Detailed Project (Task) Report for Speech Engineer Post at Saarthi.ai

Chodingala Piyushkumar K.

1. Introduction

Recently, Voice Assistants (VAs) are very user-friendly and can operate various voice-based applications, such as smart offices, smart homes, and hand-free operations. There are many smart voice assistants available, such as Apple Siri, Google Assistant, and Microsoft Cortana, and Samsung Bixby [1]. In this study, the multi-class classification system for various task on VAs is proposed. Here, I have used Mel-frequency Cepstral Coefficients (MFCC) feature to extract the acoustic information from speech signals along with Convolutional Neural Network (CNN) as back-end classifier.

2. Feature Used

The MFCC is proved to be one of the most successful feature set in wide range of speech technology applications [2]. It is also used to mimic the auditory representation. The windowed speech signal is processed through Fourier transform (FT) to produce STFT. The weighted sum is performed for each Mel scale subband filter. Then, Discrete Cosine Transform (DCT) is applied and desired number of cepstral coefficients are extracted to get MFCCs. In this project, I have used 40 Mel scale subband filters for feature extraction. *13*-Dimensional static coefficients have been extracted along with its Delta and double-Delta coefficients. Hence, the final features will be of *39*-Dimensional. The functional block diagram of MFCC extraction is given in Fig. 1. This extracted MFCC features are fed to classifier for classification task.

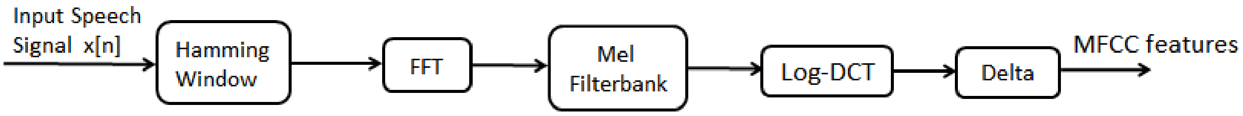

Figure 1: Functional block diagram of MFCC feature extraction.

3. Dataset and Classifier Used

The dataset partition is given in Table 1.

In this project, CNN is used as a back-end classifier. The CNN is a neural network-based architecture which consist of one or more convolutional layers followed by classification layers

Table 1: Dataset partition

Dataset	No of Files	No of Unique Labels In Action Class	No of Unique Labels In Object Class	No of Unique Labels In Location Class
Train	11566	6	14	4
Valid	3118	6	14	4

[3]. In this work, the input MFCC feature size for CNN is taken to be 39×400. This CNN architecture consist of four convolutional layers (Conv1, Conv2, Conv3, and Conv4) followed by two fully-connected layers (FC1 and FC2). Here, in the first two convolutional layers, the data is convolved using a kernel size of 5×5 with a stride of 1 and padding of 2. Furthermore, in the remaining two convolutional layers, the kernel is used of size 3×3 with the stride and padding of 1. Here, after every convolutional layer, I have used max-pool layer having kernel of size 2×2 with a stride of 2, in order to reduce the size of data and also to reduce the computation cost of CNN model. The detailed information of input and output of each layer is given in Table 2. After extraction of features from convolutional layers, the output of Conv5 is fed to FC1 layer and the probabilistic output for classification is taken from FC2. The Rectified Linear activation Unit (ReLU) function is taken as the activation function for all hidden as well as FC layers [4]. Binary cross-entropy is taken to be the loss function and for optimization of weights, the stochastic gradient descent method is used.

Table 2: Details of proposed CNN architecture. Here, N is selected according to number of unique labels in class (i.e., action=6, object=14, location=4).

Layer	**Filter/Stride**	**Output**	**# Parameters**
Conv1	5x5/1x1	16x37x398	400
batchnorm1	-	16x37x398	-
MaxPool1	2x2/2x2	16x18x199	-
Conv2	3x3/1x1	32x18x199	4608
batchnorm2	-	32x18x199	-
MaxPool2	2x2/2x2	32x9x99	-
Conv3	3x3/1x1	64x9x99	18432
batchnorm3	-	64x9x99	-
MaxPool3	2x2/2x2	64x4x49	-
Conv4	3x3/1x1	16x4x49	9216
batchnorm4	-	16x4x49	-
MaxPool4	2x2/2x2	16x2x24	-
FC5	-	1x200	153600
FC6	-	1xN	1200
Total Trainable Parameters			**187456**

4. Experimental Results

The performance metrics used in this work is F1-score [5]. For calculation of F1-score, first step is to use a classification model which make a prediction of class labels for each sample of testing dataset. The predicted labels are then compared with actual labels of testing data. The F1-score is then calculated based on the precision and recall of the classification model. The prediction of labels by classification model is divided in four parts, namely, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The F1-score is calculated from these four parts as:

$$\text{F1-score} = \frac{2 * precision * recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}. \tag{1}$$

The results of the system are not shown in this document, due to the fluctuations in the results.

5. Summary And Conclusion

In this study, the significance of the MFCC feature is shown for various task classification for VAs. Hence, the experiments are performed on given dataset of various voice commands used for VAs. Also, the CNN classifier is used for back-end classification task. In the future, other acoustic feature and classifiers can be explored for this multi-class task classification to improve further accuracy of the system.

References

[1] V. Kepuska, G. Bohouta, Next generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa, and Google home), in: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), University of Nevada, United States, 8-10 Jan. 2018, pp. 99–103.

[2] R. Vergin, D. O'Shaughnessy, A. Farhat, Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition, in: IEEE Transactions on Speech and Audio Processing, Vol. 7, 1999, pp. 525–532.

[3] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[4] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: International Conference on Machine Learning (ICML), Haifa, Israel, 21-24 June 2010.

[5] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, BMC genomics 21 (1) (2020) 1–13.