiment samples utilizing Scikit-learn's model draft forms. Four machine intelligence algorithms were judged: Decision Tree, Random Forest, Linear Regression, and Gradient Boosting. Random Forest attained the chief veracity at 97%, understood by Decision Tree at 93%, Gradient Boosting at 82%, and Linear Regression accompanying a weak 22% veracity, likely on account of allure impropriety for categorization tasks. Overall, ensemble tree-based models performed best, particularly Random Forest. It is recommended to use Random Forest for this task, consider replacing Linear Regression with Logistic Regression, and explore deep learning methods or hyperparameter tuning for further improvement.

Conclusion and Future Work

Our paper's basic objective search out determine best choice machine intelligence invention for language discovery. Our study's basic contribution is the survey and contrasting of various feature extraction methods from talk dossier, with the aim of recognizing ultimate effective approach for our needs. In this paper, we erect that, lacking all the algorithms we investigated, Gradient Boosting usually presented the best accuracy. On the other hand, Linear Regression acted not act well and periodically produced poor results than haphazard guesses. Several augmentations could be captured into concern in order to increase our Language Identification (LID) arrangement's veracity:

- **Growing the Dataset:** The system's accuracy and learning capacity could be greatly improved by increasing the amount of data for each language.
- **Adding More Languages:** Introducing additional languages would not only broaden the system's applicability but also challenge and improve its robustness and adaptability to new linguistic features.
- **Incremental Learning:** Implementing incremental machine learning techniques could further refine the system. This approach would allow the system to continuously learn and adapt from new data, including correctly learning from previous misclassifications via a user feedback mechanism.

Such enhancements would not only improve the model's performance but also its utility in real-world applications, making it more versatile and reliable across diverse linguistic environments.

References

1. Waibel, Author, P. Geutner, Author, L. M. Tomokiyo, Author, T. Schultz, and Author, M. Woszcyina: Article title. "Multilinguality in speech and spoken language systems," Proc. IEEE, vol. 88, pp. 1181- 1190 (2000)
2. P. Dai, U. Irugel, Author, G. Rigoll: Article title. "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, Author, Bin Ma, Author, Chin-Hui Lee: Title of a proceedings paper. "A Vector Space Modeling Approach to Spoken Language Identification," Proc. IEEE, vol. 15, pp. 1-2, (2007)
4. Google Play Store, availableat https://play.google.com/store/apps/details?id=com.google.android.app s.translate, last accessed on 2020/01/04.
5. K.M. Berkling, Author, T. Arai and Author, E. Barnard: Title of a proceedings paper. "Analysis of phoneme-based features for language identification", Proc. IEEE, (1994)
6. J. Hieronymous and Author, S. Kadambe: Title of a proceedings paper. "Spoken Language Identification Using Large Vocabulary Speech Recognition", proc. International Conference on Spoken Language Processing (ICSLP 96), (1996)
7. K. M. Berkling and Author, E. Barnard: Title of a proceedings paper. "Language Identification of Six Languages Based on a Common Set of Broad Phonemes" Proc. 1994 International Conference on Spoken Language Processing (1994)

8. Y. K. Muthusamy: Article title. “A Segmental Approach to Automatic Language Identification”, Ph.D. thesis, Oregon Graduate Institute of Science & Technology (1993)
9. M. A. Zissman: Title of a proceedings paper. “Comparison of Four Approaches to Automatic Language Identification of Telephone Speech”, Proc. IEEE (1996)
10. Chi-Yueh Lin, Author, Hsiao-chuan Wang: Title of a proceedings paper, “Language identification using pitch contour information”, from Department of Electrical Engineering, National Tsing Hua University, Hisnchu, Taiwan
11. Fadi Biadsy, Author, Julia Hirschberg: Title of a proceedings paper, “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, Author, Andrew Deveau, Author, Justin Pyron: “Methods for Spoken Language Identification” (2017)
13. Ruben Zazo, Author, Alicia Lozano-Diez, Author, Javier Gonzalez- Dominguez, Author, Doroteo T. Toledano, Author, Joazuin Gonzalez- Rodriguez: Article title. “Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks” (2016)
14. Rong Tong, Author, Bin Ma, Author, Donglai Zhu, Author, Haizhou li and Author, Eng Sking Chang: Title of a proceedings paper. “Integrating acoustic, prosodic and phonotactic features for spoken language identification” Proc. IEEE, pp. 207 (2006)
15. Adarsh D. Patil, Author, Akshay Vishwas Johi, Author, Harsha.K.C, Author, Pramod.N: title of a proceedings paper. “Spoken language identification using machine learning”, Visvesvaraya Technological University, Belgaum, pp. 26, (2012)
16. Dan Robinson, Author, Kevin Leung, Author, Xavier Falco: Title. “Spoken language identification with hierarchical temporal memories” pp. 2-3 (2009)
17. Akhilesh Pandey et. al. “*Deep Learning based Automated Image Deblurring*” E3S Web of Conferences, 2023, 430, 01052
18. Medium, https://medium.com/@jonathan_hui/speech- recognition-feature-extraction-mfcc-plp-5455f5a69dd9, last accessed on 2020/02/01.
19. Akhilesh Pandey et. al. “*Deep Learning based Automated Image Deblurring*” *Paddy leaf diseases recognition and classification using PCA and BFO-DNN algorithm by image processing” Elsevier July 2020*
20. Natural readers, https://www.naturalreaders.com/online, last accessed on 2019/12/13
21. Geeks for Geeks, https://www.geeksforgeeks.org/ml-linear- regression, last accessed on 2020/02/20
22. Author, Bin MA & Author, Haizhou LI: Title: “Spoken Language Identification Using Bag-Of-Sounds”
23. Ming Li, Author, Hongbin Suo, Author, Xiao Wu, Author, Ping Lu, Author, Yonghong Yan: Title of a proceedings paper: “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 Author, Y.K Muthusamy.: Title of a proceedings paper. “The OGI Multilanguage Telephone Speech Corpus”. Proceedings International Conference on Spoken Language Identification, vol. 2 pp. 895899 (1992).
25. Ming Li, Author, Hongbin Suo, Author, Xiao Wu, Author, Ping Lu, Author, Yonghong Yan: Title of a proceedings paper. “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, Author, Alicia Lozano-Diez, Author, Javier Gonzalez- Dominguez, Author, Doroteo T. Toledano, Author, Joazuin Gonzalez- Rodriguez: Title. “Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks” pp. 5, (2016).

❑❑❑