SPEECH CLASSIFICATION USING DEEP CONVOLUTION NETWORK

Aakash Agarwal (191010201)
Dept. of ECE
IIIT-NR

Himanshu Singh (191010219)
Dept. of ECE
IIIT-NR

Piysuh Rawat (191010227)
Dept. of ECE
IIIT-NR

Abstract

Speech Recognition and Classification is one of the most advanced topics. Deep Learning plays an important role in this area. The purpose of this work is to identify and classify words in English Language using Deep Learning Algorithms. We will not be recognizing a full sentence but just one word. Convolution Neural Networks (CNN) is used in this work to automatically detect words in English. Mel frequency coefficients technique is used for feature extraction purpose. The result section shows how the accuracy and the performance of the model, on seen and external noisy data as well.

Introduction

Speech Recognition Systems are very common these days. Many researchers have already done some brilliant work in the given field. Speech Recognition is widely used in day to day life to make our life easier Recently, Deep Learning which is a part of machine learning came to limelight and showed its merit in the said and various other fields. Deep Learning is a very vast topic and is used in most of the high level Artificial Intelligent Systems. We have developed a system the automatically recognizes English speeches using Mel Frequency Cepstral Coefficients for feature extraction and deep Convolutional Neural Network for classification..

Data Preparation

A set of one second long, 5 short words in English is used, created by several people with different accent, of all genders. The files are then differentiated into 5 classes, each for different word. The categories are **'left', 'go', 'yes', 'down', 'up'**, each of which is a very common English word. Table 1 shows all the classes in our dataset. We have used supervised learning which means that we have labelled our dataset for training.

1	LEFT
2	GO
3	YES
4	DOWN
5	UP

Fig1 Classes within dataset

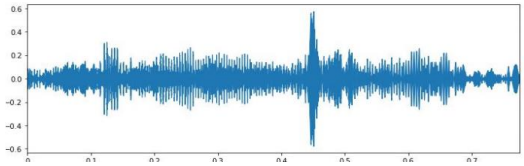

Fig. 2. Raw waveform of word "LEFT"

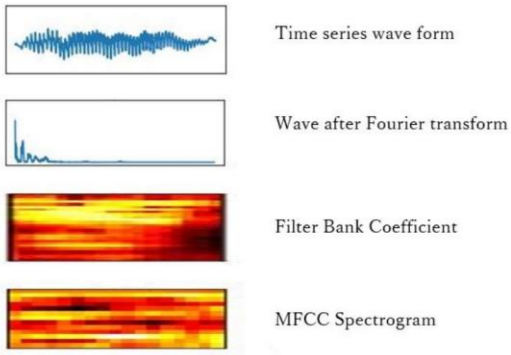

Fig 3. Different waveforms

MEL FREQUENCY CEPSTRAL COEFFICIENTS

Mel Frequency Cepstral Coefficients are a feature that are used in almost all automatic speech and speaker recognition. The Mel scale relates the perceived frequency of the tone to actual measured frequency.

The formula for converting from frequency to Mel scale is:

$$M(f) = 1125\ln(1 + \frac{f}{700})$$

And to go back to the frequency

$$M^{-1}(m) = 700(\exp\left(\frac{m}{1125}\right) - 1)$$

The following block diagram helps us to understand how we get the MFCCs.

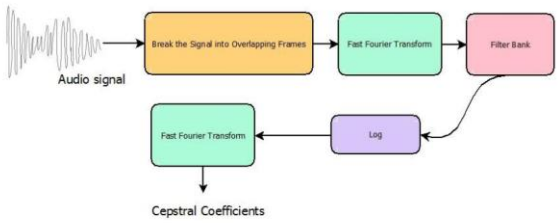

FIG. 4 Block Diagram to show the steps involved in MFCC

The following algorithms is used to create MFCC features:

1. Frame the signal into short frames.
2. For each frame calculate the periodogram estimate of the power spectrum.
3. Apply the Mel filter bank to the power spectra, sum the energy in each filter.
4. Take the logarithm of all filter bank energies.
5. Take the DCT of the log filter bank energies.
6. Keep DCT coefficients 2-13, discard the rest

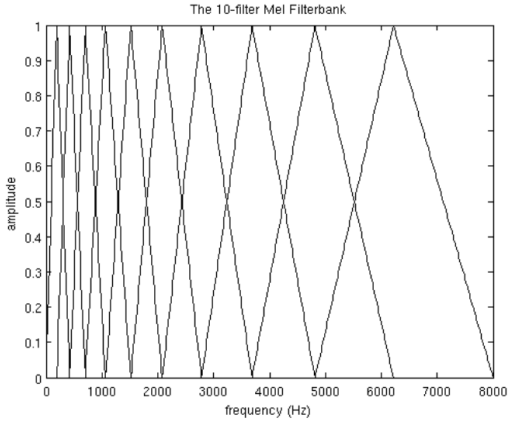

FIG 5. A Mel Filter Bank

The above figure shows a Mel filter Bank containing 10 filters. The filterbank starts at 0Hz and ends at 8000Hz.

DATA PREPROCCESSING

Real-world data contains noises and/or missing values, which cannot be directly used for our model. Data preprocessing is required for cleaning the data and making it suitable for the model which increases the accuracy and efficiency too. So, the audio data is being pre-processed before we use it in our model. Sound waves are normalized to range from -1 to +1. Sounds are translated to mono as we have used only one channel. Spectroscopy, log Mel filter banks and Mel-Frequency cepstral coefficients (MFCCs) from samples to convert the raw waveform to a time-frequency domain is used to extract features from raw waveform. These features will be given as the inputs for the neural nets. Now data is turned into a vectors/values and then split into two separate sets of training data 80% and 20% each.

DEEP NEURAL NETWORKS

Our model has used Conv1d and GRU layers to model the network. Only one dimensional convolution is done by the Conv1d layer of the neural network. Gated recurrent Unit or GRU layer is used to solve the vanishing gradient problem.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 8000, 1)	0
batch_normalization_1 (Batch	(None, 8000, 1)	4
conv1d_1 (Conv1D)	(None, 7988, 8)	112
max_pooling1d_1 (MaxPooling1	(None, 2662, 8)	0
dropout_1 (Dropout)	(None, 2662, 8)	0
conv1d_2 (Conv1D)	(None, 2652, 16)	1424
max_pooling1d_2 (MaxPooling1	(None, 884, 16)	0
dropout_2 (Dropout)	(None, 884, 16)	0
conv1d_3 (Conv1D)	(None, 876, 32)	4640
max_pooling1d_3 (MaxPooling1	(None, 292, 32)	0
dropout_3 (Dropout)	(None, 292, 32)	0
batch_normalization_2 (Batch	(None, 292, 32)	128
bidirectional_1 (Bidirection	(None, 292, 128)	124416
bidirectional_2 (Bidirection	(None, 292, 128)	198144
bidirectional_3 (Bidirection	(None, 128)	198144
batch_normalization_3 (Batch	(None, 128)	512
dense_1 (Dense)	(None, 256)	33024
dense_2 (Dense)	(None, 11)	2827

Total params: 563,375
Trainable params: 563,053
Non-trainable params: 322

FIG. 6 the Deep Neural Network

RESULT

Through this model we were able to achieve an accuracy of 93%. We also created a script that would real time audio and classify the word said into one of the five different classes (['**left**', '**go**', '**yes**', '**down**', '**up**']).

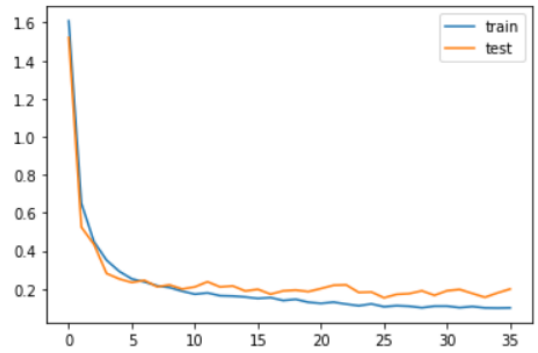

Fig. 7 The loss is decreasing as number of epochs increase

CONCLUSION

In this project, we have demonstrated the use of MFCC feat technique for feature extraction and deep Convolutional Neural Network for model training of the model for one isolated word in the English language speech recognition. The work that we have done is very basic and is done for the analysis of the native English isolated word, and further work is to be done on noisier dataset. In addition, we used only some number of English words (5 to be precise) which is a very small number. However, the implementation of this work on a large scale requires a very large collection of audio files that is a very large dataset. Accuracy of this model on HINDI language is also need to be done but currently there is no dataset of the same. We plan to manually generate one soon.

REFERENCES

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent PreTrained Deep Neural Networks for Large-Vocabulary Speech Recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 30–42, Jan. 2012, doi: 10.1109/TASL.2011.2134090.

[2] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 22, no. 10, pp. 1533–1545, Oct. 2014, doi: 10.1109/TASLP.2014.2339736.

[3] S. Malekzadeh, M. H. Gholizadeh, and S. N. Razavi, "Persian Vowel recognition with MFCC and ANN on PCVC speech dataset," arXiv preprint arXiv:1812.06953, 2018.

[4], "A Letter to Sound System for Farsi Language Using Neural Networks," in 2006 8th international Conference on Signal Processing, Nov. 2006, vol. 1, doi: 10.1109/ICOSP.2006.345518.

[5] S. Malekzadeh, "Phoneme-Based Persian Speech Recognition," arXiv:1901.04699 [cs, eess], Jan. 2019, doi: 10.13140/RG.2.2.32856.96007.