An Improved Approach to Open Set Text-Independent Speaker Identification (OSTI-SI)

ShrutiSarika Chakraborty
School of Education Technology
Jadavpur University
Kolkata, India
shrutisarikachakraborty@gmail.com

Ranjan Parekh
School of Education Technology
Jadavpur University
Kolkata, India
rparekh@school.jdvu.ac.in

***Abstract*—This paper focuses on open set text independent speaker identification which is one of the most challenging subclass of Speaker recognition. The initial stage is similar to closed set speaker identification, where the distortion for each test voice against all train voices are determined. The distortions after normalization is set as decision criteria which eases the process of thresholding. The threshold variation which is mostly independent of dataset but dependent on the size of train data set and its values are quite similar for three datasets. The identification rate with balanced False Acceptance Rate(FAR) and False Rejection Rate(FRR) is 73-86%.**

Keywords--- Open set text-independent speaker identification; GMM-UBM (Gaussian mixture model-Universal Background model); Threshold; Vector Quantization (VQ);

I. Introduction

Speaker Recognition is the process of recognizing individual based on his/her voice. It can be classified into two types, speaker verification and speaker identification. Speaker verification is the process of verifying a speaker's claimed identity based on his/her already registered voice whereas speaker identification involves identifying whether a speaker's voice matches or not with any member of several registered voices [2]. Speaker verification is therefore a one to one matching process whereas speaker identification typically involves performing one to many matches. Both of these can either be text-dependent or text-independent. In the former case, a fixed and pre-defined text string is provided to the speaker based on which the voice patterns are compared, while in the latter case the text is arbitrary and typically unknown [3]. Speaker identification can again be of two types: open-set and closed-set. In the closed-set case, it is assumed that the test voice pattern is already present in the database and simply needs to be identified. In the open-set case, it is not known from beforehand whether the test voice pattern is actually present in the database or not [1]. Open-set matching is therefore more challenging [4] as it not only involves a comparison technique but also requires appropriate thresholds to prevent false matching of new voices with existing voices. Typical applications of closed-set speaker verification include voice based authentication systems while open-set speaker identification is required in surveillance and criminal investigations e.g. ransom callers.

The focus of this paper is to improve on existing techniques of open-set text-independent speaker identification (OSTI-SI). There are two main challenges to deal with. First, since because of text independence voice patterns do not have any text references for comparisons, this makes each voice pattern quite arbitrary which makes their comparisons difficult. Even the same speaker speaking different words can sound different and present radically changed voice waveforms. Secondly, because of the open-set database, it is not known whether the test voice pattern is actually present or not and can result in false matches. This requires determining a threshold for decision making, which is not fixed and will change depending on the database. A OSTI-SI based system will therefore need to be robust enough for comparing arbitrary voice patterns with each other and also incorporate adaptive thresholds that change depending on the database. The paper is organized as follows: section 2 provides a literature survey of existing approaches, section 3 outlines the proposed methodology, section 4 provides experimentations and results for testing the proposed system, section 5 analyses the system vis-à-vis other contemporary approaches, section 6 brings up the final conclusions with future scopes.

II. Literature Survey

The past two decades have witnessed researches focusing on speaker recognition and its development. However, at present speech recognition technology is more implemented than speaker recognition technology [11]. Speaker identification is accomplished by proper feature extraction, choosing suitable feature for the purpose followed by feature matching [3]. Feature extraction deals with converting the speech data into acoustic vectors to facilitate feature matching. MFCCs are widely utilized as features in speech processing tasks like language identification, speaker identification, emotion recognition for it serves the purpose quite efficiently. [2] Feature matching employs quantization techniques like Vector Quantization (VQ) and/or speaker modeling techniques like GMM to achieve desired classification of test data [11]. In recent years, studies indicate that the energy distribution of human voice signals follows a Gaussian Model which is why GMM is more dominant in the field of speaker identification. For the past twenty years, GMM-UBM is identified as one of the major approaches, in field of speaker identification.

978-1-5386-1931-5/17/$31.00 ©2017 IEEE

[16,17]. The feature matching procedure is accompanied with the calculation of distance between the test speaker's acoustic model with all registered speakers' acoustic models. The test speech is then mapped with that registered speaker with which it processes minimum distance [2, 7]. Recent researches focuses on creating a thresholding method effective for OSTI-SI. The proposed method was quite efficient to eliminate all untrained or unknown data but not very accurate to authenticate known data [5]. OSTI-SI has also been tried upon cohort model and UBM based approach. It has been observed that when UBM based approach when unified with cohort model in a projective framework, improves accuracy [12]. There are other alternative approaches for open set speaker identification such as GMM-UBM supported by score normalization and i-vector method. A comparative study on the effectiveness of two methods had been made [1]. The test conducted on a subset of NIST-SRE 2008 database containing 400 registered speakers and 200 out of set speakers. It had been observed that when the test was conducted on clean data 39.5% accuracy was obtained for GMM-UBM while 42.5% for GMM-UBM with TZ(test normalization, zero normalization) norm and 49.5% with i-vector method. The i-vector method [15] has relation to the GMM-UBM technique as a single i-vector is said to be the consolidated representation of an adapted GMM. The other subset of speaker recognition which is speaker verification has gained prominence recently. Speaker identification suffers low accuracy as the increase in population of registered speakers increases, the variability and the chance of false acceptance of unknown speaker or false rejection of registered speaker increases [4]. It also becomes increasingly difficult to tune the threshold in such cases. Speaker verification is a simpler method involving binary classification of determining whether the speaker's claimed voice, matches with the already registered voice of that speaker in the database [15]. Actually, it is the task of deciding, if the given speech utterance is provided by the hypothesized speaker S or not [19]. The binary classifier can be formulated as follows, where T(x) is denoted as the test ratio (for speaker verification systems using GMM is the likelihood ratio) and η is the threshold value [19].

H0: x is from the hypothesized speaker.

H1: x is not from the hypothesized speaker.

Then the decision in an optimal manner is:

$$T(x) = \frac{f(H0|x)}{f(H1|x)} \geq \eta, accept, T(x) = \frac{f(H0|x)}{f(H1|x)} \leq \eta, reject$$

III. Proposed Methodology

It has been observed that OSTI-SI lacks implementation in day-to-day applications due to problems such as low accuracy and indefinite thresholding method, which is to be employed irrespective of dataset to obtain a reliable result. Also, it is quite difficult to define a threshold which will result in equal error rate. The other problem lies with the choice of decision to be used. Due to intra-person variability of speech more problems arises [4]. The more populated the registered dataset is, the more prone to error it is. This study focuses on building an improved method of OSTI-SI. The method consists of two stages. In the first stage, the MFCC features are extracted from the speech samples and then they are vector-quantized based on LBG (Linde-Buzo-Gray) algorithm. Then the Euclidean distance is measured for all test data against all trained samples. The train sample with which there is minimum Euclidean distance is considered as the best match for that particular test data. The Euclidean distances are normalized to obtain the data in range of 0-1. In the second stage, the normalized minimum Euclidean distance or the distortion of each test data with respect to its best matched train data is considered as the decision criteria. Then a threshold is determined based on observations which balances the FAR and FRR. The threshold tends to decrease with increase in population. The same method with the same threshold with only slight changes, for respective length of population has been compared against other datasets. The accuracy ranges between 73 – 86 % for all datasets. The block diagram of the process is given below in Fig.1.

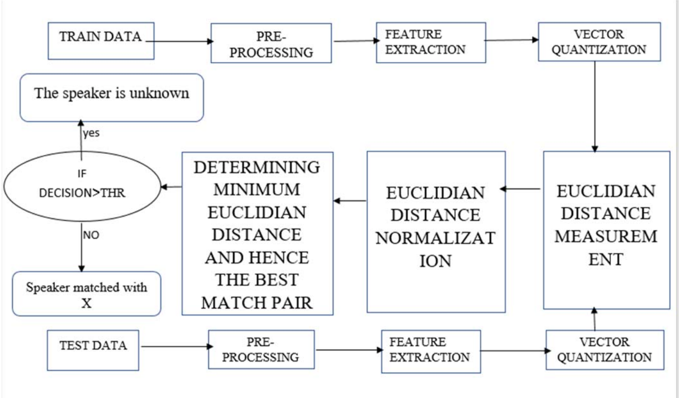

Fig.1 . Block diagram

A. Pre-Processing.

All voice samples are sampled at 16000Hz and the voices from different sessions of the same speaker have been merged to produce samples of length 60-150 sec. It is quite certain that lengthier the train data is, better is the accuracy. Also, clean noise free data are used for this paper.

B. Feature Extraction

For the purpose of speaker recognition, the most dominating features are mel frequency cepstral co-efficient [2]. MFCC involves sensitivities of human perception with respect to frequencies under consideration. The process of extracting MFCC features in given below in Fig.2.

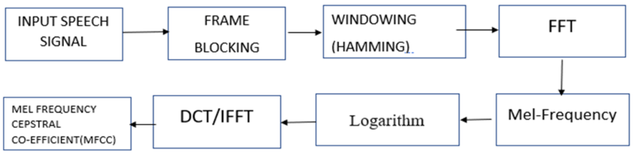

Fig.2 . MFCC

C. Framing

Entire speech sample is segmented into small frames before further operations. The length generally varies between 20 to 30 ms [7]. Adjacent frames are separated by M $(M < N)$ in a voice signal which is segmented into frames of N samples. Regular values for $N$and $M$are $N = 256, M = 100$.

D. Windowing

After frame blocking, each frame is windowed with a Hamming window in order to taper the first and last points of the frames to reduce signal discontinuities. In a typical case, the signal in a frame is denoted by (n), where $n = \{0, \dots, N-1\}$, and the signal after windowing is given by $s(n) * t(n)$, where $t(n)$ is the representation of Hamming window defined by (1) [3]

$$t(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \; ; 0 \le n \le N-1 \qquad (1)$$

E. Discrete Fourier Transform

The Fast Fourier Transform (DFT) transforms each frame of Nsamples from the time domain to the frequency domain. The DFT operation is defined by the following:

$$X_k = \sum_{i=0}^{N-1} x_i . e^{-j.\frac{2\pi k i}{N}} \qquad (2)$$

F. Mel Frequency Wrapping

Human perception of the contents of frequency of sound for speech signals does not follow a linear scale. This fact has been proven from psychological studies. For each tone with an actual frequency f measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. This *mel-frequency* scale follows a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz [11] as given by the following formula:

$$M = 2595.\log_{10}\left(1 + \frac{f}{700}\right) \qquad (3)$$

The number of *mel* spectrum coefficients, *K*, is generally chosen to be 20. This filter bank is applied in the frequency domain, which is equivalent to applying the triangle-shape windows to the spectrum [3]. To smooth the magnitude spectrum and to minimize the size of the features are two major reasons of using triangular bandpass filters [3].

G. Discrete Cosine Transform

It is now required to transform the log Mel spectrum into time domain for which inverse Discrete Cosine Transform (DCT) is used [7]. The result of the transformation gives us Mel Frequency Cepstral Coefficient (MFCC) which represents acoustic vectors. So each input utterance is transformed into a sequence of the acoustic vector. Forward DCT is defined by the following where $\propto$ is a constant dependent on N

$$X_k = \propto . \sum_{i=0}^{N-1} x_i . \cos\left\{\frac{(2i+1)\pi k}{2N}\right\} \qquad (4)$$

H. Vector Quantization

Vector Quantization (VQ) is a data reduction method. It is useful for reduction of dimension so as to reduce redundancy of data. It may be regarded as a process by which vectors from a large vector space is consolidated into the limited number of regions present in that space [3]. The centre of each region so obtained (known as clusters) is called a codeword. A codebook is the collection of all codewords of the vectors. VQ technique is implemented through LBG algorithm. Feature vectors of both train and test data extracted from MFCC is applied to VQ [5].

I. Classification

Distortion measure is the parameter which represents the similarity between the input feature vectors and a codebook. The smaller the distortion, the higher the similarity. The distortion is shown below [5].

$$D_Q(I, V) = \sum_{t=1}^{T} \min d(i_t, v_k) \;\; ; \; 1 \le k \le K \qquad (5)$$

Here D_Q is the distortion of the input feature vectors I, which consist of T vectors, to the codebook V, which consist of k centroids, i_t is the t^th^ input vector and $d$is the Euclidean distance. For two n-dimensional vectors $P = \{p_1, p_2, \dots, p_n\}$ and $Q = \{q_1, q_2, \dots, q_n\}$ Euclidean distance is defined as:

$$d(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad (6)$$

The sum of the Euclidean distances between each feature vector to the nearest centroid in a codebook is the distortion lying between a codebook and a group of feature vectors [5]. The Euclidean distance is measured for each test sample against all train samples to obtain distortion. The Euclidean distances obtained for a certain test sample when compared against all train samples is normalized. Normalization sets all Euclidean distances within the range of 0-1. This simplifies the thresholding process. Let $a = \{a_1, a_2, \dots, a_n\}$ denote the Euclidean distances of a certain test sample against n train data. Then Normalized Euclidean Distance is given by:

$$D = \frac{a_1 + a_2 + \cdots + a_n}{\sqrt{\sum_{i=1}^{n}(a_i)^2}} \; ; 1 \le i \le n \qquad (7)$$

Minimum Euclidean distance is the distortion which defines the similarity measure. The train data against which the test sample processes minimum Euclidean distance in a set of n train samples is the best match for that test sample. In closed set speaker identification, the test sample is matched against that train data. But in open set the scenario is different. The decision criteria is the minimum normalized Euclidian distance. If it is greater than the threshold the voice is rejected as unknown else it is matched with its best match.

$$\begin{aligned} D(X,S_i) > \sigma_i \ \rightarrow reject \\ D(X,S_i) < \sigma_i \ \rightarrow accept \end{aligned} \tag{8}$$

Here S_i is the set of train data/registered speakers stored in the database and X is the current test sample, D denotes normalized Euclidian distances, σ_i are the threshold values chosen to obtain equal error rate. The error which arises in first stage of OSTI-SI is the error of mismatch. The cause of it can be many but not decision or threshold. So, this error is overlooked in this paper. The second error is False acceptance error(FA), where an unknown voice is mistakenly matched against a train data. The third error is False Rejection error(FR), where the known voice is mistakenly rejected as unknown. These errors rise either due to decision or threshold. In order to gain an equal error rate, where the number of false acceptance rate (FAR) and false rejection rate (FRR) error is balanced the threshold needs to be correct.

TABLE I. FAR AND FRR

		Predicted	
		Registered Speaker	Unknown voice
Actual	Registered speaker	No error	FR
	Unknown voice	FA	No error

FAR is directly proportional to the decision threshold σ_i , while FRR is inversely proportional to the same. FAR and FRR when plotted against the decision threshold the point of intersection of these two curves is defined as the Equal Error Rate (EER). The FAR and FRR are equal at EER [6]. It has been observed that the threshold values adopted by trial and error method to obtain EER is more or less same for all the datasets. The threshold is studied with gradual increase in length of train and test data and it decreases with increase in population which is quite obvious. The determination of threshold is simplified by normalization of Euclidian distances which otherwise would had been too difficult to study.

This approach is independent of external variations, which may affect accuracy negatively at times. In GMM-UBM,UBM is a model trained from voices of non-target speakers using expectation maximization (EM) may hamper system accuracy. But that is the inevitable part of the approach. Proper choice of UBM set is necessary to get optimum accuracy. The choice of UBM set is decided by trial and error method which is again time-consuming. Cohort model fails for speakers sounding similar to each other. This method, with proper threshold is free from such errors. The results obtained from Uyghur dataset, where the similarity of voices of different persons is very high, demonstrates that. It also balances FAR and FRR unlike methods that gives unbalanced FAR and FRR, the latter being very high, compared to former. It is required to make further studies on the thresholding method for this method, so that an adaptive threshold can be employed to overcome difficulties in thresholding.

IV. EXPERIMENTATIONS AND RESULTS

Two datasets used to test the performance and efficacy of the proposed approach. The first is the Uyghur dataset. It contains the voice of 200 speakers equally divided into male and female class. It is a subset of THYUG-20 SRE [9]. It contains 3003 recorded voices of 200 people. The environment is noise free. For each speaker, his/her train data is formed by merging his/her voice from multiple trials producing a speech signal with the length varying between 60-150 sec. The test data is kept constant to 10 sec. The male and female datasets are further subdivided into two to conduct the experiment. The number of train data at each trial is kept exactly half of test data to test open set accuracy. The number of train data and test data is gradually increased as observed in Figures 3,4,5 to note the gradual change in threshold. The second dataset is the Librispeech dataset formed from Librispeech recordings [10]. All samples are clean data i.e. without noise. The train dataset contains 25 speech signals while the number of test data is 50. The train data are of 60-150 sec in length while test data is of 8-10 sec in length. After normalization, the distortions lie between 0-1. For simplicity threshold has been multiplied with 100 so as to ease the process of observation. In case of GMM-UBM, the UBM model used here is another set of Librispeech recordings totally different from the previous one. It also consists of a few custom recordings recorded in noise free environment. A total of 50 speech signals are used for training the UBM model and each has a duration of 30-60 sec.

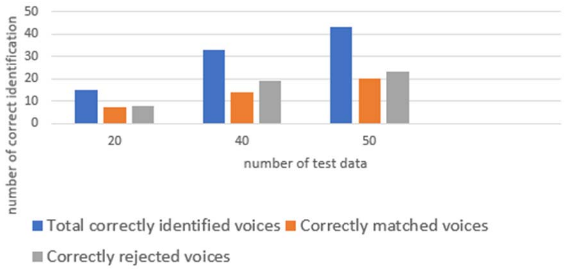

Fig. 3. Open set identification of Librispeech dataset

Fig 3 plots number of correctly identified voices of Librispeech recordings with 20, 40 and 50 test data respectively. Fig 4 and 5 plots open set identification of Uyghur male and female dataset with 40, 60, 80 and 100 test data respectively. Each voice is sampled to 16000Hz and all voices used here are clean data with no environmental noise. Figures below contain the result obtained by applying proposed method on the above-mentioned datasets where the number of voices at each trial increases gradually.

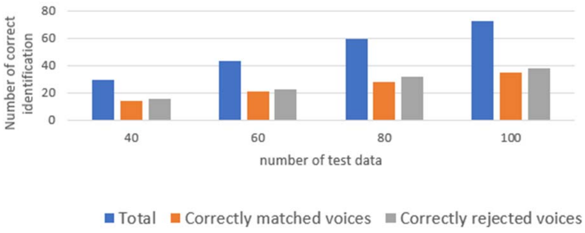

Fig. 4. Open set identification of Uyghur male dataset

In each case number of train data is exactly half of test data. The threshold value decreases with increase in number of train data as the denominator increases (equation 7) the normalized distortion falls and this explains the change in threshold value. Value of thresholds for different datasets do not differ much.

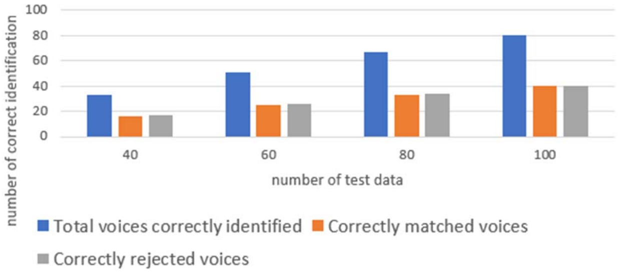

Fig. 5. Open set identification of Uyghur female dataset

It can be observed from the figures 3,4,5 that the identification rate of the system is in between (73-86)%. A balance is achieved between FAR and FRR. It can be seen from Fig 6 that the thresholds used in all three different datasets are more or less equal. It gradually decreases with increase in number of train data and change of threshold follows a similar pattern.

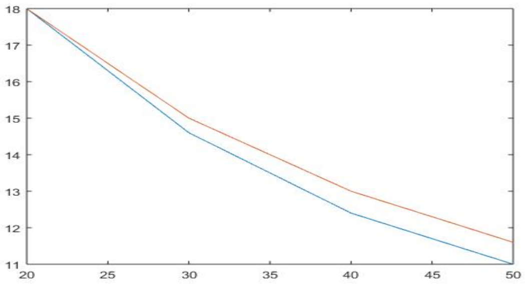

Fig. 6. Variation of threshold values with number of train data for Uyghur female and male dataset

For Uyghur, male dataset the closed set speaker identification is 93/100 test voices. The number of correct identification of unknown voices is 38/50 (76%), while the correct identification of known voices and matching it to its best match is 35/50 (70%). The overall accuracy is 73%. The increase in number of voices increases the variability and hence accuracy falls. For Uyghur, female dataset the closed set speaker identification is 91/100. The number of correct identification of unknown voices is 40/50 (80%) while the correct identification of known voices is 40/50 (80%). The overall open set identification is 80%. The threshold used is 11.6 while for male dataset it is 11.0. The threshold is not subject to huge changes in order to get enough accuracy. For Librispeech recordings the closed set speaker identification is 44 out of 50 voices. The overall correct identification is 43 (86%) while the correct identification of known voices is 20/25 (80%). The number of unknown voices correctly rejected is 23/25 (92%).It is more due to less number of data. In all these cases half of test samples are unknown and half of them are known. Hence the train set contains half the number of voice samples present in test data set in all cases.

V. Analysis

In closed set identification, it is assumed that the test voice will belong to one of the registered voices and it is matched to its best match. There is no provision of decision or threshold to reject an imposter voice. In OSTI-SI, determination of correct decision and a good threshold is prioritized. Hence closed set identification technique needs further improvement for open set identification. Recent researches focusing on creating a thresholding method was efficient to nullify FA but suffered from a considerable amount of FR [5]. A threshold for each codebook was trained after the generation of the codebook of certain person or speaker. The threshold which consisted of minimum and maximum distortion value, was obtained by authenticating several voices of the same speaker [5]. In the verification phase, the resulting lowest distortion is validated [2, 7]. But it is desired to obtain a balance between FA and FR, so the need for improved method arises. Although cohort model performs well in situation where unknown voices belongs to casual imposters but cohort based speaker model becomes more vulnerable to attack by speakers sounding more similar to registered speakers [13]. The speaker verification is efficient for 1:1 match. This idea was extended into GMM-UBM where GMM is trained for each train samples in the training set. A speaker-independent model, or UBM, is trained from the out-of-set speakers using the EM algorithm [12]. Every speaker has a model which is represented with certain parameters using a GMM.A score that decides whether a given utterance $Q = \{q_1; q_2; \dots, q_n\}$ originates with speaker 'a' using the mean log-likelihood 1/n log(P(Q|θa)) where N is the number of models in the mixture, θa ={μa,$\sum_{a,}$$\alpha_a$}.The decision function for a UBM is given as follows, where T_θ is the threshold [12].

$$\frac{1}{n}\log[P(Q|\theta a)] - \frac{1}{n}\log[P(Q|\theta_{UBM})] > T_\theta \qquad (9)$$

The decision adopted for this paper is far simpler considered to GMM-UBM or i-vector approach. The method used for speaker classification is based on Bayes Classifier. Bayes decision rule is the optimal decision rule [18].

$$P(w|x) = \frac{P(x|w).P(w)}{P(x)} \qquad (10)$$

In words, posterior = (likelihood × prior) / evidence. To get maximum accuracy in classification, the posterior probability should be maximum [18] and so the distortion must be minimum. Hence, we are employing this criteria as the decision criteria unlike GMM-UBM where we take the likelihood ratio as the deciding factor. In this study, comparison of proposed method against GMM-UBM has been done, where GMM-UBM classifier is given by,

$$\log[P(X|\theta)] - \log[P(X|\theta_{UBM})] > T_{\theta} \quad (11)$$

Also in case of GMM-UBM an additional set of Universal background model is needed, consisting of voices of unknown speakers to be compared against the registered data set. The change in UBM dataset also changes the accuracy for the test set. Hence for a particular dataset a particular UBM is needed for optimum accuracy. Fig.7 indicates comparison between GMM-UBM and the proposed method. The threshold value is inversely proportional to number of train data in the database.

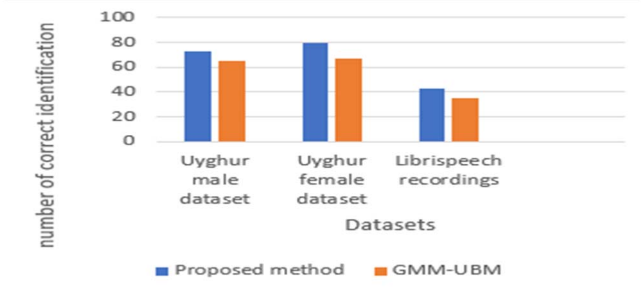

Fig. 7. Comparisons

Hence this study provides an improved approach for which (i) The accuracy of proposed method is more than GMM-UBM method which is one of the most dominating method [1] in the field of OSTI-SI. (ii) Unlike GMM-UBM it is independent of UBM model. For GMM-UBM,we had to choose an appropiate UBM for better accuracy (equation 11). (iii) The method provides a simpler approach to speaker identification, where the threshold can be determined from the rate of its decrease with increase in train data and vice versa. Due to normalization, the threshold value lies within 0-1 and it’s value for different datasets are quite similar to each other.(iv)It works well for similar sounding,but different voices arising from different sources unlike cohort models.

VI. Conclusions and Future Scopes

OSTI-SI is the most challenging subset of Speaker identification. The paper proposes a method better and faster than the predominant GMM-UBM technique. It is not dependent on external factors like UBM as in case of GMM-UBM whose accuracy differs with different UBM .It produces good result with similar sounding but different voices, unlike cohort model. The thresholding process is also simpler due to normalization of the distortion, which is treated as decision. The threshold variation is inversely proportional to the size of train data-set.The FAR and FRR is also balanced. It is needed to create a adaptive threshold. Also, it is needed to apply revised method in order to increase the accuracy of the process.

References

[1] R. Karadaghi, H. Hertlein and A. Ariyaeeinia, "Effectiveness in open-set speaker identification," 2014 International Carnahan Conference on Security Technology (ICCST), Rome, 2014, pp. 1-6

[2] A. K. Singh, R. Singh and A. Dwivedi, "Mel frequency cepstral coefficients based text independent Automatic Speaker Recognition using matlab," International Conference on Reliability Optimization and Information Technology (ICROIT), Faridabad, 2014, pp. 524-527.

[3] N. M. AboElenein, K. M. Amin, M. Ibrahim and M. M. Hadhoud, "Improved text-independent speaker identification system for real time applications," Fourth Int. Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC), Cairo, 2016, pp. 58-62

[4] A. M. Ariyaeeinia, J. Fortuna, P. Sivakumaran and A. Malegaonkar, "Verification effectiveness in open-set speaker identification," in IEE Proceedings - Vision, Image and Signal Processing, vol. 153, no. 5, pp. 618-624, Oct. 2006.

[5] R. A. Sadewa, T. A. B. Wirayuda and S. Sa'adah, "Speaker recognition implementation for authentication using filtered MFCC — VQ and a thresholding method," 3rd International Conference on Information and Communication Technology (ICoICT), Nusa Dua, 2015, pp. 261-265.

[6] H. B. Kekre and V. Kulkarni, "Closed set and open set Speaker Identification using amplitude distribution of different Transforms," 2013 International Conference on Advances in Technology and Engineering (ICATE), Mumbai, 2013, pp. 1-8.

[7] F. K. Soong, A. E. Rosenberg, B. H. Juang and L. R. Rabiner, "Report: A vector quantization approach to speaker recognition," in AT&T Technical Journal, vol. 66, no. 2, pp. 14-26, March-April 1987

[8] S. Dey and K. Kashyap, "A dynamic-threshold approach to text-dependent speaker recognition using principles of immune system," IEEE India Conf (INDICON), New Delhi, 2015, pp. 1-6.

[9] A. Rozi, Dong Wang, Zhiyong Zhang and T. F. Zheng, "An open/free database and Benchmark for Uyghur speaker recognition," Int Conf. Oriental COCOSDA held jointly with conf. on Asian Spoken Language Research and Evaluation, Shanghai, 2015, pp. 81-85.

[10] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpusbased on public domainaudiobooks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, 2015, pp. 5206-5210.

[11] F. Y. Leu and G. L. Lin, "An MFCC-Based Speaker Identification System," IEEE 31st International Conference on Advanced Information Networking and Applications (AINA), Taipei, 2017, pp. 1055-1062.

[12] A. Brew and P. Cunningham, "Combining Cohort and UBM Models in Open Set Speaker Identification," Seventh International Workshop on Content-Based Multimedia Indexing, Chania, 2009, pp. 62-67.

[13] V. Prakash and J. H. L. Hansen, "In-Set/Out-of-Set Speaker Recognition Under Sparse Enrollment," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2044-2052, Sept. 2007.

[14] I. Magrin-Chagnolleau, F. Bimbot, and R. IRISA. Indexing telephone conversations by speakers using time frequency principal component analysis.. IEEE Int. Conf. on Multimedia and Expo, 2000.

[15] N. Dehak, P. Kenny, R. Dehak et al., “Front-End Factor Analysis for Speaker Verification,” IEEE Transactions on Audio, Speech, and Language Processing,,vol. 19, no. 4, pp. 788-798, 2011.

[16] D. A. Reynolds, R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Transactions on,Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, 1995.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” Digital Signal Processing, vol. 10, no. 1–3, pp. 19-41, 2000.

[18] R. Duda, P. Hart,D. Stork. Pattern Classification,2nd edition, Wiley, New York

[19] F. Răstoceanu and M. Lazăr, "Score fusion methods for text-independent speaker verification applications," 6th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Brasov, 2011, pp. 1-6.