Text Dependent Speakers Pattern Classification with Back Propagation Neural Network

N K Kaphungkui, Gurumayum Robert Michael, Aditya Bihar Kandali

Abstract: Speaker Recognition is the procedure of validating a speaker’s claimed identity using his/her speech characteristics which is unique to each individual. The primary objective of all speech recognition systems is a man-machine interface which grants access into the system with the voice characteristics. This will serve as a highly secure biometric system where security is the primary concern. The primary aim of this paper is to classify each speaker accurately with MFCC and Back Propagation Neural Network. Scaled conjugate gradient training function is used for back propagation neural network. A small database of 10 people is created from a group of five male and five female uttering the same sentence five times repeatedly. The sentence consists of five different words. The numbers of data set for classification is 22182. The accuracy obtained from the classification is 92.1% with small percentage of 7.9% misclassification which is acceptable good. The tool for simulation is MATLAB.

Keywords: Speaker recognition, MFCC, text dependent, confusion matrix, training, validation, testing.

I. INTRODUCTION

Speaker recognition system is broadly classified into different types as shown in Fig.1 [12]. The system having many trained speakers is an open set system. Whereas a closed set recognition system is the system which has specified registered speakers in the system. Again in Text-Dependent type (constrained mode), the same test utterance (words or sentences) is used during training and testing session. As such, the speaker has prior knowledge of the system and the system need cooperation from the speaker. But in text independent system (unconstrained mode) the test speakers have no information about the words or sentences used for training the system and have the liberty of speaking anything [15], [11].

Finally Speaker identification is the process of figuring out that a given articulation belongs to which of the enlisted speaker and Speaker verification is a framework of either accepting or dismissing the personality claimed of as a speaker. It is a 1:1 mapping process [3]. System’s training and testing are two important phases which governs all types of speaker recognition system [10], [14]. The various Techniques which are used for unique speech features extraction are MFCC, LPCC, RCC, LPC and PLPC. Likewise for Speech classification various classifiers such as Neural Network DWT, HMM, GMM, SVM and VQ are used. Researcher have reported that when 70% of total samples are used for training and 30% of total samples are used for system testing, the highest efficiency of 81.8% for 10 person is obtained using combination of MFCC, pitch and rms with feed forward neural network (FFNN) [16]. Other have proposed that ANN can give results better than fuzzy logic based systems in terms of accuracy rate with the condition that the samples is recorded in a controlled and clean environment. 74% accuracy is obtained with ANN against 72% accuracy rate with fuzzy logic [17]. Mel Frequency Cepstrum Co-efficients) Based Text-Dependent Speaker Identification Using BPNN is already presented. The highest accuracy of 92% is achieved for 10 users [18]. A new method where both LPC and MFCC are used in parallel is also proposed for Assamese Speaker Recognition Using Artificial Neural Network for 10 speakers where each single word is uttered twenty times by each speaker. A moderately high accuracy is obtained [19]. This work will be carried out with basic 13 coefficients MFCC to represent speech features and BPNN as classifier.

II. MEL FREQUENCY CEPSTRAL COEFFICIENT

In recent years many have reported that speech features extraction based on MFCC method is becoming popular and successful as it is being modeled as human auditory system with high accuracy rate [9], [10].
Some have already used a new way of using weighted MFCC for speaker recognition and found that the recognition rate is superior to non-weighted MFCC [2]. A 16-order MFCC coefficient with a combination of LPCC as classifier for speaker recognition is also reported [7]. Mel Frequency Cepstral Coefficients (MFCC) algorithm is more preferred to extract speech features for performing voice recognition due to the generation of unique coefficients from the user’s voice [8]. Fig. 2 shows the basic steps involving for speech feature extraction based on MFCC [1].

![Fig. 2 MFCC blocks](image)

This work’s aim is the extraction of the speech features with MFCC algorithm from any speech wav file and represent with a unique 13 coefficients for each speakers. The MATLAB code parameters for speech feature extraction is set as follows:

- a) Frame duration = 26msec
- b) Frame shift = 10msec
- c) Pre-emphasis coefficient = 0.97
- d) Number of filter bank channel = 20
- e) Filter’s Lower cut off frequency= 300Hz
- f) Filter’s Upper cut off frequency = 3750 Hz.

A five words sentence “This voice is my password” is uttered five times repeatedly. The sampling of the voice is 48 KHz with a 16 bit, bit depth. The resultant MFCC features are in matrix form which consists of fix 13 rows and having different number of column for different speaker. It can be represented as M x N where M is fix rows and N is variable columns.

A small database is created for five female F1, F2, F3, F4, F5 and five male M1, M2, M3, M4 and M5 in a quite noise free environment with the same recording device. The length of the speech will be different for different person as it depends upon how fast or how slow the person utters the sentence. The more the time is taken while uttering the sentence the more the speech length is and vice versa. The resultants 13 coefficients representing the speech are then used for training the neural network. The more the number of input data is used for training, the better the performance of the network will be. The pictorial representation of the simulated speech waveform and its cepstrum for a particular person is shown in Fig 3.

III. BACK PROPAGATION NEURAL NETWORK

The structure of multi layer Perceptron neural network is shown in Fig. 4. There are two basic steps that governs the Back propagation algorithm i.e. feed forward direction which propagates from inputs towards output nodes and back propagation which propagates from output towards input nodes. A back propagation algorithm is a method for training Multi-layer Artificial Neural Networks. Initially while neural network is designed, random values of weight are assigned for the model to predict the correct output. If there is a huge variation between the output and the actual target, the weights are updated to minimize the model error. The solution of the learning problem lies with the weight that minimizes the error function. The basic steps of the algorithm is as follows:

- Error calculation – how much difference between network output from actual output
- Error minimizing – model error minimum or not
- Change the network parameters – if there is huge error, update the weights and bias and check the error again. The whole process is repeated to obtain minimum error.
- Network ready for prediction – after obtaining least error, the training stop and the model is ready for the inputs to predict the actual output.

Minimizing the value of the mean square error function so that the output is closely following the desired targets by updating the weights of the network is the algorithm of back propagation in machine learning. [4], [5], [6]. The weights minimizing the error function is the solution to the neural network learning problem. The more the number of training data set, high accuracy is obtained and the network’s performance will improve [13].
The Back Propagation Algorithm is shown in Fig. 5.

IV. CONFUSION MATRIX

Confusion matrix is the matrix which is used for different pattern classification as shown in Fig. 6. The sum of a particular row represents the total number of test examples of that class i.e. the sum of true positive (TP) and false negative (FN) of that class. For example, total number of test examples of class 1 is the sum of TP1+E12+E13+E14+…+E1n. The total number of FN’s of a certain class is the sum of all the errors in a row excluding the TP. Likewise the total number of FP’s of a particular class is the sum of all the errors in a column excluding the TP. And the total number of TN’s for a specific class will be the sum of all the rows and columns by excluding or masking that class’s row and column.

The various parameters of confusion matrix namely Accuracy, Precision, Sensitivity and Specificity are expressed as under.

Accuracy is defined as how close the predicted values are to the actual or true values. Accuracy of the confusion matrix is given by the sum of correct classifications or the true positive values divided by the total number of classifications.

\[
\text{Accuracy} = \frac{\text{Sum of diagonal elements}}{\text{Sum of all elements}}
\]

Precision refers to how close two or more measurements are to each other. Precision of a certain class is calculated as the ratio of true positive value of that class to the corresponding sum of true positive value and the false positive value of that class.

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

Sensitivity (which is also known as a true positive rate, TPR) is the proportion of data’s that are genuinely positive and are correctly classified. Sensitivity or recall is calculated as the ratio of true positive value of that class to the corresponding sum of true positive value and the false negative value of that class.

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

Specificity (also called as the true negative rate, TNR) is the proportion of actual negative data’s that are correctly classified. Specificity is given by the ratio of true negative value to the sum of true negative value and false positive value.

\[
\text{Specificity} = \frac{TN}{TN + FN}
\]

V. IMPLEMENTATION AND SIMULATION RESULT

The model is built with a multi-layer perceptron neural network consisting of 13 inputs nodes, 350 hidden neurons with 10 output node for ten class classification. The basic 13 coefficient features obtained from MFCC computation of ten person are concatenated and given as inputs to the neural network. The inputs given to the neural network is 13x22182 and the target is set at 10x22182.
The total number of data set for classification is 22182. The simulation result of ten class’s pattern classification is shown in the confusion matrix of Fig. 7.

![Confusion Matrix](image)

**Fig. 7 Simulation result of pattern classification**

Accuracy of the classifications is found to be 92.1% and the corresponding Precision and Sensitivity of various classes is also summarized in the table as shown in Fig.8. Precision varies between 88.7% and 94.1% and sensitivity between 89.5% and 95.6%.

![Table of Accuracy, Precision and Sensitivity](image)

**Fig. 8 Table of Accuracy, Precision and Sensitivity**

The specificity of different classes is also calculated and listed as shown in the Fig.9. It varies from 98.5% to 99.5% which is acceptable good.

![Specificity of ten classes](image)

**Fig. 9 Specificity of ten classes**

The ROC is a graph which summarizes the classifier performance and it is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis). The more the results lean towards the true positive rate the better is the classification. The simulated result of all the ten classes classification scoring the maximum true positive rate and the validation performance at 832 epochs is also shown in Fig.10.

![Receiver Operating Characteristic (ROC) and Validation Performance](image)

**Fig. 10 Receiver Operating Characteristic (ROC) and Validation Performance**

VI. CONCLUSION

When more number of data set is given to the network for adequate training, the network performance will improve and achieved high accuracy. 75% of the data’s are use for training, 15% for validation and another 15% for testing the network. Network classification performance in terms of accuracy, precision, sensitivity and specificity all score above 90% with less percentage of misclassification i.e only 7.9%. This model is built for a small number of database and it can be extended for N number of users applying the same methodology. The simulation result of various parameters and the calculated values are all shown in the implementation and simulation section.

REFERENCES


14. “Speaker Recognition Based on MFCC and BP Neural Networks” Yi Wang, Dr. Bob Lawlor, 28th Irish Signals And systems Conference Year: 20-21 June 2017, Pages: 1 - 4, IEEE Conference publication


