International Journal of Engineering Sciences & Emerging Technologies, Dec. 2024
ISSN: 22316604 Volume 11, Issue 4, pp: 902-906 ©IJESET

ON FLY CONVERSION

Anurag Malik[1], Sarthak Srivastava[2], Sahil Rastogi[3,], Om Bajpai[4], Pratham Vyas[5]

[1]Associate Prof. , CS&E Department, Moradabad Institute of Technology, India
anurag_malik@rediffmail.com
[2,3,4,5] B.Tech Scholar, CS&E Department, Moradabad Institute of Technology, India
[2]sarthak.0591@gmail.com, [3]sahilrastogi817@gmail.com, [4]ombajpai911@gmail.com,
[5]prathamvyas2010@gmail.com

ABSTRACT

The proposed method focuses on understandingaudio queries that have both Kannada and English words. It uses a special Word Prediction model along with a Deep Learning speech recognition model to accurately convert these audio queries into text. Another method suggested is to use cosine similarity to quickly and accurately recognizethe words in different languages. Since there were no existing datasets with English, Marathi, Hindi, Bengali, Punjabi, Malayalam, etc sentences, the authors created their own dataset. When tested, this method achieved a 71% accuracy rate, outperforming other translation and recognition solutions. Multilingual translation and recognition are challenging because people often mix languages when they speak. Solving this problem could helpbreak down language barriers and make communication easier between people.

KEYWORDS

Deep learning, LSTM, word predictor, LAS

1 INTRODUCTION

In countries like India, where many languages are spoken,people often mix their local language with English when they talk. It's important to be able to understand and translate these mixed-language sentences to communicate well. However, current translation tools, like the Google API, have some issues. For example, if you ask a questionin a mix of Kannada and English, the API might try to match the Kannada words to similar-sounding Englishwords, leading to mistakes. So, choosing the right languageis crucial for the API to work well.

To solve this problem, we're proposing a method that looks at the context of the sentence to understand which language is being used. This helps avoid mistakes in translation. But one big challenge is that there aren't enough datasets available, especially for sentences mixing local languages (Hindi, Begali, Marathi, Malayalam, Punjabi etc) with English. Also, it's important to turn spoken queries into written text accurately for better translation and understanding.

2 LITERATURE REVIEW

[1] This paper talks about a Speech Recognition model called LAS. It's really good at understanding languages it was trained on, even better than other models that focus on just one language. It learns from each language separately and then combines that knowledge to understand sentences in different languages. Although the paper doesn't directly talk about recognizing multiple languages in one sentence like we need, it gives us some good ideas on how to use similar models to solve our problem. However, it's important to note that this model relies on understanding the specific details of each language, which might not work well for all languages.

[2] This paper discusses software that helps people who have trouble reading by converting text to speech. It helped us understand the importance of recognizing languages for our voice assistant model. The software focuses on converting text to speech in multiple languages, but changing languages in the software can be hard for visually impaired users. So, our goal is to build a single model that can understand and process multilingual queries without the need for users to manually switch languages.

[3] In this paper, researchers looked into using Deep Neural Networks to identify languages. They found that their model, which analyzes short speech clips, works better than other methods. They also found that having a large dataset helps the model identify languages even better.

3. DATA GENERATION

Originally, we came up with 184 common English questions, but we only picked 131 questions related to finding directions for this study. For each of these questions, we created possible sentences in multiple languages, resulting in 412 sentences. We then labeled these sentences with different parts of speech (like nouns, verbs, etc.) in English and Kannada. We focused on the most commonly used sentences to create recordings. We chose 64 words and had three different people record each word ten times, giving us a total of 1920 recordings.

4. METHODOLOGY

Here's how we recognize and translate multilingual audio queries:

1. We take the audio file containing the query and break it down into separate files, each representing a single word.
2. These files are then fed into a prediction model that uses deep learning to turn the audio into text, giving us the words in the query.
3. To make sure we get the words right, we use two prediction models: one predicts the next word in the sequence, and the other predicts the part of speech (like noun, verb, etc.) for each word.
4. These prediction models use a type of neural network called Recurrent Neural Networks (RNNs) to make guesses about what the next words or parts of speech could be.
5. We use the predictions from these models to help improve the accuracy of the speech-to-text conversion.

After we have the text of the query, we send it to the Google Translation API to translate it into a single language. Then, we send this translated query to a search engine to find relevant results. This whole process is designed to be user-friendly, from recording the audio query to showing the search results.

One challenge we face is that we need a lot of examples to train the model to understand different languages well. Getting such a large dataset can be difficult, but we can try to overcome this by collecting recordings from a diverse group of people, including different ages, genders, and dialects.

5. IMPLEMENTATION

5.1. Preprocessing:

- We converted audio files into arrays, where each element represents the loudness of the audio.

- To make sure all voices are consistent, we adjusted the array to have values between -1 and 1.

- We removed any silent parts at the beginning or end of the audio.

- Since people speak at different speeds, we standardized the length of audio files to 20,000 units.

5.2. Splitting of Sentence:

- We observed that each word in the audio typically takes up between 15,000 to 25,000 units in the array.

- We used the change in loudness to separate each word in the audio.

- By smoothing the audio and identifying low points, we successfully isolated individual words.

5.3. Word Predictor:

- *We used a model called Long Short-Term Memory (LSTM)to predict the next likely words based on the sequence of words.*

- *The sentences were divided into sequences of words, and allpossible combinations (n-grams) were generated.*

- *These sequences were fed into the LSTM model to predict the next word.*

5.4. Voice tp Write:

5.4.1. Methodology 1: Deep Learning:

- The word predictor model gives a list of potential next words,which is used to classify the input segment.

- We trained the model using pre-processed audio files and Mel Frequency Cepstral Coefficient (MFCC) features.

- The model architecture includes input, hidden, and output layers, with each layer having a specific number of nodes.

- The Softmax activation function is used to determine the probability distribution over possible word classes.

5.4.2. Methodology 2: Based on Similarity of Signal:

- By comparing the input audio with the training dataset usingcosine similarity, we determine the class (or word) that appears most frequently among the most similar recordings.

6. RESULTS AND DISCUSSIONS Here's a simplified explanation of the results:

6.1. Splitting of Sentence:

- Out of 30 sentences in the training set, the algorithm correctly split 28 sentences.

- For the remaining 2 sentences, it successfully separated thewords after recording, ensuring there were enough gaps between them.

6.2. Word Prediction:

- The Word Predictor model had an accuracy of 90%.

- When asked to predict the top 5 potential words, the results matched expectations.

6.3. Voice to Write:

6.3.1. Methodology 1: Deep Learning:

- Model accuracy was calculated by looking at how well it predicted words in each sentence

and averaging these accuracies across all sentences.

- Accuracy is measured by the ratio of correctly predicted words to the total number of words in the dataset.

6.3.2. Methodology 2: Based on Similarity of Signal:

- For the first method, we calculated the average similarityfor each class based on recordings. The class with the highest average similarity was chosen as the predicted word.

- In the second method, we looked at the top 20 similar recordings and chose the class with the highest similarity asthe prediction.

- Model accuracy was determined by the ratio of correctly predicted words to the total number of words in the dataset.

7 NOVELTY APPROACH

Combining a multilingual Next Word Prediction model with a DL translator is a new way to improve translation. The Word Predictor looks at previous words to guess the next five words in a sentence. These guesses help the DL translator figure out what the words in the sentence are, without having to check every word in the dictionary. This makes translation faster and more accurate because it focuses on the words that are most likely to be used together.

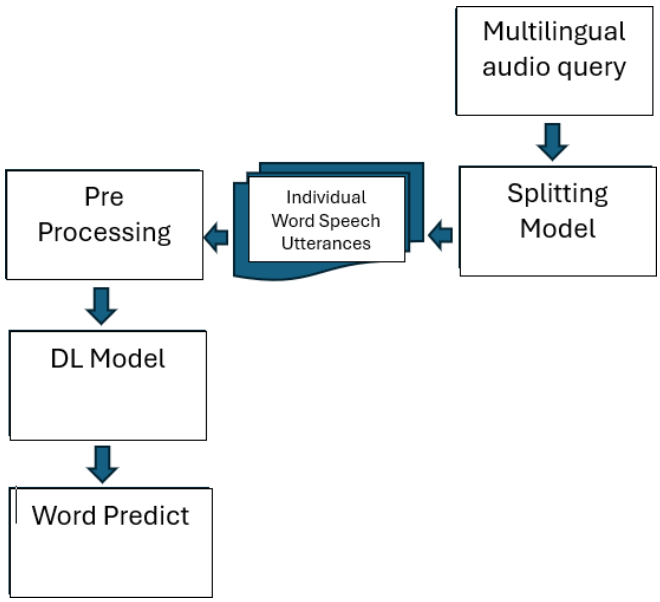

Figure 1. Process Flow

8 CONCLUSIONS

The method we used in this study is a big step forward. It relieson the model learning by itself to understand and translate multilingual questions into single language ones. By using the top predictions from the Word Predictor model, our DeepLearning model can focus on the most likely words, which makes it faster and more accurate at translating audio queries.

When we tested it with our multilingual dataset, the Deep Learning model was 85% accurate. But when

real users tried it,the accuracy dropped slightly to 71%.

Comparatively, another model using cosine similarity had an average accuracy of 0.59. However, when it looked at the mostsimilar recordings, its accuracy improved to 0.64.

9. LIMITATIONS

A big problem with our plan is that the Deep Learning model has to run every time a new word is guessed. This can take a lot of time, especially when trying to understand just one sentence. Also, our strategy relies a lot on how well the Word Predictor model works. If it doesn't predict words accurately, it can mess up the Deep Learning model's work. Plus, because we don't have probabilities for the predicted words, we can't tell how sure we are about them being right in the sentence.

10 REFERENCES

[1] Toshniwal, Shubham, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. "Multilingual speech recognition with a single end-to-end model." In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4904-4908. IEEE,2018.

[2] Fogarassy-Neszly, Paul, and Costin Pribeanu. "Multilingual text-to-speech software component for dynamic language identification and voice switching." Studies in Informatics and Control 25, no. 3 (2016): 336.

[3] Vaceslavovic, Belousov Urij. "Automatic language identification using deep neural networks." (2019).

[4] Wanli, Zhang, and Li Guoxin. "The research of feature extraction based on MFCC for speaker recognition." In Proceedings of 2013 3rd International Conference on Computer Science and Network Technology, pp. 1074-1077. IEEE, 2013.

[5] Sundermeyer, Martin, Ralf Schlüter, and Hermann Ney. "LSTM neural networks for language modeling." In Thirteenthannual conference of the international speech communication association. 2012.