Multiple Feature based Speaker Authentication System

Sanjeevakumar. M. Hatture*
* Department of Computer Science and Engineering,
Basaveshwar Engineering College Bagalkot,
Karnataka State, India

Reshmabanu M. N**
** Department of Computer Science and Engineering,
Basaveshwar Engineering College Bagalkot,
Karnataka State, India

Abstract:- **Speaker authentication systems have been governed by voice features like spectral, prosody, phonemes etc. features with a proper transformation and extraction of the voice signal of the speakers in small range of time. Yet, there is proof that voice signal contains information related to specific speaker as multiple short and long-term characteristics. The speaker recognition system with specific feature may fail due to variations in the voice signal. To enhance the accuracy of the speaker recognition system multiple features are to be extracted and combined. In this paper a model with multiple feature based speaker authentication system is presented. An overview of the features of voice signal uttered by the user and classifier employed for speaker authentication system is described. The proposed model will provide better speaker authentication with multiple features.**

Keywords — Gammatone filter, PNCC, PLP, Prosody features, SVM classifier.

I. INTRODUCTION

Voice biometrics becomes a sophisticated security tool nowadays due to the development of new techniques and applications. This biometrics is the only biometrics which performs acoustic information. Speaker authentication system identifies who is speaking but not what he/she has spoken. It is hard to forge speaker's voice. Speaker recognition pick out the characteristics of voice such as pitch, dialect, tone and speed etc. which are unique to specific speaker and generates a voiceprint which cannot be regenerated. A voiceprint is a secure method for authenticating a person's identity which cannot be forgotten, duplicated stolen. There are two broader way to categorize biometric identifiers. One is Physiological Biometric Traits these features do not change by influence of psycho emotional state. These traits are more reliable. For example palm print recognition, fingerprints, DNA, hand geometry, iris, face recognition. Second one is Behavioural Biometric Traits these features change by influence of psycho emotional state. These traits are not reliable. For example typing rhythm, gait etc. Voice is neither physiological nor behavioural biometric trait. Speaker authentication is performed in two ways either text-dependent or text-independent. In text-dependent system speaker has pre-determined word to be uttered but in text- independent there is no such restrictions. There is one more category know as text-prompted where the system will generate the phrase to be uttered.

Following are the different ways of text dependent system – i. There is fixed word / sentence (phrase) which is uttered by speaker. ii. Speaker to choose phrase of his own choice which is uttered every time for authentication. iii. There is fixed vocabulary set for utterance. Following is the way of using text-independent system, Speaker is free to utter any phrase.

To enhance the accuracy of the speaker recognition system multiple features are to be extracted and combined. In this paper a model with multiple feature based speaker authentication system is discussed. Multiple features means; instead of extracting only single feature from voice signal two or more than two features will be extracted. Features used here are PNCC (Power-Normalized Cepstral Coefficients), PLP (Perceptual Linear Prediction) and prosodic features like pitch and loudness. In the proposed model for speaker recognition the input voice signal is pre-processed where noise and unwanted information is removed. The feature of voice viz. PNCC, PLP and prosodic features like pitch and loudness are to be extracted to build the individual speaker model. Further, speaker recognition/classification is performed using Multiclass Support Vector Machine (SVM) classifier and decision logic needs to be implemented. Later the performance of individual feature and their combination are examined by conducting the experiments separately. The improvement in the efficiency of recognition, reduction in the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are expected.

In section II, literature survey on different speaker authentication technologies are discussed. Section III gives detailed description of the model proposed in this work. Section IV gives brief explanation on different algorithms used. Section V gives the conclusion of this paper.

II. LITERATURE SURVEY

Some of the technologies used for speaker recognition systems are summarized in this section.

The MFCC (Mel Frequency Cepstral Coefficients) features, K-means and GMM (Gaussian mixture model) classifiers are employed. The training and testing time is considered as a parameter for the evaluation of performance of both the classifiers. On comparison GMM has training and testing time 454.5678 secs. 25.1271 secs. respectively. Similarly for K-Means 435.6854 secs. 23.7178 secs. respectively. It is asserted that K -means is a better classifier in terms of the training, testing and recognition time which tends to be minimum compared to the GMM in [1].

In [2] three techniques are used namely, Nearest-neighbour Discriminant Analysis (NDA) approach that is formulated to alleviate the limitations associated with the

conventional Linear Discriminant Analysis (LDA) that assumes Gaussian class-conditional distributions. Second is the application of speaker and channel-adapted features which are derived from an automatic speech recognition system, and last one is the use of a Deep Neural Network (DNN) acoustic model with a large number of output units to compute the frame-level soft alignments. NIST 2010 database is used. The experiment results show that NDA provides better speaker recognition performance across all three metrics. For the GMM based system, a relative improvement of 35% in Equal Error Rate is achieved with NDA over LDA, while for the DNN based systems with MFCCs and fMLLR (Featurespace Maximum Likelihood Linear Regression) features relative improvements of 25% and 18% are achieved, respectively.

GMM-SVM dual modeling framework for speaker verification applied in [3]. Here two effective approaches are used to combine GMM and SVM classifiers, the parallel-GMMSVM and the serial-GMMSVM. Both of the methods enhanced the conventional verification scheme and effectively improve recognition performance. Parallel-GMMSVM: the parallel-GMMSVM is a parallel-style combination of GMM and SVM classifiers. The utterance from a test speaker is fed into both GMM and SVM classifiers, and a voting scheme is designed to consider all classification results of these two classifiers. Serial-GMMSVM: It is a serial-style combination of GMM and SVM classifiers. Within the serial-GMM-SVM recognition framework, the recognition operation is divided into two stages: GMM speaker identification; and SVM speaker verification. The results of the experiment are like this, serial-GMMSVM approach demonstrated the highest recognition rate of 77.41 %, followed by parallel-GMMSVM at 76.27 %.

Attacks against Machine based and Human-based Speaker Verification systems are studied and an effective speaker verification system is built. In both systems, audio clips are categorized into three groups, namely, original speaker, attack of different speaker and attack of conversion. Two speech corpuses are used namely MOBIO and Voxforge datasets. This attack system comprises three phases. The first phase contains the set of voice samples OS (victim). The second phase focuses on the creation of the victim's voice based on the set of voice samples in the first phase. The last phase contains the use of voice reproduction capability to compromise any application or context that utilizes the victim's voice. Speaker verification algorithms in [4] showed 80–90% success rate.

A verification of the mimicry voice signal with the two-stage testing is presented in [5]. The first stage uses GMM for speaker identification. The second stage of testing uses MFCC and GMM techniques. Own database is used. The first stage of comparison to the voice recognition system is automatic task whereas the second stage does it manually. If the level of acceptance is set to 1.5, the person-A mimicry voice matches with person- A in presented voice model. If the acceptance level is set to 1.0 then the person-A mimicry voice matches with person-B as well as person-A. These observations are made through experiments.

Speaker recognition system which uses semi-fragile technique and blind digital speech watermarking technique is discussed in [6]. Other techniques which are used in this work are discrete wavelet packet transform (DWPT) and quantization index modulation (QIM). They make possible to insert watermark within wavelet's sub-bands. MOBIO, MIT, and TIMIT speech corpuses are used. For verification two speaker verification systems are used: the i-vector and GMM-UBM Systems with LPRC (Linear Predictive-Residual Cepsrum Coefficients) and MFCC features. This technique easily identifies various attacks, like additive noise, compression attacks, cut sampling and filtering. There is negligible effect on the performance of speaker recognition system due to this technique. Tampering threshold is set to identify the alteration in watermarked speech signal which may be occurred with intent or by accident. Experiment shows, TIMIT has the best recognition rates. It is a clean speech database so the results produced are good. Other voice corpuses incorporated with disturbances in environment, microphone and mismatches in the channel which results in the poor recognition rates. The total degradation effect of this technique is calculated as 0.97 %.

MLPNN (Multi-Layer Perceptron Neural Network) has two sub-neural networks. Voice corpuses like YOHO, BANCA and XM2VTS are used. Wavelet Transform is used to extracted the features. Various parameters are used to measure the performance of classifiers like Average Recall, Average Precision, Classification Accuracy and Root Mean Squared Error (RMSE). These parameters are calculated for MLP and other classifiers then these measurements are compared with MLP classifier. The classification accuracy of MLPNN improved by 4.58 %, 7.09% and 2.75 % when compared with feature selection by wrapper based selection with MLP, Kernel PCA with MLP and PCA (Principle Component Analysis) with MLP respectively. RMSE of introduced Neural Network decreases by 10.91%, 6.96% and 12.49 % when compared with feature selection by PCA, Kernel PCA and wrapper based selection respectively. Introduced approach in [7] shows the improvement in

classification accuracy equal to 7.09 %, average recall equal to 7.09 % and average precision equal to 7.58 %.
A forensics speaker recognition which uses Bayesian framework and it is a data-driven evaluation methodology. It is a logical way of assessing speech. It presents the biometric speech evidence of questioned recording. Bayesian framework methods are rational means of evaluating and quantifying the value of biometric voice evidence. This framework includes calculation of likelihood ratios based on automatic pattern recognition methods. [8] Showed that this approach is the rational way of presenting voice evidence and assessing voice signal.
Spear is an open source and extensible toolbox for latest speaker recognition. Spear implements a set of tool chains related to speaker recognition. It incorporates complete working set of speaker recognition such as feature extraction, template generation, classification and estimation of voice recognition. Several latest modeling techniques are explored in [9] and those are inter-session variability, GMM, Joint Factor Analysis (JFA) and Total variability (i-vectors). MOBIO corpus is used for experimental comparison of different modeling techniques. The results shows that, by combining the three single systems substantially boosts the system performance Half Total Error Rates are 14.7% and 7.9% for Female and Male speakers.
Speaker Recognition for Mobile User Authentication uses MFCC feature and VQ classifier. It is based on the LBG (Linde,Buzo,Gray) Algorithm. Sphinx and GREYC databases are used. The system verifies whether the user is the mobile's owner. For Vector Quantization (VQ) based modeling, the recognition test is classically performed through Euclidean distance computation between the reference template and the new captured template. Results show that it is good to use 45 Mel filters for the Sphinx database and 50 when the phone's microphone is used for GREYC Database. In this second case, the obtained Equal Error Rate (EER) is equal to 4.22%. Comparison of the obtained EER is made for the considered databases with other databases, in function of different numbers of centroids. The EER value is lower with 256 centroids for the Sphinx database and 128 centroids for the GREYC database in [10].
Brief summarization of all the technologies in speaker recognition system is made here. Most of the works used the MFCC feature extraction method. It is one of the most powerful feature extraction methods. Other works used fMLLR, NDA, speaker and channel-adapted features, acoustic features, LPRC and Wavelet Transform feature extraction methods. It is important to use classifiers for pattern matching in speaker authentication system. In the mentioned literature survey four of them used GMM classifiers. It is used as single classifier as well as in combination with other classifiers like SVM and GMM-UBM. There are other classifiers also used such as K-means, DNN, DWP, QIM, MLPNN and VQ. Number of speaker corpuses is used such asNIST-SRE, NIST 2010, MOBIO, Voxforge, YOHO, BANCA, XM2VTS, Sphinx, GREYC and some used their own speaker corpus. The detailed description of the proposed model is given in the next section.

III. PROPOSED MODEL

The proposed methodology for the speaker authentication system is shown in Fig.1. In this block diagram there are two phases. One is training phase and second is testing phase. Each phase contains different functional module. Training phase has three modules namely, Pre-processing module, Feature Extraction module and Speaker Modeling module. Similarly Testing phase contains same modules as training phase in addition to that it contains feature matching module. There are totally four modules in this phase. It is text-prompted system where the phrase to be uttered by the speaker is generated by the system itself. In the initial stage the voice samples are recorded through the audio recorder device. Volunteers are asked to utter the numbers 0 to 30 (in English) and each utterance is recorded in individual file. Twenty different volunteers' voice data are recorded. For every individual an identity is given and in training phase these samples will be trained. Further template will be generated and stored in knowledgebase. Detailed explanation for each module is given below.

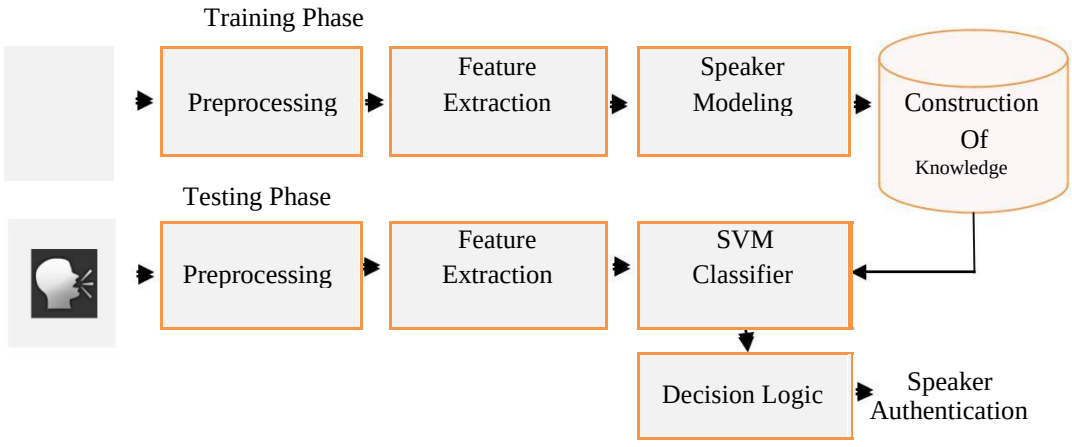

Fig. 1: Steps involved in proposed Speaker Authentication System.

Pre-processing: In this module silence, noise and unwanted signal from input voice is removed.

Feature Extraction: Efficient feature extractions methods are applied over the filtered signal thus unique features from the signal of individual speaker are extracted. In proposed methodology PNCC, PLP and Prosodic features like pitch and loudness are going to be extracted. PNCC feature is extracted and extraction of the other two features is still in progress.

Speaker Modeling: This module generates the voice temples to be stored in the knowledgebase.

Knowledgebase: It stores the voice templates generated by training phase. Own knowledgebase is used here.

SVM Classifier: In this module classifiers are applied on the voice signal. SVM (Support Vector Machine) classifier is applied in this work. The classification task computes a match score of the similarity between input features extracted from the testing voice and the template stored in the knowledgebase.

Decision logic: It will decide whether test voice signal is matched with template stored in knowledgebase. If the features are matched with the template then speaker is identified otherwise speaker is rejected. Threshold value is set to reduce the FAR and FRR.

IV. ALGORITHMS USED

A. *Silence Removal*

As in [13] the algorithm is developed for removal of silence part in voice signal. It mainly calculates the following features.

- Signal Energy: voiced segments have the larger energy than the energy stored in the silent segments. Signals with lower energy are removed.
- Spectral centroid : unvoiced segments which contain environmental sounds are eliminated using spectral centroid. Voiced segments have larger spectral centroid than the noisy sound segments as they have tendency of lower frequencies and they possess lower spectral centroid values.

The energy sequence and the respective threshold, the spectral centroid sequence and the respective threshold and the audio signal are present in the output signal. The results are plotted in Fig. 2a and 2b.

B. *Gammatone filter*

Gammatone filter (GT) is linear filter. In this filter the audio signal is fragmented into short frames. These frames facilitate to consider non-stationary voice signal into stationary voice signal. Fast Fourier Transform (FFT) is applied to emphasize the perceptual meaningful sound signal frequencies. The algorithm includes GT filter order, total filter bank bandwidth, Equivalent Rectangular Bandwidth (ERB) model and number of filters. Lastly, the log function and the discrete cosine transform are applied.

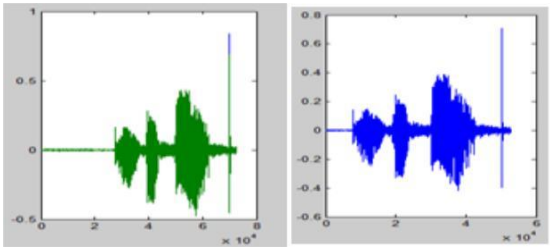

Fig.2a Input voice signal, Fig.2b Output voice signal

C. *Power-Normalized Cepstral Coefficients (PNCC)*

PNCC processing, as in [11], contains the major components listed below. PNCC features can be accommodated to any sampling frequencies.

- Initial processing
- Temporal integration for environmental analysis
- Asymmetric noise suppression
- Temporal masking
- Spectral weight smoothing
- Mean power normalization

Fig.3a shows the original voice signal. Fig.3b shows the output of Gammatone filter and Fig.3c shows the output of PNCC Features.

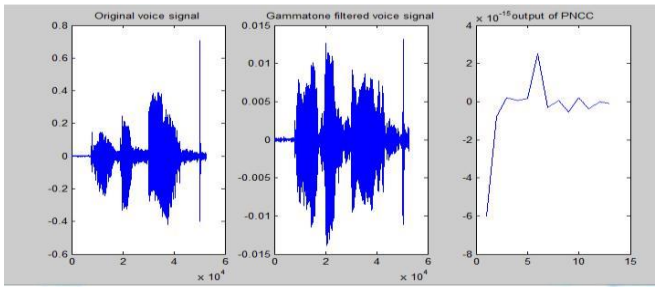

Fig.3a. Original voice Fig.3b. Gammatone filtered voice Fig.3c. Output of PNCC Feature extraction.

D. *Perceptual Linear Prediction (PLP)*

It models the voice signal according to the concept of psychophysics of hearing. It throw-outs unwanted information in the voice signal. This is the reason for improved voice recognition rate in this method. It is similar to LPC but the difference is that here spectral characteristics have been transformed to match characteristics of human auditory system. The flow of processing is shown in Fig 4. It approximates three perceptual aspects as shown in the figure below.

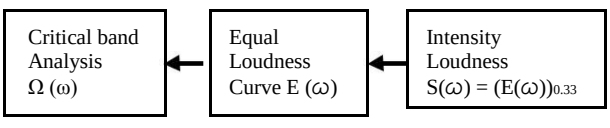

Fig 4. Block Diagram of PLP Processing.

E. Prosodic Features

As in [12], Prosodic features are suprasegmental features which will be used to distinguishing voice signal from the segment boundaries. They are of two types.

- Pitch: It is a perceived quantity which is related to the fundamental frequency of vibration of the vocal cords over some duration.
- Loudness: It is a perceived quantity which is a function of the intensity of vibration of the vocal cords over some duration and pitch.

F. Support Vector Machine (SVM)

SVM is supervised learning algorithm and a binary classification method. It identifies the optimal linear decision surface which is a weighted combination of elements of a training set. Identification of the optimal linear decision surface based on the theory of structural risk minimization is significant task carried out here. These elements in decision surface are called as support vectors, which describe the boundary between the two classes. Advantage of this classifier is that the training is fairly simple. It is scalable to high dimensional data.

V. CONCLUSION

Speaker authentication system has large scope in research area. Feature extraction and classifiers play vital role in this system. There are number of methods exists among them MFCC is powerful one. In this work effort is made to explore efficient features other than existing one. The proposed model will provide better speaker authentication with multiple features. Performance of individual features and combination of those features like PNCC, PLP, Prosodic features such as pitch and loudness will help in analysing results easily. It helps us to identify efficient feature extraction methods which can improve recognition rate and minimize the error rates.

REFERENCES

[1] O.O Adeosun and A.O Folowosele ", Performance Evaluation of Voice Classifier Algorithms for Voice Recognition Using Hidden Markov Model" ,Computer Engineering and Intelligent Systems, Vol.7, No.1, pp.57-63, 2016.

[2] Seyed Omid Sadjadi, Sriram Ganapathy, Jason W. Pelecanos, "The IBM 2016 Speaker Recognition System",Odyssey,Bilbao, Spain 2016.

[3] Ing‑Jr Ding1 · Zih‑Jheng Lin1 ," Identity authentication by sensed acoustic voices from a speaking person using an efficient GMM‑SVM dual modeling framework", Springer-Verlag Berlin Heidelberg 2016.

[4] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena," All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines" ,Springer International Publishing Switzerland ESORICS 2015, Part II, LNCS 9327, pp. 599–621, 2015.

[5] Mohammad Ali Nematollahi, Mohammad Ali Akhaee, S. A. R. Al-Haddad and Hamurabi Gamboa- Rosales,"Semi-fragile digital speech watermarking for online speaker recognition" ,sringeropen journal ,pp.1-15,2015.

[6] V. Srinivas, Dr. Ch. Santhi rani and Dr. T. Madhu , "Neural Network based Classification for Speaker Identification ",International Journal of Signal Processing, Image Processing and Pattern Recognition ,Vol.7, No.1, pp.109-120 ,2014.

[7] Anand Vardhan Bhalla,Shailesh Khaparkar (Asst. Prof.), Mudit Ratana Bhalla , "Performance Improvement of Speaker Recognition System" International Journal of Advanced Research in Computer Science and Software Engineering , Vol.2, No. 3, 2012.

[8] Andrzej Drygajlo ,"From Speaker Recognition to Forensic Speaker Recognition", Springer International Publishing Switzerland ,LNCS 8897, pp. 93–104, 2014.

[9] Elie Khoury, Laurent El Shafey, S´ebastien Marcel, "Spear: An Open Source Toolbox For Speaker Recognition Based On Bob" ,IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP) ,pp.1674-1678,2014.

[10] Kevin Brunet, Karim Taam, Estelle Cherrier, Ndiaga Faye, Christophe Rosenberger, "Speaker Recognition for Mobile User Authentication: An Android Solution" , HAL , HAL Id: hal-00848318,2013.

[11] Chanwoo Kim, Richard M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition ",IEEE/ACM Transactions on Audio, Speech, and Language Processing, VOL. 24, Issue: 7, pp. 1315 – 1329, 2016 .

[12] Homayoon Beigi, "Fundamentals of Speaker Recognition", springer,2011.

[13] Theodoros, Giannakopoulos ,"A method for silence removal and segmentation of speech signals, implemented in Matlab ",2008.

Sanjeevakumar. M. Hatture received the Bachelor's Degree in Electronics and Communication Engineering from Karnataka University, Dharwad, Karnataka, India, and the Master Degree in Computer Science and Engineering from Visvesvaraya Technological University, Belagavi, Karnataka, India, and currently pursuing the Ph.D Degree in the Department of Computer Science and Engineering at Basaveshwar Engineering College, Bagalkot under Visvesvaraya Technological University, Belagavi, Karnataka, India. His research interests include biometrics, image processing, pattern recognition, soft-computing and network security. He is life member of professional bodies like IEI and ISTE. He has published 14 papers in International journals and 5 papers in conferences.

Reshmabanu M Nadaf received Bachelor's Degree in Computer Science and Engineering from Basaveshwar Engineering College Bagalkot under Visvesvaraya Technological University, Belagavi, Karnataka, India, and currently pursuing the Master Degree in Computer Science and Engineering from Basaveshwar Engineering College, Bagalkot, under Visvesvaraya Technological University, Belagavi, Karnataka, India