Development of Indian Spoken Language Identification System for Two Languages using MFCC Feature with Deep Neural Network

Priyank S. Yadav, Kiran R. Trivedi

Abstract: ***Language is the ability to communicate with any person. Approximate number of spoken languages are 6500 in the world. Different regions in a world have different languages spoken. Spoken language recognition is the process to identify the language spoken in a speech sample. Most of the spoken language identification is done on languages other than Indian. There are many applications to recognize a speech like spoken language translation in which the fundamental step is to recognize the language of the speaker. This system is specifically made to identify two Indian languages. The speech data of various news channels is used that is available online. The Mel Frequency Cepstral Coefficients (MFCC) feature is used to collect features from the speech sample because it provides a particular identity to the different classes of audio. The identification is done by using MFCC feature in the Deep Neural Network. The objective of this work is to improve the accuracy of the classification model. It is done by making changes in several layers of the Deep Neural Network.***

Keywords: Mel frequency cepstral coefficients, Convolutional Neural Network, Language Identification System.

I. INTRODUCTION

When a person meets another person speaking different language, it is hard to identify in which language he/she is talking. There are many applications to recognize a speech like spoken language translation, multilingual speech recognition, in which the system must be able to identify the language spoken by a person. In the Speech to text conversion application, the system is only able to recognize English language, by having the spoken language identification function, the system will be able to recognize the language in the speech sample. That motivates to develop Spoken Language Identification system for Indian languages. By using deep learning, computers are able to analyze complex data that can provide more accurate results.[6] Objective of this work is to identify two Indian languages Hindi and Kannada in a spoken sentence using deep neural network approach, collect database for training, extraction of speech features using MFCC, design neural network using Convolutional Neural Network (CNN), training the deep neural network and to detect languages with good accuracy.

Revised Manuscript Received on May 20, 2020.

Priyank S. Yadav, Department of Communication Systems Engineering (E.C.), SSEC, Bhavnagar, Gujarat, India.

Dr. Kiran R. Trivedi, Associate Professor, Department of Communication Systems Engineering (E.C.), SSEC Bhavnagar, Gujarat, India.

II. PROPOSED WORK

The system is implemented in the laptop. This work is mainly divided into two parts, first is to generate acoustic features from the dataset and second is to create modelling of the deep neural network using a classifier. After designing a deep neural network classifier for this work, a speech sample is to be predicted on the saved model. The Fig 1 shows the block diagram of proposed work. That is explained further.

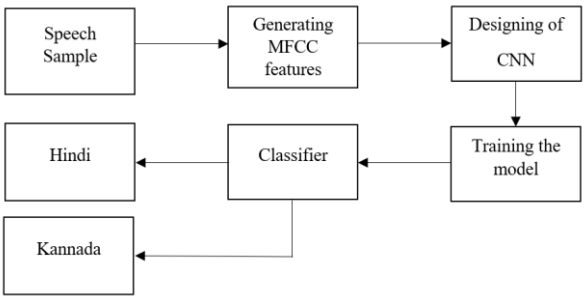

Fig 1. Block Diagram of proposed work

A. Dataset

This work is for dedicated for two Indian languages (I) Hindi and (II) Kannada. So, the Indian news channels audio is chosen to use for preparation of dataset. Because news channels' audio data is clear without any background noise. Both female and male news anchor's spoken sentences are recorded for this work. So, that the model will be independent from gender of the speaker. The audio file is recorded by using a free software jetAudio. It has a recording feature that allows to record stereo audio within the pc as the video is played in the internet browser. The recording window is shown in Fig 2.

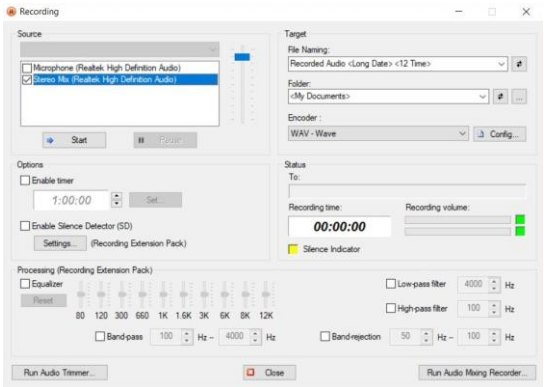

Fig 2. jetAudio Recording Window

DD News and DD Chandana channels on the YouTube provide the news in the following languages Hindi and Kannada. These recordings are saved in the pc as wav files by jetAudio automatically. After recording several files, these audio data are split into 10 seconds using a free software called NCH WavePad Sound Editor as shown in Fig 3. Total 500 samples are prepared for each language. The list of all the prepared samples is now created in a csv file. In the csv file the sample is labelled with the correct language of that sample.

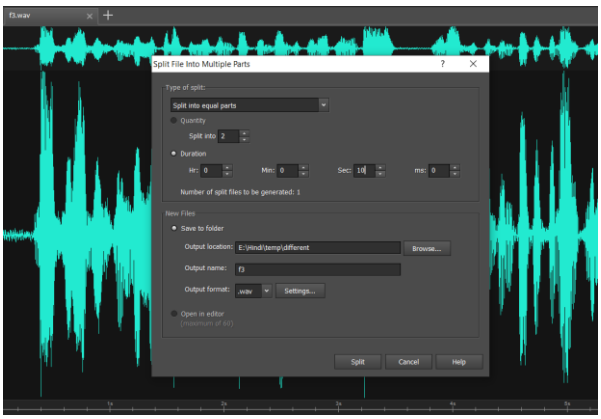

Fig 3. Split wav file into multiple parts using NCH WavePad Sound Editor

B. Generation of MFCC features

In any speech recognition techniques, MFCC is generally used for feature extraction. It mimics parts of speech perception and production to extract features having details about the linguistic message spoken by the speaker. It tries to eliminate the speaker dependent characteristics by removing their harmonics.

Since speech is a non-stationary signal, its frequency contents are continuously changing with time. For the analysis purpose speech signal is divided into 20-30 milliseconds, because the shape of our vocal tract is unvarying for small intervals of time. This process is known as framing. For each frame calculate the periodogram estimate of the power spectrum. The periodogram spectral estimate still contains a lot of information not required for Speech Recognition. A mel-filterbank is to be applied on the periodogram. The Fig 4 shows 10 filters for frequencies 0 to 8000 Hz.[1]

Take the log of each of the 26 energies from step 3. This leaves with 26 log filterbank energies. Take the Discrete Cosine Transform (DCT) of the 26 log filterbank energies to give 26 cepstral coefficients. For ASR, only the lower 12-13 of the 26 coefficients are kept.

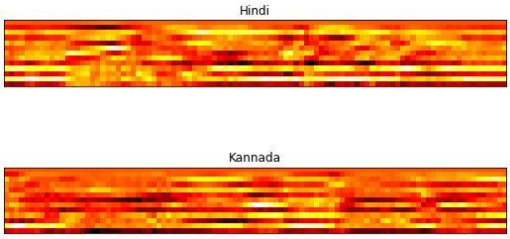

Fig 5. MFCC feature generated from a random speech sample

Keep DCT coefficients 2-13, discard the rest. The resulting features (12 numbers for each frame) are called Mel Frequency Cepstral Coefficients. These MFCC features are generated using the python_speech_features library in the python directly from the wav file that is created in the dataset. So, in this step by using this library mfcc features are generated for the input dataset of speech sample for both languages Hindi and Kannada. Fig 5 is showing a single mfcc feature generated from a speech sample for both languages.

C. Classification

After creating the random feature using the mfcc from the python_speech_features library. The next important task is to create a convolutional neural network [4] for training the dataset. The convolutional neural network is designed at this stage. In the convolutional neural network model complexity can be increased as per the requirement.[5] Here validation split is 30%. Batch size is 32. Number of epochs is 30 for training the model in the proposed work.

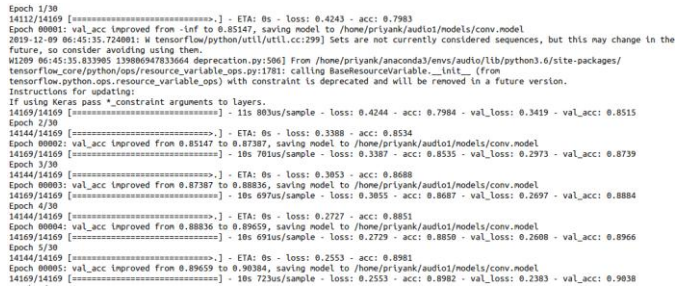

Fig 6. Training of the model

As shown in the above Fig 6 the training accuracy is 79%, validation accuracy is 85% and loss is 42% in the first epoch. After several epochs it should increase the accuracy and loss should be reduced. In the next epochs, loss and validation loss is decreasing. The maximum validation accuracy got in the last epoch is 93%. So, the model is trained well on the given dataset. The following step is to test the model on the random speech sample and it should predict the correct class in between Hindi and Kannada.

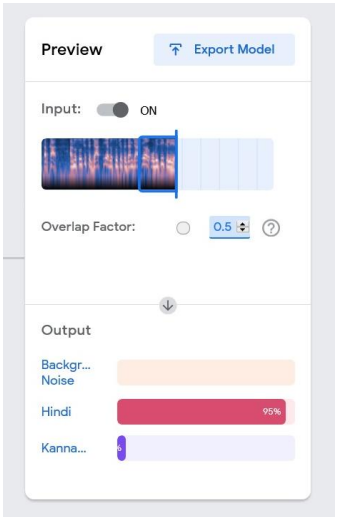

Fig 7. Testing of the model

By testing the trained model, test accuracy is achieved 95% as shown in Fig 7.

III. CONCLUSIONS

In this work the main objective to get the high accuracy for training the deep neural network is achieved. The main role of the MFCC feature has greatly helped the convolutional neural network to achieve high accuracy. Furthermore, number of Indian languages can be tested for this model.

ACKNOWLEDGMENT

The authors would like to thank everyone who helped us during the work by leading towards the accomplishment.

REFERENCES

1. Mel Frequency Cepstral Coefficient (MFCC) tutorial. (n.d.). Retrieved from practicalcryptography: accessed on 23 August 2019 http://www.practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/
2. S. D. Joge and A. S. Shirsat, "Different language recognition model," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, 2016, pp. 1096-1101. doi: 10.1109/ICACDOT.2016.7877756
3. Mwiti, D. (2018, May 8). Convolutional Neural Networks: An Intro Tutorial. Retrieved from heartbeat: accessed on 23 August 2019 https://heartbeat.fritz.ai/a-beginners-guide-to-convolutional-neural-networks-cnn-cf26c5ee17ed
4. Prabhu. (2018, March 4). Understanding of Convolutional Neural Network (CNN) — Deep Learning. Retrieved from Medium: accessed on 23 August 2019 https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148
5. Rosebrock, A. (2018, December 31). Keras Conv2D and Convolutional Layers. Retrieved from pyimagesearch: accessed on 28 August 2019 https://www.pyimagesearch.com/2018/12/31/keras-conv2d-and-convolutional-layers/
6. Hargrave, M. (2019, April 30). Deep Learning. Retrieved from Investopedia: accessed on 23 August 2019 https://www.investopedia.com/terms/d/deep-learning.asp

AUTHORS PROFILE

Priyank S. Yadav B.E. in Electronics and Communication Engineering at GEC Bhavnagar, pursuing M.E. in Communication Systems Engineering (E.C.) at SSEC, Bhavnagar.

Dr. Kiran R. Trivedi Associate professor at SSEC Bhavnagar. 23 years teaching experience. 40 publications. IETE membership