STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 5, Issue 3
www.stmjournals.com

Performance of Speaker Recognition System in Mismatching Speaking Style

Pinky J. Brahmbhatt[1,*], K.G. Maradia[2]
[1]Department of Electronics and Communication Engineering, L.D. College of Engineering, Navrangpura, Ahmedabad, Gujarat, India
[2]Department of Electronics and Communication Engineering, Government Engineering College, Gandhinagar, Gujarat, India

Abstract

Analysis of speaker recognition system under different speaking style and mismatch conditions is presented. Generally, in text independent speaker recognition system, the way of speaking style used remains the same in training and testing phase. Whisper speech is generally used quietly to convey secret information or to avoid disturbing others in quiet place, so hearing of speech is limited to nearby listener only. Source of excitation and vocal tract system contains speaker specific information. Vocal folds do not vibrate in whisper speech, so source excitation related information is not present. Fast speech is a tendency to speak rapidly, as if motivated by urgency unobvious to listener. The CHAINS speech corpus: CHAracterizing INdividual Speakers database in whisper, solo and fast speaking style is used for performing experiments with Gaussian Mixture Modeling- Universal Background Modeling (GMM-UBM) approach with the most widely used feature Mel Frequency Cepstral Coefficients (MFCC). The mismatch condition in training and testing speaking style is observed from the experiments and the performance degradation is found to be high when mismatch of train-test condition is considered.

Keywords: *Whisper speech, fast speech, CHAINS database, GMM-UBM, MFCC*

****Author for Correspondence E-mail:*** *pjbrahmbhatt@ldce.ac.in*

INTRODUCTION

Speech signal not only conveys the message but also information of speaker, language, accent, emotions, gender, dialects and paralinguistic. Large amount of data is generated during speech production process in which all of them not containing useful information for speaker recognition. So, from this data speaker specific information is extracted to identify who is speaking. Speaker recognition is used to identify the speaker or verify a person's claimed identity from their speech. *Automatic speaker verification* (ASV) and *automatic speaker identification* (ASI) are the two different task of speaker recognition. Both systems use a stored database of reference models for speakers. ASV requires evaluation of test pattern with one reference model and takes binary decision whether the test speech matches the model of the claimant. On the other hand, ASI requires choosing one of the speakers among reference models whose voice matches best with the test voice [1]. Speaker recognition includes application in forensic, banking, security, and access control. Speaker recognition system extracts features from the speech; trains speaker's model and identify the speaker. Feature extraction is the process of deriving sequence of feature vectors from the raw signal in which unique information of speaker information are accented [2]. Various features are proposed and categorized mainly in four categories: (i) Spectral features like MFCC [3], LPCC (Linear Prediction Cepstral Coefficient) [4] (ii) voice source feature like residual phase (iii) dynamic features like time derivatives of spectral features (iv) Prosodic features like fundamental frequency (F0), energy distribution, duration (v) high level features like idiolect-speaker's characteristic vocabulary is used to characterize the speaker. Among all the features MFCC is the most widely used feature for speaker recognition. Figure 1 shows the conventional automatic speaker verification/identification system with GMM-UBM approach which has two phases 1. Speaker enrollment 2. Speaker recognition.

Speaker specific model is first derived from background model which is prepared with the speaker independent data. At the time of enrollment of new speaker to the system, parameters of the background model are altered according to the new speaker's feature distribution. It has been found that adapting the means only work well rather that adapting mean, weight and variance of the background model [5]. In the recognition mode, both the hypothesized model and the background model are matched, and the match score is calculated using average log likelihood ratio of the target model and the background model.

In this paper, speaker recognition system is developed with GMM-UBM [5] approach with the state-of-the-art feature MFCC. Speaker verification and identification performance evaluation is done on CHAINS whisper, solo(normal) and fast speech database and resultant error rate degradation is observed for different train-test conditions like solo-solo, solo-whisper, whisper-solo, whisper-whisper, solo-fast, fast-solo and fast-fast. The paper is structured as follow: Section 2 describes the analysis of normal, whisper and fast speech and Section 3 discusses experimental setup. Finally, conclusions are presented in section 4.

ANALYSIS OF SPEECH

Speech is the most common and powerful way to convey the information from one person to another. Speech is generated when pressure variations from the mouth of a speaker take place. When these pressure variations propagate through air and enter the listener's ear, the ear perceives the stimuli, analyses the stimuli and transmits them to the brain. Brain decodes the received waves into the message. From the received message, listener can identify the language, gender, speaker, emotion and paralinguistic information. According to source-filter model for speech, human speech production system can be divided into source of excitation and vocal tract system or filter.

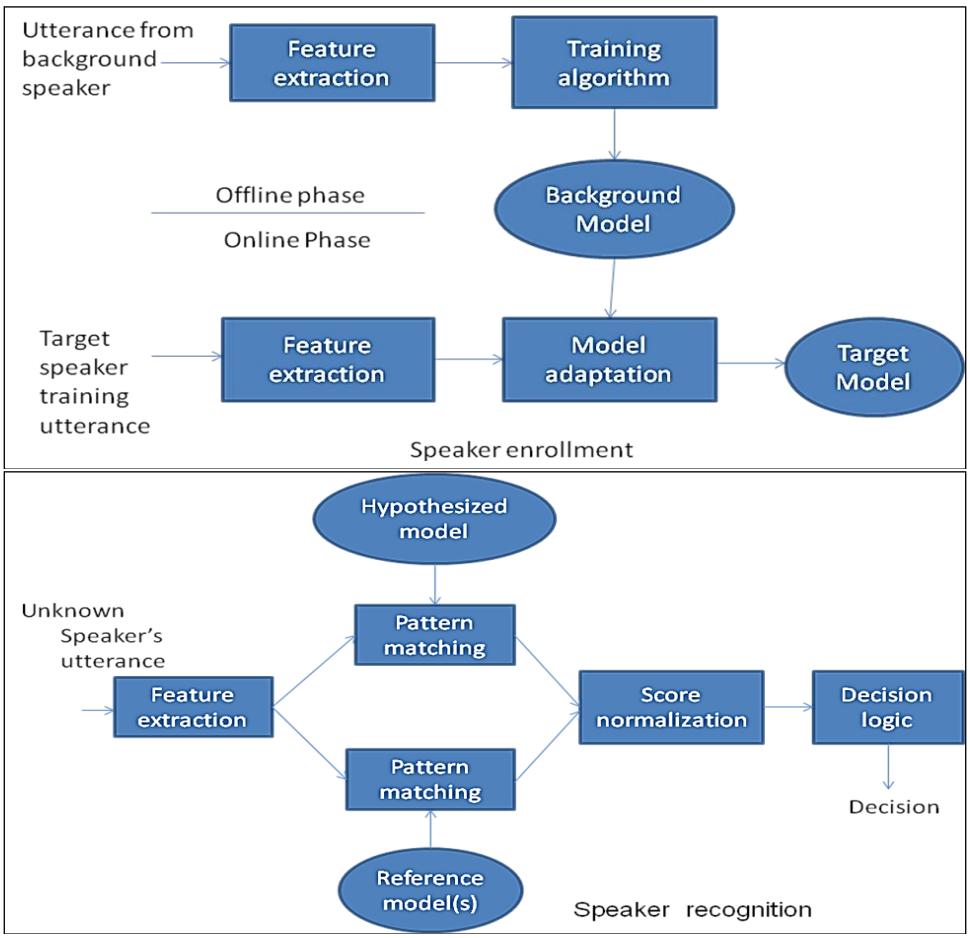

Fig. 1: *Speaker Verification/Identification System [2]*

Whisper Speech

Whispered speech is generally used in a situation when the speaker wants to convey some secrete information or if speaker wants to pass on the information to target listener only. Whisper speech is employed in public situations in order to protect personal information such as password, PIN number, bank account balance and contact number. Whisper speech differs from normal speech in many ways. In the perception of whisper speech listener has to come nearer to speaker because of its weak amplitude level. As shown in the Figure 2, pitch is almost absent in the whisper speech.

Glottis is open and there is no vocal fold vibration in whisper. Turbulent flow is produced, however, at the glottis, rather than at a vocal tract constriction [6].

From Figure 2, it is observable that whisper speech does not contain any harmonics, lower formant frequencies are shifted, low energy and change in duration compared to normal speech. Due to presence of noise and low amplitude in whispered speech, it deprives the speaker specific information present in it.

Fast Speech

Variability in speaking rate affects the performance of speaker recognition system. Many acoustic properties changes with speaking rate such as duration of the speech, F0 range, higher order harmonics. Fast speech contains shorter pauses between or within sentences. Figure 3 represents the speech waveform and spectrogram of normal speech and fast speech for the same sentence. It is depicted that for the same sentence, the duration becomes short, higher order formants are not present and most of the energy is concentrated in lower frequency. The duration of the train and test speech plays important role in speaker recognition [7]. The performance of the system will be degraded if the limited data condition is used.

EXPERIMENTAL SETUP

In the CHAINS (CHAracterizing Individual Speakers) corpus the recordings of 36 speakers was carried out in two different sessions at the time separation of about two months. The first recording session of natural speech was done in a professional recording studio using Neumann U86 condenser microphone. The second recording session of whisper and fast speech was done in a quiet office environment using an AKG C420 headset condenser microphone. Four short fables used for training were uttered in normal, whispered and at fast rate by the subject and thirty-three individual sentences of which nine selected from the CSLU speaker identification corpus and twenty-four from the TIMIT corpus were used for testing the system. In the recording of FAST speech, two recordings of same sentence were played to each subject in which one was read a comfortable rate and the second was at a substantially greater rate of articulation to illustrate the degree of rate increase they should aim for, and subjects then read all texts are this self-controlled rate.

The speaker recognition system is developed using adaptive GMM-UBM approach with CHAINS [8] normal, whisper and fast data rate. The database contains 144 train utterances and 1188 test utterances are recorded from 36 different speakers of which 16 female speakers and 20 male speakers. 12 dimensional MFCC features are extracted for train speech data and given as input to GMM-UBM system. UBM model is developed using close set speaker identification approach. Speaker specific models are adapted by adapting mean of UBM to develop adaptive GMM models for 36 speakers. Each test utterance is tested against 36 speaker model which gives total 42768 trials out of which 1188 genuine trials and 41580 imposter trials. To measure the score of each segment against all training model, log likelihood ratio is calculated in speaker identification. The model is selected as identified speaker which gives maximum score among all 36 training models. True score ratio and False score ratios are calculated to plot DET (Detection Error Trade off) curve. We have used MSR identity open source toolkit [9] to develop the system.

RESULTS AND DISCUSSION

Figures 4 and 5 show the results of the experiments. In the first experiment, the system is tested with normal and whispered speech database for different train-test condition.

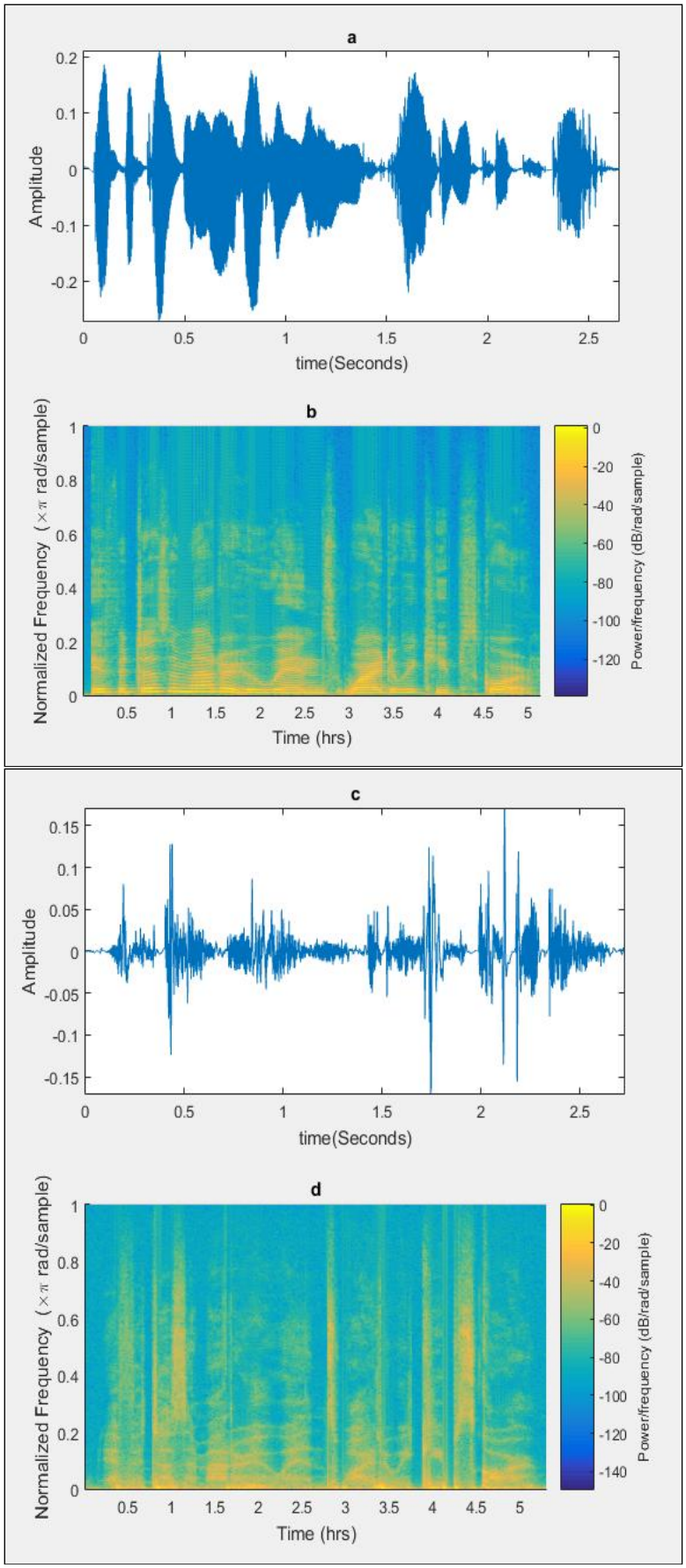

Fig. 2: *Speech Waveform and Spectrogram of (a, b) Normal Speech (c, d) Whispered Speech.*

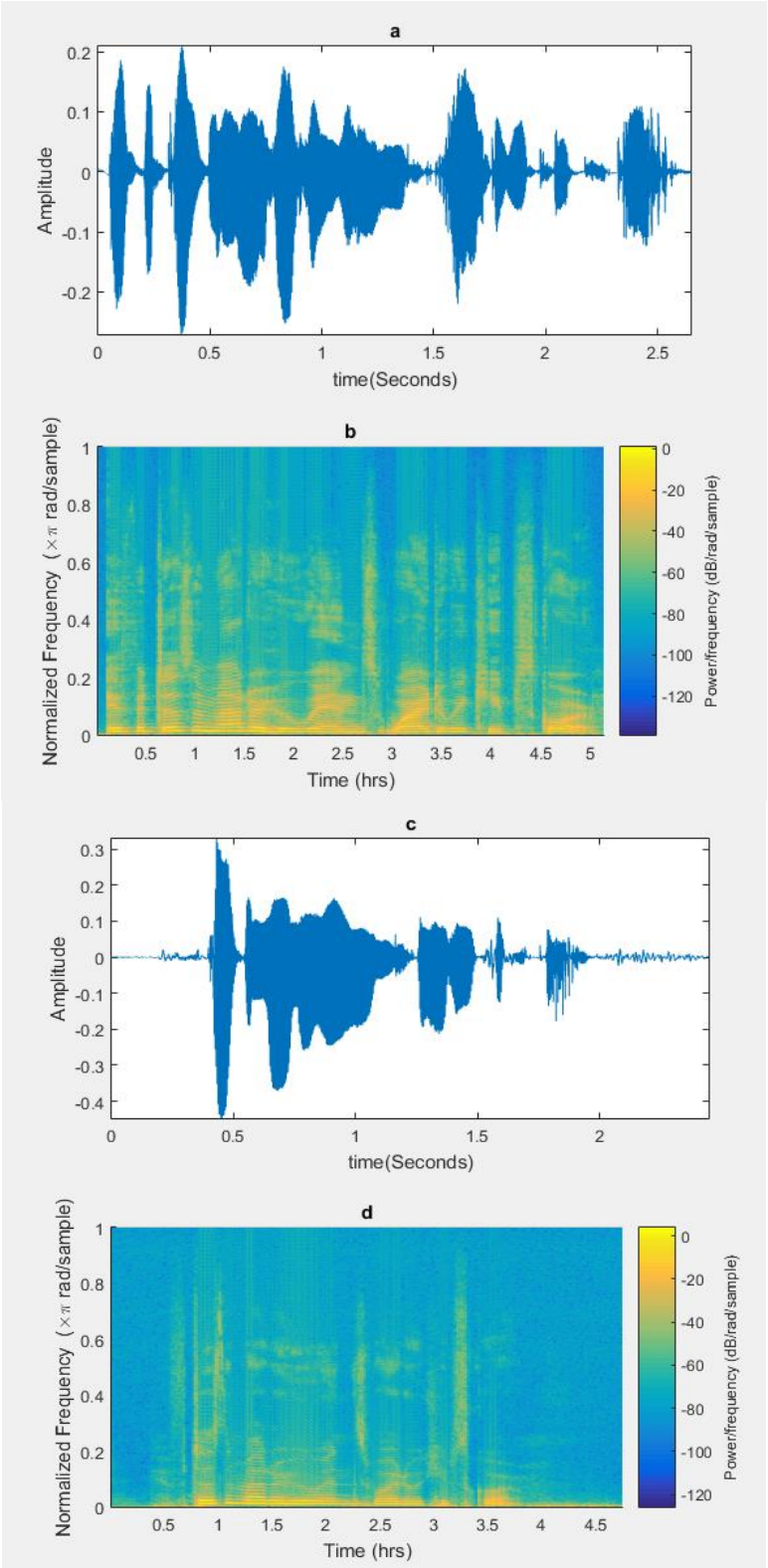

Fig. 3: *Speech Waveform and Spectrogram of (a, b) Normal Speech (c, d) Fast Speech.*

Four different train-test conditions like solo-solo, whisper-solo, solo-whisper and whisper-whisper are implemented to check the performance of the system. From the Table 1, it is cleared that when the system is trained and tested with same type of database, the identification rate is high and Equal Error Rate (EER) is very low but when the mismatch of database is taken during training and testing the performance of the system is degraded

considerably. Among four different conditions when system is trained with Normal speech and tested against whisper speech the identification rate is 5.39% and EER is 38.38% which is quite poor compared to same training and testing condition [10].

In the second experiment, performance of speaker recognition system is evaluated with normal and fast speech. Four different train-test conditions like solo-solo, fast-solo, solo-fast and fast-fast are considered to evaluate the performance of the system. Table 2 shows the result when the system is trained and tested with different speech data i.e. normal and FAST speech. When the system is trained and tested with either normal or FAST speech the performance of the system is excellent but when the mismatch of train and test condition is taken the identification rate and EER are degraded. When the system is trained with Fast speech and tested against normal speech the identification rate is 27.69% and EER is 24.33% which is quite poor compared to same training and testing condition [11].

From both the experiments, it is observed that when mismatch of normal and whisper speech is considered, the EER is 38.38% and the identification rate is only 5.39% which is poor compared to mismatch of normal and fast speech. It is also noticed that train-test condition normal-normal, fast-fast and whisper-whisper is giving better results while whisper-normal, normal-whisper, fast-normal and normal-fast gives very less identification rate. Hence it shows that training and testing of same speech gives better recognition performance [12].

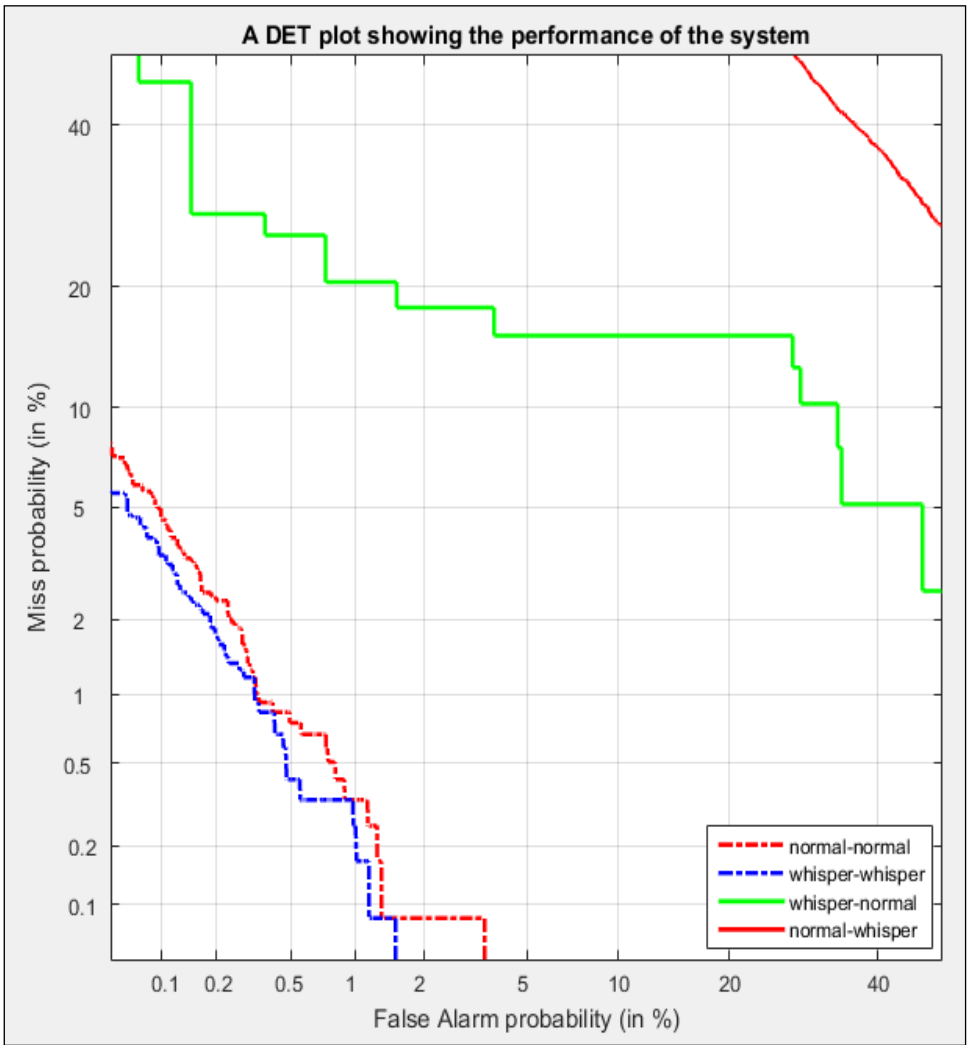

Fig. 4: *DET Curve for CHAINS Normal-whisper Database for Different Train-Test Conditions.*

Table 1: *% Identification and EER for different Train-Test Conditions on CHAINS database (Normal-Whisper).*

Train-Test	%Identification Rate	%EER	DCF
Normal-Normal	99.75	0.67	0.0061
Whisper-Normal	8.33	15.38	0.0960
Normal-Whisper	**5.39**	**38.38**	**0.3806**
Whisper-Whisper	99.92	0.51	0.0044

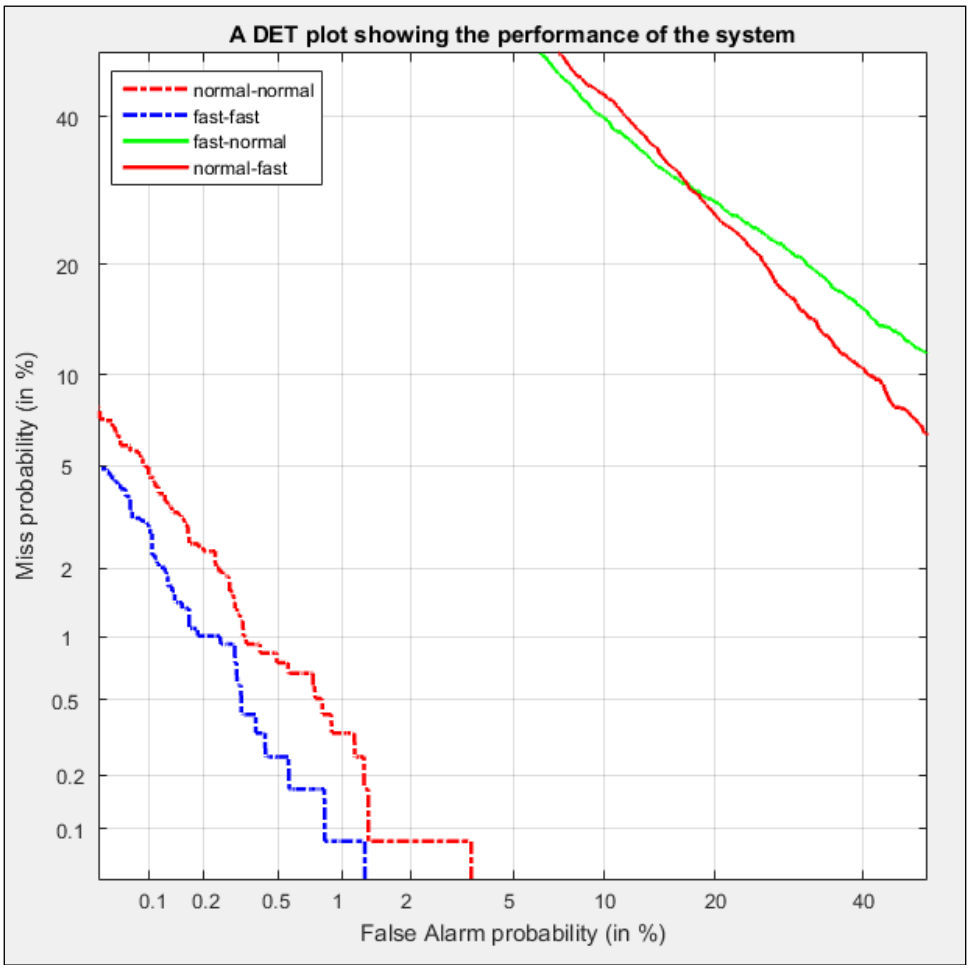

Fig. 5: *DET Curve for CHAINS Normal-fast Database for Different Train-test Conditions.*

Table 2: *% Identification and EER for Different Train-Test Conditions on CHAINS Database (Normal-Fast).*

Train-Test	%Identification Rate	% EER	DCF
Normal-Normal	99.75	0.67	0.0061
Fast-Normal	**27.69**	**24.33**	**0.2332**
Normal-Fast	25.34	23.15	0.2265
Fast-Fast	100	0.42	0.0034

CONCLUSIONS

This paper investigated the effects of speaking style and speaking rate on speaker recognition system. Generally, the speaker recognition system is tested with the same type of speech. It is concluded that, when the system performance is evaluated with same database, the performance is excellent. System

performance can be further improved by considering delta and delta-delta MFCC features. We propose to analyze the system with other features and combine it with MFCC feature to improve the performance of the system.

REFERENCES

1. Douglas O'Shaughnessy. *Speech Communications: Human and Machine*, second edition. New Jersey: Wiley-IEEE Press; 2009.
2. Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*. 2010; 52 (1): 12-40p.
3. Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transaction on Acoustics, Speech, Signal Processing*. 1980; 28: 357-366p.
4. Huang X, Acero A, Hon H. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. New Jersey: Prentice-Hall; 2001.
5. Douglas A Reynolds, Quatieri T, Dunn R. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 2000; 10(1): 19-41p.
6. Thomas F Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ, USA: Prentice Hall Press; 2001.
7. Krishnamoorthy P, Jayanna HS, Prasanna SRM. Speaker recognition under limited data condition by noise addition. *Expert Systems with Applications*. 2011; 38: 13487-13490p.
8. Cummins, Fred. CHAINS: Characterizing Individual Speakers. [Online] CHAINS database, available from http://chains.ucd.ie/ [Accessed December 2018].
9. Microsoft Research. (2013). Emerging Technology, Computer, and Software Research. [Online] MSR identity toolkit, Microsoft research, available from http://research.microsoft.com/ [Accessed December 2018].
10. JPC Jr. Speaker recognition: a tutorial. *Proceedings of the IEEE*. 1999; 85 (9): 1437–1462p.
11. Thomas F Quatieri, Robert B Dunn, Douglas A Reynolds. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*. 2000; 10: 19-41p.
12. Douglas Reynolds. Speaker identification and verification using Gaussian Mixture Models. *Speech Communications*. 1995; 17: 91–108p.

Cite this Article

Pinky J. Brahmbhatt1, K.G. Maradia, Performance of Speaker Recognition System in Mismatching Speaking Style. *Journal of Artificial Intelligence Research & Advances*. 2018; 5(3): 58–65p.