MULTILINGUAL SPEECH TO TEXT USING DEEP LEARNING BASED ON MFCC FEATURES

P Deepak Reddy, Chirag Rudresh and Adithya A S

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

ABSTRACT

The proposed methodology presented in the paper deals with solving the problem of multilingual speech recognition. Current text and speech recognition and translation methods have a very low accuracy in translating sentences which contain a mixture of two or more different languages. The paper proposes a novel approach to tackling this problem and highlights some of the drawbacks of current recognition and translation methods.

The proposed approach deals with recognition of audio queries which contain a mixture of words in two different languages - Kannada and English. The novelty in the approach presented, is the use of a next Word Prediction model in combination with a Deep Learning speech recognition model to accurately recognise and convert the input audio query to text. Another method proposed to solve the problem of multilingual speech recognition and translation is the use of cosine similarity between the audio features of words for fast and accurate recognition. The dataset used for training and testing the models was generated manually by the authors as there was no pre-existing audio and text dataset which contained sentences in a mixture of both Kannada and English.

The DL speech recognition model in combination with the Word Prediction model gives an accuracy of 71% when tested on the in-house multilingual dataset. This method outperforms other existing translation and recognition solutions for the same test set.

Multilingual translation and recognition is an important problem to tackle as there is a tendency for people to speak in a mixture of languages. By solving this problem, the barrier of language and communication can be lifted and thus can help people connect better and more comfortably with each other.

KEYWORDS

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

1. INTRODUCTION

In growing countries like India, which is a land of over 22 different languages and numerous other dialects, the people have developed the habit of speaking with a mixture of their local language with English. The ability to recognize the languages used in the multilingual sentence and translating them into a single language is an essential task for better communication. The existing solutions for translation cannot effectively differentiate the various languages used in a multilingual sentence and have certain drawbacks to their functionality.

We have established a problem statement that addresses these constraints of multilingual speech recognition which aim to resolve the drawbacks of current solutions. The general intuition behind tackling this problem is as follows:

- Recognize that the sentence contains more than a single language and distinguish the various words present in the sentence for ease of translation.
- Converting the audio query to a textual sentence and then recognising and translating the multilingual textual query.

A popular speech and text recognition and translation tool present is the Google API. There were several drawbacks and constraints observed when this tool was tested on the in-house multilingual dataset and general multilingual speech queries. The performance of the API is heavily dependent on the language chosen for translation and recognition. For example, when the language selected was English, when an audio and a text multilingual query was provided as an input which contained a mixture of Kannada and English words, the API tried matching the Kannada words to the closest sounding English word. For instance, 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'. Thus the selection of language is an essential criteria for the performance of the Google API.

The proposed method for language recognition uses the language context by using prior information of words present and their meaning to the sentence. This eradicates the problem of recognizing and translating out-of-context words for the multilingual speech query.

A major hurdle for tackling this problem is the scarcity of available datasets, especially local languages in a mix with English. Another important task is the conversion of audio query to textual representation for better translation and recognition.

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

Initially, 184 of the most common English queries were created, but only 131 navigation-related queries were chosen for this study. All potential multilingual sentences were created for each of these English queries, resulting in a total of 412 multilingual sentences that were then POS tagged with 7 classes from English and 7 classes from Kannada. The most frequently used sentences were chosen from this group to generate the speech data. A total of 64 words were chosen and recorded by three different people, with each word being recorded ten times by each person, totaling 1920 recordings.

4. METHODOLOGY

The following is the method adopted for recognising and translating multilingual audio queries:

- 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 sent into a prediction model, which use a deep learning model to translate the audio to text and generate text output for the words.
- A next-word prediction model and a POS tag prediction model are used to improve the accuracy of the speech-to-text 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.
- For improved outcomes, these words and tags are utilised to reduce the search space for the speech-to-text conversion model.

The multilingual text query is sent to the Google Translation API, which returns a monolingual query, which is then sent to the Search Engine, which returns the relevant result. The entire process is integrated into a user-friendly application, from recording the audio query to displaying the findings.

The fundamental limitation of this method is that the training audio and textual multilingual query dataset required for the speech-to-text conversion model is enormous (hundreds of thousands), which is impossible to achieve given the team size and time constraints. However, this can be addressed by increasing the dataset by recording people of all ages, genders, and dialects.

The accuracy of the translation and output is also dependent on the performance of the Google translation API and the Search Engine. The application is dependent on the database's storage constraints as well as the maximum number of audio and textual queries that can be stored.

5. IMPLEMENTATION

5.1. Preprocessing

The audio files were transformed into an array where each value of the array represents the amplitude of the audio file. Since users can speak in different words, this array was normalised between -1 to 1 so that all the voice amplitudes will be of the same level. There was a possibility of a silence factor existing between the start of the audio file and the end which was also removed. Since people can speak at different speed levels, the speed of audio files were changed by making all the arrays to a size of 20,000

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

The word predictor model provides a list of next probable words (classes) to this module, which it uses to classify the input chunk into one of these words. To train the model, pre-processed wav files from these classes were used. The model was trained using the extracted mfcc (Mel Frequency Cepstral Coefficient) features. An input layer, two hidden layers, and an output layer

make up the neural network. The input layer has 100 layers, the first hidden layer has 200 ReLU-activated neurons, the second hidden layer has 100 ReLU-activated neurons, and the final output layer has five neurons, which is the number of possible words after that (given by the next word prediction module). On the last layer, the Softmax activation function is utilised.

5.4.2. Methodology 2 : Based on Similarity of Signal

Taking cosine similarity of the input chunk with the training data set, the highest occurring class among the top 20 most similar recording was predicted as the next occurring word

6. RESULTS AND DISCUSSIONS

6.1. Splitting of Sentence

Each of the 30 sentences in the training data set was tested on the algorithm out of which 28 sentences were splitted correctly. The other 2 sentences after recording with sufficient gaps in between the words was also splitted properly by the algorithm

6.2. Word Prediction

The model accuracy of Word Predictor was 90% and when asked to predict the top 5 possible words, the results were as expected

6.3. Speech to Text

6.3.1. Methodology 1 : Deep Learning

The Model accuracy was derived by taking the accuracy of the model for each sentence and then averaging the same for all the sentences tested. Prediction accuracy was taken as the number of words correctly predicted divided by the total number of words present

6.3.2. Methodology 2 : Based on Similarity of Signal

Each class had 6 types of recordings and average similarity was taken for that particular class. Highest average among all the classes was predicted as the next occurring word. In the second method top 20 similar recordings were taken and the class having highest similarity was predicted as the next sentence. The accuracy of the model was derived by taking the number of correctly predicted words to the total number of words present.

Accuracy of 59% was achieved by taking average similarity of each class and 64% was achieved by taking the class having highest frequency among top 20 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

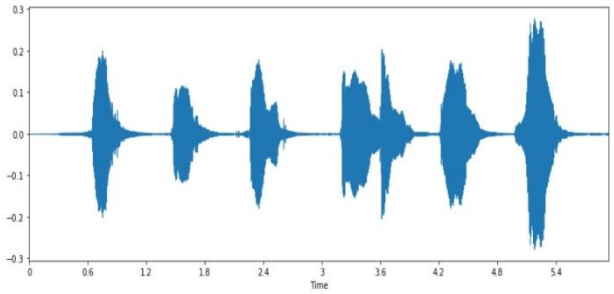


Figure 1: wav file of audio query recording

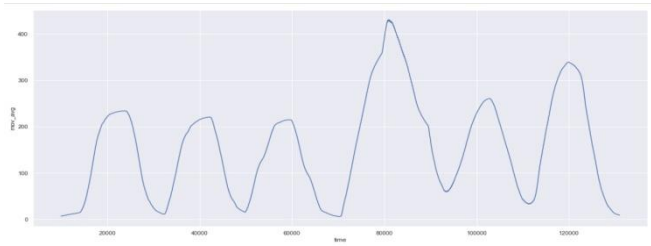


Figure 2. Wav file after smoothening which helps clearly identify individual words present in the audio query

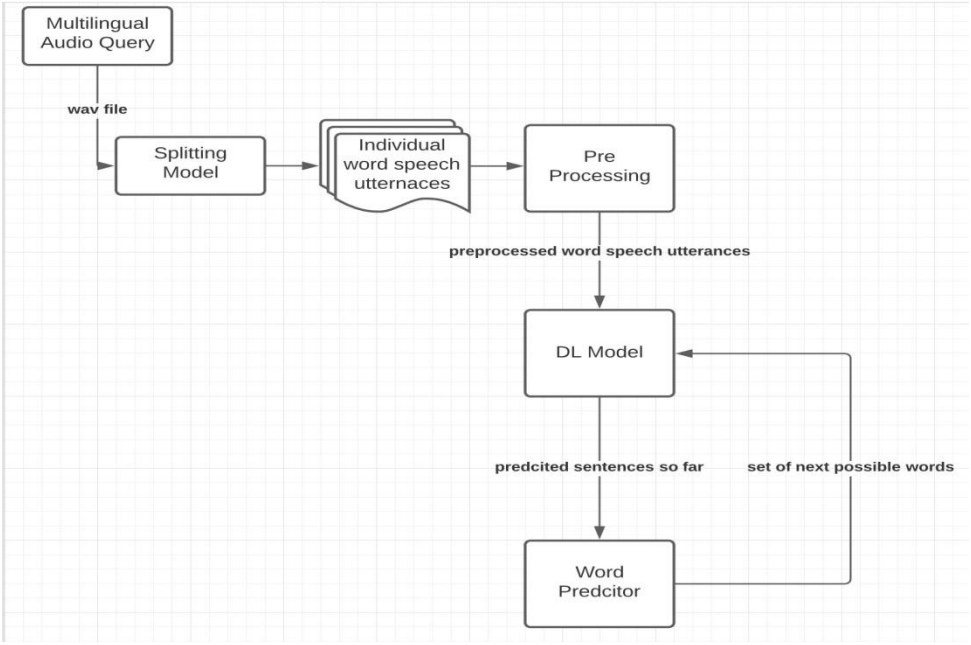


Figure 3. Flow chart of proposed methodology

Table 1. Test results of splitting module

Tested Sentence	Actual Number of Words	Predicted Number of Words
When does the nearest udhyaana open	6	6
hathiradha udhyaanavu yaavaaga open aagothe	5	6
Nanna hathira iro plumbing service	5	5
Hathira iro kolaayi service	4	4
Hathira iro plumbing service	4	5
hathiradha job openings	3	3
Udhyogavakaasha near me	3	3
Kelasa openings near me	4	4

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

Actual Sentence	Predicted Sentence	Average Accuracy
hathiradha park yaavaaga thereyuthadhe	hathira park yaavaaga thereyuthadhe	0.95
when does the nearest udhyaana open	when there the nearest open openings	0.84
hathiradha udhyaanavu yaavaaga open aagothe	are udhyaanavu yaavaaga open aagothe	0.86
nanna hathira iro plumbing service	hathira traffic iro ground service	0.85
hathira iro kolaayi service	nanna iro kolaayi service	0.87
hathira iro plumbing service	are iro plumbing service	0.84
hathiradha job openings	hathira park openings	0.87
udhyogavakaasha near me	udhyogavakaasha needalu me	0.94
kelasa openings near me	what openings me me	0.86

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
bheti needalu hathiradha good places yavuvu	nanna near hathiradha good turn yavuvu	hathiradha near hathiradha good places yavuvu	0.5	0.67
hathiradha park yaavaaga thereyuthadhe	what park anila thereyuthadhe	hathiradha park any udhyaana	0.5	0.5
when does the nearest udhyaana open	when are the iro udhyaana open	when are the iro udhyaana open	0.67	0.67
hathiradha udhyaanavu yaavaaga open aagothe	are udhyaanavu yaavaaga grounds aagothe	are udhyaanavu yaavaaga grounds aagothe	0.6	0.6
nanna hathira iro plumbing service	nanna hathira iro plumbing aagothe	nanna hathira iro plumbing aagothe	0.8	0.8
hathira iro kolaayi service	hathira iro kolaayi yavdu	hathira iro kolaayi yavdu	0.75	0.75
hathira iro plumbing service	hathira iro plumbing service	hathira iro plumbing service	1	1
hathiradha job openings	are udhyaanavu openings	are job openings	0.33	0.67
udhyogavakaasha near me	are near near	are near me	0.33	0.67
kelasa openings near me	nanna openings near near	nanna openings near me	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 suggested in this study is a completely innovative strategy that relies on a model's self-learning skills to recognise and effectively translate multilingual queries to monolingual questions. 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 key disadvantage of our suggested strategy is that the DL model runs every time a new word is predicted and given as input to the DL model, making it rather time consuming to recognise a single sentence. Another drawback of our strategy is that it is largely reliant on the Word Predictor model's performance and accuracy, as the output of the next probable words is fed into the DL model. Because the predicted words do not have a probability associated with them, a metric for the certainty of occurrence of the predicted words in the phrase cannot be calculated.

11. FUTURE WORK

There are a few adjustments that might be made to our suggested model to improve its accuracy even more:

- Increasing the data set, both audio and textual, by varying the voice in terms of age, gender, noise level, and other factors.
- Instead of comparing the wav forms, the similarity methodology can be improved by comparing the spectrograms of the words using a similarity index comparison.
- Prediction of words and their parts of speech in a given sentence based on a variety of more useful language aspects

ACKNOWLEDGEMENTS

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

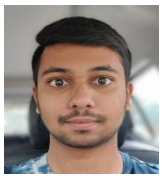
REFERENCES

[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.