IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 2, February 2016

Automatic Speaker Recognition System

Mandeep Kaur[1], Simrat Kaur[2]

Student, Dept. of CSE, BBSBEC, Fatehgarh sahib, Punjab, India[1]

Assistant Professor, Dept. of CSE, BBSBEC, Fatehgarh sahib, Punjab, India[2]

Abstract: This paper describes the development and implementation of an automatic speaker recognition system. The system uses text independent speaker verification with MFCC features and GMM based speaker modelling for authenticating the user. The developed system has been used by the 52 samples of students on a regular basis. The system tends to have a very high accuracy of 99% with only 7.6 seconds per sample response time for a 10 second voice. The system only needed 10-20 seconds sample to accurately predict the user and also verify if the sample is from known voice or unknown voice also.

Keywords: User Identification, Speaker Identification, GMM.

I. INTRODUCTION

The speech conveys several levels of information. Primarily, the speech signal conveys the words or message being spoken, but on a secondary level, the signal also conveys information about identity of the talker. While the area of speech recognition is concerned with extracting the underlying linguistic message in an utterance, the area of speaker recognition is concerned with extracting the identity of the person speaking the utterance [21].

In automatic speaker recognition methods, the speaker to be recognized is usually required to speak the same utterance which was used to obtain the reference pattern for that speaker. However, such a restriction is not generally necessary for speaker recognition by humans [2] [15].

In speaker recognition, the basic requirements are extraction of a few features from the speech and then cluster the speech features of same user in one group and the different users in n different groups. An estimation maximization algorithm is then applied on a new test speech to identify the users. Several problems exist in a class based speech model [11]. Since there are N numbers of classes, an unknown speech is likely to match to one of them if its probability is maxima of all the probability, however small it may be. Simple techniques like probability thresholding can be applied to make sure unknown speech is not matched to any existing speech in the database [9].

Several techniques exist for modeling the speech of a user. One of the well known methods is Gaussian mixture model (GMM) [14]. The use of Gaussian mixture models for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities [1].

Another very important method for speech modeling is VQ or Vector Quantization. Vector Quantization is a lossy data compression method based on the principle of block coding [18] [19]. VQ can be thought of as a process of redundancy removal that makes the effective use of nonlinear dependency and dimensionality by compression of speech spectral parameters [3].

This paper describes: Section II provide the summary of Speaker Identification System. Section III gives the overview of experimental analysis. Section IV presents the performance analysis of system. Finally, Section V describes the conclusion.

II. SPEAKER IDENTIFICATION SYSTEM

A Speaker identification system has two phases: Training Phase and Testing Phase. In training phase, all the speech samples available in the speech database are pre-processed and MFCC feature vectors are obtained. These features are then modeled using GMM-UBM. In testing phase, the test speech sample is pre-processed and its features are extracted and Log likelihood ratio is taken between the test speaker and all available speaker models in the database.

The speaker is identified as the one who has maximum likelihood ratio as shown in Fig.1

Analog speech signals acquired through the microphone which gives digital representation of speech signal. Speech recording is the first step of implementation. Recording has been done by native speaker.

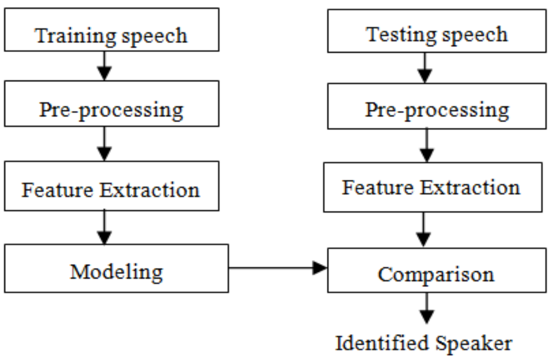

Fig. 1 Implementation of proposed approach

A. Pre-Processing

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. There are four steps in Pre-processing [16].

• *Background Noise Removal*

First step in the signal processing is Removal of Background Noise. By this process, background noise is removed from the data so that speech samples free from noise can be used as input for the further processing.

The starting and ending of any speech signal consists of considerable amount of silence region as show in Fig.2. These regions do not provide any speaker specific characteristic information. It also occupies considerable amount of space. Elimination of these regions helps to improve the performance of the system and also it helps to identify the starting and ending point of speech [6]. In this paper, silence discrimination is done based on energy thresholding [8] [21].

A signal varies in amplitude with time. The amplitude of voiced signal is much higher as compared to that of unvoiced signal [8]. The energy of signal is a representation of amplitude variation. The energy of voiced speech signal is much higher than the energy of unvoiced one. Similarly, threshold value is predetermined for energy speech signal.

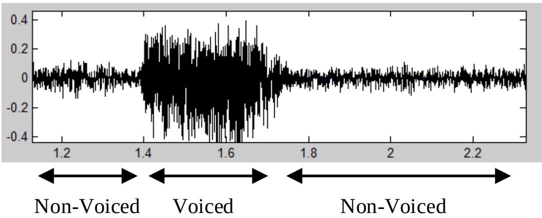

Fig.2: Amplitude variation in voiced and unvoiced signal

• *Pre-emphasis*

The digitized speech waveform has a high dynamic range and suffers from additive noise. In order to reduce this range and spectrally flatten the speech signal, pre-emphasis is applied. First order high pass FIR filter is used to pre-emphasize the higher frequency components [22]. The transfer function of the pre-emphasis digital filter is given by the following

$$Hp(z) = 1 - a^{-1} \qquad .....(1)$$

Where a is a constant, which has a typical value of 0.97. [22] [16]

• *Frame Blocking*

The speech signal is split into several frames such that each frame can be analyzed in the short time instead of analyzing the entire signal at once. The frame size is of the range 0-20ms [4]. Overlapping is done because on each individual frame, hamming window is applied. Hamming window gets rid of some of the information at the beginning and end of each frame. Overlapping reincorporates this information back into our extracted features [3].

• *Windowing*

Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. The choice of the window is a tradeoff between several factors [16]. In speaker recognition, the most commonly used window shape is the hamming window. The hamming window can be defined as equation (2)

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

Where *N* is the window length in samples [16].

B. Feature Extract

In this paper, GMM (Gaussian Mixture Model) is used for feature selection after we extract the feature. Better selection of feature vectors in GMM (Gaussian Mixture Model) is important from the training point of view [20] [11]. The robustness of the system in speech samples of considerable diversity of noise is important. GMM is a probabilistic model used for density clustering and estimation. GMM will help differentiation of feature vectors from the parent set of feature vectors. GMM is based on the hypothesis that all vectors are independent [8].

For the GMM method, the distribution of feature vectors extracted from a speaker's utterance X= $\{x_t\ 1 \leq t \leq T\}$ is modeled by a weighted sum of *K*mixture components which can be defined as [12].

$$P(x_t/\lambda) = \sum_{k=1}^{k} C_k N(x_t, u_k, \sum k) \qquad .... (3)$$

Where λ is the brief notation of the GMM parameters λ = $\{C_k, u_k, \sum k\}$1≤k≤K; and the mixture component N $(x_t, u_k, \sum k)$ denotes a Gaussian density function with the mean vector u_k and covariance matrix Σk defined as (4)

$$(x_t, u_k, \sum k) = \frac{1}{(2\pi)^{D/2} \sum K^{1/2}} \exp\left[-\frac{1}{2}(x_t - u_k) \times^{-1}_{k} \sum (x_t - u_k)\right] \qquad .....(4)$$

Where prime denotes vector transpose and *D* is the dimension of the vector x_t. [17] [11] [6]

C. Pattern Classification

In the Proposed System classification process are using Eigen vector values in PLDA (Probabilistic linear Discriminant analysis) [10] for better classification. The frame work using in which Gaussian Probabilistic linear Discriminant analysis is used for transformation of Eigen vectors of features reduces the undesired parts in the vectors [7]. Using Gaussian Probabilistic linear Discriminant analysis clearly helps to get better speaker discrimination.

D. Performance Parameters

Accuracy and Speed are the criterion for measuring the performance of an automatic speech recognition system which is described below:

• *Accuracy Parameters*

Word Error Rate (WER): The WER is calculated by comparing the test set to the computer-generated document and then counting the number of substitutions (S), deletions (D), and insertions (I) and dividing by the total number of words in the test set. [23][24]

$$\text{Formula: WER} = \frac{S+D+I}{N} \qquad .....(5)$$

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 3, Issue 2, February 2016

Word Recognition Rate (WRR): It is another parameter for determining accuracy. [23][24]

Formula: WRR=1-WER …… (6)

Single Word Error Rate (SWER) and *Command Success Rate* (CSR) are two more parameters to determine accuracy of a speech recognition system [23] [24].

- *Speed Parameter*

Real Time Factor is parameter to evaluate speed of automatic speech recognition [23][24].

Formula: $RTF = \frac{P}{I}$ .... (7)

Where P: Time taken to process an input Duration of input I e. g. RTF= 3 when it takes 6 hours of computation time to process a recording of duration 2 hours. RTF $\leq$ 1 implies real time processing.

III. EXPERIMENTAL ANALYSIS

A. *Database*

This system uses speech database of 52 speakers. Every speaker utters spontaneous speech of minimum 60 second duration and is recorded under open environment in the presence of various noisily. In order to show robustness of the proposed system, speakers are recorded at different days, hence the voice and channel characteristics may vary at different days. Hence, a total of 52 speech samples are recorded at a sampling rate of 16000Hz with the help of a condenser microphone.

B. Implementation

This system is developed in MATLAB Software. 52 samples of students were collected in phase one and then in phase two these samples is used to train a GMM-UBM model. The universal background model is trained by implementing the GMM-UBM algorithm in MATLAB.

Then, each student's model is created and labeled. In the current experiment, the voice sample is recorded individually in a wav file format and stored id is same as filename of the sample. Model is created using GMM and PLDA based technique. The whole system is text independent.

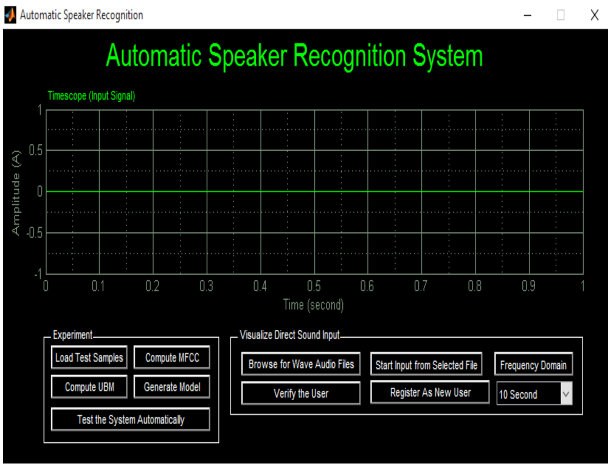

Fig.3: The GUI Interface of the "Automatic Speaker Recognition System

The Fig.3, show the GUI of the Automatic Speaker Recognition System. The system has two sections as shown "Experiment" and "Visual Direct Sound Input". In the "Experiment" section, Load the Test data, Compute MFCCs, Compute UBM, Generate each sample's model and Test the system automatically.

In the "Visual Direct Sound Input" section, the system can load a sample and test who the user is or register a new user if it does not exist in the database.

IV. PERFORMANCE ANALYSIS OF THE SYSTEM

The developed voice biometric system has been used by a group of 110 (100 male and 10 female) students on a regular basis. These students include undergraduate students enrolled for a particular course, research scholars and postgraduate students and are of the age group 20-35. The Table I shows the performance of the system for 2704 trials in terms of recognition rate and EER for various SV approaches. For contrast purpose, the performance of an offline GMM-UBM based SV system on the same training and test data is also shown in the Table I.

Table I. Performance comparison in terms of recognition rate and EER for various SV approaches

System	**Rec. rate**	**EER (in %)**
Automatic Speaker Recognition System	99%	1.125
SV System (i-vector based SV)	94.2%	7.65
Offline System (GMM-UBM based SV)	87.2%	13.2

The Proposed System resulted in obtaining an absolute improvement of 5 % compared to that of the i-vector based SV approach and GMM-UBM approach

A. *Effect of test data duration on the system performance*

The duration of test data required is an important parameter for any practical SV system. It is desirable to have a high per-forming system with the uses of minimum amount of test data. To analyse the effect of test data duration on the developed automatic Speaker recognition system, truncated the already recorded test data with different duration and re-run the evaluation. The performance of the system in terms of EER and recognition rate is given in the Table II. It can be observed from the table that the performance of the system is largely affected by the reduction in test data duration.

Table II. Performance of the SV System for Different Test Data Durations.

Duration(s)	**Rec.Rate (in %)**	**EER (in %)**	**Rec. rate (in %)**	**EER (in %)**
	Proposed System		**SV System**	
10	98.82	1.10	88.22	11.60
20	99.07	0.92	89.34	9.2
30	98.89	1.09	90.86	8.2
40	99.1	0.88	92.01	8.56
50	99.07	0.92	94.21	7.90

DOI 10.17148/IARJSET.2016.3214

B. Response time for the system

The response time is the time taken by the system to authenticate a person. It is noted that the developed system takes about 7.6 seconds to authenticate a person using test data of 10 seconds duration. Shows the break up of the total response time for various computational modules of the system.

III.CONCLUSION

This paper describe improved upon the previous work of “Speaker Recognition and Verification using i-vectors approach” and implemented a “GMM + PLDA” based system which is fast and very efficient system. The system is implemented in MATLAB and tested the results. This system is very fast. In the base work, the system took about 26 seconds to respond. This system took only 7.6 seconds to respond which is fairly fast. Also, the old approach required about 50 seconds of sample to reach a viable accuracy. Describe system needed 10-20 seconds sample to accurately predict the user and also verify if the sample is from known voice or unknown voice also.

This system tends to have a very high accuracy of 99% with only 7.6 seconds per sample response time for a 10 second voice

ACKNOWLEDGMENT

The authors would like to thank Assistant Professor **Simrat Kaur** and all my friends, reviewers and Editorial staff for their efforts in preparation of this paper.

REFERENCES

[1] Reynolds, Douglas, and Richard C. Rose. "Robust text-independent speaker identification using Gaussian mixture speaker models." Speech and Audio Processing, IEEE Transactions on 3.1 (1995): 72-83.
[2] Atal, B. S. "Text-Independent Speaker Recognition." The Journal of the Acoustical Society of America 52.1A (1972): 181-181.
[3] Gill, Manjot Kaur, Reetkamal Kaur, and Jagdev Kaur. "Vector quantization based speaker identification." International Journal of Computer Applications 4.2 (2010): 0975-8887.
[4] Matsui, Tomoko, and Sadaoki Furui. "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's." Speech and Audio Processing, IEEE Transactions on 2.3 (1994): 456-459.
[5] Reynolds, Douglas A., and William M. Campbell. "Text-independent speaker recognition." Springer Handbook of Speech Processing. Springer Berlin Heidelberg, 2008. 763-782.
[6] Gish, Herbert, and Michael Schmidt. "Text-independent speaker identification." Signal Processing Magazine, IEEE 11.4 (1994): 18-32.
[7] Zong, Feng. "Speaker Recognition Techniques." Applied Mechanics and Materials. Vol. 599. 2014.
[8] Furui, Sadaoki. "Speaker-dependent-feature extraction, recognition and processing techniques." Speech Communication 10.5 (1991): 505-520.
[9] Rose, Richard C., and Douglas Reynolds. "Text independent speaker identification using automatic acoustic segmentation." Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on. IEEE, 1990:293-296.
[10] Park, Alex, and Timothy J. Hazen. "ASR dependent techniques for speaker identification." INTERSPEECH. 2002: 1337-1340
[11] Kinnunen, Tomi, Evgeny Karpov, and Pasi Franti. "Real-time speaker identification and verification." Audio, Speech, and Language Processing, IEEE Transactions on 14.1 (2006): 277-288.
[12] David, Petr. "Experiments with speaker recognition using GMM." Proc. Radioelek-tronika (2002): 353-357.
[13] Kinnunen, Tomi, et al. "Comparing maximum a posteriori vector quantization and Gaussian mixture models in speaker verification." Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009:88-92
[14] M. Nishida, Masafumi, and Tatsuya Kawahara. "Speaker indexing and adaptation using speaker clustering based on statistical model selection." Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on. Vol. 1. IEEE, 2004:353-356.
[15] Nishida, Masanori, and Tatsuya Kawahara. "Unsupervised speaker indexing using speaker model selection based on Bayesian information criterion." Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on. Vol. 1. IEEE, 2003:172-175
[16] Hemakumar G, Punitha P “Speech Recognition Technology: A Survey on Indian Languages” International Journal of Information Science and Intelligent System, Vol. 2, No.4, 2013:1-38
[17] Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." Digital signal processing 10.1 (2000): 19-41.
[18] Gray, Robert M. "Vector quantization." ASSP Magazine, IEEE 1.2 (1984): 4-29
[19] Soong, Frank K., et al. "Report: A vector quantization approach to speaker recognition." AT&T technical journal 66.2 (1987): 14-26.
[20] Zivkovic, Zoran. "Improved adaptive Gaussian mixture model for background subtraction." Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 2. IEEE, 2004:28-31.
[21] S. Dey, S.Barman, R.Bhukya, R.Das, Haris, S.Prasanna and R. Sinha,“ Speech Biometric Based Attendance System”, IEEE Twentieth Nation’al Conference on Communications (NCC), 2014, Page(s) 1 – 6.
[22] Jianhang Qiu,” Audio signal analysis with filter by using FIR filter and detection of signal in noise” EECS451 Project Report, 2012, Page(s) 1-5
[23] H. B. Chauhan, Prof. B. A. Tanawala,” Comparative Study of MFCC And LPC Algorithms for Gujrati Isolated Word Recognition”, International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2015,Page(s) 822-826
[24] Vivek Sharma , Meenakshi Sharma,” A quantitative study of the automatic speech recognition technique”, International Journal of Advances in Science and Technology (IJAST) Vol I Issue I (December 2013), page(s) 34-39

BIOGRAPHY

Ms.Mandeep Kaur studied her B.Tech (Information Technology) from Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib and curently pursuing her M.Tech (E-Security) from Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib. Her research area includes Speaker Recongnition. She has presented a paper in International Conference.