MULTILINGUAL SPEECH RECOGNITION METHODS USING DEEP LEARNING AND COSINE SIMILARITY

P Deepak Reddy, Chirag Rudresh and Adithya A S

Department of Computer Science Engineering,
PES University, Bengaluru, Karnataka, India

ABSTRACT

The paper includes research on discovering new methods for multilingual speech recognition and comparing the effectiveness of the existing solutions with the proposed novelty approaches. The audio and textual multilingual dataset contains multilingual sentences where each sentence contains words from two different languages - English and Kannada.

Our proposed speech recognition process includes preprocessing and splitting each audio sentence based on words, which is then given as input to the DL translator (using MFCC features) along with next word predictions. The use of a Next Word Prediction model along with the DL translator to accurately identify the words and convert to text. Similarly the other approach proposed is the use of cosine similarity where the speech recognition is based on the similarity between word uttered and the generated training dataset. Our models were trained on an audio and textual dataset that were generated by the team members and the test accuracies were measured based on the same dataset.

The accuracy of our speech recognition model, using the novelty method, is 71%. This is a considerably good result compared to the existing multilingual translation solutions.

Communication gap has been a major issue for many natives and locals trying to learn or move ahead in this tech-savvy English-speaking world. To communicate effectively, it is not only essential to have a single language translator but also a tool that can help understand a mixture of different languages to bridge the gap of communication with the non-English speaking communities. Integrating a multilingual translator with the power of a smart phone voice assistant can help aid this process.

KEYWORDS

Natural Language Processing, Deep Learning, Multilingual Speech Recognition, Machine Learning, Speech to Text.

1. INTRODUCTION

In the developing country of India, it has become a common trait amongst the people, to speak in a mixture of their native language and English. The domain of recognising and translating multilingual speech is still a very less researched topic but there are a few top contenders who lead the market in translating and recognising monolingual sentences. Apart from the fact that these tools are not currently able to efficiently and accurately handle a mixture of multiple

David C. Wyld et al. (Eds): DMML, SEAS, ADCO, NLPI, SP, BDBS, CMCA, CSITEC - 2022
pp. 15-24, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120702

languages in a single speech or text query, there are other drawbacks to the current models and tools.

In order to accurately solve the above issues and drawbacks, we have formulated a problem statement to tackle these limitations and solve the problem of multilingual speech recognition which is mentioned below.

- Recognizing the various languages used in a multilingual sentence and translating it into a single language sentence.
- Developing a model that comprehends multilingual voice navigation queries by recognizing the various languages used in a multilingual sentence and reply accordingly with a single language sentence.

India has 22 official languages and most of these are constrained to a specific state and are not spoken much outside these states. Hence, these languages are low-resourced and it is not easy to come across speech data for these languages. Adding to this, there does not exist any readily available multilingual dataset in combination with English for these languages.

One of the most popular translation and speech recognition tools currently available is the google api and its performance was tested on multilingual speech query containing Kannada and English. When the input language is chosen as English, the translator aims to translate the words using the English dictionary vocabulary. Thus, when another language (for example Kannada) is used in the sentence, that foreign word is mapped to the closest sounding English word irrespective of the meaning. For example: When the input speech is ‘How to go to Shaale’ all the English words are recognised correctly but the Kannada word ‘Shaale’ is mapped to the closest sounding English word ‘Charlotte’.

In our proposed approach, the model predicts each word based on the context and hence the entire sentence predicted so far is taken into consideration rather than recognising each word as a discrete component.

The challenges of multilingual speech recognition include the scarcity of data set for effective training and testing of the model. Limited existing literature for understanding the domain of multilingual speech recognition. One of the main obstacles is to convert the audio query to multilingual text.

2. LITERATURE REVIEW

[1]The paper is related to Speech Recognition using the LAS model, which uses the internal representations of the languages learnt by the model during training. The paper describes this method to perform better than existing single language translation and recognition models, as it combines the inferences drawn from training each language separately and then combining them to recognize monolingual sentences of various languages.

Although the paper is not directly related to our problem statement of multilingual speech recognition, the methodology used for combining multiple trained model pipeline gives us an idea of how to use DL models to train and test based on multiple language sentences. The paper has scope with respect to research on performance and working of existing speech recognition tools like Google API, Python libraries, etc and can further extend the use case towards solving the problem of multilingual translation and recognition using the same described model with a few tweaks. One major drawback of the paper’s described method, is that the models use the internal representations of each language to recognise the words spoken, whereas in reality the

languages vary in script and dialogue which are more practically applied for differentiating and recognizing the words.

[2]The paper is related to dynamic language identification and focused on the use case of a software that will help in the text-to-speech feature for applications that are developed for people who are visually challenged or have reading disabilities. This helped us in formulating the use case for our model which is regional voice assistant that can convert multilingual audio query to a single language query. In order to achieve this it was understood that language recognition is an important feature that is required for multilingual text-to-speech conversion. It is because the algorithms used in this process are different from those used in automatic language detection, since the recognition is done non synchronously on a continuous stream of texts. It mainly focused on the software component of multilingual text-to-speech. The results further indicated that for language detection algorithms, fragmentation of a piece of text is an important parameter. Tri grams provided better accuracy in language recognition as compared to single or bigram. But the limitation of this approach of changing language for another text is that since most of the users of this application are visually challenged, manually changing voice in the audio menu by following voice guidance was difficult and really time-consuming. So our proposed solution intends to build a single model that can understand the multilingual language queries.

[3]Paul Fogarassy and Costin Pribeanu in their paper 'Automatic Language Identification Using Deep Neural Networks, explored the performance of deep neural networks on the problem of Language identification. This deep neural network model works on the features extracted from short speech utterances. It was found that the proposed model using extracted form of short speech utterances outperforms the current state-of-the-art i-vector based acoustic model. From the research it was found out that when the data-set is large, the deep neural networks perform the language identification better.

The DNN outperforms the state-of-the-art models in most cases. This is when the training data for each language is more than 20 hours.

Similar approaches for our research problem may not work as desired as it is found that it is better to directly recognise the next word instead of trying to identify the language and then recognise the word.

3. DATA GENERATION

184 most common English queries were first generated but only 131 navigation related queries are chosen for this research purpose. For each of these English queries, all the possible multilingual sentences were framed resulting in a total of 412 multilingual sentences which were then POS tagged with 7 classes from English and 7 classes from Kannada. Most commonly used sentences were identified from this set for generating the speech data. A total of 64 words are picked and were then recorded by 3 different individuals and each word is recorded 10 times by each individual resulting in a total of 1920 recordings.

4. METHODOLOGY

The approach used for recognition and translation of multilingual audio query is as follows:

- The input audio wav file containing the query is split into individual wav files each containing the individual words of the query.

- These wav files are then passed to a predictive model that uses a deep learning model to map the audio to text and generate the text output for the corresponding words.
- The accuracy of the speech-to-text model is further increased using a next-word prediction model and a POS tag prediction model.
- Both these prediction models take in a sequence of words, tags respectively and use a RNN to generate the next possible 'n' words, tags that follow.
- These words, tags are then used to decrease the search space for the speech-to-text conversion model for better results.

The multilingual text query is passed through to Google Translation API to get the corresponding monolingual query which is then passed to the Search Engine to get the appropriate output. The entire process of recording the audio query to display the results is integrated into a user-friendly application.

The main constraint of this method is that the training audio and textual multilingual query dataset needed for the speech-to-text conversion model is very high (in terms of hundreds of thousands), which is not feasible with the team size and the time constraint. But this can be overcome, by expanding the dataset by generating recordings of different age groups, gender and language dialects.

This methodology also depends on the performance of the Google translation API and the Search Engine for the accuracy of the translation and output. The application depends on the storage constraints of the Database used as well as the limit on the amount of audio and textual queries that can be stored.

5. IMPLEMENTATION

5.1. Preprocessing

The wav files were converted to a numpy array using librosa where each value in the array represents the amplitude and the array was normalised between values -1 to 1. The silence factor that existed due to delay in pressing the record audio start/stop button in the beginning and end of the audio file was removed. The speed of audio files were changed by changing its frame rate such that the size of each file is 20000 numpy array length when read by librosa since each user can speak a particular word at different speeds.

5.2. Splitting of sentence

After careful analysis of a few recorded sentences, it was found that each individual word utterance was between 15,000 and 25,000 array length. It was observed that the amplitude is low between each utterance of words. Hence amplitude is used as a factor to split the audio file. Moving Average with a window size of 10,000 is used to smoothen the wav file and to clearly identify the minimas. Then, minimas were found in the smoothened signal at a window size of 15,000 array length. Thus when the original signal is sliced at these minimas, individual word utterances are obtained.

5.3. Word Predictor

[5] The Word Predictor model uses the concept of LSTM to take bags of words as input and predict the next possibly occurring words. LSTM uses the memory of previously occurring words and learns the weights of next occurring words, thus using this knowledge the model is able to

deduce the next possible words from the trained vocabulary. Sentences were tokenized and all n-gram (n=4) sequences were generated. The first three tokens were considered as features which were used to predict the fourth word. These sequences were passed to the LSTM model as input to generate the next top 'k' models. Since there was an ambiguity of prediction of first and second word, similar LSTM models with n=2 and n=3(bigram and trigram) predictors were also built.

5.4. Speech To Text

5.4.1. Methodology 1 : Deep Learning

This module receives a set of next possible words (classes) from the word predictor model and classifies the input chunk into one of these words. Pre-processed wav files of these classes were selected to train the model. The extracted mfcc (Mel Frequency Cepstral Coefficient) features were used for training the model. The neural network contains an input layer, two hidden layers and an output layer. The input layer contains 100 layers, the first hidden layer contains 200 neurons activated with ReLU, the second hidden layer contains 100 neurons activated with ReLU and the final output layer contains five neurons which is equal to the number of next possible words (given by the next word prediction module). Softmax activation function is used on the final layer.

5.4.2. Methodology 2 : Deep Learning

Cosine similarity is found between the input chunk and among all the recordings of next possible words and highest occurring class among the top 20 most similar recordings is then predicted as the next occurring word.

6. RESULTS AND DISCUSSIONS

6.1. Splitting of sentence

The splitting algorithm was tested on 30 sentences out of which 28 were correctly splitted into respective words as shown in the table. The incorrectly splitted sentences were re-recorded with sufficient gaps between words after which the splitting was done accurately.

6.2. Word Prediction

Word Predictor model gave 90% accuracy and when predicted top 5 possible next words for a current sentence, the desired word was present in this predicted possible words.

6.3. Speech to text

6.3.1. Methodology 1: Deep Learning

The average accuracy for each sentence was calculated by taking the accuracy of models while predicting each individual word of that corresponding sentence. Model Accuracy is the average accuracy of all the sentences tested. Prediction Accuracy is the number of words correctly predicted divided by the total number of words present.

6.3.2. Methodology 2 : Based on similarity of signal

Average similarity of each class is also found and the class with highest average is then predicted as the next occurring word. The second approach was to find similarity between the input chunk and among all the recordings of next possible words. Highest occurring class among the top 20 most similar recordings is then predicted as the next occurring word. The accuracy for both the methods is as shown in the table where it was calculated by taking the ratio of number of correctly predicted words by total number of words in the sentence.

The average accuracy of 0.59 was achieved when prediction was done using average similarity of each class and 0.64 was achieved when using the highest occurring class among the top 20 most similar recordings.

7. NOVELTY APPROACH

The concept of using a multilingual Next Word Prediction model in accordance with the DL translator is a novel approach used to tackle the problem of translation.

The input to the Word Predictor is a sequence of textual multilingual words, that is used to train the predictor to analyse and predict the next possible 5 words using the knowledge of prior occurrence of words in sentences. These 5 words are then provided as input to the DL method, which uses these 5 words to compare and figure out the word utterance rather than comparing it with the entire vocabulary.

This concept helps decrease the time for translation by reducing the DL translator search corpus and also increase the accuracy of the translation.

8. FIGURES AND TABLES

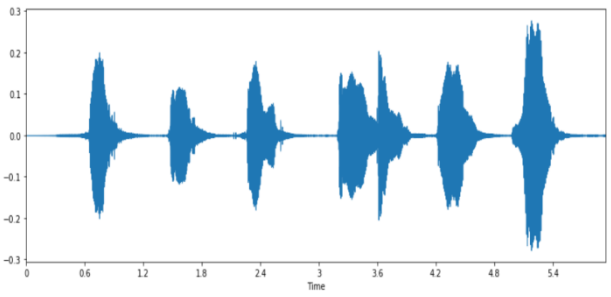

Figure 1. wav file of audio query recording

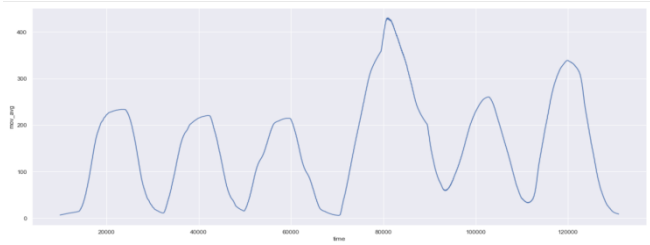

Figure 2. Wav file after smoothening which helps clearly identify individual words present in the audio query

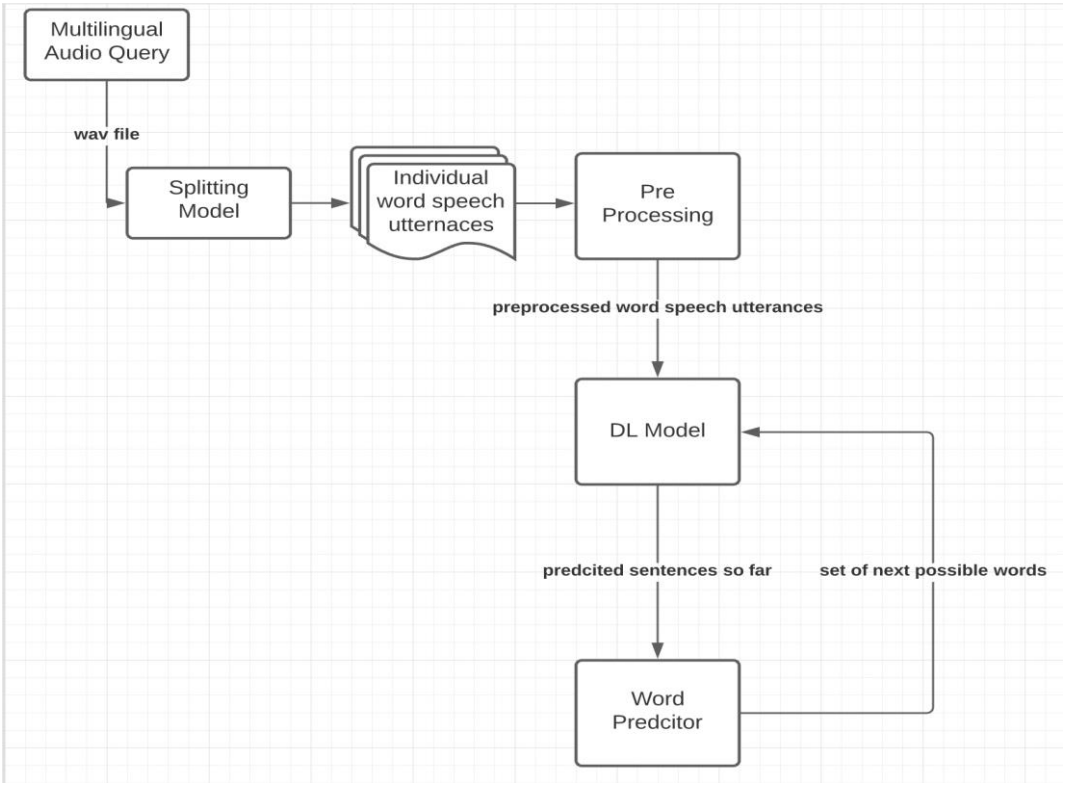

Figure 3. Flow chart of proposed methodology

Table 1. Test results of splitting module

Tested Sentence	Actual Number of Words	Predicted Number of Words
nanna next turn yenu	4	4
what is my next saradi	5	5
what is my mundina saradi	5	5
how is the datthane ahead	5	5
munde traffic hegidhe	3	3
how is the datthane munde	5	5
what are the nearby anila stations	6	6
what are the nearby gas kendragalu	6	6

Table 2. Test results of speech to text conversion using deep learning

Actual Sentence	Predicted Sentence	Average Accuracy
nanna next turn yenu	nanna next turn yenu	0.83
what is my next saradi	what is my next saradi	0.71

what is my mundina saradi	what aagothe any stalagalu saradi	0.68
how is the datthane ahead	how service the yaavaaga ahead	0.73
munde traffic hegidhe	munde traffic hegidhe	0.78
how is the datthane munde	how service the yaavaaga ahead	0.79
what are the nearby anila stations	what does the nearby anila stations	0.84
what are the nearby gas kendragalu	what aagothe the nearby gas kendragalu	0.84
what are the nearby anila kendragalu	what aagothe the nearby good kendragalu	0.85

Table 3. Test results of speech to text conversion using similarity of signal

Actual Sentence	**Similarity Predicted Sentence**	**Weights Predicted Sentence**	**Similarity Accuracy**	**Weights Accuracy**
nanna next turn yenu	nanna next turn yenu	nanna next turn yenu	1	1
what is my next saradi	what is any next yenu	how is my next saradi	0.6	0.8
what is my mundina saradi	are is any mundina yenu	what is my mundina yenu	0.4	0.8
how is the datthane ahead	what is some yaavaaga ahead	what aagothe my nearest ahead	0.4	0.2
munde traffic hegidhe	are traffic hegidhe	are traffic hegidhe	0.67	0.67
how is the datthane munde	are is my nearest munde	what is some nearest munde	0.4	0.4
what are the nearby anila stations	what is the nearby gas stations	how is some nearby gas kendragalu	0.67	0.17
what are the nearby gas kendragalu	what is some nearby gas kendragalu	what is some nearby gas kendragalu	0.67	0.67
what are the nearby anila kendragalu	what is the nearby gas kendragalu	what is the nearby gas kendragalu	0.67	0.67
hathira gas stations yavdu	what iro stations yavdu	what gas stations yavdu	0.5	0.75

Table 4. Accuracy of pre-existing models

Pre existing models	**Accuracy**
CMU Sphinx(HMM model trained and tested with our data)	57%
Similarity measure using Neural Network	64%
Google Translate(Kannada words recognition)	35.4%

Table 5. Accuracy of our deep learning and similarity models

Our Proposed Models	Accuracy
Deep learning model(using MFCC features)	71%
Deep learning model(using MFCC features, for kannada words recognition)	66.6%
Similarity model(using highest average similarity of each class)	0.59%
Similarity model(using the highest occurring class among the top 20 most similar recordings.)	0.64%

9. CONCLUSIONS

The methodology proposed in this paper is a completely novel approach which uses the self-learning abilities of a model to recognize and translate the multilingual queries to a monolingual query, as accurately as possible. The Deep Learning model uses the top predictions from the Word Predictor model to reduce the search space while identifying and translating each word input from the audio query. The new deep learning model, when tested on the generated multilingual dataset, gives an accuracy of 85%. And when it is tested live by the user, it gives an accuracy of 71%.For cosine similarity model The average accuracy of 0.59 was achieved when prediction was done using average similarity of each class and 0.64 was achieved when using the highest occurring class among the top 20 most similar recordings.

10. LIMITATIONS

The main drawback of our proposed method is that, the DL model runs every single time that a new word is predicted and given as input to the DL model, thus for recognising one sentence the model can be quite time consuming. Another limitation of our method is that it is heavily dependent on the performance and accuracy of the Word Predictor model, since the output of the next possible words is given as the input to the DL model. Since the predicted words do not have a probability attached to them, it is not possible to obtain a metric about the certainty of occurrence of the predicted words in the sentence.

11. FUTURE WORK

There are a few improvements that can be added which could further increase the accuracy of our proposed model :

- Increasing the data set, both audio and textual, with a variation in voice such as age, gender, noise level, etc.
- The similarity approach can be improved using similarity index comparison of the spectrograms of the words, instead of the previously mentioned method which compares the wav forms.
- The prediction of words and their parts of speech in a particular sentence by leveraging different and more useful language features

ACKNOWLEDGEMENTS

We would like to express our gratitude to PES University for providing us with continuous support and encouragement.

REFERENCES

[1] Multilingual Speech Recognition with a single end-to-end model - Shubham Toshniwal, 15th February 2018
[2] Multilingual Text-to-Speech Software Component for Dynamic Language Identification and Voice Switching ,September 2016 Paul Fogarassy,Costin Pribeanu
[3] Automatic Language Identification Using Deep Neural Networks,2016, Ignacio Lopez-Moreno, Javier Gonzalez , Dominguez, Oldrich Plcho.
[4] The research of feature extraction based on MFCC for speaker recognition,2014,Zhang Wanli,Li Guoxin
[5] LSTM Neural Networks for Language Modelling,2012,Martin Sundermeyer, Ralf Schlu ̈ter, and Hermann Ney

AUTHORS

P Deepak Reddy
Final Year Student Engineering studying at PES University, Bengaluru.

Chirag Rudresh
Final Year Student Engineering studying at PES University, Bengaluru.

Adithya A S
Final Year Student Engineering studying at PES University, Bengaluru.