SPEAKER IDENTIFICATION SYSTEM USING CLOSE SET

Shweta Bansal[1], Ankur Hooda[2], Anima[3]

[1, 2, 3] *ECE, KIIT College of Engineering, Gurgaon,* ***bansalshwe@gmail.com, ankur.hooda45@gmail.com, animadahiya806@gmail.com***

Abstract

The present paper describes experiments conducted to evaluate the performance of speaker recognition. The experiments conducted using Neural Network shows that the complexity of speaker recognition increases when the numbers of speakers to be identified are large in numbers in the text independent situation. In the first experiment error rate was zero for 10 speaker's classification and performance was good. By increasing the number of speakers the error rate increased, classification and performance were poor. After 25 speakers, the error rate was very high. For 100 speaker's classification MATLAB NN tool did not support for display the confusion matrix. To overcome this problem the second experiment has been done. In this experiment a close set of 10 groups of 100 speakers (each group of 10 speakers) in terms of cell array in MATLAB has been defined and we observed that the best result of speaker identification was 100% in 20 continuous features of speaker's voice, but it increased time complexity. In the third experiment speaker's dialect and regions were also identified and classification performance was 100% at 97 epochs, validation performance was 0.0035046 at 91 epochs and the error rate was zero has found.

Index Terms: *text dependent, text independent, speaker identification, Neural Network, close set, MFCC, Speaker identification, close set.*

----------***----------

1. INTRODUCTION

Speaker recognition can be performed in two ways. In first way Text-dependent and Text-independent conditions. In addition there are two different types of speaker identification, the closed-set and the open-set identification. In closed-set identification, the sample of the test speaker is compared against all the available speaker samples and the speaker ID of the sample with the closest match is identified. In open-set the test speaker's sample is compared with the large number of reference samples which may or may not contain the test sample. It verifies that a given speaker is the one whom he claims to be. If it matches the set threshold then the identity claim of the user is accepted otherwise rejected. The identification strategy is based on matching the set threshold value for accepting or rejecting the speaker. [1][2][3].

Speaker recognition using Neural Network techniques are very complex for large number of speakers. In the present study three experiments have been done to find the performance of speaker recognition versus complexity of computation. In the first experiment the study of classification of 100 speakers uttered 10,000 sentences have been done by Neural Network (NN) using back propagation adaptive learning (Scaled conjugate gradient) method with two hidden layers. For classification of speakers 70% voice data for training, 15% for validation and 15% for testing the voice data have been used for getting the best results. The performances were better with less number of speakers but the complexity was increased by increasing number of speakers. To overcome this problem the second experiment was performed by grouping speakers in close sets of 10 groups for 100 speakers (10 speakers in each group) MATLAB has been used to conduct this experiment. This produce increased the performance of speaker identification system, but complexity was high. And in the third experiment dialect and regions of three non-native (South Indian, Punjabi and Haryanvi) speakers who uttered 150 text dependent Hindi sentences in NN [4] have been classified with 100% performance.

2. VOICE DATABASE COLLECTION

In the first and second experiments a corpus of 10,000 Hindi sentences uttered by 100 speakers (each speaker uttered 100 text independent sentences) was created. These sentences were taken from mobile communication database for the analysis of speaker's voice. In the third experiment, a corpus of 450 text dependent Hindi sentences uttered by three non native speakers i.e. 150 sentences uttered by South Indian, 150 by Panjabi and 150 by Haryanvi native speakers. All speakers were within 18years-60years age groups. Voice data were recorded using a head held moving coil microphone in a quite room. "PRAAT" software with sampling rate 16 kHz/16 bit has been used for the analysis of samples.

3. VOICE FEATURE EXTRACTION

Experiments are based on the average information about the parameters of one second speech utterance by an speaker.193 Mel Frequency Cepstral Coefficients (MFCC) were extracted as voice features with sampling rate was 100.The total number of filters used were 40(13linear filters+ 27 log filters) and the linear spacing between the filters was 66. 5 prosody related features consisted of average value of pitch(f0),intensity, duration, RMS value of sound pressure and power, these were extracted using PRAAT tool. MFCC values were calculated by MATLAB [8].

4. EXPERIMENT 1

4.1 Classification of 100 speakers using NN

A multilayer feed forward neural networks with one input layer, one output layer and two hidden layers have been used with 20 hidden neurons. The training method was Scaled Conjugate Gradient descent with momentum, Neuron Transfer Function for hidden layer and output layer were considered Log sigmoid Transfer Function. In this experiment 70% voice data have been used for training whereas 15% data each for validation and testing.

To begin with, speaker recognition was conducted with first 10 test speakers out of 100 speakers. The classification was 80% and the performance was 0.082485 at 36 epochs with zero error rate. By increasing the number of speakers i.e.by 15 the classification was 55% and the performance was 0.063405 at 43 epochs with 11.76% error rate .After 25 speakers it was very difficult to trace the confusion matrix for classification and error rate was very high. For 100 speaker's classification MATLAB NN tool did not support for display the confusion matrix. The variations in the results by increasing the number of speakers are shown in table 1:

Table 1: variations in performance of 100 speaker's Classification

Number of Speakers	Number of Samples	Number of Inputs	Confusion Rate(%)	Performance	Epochs	Rate of Error (%)
10	1000	199	80	0.082485	36	0
15	1500	199	55	0.063405	43	11.76
20	2000	199	20	0.04947	37	63.85
25	2500	199	7.6	0.0399	27	78.16
30 . . .	3000	199	Not Display	0.022525	21	91.34
100	10000	199	Not display	Not display	Not display	Not display

5. Experiment 2

5.1 Close Set Speaker Identification

To overcome this problem of recognising large number of speakers using NN technique, our experiment has been done by grouping speakers in closed sets . A close set of 10 groups of 100 speakers (a group of 10 speakers) in terms of cell array in MATLAB has been defined. Each cell represents a single group of close set. An associative memory for all 100 speakers' voice features in MATLAB Cell array in terms of close set [11] has been used.

We have followed the speaker identification algorithm shown in figure 1. This algorithm uses the close set of associative memory. All the voice features of 100 speakers have been taken into account in the associative memory [11][12]. The task is done by using the extracted feature of all speakers matching with the test speaker's voice features. If the match is found then test speaker is identified. Matching performed in the order of selection choosing 3, 5,10,15,20 continuous features of test speakers' voice.

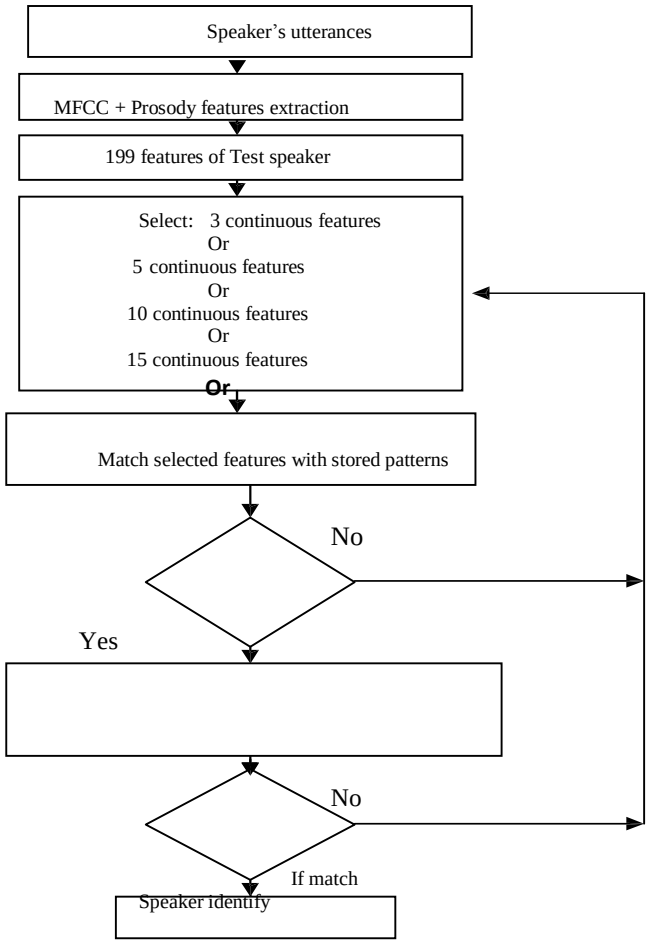

Figure 1: Function diagram of speaker recognition

shown in figure 1. This algorithm uses the close set of associative memory. All the voice features of 100 speakers have been taken into account in the associative memory [11][12]. The task is done by using the extracted feature ofall speakers

In the first iteration (i, i+1, i+2) match the features of test speaker with all speakers features, if the match found then select all those speakers. Otherwise go to the next iteration (i+1, i+2, i+3) up to the nth iteration. The same rule follow for the 5, 10, 15, 20 continuous features. The cell structure of the test speaker is shown in figure 2

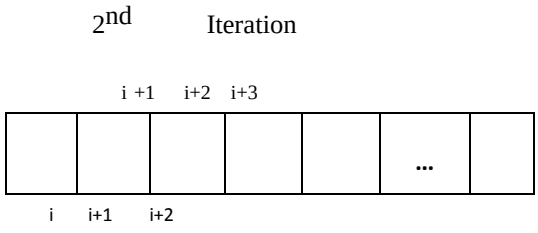

Figure 2: Selection of 3 continuous features

We have done 5 experiments. In the first experiment 3 continuous feature of test speaker have been taken as input. The identification gave the 52% performance and reduced time complexity because no more feature had to be searched from the associative memory for the selection of the speakers.

In second experiment (for 5 continuous features) the performance was 67% with increased time complexity. In further experiments we have increased the number of continuous features (i.e. 10, 15, and 20) of test speakers and the performance was 79%, 89% and 100% respectively. We noticed that the best result of speaker identification is 100% in
20 continuous features of speaker's inputs, but it increased the value of time complexity. The observations from the all experiments are shown in table 2 and performance graph in figure 3

Table 2: Performance of speaker identification algorithm

Number of continuous Feature	3 continuous Features	5 continuous Features	10 continuous Features	15 continuous Features	20 continuous Features
Time Complexity	$O(n^3)$	$O(n^5)$	$O(n^{10})$	$O(n^{15})$	$O(n^{20})$
Speaker Identification	52%	67%	79%	89%	100 %

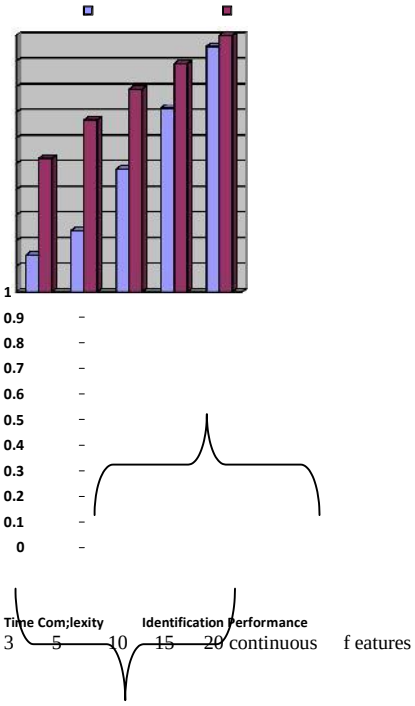

Figure 3: Performance graph

6. Experiment 3

6.1 Region and Dialect identification

One hundred fifty text dependent Hindi sentences recorded by three non native speakers i.e. 450 utterances were used for this study. 199 extracted features have been taken as inputs with one input layer, one output layer had three units which represented output categories i.e. South Indian, Punjabi & Haryanvi. Two hidden layers with 20 neurons have been used. The architecture of Multilayer feed forward network for three non native speaker's classification is shown in figure 4

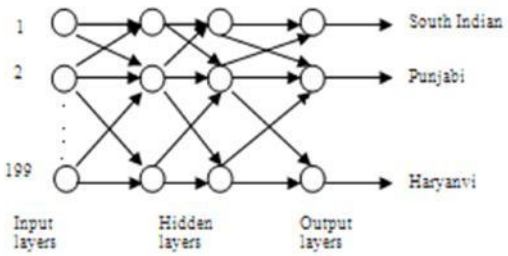

Figure 4: Multilayer feed forward network of three non native's speaker's classification

From the total voice database 314 sentences have been used for training, 68 sentences for validation and 68 sentences for testing. We observed that the classification performance was 100% at 97 epochs, validation performance was 0.0035046 at 91 epochs and the error rate was zero. The best performance was 100% because the numbers of speakers were very less i.e. three and the database was text dependent .The MFCC +

prosody features were differ for each and every non native speakers because of their dialects.The best validation performance for Region and Dialect classification is shown in figure 5

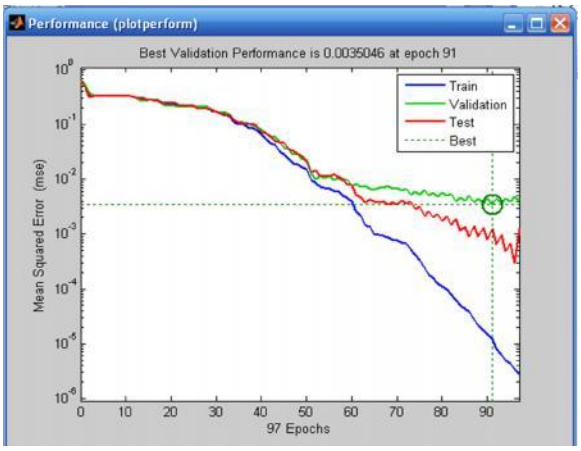

Figure 5: Best validation performance for Region and Dialect classification

CONCLUSIONS

In the first experiment speaker's classification can give better results with less number of speakers using Neural Network but the performance decreases by increasing the number of speakers. It will take either too much time for execution or the system may not response. For better results experiment 2 has been done with close set of speakers. That has produced better results but in more complex manner. In the third experiment dialect and region identification performance was 100 %, because only 3 non native speakers were there to be classified which were very less and the voice database was text dependent. Future improvement can be done as given below

I. Define the separate close set for male and female speakers.
II. Define the cluster of nearest distance of voice features from the neurons can placed inside the same cluster.
III. Used text dependent sentences for better results as we have discussed in our experiment 3.

In this order to define cluster, the result for recognition of speaker will be more accurate and also increase the performance with less time complexity, because the similar type of speakers' voice placed inside the same group. That overcomes the more searching of speakers. In our experiment we have designed the associative memory for voice features In earlier experiments the associative memory was used for pattern segmentation and large vocabulary for continuous speech recognition [11][12].

ACKNOWLEDGEMENTS

We would like to thanks Ms. Shweta Sinha, Shipra Arora and Minakshi Mahajan our team members for their valuable suggestions and support. Our thanks are due to Mr. A.K. Dhamija, for help in recording as well as maintaining the speech data. Also our thanks are extended to the supporting staff Mr. Gautam Kumar, for his continuous effort in the maintenance and development of text and speech corpora. Our very special thanks are to KIIT management Dr. Harsh Vardhan Kamrah and Ms. Neelima Kamrah for providing their constant encouragement, support and environment that was conducive for this work.

REFERENCES

[1] L.R.Rabiner and B.H.Juang"Fundamental of speech recognition"Prentice Hall Englewood cliffs,N.J.1993.
[2] L.R.Rabiner and R.w.schafer "Digital processing of speech signal" Prentice Hall Englewood cliffs,N.J.1978.
[3] Furui, "An overview of speaker recognition technology", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.
[4] B.Atal"automatic recognition of speakers from their voices"proc.IEEE,vol.64,pp.460-475April1976
[5] S.S.Agrawal"Emotion in Hindi speech analysis,perception and recognition"Oriental COCOSDA 2012.
[6] Samudravijaya K, Speech and Speaker Recognition: A tutorial, Proc. Int. Workshop on Tech. Development in
[7] Indian Languages, Kolkata, Jan 22-24, 2003.
[8] Samudravijaya K and Anshu Gupta, Text prompted speaker verification, J.Acoust. Soc. India, vol.30, pp. 214-217, 2002
[9] Vibha Tiwari"Mfcc & its application in speaker recognition" International journal merging technologies1(1):19-22(2010)
[10] Ran D. Jilka "Text independent speaker verification using covariance Modeling" IEEE Signal Processing Letters Vol 8, Page 97{99,2001}.
[11] ."Self organizing map and associative memory model hybrid classifier for speaker recognition" by M Inal, Y S Fatihoglu 6thSeminar on neural Network Applications in Electrical Engineering Neurel2002714 (2002)
[12] Deliang Wong, Joachim Buhamann "Pattern
[13] Segmentation in Associative Memory" Massachusetts Institute of Technology, Neural Computation 2,94-106 (1990)
[14] Z K Kayikci "Large vocabulary continuous speech recognition using associative memory and hidden Markov model" Proceeding SSIP'08 Proceedings of the 8th conference on Signal, Speech and image processing 2008