International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 6 Issue 9|| September 2017 || PP. 19-25

Speaker Recognition Using MFCC &Evolution with Different Classification Techniques

*Pravin G. Sarpate[1], Ramesh R. Manza[2]
[1](Department of CSIT, Dr. Babasaheb Ambedkar Marathwada University, India)
[2](Department of CSIT, Dr. Babasaheb Ambedkar Marathwada University, India)
Corresponding Author: Pravin G. Sarpate

Abstract : *This research paper is designed a unimodal biometric system based on speaker recognition. TheMel-Frequency Cepstral Coefficients (MFCC)is used for extracting the voice features. We applied this technique for identification of a person on KVKRG voice database. For this experiment, 500 phonemes of 50 subjects from KVKRG voice database were used. The database consists of phoneme in language spoken in multimodal biometrics research lab, Dept. of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University Aurangabad of Maharashtra. The phoneme of each subject is collected five times i.e. English alphabets (a – z) and five times i.e. digits (0 – 10).Three samples used for training and the rest two used for testing. Different classifiers were used like Linear Discriminate, SVM, k-NN, Ensemble Subspace Discriminate. The recognition rate for (a-z) is 98% and (0-10) is 96%. Experimental result shows that biometrics system gives improvement in the overall system performance, results quickly and accurate.*

Keywords: *Biometrics, Histogram of Oriented Gradients, MFCC, Multimodal biometrics, Verification*

Date of Submission: 25-08-2017 Date of acceptance: 09-09-2017

I. INTRODUCTION

Speaker recognition is the process of identifying a person on the basis of speechalone. It is a known fact that speech is a speaker dependent feature that enables us torecognize friends over the phone. Speaker recognitionmake it possible to verify the identity of persons accessing systems; allow automatedcontrol of services by voice, such as banking transactions; and control the flow of privateand confidential data.Today all over the world every person wants security of data, physical access etc. To solve security problem biometrics is solution. Biometrics means automatics identification of a person based on his/her physiological or behavioral characteristics.The voice is the combination of physical and behavior biometrics. There are two types of biometrics system i.e. Uni-modal and Multi-modal biometric system.Substantial progress has been achieved in voice-based biometrics in recent times but a variety of challenges remain for speech research community.Here we develop biometric system for speaker recognition.

The proposed method will explain in brief on section two. The voice recognition processes will describe in the section three. The implementation and result will discuss in the section four and five respectively. The section six will contain the conclusion, at the end, the acknowledgment and the references.

II. PROPOSED METHOD

Several of biometric characteristics exist and are use in various applications. Each biometric has its advantage and disadvantages and the choice depends on the application. The proposed methodology is employed to extract the voice feature from Mel-Frequency Cepstral Coefficients. The length of voice feature vector is 13 which are sufficient for the voice recognition. The proposed system has four stages: preprocessing, feature extraction, store in database and apply different classifications. The following fig. 1 shows the detail about proposed method.

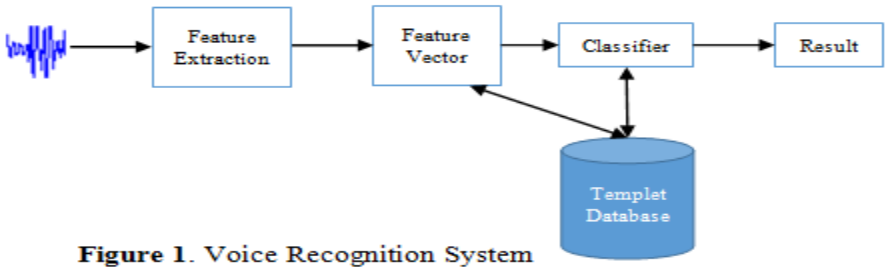

Figure 1. Voice Recognition System

III. VOICE BIOMETRIC

Speaker recognition is one of the most acceptable biometric because it is one of the most common methods of identification which humans use in their voice interactions. In recent years, speaker recognition has attracted much attention and its research has rapidly expanded by not only engineers but also computer scientist, since it has many potential applications in communication and automatic access control system. Especially, textual language has become extremely important in modern life, speech has dimensions of richness that text cannot approximate. Speech carries information on the several levels viz. speaker specific information, message being expressed as a sequence of words or phrases and information about the acoustic environment in which it was recorded. The speaker specific information can be the identity/sex/language or dialect, his/her attitude and possibly physical or emotional conditions of the speaker. The component which are related for speaker unique information are birth place, education place, economical position, social prestige, personality and anatomical structure of vocal apparatus [1-5].

1.1 VOICE RECOGNITION PROCESS AND ALGORITHM DETAIL

In general, the voice recognition system divided in three chores feature extraction, identification/verification and detection.

3.1.1 Acquisition

KVKRG Voice Database collected using the Windows 7 Operating system, MATLAB R2013a, Praat Version 5.3.01 size 9.35 MB (9,805,824 bytes), Text Editor Microsoft Office 2010. Hardware used Processor Intel(R)Core(TM)2Duo CPU T6600 @ 2.20GHz 2.20 GHz, installed memory (RAM) 3.00 GB, dynamic stereo headphone 40mm ferrite drive unit's frequency response 20-20,000Hz impedance 32 Ohms Sensitivity105db/mw Rated power 100 mw Power handing capacity 1000mw. Sampling frequency16000H, in the data recording.

3.1.2 FeatureExtraction*:*

The raw speech signals are complex and could not be suitable for input to the speaker recognition system therefore the requirement for a good front-end ascends. Acoustic model information is in compact form [9]. For extracting the compact features,the noisy information is removed by applying pre-processing [10] [11]. Though characters which show small amount of information are maximally close classes. All feature extraction techniques used similar Pre-processing steps [6].

3.1.3 Identification /Verification*:*

It is used to determine which speaker out of a group of known speakers produces the input voice sample. Speaker verification is used to determine who (he or she) the person is & it claims to determine according to his/her voice samples. This task is also known as voice verification or authentication or speaker authentication, talker verification or authentication [7-18].

3.1.4 Speakerdetection*:*

In this considered as a true or false unary decision problem, means only one result obtained [13].

3.1.5 Mel Frequency Cepstral Coefficients (MFCC)

Step 1: Pre–emphasis

This pre-emphasis is done by using a filter. This pre-emphasis filter is a first-order high-pass filter. In the time domain, with input x[n] and 0.9 _a _ 1.0, the filter equation is

$$Y[n] = X[n] - 0.95\,X[n-1] \quad (1)$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

Step 2: Framing

The process of segmenting the speech samples into a small frame with the length within the range of 20 to 40 msec. Adjacent frames are being separated by M (M<N). Typical values used are M = 100 and N= 256.

Step 3: Windowing

Speech signal are non-stationary statistical properties are always varying time to time. Here to extract spectral features from a small window of speech that characterizes a sub-phone and for which it can make the (rough) assumption that the signal is stationary. For extracting the waveform from the window, we used non-zero regions. If the window is defined as

W (n), 0 _ n _ N-1 where,

N = number of samples in each frame

Y[n] = Output signal

X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y(n) = X(n) * W(n) \quad (2)$$

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) 0 \le n \le N-1 \quad (3)$$

Step 4: FFT

It is used to convert frame of glottal pulses U[n] N samples from time domain into frequency domain and the vocal tract impulse response H[n] in the time domain:

$$Y(w) = FFT[h(t) * X(t)] = H(w) * X(w) \quad (4)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively

Step 5: Mel Filter Bank Processing

The voice signals come from FFT spectrum are does not follow linear scale because it is in a wide format. As shown in Fig.2, the bank of filters allowing to Mel scale.

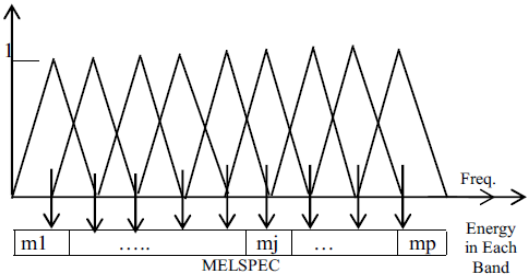

Figure2. Mel scale filter bank

As shown in Figure 1, a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. The triangular shape shows magnitude filters of each frequency and all are equal to center frequency which decreases linearly of two adjacent filters [14]. Then, each filter output is the sum of its filtered spectral components. After that the following equation is used to compute the Mel for given frequency f in HZ:

$$F(Mel) = [2595 * log\,10\,[1 + f]\,700] \quad (5)$$

Step 6: Log

Mel spectrum value magnify by using Log i.e. $\log 10\,[1 + f]$

Step 7: Discrete Cosine Transform

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called MFCC. Each set of input utterance is a sequence of transformed coefficient of acoustic vector.

Step 8: Delta Energy and Delta Spectrum

Cepstral features change over a time which indicates slope of frame change and its transition. Energy or velocity of signals X is a time series t window, features are 13 delta (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added is represented

$$\text{Energy} = \sum X2\,[t] \quad (6)$$

Each of the 13 delta features represents the change between frames as shown in equation 7 corresponding cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

$$d(t) = \frac{c(t+1) - c(t-1)}{2} \quad (7)$$

After adding energy, and then delta and double-delta features to the 12 cepstral features, we end up with 39 MFCC features.

IV. EXPERIMENT RESULT

4.1 Experimental Setup

For the speaker recognition technique experiments has been done using well-known feature extraction algorithm MFCC on KVKRG Voice Database.

Database: KVKRG VOICE DATABASE

Source: This database is developed by Multimodal Biometric Research Lab under the UGC SAP project, in the Department of Computer Science & Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India.

Table 1. KVKRG Voice database properties descriptions

Properties	Descriptions
# of subjects	50
# phoneme (speech)	500
Language	English (Indian)
Data type	Wave
Speech type	Spoken Speech
Gender	Male, Female
Age	20-40
Recording Condition	Normal
Sampling Frequency	16000Hz
Subject region	Maharashtra

This database contains human (male & female) 20 to 40 years age group speech (phoneme). The phoneme for each subject collected five times i.e. English alphabets (A - Z) and digit(0-10).

V. RESULT& DISCUSSION

1.2 Experiment 1 KVKRG Voice (A –Z)

Table 2. Different classifier, accuracy and time in sec for a-z

Sr. No.	Name of Classifier	Accuracy in Percentage	Training Time in Sec
1	Linear Discriminant (Linear Discriminant)	98	1.8937
2	SVM (Linear SVM)	86	22.776
3	SVM (Cubic SVM)	88	20.744
4	SVM (Medium Gaussian SVM)	92	18.804
5	SVM (Coarse Gaussian SVM)	83	18.83
6	KNN (Fine KNN)	95	1.6249
7	KNN (Weighted KNN)	94	0.13303
8	Ensemble (Bagged trees)	83	1.8626
9	Ensemble (Subspace Discriminate)	98	1.5106
10	Ensemble (Subspace KNN)	97	1.1251

For this experiment,the database is contained of 50 subjects and 5 samples of each. Here we have taken 3 samples for training i.e. 150and 2 samples for testing i.e. 100.First,the Ensemble (Bagged Trees) and SVM (Coarse Gaussian SVM) classifiers were applied which are obtained the same recognition rate 83%. Second time the SVM (Linear SVM) is applied and it got the RR 86% which is increase by 3%. Third time the SVM (Cubic SVM) is appliedand it got the RR 88%. It is increase by 2%. Fourth time theSVM (Medium Gaussian SVM) gives RR 92%. Fifth time the KNN (Weighted KNN) and (Fine KNN) give RR 94% and 95% respectively.The Ensemble (Subspace KNN) and (Subspace Discriminate) give RR 97% and 98% respectively. Then,the Linear Discriminant gives RR 98%. The highest recognition rate is obtained by Linear Discriminant and Ensemble (Subspace Discriminate). The following Fig. 3shows the comparison between different classifier accuracy and training time for KVKRG voice (a-z) database, Fig. 4 shows scatter plot of Linear Discriminate and Fig. 5 shows parallel coordinates plot of Linear Discriminate.

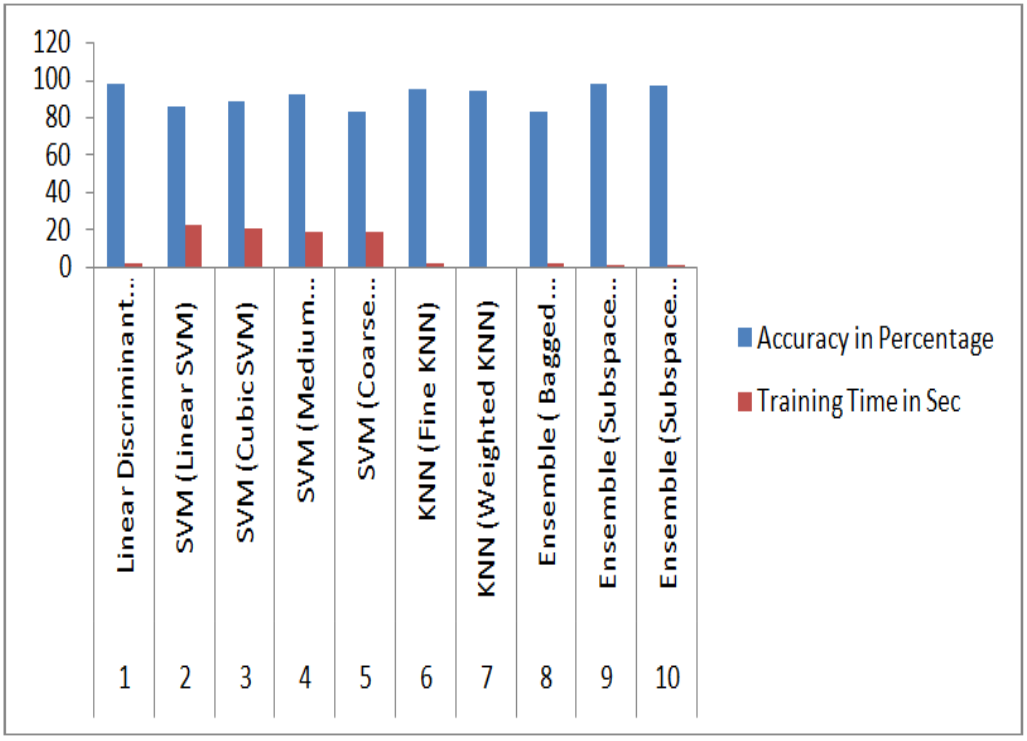

Figure3. Comparison between different classifier accuracy and training time for a-z

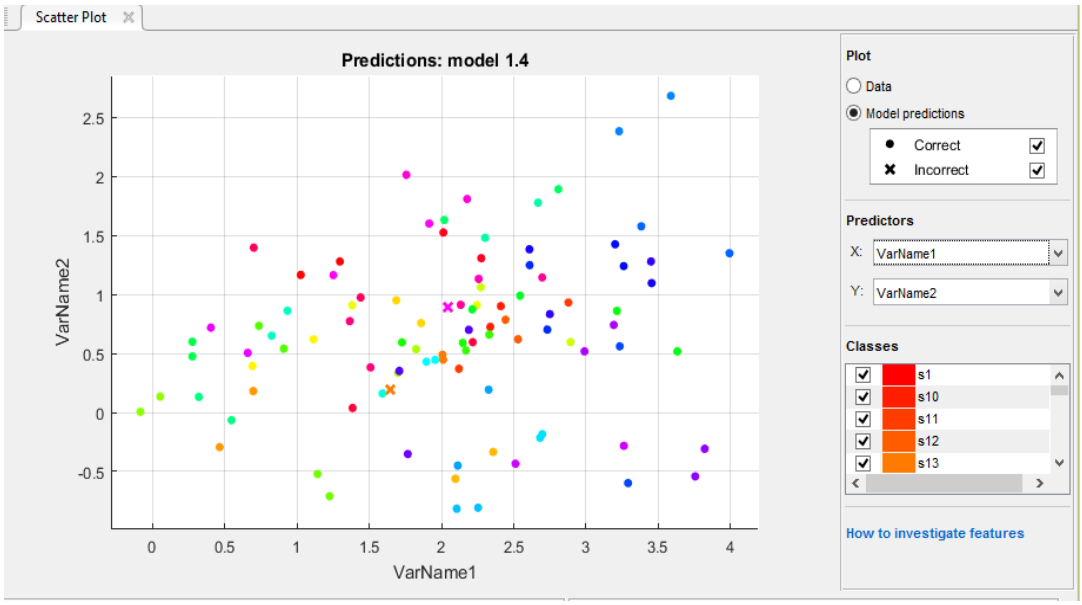

Figure4. Scatter plot of Linear Discriminant for a - z

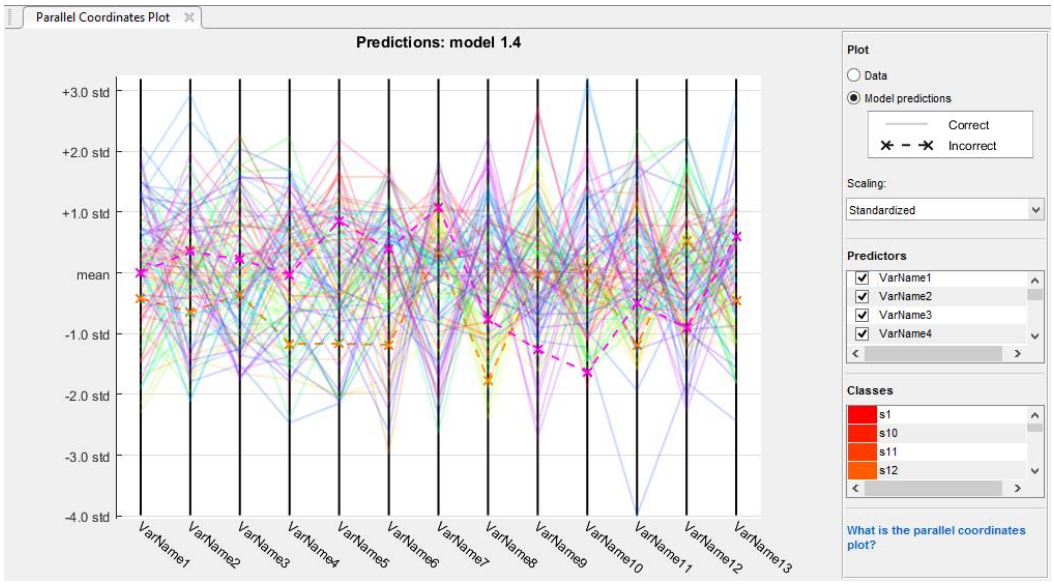

Figure5.Parallel coordinates plot of Linear Discriminant for a - z

5.2 Experiment 2 KVKRG Voice (0 –10)

Table 3. Different classifier, accuracy and time in sec for 0-10

Sr. No.	Name of Classifier	Accuracy in Percentage	Training Time in Sec
1	Linear Discriminant (Linear Discriminant)	95	1.593
2	SVM (Linear SVM)	78	23.952
3	SVM (Quadratic SVM)	76	20.315
4	SVM (Cubic SVM)	74	21.108
5	SVM (Medium Gaussian SVM)	88	20.528
6	SVM (Coarse Gaussian SVM)	77	20.279
7	KNN (Fine KNN)	92	0.5348
8	KNN (Weighted KNN)	91	0.11752
9	Ensemble (Subspace Discriminate)	96	1.5126
10	Ensemble (Subspace KNN)	84	1.1039

In this experiment,the database is contained 50 subjects and 5 samples of each. Here, three samples were taken for training i.e. 150 and 2 samples for testing i.e. 100. First,the SVM (Cubic SVM),(Quadratic SVM), (Coarse Gaussian SVM), (Linear SVM), Ensemble (Subspace KNN), (Medium Gaussian SVM)were applied and it obtained 74%, 76%, 77%, 78%, 84%, 88% as recognition rate respectively. Then,theKNN

(Weighted KNN), (Fine KNN), Linear Discriminant, Ensemble (Subspace Discriminate)were applied and it obtained the RR as 91%, 92%, 95%, 96%. The highest recognition rate is obtained by Ensemble (Subspace Discriminate). The following Fig. 6 shown the comparison between different classifier accuracy and training time for KVKRG voice (0-10) database, Fig. 7 shown scatter plot of Linear Discriminate and Fig. 8 shown parallel coordinates plot of Linear Discriminate.

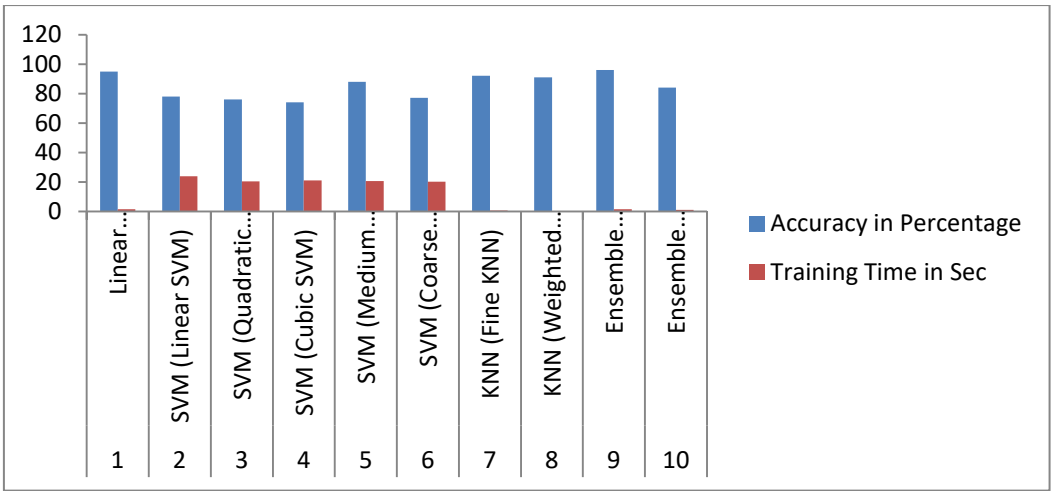

Figure6. Comparison between different classifier accuracy and training time for 0-10

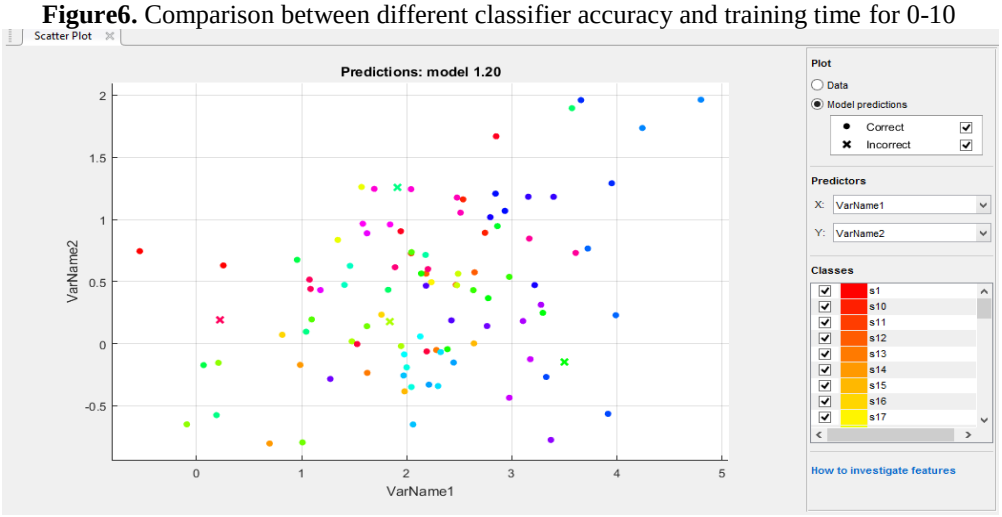

Figure7. Scatter plot of Ensemble (Subspace Discriminate) for 0 - 10

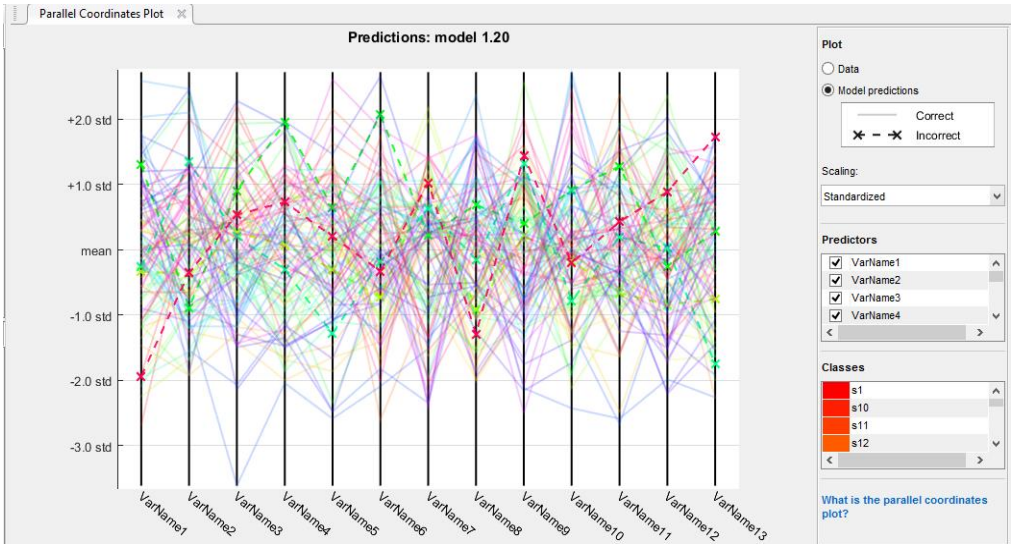

Figure8. Parallel coordinates plot of Ensemble (Subspace Discriminate) 0 - 10

VI. CONCLUSION

As we gone through the experimental results, we found that recognition rate for some classifier increase sometimes and decrease, sometimes same. For taking the result first we extract the features of KVKRG voice database (a–z) with the help of MFCC and apply the Linear Discriminate & Ensemble (Subspace Discriminate) classifiers we got highest recognition rate is 98%. When we applied Ensemble (Subspace Discriminate) classifiers on KVKRG voice database (0-10) we got highest recognition rate is 96%.

Therefore, it is concluded that speaker recognition rate is depend on the which database isused, on which condition it is created, no. of samples per subject and more important thing which classifier is used to classify the data. In our result, it is observed and concludes that, the Linear Discriminate & Ensemble (Subspace Discriminate) is more suitable for KVKRG voice database (a-z) and Ensemble (Subspace Discriminate) is for KVKRG voice database (0-10) as compare to other classifiers.

ACKNOWLEDGMENT

We are highly thankful to UGC SAP-II DRS PHASE-I:BIOMETRIC MULTIMODAL SYSTEM DEVELOPMENT, Research Laboratory, Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS), India, for our research work and KVKRG Voice database.

REFERENCES

[1] Bansod,N.S.,Dadhade, S.B,Kawathekar, S.S. Kale, K.V., Speaker Recognition Using Marathi (Varhadi) Language Intelligent Computing Applications (ICICA), International Conference, 2014,pp.421,425, 6-7.
[2] S. Furui, Cepstral analysis technique for automaticspeaker verification,IEEE Trans. Acoust,Speech,Signal and Processing, vol. 29, Biometrics ,Foundation Documents, NationalScience & Council,1981,pp.254-272.,P-92.
[3] N.S. Bansod, S.B. Dabhade,et.al,Review Of DifferentTechniques For Speaker Recognition System, Advancesin Computational Research, ISSN: 0975-3273 & EISSN:0975-9085, Volume 4, Issue 1, 2012,pp.-57-60.
[4] A. K. Jain, et al. Biometrics Recognition: Security and Privacy Concerns, IEEE Security & Privacy, 2003, pp. 33-42.
[5] A. N. Mishra, et.al., Robust Features for Connected Hindi Digits Recognition, International Journal of Signal Processing, Image processing and Pattern Recognition, 2011, Vol. 4, No. 2.
[6] Pawan Kumar, et.al., Spoken Language Identification Using Hybrid Feature Extraction Methods, Journal Of Telecommunication, 2010, Volume 1,Issue 2.
[7] Pawan Kumar, Mahesh Chandra, Speaker Identification Using Gaussian Mixture Models, MIT International Journal of Electronics and Communication Engineering, 2011, Vol. 1, pp. (27-30), ISSN 2230-7672©MIT Publications.
[8] Zaidi Razak, Noor Jamilah Ibrahim, et.al., Feature extraction using mel frequency cepstral coefficient (MFCC), Universitiy Malaya, 2010, ISSN 2151-9617.
[9] A. K. Jain, et al. Biometric Template Security, EURASIP Journal on Advances in Signal Processing, vol.2008:578416
[10] AzizahAbdManaf, AkramZeki, MazdakZamini, SuriayatiChupratEyas EL-Qawasmeh (Eds.), Informatics Engineering and Information Science, International Conference , 2011, ICIEIS, Proc. Part-I P No. 49.
[11] Lindasalwa Muda, MumtajBegam and I.Elamvazuthi, Voice Recognition Algorithms using MFCC and DTW techniques, journal of computing, vol.2, issue 3, 2010, ISSN 2151-9617.
[12] Anjugam M, Kawita M, et.al., Design and Implementation of Voice Control System For Wireless Home Automation Networks,ICCCE, 2012, ISBN 978-1-4675-2248-9.
[13] Matthew A. Turk and Alex P. Pentland, Face Recognition Using Eigenfaces, Vision and Modeling Group, The Media Laboratory Massachusetts Institute of Technology 1991.
[14] A. Ross, K. Nandakumar, and A. K. Jain, Handbook of Multimodal biometrics, Springer-Verlag, 2006.
[15] Peng Yang, Shiguang Shan, Face Recognition Using Ada-Boosted Gabor Features, IEEE International Conference on Automatic Face and Gesture Recognition, Korea, May 2004, 356-361.
[16] Anissa Bouzalmat and ArsalaneZarghili, Facial Face Recognition Method using Fourier Transform Filters Gabor and R_LDA, International Conference on Intelligent Systems and Data Processing (ICISD)2011.
[17] Chengjun Liu and Harry Wechsler, Independent Component Analysis of Gabor Features for Face Recognition, IEEE Transactionson Neural Networks, July 2003, Vol. 14, No. 4.
[18] R. Jarina, J. Polacky, P. Pocta and M. Chmulík,Automatic speaker verification on narrowband and wideband lossy coded clean speech, IET Biometrics, vol. 6, no. 4, pp. 276-281,72017.doi: 10.1049/iet-bmt.2016.0119.

International Journal of Engineering Science Invention (IJESI) is UGC approved Journal with Sl. No. 3822, Journal no. 43302.

Pravin G. Sarpate. “Speaker Recognition Using MFCC &Evolution with Different Classification Techniques .” International Journal of Engineering Science Invention (IJESI), vol. 6, no. 9, 2017, pp. 19–25.