International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 4, Issue 5, Aug - Sep 2015, 7-16

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

TEXT- INDEPENDENT MULTI-SENSOR SPEAKER VERIFICATION SYSTEM

KSHIROD SARMAH[1] & UTPAL BHATTACHARJEE[2]

[1]Department of Computer Science, NERIM Group of Institutions, Guwahati, Assam, India

[2]Department of Computer Science and Engineering, Rajiv Gandhi University, Itanagar, Arunchal Pradesh, India

ABSTRACT

The performance of common speaker verification (SV) system vastly affected when speaker model training is done in the speech sample that recorded done by one device and the testing is done in another device. It is a major problem of speaker verification system in multi-device environment. In this paper we report the experiment carried out on the recently collected speaker recognition database Arunachali Language Speech Database (ALS-DB) a multilingual and multichannel database to study the impact of device variability on speaker verification system. The collected database is evaluated with Gaussian mixture model and Universal Background Model (GMM-UBM) and Mel - Frequency Cepstral Coefficients (MFCC) combined with prosodic features as a front end feature vectors based speaker verification system. The impact of the device both matching and mismatch in training and testing has been evaluated in text independent manner. For matching condition of device we have found Equal Error Rate (EER) **7.50**% with minimum Detection Cost Function (MinDCF) value **0.1062** and for mismatching condition of devices that of **18.70**% with MinDCF value **0.3425**. The performance of the SV system has degraded approximately **11.00**% due to mismatching condition of devices in text independent speaker verification system.

KEYWORDS: GMM-UBM, MFCC, Multi-Sensor, Prosodic, Speaker Verification

INTRODUCTION

Automatic Speaker Recognition (ASR) refers to recognizing persons from their voice. The sound of each speaker is identical because their vocal tract shapes, larynx sizes and other parts of their voice production organs are different. ASR System can be divided into either (1) Automatic Speaker Verification (ASV) or (2) Automatic Speaker Identification (ASI) systems. Speaker verification aims to verify whether an input speech corresponds to the claimed identity. Speaker Verification is the task of determining whether a person is who he or she claims to be (a yes/ no decision).Since it is generally assumed that imposter (falsely claimed speaker) are not known to the system, so it is also referred to as an Open-Set task [1].

The state-of-art speaker verification system use either adaptive Gaussian mixture model (GMM) [2] with universal background model (UBM) or support vector machine (SVM) over GMM super-vector [3]. Currently, SVM is one of the most robust classifiers in speaker verification, and it has also been successfully combined with GMM to increase accuracy [4, 5]. Mel-frequency Cepstral coefficients are most commonly used feature vector for speaker verification system. Supra-segmental features like – prosody, speaking style are also combined with the cepstral feature to improve the performance [6].

Till date, most of the speaker verification system operates only in text dependent as well as a single- sensor (device) environment. Multi-channel and multilingual speaker recognition is the key to the development of spoken dialogue systems

that can function in multi-device environments robustly. In the multi-channel speaker verification system the channel factors and speaker factors play important but different roles which are combined in Joint Factor Analysis (JFA).Multilingual speaker verification (MSV) system also very important area of research like India, one of the most favorable multilingual countries of the world. The performance of MSV found a little degrades due to mismatching phonetic structure of different languages spoken by the same speaker [7]. In this paper we concentrates only on channels affect in text independent speaker verification system.

Channel compensation in the front-end processing addresses linear channel effects, but there is evidence that handset transducer effects are nonlinear in nature and are thus difficult to remove from the features prior to training and recognition [8]. Because the handset effects remain in the features, the speaker's model will represent the speaker's acoustic characteristics coupled with the distortions caused by the handset from which the training speech was collected. Speaker same likelihood the same speaker the effect is that log-likelihood ratio scores produced from different speaker models can have handset-dependent biases and scales. This is especially problematic when trying to use speaker-independent thresholds in a system, as is the case for the NIST SREs [3].

Poor-quality microphones introduce nonlinear distortion to the true speech spectrum. Quatieri & al. [8] demonstrate, by comparing pairs of same speech segment recorded with good- and poor-quality microphones, that poor-quality microphones introduce several spectral artifacts, such as phantom formants that occur at the sums, multiples and differences of the true formants. Formant bandwidths are also widened and the overall spectral shape is flattened which affect in the speech features in speaker recognition system.

The A/D converter adds its own distortion, and the recording device might interfere with a mobile phone radio-waves. If the speech is transmitted through a telephone network, it is compressed using lossy techniques which might have added noise into the signal. Speech coding can degrade speaker recognition performance significantly [10, 11].

To evaluate the text independent speaker verification system in multi-sensor environment, a multi-lingual and multi-sensor speaker recognition database has been developed and initial experiments were carried out to evaluate the impact of language variability on the performance of the baseline speaker verification system [12, 13]. In this work, we are going to discuss how device variability affected the performance of a SV system.

SPEAKER RECOGNITION DATABASE

In this section we describe the recently collected a Multi-devices and Multilingual speech corpus namely, Arunachali Language Speech Database (ALS-DB) [13]. Arunachal Pradesh of North East India is one of the linguistically richest and most diverse regions in all of Asia, being home to at least thirty and possibly as many as fifty distinct languages in addition to innumerable dialects and subdialects thereof [14].To study the impact of device variability on speaker recognition task, ALS-DB is collected in multi-device environment. Each speaker is recorded for three different languages – English, Hindi and a local language, which belongs to any one of the four major Arunachali languages - Adi, Nyishi, Galo and Apatani. Each recording is of 4-5 minutes duration. Speech data were recorded in parallel across four recording devices, which are listed in Table 1.

Table 1: Different Type of Devices and Recording Specifications

Device Sl. No	Device Type	Sampling Rate	File Format
Device 1	Table mounted microphone	16 kHz	wav
Device 2	Headset microphone	16 kHz	wav
Device 3	Laptop microphone	16 kHz	wav
Device 4	Portable Voice Recorder	44.1 kHz	mp3

The speakers are recorded for reading style of conversation. The speech data collection was done in laboratory environment with air conditioner, server and other equipments switched on. The speech data was contributed by 100 male and 98 female informants chosen from the age group 20-50 years. During recording, the subject was asked to read a story from the school book of duration 4-5 minutes for twice and the second reading was considered for recording. Each informant participates in four recording sessions and there is a gap of at least one week between two sessions.

FEATURE VECTORS

MFCCs Computation

The computation of the MFCCs is consists of several stages. The first stage is pre-emphasis followed by the short-time Fourier analysis on an overlapping Hamming window. After that, we can extract either the power or the magnitude of the Fourier coefficient. Afterwards, a filterbank transformation is applied to transform the signal into a smooth spectrum representation close to the envelope of the speech signal. The output of the filterbank then transform to the log domain. Finally , to decorrelate and produce the cepstral coefficients we apply DCT. The filterbank can be either linear or mel scale. Mel scale that resembles the way a person hears is applied here.

If the output of an M –channel filterbank as Y(m), m=1,2……..,M, Then MFCCs are obtained as follows:

$$C_n = \sum_{m=1}^{M} [logY(m)] cos\left[\frac{\pi n}{M}(m - \frac{1}{2}\right] \quad (1)$$

Here n is the index of the cepstral coefficient. The final MFCC vector is obtained by retaining about 12-15 lowest DCT coefficients.

In the final step, the Mel-spectrum plot is converted back to the time domain by using the following formula:

$$Mel(f) = 2595*\log10(1+f/700) \quad (2)$$

Where f is linear frequency.

To emphasize the dynamic features of the speech in time, the time derivative (Δ) and the time –acceleration (ΔΔ) are usually computed. It is common to compute 12 MFCC, one Energy coefficient and its corresponding (Δ) and (ΔΔ).

Prosodic Features

Prosodic features are the rhythmic and in intonational properties in speech, examples are voice fundamental frequency (F0), F0 gradient, intensity and duration. Prosody refers to non-segmental aspects of speech, including for instance syllable stress, intonation patterns, speaking rate and rhythm. One important aspect of prosody is that, unlike the traditional short-term spectral features, it spans over long segments like syllables, words, and utterances and reflects differences in speaking style, language background, sentence type, and emotions to mention a few. A challenge in

text-independent speaker recognition is modeling the different levels of prosodic information (instantaneous, long term) to capture speaker differences; at the same time, the features should be free of effects that the speaker can voluntarily control.

The most important prosodic parameter is the fundamental frequency (or F0). Combining F0-related features with spectral features has been shown to be effective, especially in noisy conditions. Other prosodic features for speaker recognition have included duration (e.g. pause statistics, phone duration), speaking rate, formants, pitch and energy distribution/modulations among others [16, 17,18]. In that study, it was found out, among a number of other observations, that F0-related features yielded the best accuracy, followed by energy and duration features in this order.

Prosody features have also proven to be robust in the noisy and multi-channel environment. Therefore, these features show very great potential for the speaker verification tasks.

GMM-UBM AS A CLASSIFICATION METHOD

The GMM-UBM approach for speaker verification system can be considered primarily as a four phase process. At the first phase, a gender independent UBM model is generated which is a GMM that built based on the Expectation-Maximization (EM) algorithm and using utterances from a very large population of speakers [3]. The target speaker specific models are then obtained through the adaptation of mean from the UBM using the speaker's training speech and a modified realization of the maximum a posteriori (MAP) approach [3]. In the testing phase, a fast scoring procedure is used in order to reduce the number of computations [3]. This involves determining the top few scoring mixtures in the UBM for each feature vectors and then computing the likelihood of the target speaker model using the score for its corresponding mixtures. The scoring process is then repeated for all the feature vectors in the test utterance to obtain the average log likelihood score for each of the UBM and the target speaker model. Finally, UBM-based normalization is performed by subtracting the log likelihood score of the UBM from that of the target speaker model. This is firstly to minimize the effect of unseen data, and secondly to deal with the data quality mismatch [3].

A GMM is a probabilistic model for density estimation using a mixture distribution and is defined as a weighted sum of multi-variate Gaussian densities.

A GMM is a weighted sum of M component densities is given by the form

$$P(x|\lambda) = \sum_{i=1}^{M} w_i \, b_i(\mathrm{x}) \tag{3}$$

Where x is a dimensional random vector, b_i(x), i =1,2……M, is the component densities and w_i i=1,2,….,M, is the mixture weights.

The Gaussian Function can be defined of the form

$$b_i(\mathrm{x}) = \frac{1}{(2\pi)^{\frac{\mathrm{D}}{2}} |\sum_{\mathrm{i}}|^{1/2}} \exp\left\{-\frac{1}{2}\left(x-\mu_i\right)' \sum_i^{-1}\left(x-\mu_i\right)\right\} \tag{4}$$

with mean vector μ_i and covariance matrix $\sum_i$. The mixture weight satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weight from all component densities.

These parameters can collectively represented by the notation:

$$\lambda = \{ w_i, \mu_i, \sum_i \} \quad \text{for i= 1,2 ……, M.} \tag{5}$$

In speaker verification system, each speaker can be represented by such a GMM and is referred to by the above model λ.

For a given training vectors and a GMM configuration, we have to estimate the parameters of the GMM, λ, for the best matches the distribution of the training feature vectors. The most popular and well-known method is maximum likelihood (ML) estimation.

The main purpose of ML estimation is to find the model parameters which maximize the likelihood of the GMM given the training data. For a sequence of T training vectors X= $\{x_1, x_2, x_3, \ldots\ldots x_T\}$ the GMM likelihood can be defined as

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda). \tag{6}$$

The speaker-specific GMM parameters are estimated by the Expectation-Maximization (EM) algorithm using training data spoken by the corresponding speaker. The basic idea of the EM algorithm is, beginning with an initial language model λ, to estimate a new model λ' such that $P(X|\lambda') \geq P(X|\lambda)$. The new model then becomes the initial model for the next iteration and the process is repeated until some convergence threshold is reached [19].

On each EM iteration, the following re-estimation formulas are used which guarantee a monotonic increase in the model's likelihood value,

Mixture Weights:

$$w_i = \frac{1}{T} \sum_{t=1}^{T} pr(i|x_t, \lambda) \tag{7}$$

Means:

$$\mu_i = \frac{\sum_{t=1}^{T} pr(i|x_t,\lambda) x_t}{\sum_{t=1}^{T} pr(i|x_t,\lambda)} \tag{8}$$

Variance (diagonal covariance):

$$\sigma_i^2 = \frac{\sum_{t=1}^{T} pr(i|x_t,\lambda) x_i^2}{\sum_{t=1}^{T} pr(i|x_t,\lambda)} - \mu_i^2 \tag{9}$$

The a posteriori probability for component i is given by

$$pr(i|x_t, \lambda) = \frac{w_i b_i(x)}{\sum_{k=1}^{M} w_k b_k(x)} \tag{10}$$

There are lots of reasons to consider in contrasting one of the standard MAP approaches to its iterative form. The standard MAP technique is simply a single iteration while EM based result is iterative. A single iteration assumes that the mixture mean components vary in a completely independent manner [20], and consequently, one a single iteration would be required to solve the MAP solution.

BASELINE SYSTEM

In this works, the baseline system, a speaker verification system was developed using Gaussian Mixture Model with Universal Background model (GMM-UBM) based modeling approach. A 39-dimensional feature vector was used, made up of 13 mel-frequency cepstral coefficient (MFCC) and their first order derivatives as well as second order

derivatives. The first order derivatives were approximated over three samples. The coefficients were extracted from a speech sampled at 16 KHz with 16 bits/sample resolution. A pre-emphasis filter $H(z) = 1\text{-}0.97z^{-1}$ has been applied before framing. The pre-emphasized speech signal is segmented into frame of 20 ms with frame frequency 100 Hz. Each frame is multiplied by a Hamming window. From the windowed frame, FFT has been computed and the magnitude spectrum is filtered with a bank of 22 triangular filters spaced on Mel-scale and constrained into a frequency band of 300-3400 Hz. The log-compressed filter outputs are converted to cepstral coefficients by DCT. The 0^{th}cepstral coefficient is not used in the cepstral feature vector since it corresponds to the energy of the whole frame [14], and only 12 MFCC coefficients have been used. To capture the time varying nature of the speech signal, the first order and second order derivative of the Cepstral coefficients are also calculated. Combining the MFCC coefficients with its first order and second derivatives, we get a 36-dimensional feature vector.

In the next phase, 6 dimensional prosodic features vector consist of pitch, short time energy and its first and second order derivatives (Δpitch, Δenergy, ΔΔpitch and ΔΔenergy) that have been combined with the 36 dimensional MFCC features vector. As a result, we got a 42-dimension feature vectors.

Cepstral Mean Subtraction (CMS) has been applied on all features to reduce the effect of channel mismatch. In this approach we apply Cepstral Variance Normalization (CVN) which forces the feature vectors to follow a zero mean with unit variance distribution in feature level solution to get more robustness results.

The Gaussian mixture model with 1024 Gaussian components has been used for both the UBM and speaker model. The UBM was created by training the speaker model with 50 male and 50 female speaker's data with 512 Gaussian components each male and female model with Expectation Maximization (EM) algorithm. Finally UBM model is created by pulling the both male and female models and finding the average of all these models [20]. The speaker models were created by adapting only the mean parameters of the UBM using maximum a posteriori (MAP) approach with the speaker specific data.

The detection error trade-off (DET) curve has been plotted using log likelihood ratio between the claimed model and the UBM and the equal error rate (EER) obtained from the DET curve has been used as a measure for the performance of the speaker verification system. Another measurement Minimum DCF values has also been evaluated.

EXPERIMENTS AND RESULTS

All the experiments reported in this paper are carried out using the database ASL-DB described in section 2. An energy based silence detector is used to identify and discard the silence frames prior to feature extraction. Data from the all devices have been considered in the present study. All the four available sessions were considered for the experiments. Each speaker model was trained using one complete session. The test sequences were extracted from the next three sessions. The training set consists of speech data of length 120 seconds per speaker. The test set consists of speech data of length 15 seconds, 30 seconds and 45 seconds. The test set contains more than 3500 test segments of varying length and each test segment will be evaluated against 11 hypothesized speakers of the same sex as segment speaker [22].

Experiments

In this experiment any type of language has been considered for training the system from the speech data from session1 recorded that of device1, device 2,device 3 and device 4 separately and all four devices have been considered

separately for testing the system taken the testing data from the second, third or fourth session. The result of the experiments has been summarized in Table 2. Figure 1, Figure 2, Figure 3 and Figure 4 show the DET curves obtained for the four devices in the speech database.

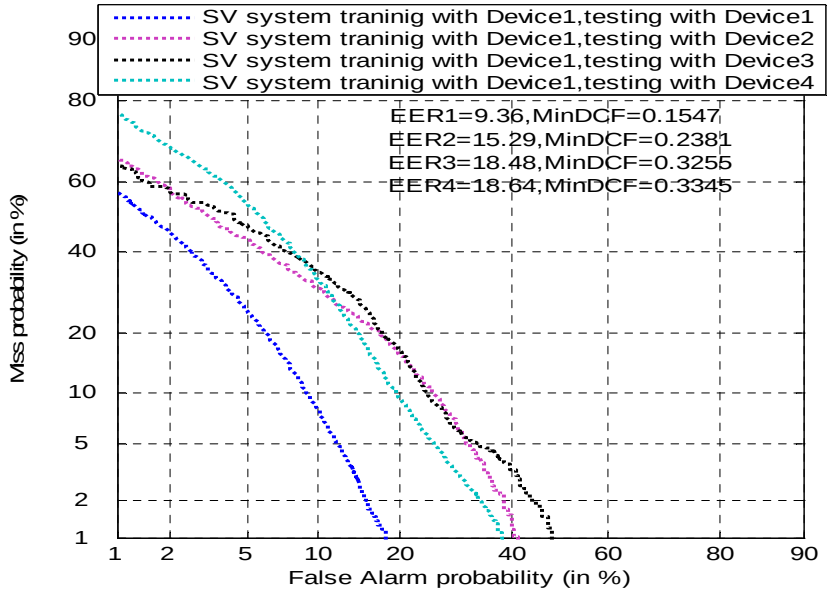

Figure 1: DET Curve for the Speaker Verification System for Training with Device 1 and Testing with Device 1, Device 2, Device 3 and Device 4 Respectively

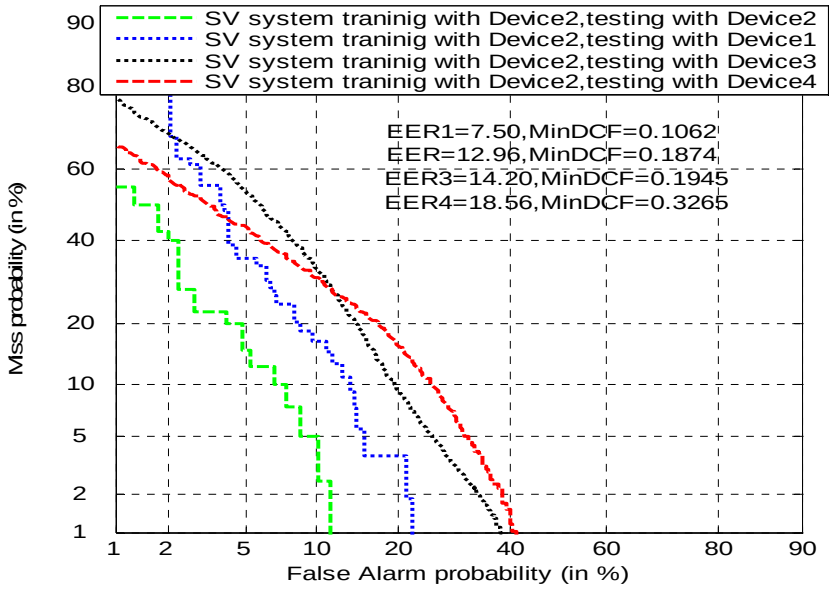

Figure 2: DET Curve for the Speaker Verification System for Training with (a) Local Language of Device 2 and Testing with the Same Language with Device 1, Device 2, Device 3 and Device 4 Respectively

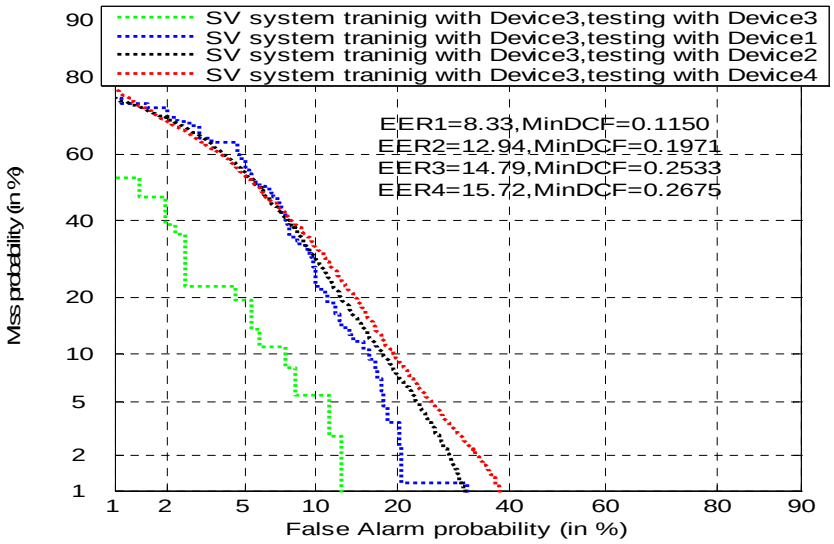

Figure 3: DET Curve for the Speaker Verification System for Training with (a) Local language of Device 3 and Testing with the Same Language with Device 1, Device 2, Device 3 and Device 4 Respectively

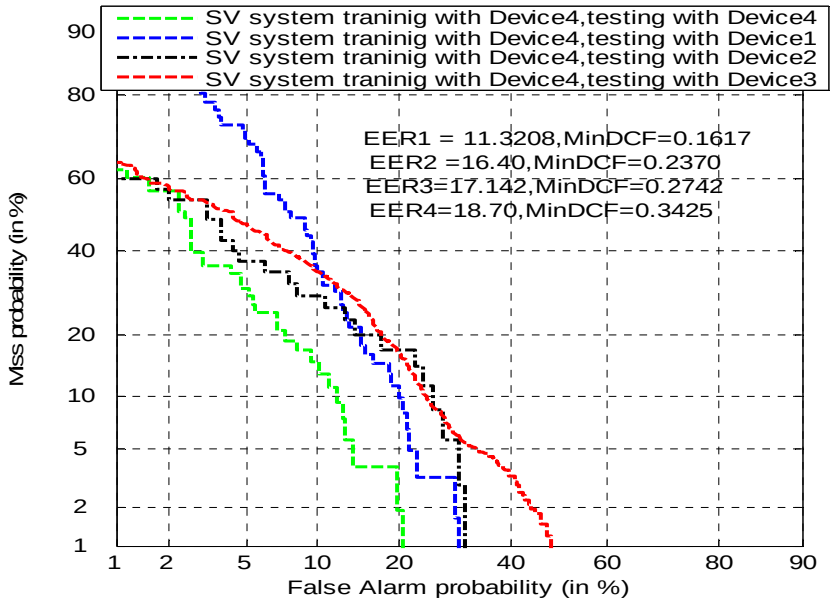

Figure 4: DET Curve for the Speaker Verification System for Training with (a) Local Language of Device 4 and Testing with the Same Language with Device 1, Device 2, Device 3 and Device 4 Respectively

Table 2: EER and Min DCF Values for Speaker Verification System for Training and Testing with Mismatching of Devices

Training Devices	Testing Devices	ERR%	Verification Rate%	Min DCF
Device 1	Device 1	**9.36**	**90.64**	**0.1547**
	Device 2	15.29	84.71	0.2387
	Device 3	18.48	81.52	0.3255
	Device 4	18.64	81.36	0.3345
Device 2	Device 2	**7.50**	**92.50**	**0.1062**
	Device 1	12.96	87.04	0.1874

	Device 3	14.20	85.80	0.1945
	Device 4	18.56	81.44	0.3265
Device 3	Device 3	**8.33**	**91.67**	**0.1150**
	Device 1	12.94	87.06	0.1971
	Device 3	14.79	85.21	0.2533
	Device 4	15.72	84.18	0.2675
Device 4	Device 4	**11.32**	**88.68**	**0.1617**
	Device 1	16.40	83.60	0.2370
	Device 2	17.14	82.86	0.2742
	Device 3	**18.70**	**81.30**	**0.3425**

CONCLUSIONS

From the experimental point of view we come to conclude that the performance of the speaker verification system was better in matching condition of the device. But due to mismatching of channels that recording in both training and testing condition the performance of the SV system was degraded. For matching condition of channels we found EER rate **7.50**% with minimum DCF value **0.1062** and for mismatching condition of channels we found EER rate **18.70**% with minimum DCF value **0.3425**. The performance of the SV system has degraded approximately **11.00%** due to mismatching condition of sensors.

ACKNOWLEDGEMENTS

This work has been supported by the ongoing project grant No. 12(12)/2009-ESD sponsored by the Department of Information Technology, Government of India.

REFERENCES

1. D. A. Reynolds, : An overview of Automatic Speaker Recognition Technology, MIT Lincoln Laboratory, 244 wood St. Lexinton, MA 02140,USA, *IEEE* (2002)
2. D. A. Reynolds,: Robust text-independent speaker identification using Gaussian mixture speaker models, *Speech Communications*, vol. 17, pp. 91-108 (1995)
3. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10(1–3), pp. 19-41, (2000)
4. E. Shriberg et al., :Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Commun*,vol.46, no.3-4, pp 455-472, (2005)
5. L. Ferrer et al. : Parameterization of Prosodic Feature Distribution for SVM Modeling in Speaker Recognition, *Proc. ICASSP*,vol.4,2007, pp.233-236.
6. B.C. Haris, G. Pradhan, A. Misra, S. Shukla, R. Sinha and S.R.M. Prasanna, "Multi-variability Speech Database for Robust Speaker Recognition", *In Proc. NCC*, pp. 1-5, (2011)
7. U. Bhattacharjee And K.Sarmah, "GMM-UBM Based Speaker Verification in Multilingual Environments", *International Journal of Computer Science Issues (IJCSI).*Vol. 9,Issue 6,No.2, ISSN:1694-0814, November 2012, pp.373-380.

8. D. A. Reynolds, M. Zissman, T. F.Quatieri, G.O'Leary, and B.Carlson, " The effects of telephone transmission degradations on speaker recognition performance". *In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, May pp. 329–332 ,(1995)

9. T.F.Quatieri, D.A. Reynolds, and G.O'Leary, "Estimation of handset nonlinearity with application to speaker recognition". *In IEEE Trans. on Speech and Audio Processing* 8, 5 pp.567–584, (2000)

10. M. Phythian, J.Ingram, and S.Sridharan, "Effects of speech coding on text-dependent speaker recognition", *In Proceedings of IEEE Region 10 Annual Conference on Speech and Image Technologies for Computing and Telecommunications* (TENCON'97) pp. 137–140, (1997)

11. L. Besacier, S. Grassi, A. Dufaux, M.Ansorge, and F. Pellandini, "GSM : speech coding and speaker recognition". *In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2000)* (Istanbul, Turkey, pp. 1085– 1088, (2000)

12. U.Bhattacharjee, and K. Sarmah, "A Multilingual Speech Database for Speaker Recognition", *In Proc. IEEE, ISPCC*, March (2012)

13. U.Bhattacharjee, and K.Sarmah, "Development of a Speech Corpus for Speaker Verification Research in Multilingual Environment", *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307, Volume-2, Issue-6, January (2013)

14. Arunachal Pradesh, http://en.wikipedia.org/wiki/Arunachal_Pradesh.

15. Z.Xiaojia, S.Yang and W. DeLiang, "Robust speaker identification using a CASA front-end, Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp.5468-5471, (2011)

16. A.Adami, R. Mihaescu, D.A. Reynolds, and J.Godfrey, " Modeling prosodic dynamics for speaker recognition", *In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)* (Hong Kong, China, April 2003), pp. 788–791 (2003)

17. K.Bartkova, D.L.Gac, D. Charlet, and D. Jouvet, "Prosodic parameter for speaker identification", *In Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)* (Denver, Colorado, USA, September 2002), pp. 1197–1200, (2002)

18. D.A. Reynolds, W. Andrews, J.Campbell, J. Navratil, B.Peskin, A. Adami, , Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J.Godfrey, D. Jones, and B. Xiang, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition". *In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)* (Hong Kong, China, April 2003), pp. 784–787, (2003)

19. D. A. Reynolds, "Gaussian Mixture Models*", MIT Lincoln Laboratory, 244 wood St. Lexinton, MA 02140,USA.

20. J.Pelecanos, R.Vogt, and S. Sridharan, "A study on standard and iterative MAP adaptation for speaker recognition", *Proceeding on the 9th Australian International Conference on Speech Science & Technology* Melbourne, December 2 to 5 (2002)