A Review On Speaker Verification: Challenges And Issues

Sujiya Sreedharan, Chandra Eswaran

ABSTRACT: Personal Voice based verification is an essential requirement for protecting and controlling various confidential resources in the present technological world. Security key codes like passwords and Personal identification number and other traditional passwords can be stolen and used without the permission of the legitimate user, which resulting in loss of integrity leading to great threat to security. Hence to overcome such security issues high-tech and consistent biometric authentication technique is required in verifying identity claim of an individual from his/her voice with enhanced security measures. Recent technologies focused towards biometric features which is the emerging development of mobile technology. This article gives an overview on speaker verification biometric technology with the issues and present scenario and various applications on voice processing technology.

Index Terms— Speaker Recognition, Speaker Verification, Application on speaker verification, Issues, Challenges

1 INTRODUCTION

SPEECH signal of a human contains linguistics and physiological information for recognizing a speaker. Speaker recognition as a security system is a broad research area in signal processing and pattern recognition. Voice recognition has become popular technology for remote authentication especially in the advancement of telecommunications and networking [1]. Speaker Verification is used in many popular on-site applications like access control, to car, home, warehouse, computer terminals. It is implemented in remote applications such as telecom network, databases, web sites, e-trade, banking transaction and other confidential transaction and speaker verification is used popularly in forensics investigation and personalization[2][3]. Speaker recognition is a popular research area in speech processing which are further classified into speaker identification and verification where independent researches are concentrated in the specified area of division. Speaker identification is a task of comparing one individual unknown voiceprint with n number of individual enrolled in the speech database whereas in Speaker Verification the claimed identity is compared with the voiceprints present in the database and verification is performed based on the threshold value of the claimed speaker[4][6]. Speaker recognition is performed in three approaches. The first approach is done using long term analysis of acoustic feature concentrating on speaker dependent components which represents the vocal track space of an individual [5]. The main highlight of this approach is it discards much speaker dependent components and acquire information that is (>20 seconds) of speech utterance[6][7]. The second approach is modeling the speaker model based on the individual phonetic sounds that is composed of utterances. A comparative analysis from phonetic sounds from acoustic features is tested for finding the similarity of phonetic sounds from an individual [8]. In the third approach, rather than training the individual model of particular individual discriminative neural networks are capable of training the models from a known environment for recognizing a speaker[9][10]. This approach produces good recognition performance when compared to traditional model of training [11]

Identification of an individual from the total number of unknown speaker is literally known as Speaker Identification represented in (Figure 1) is a process of identifying an individual from the voice prints. A task of identifying an individual by a machine. Two distinct operations namely training and testing are acquired for building a speaker model. Speaker Identification is carried out offline before system deployment. The utterance of an unknown speaker from the trained speaker is done in the phase of testing scenario to quantify the performance of the system [12]. Speaker identification and verification works both on open-set and close-set environment for recognizing the unknown and claimed speaker. A pool of individual group from a known workspace is a closed set environment for example a corporate office, whereas open set identification where the system is not aware of the speaker voice where the voice print comes from a general population example forensic investigation.

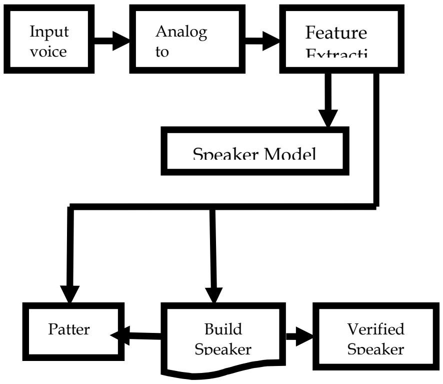


Figure 1: *Speaker Identification System*

- *Sujiya Sreedharan is currently pursuing PhD degree program in Computer Science in Bharathiar University, India,*
  E-mail: suji.sreedharan@gmail.com
- *Chandra Eswaran is currently working as a Professor in the department of Computer Science in Bharathiar University, India,.*
  E-mail: crcspeech@gmail.com

2. SPEAKER VERIFICATION

Speaker verification referred to figure 2 is a task of verifying an individual based on the features of the claimed identity. During initial stage of speaker identification the speech sample from the individual has to be provided to the system for training the model for the registered speaker. Verification or authentication is performed on testing phase to validate whether the claimed person is fake or true to deny or allow the individual to access the service. Threshold value based decision is the task of speaker verification. In verification system two key performances is done namely False Acceptance Rate (FAR) and False Rejection Rate (FRR) for testing the system robustness. However, uncertainty, phone content and channel variability are the challenging task in the area of speaker Verification for building an accurate model for recognition. [1,13]. The characteristic extracted from the speaker voice are used in both training model and to build up a reference representation of the claimant. In many application, service accessibility is active after the verification task is performed. During verification task the voice is compared with the characteristics of the speaker with previously stored voiceprint for accepting/rejecting of the claimed identity. In speaker verification good performance is achieved by more number of training and testing utterances. Basically in speaker verification system two phases of task are performed before forming up the entire verification system process [14,15, 18].

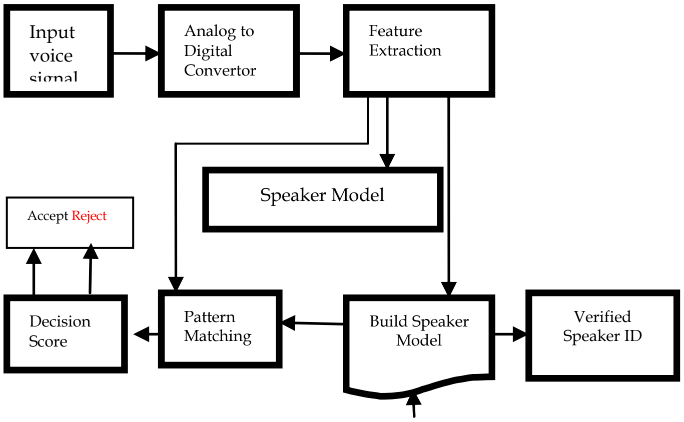


Figure 2: *Speaker Verification System*

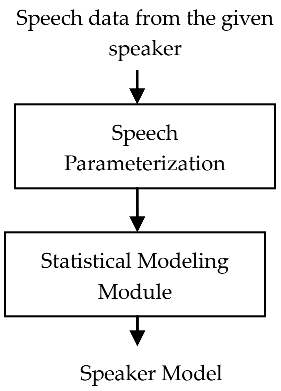


Figure 3: *Speaker Verification: Training Phase*

2.1 Training Phase

The initial step is used to extract parameters by using various feature extraction techniques for extracting speaker related attributes from the speech signal for verification process. In the second step the extracted parameters are used to obtain the statistical model of the speaker. In the final step the training is performed on the background model for recognition a speaker based on the features [16]

2.2 Testing Phase

The testing phase also called as operational session. A comparative analysis between the individual voiceprint and the reference model of each individual is performed for decision support under matched condition[16][17]

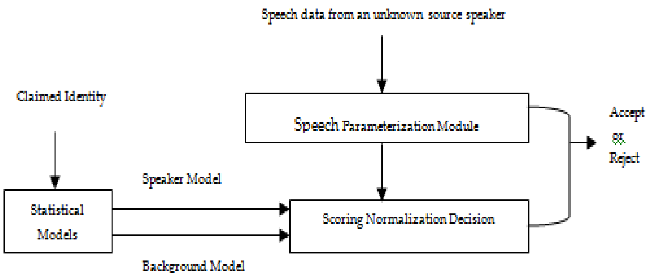

Speech data from an unknown source speaker
Claimed Identity
Speech Parameterization Module
Accept
or
Reject
Speaker Model
Statistical
Models
Scoring Normalization Decision
Background Model

Figure 4: *Speaker Verification: Testing Phase*

In testing phase an individual initiates by making a claim as to who he/she is, which leads the system for verifying the claimed identity is true or false. In speaker verification system as an analyzer compares the speech against the background model, individual voiceprints and claimed identity for decision making. The ratio analysis is then concluded and compared to a threshold value for acceptance or rejection of the requested claim [18,19]

2.3 Speech Parameterization

Representation of parameters in more suitable way for statistical modelling of the voice signal is transformed to feature vector representation in more compact, less redundant for a score distribution. Cepstral representation of the speech is the important clue for speech parameterization in any recognition and verification system. Discrimination between different speech sounds is performed based on the extracted feature information. The information is represented as sequence of parameter vectors. Robust speech parameterization can be achieved by representing speech that is invariant to the changes of acoustic environment [1, 10, 13, 20]

2.4 Statistical Modeling

Statistical modelling techniques recently have more progress in speaker recognition. The scope of pattern recognition techniques leads path for standard classification and clustering of features in speaker verification system. In recent years are Gaussian Mixture model build with background model, Vector Quantization model, Support Vector Machine, Joint Factor Analysis are used in the research area of speaker verification [21]

2.5 Normalization

Normalization is the final stage where decision of acceptance or rejection of a claimed speaker is done based on the identity. In normalization the threshold based decision method is performed by comparing the identity claimed and initiated voice prints. Acceptance or rejection decision of the claimed identity is performed based on the threshold value the higher value of the threshold value is acceptable where the low value of threshold is considered for rejection of the claimed identity, else the claimed identity is rejected [22, 27]. During the decision stage, the noise reduction method in score analysis is stabilized during decision making stage using log likelihood score. The main focus of normalization is to stabilize the mismatch condition between training and testing phase. This mismatch is reduced by adaption of score distribution to test environments. Normalization is considered to be best noise reduction method in the literature point of view which is used to minimize the mismatch condition that exists between the claimed speaker and its impostors which are applicable for both speaker identification and verification [1, 14, 15, 23]

3. Speaker verification technology

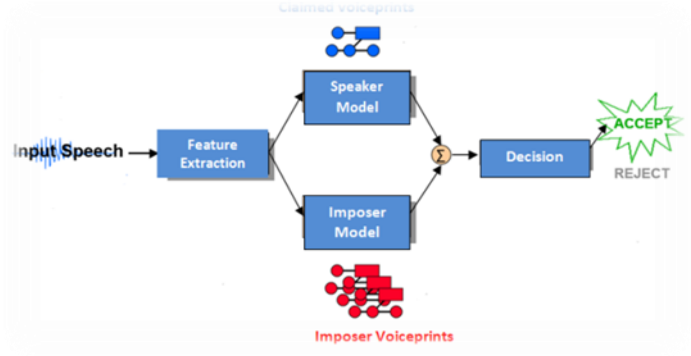


***Figure 5**: Speaker verification model*

Automatic speaker verification is categorized into various phases which are discussed in the oncoming sections

3.1 Digital speech data acquisition

Speech data acquisition plays a vital role in developing a speech recognition system. The relevant data of the speech samples are statistically analysed and the variables are treated independently by acquisition method. A high quality of acquisition device is used to capture acoustic signals in more reliable and useable manner. Current accepted acquisition methodology mainly focuses on recorded environment, regardless of hardware /software configuration. Upcoming voice technology advancement as improved in accessibility and opportunities for research in speaker recognition. Recording with Data Acquisition Toolkit (DAT), minidisk recorder (MD) or director connection between a microphone and a audio board are used for processing a speech to the computer with either an analog or digital connection[1- 24].

3.2 Feature Extraction

For speaker recognition purposes, optimal feature such as high variation in inter-speaker, low variation in intra-speaker, easy to evaluate, effective against disguise and mimicry, robust against distortion and noise, should be focused for developing a robust speaker verification system. Various features are extracted for recognizing a speaker accurately which are mentioned in the following section [25].

- Spectral features
- Source features
- Dynamic features
- Supra segmental features
- high level features

Physical characteristics of the speech source which are represented within short-term duration is called spectral. Time evolution of the spectral features denotes the dynamic features. Source features represents glottal voice feature vector. Supra-segmental features represent patterns of intonation, rhythm, stress, prosody etc that span over various segments. Finally, symbolic type of representation of speech information are denoted as high level features [26, 27].
Feature extraction is important for extraction various information. First, speech carries several features which are highly complex in nature. In speaker recognition correlation persists between behavioral and physiological characteristic features of the speaker. Various other important measures should also be considered before extracting features in undesirable noise condition whose effect must be minimized. Feature extraction techniques should extract features by concentrating on the following parameters [12, 28]

- Speaker duration variability
- Concentration on impersonation /mimicry
- Health issues which leads to variations in voice
- Speech should occur frequently and naturally
- Focus on noise cancellation and distortions

3.3 Speaker Modeling

Speaker modeling is an important task in speaker verification where separate model for each individual is created for the enrolled speaker during enrolment phase. Speaker modeling is performed of the extracted speech parameter which plays a major role in creating a speaker model. [15] Template Matching: A sequence of feature vector in template based in a fixed phrase manner. Generation of match score by Dynamic Time Warping(DTW) which is evaluated during the verification process. Evaluation analysis is done to calculate the similarity score between the test data and speaker template. Template matching techniques are used popularly for text dependent application [29] Nearest Neighbor: Dependent on all feature vectors starting from enrolment of speech regardless on any explicit model which retains to correspond to the speaker. Match score generation depends on the distance score between test features to its corresponding feature vector to its k-nearest neighbors' during verification process. Feature pruning techniques are combined with this model to overcome storage and computational complexity. [30] Neural Networks: Techniques like multi-layered perceptions or radical basis function are some of the popular techniques used in speaker verification. This model is explicit based training model which is applicable for discriminating between the modeled speakers with the alternative speakers. This model computationally expensive for training data and the limitation is where the model is not generalized [30] Hidden Markov Models: Statistical model efficiently used for encoding the temporal evolution of features which repents how a speaker sound is produced. During verification, generation of test feature sequence is computed based on the likelihood score of the test feature sequence against speaker's HMMs. Speaker verification based on text-dependent applications of the entire phonemes may be modeled using multi-state left-to right HMMs. For text-independent applications, Gaussian Mixture Models (GMMs) are applicable which rely on single state HMMs. According to the review literature Hmm based speaker verification model provides better performance when compared to other techniques in noisy conditions [23].

3.4 Imposer model

Popular model used in biometric verification system for the representation of person independent feature characteristics comparative to the model of person-specific feature characteristics which leads a pathway for making an acceptance or rejection decision of the claimed individual. In case of speaker independent verification model the speech samples are trained with large set of speakers for the representation of general speech characteristics. In Gaussian Mixture Model (GMM) generation of likelihood score on the unknown speaker is performed based on match score obtained from the speaker oriented model. Prior model based representation like Universal Background Model with MAP parametric Estimation are the state-of-art in the area of speaker verification. [11,21].

3.5 Likelihood Ratio Test

Consider an observation parameter O, and a hypothesized individual P, the verification process has to confirm that the value of O was from P. The hypothesis test is represented by using the following procedure

H0: O is from individual P

H1: O is not from individual P

By using statistical pattern recognition techniques, from the two hypothesis scenario the optimum decision iis calculated using the likelihood ratio using the following formula

$P\ (O \mid H0)\ p\ (O \mid H1) \{\geq \theta$ Accept H0 $< \theta$ Reject H0 …… (1)

where $p(O \mid Hi)$, i = 0, 1 is the probability density function for the hypothesis Hi evaluated for the measurement Y , also referred to as the "likelihood" of the hypothesis Hi given the measurement . The decision threshold for accepting or rejecting H0 is θ. The basic aim in developing a verification system is to determine techniques to compute this likelihood ratio function, usually by finding method to represent and model the two likelihoods, $p(O \mid H0)$ and $p(O \mid H1)$[13, 20].

3.6 Decision logic

In speaker verification process performance is evaluated based on False Acceptance Rate (FAR) and False Rejection Rate (FRR) with the following formulation

$$False\ Reject\ Rate = \frac{number\ of\ rejective\ true\ speaker}{total\ number\ of\ true\ speaker}$$

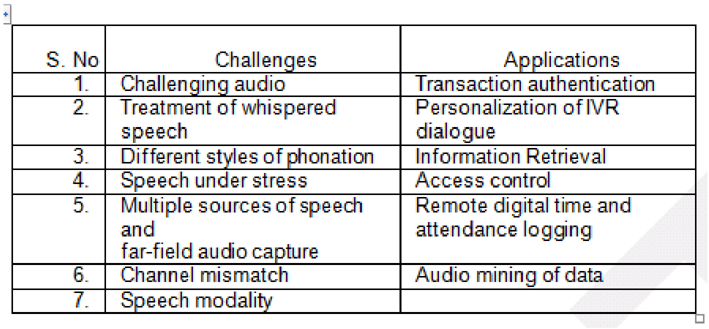
$$False\ Acceptance\ Rate = \frac{number\ of\ accepted\ imposter}{total\ number\ of\ imposter}$$

EER = $False\ Reject\ Rate = False\ Acceptance\ Rate$

False Acceptance: Also called as missing probability ratio in which the number of verified identities for which the test speaker varies from the target speaker normalized against the total number of acceptances. [10]

False Rejection rate: Also called as false alarm probability in which the number of identities which were not verified for which the test speaker was the same as the target speaker normalized against the total number of rejections. [30]

TABLE 1: CHALLENGES & APPLICATIONS OF SPEAKER VERIFCATION [1-9]



S. No	Challenges	Applications
1.	Challenging audio	Transaction authentication
2.	Treatment of whispered speech	Personalization of IVR dialogue
3.	Different styles of phonation	Information Retrieval
4.	Speech under stress	Access control
5.	Multiple sources of speech and far-field audio capture	Remote digital time and attendance logging
6.	Channel mismatch	Audio mining of data
7.	Speech modality	

5. Summary and Conclusion

From the literature point high level features are to be focused than low level feature spectrum in order to improve accuracy. Voices based biometrics is the ability to authenticate direct physical access to the real-time application. In forensics investigation application the voice based result offers as evidence in judicial trials. Speaker Verification technology has certain limitation and challenges to overcome to attain a good recognition system. Compared to other physical traits recognition voice recognition technology face issues mostly by physical and emotional stress. Much concentration has to be focused on feature extraction part which plays a crucial role in recognizing a speaker accurately.

References:

[1] J. P. Campbell et.al., "Speaker recognition: A tutorial," Proceedings of the IEEE, vol. 85, no. 9, pp. 1437–1462, (1997)

[2] Frederic Bimbot et.al., "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal Processing:4, 430–451 Hindawi Publishing Corporation, 2004.

[3] D. Petrovska et.al., "Segmental approaches for automatic speaker verification," Digital Signal Processing, vol. 10, no. 1–3, pp. 198– 212, 2000.

[4] Dr.Mahesh.et.al., "Speaker Features and Score Normalization for Multimodal Recognition Systems", Journal of Information Assurance and Security Recognition Techniques: A Review, International Journal Of Computational Engineering Research / ISSN: 2250–3005, 2014.

[5] Kinnunen et.al., "An overview of text independent speaker recognition: from features to supervectors", Speech Communication.,, 52, (1), pp. 12–40,2010.

[6] JayannaH. et.al., "Analysis of feature extraction, modeling and testing techniques for speaker recognition", IETE Tech. Rev.,26, (3), pp. 181–190, 2009 .

[7] Kinnunen T. et.al., "Real-time speaker identification and verification", IEEE Trans. Audio Speech Lang. Process. 14, (1), pp. 277–288, 2006

[8] ManamA.B. et al, " Speaker verification using acoustic factor analysis with phonetic content compensation in limited and degraded test conditions". Proc. TENCON, pp. 1402–1406.

[9] AndoA. et al., '"Speaker recognition in duration-mismatched condition using bootstrapped i-vectors". Proc. APSIPA, pp. 1– 4.,2011.

[10] MaJ.Sethu et al., ”Duration compensation of i-vectors for short duration speaker verification”, Electron. Lett., 53, (6), pp. 405– 407, 2017 .

[11] Kanagasundaram.A. et al, “Dnn based speaker recognition on short utterances”, preprint arXiv:161003190, 2017.

[12] ChenY.TangZ.et.al., “Speaker recognition of noisy short utterance based on speech frame quality discrimination and three-stage classification model”, Int. J. Control Automation., 8(3), pp. 135–146, 2015.

[13] Auckenthaler, et al., “Score normalization for text-independent speaker verification systems”. Digital Signal Processing Vol. 10, pp. 42-54, 2000.

[14] Jain et.al., “An introduction to biometric recognition,” IEEE Trans. Circuits Systems Video Technol., vol. 14, no. 1, pp. 4–20, 2004.

[15] D. Reynolds “An overview of automatic speaker recognition technology,” in Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP), vol. 4, pp. 4072–4075, 2002.

[16] T.Kinnunen et.al., “An overview of text-independent speaker recognition” Speech Communication ., vol. 52, no. 1, pp. 12–40,2011.

[17] H. Beigi, Fundamentals of Speaker Recognition. New York, NY: Springer, 2011. D. A. Reynolds et.al., “Speaker verification using adapted Gaussian mixture models,” Digital Signal Process., vol. 10, no. 1–3, pp. 19–41, 2000.

[18] X. Fan et.al., “Speaker identification within whispered speech audio streams,” IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 4, pp. 1408–1421, 2011.

[19] Amin Fazel et.al., “Statistical Pattern Recognition Techniques for Speaker Verification”, IEEE Circuits and Systems Magazine,vol 2,pp 62-81, 2011.

[20] Xing Fan et.al., Speaker Identification within Whispered Speech Audio Streams”, IEEE Transactions On Audio, Speech and Language Processing, vol. 19, no. 5,2011.

[21] Luciana Ferrer, “A comparison of approaches for modeling prosodic features in speaker recognition” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 7, pp. 2095–2103, 2010.

[22] Gregory Ditzler et.al, “Fusion Methods for Boosting Performance of Speaker Identification Systems”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 1,2010.

[23] H. Beigi, "Fundamentals of Speaker Recognition,” VDM Verlag, Saarbrücken. Farrùs, “Prosody in Automatic Speaker Recognition: Applications in Biometrics and Voice Imitation,” VDM Verlag, Saarbrücken, 2011.

[24] M. Sahidullah “Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition,” Speech Communication, Vol. 54, No. 4, pp. 543-565, 2010.

[25] Tomi H. Kinnunen, “Optimizing Spectral Feature Based Text- Independent Speaker Recognition” A Phd Thesis UNIVERSITY OF JOENSUU ,2005.

[26] R. Prabhavalkar et.al., “Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brisbane, Australia, , pp. 4704–4708, 2015.

[27] P. Kenny, “Bayesian speaker verification with heavy-tailed priors,” in Proc. Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic, 2010.

[28] N. 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.

[29] H. Aronowitz.,et.al., “Text-dependent speaker verification using a small development set,” in Proc. Odyssey Speaker and Language Recognition Workshop, Singapore, 2012.

[30] P. Kenny et.al., “Joint factor analysis versus eigen channels in speaker recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 1435–1447, 2007.