1/3/2018

Speaker Verification

Deep Learning Project Report

Submitted By :

M.Khalid, i14-1636

Zeeshan Ali, i14-1623

Speaker Verification Using Deep CNN

Abstract — Recently deep convolutional neural network has been used to learn features of different speakers but the quality of learned-features is not sufficient good so a backend model is needed for extraction of speaker's features. In this report we have used CNN model to get speaker's features that are more robust. In 1st phase of project we have built an underlying model of CNN to classify speaker at the utterance level and in 2nd phase we used the trained model to get speaker's features the extracted features from the test set will be compared to stored speakers features to verify the claimed identity of a speaker. We used KING and NKING data set for our speaker verification model.

Index --- Speaker Verification, CNN, Speaker features

1. Introduction

The speaker verification is the task of verifying the claimed identity of a speaker by recording their voice and extracting features from that voice. The major technical challenge is to get speaker voice representation model that can be used later for authenticating speaker. In most systems i-vector is used as a feature representation of speakers that is calculated by Gaussian Mixtures Model-Universal Background Model (GMM-UBM) frame work. After that different approaches were investigated like SVM model and PLDA (Probability Linear Discriminant Analysis) as a supervised learning method. In our case we used convolutional neural network which has often been utilized for 2D inputs but we are efficiently using 3D CNN as well to learn speaker features. Generally there are three phases of our overall model: In phase 1 the back ground model is created for speaker representation, in phase 2 the speakers model are created by using trained model and the test features set are used to compare them with already created speaker's model to prove/disprove the claimed identity of the speaker.

2. Methodology

In this section we describe the general process of model. There are total three phases of the project in which whole model is carried out. Each new phase is dependent on previous phase implementation.

2.1 Phase 1:

In this phase a background model is built for speaker representation from the speaker utterances generated by the model. The input features for training the model are extracted from the audio files Of the speakers. Each extracted frame consist of 0.8s sound sample at frame level with 40 mfcc features coefficients So total 80 temporal frames and their 40 mfcc coefficients are collected and overlapped 20 times. Each features map is in format of 80x40x20.The procedure of generation of features map is given below:

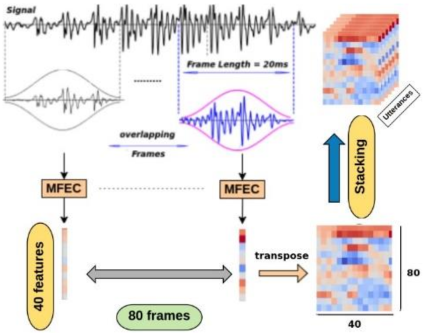

Fig-1 mfcc's of audio signals

After getting features map, we tried to reduce softmax loss by using 3D and 2D CNN Model. The trained Model will be used in 2nd phase of the project. The architecture diagram of the 2D CNN model is shown below but we also have trained 3D CNN Model as well.

2.2 Phase 2:

Once background model for speaker representation is trained then weights are fixed and then we feed forward the speaker features maps and get final features of the speakers and averaged the logits produced by the model for each speakers. Then we calculated the cosine similarity of test set speakers with the already stored speaker's models. In this setup, false rejection and false acceptance rates are investigated as the main error indicators. The false rejection/acceptance rates depend on the predefined

Threshold.

Layer	Input	Output	Kernel	Stride
Conv1	80x40x20	80x40x16	16x(3x3)	1
BN, DO, ReLu, MP1	80x40x16	40x20x16	-, 0.5, -, 2x2	2
Conv2	40x20x16	40x20x32	32x(3x3)	1
BN, ReLu	40x20x32	40x20x32	-	-
Conv3	40x20x32	40x20x32	32x(3x3)	1
BN, ReLu, MP2	40x20x64	20x10x64	-, 0.5, -, 2x2	2
Conv4	20x10x64	20x10x128	128x(3x3)	1
BN, ReLu	20x10x128	20x10x128	-	-
Conv5	20x10x128	20x10x128	128x(3x3)	1
BN, ReLu, MP3	20x10x128	10x5x128	-, 0.5, -, 2x2	2
Conv6	10x5x256	10x5x256	256x(3x3)	1
BN, ReLu	10x5x256	10x5x256	-	-
Conv7	10x5x256	10x5x256	256x(3x3)	1
BN, ReLu, MP	10x5x256	5x3x256	-, 0.5, -, 2x2	2
Conv10	5x3x256	1x1x256	256(5x3)	1
FC-1	1x1x256	1x512		
FC-2	1x512	1x1024		
FC-3	1x1024	1x51		

Table-1 CNN Architecture

3. Experiments

3.1 Data Set

We used KING and NKING Data which contains 51 Males' Speaker's recorded voice and each speaker's audio files are consist of 10 sessions. After getting mfcc features maps the overall data set is almost 24,000 images of shape 80x40.We split the data into training and testing sets with test ratio 30%.

3.2 Accuracy / Error

The # of utterances in each example has very significant role in accuracy and final error. We found that at 20 # of utterances we get good results.

# Utterances	Model	Accuracy
1	2D-CNN-1	48.0%
20	2D-CNN-1	27.0%
20	2D-CNN-2	60.0%
20	3D-CNN	45.0%

3.3 Hyper Parameter

After Cross validation technique, we found that the batch size of 200 image was good for our model and learning rate was 0.001 initially. More Over we used Xavier weight initialization for our model and # of neurons for last 3 fully connected layers are 512, 1024 and 51 which gives goods results. We used Adam Optimizer with soft max cross entropy loss. The batch sized that is suitable for this data is 200.

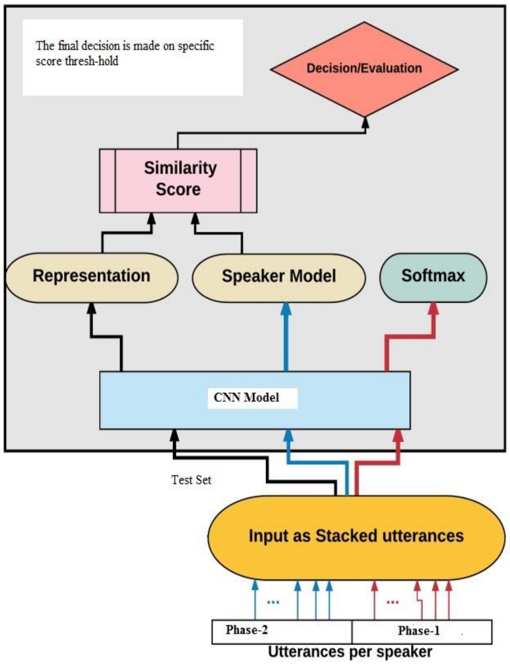

Fig-2 Model Architecture

4. Conclusion

Our investigation presented CNN underlying model for speaker representation to prove or disprove the speaker claimed identity. The features learned from the model are more robust due to overlapping utterance of same speakers. Our experimental results are good and we hope this model can be used for Bio authenticating user application as a future feature enhancement.