International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. **06** Special Issue **01** | **2021**

Speaker Verification using Machine learning

Vipin J Gawande[1] M. U. Karande[2]

ME,Student,CSE,VBKCOE,Malkapur,Maharastra,India
Assistant Professor,CSE,VBKCOE,Malkapur,Maharastra,India

ABSTRACT

Speaker verification is the process of identifying identity of a person from speech signal. It works on principle that every person has unique voice characteristics which can be used to authenticate them. Speaker verification system consists of preloaded database of the person's voiceprint (template).At the time of verification the system compares input voice print with the pre stored voiceprint of that person in database and verifies the person. Speaker identification and speaker verification both systems identifies the speaker identity but the fundamental difference between the two is that in verification its 1:1 match i.e. the system compares input voice with one specified template voice print whereas identification refers to 1:N match where system compares input with all stored templates. Here is proposed a speaker verification system in which it authenticates the speaker. There are many approaches towards developing a speaker verification system like GMM-UBM, i-vector based, deep learning algorithms etc. Proposed here is i-vector based speaker verification system.

1. INTRODUCTION

Speaker verification is the process of verifying the claimed identity of a speaker from the speech signal from the speaker (voiceprint). This is also known as speaker authentication or detection. In speaker verification task, the user claims an identity and the system's task is to accept or reject the claimed identity, Speaker authentication has 2 main parts:

1) **Enrollment:** It is the process which creates and memorizes the speaker's voiceprint or speech signal. Enrollment is carried out just once, in the initial phase, and involves the recording of words pronounced by the user.

2) **Verification:** In this phase, the identity claim is accepted or rejected. Verification is performed each time access to the service is required, and it compares the voice characteristics of the speaker undergoing authentication with the voiceprint previously created for that identity. There are two types of speaker verification systems. One is text-dependent speaker verification – The text to be spoken is already known to the system. Knowing the text can improve system performance. E.g. prompted phrase, password phrase, fixed phrase. Other is Text-independent speaker verification – The uttered text is unknown to the system. Hence, there are less restrictions on user so as to what to speak. E.g. User initiated phrase, conversational speech.

2. DATABASE

The project involves the use of the TIMIT database. This is a database that contains speech signals recorded from 630 speakers containing eight major dialects of American English, each reading 10 phonetically rich sentences. So we intend to use a certain number of sentences spoken by certain number of speakers and identify all the Mel Frequency Cepstral Coefficients of the speaker and form the feature matrix using these coefficients. Then these sentences can be compared with various test cases to verify the speaker. The system uses the GMM-UBM method for training and testing of the data. GMM-UBM uses the combination of Gaussian Mixture Model with the Universal Background Model. This is the model used by most of the speaker verification models as it is robust, and its implementation is pretty easy. GMM has become the standard classifier for text independent speaker verification systems and UBM is a model in a speaker verification system to represent general, person independent, channel independent feature characteristics to be compared against a model of speaker-specific feature characteristics when making an accept or reject decision. This project can be easily implementable in offices, banks laboratories etc. as a security measure.

3. PROPOSED METHODOLOGY

Speaker verification, or authentication, is the task of confirming that the identity of a speaker is who they purport to be. Speaker verification has been an active research area for many years. An early performance

breakthrough was to use a Gaussian mixture model and universal background model (GMM-UBM) on acoustic features (usually mfcc). For an example, see Speaker Verification Using Gaussian Mixture Model. One of the main difficulties of GMM-UBM systems involves intersession variability. Joint factor analysis (JFA) was proposed to compensate for this variability by separately modeling inter-speaker variability and channel or session variability. However, discovered that channel factors in the JFA also contained information about the speakers, and proposed combining the channel and speaker spaces into a *total variability space*. Intersession variability was then compensated for by using backend procedures, such as linear discriminant analysis (LDA) and within-class covariance normalization (WCCN), followed by a scoring, such as the cosine similarity score. proposed replacing the cosine similarity scoring with a probabilistic LDA (PLDA) model and proposed a method to Gaussianize the i-vectors and therefore make Gaussian assumptions in the PLDA, referred to as G-PLDA or simplified PLDA. While i-vectors were originally proposed for speaker verification, they have been applied to many problems, like language recognition, speaker diarization, emotion recognition, age estimation, and anti-spoofing. Recently, deep learning techniques have been proposed to replace i-vectors with *d-vectors* or *x-vectors*. Audio Toolbox in MATLAB provides ivector System which encapsulates the ability to train an i-vector system, enroll speakers or other audio labels, evaluate the system for a decision threshold, and identify or verify speakers or other audio labels. In proposed approach you develop a standard i-vector system for speaker verification that uses an LDA-WCCN backend with either cosine similarity scoring or a G-PLDA scoring.

TIMIT corpus dataset is used for development of speaker verification system.TIMIT consists of speech data collected from 630 speakers, speaking English, and belonging to 8 major dialect regions of the United States. In this, 70% of the speakers are male and the remaining 30% are female. 10 sentences are spoken by each speaker and hence they have a total of 6300 sentences. These 10 sentences represent around 20-30 seconds of speech material per speaker. The speech was digitalized at a sample rate of 20 Hz using Digital Sound Corporation DSC 200 with anti-aliasing filter at 10 kHz. The speech was then digitally filtered and down-sampled to 16 kHz.

In order to make speaker verification system robost two concepts were taken into consideration. One is noise and other is imposters. Noizeus dataset consists different types of noise at various SNRs. Airport 0db, Station 5db and Restaurant 10 db noise dataset was taken from noizeus and added to TIMIT train and test dataset at 0db, 5db, 10db and 15 db SNR. Speaker verification system was trained using clean speech files and tested using noisy speech files to check its robostness. Another important concept is imposter handling. Imposters are the one who try to mimic the actual speaker or the one whose voice prnts are similar to that of actual speaker. The proposed system was tested using imposters of the enrolled speakers along with the speech files of actual speaker to check its robostness. The results shows EER obtained using noisy dataset as well as using imposters for testing.

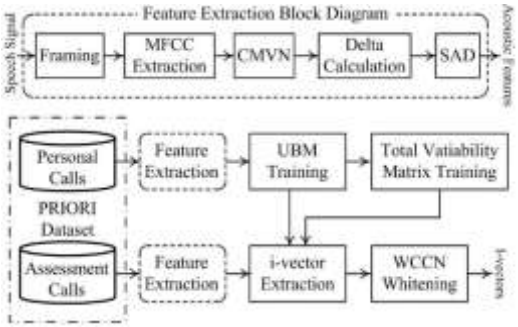


Figure 1: I Vector Framework for Speaker Verification

Speaker identification: It is the process of determining which speaker is speaking the given utterance.
Speaker verification: It is the process of accepting or rejecting the identity claim of a speaker. [17]

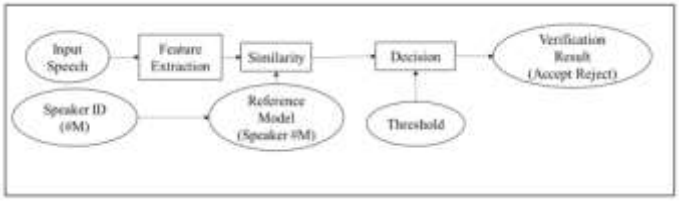


Figure 2: Speaker verification System

Results and Discussion: - The TIMIT database in hand has up to 6300 spoken sentences from speakers belonging to 8 different dialect regions of the United States. This database is broken down to a subset A total of 39 speakers were utilized from this database for the implementation for the work. The dialect distribution of these speakers according to the dialects is as shown in Table 1.

Dialect Region		**Number of male**	**Number of female**	**Total number of speakers**
Name of the region	**Code (DR)**			
New England	1	31(63%)	18(27%)	49(8%)
Northern	2	71(70%)	31(30%)	102(16%)
North Midland	3	79(67%)	23(23%)	102(16%)
South Midland	4	69(69%)	31(31%)	100(16%)
Southern	5	62(63%)	36(37%)	98(16%)
New York City	6	30(65%)	16(35%)	46(7%)
Western	7	74(74%)	26(26%)	100(16%)
Army Brat	8	22(67%)	11(33%)	33(5%)
Total number of speakers		438(70%)	192(30%)	630(100%)

The above 39 speakers are divided to fall into the following categories: UBM set, Enrollment set and Test set. As each speaker consists of 10 utterances, all these 10 utterances are present in the UBM set, 5 of these are in Enrollment set and 2 in Test set. We have performed concatenation of the 5 utterances in Enrollment set together, and that of 2 files in Test set. Concatenation helps in increasing the duration of the speech data and thus, in turn, increases the MFCC features obtained later during the processing of the data. The UBM building for 39 speakers was a lengthy process which took up more than 2 hours to store the features and projector file in hdf5 format on the computer drive. However, UBM building took up 13 seconds. The enrollment was carried out. 39 speakers (17 males and 22 females) are enrolled. A test speaker is considered and is tested against all speakers, giving a similarity score. In this case, there are some true positive and true negative speakers.

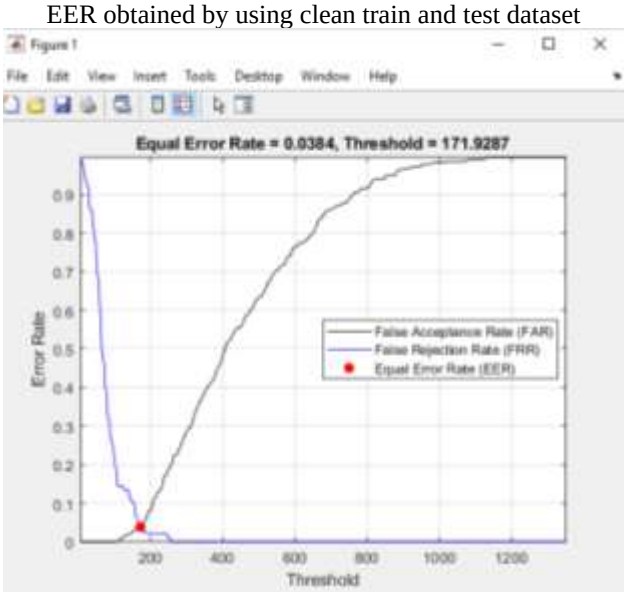
Comparison of EER obtained using different features

	Features	EER
1	linearSpectrum	0.2016
2	melSpectrum	0.2484
3	linearSpectrum + melSpectrum	0.2046
4	barkSpectrum	0.2541
5	erbSpectrum	0.2701
6	gtcc + gtccDelta + gtccDeltaDelta	0.0426
7	spectralcentroid + spectral crest	0.3507
8	Decrease + Flux + Entropy + Flatness	0.2997
9	Kurtosis + Rolloff + Skewness + slope + spread	0.2997
10	mfcc + mfcc delta + mfcc delta delta	0.0312
11	pitch	0.2545
13	mfcc + mfcc delta + mfcc delta delta + pitch	0.051
14	mfcc + gtcc	0.0297

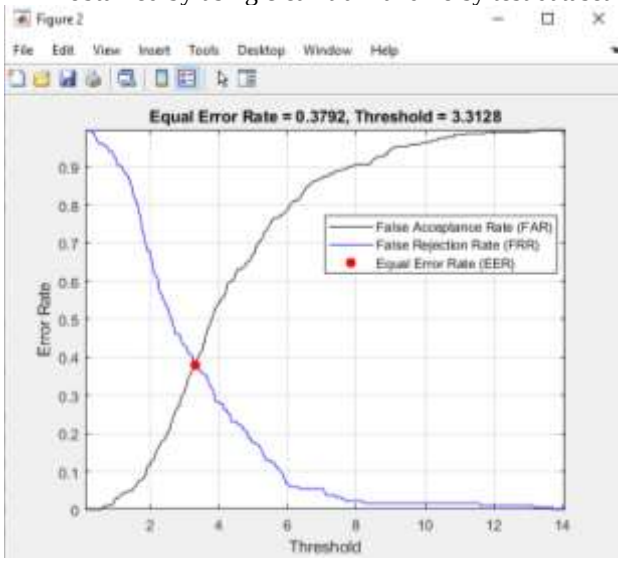
Results obtained using mfcc, mfccDelta, mfccDeltaDelta features with noisy test dataset

airport0				station5				restaurant10			
0	5	10	15	0	5	10	15	0	5	10	15
0.2971	0.3657	0.2686	0.2938	0.3799	0.3716	0.2686	0.3172	0.4201	0.2201	0.2828	0.2544

EER obtained by using clean train and test dataset



EER obtained by using clean train and noisy test dataset



4. **CONCLUSION**

The number of mixtures required in GMM is one of the most important parameters and the initialization of GMM is a very critical step for getting successful results ahead. This initialization can be random or derived from a prior knowledge, as done in the project. The choice of the clustering algorithm is also vital. It is, therefore, enough that the data is modeled by any reasonably good clustering algorithm, as far as the number of Gaussians is enough for the given training data. Using the UBM based modeling is seen to be very efficient. The GMM-UBM based system also depended on the quality of the data provided to it and its recording environment (which in this case is noise-free environment). In this project, we considered certain number of speakers from the TIMIT database having variations in the gender, dialect regions, accents and pronunciations. Initially, we started with 4 speakers and eventually kept on increasing the number of speakers. We also see that on increasing the number of speakers in UBM set, the EER of the system reduces. Hence, greater the data in the UBM set, lower is the EER of the system. The equal error rate was verified in both MATALB R2020.

5. REFERENCES :

[1] *Ritesh Tiwari* ***"Speech Recognition Market Players (7 Global, 10 Chinese) Review and Analysis"*** blog posted on www.prnewswire.com *April 2016*
[2] MIT Lecture #5 Session 2003 on ***"Speech Signal Representation"*** *spring 2003*
[3] *Homayoon Beigi* ***"Fundamentals of Speaker Recognition"*** DOI 10.1007/978-0-387-77592-0 Springer, *December 2011*
[4] *J. Campbell,* ***"Speaker recognition: a tutorial"*** Proc. IEEE **85**(9), 1437–1462 (1997)
[5] *Idiap Research Institute,* ***"Spear: An open source toolbox for speaker recognition based on Bob"*** in Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing · *May 2014*
[6] *Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn,* ***"Speaker Verification Using Adapted Gaussian Mixture Models"*** in Digital Signal Processing 10, 19–41 (2000)
[7] *Davis, S. Mermelstein, P. (1980),* ***"Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences."*** In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366
[8] *X. Huang, A. Acero, and H. Hon.,* ***"Spoken Language Processing: A guide to theory, algorithm, and system development."*** Prentice Hall, *2001*.
[9] *HTK,* ***"Model Adaptation using MAP"*** ICSI, Berkeley.
[10] *APSIPA Distinguished Lecture on* ***"Speaker Verification - The present and future of voiceprint based security"*** April 2017.
[11] *Hussein Sharafeddin, Mageda Sharafeddin, Haitham Akkay,* ***"Voice verification system for mobile devices based on ALIZE/LIA-RAL"*** 4th International Conference on Pattern Recognition Applications and Methods, At Lisbon, Portugal, *January 2015*
[12] *D.A. Reynolds,* ***"Experimental evaluation of features for robust speaker identification."*** IEEE Trans. Speech Audio Process. **2**(4), 639–643 *(1994)*
[13] *B.G.B. Fauve, D. Matrouf, N. Scheffer, J.F. Bonastre, J.S.D. Mason,* ***"State-of-the-art performance in text-independent speaker verification through open-source software."*** IEEE Trans. Audio Speech Lang. Process. **15**(7), 1960–1968 *(2007)*
[14] *Mitchell McLaren, Luciana Ferrer, Diego Castanl, Aaron Lawson,* ***"The Speakers in the Wild (SITW) Speaker Recognition Database"*** INTERSPEECH 2016, September 8–12, 2016, San Francisco, USA
[15] *Dehak, N., Kenny, P., Dumouchel. P., Dehak, R., Ouellet. P.,* ***"Front-end factor analysis for speaker verification"*** IEEE Transactions on Audio, speech and Language Processing 2011
[16] *Kenny, P., Ouellet, P., Dehak, N., Gupta, V. and Dumouchel, P.,* ***"A Study of Inter-Speaker Variability in Speaker Verification"*** IEEE Transactions on Audio, Speech and Language Processing, 16 (5) July 2008 : 980-988
[17] *Najim Dehak and Stephen Shum, tutorial on* ***"Low-dimensional speech representation based on Factor Analysis and its applications"*** Spoken Language System Group MIT Computer Science and Artificial Intelligence Laboratory
[18] *Xugang Lu *, Jianwu Dang,* ***"An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification"*** ScienceDirect, Speech Communication 50 (2008) 312–322
[19] *Jagannath H Nirmal1*, Mukesh A Zaveri2, Suprava Patnaik1 and Pramod H Kachare3,* ***"A novel voice conversion approach using admissible wavelet packet decomposition"*** EURASIP Journal on Audio, Speech, and Music Processing 2013, 2013:28
[20] *Ahilan Kanagasundaram thesis on* ***"Speaker Verification using I-vector Features"*** Doctor of Philosophy, Queensland University of Technology