e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

A REVIEW ON SPOKEN LANGUAGE IDENTIFICATION

Aiswarya PA*1, Ms. Arya KA*2

*1Department Of Computer Science Vimala College (Autonomous) Thrissur, Kerala, India.

*2Assistant Professor, Department Of Computer Science Vimala College (Autonomous) Thrissur, Kerala, India.

DOI : https://www.doi.org/10.56726/IRJMETS45611

ABSTRACT

Speech along with other utterances that carry meaning are referred to as spoken language. Most people communicate in at least one native language that they learned as children and additional languages that they have learned throughout the course of their lifetime. Different languages have different rules for how consonant and vowel sounds can be combined into syllables in words. The vocal and auditory abilities of humans allow for a very wide variety of articulations and subsequent sounds, all of which are used in spoken language. Each spoken language has a slightly varied range, which contributes to the challenge of learning a foreign language and expressing it "with an accent". Spoken language identification (SLID) is the method of identifying language from an audio clip by an unknown speaker, regardless of gender, speaking style, or age. In this study, numerous spoken language identification models have been developed by utilizing various deep learning techniques, machine learning techniques, datasets, and performance metrics. Data gathering, feature elimination, and language classification are the three core components of the SLID system.

Keywords: Spoken Language Identification(SLID), CNN, CRNN, LSTM, HMM, Language Identification, Hybrid Feature Extraction Techniques, Mel-Frequency Cepstral Coefficient, Arabic Digits, Na¨Ive Bayes, SVM, Word Embedding, Log-Mel Spectrograms, Relu, RNN, Automatic Speech Recognition System, Amazigh Digits, Back Propagation Neural Network.

I. INTRODUCTION

In spoken language identification, an audio sample is used to identify the language being spoken by an anonymous speaker. Over 5,000 languages are spoken worldwide, and each has unique characteristics ranging from acoustics to meanings. Western nations have advanced significantly in their use of programs that use spoken language recognition. It can be difficult to distinguish spoken language from that of other age groups, different genders, and different accents. As such, voice utterances cannot be properly processed, and grammatical rules cannot be used in the absence of automatic language detection. The most widely used methods for identifying spoken audio languages and dialects include acoustic data, phonotactic and prosodic approaches, and other methods. A variety of multilingual speech processing applications, including spoken language translation, multilingual voice recognition, and spoken document retrieval, have made extensive use of S-LID, an enabling technology[1]. S-LID serves as the first stage in speech-based assistants, such as Apple Siri and Amazon Alexa, in which the corresponding grammar [1] is selected from a list of possible languages in order to analyze the languages' semantics further. S-LID has always been a difficult subject because of the variety in speech input types and the lack of understanding of how people understand and interpret speech under various circumstances. This elevates it to a crucial research area in the field [1] of voice signal processing.

II. RELATED WORKS

Referring to the research done by a number of authors who examined various techniques for spoken language identification.

A. Spoken Language Identification Using Deep Learning [2021]

The issues with the acoustic phonetics technique for automatic language identification have been addressed in various ways [2]. Researchers are now able to apply GANs for language identification for robustness on unsupervised and semisupervised tasks because of encouragement in the deep learning sector. Short utterances [3] do not function well for Support Vector Machines (SVM) classifiers, which results in lower

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

accuracy. Speaking language processing activities are provided by ineffective conventional identification methods on i-vector systems. The spectrograms of audio snippets, which can record or save the frequency of certain auditory utterances, are created using a log-Mel spectrum to tackle the issues outlined above.

The convolutional neural network (CNN) technology can be used to identify several languages, and it is effective and quick. Deep learning and the spectrum technique were used to complete it. The language [3] is quite difficult to distinguish in several of the audio clips because of the background noise. It is suggested to extract the properties using a deep learning CNN technique. The proposed framework for identifying spoken languages is shown in phases in Figure 1. Using audio recordings, this model creates spectrogram visuals from the audio files. [3] It utilizes a convolutional neural network (CNN) to highlight key characteristics or traits that make output easy to detect. [4] One of the main goals is to identify languages other than English, French, Spanish, and German, such as Estonian, Tamil, Mandarin, Turkish, Chinese, Arabic, Hindi, Indonesian, Portuguese, Japanese, Latin, Dutch, Portuguese, Pushto, Romanian, Korean, Russian, Swedish, Tamil, Thai, and Urdu.

[3]The proposed deep learning-based framework for spoken language identification transforms audio utterances into spectrograms based on their frequency and timing. After that, characteristics from photos are extracted using a convolutional neural network (CNN) in order to classify them [4]. In the end, multilingual categorization is accomplished using the softmax activation function. Data augmentation is performed at the preprocessing stage to address issues with class imbalance. The amount of data can be increased with the use of data augmentation by adding adjustments to already existing data, such as crop, rotate, flip, shearing, and many other effects. [4] The dataset is split into two directories: train, which comprises (73080) samples, and test, which contains (540) samples, with three languages—English, German, and Spanish.

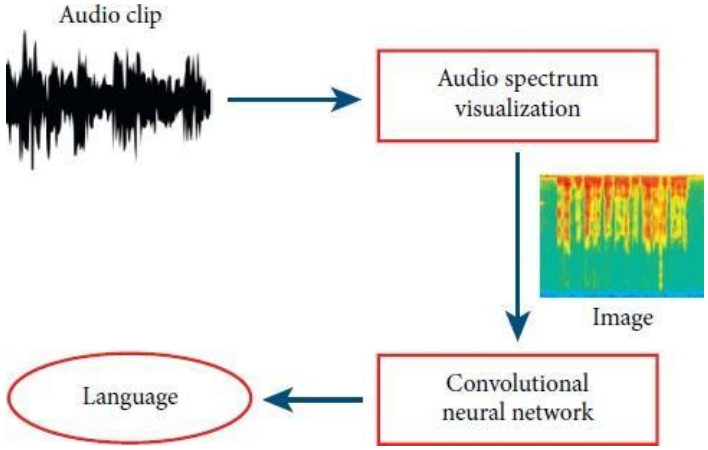

Fig. 1. Phase of spoken language identification using spectrograms

The term ”spectrogram” refers to an image that displays the frequencies that are present [4] on a signal over time. Using the Fast Fourier Transform (FFT), a time series signal of data points can be used to determine the signal’s frequency. Time-series data can be transformed using the Fast Fourier Transform (FFT) to determine the magnitude of the frequency at a specific point in time. We can observe how quickly [4] the frequencies improve because the time-series data is first windowed, typically in small chunks, and the FFT data is stored together to create spectrogram images. Figures 2 and 3 display the conversion frequencies from f hertz type to m mels.

With its various hyperparameters, the 2D ConvNet model specifies an explanation layer by layer. The word embedding model [4]is a pre-trained model used in experiments by Keras to identify languages. There are 22 languages contained within the dataset, and each language has 1000 rows. [4]The first step is to download the stop words using the nltk toolkit. Preprocessing complete, divide the data further into the train and test sets. Since there are 22 classes, they employ softmax when applying the word embedding model to the dataset at the output layer. The accuracy that was attained with this model was 95%. This method extracts words from a given dataset using Fig. 2. Generated spectrogram from an English audio file Fig. 3. Wavelength of audio Bernoulli Na¨ıve Bayes machine learning. [4]All the data are divided into X and Y in a preprocessing stage, after

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

which the data are encoded using a label encoder package. Following that, clean up the data by making all the sentences lowercase. Once the model is fitted, the Na¨ıve Bayes method is used, requiring 29.7 seconds and providing an accuracy of 93%.

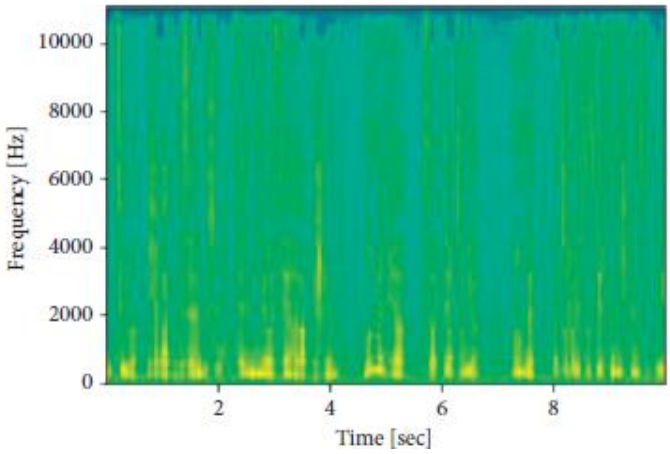

Fig. 2. Generated spectrogram from an English audio file

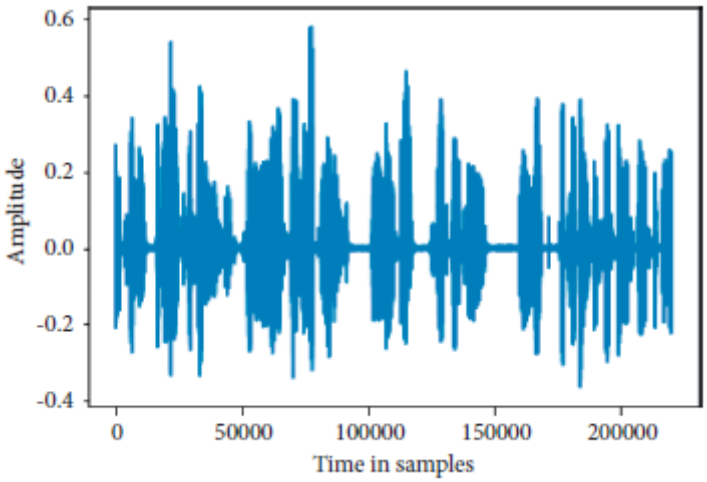

Fig. 3. Wavelength of audio

The four datasets— spoken language identification, language identification dataset, common voice Kaggle, and Mozilla common voice dataset [4] are used to evaluate different approaches. The spoken language identification dataset consists of 540 test samples and 73080 train samples in English, German, and Spanish. It includes recordings of both men and women. The audio files have FLAC (Free Lossless Audio Codec) extensions. [4]The language identification dataset comprises 22000 train examples in the following 22 languages: Arabic, Chinese, Dutch, English, Estonian, French, Hindi, Indonesian, Japanese, Korean, Latin, Persian, Portuguese, Pushto, Romanian, Russian, Spanish, Swedish, Tamil, Thai, Turkish, and Urdu. A CSV file containing 1000 rows per language holds all the information. [4] The common voice Kaggle dataset consists of 354785 audio samples that are further divided into six folders with 16 main languages: English spoken in the United States, Australia, England, Canada, the Philippines, Hong Kong, India, and South Asia; Irish; Malaysian; New Zealand; Scottish; Singaporean; South Atlantic; South African; Welsh; West Indies; and Bermuda. All languages are stored as mp3 files. It includes audio from both men and women. [4] While running the convolution neural network, word embedding Keras, and Na¨ıve Bayes, the trial and error method is utilized. Another definition of an NPcomplete task is the choosing of a hyperparameter. [4]In CNN, the epochs are set to 60, the batch size to 32, and the activation function to ReLU. With Adam Optimizer, dropouts are used. The softmax function is applied at the output layer. It is a pre-trained Keras model for word embedding. When applying categorical cross-entropy loss with Adam optimizer and 25 epochs, it is employed. [4]Utilising word embedding, SVM, logistic regression, VGG16, ResNet50, and random forest classifier, [3]the performance of different spoken language detection

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

algorithms is compared [9]. Three languages are included in Spoken Language Identification, and test and train folders also contain audio samples of several languages. Using CNN, the accuracy was successfully attained at 100% with good precision, recall, and F1 score [3]. Accuracy of 98% was achieved by 2D convolutional neural networks. A pre-trained model was used to achieve a 95% accuracy in word embedding on another dataset of CSV files. On a dataset of 22 languages, using the Bernoulli Na¨ıve Bayes technique achieved an accuracy of 93%. Furthermore, using word embedding, a pre-trained Keras model, to apply a different strategy to this dataset. It is a little bit quicker and more accurate than Na¨ıve Bayes. [4]The SVM and random forest classifier model, when applied to the 16-language data set, respectively, attained accuracy of 82.88% and 72.42%.

B. CLIASR: A Combined Automatic Speech Recognition and Language Identification System [2020]

This study, describes an approach called the Combined Automatic Speech Recognition and Language Identification System, which combines ASR and LI technologies to recognize spoken digits after identifying the language in which they are pronounced [5]. For speech-based multilingual identification and speech recognition tasks, an internal corpus made up of bilingual digits sounds with 10 digits generally pronounced in Modern Standard Arabic (MSA) and the Amazigh Moroccan dialect was used. The most well-known studies on the identification of spoken audio languages and dialects have been undertaken using acoustic data, phonotactic and prosodic approaches, and other techniques. [5]The simplest level of features that can identify a speech waveform is acoustic information, which is also the lowest level.

The authors employed a phonotactic technique to automatically identify Arabic dialects by initially modeling dialects using trigram models and subsequently a phone recognizer [5]. They also examined how prosodic features, such as rhythm and intonation, can be utilised to differentiate distinct dialects from the Gulf, Iraq, Levant, and Egypt. Voice-based language identification and voice recognition systems must distinguish between languages and recognize them, respectively [5]. A portion of the Moroccan Amazigh dialect is used in one corpus of spoken numbers, and English is used in the other.

Two experiments will be conducted, the first of which involves creating a recognition system [5] with a mixed model to recognize digits in both MSA and the Moroccan dialect. Prior to employing the ASR system, a language identification system must be implemented and maintained automatically. [5]This makes it possible to assign each spoken digit to the appropriate model. In order to account for varied combinations of speakers, they designed a system using the previously stated corpus and ten different learning and test sets, which produces four speakers representing 70% of the learning stage and two speakers representing the remaining 30% of the test phase for each set. A bilingual identification system trained on [5] standard Arabic and Moroccan Amazigh corpora will recognize the uttered digit. The output will be assigned to the appropriate Decoder (MSA or Amazigh) based on an HMM with three states and sixteen GMMs each. They have used a framework for the Language Identification System that is based on Librosa and comprises acoustic, spectral, and rhythmic characteristics. It provides features with a total of 34 components. Unknown spoken digits language is determined for the evaluation procedure by comparing them to the learning model using a similarity computation. The outcomes demonstrate that the system performs quite well, with an accuracy of 95.71%, even though it was trained on a limited corpus. With the HMM-based CMU Tool, they create a voice recognition system. Amazigh and MSA speech data require independent training of acoustic and linguistic models in order to represent both Arabic and Amazigh sounds in the CMU Sphinx system. Multiple speakers have been used to analyze the MSA model. Two speakers were used for the evaluation stage and four speakers were used to develop the learning model. Ten Arabic numbers from 0 to 9 are asked to be pronounced similarly by both genders. The results demonstrate that, even with limited data input, the system performs admirably at [5]76%, while for the Amazigh dialect, the results demonstrate a high rate of Amazigh digit recognition (89.5%).

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

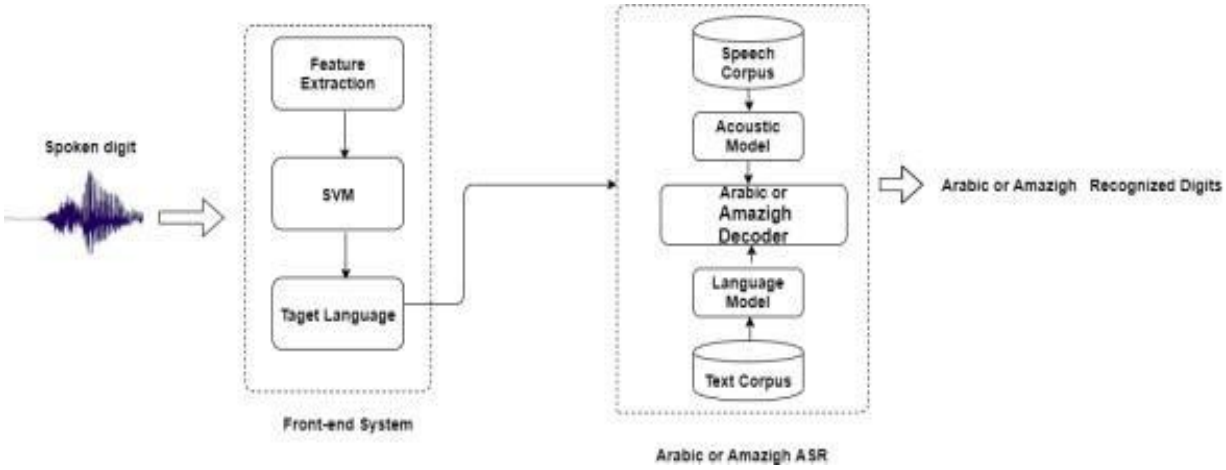

Fig. 4. CLIASR system architecture

The proposed system combines two key phases: Automatic Language Identification, which identifies the voice utterance's language (for example, Arabic or Amazigh), and an ASR system, which recognises the uttered words (digits). The proposed system's architecture is shown in Figure 4. A few instances of the scenarios taken into account are the development of an ASR trained on both Arabic and Amazigh corpora, which is regarded as a baseline system, and the combined system that consists of both LI and ASR. The results show that the system was able to achieve a rate of 79.2% while using a very small corpus size, which is an improvement of over 33%.

C. A Language Identification System using Hybrid Features and Back-Propagation Neural Network [2020]

For speech recognition, humans are the most reliable language identification systems [6]. Within seconds of hearing speech, people can tell if it is being spoken in a known language or not. Language Identification [7] is the process of accurately identifying a language that is unknown by comparing the speech biometrics of a test speech sample with language models amassed beforehand. For spoken language identification (LID) systems, this study proposes and promotes the usage of hybrid robust feature extraction algorithms. In neural networks, backpropagation is the abbreviation for "backward propagation of errors." [8]It is a common technique for developing artificial neural networks. About each weight in the network, this technique aids in calculating the gradient of a loss function. Different methods, including Mel frequency cepstral coefficients (MFCCs), perceptual linear prediction features (PLP), and relative perceptual linear prediction features (RASTA-PLP), are used individually throughout the feature extraction step. Later, the performance of the LID system based on a variety of hybrid feature combinations—including MFCC, PLP, combined with their first order derivatives, MFCC + RASTA-PLP, and MFCC + SDC (Shifted delta cepstral coefficients)—is examined.

The database includes 50 utterances of each language— Tamil, Malayalam, Hindi, and English—spoken by speakers of similar numbers of both genders to account for speaker and gender heterogeneity. The audio recordings are produced with the use of expert acoustic systems. Sennheiser dynamic microphones are used to acquire voice signals. Using specialized Sound Forge software, speech signals for each isolated word are recorded and organized into distinct files. The mono-wave file type is used to record the speech samples. This is done at a sampling rate of 16 kHz. Thus, an experimental corpus is created by following this process for documenting and analyzing the suggested strategy. On the 200 input voice signals (isolated words), all experiments were performed. The training, testing, and validation data sets are randomly split into this user-defined dataset of several languages in the ratio of 2:1:1.

The previously mentioned feature extraction algorithms are used alone first, and then various combinations are tested. [6]Pre-emphasizing the speech signals is followed by the use of individual feature extraction techniques like MFCC, PLP, and RASTA-PLP as well as hybrid forms that combine several features like MFCC + DMFCC, PLP + DPLP, MFCC + RASTAPLP, and MFCC + SDC. 40 filters are utilized in the MFCC filter bank, with the first [6] 13 filters being linear and the remaining filters being logarithmic. Then, for each isolated word, a thirteen-coefficient matrix of the retrieved features is produced. The identical matrix with 13 coefficients is also constructed for Delta MFCC (DMFCC). Later, a single matrix is created for each isolated word by concatenating

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

the acquired matrices from MFCC and DMFCC. The PLP and DPLP hybrid feature extraction algorithms use the same methodology. Finally, MFCC features along with SDC and RASTA-PLP features are incorporated [6]. Numerous studies support the best M, D, P, and K parameter settings of 7-1-3-7 for the SDC feature extraction approach. By utilizing this common configuration of SDC settings in this paper. The resulting hybrid feature vectors are then given into the FFBPNN classifier, which is described in more detail in the following section.

A matrix is used to store the feature vectors that were obtained during the feature extraction phase. The FFBPNN is then given this matrix as an input. [6]FFBPNN is trained in the supervised mode for this investigation. The MATLAB Neural Network Toolkit is used to evaluate performance. The Levenberg Marquardt ”trainlm” method and the Scaled Conjugate Gradient ”trainscg” learning algorithm are used to calculate the statistical measures for performance evaluation, including training, validation, testing error, and correctness. [6]The hidden neurons in the ”trainlm” method are set to 100. The hidden layer uses the Sigmoid activation function, and the output layer uses the SoftMax activation function. For the remaining neural network parameters, default values are used. The scaled conjugate gradient's ”trainscg” function is used to choose the hidden neurons, which are 30 in number. There are between 30 and 60 different epochs.

Extraction of feature vector matrices for every isolated word in the training corpus. Create a single matrix with all the acquired features for each isolated word in the training database. Carry out the aforementioned procedures for each isolated word in the test and validation corpora. The FFBPNN classifier is further fed the three matrices that were obtained for training, validation, and testing. To get the least amount of error and the best identification accuracy, the FFBPNN parameters are selected and calculations are done. [6]The training, validation, and testing performance data are recorded in the last step.

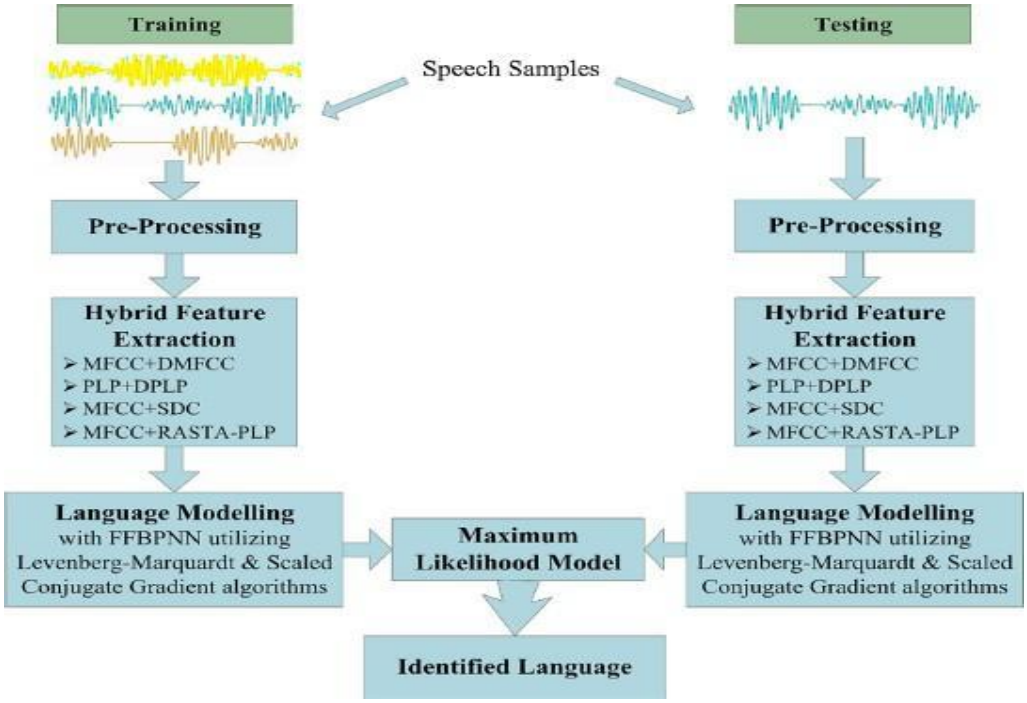

Fig. 5. Process flow diagram of the proposed method

Figure 5 shows the suggested method's overall process flow diagram. In this study, [6]two learning algorithms—”trainlm” and ”trainscg” separately with the FFBPNN classifier—are used to assess the performance of the LID system by merging the various features or hybrid features on the identification rates. Feature extraction is crucial to this task. This work makes use of several hybrid feature extraction algorithms. The 26-dimensional input feature vector applied to the FFBPNN with the ”trainlm” learning function results in the best performance. With the MFCC + RASTA-PLP feature extraction strategy, the highest identification rate of 94.6% is accomplished. Later, a three-layer BPNN is used to classify data, and the impact of increasing the epochs is examined. The results show that as the number of epochs is increased, identification accuracy increases. Similarly to this, efficient training is crucial because it gives access to a reliable prediction output [6].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

The criteria for stopping training must therefore be carefully considered. The system weights are then continuously adjusted after this stage, greatly reducing the discrepancy between the desired and expected outputs. Error is defined as a difference between the desired and projected results. An error matching the training error is produced to execute the trained model on the training data to update the weights. [6]Several simulations have been done using an isolated word dataset of words from different languages to assess and demonstrate the influence of the hybrid features on language identification.

D. Spoken Language Identification System Using Convolutional Recurrent Neural Network [2022]

This approach for identifying spoken languages is based on the arrangement of feature vectors [9]. For spoken language identification [9] on seven languages, including Arabic, selected from subsets of the Mozilla Common Voice (MCV) corpus, the proposed system uses a hybrid Convolutional Recurrent Neural Network (CRNN), which combines a Convolutional Neural Network (CNN) network with a Recurrent Neural Network (RNN) network. The CRNN architecture is created by the suggested system by combining the best features of the CNN and RNN architectures. The objective of this paper [9] is to introduce an innovative spoken LID system for seven languages, including Arabic, that examines and analyses spoken LID issues using the most advanced techniques now in use.

It contrasts the Mel Frequency Cepstral Coefficient (MFCC) feature and the Gammatone Cepstral Coefficient (GTCC) feature, as well as a mixture of both, during the feature extraction step [9]. The suggested system's results show that when GTCC and MFCC characteristics are combined, they perform better than when they are employed alone. [9] Classifying spoken language from a given audio sample is the aim of spoken LID systems. The primary contribution of this research is the consideration of spoken language identification for seven languages, including Arabic, utilising public audio speech corpora and the combined application of the MFCC and GTCC within the CRNN framework. [9] It evaluates the impact of fine-tuning the features. conduct extensive tests and experiments with the proposed architecture, demonstrating its applicability in a variety of contexts and its extension to new languages; and used system errors to infer the degree of similarity in considered languages [9]. The hybrid CRNN combines a CNN's descriptive capabilities with an LSTM architecture's capacity to capture sequence data.

The publishers regularly add new voices and languages; the most recent release has 87 different languages. [9] The MCV project relies on crowdsourcing for data validation and collection in order to scale and be sustainable. The spoken LID experiments employing subsets of MCV [9] for seven target languages—Arabic, German, English, Spanish, French, Russian, and Chinese—are shown as an illustration of how MCV might be used. For the majority of these languages, these are the initial spoken LID experiments. The goal was to evaluate Arabic's spoken LID performance in comparison to other languages. [9] There have been almost 200,000 male and female speakers who have contributed, totaling 18,243 hours of audio. By identifying speech boundaries, data was processed. Specifically, using R2020b-MATLAB's identify speech function, voice signals were segmented, and silent areas were eliminated to correspond with speech boundaries. To avoid the silent parts affecting the categorization work, this was done. It was proposed to bring together MFCC and GFCC characteristics in the spoken LID system to decrease signal redundancy and increase accuracy.

[9] This article suggests a CRNN architecture that combines CNN and LSTM. While the LSTM has demonstrated its capacity to detect the language in sequence-to-sequence series, CNN is a great network for feature extractions. [9] The new aspect of the proposed CRNN design is the use of three layers—a sequence folding layer, a sequence unfolding layer, and a flattened layer—to transform the kind of input data based on the networks utilized in CRNN. For CNN to extract spatial characteristics from the input array, the sequence folding layer turns the image sequences into an array of images. The flattening layer [9] transforms image sequences into feature vectors for input to the LSTM layer, and the sequence unfolding layer transforms this array of images back into image sequences. In this paper [9], two methods for performance evaluation of the system are proposed. The first is an evaluation that was implemented at the sequence level, and the second was done at the file level. The correctness of each sequence was determined to calculate the overall speech file accuracy. This study [9] evaluated the performance of the CRNN architecture for the spoken LID problem using four performance metrics: accuracy (A), precision (P), recall (R), and F-measure (F1). TP stands for all true positives, TN for all true negatives, FP for all false positives, and FN for all.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

A comparison of the precision of the proposed system's spoken LID using GTCC and MFCC was conducted. The accuracy of GTCC in comparison to MFCC was assessed for all files utilising a taken-into-account MCV corpus [9] with the execution of five runs. Table 5 displays the experimental findings for the GTCC, MFCC, and combined characteristics. When GTCC and MFCC were combined as system features, the highest average testing accuracy was attained. This work resulted in better Experiments in terms of the accuracy of the proposed system, whereas one of the Experiments surpassed another one in terms of execution speed. Therefore it is 27% faster but approximately 0.9% less accurate. Arabic was the language with the highest accuracy in terms of identification, followed by Russian. Spanish had the weakest language identification accuracy. Due to the similarities between English and Spanish in particular [9], two standard architectures also demonstrated difficulty in distinguishing between the two languages. German and Spanish have a lot in common linguistically. French, on the other hand, is comparable to Spanish. Although English has a somewhat stronger predisposition towards French, German, and English are more likely to be confused [9]. Overall, each language's acquired representations of the architecture are quite distinct.

In Arabic, the GTCC's average spoken LID precision was 94.88%, compared to 94.20% for Chinese, 91.30% for English, 90.58% for French, 74.84% for German, 92.98% for Russian, and 89.86% for Spanish. [9]In contrast, the average MFCC precision for spoken LID in one experiment was 94.82% for Arabic, 94.88% for Chinese, 89.76% for English, 92.14% for French, 76.78% for German, 95.32% for Russian, and 91.72% for Spanish. The findings show that GTCC beat MFCC in spoken LID in Arabic and English [9]. The system achieved the highest overall average accuracy of 92.81% for detecting the seven spoken languages taken into account in the selected corpus, through experiments with various settings.

[9]The average precision for spoken LID in Arabic, Chinese, English, French, German, Russian, and Spanish while combining GTCC and MFCC was 95.84%, 93.85%, 92.64%, and 81.87%. Furthermore, the tests show that the suggested approach distinguishes between related languages, provides the best results for detecting Arabic and Russian, and is expandable to other languages. Additionally, this paper serves as a useful resource for those looking to create ASR system that deal with Arabic.

III. COMPARATIVE STUDY

Spoken language identification (SLID) is the method of identifying language from an audio clip by an unknown speaker, regardless of gender, speaking style, and distinct age speaker. It is challenging to identify the characteristics that can effectively and clearly distinguish different languages.

The first article offers two contributions to the study of spoken language identification [10]. To identify languages from audio-generated images, first employ the deep learning architecture for image classification. Short files that require less preprocessing can deliver powerful performance. This method produced positive outcomes with a 98% accuracy rate. Secondly, applying the Bernoulli Na¨ıve Bayes method on a dataset of 22 languages for language identification [3]. In terms of model fitting data, it takes a little bit longer than CNN. With this method, accuracy is 93%. Additionally, using a word embedding pre-trained model by Keras to apply a different strategy to this dataset. Compared to Na¨ıve Bayes, it is a tiny bit faster and more precise. This method had a 95% accuracy rate. By eliminating the audio noises, log-Mel spectrogram performance can be further improved [3]. The provided data can be enhanced by applying numerous techniques including pitch shifting, cropping, rotating, flipping, adding random noise, changing audio speed, and other techniques [3]. They aid in strengthening neural networks' resistance to changes that could be prevalent in real-world circumstances. Various feature extraction methods, such as Constant-Q transform and Fast Fourier Transform, and their effects on language recognition are frequently observed or reviewed in more detail [3]. They are known to have an advantageous effect on how well convolutional neural networks operate.

Compared to the first paper, the presented system in the second paper combines [5] two ASR and LI systems— one for language identification and the other for speech recognition— and runs on a bilingual corpus (Amazigh Moroccan dialect and MSA) to efficiently recognize the spoken digits by switching to a model that has been specifically trained on that language. The design of an ASR trained on both Arabic and Amazigh corpora [5], which is regarded as a baseline system, and the combined system that comprises both LI and ASR were both taken into consideration in trials. The results demonstrate that despite the limited corpus size employed in the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

system, were still able to reach [5] a rate of 79.2% (an improvement of about 33% i comparison to the reference system). Although the method is excellent, it still requires a lot of time and resources, especially [5] when the complexity of the data is significant.

The third study developed a hybrid feature extraction system-based automatic LID system. The classifier employed in this study is FFBPNN, which identifies languages using [6] the two learning algorithms ”trainlm” and ”trainscg.” The results obtained demonstrate performance improvement with the usage of hybrid features in terms of accuracy and test error rate. Additionally, it has been demonstrated that the LID system performs [6] better with the ”trainlm” learning algorithm than with the ”trainscg” learning method. The greatest classification rates achieved [6] with MFCC + RASTA-PLP hybrid features utilising the ”trainlm” learning function are 94.6 with a minimal test error rate of 0.10. The LID system’s performance is evaluated using a user-defined database of four languages: Tamil, Malayalam, Hindi, and English rather than being estimated across several datasets. The suggested study will assist the researchers in carrying out comparable speech-processing tasks using far bigger language datasets that were able to use due to the limitations of the available computer resources. Because equivalent LID frameworks can be constructed [6] and evaluated with standard datasets with less computation time when GPU and other sufficient computational resources are available. This LID system can play a significant role in several activities, including spoken language translation, travel guidance, emergency assistance, and linguistic interpretation [6].

In the fourth paper [9], it employs a spoken LID system with a CRNN architecture that utilises the combined GTCC and MFCC properties of speech signals. The features from the GTCC and MFCC, as well as a combination of both, were compared during the feature extraction step. In Arabic, the GTCC’s average spoken LID precision was 94.88%, compared to 94.20% for Chinese, 91.30% for English, 90.58% for French, 74.84% for German, 92.98% for Russian, and 89.86% for Spanish. [9] In contrast, the average MFCC precision for spoken LID in one experiment was 94.82% for Arabic, 94.88% for Chinese, 89.76% for English, 92.14% for French, 76.78% for German, 95.32% for Russian, and 91.72% for Spanish. [9] The findings show that GTCC beat MFCC in spoken LID in Arabic and English. The system achieved the highest overall average accuracy of 92.81% for detecting the seven spoken languages taken into account in the selected corpus, according to the experiments with various settings. [9] The average precision for spoken LID in Arabic, Chinese, English, French, German, Russian, and Spanish while combining GTCC and MFCC was 95.84%, 93.85%, 92.64%, and 81.87%. Furthermore, the tests show that the suggested approach distinguishes between related languages, provides the best results for detecting Arabic and Russian, and is expandable to other languages. Additionally, this paper serves as a useful resource for those looking to create ASR systems that deal with Arabic.

IV. CONCLUSION

Regardless of the speaker’s gender, speaking style, or age, spoken language identification (SLID) is a technique for recognising a language from an audio sample. By combining a variety of deep learning, machine learning, datasets, and performance indicators, multiple spoken language identification models have been created in this work. Applications for spoken language identification are becoming more and more important in our daily lives, particularly in the area of multilingual speech recognition. The SLID system’s three essential elements are data collecting, feature elimination, and language categorization.

Convolutional neural networks (CNN) have produced outcomes with an average accuracy of 98%, and Convolutional Recurrent Neural Networks had an average accuracy of 92.81%. For a given multilingual mixed speech corpus, the proposed LI-ASR system achieves a rate of 79.2%, which is 33% better than an ordinary ASR and with an overall accuracy of 94.6% and a minimal test error rate of 0.10, the MFCCRASTA- PLP hybrid feature extraction technique outperforms the other hybrid feature extraction techniques.

Spoken language systems have been thoroughly researched in terms of feature extraction and classifier [9] learning using a variety of architectures. However, they are limited to the majority of human languages. The main problem is that, within the group of languages, Arabic has received very little attention from the spoken LID research community, given the fact that many academics have dealt with spoken LID problems successfully using a variety of strategies. In the future, more approaches should be brought to identifying the languages that are more difficult to distinguish.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

ACKNOWLEDGEMENT

We express our sincere gratitude to all the teaching staff of the Department of Computer Science, Vimala College Autonomous Thrissur, for their valuable guidance and support at each stage of the term paper.

V. REFERENCES

[1] S. Guha, A. Das, P. K. Singh, A. Ahmadian, N. Senu, and R. Sarkar, "Hybrid feature selection method based on harmony search and naked mole-rat algorithms for spoken language identification from audio sig nals," IEEE Access, vol. 8, pp. 182868–182887, 2020.

[2] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, and M. Masud, "Spoken language identification using deep learning," Computational Intelligence and Neuroscience, vol. 2021, 2021.

[3] "Liverpool johnmoores university." https://www.hope.ac.uk/clearing/?gclid=EAIaIQobChMI5r-bx5vPgAMVpYRLBR0h9AMOEAAYASAAEgKEGfDBwEgclsrc=

[4] "National library of medicine." https://www.ncbi.nlm.nih.gov/ .

[5] K. Lounnas, H. Satori, M. Hamidi, H. Teffahi, M. Abbas, and M. Li chouri, "Cliasr: A combined automatic speech recognition and lan guage identification system," in 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), pp. 1–5, 2020.

[6] D. Deshwal, P. Sangwan, and D. Kumar, "A language identification system using hybrid features and back-propagation neural network," Applied Acoustics, vol. 164, p. 107289, 2020.

[7] "Researchgate." www.researchgate.net , 2008.

[8] "Submission by the republic of maldives, mal-es-doc."
https://www.un.org/depts/los/clcsnew/submissionsf iles/mdv5310/MAL– ES – DOC.pdf.

[9] A. A. Alashban, M. A. Qamhan, A. H. Meftah, and Y. A. Alotaibi, "Spoken language identification system using convolutional recurrent neural network," Applied Sciences, vol. 12, no. 18, 2022.

[10] P. Bam, S. Degadwala, R. Upadhyay, and D. Vyas, "Spoken language recognization based on features and classification methods: A review," in 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 868–873, 2022.

[11] G. Choueiter, G. Zweig, and P. Nguyen, "An empirical study of automatic accent classification," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4265–4268, 2008.

[12] J. Oruh, S. Viriri, and M. F. Ijaz, "Deep learning-based classification of spoken english digits," Intell. Neuroscience, vol. 2022, jan 2022.

[13] H. Qu, X. Chen, J. Hong, Y. Xu, C. Li, Z. Li, and Y. Liu, "Experimental and 3D Numerical Investigation on Proppant Distribution in a Perforation Cluster Involving the Artificial Neural Network Prediction," SPE Journal, pp. 1–26, 02 2023.

[14] J. Chen, T. Okamoto, and S. Belkada, "Interactive tutoring on communication gaps in a communicative language learning environment," in International Conference on Computers in Education, 2002. Proceed ings., pp. 445–446 vol.1, 2002.

[15] D. Kimothi, P. Biyani, J. M. Hogan, A. Soni, and W. Kelly, "Learning supervised embeddings for large scale sequence comparisons," bioRxiv, 2019.

[16] B. Karlik, "Machine learning algorithms for characterization of emg sig nals," International Journal of Information and Electronics Engineering, vol. 4, 01 2014.

[17] W. Kang, M. J. Alam, and A. Fathan, "Deep learning-based end-to-end spoken language identification system for domain-mismatched scenario," in Proceedings of the Thirteenth Language Resources and Evaluation Conference, (Marseille, France), pp. 7339–7343, European Language Resources Association, June 2022.

[18] H. S. Das and P. Roy, "Chapter 5 - a deep dive into deep learning techniques for solving spoken language identification problems," in In telligent Speech Signal Processing (N. Dey, ed.), pp. 81–100, Academic Press, 2019.