LANGUAGE IDENTIFICATION AND TRANSLATION SYSTEM USING DEEP LEARNING

Abstract

Language Identification is the process of identifying the language being spoken by an anonymous speaker from an audio clip. Machine and Deep Learning techniques such as Artificial Neural Network, Long Short-Term Memory are used for language identification and then comparing these models for evaluating the performance. The objective is to identify the language effectively from spectral features extracted using Mel Frequency Cepstral Coefficient from pre-processed audio samples. Language Translation is the process of translating from one language to another. The goal of language translation is to make it easier for groups or individuals who speak various languages to communicate and comprehend one another and exchange ideas. The solution for this problem is by introducing intermediate language which is understood by both of us in conversation. We used English as an intermediate language which is widely spoken all over the world and it follows strong grammatical rules that make it suitable for use as an intermediate language. Thus, the first step is to translate input text into an intermediate language and then translate into target language text from intermediate language text. It is feasible to increase the precision and quality of language translations by applying Deep Learning techniques, such as Neural Machine Translation. The main goal of this work is to maintain ordering of words in a sentence which is accomplished through Positional Embeddings mechanism and to preserve contextual meaning of words in a sentence which is done by performing a Multi headed attention mechanism.

Authors

Santhi. S
Department of CSE
Mepco Schlenk Engineering College
Sivakasi, India.
santhicse@mepcoeng.ac.in

Selva Nidhyananthan. S
Department of ECE
Mepco Schlenk Engineering College
Sivakasi, India.
nidhyan@mepcoeng.ac.in

Menaga Devi. R
Department of CSE
Mepco Schlenk Engineering College
Sivakasi, India.
menagadevi_cs@mepcoeng.ac.in

Mahalakshmi. K
Department of CSE
Mepco Schlenk Engineering College
Sivakasi, India.
mahalakshmi2002_cs@mepcoeng.ac.in

Keywords: Language Identification, Language Translation, Neural Machine Translation, Positional Embedding, Encoder-Decoder model, Long Short-Term Memory.

I. INTRODUCTION

Language is the primary means of human interaction. Each language has a unique syllable structure, pronunciation pattern, and usage pattern, which makes it challenging to train a machine to distinguish between them. Social media networks utilize language identification and translation to convert user-generated material from one language to another. One of the challenging tasks that supports many real-time apps that facilitate communication between people who are unable to understand the language of others which is solved by good translation mechanism. Such translation requires one of the significant steps to be handled prior is Language Identification. Such Identification is determined from speech in advance and it is an important task for the translation system to select the ideal translation model for the language it is translating.

People from different regions find it difficult to grasp each other language which leads to communication barrier among them. Intermediate Language Translation is a preferable strategy for this challenging task. As long as both parties can agree on an intermediate language, it is more reliable solution to translate from the source language into the intermediate language, which can then be translated into the target language. Word order, word structure, and contextual meaning preservation are all the significant criteria to consider while translation.

An intermediate language serves as a link between two dissimilar source and target languages which is a standardized, machine-readable language. The use of an intermediate language can increase the effectiveness and precision of language translation because it enables a more organized and consistent contextual representation. English is used widely used all over the world. The material made available on the web has a significant influence on English. English is the language of communication for 20% of the world's population. To bridge this enormous linguistic gap with the least amount of human intervention, operational and precise computational solutions are required. English, a second language used by a large population and it also featured as grammatical rules that make it a good starting point for learning other languages. Such translation process divided into two steps,

- Language Identification from audio sample
- Language Translation (Source → Intermediate → Target)

1. **Language Identification:** Many speech signal processing applications currently depend heavily on the precise language identification of the speech instance. Due to individual speakers' differing in word pronunciations, acoustic factors, and regional accent variations, such audio signals are highly varied. Speech that is noisy alters its spectral characteristics, which could result in incorrect classification. The pre-processing approaches should be used effectively in order to maximize the classification performance metric. Spectral features are used to train the neural network layer which enriches identification performance. The purpose of this study is to observe how well a deep learning framework identifies and categorizes language based on spectral properties of an audio stream.

2. **Language Translation:** In tradition, machine translation systems relied on statistical models or rule-based methodologies which do not focus on the context of the text. Translations as a result were frequently incorrect and even challenging to understand. On the other hand, Sequence to Sequence models make use of a neural network architecture that can figure out the context of the text being translated. In particular for complex languages with various sentence structures, word ordering, and inflections, this enables more accurate translations. In this work, Spanish-English, French- English, Italian-English and also Indian language Hindi-English language pair were used. Neural Machine Translation with attention is used in this work, which is the process of training the neural network with enormous amounts of text data corpus. The neural network predicts the next word by analyzing the context of the preceding word, which is done via sequence-to-sequence prediction

II. RELATED WORKS

Deshwal et al. [1] introduced a Language Identification System that utilizes a combination of feature extraction techniques such as Mel Frequency Cepstral Coefficients + RASTA-Perceptual Linear Prediction (MFCC+RASTA+PLP) and a Language Modelling Technique called Feed Forward Back Propagation Neural Network, trained with "trainlm". The system performs best with a 26-dimensional input feature vector and the "trainlm" learning function, although achieving high identification accuracy requires running multiple epochs. Boussaid et al. [2] developed a system for identifying individual Arabic words using a hybrid approach to feature extraction. The PLP, RASTA-PLP, and MFCC techniques, along with their first order derivative, were used in conjunction with "trainscg" learning and a classifier for the FFBPNN neural network. Although the Support Vector Machine (SVM) classifier works effectively in spaces with multiple dimensions when there is a distinct margin of distinction between groups, this paper finds that it was not suitable for large data sets with more noise and lacks a probabilistic explanation for classification. Sekkate [3] proposed a Gamma tone Frequency Cepstral Coefficients Hybrid feature extraction method with SVM classifier for classification and feature extraction. A system for identifying single Arabic words was created by Boussaid et al. [4] utilizing a hybrid feature extraction approach. They employed the PLP, RASTA-PLP, and MFCC techniques, as well as their first order derivative, in combination with "trainscg" learning and a classifier for the FFBPNN neural network. While the SVM classifier performs well with a clear separation margin between classes and is more efficient in high-dimensional spaces, this study concludes that it was unsuitable for large datasets with significant noise and did not provide a probabilistic explanation for classification.

Hassine M et al. [5] achieved recognition rates of 98.3% using FFBPNN and an accuracy of 97.5% with SVM. However, FFBPNN outperformed SVM in terms of performance despite requiring a longer computation time. Ankur Maurya [6] proposed a Speaker Recognition system for Hindi Speech Signals that utilizes MFCC for feature extraction and Gaussian Mixture Model (GMM) for classification. The study achieved an 86.27% accuracy for text-independent recognition using the MFCC-GMM approach for Hindi speech samples. Eslam Mansour Mohammed [7] evaluated the use of Linear Predictive Coding (LPC) and MFCC features with artificial neural networks for spoken language identification, which was crucial for multilingual services. The study demonstrated the effectiveness of these features and provides valuable insights for improving spoken language identification. Muhammad Bagus Andra [8] presented a method for generating text

transcripts and recognizing speakers in concurrent Bahasa Indonesian conversations. Pitch-aware speech separation and the Reinforced Learning (RL) Model are combined in the suggested method. Recurrent Neural Network (RNN), and external language and spelling correction models. This approach could be beneficial for online discussions and remote conferences involving simultaneous speech. Priyank Mathur et al. [9] developed the Stanford Language Identification Engine (SLIDE), which was used to extract meaningful information from large amounts of raw data for language identification automatically. SLIDE was achieved by using DSL, SLIDE distinguished among comparable dialects with a 95.12% efficiency.

Marco Lui et al. [10] proposed a method for automatic detection and language identification of multilingual documents. It overcomes cutting-edge technology methods for language identification in multilingual documents and is able to identify multilingual documents based on actual data. Saha et al. [11] proposed a new end-to-end system that uses feature-based and deep learning-based methods, including Generative Adversarial Network (GAN) and Convolutional Neural Network (CNN), to detect and identify the language of scene text regions in images. This model was tested on various datasets, including an in-house multi-lingual Indic scene text dataset, and achieves satisfactory results. Zhang et al. [12] proposed a common structure for multilingual automatic speech recognition and language identification to address the issue of recognize speech in several languages. Kano et al. [13-14] proposed Transcoder, a transformer-based model for direct speech-to-speech translation, which eliminates the need for separate components in traditional approaches. JoeyNMT and Transformer Neural Machine Translator (NMT) with self-attention were and assessed using the Bilingual Evaluation Understudy (BLEU) score. Language identification models were developed using multinomial naive Bayes (MNB) and logistic regression. Pavan et al. [15] explored language identification techniques for multilingual machine translation, specifically for Hindi, Marathi, and Sanskrit. The authors found that the most accurate technique is a support vector machine-based language identifier. They also found that including language identification improves translation quality. Bahar et al. [16] proposed grid based bilateral translation with two dimensions and that translated in both directions. The paper included experimental results on German to Turkish and English to translating duties into English, which substantiate the claimed model generates good quality translations in both directions. Nakamura et al. [17] presented the ATR Multilingual Speech-to-Speech Translation System for translating between English and Asian languages (Japanese and Chinese). The system's three key components are text-to-speech synthesizing, machine text-to-text translation, and the continuous recognition of voices. Zhang et al. [18] proposed a new approach for identifying and transforming comparative sentences in patent Chinese-English machine translation. The approach was based on Hyper Netted Chain theory, which uses rules instead of statistical methods. Pradeep et al. [19] suggested examining a hybrid method of extracting features for identifying English handwriting. George et al. [20] proposed Context Dependent-Deep Belief Network-Hidden Markov Model (CD-DBN-HMM) for Large Vocabulary Continuous Speech Recognition (LVCSR) with CD-DBN-HMMs However, the pre-training of a five-layer DBN-HMM took about 62 hours and fine-tuning took over 16.8 hours. Jayanti et al. [21] proposed a dual translation model that uses Neural Machine Translation with attention mechanism to translate both international and Indian regional languages. Vaswani et al. [22] proposed hyper parameters for the seq2seq model by Google Brain. Meanwhile, the Angla Bharati-I Computer Perception Framework [23] was developed by researchers at IIT Kanpur to provide a practical interpretation system for translating English into Hindi. Sumit Kushwaha [24] mentioned the futuristic view of

Artificial Intelligence and Machine Learning algorithms in financial and healthcare industries. Sumit Kushwaha [25] surveyed Artificial Intelligence and Human Computer Interaction in most popular fields.

III. DATASET PREPARATION

Working with neural networks requires a great deal of knowledge. Learning is more reliable when neural networks learn with enough training data. The dataset for Language Identification consists of 250 audio samples of each language namely: Hindi, English, Spanish, French, Italian each with duration of three to five seconds. Audio recordings were collected from https://www.commonvoice.com/ with extension in wav format. The dataset collected for language translation from http://www.manythings.org/anki/ which are the pair of language such as English-French, English-Italian, English-Spanish, English-Hindi each dataset consists of nearly around 50,000 texts. Text samples which taken for training around 5000 texts for each of the language pairs. These corpora are not in pre-processed format, which are then pre-processed by using NLP techniques.

IV. PROPOSED WORK

1. **Pre-Processing:** Pre- processing is used to clean the data before building the model and to get good results in the feature extraction stage. The pre-processing approaches should be carried out effectively in order to maximize the performance metrics.

   - **Audio Pre-Processing:** *Pre*-processing includes identifying and eliminating extraneous noise signals that occurred in voice. Erroneous rate in feature extraction is brought on by noisy signals. Speech that is noisy significantly alters its spectral characteristics, which could result in incorrect classification.

   - **Algorithm : Removal of Noisy Signal**

     Input: Noisy Input speech Signal
     Output: De-noised Speech signal
     Steps:
     - Over the noisy audio clip, a Fast Fourier Transform (FFT) is calculated.
     - Statistics are computed using the FFT.
     - Based on the noise's statistical distribution, a threshold is determined.
     - By comparing the signal FFT to the threshold, a mask is identified.
     - A smoothen filter is used to averaging the mask over frequency and time.
     - The mask is inverted.

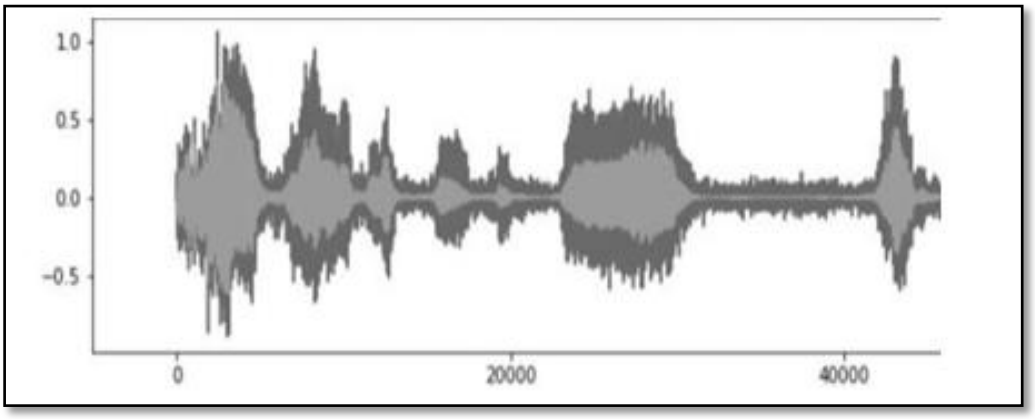


Figure 1: Noised Audio Signal

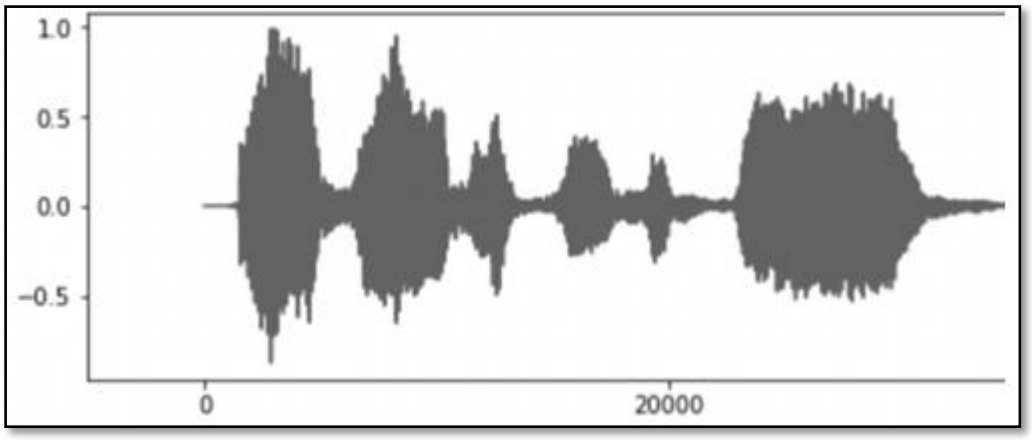


Figure 2: Denoised Audio Signal

Figure1, illustrates the separation of the clean audio signal from the noisy signal, Figure 2, illustrates denoised audio signal. Next pre-processing step is Voice Activity Detection (VAD). It is used to reduce the data load by removing the unnecessary silence signal from the input audio which is determined by creating buffer sized window capacity of 1000 and fix the threshold which is predetermined as 0.012. Using this buffered window on all over the audio signal and calculate root mean square for each of the frames and for less than of threshold signal frames gets overlap on the previous frames. By repeating this process and VAD signal is determined as follows:

$$tot_s = \frac{S_i}{Buffer\ Size} \quad (1)$$

where tot_s is the length of fragmented samples which is determined by dividing audio sample S_i and capacity of the buffer. Figure 3 illustrates presence of silence in audio signal Figure 4 represents removal of silence to reduce data load.

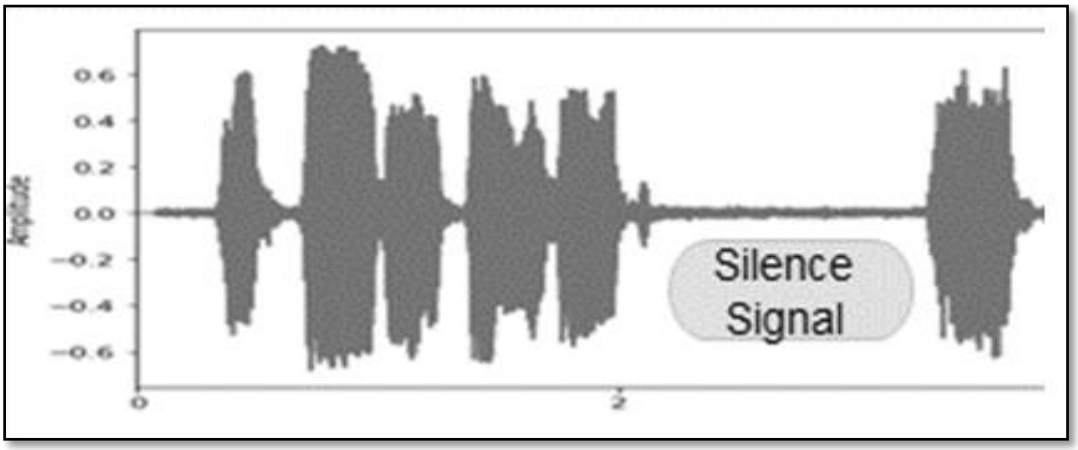


Figure 3: Presence of Silence Audio Signal

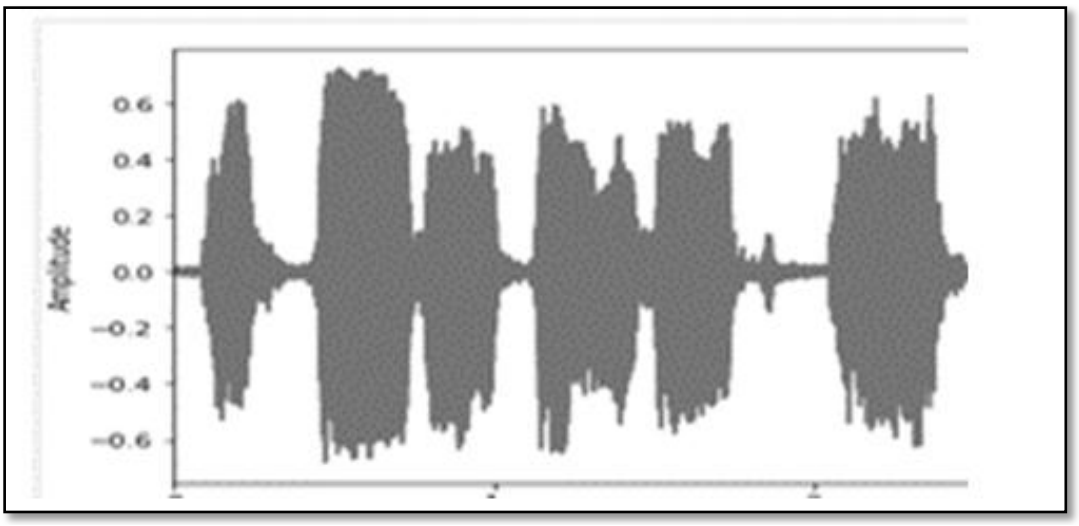


Figure 4: Removal of Silence Audio Signal

2. **Text Pre-Processing:** Pre-processing is the foremost step and essential in cleaning up the raw input text and converting it into a suitable format so that machine learning algorithms can work effectively on it. Pre-processing helps the input text to be cleaned up and any extraneous or irrelevant information removed. For the purpose of such processing and analyzing the text corpus, Natural Language Processing (NLP) offers a set of tools and techniques that are essential for text preparation. It is the process of removal of redundant and irrelevant information such as punctuation, special characters which makes the data noisy. Breaking the text corpus down into parts such as word tokens to identify unique words and maintain a dictionary for each of the tokenized chunk of words is one of the fundamental processes which helps to identify the vocabulary size of the text corpus. In deep learning, Keras is an effective library which enables one to convert a corpus of text into a sequence of tokens, each of which stands for a unique word within the corpus. Tokenization applies to input sentences to separate out unique words. Moreover, the Tokenizer is used for techniques for converting tokens into numerical values that can be fed into deep learning models. The next step, padding, ensures the model operates effectively by adding the padded sequence to make all of the text corpus in fixed length

because varying length in sequence causes issues in performance. Deep learning models cannot process inputs of varying lengths since they are made to operate with inputs that have a constant length. In order to make the lesser inputs equal in length to the longest input in the dataset, padding is used to address this issue. Word embedding is a technique used to represent words into numerical format and find patterns, relationships in text data that would be challenging to find using conventional text processing techniques. A neighboring vector in the embedding space has handle out-of-vocabulary (OOV) words, or words that are absent from the training data. The primary goal of word embedding is to represent each word in a high-dimensional vector space, where words with related meanings are clustered together. These pre-processing techniques tackle the specific difficulties brought on by multilingual text data and can aid in enhancing the precision and effectiveness of language translation mechanisms.

V. FEATURE EXTRACTION

Due to complexity of audio signal, the raw audio samples could not feed immediately to the model. Mel Frequency Cepstral Coefficient (MFCC) spectrum characteristics are the most significant and frequently used features extracted from audio samples. In order to extract spectral features, some of the steps to be followed. Fast Fourier Transform (FFT) technique is an essential step of extracting MFCC features. As lower frequencies carry more energy than high frequencies and that high frequencies have a tendency to decay much more rapidly than lower frequencies, a pre-emphasis filter is used. This flattens the spectrum of the signal and makes it more consistent. The first order filter is given in Equation (2) of pre-emphasis filter to a signal s.

$$Pre\_emp(t) = s(t) - \alpha s(t-1) \quad (2)$$

Pre_emp(t) is pre pre-emphasized signal at t time, S(t) is a input signal at time t ,α is pre - emphasis coefficient typical value as 0.95. The next stage is to block the frame over pre-emphasize filtered audio samples. Each frame is of size 0.025, and each stride is of 0.01. Apply Hamming windowing function to each frame after splitting the signal into frames are given in Equation (3).

$$w(t) = 0.5\,(1 - \cos(2\pi t N)),\ 0 \le t \le N \quad (3)$$

where N is the window length. Later, Fourier transform is calculated to analyses frequency domain and then a triangular Mel-filter is used to determine the logarithm of the Mel filter bank energies. The extracted Mel scaled features are numerical vectors which is then fed into model for classification

VI. METHODOLOGY

1. Language Identification

- **Machine Learning Techniques:** The earlier techniques of classification such as Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Deep Belief Network (DBN) are trained to obtain accuracy around 88%. In speech recognition systems, GMMs are used to determine the probability distribution of audio samples. As a result, when given an input audio, the system is able to determine the most

similarity of given word but the methodology is expensive, both in terms of memory and computational time. DBN is probabilistic based model which provides observable data and labels with probability distribution but they failed to consider for two-dimensional data. SVM works for well separated margin of separation and its more effective in high dimensional data but this is not suitable for handling mass amount of data set which consists of more noisy part makes inaccurate classification. Artificial Neural Network model (ANN) are used in LID system used to perform classification or identifying the given utterance of audio samples. It consists of one input layer with 100 computation neuron units and two concealed layers with RELU activation function and output layer with SoftMax classifier to classify the language using Adam optimizer shown in Figure 5.

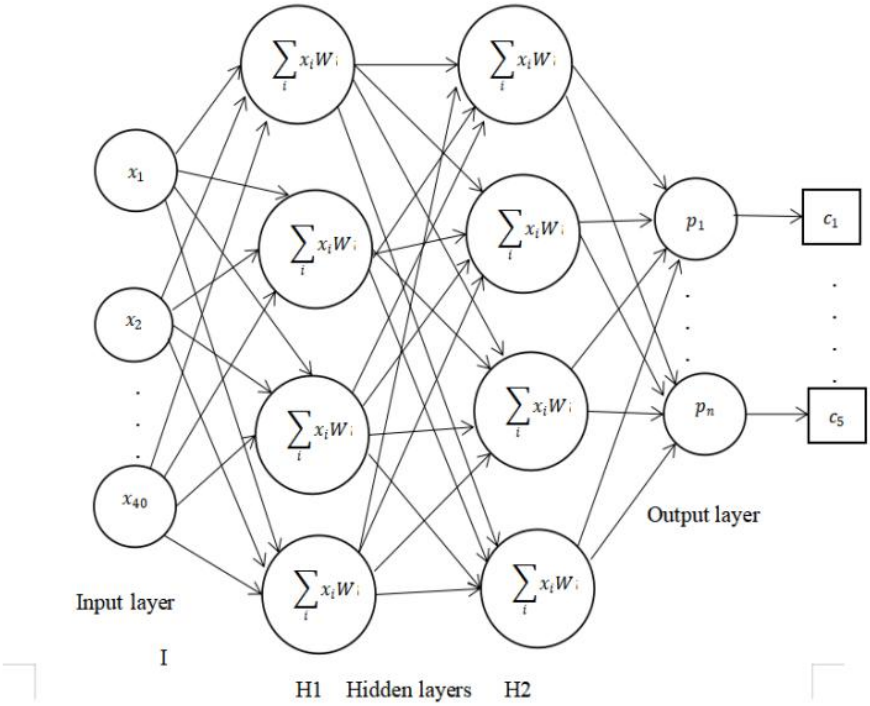


Figure 5: Artificial Neural Network Architecture

- **Deep Learning Techniques:** Long Short Term Memory (LSTM) is one of the deep learning frameworks, variant of Recurrent Neural Network (RNN) and primarily used for training, processing, and classification of sequential data. LSTMs have a memory cell that helps them remember information for longer periods of time and it resolves vanishing gradient problem effectively. The deep learning model built with one input layer with 256 neurons, two hidden dense layer each of which 128 hidden computation unit neurons with RELU activation function and one output layer which is for classifying the language shown in Figure 6. A training set of spectral feature vectors of 40 fed as input to the model which then trained and built as language classifier.

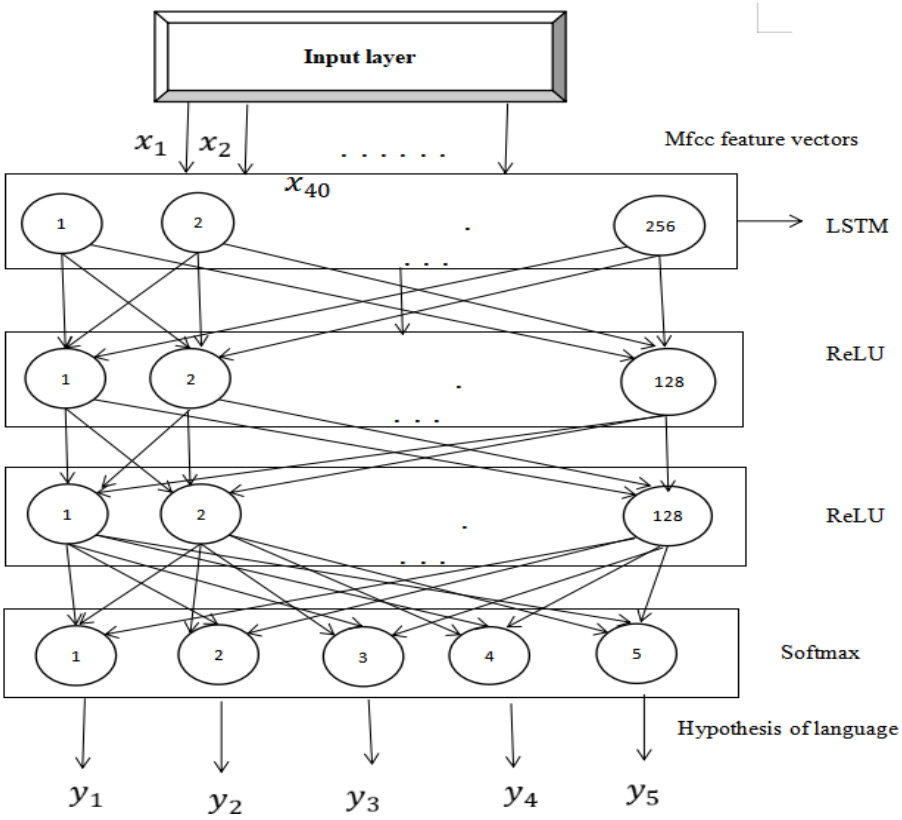


Figure 6: LSTM architecture for Language Identification

2. **Language Translation:** Neural machine translation well suited to create and train a large neural network which reads a sentence and produces a suitable translation. Most of the neural machine translation models suggested employ an encoder-decoder architecture, wherein there is an encoder and decoder for each language. These language-specific encoders are employed for each phrase. A fixed-length vector is created by reading and encoding a source text using an encoder neural network. Normally, encoder-decoder performance does slightly decrease as the sequence length of an input increases. The proposed model extends the encoder-decoder paradigm that independently learns to concentrate on contextual words during translation in order to overcome this issue. The proposed model embeds with set of position information with an input phrase which helps to understand the key information to be focused each time a generation of word during translation which is accomplished through Positional Embedding (PE).

   - **Positional Embedding (PE):** Information of each word is located in a sequence by appending positional encodings to the input embeddings. The positional encodings and word embeddings both have the same dimension, allowing the two to be combined. Sine cosine function is used to maintain periodicity of generating similar values for contextual words. The first step is to assign unique position number for each word at each time step within range [0,1] then calculate relative distance using equation 4, 5 and applies for alternate sequence.

$$PE(pos, i) = sin\ sin\left(\frac{pos}{10000^{\frac{i}{n}}}\right), \text{when i is even} \quad (4)$$

$$PE(pos, i) = cos\ cos\left(\frac{pos}{10000^{\frac{i}{n}}}\right), \text{when i is odd} \quad (5)$$

where i is the index of dimension, pos is the position of word in the input sequence. The model predicts a next word of target sequence based on all words which is generated previously and also with contextual vectors of sequence

- **Multi-Head Attention Mechanism:** NMT method consists of two components first is to encodes the source sentence (x_i) into internal representation and the second is to decode the target sentence (y_i) in a sequential order. The encoder consists of two residual connected sub layers are multi-head self-attention mechanism and then feed-forward network. Context vector ctxt and all of the predicted words as y_1, $y_2 \ldots y_{t-1}$ shown in Equation 6, the decoder is trained to predict the subsequent word, y_t.

$$p(y) = \prod_{t=1}^{T} p(y_t \mid \{y_1, \cdots, y_{t-1}\}, ctxt) \quad (6)$$

Each of the hidden state h_i includes details on the entire information of input sequence which is then multiplied with corresponding weights. Equation 7 represents the weighted sum of h_i is calculated to determine the context vector $ctxt_i$.

$$ctxt_i = \sum_{j=1}^{Tx} \propto_{ij} h_j \quad (7)$$

The weight α_{ij} of each h_j is computed in Equation 8:

$$\alpha_{ij} = \frac{exp\ (e_{ij})}{\sum_{k=1}^{Tx} exp\ (e_{ik})} \quad (8)$$

Where, α_{ij} represent the probability that the target word y_i which is derived from the source word x_j. Thus implements an attention mechanism in the decoder. The decoder chooses which part of the original sentence to be concentrated on. Thus, it makes the encoder with responsibility to encode all significant information of the source sequence into a fixed length embedded vector. As a result, the decoder's locations should pay attention to every position in the input sequence with the current word its decoding.

VII. EXPERIMENTAL RESULTS

It shows that deep learning framework such as LSTM classifier gives the best result. The Table-1 detailed with trained accuracy of 98.5% and a minimum trained loss of 0.04 at epoch 5 and batch size 12 whereas Machine learning framework such as ANN classifier obtains accuracy of 94.6 % and with loss of 0.18% by loop through for 10 epochs with batch size is of 32.

Table 1: Audio Processing and Its Classification Results

Research Work	Area	Classifier	Accuracy (%)
[13]	Language Identification	GMM	88.7
[14]	Speaker Identification	SVM	92.6
[15]	Speech Emotion Recognition	DBN	85.3
Proposed Approach1	Language Identification	ANN	94.6
Proposed Approach2	Language Identification	LSTM	98.5

As attention all you need [13] by google suggests input dimension of model to be 512 can obtain good accuracy which is then compared with dimension 256, the results for language pair shown in Table-2, which represents that gradually increase in accuracy and also reduce in training time by choosing embedding dimension as 512.

Table 2: Accuracy Obtained For Language Pair

Language pair	Embed-dim (256) Accuracy (%)	Embed-dim (512) Accuracy (%)
French-English	85.8	**90.1**
Spanish-English	85.1	**90.7**
Italian-English	85.2	**94.9**
English-French	76.1	**81.2**
English-Spanish	75.2	**78.6**
English-Hindi	86.2	**87.4**
Hindi-English	88.4	**93.1**
English-Italian	76.6	**78.6**

BLEU (Bi-Lingual Evaluation Understudy) is a measure utilized to assess the quality of machine-translated text, quantified as a numerical score ranging from zero to one. This score indicates how closely the machine-translated text aligns with a collection of exemplary reference translations. To determine similarity between the original sentence and the produced sentence, BLEU makes use of the fundamental principles of n-gram accuracy. The associated condition given in Equation 9 & 10 can be used to process the BLEU score:

$$\text{Precision score} = \frac{number\ \ of\ n-gram\ \ in\ candidate\ \ sentence\ \ appears\ \ in\ reference\ \ sentence}{total\ \ number\ \ of\ n-gram\ \ in\ candidate\ \ sentence} \quad (9)$$

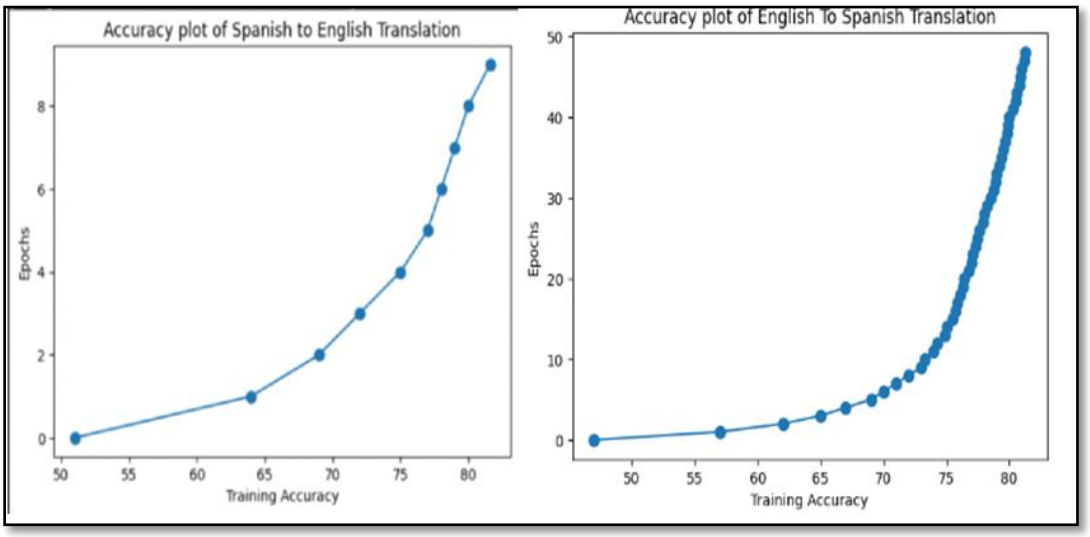
$$\text{Geometric mean precision} = (p_1 \times p_2 \times \ldots p_n)^{1/n} \quad (10)$$

P_i - precision score of i^{th} n-gram, n- total number of n-gram(1- 4). The findings of the result are shown in Table 3.

Table 3: BLEU n-Gram Score for Each Language Pair

Language Pair	One gram	Two grams	Three grams	Four grams
Spanish-English	**86.5**	80.2	75.9	72.1
French-English	**85.4**	76.9	71.9	67.7
Italian- English	**92.5**	88.1	84.8	81.8
English-Spanish	**81.5**	73.1	65.8	60.0
English-French	**81.0**	72.6	71.9	67.7
English- Italian	**87.3**	81.9	76.8	71.8
Hindi-English	**96.3**	93.9	92.5	91.2
English-Hindi	**91.5**	88.8	86.8	84.7

This proposed model was built by training with epoch of nearly around 50. As a visual plot shows that training of the translation model improves by increasing the epoch which is visually plotted in Figure 7 for all language pair translation model and Figure 8 represents the results of the proposed model which translates from source language as Spanish to intermediate language and then translate from intermediate language to target language as Italian.



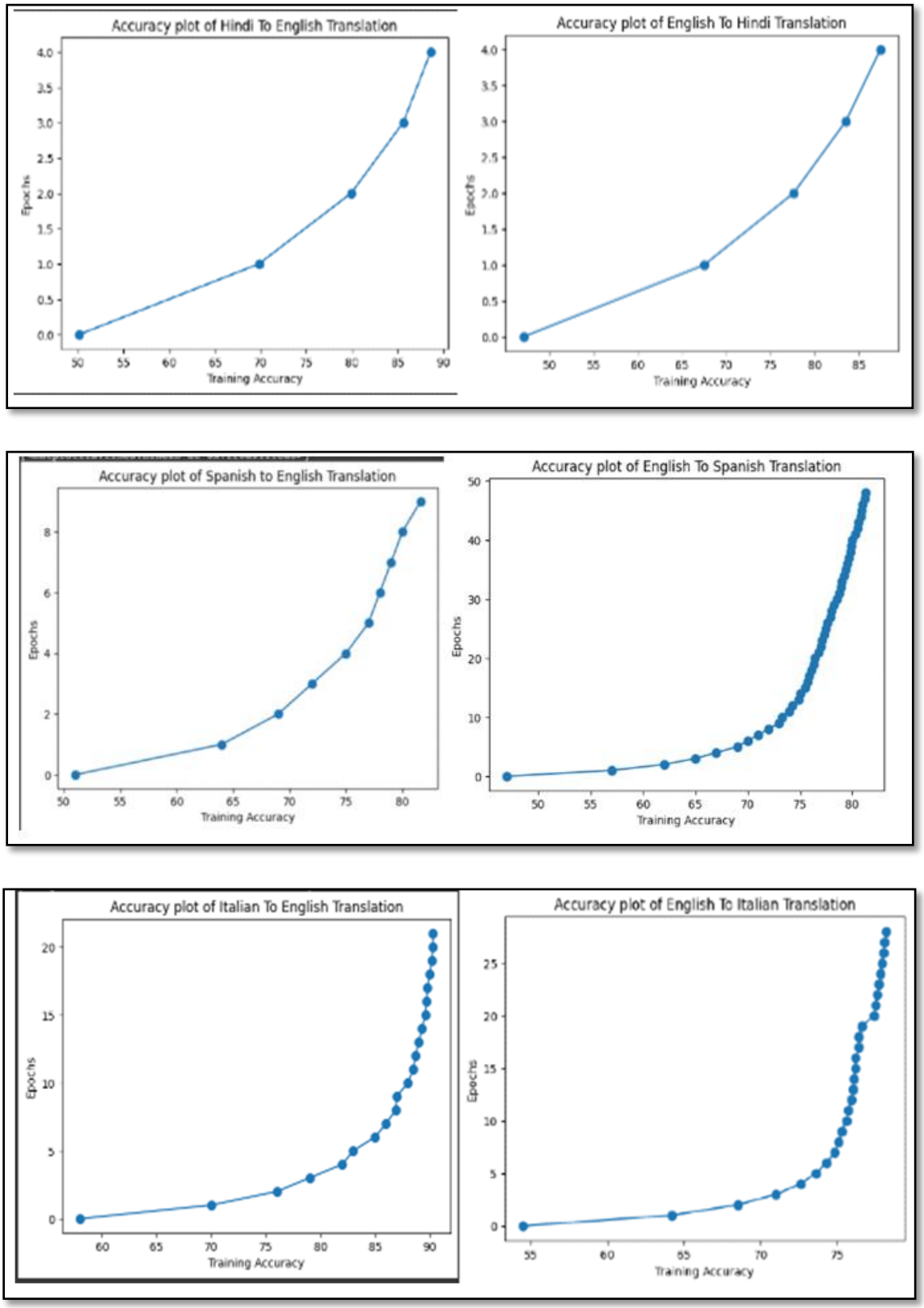


Figure 7: Accuracy Plot for Various Language-Pair Model

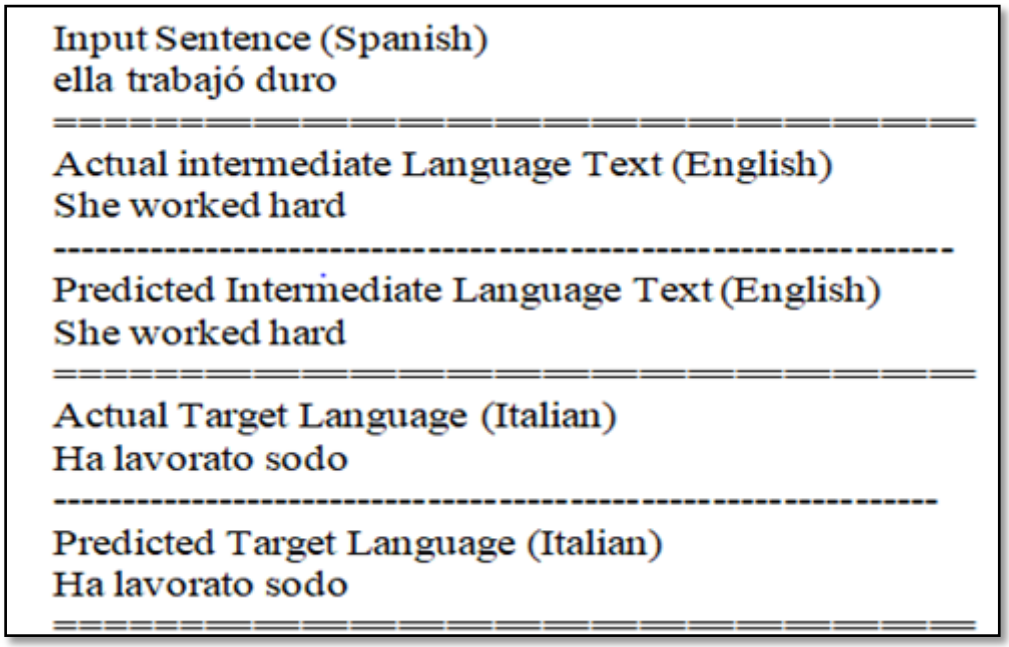


Figure 8: Translation from Spanish – English - Italian Sentence

VIII. CONCLUSION

This proposed methodology of LID for 5 different languages with extracted spectral features and classification done by LSTM, a deep learning framework with highest accuracy obtained of around 98% with loss of 0.04%. In further, experiment can be done to analyses the performance for varying speed of the audio and improve the performance of the model. Then for translation, processing can be completed quickly and effectively using conventional machine translation techniques. They have been demonstrated to be difficult in producing excellent results while having a limited operational capability. However, they have issues for the appropriate language, which is natural and in need of human intervention. Neural machine translation overcomes the limitations of conventional machine translation methods. Recent NMT models, such Seq-2-Seq, have produced the needed language with excellent results. Although the model dramatically affects its accuracy in several real-time situations particularly when it must run through unusual words. The most current NMT models, which create a context vector via an attention mechanism, overcome this problem. The accuracy of the model will remain unchanged even when it encounters unidentified words.

REFERENCES

[1] Pardeep Sangwan, Deepti Deshwal and Naveen Dahiya "Performance of a language identification system using hybrid features and ANN learning algorithms," Applied Acoustics, vol. 175, no. 8, pp. 107815, April 2021.

[2] Boussaid L, Hassine M "Arabic isolated word recognition system using hybrid feature extraction techniques and neural network," International Journal of Speech Technology, vol. 21, iss. 1, pp. 29-37, March 2018.

[3] Sekkate S., Khalil M. and Adib A. "A feature level fusion scheme for robust speaker identification," Proc. International Conference on Big Data, Cloud and Applications, Springer, pp. 289-300, August 2018.

[4] Bhanja C C, Laskar M A, Laskar R H "A pre-classification-based language identification for Northeast Indian languages using prosody and spectral features," Circuits Systems and Signal Processing, vol. 38, pp. 2266–2296, 2019.

[5] Hassine M, Boussaid L, Messaoud H. Maghrebian "Dialect recognition based on support vector machines and neural network classifiers," International Journal of Speech Technology, vol. 19, iss. 4, pp. 687–695, 2016.
[6] Maurya, Kumar, Agarwal "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach," Procedia Computer Science, vol. 125, pp. 880-887, 2018.
[7] Moselhy A M and Abdelnaiem A A "LPC and MFCC performance evaluation with artificial neural network for spoken language identification," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 6, no. 3, pp. 55-66, 2013.
[8] Andra and Usagawa "Improved Transcription and Speaker Identification System for Concurrent Speech in Bahasa Indonesia Using Recurrent Neural Network," IEEE Access, vol. 9, pp. 70760-70770, 2021.
[9] Priyank Mathur, Arkajyoti Misra, and Emrah Budur "Language Identification from Text Documents," Stanford University, 2016.
[10] Lui M, Lau J H, & Baldwin T "Automatic Detection and Language Identification of Multilingual Documents," In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1-10, 2014.
[11] Saha S, Chakraborty N, Kundu S, Nasipuri M. "Multi-lingual scene text detection and language identification using deep learning techniques," Pattern Recognition Letters, vol. 138, pp. 16-22, 2020.
[12] Zhang Y, Li J, Li X, and Liu X "A unified framework for multilingual automatic speech recognition and language identification," Journal of Signal Processing Systems, vol. 92, no. 7, pp. 1075-1084, 2020.
[13] Kano T, Sakti S and Nakamura "Transformer-based direct speech-to-speech translation with Transcoder," In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6224-6228, 2018.
[14] Tshephisho Joseph Sefara et al. "Transformer-based Machine Translation for Low-resourced Languages embedded with Language Identification," 2021 Conference on Information Communications Technology and Society, pp. 1-7, 2021.
[15] Pavan K, Tandon N and Varma, V "Addressing challenges in automatic language identification of Romanised text," 8th International Conference on Natural Language Processing, pp. 1-8, 2010.
[16] Bahar P, Brix C and Ney H "Two-Way Neural Machine Translation: A Proof of Concept for Bidirectional Translation Modeling Using a Two-Dimensional Grid," IEEE Spoken Language Technology Workshop, 2021.
[17] Nakamura S et al. "The ATR Multilingual Speech-to-Speech Translation System," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 2, pp. 365-376, 2006.
[18] Zhang R and Jin Y "Identification and Transformation of Comparative Sentences in Patent Chinese-English Machine Translation.," 2012 International Conference on Asian Language Processing, 2012.
[19] Pradeep J, Srinivasan E, Himavathi S "Performance analysis of hybrid feature extraction technique for recognizing English handwritten characters," World Congress on Information and Communication Technologies, 2012.
[20] G. Dahl, D. Yu, L. Deng and A. Acero "Context-dependent DBN HMMs in large vocabulary continuous speech recognition," Proc. ICASSP, 2011.
[21] N Jayanthi, Ch Suresh Kumar Raju "Dual Translation of International and Indian Regional Language using Recent Machine Translation," 3rd International Conference on Intelligent Sustainable Systems, 2020.
[22] A. Vaswani et al. "Attention is all you need," in Advancesin Neural Information Processing Systems, pp. 5998–6008, 2017.
[23] S. K. Dwivedi and P. P. Sukhadeve "Machine translation system in Indian perspectives," Journal of computer science, vol. 6, no. 10, 2010.
[24] Sumit Kushwaha "A Futuristic Perspective on Artificial Intelligence," 2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development, 2023.
[25] Sumit Kushwaha "Review on Artificial Intelligence and Human Computer Interaction," 2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development, 2023.

