Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Multligual Speaker Verification with Different Normalization Techniques

Kshirod Sarmah[1], Utpal Bhattacharjee[2]

Department of Computer Science and Engineering,
Rajiv Gandhi University, Rono Hills, Doimukh,
Arunachal Pradesh, India, Pin-791112,

Abstract— A Multilingual Speaker verification (MSV) system has shown very poor performance when speaker model training is done in one language while the testing is done in another. It is because of mismatching of phonetic contents of speech utterances, construction of dialects variability, different speaking styles, accents and other language dependent attributes of different languages, which is a major problem. In this paper we report the experiment that carried out on the recently collected multilingual speaker recognition database Arunachali Language Speech Database (ALS-DB) and developed a baseline system for speaker verification in multilingual environment applying different normalization techniques. The collected database is evaluated with Gaussian Mixture Model-Universal Background Model (GMM-UBM) and Mel- Frequency Cepstral Coefficients (MFCC) with its first and second order derivatives combined with Prosodic features as a front end feature vectors. Typically, the speaker model has been constructed by applying MAP adaptation algorithm from the UBM. The performance of the speaker verification system has been improved by applying Cepstral Mean Normalization (CMN) and Cepstral Variance Normalization (CVN) at the feature level and score normalization technique Zero-normalization (Z-Norm), test-normalization (T-norm) in the score level as well as distance-normalization (D-norm) at speaker model level. In this works, it has been observed the performance of the MSV system in terms of EER of 11.08%,10.30%,10.00%, and 8.90% for the GMM-UBM + Z-Norm, GMM-UBM+ T-Norm, GMM-UBM+ D-Norm and GMM-UBM + T-Norm + D_Norm respectively for language matching and mismatching environments . It has been observed that for language mismatching condition D-Norm shows better performance than T-Norm and Z-Norm. Combining T-Norm and D-Norm the performance of MSV system improved by approximately 2.00% of its recognition rate. Similarly, for language matching conditions T-Norm shows better performance than that of Z-Norm and D-Norm. The performance of the MSV system enhanced up to 95.00% accuracy of recognition rate, while applying the combined T-Norm with D-Norm in the same baseline system.

Keywords— Speaker Verification, GMM-UBM, MFCC, Prosodic, Z-Norm, T-Norm, D-Norm.

I. INTRODUCTION

A speaker verification (SV) system needs to determine whether or not a person is indeed who he or she claims to be, based on one or more spoken utterances produced by that individual. In a text-dependent setup, a predetermined group of words or sentences are used to enrol a set of speakers, and these words or sentences are then used to verify the speakers [1]. In a text-independent application, there is no prior knowledge by the system of the text to be spoken by speaker [2].

In text independent speaker verification applications, the principal state-of-the-art approach is based on Gaussian Mixture Models (GMM) [3]. The generative model is generally trained using maximum likelihood (ML) principle. The main disadvantage of the ML approach is that it doesn't generalize well to unseen speech data with finite amount of training material. To solve this problem Maximum a posteriori (MAP) approach of training is sufficient which is also known as universal background model (UBM) [3]. Speaker verification is based on a likelihood ratio calculated using a Maximum A-Posteriori (MAP) adapted GMM from a Universal Background Model (UBM). In MAP approach, prior knowledge of the distribution of model parameters is incorporated into modeling process [4].

Speaker Verification is one of very complex task with many factors that speaker identity, recording environment, transmission channel, utterance length, utterance type, gender, session, speaking style, speaker traits (like dialect, accent, stress), phonetic contents etc. SV technology also based on statistical pattern recognition tasks which represent the identity of speakers. For these above factor's variability and mismatching between training and testing in any of these acoustic variables results in performance degradation of SV system.

In a MSV system, each speaker speaks more than one or two languages (native language or secondary languages) which are not necessarily the same language in both training and testing. For mismatching of phonetic contents of different languages for the same speaker in different sessions highly degrades the performance of MSV system. The performance of a GMM-UBM based speaker verification system degrades considerably with change in training and testing language [22]. Therefore, some compensation techniques are must needed to cope with speech and speaker variability to improve the performance in MSV system. These successful compensation techniques are also well known normalization techniques. Normalization techniques have been proposed and applied basically in three levels, that is at feature level [23] ,at model level [17] and at score level [3,16,19].

Feature domain compensation is aimed at removing the channel effects and unwanted noises from the feature vectors prior to speaker model training [18] e.g., Cepstral Mean Subtraction (CMS) or Cepstral Mean Normlaization (CMN) and Cepstral Variance Normalization (CVN), etc. In model level normalization technique it is tried to modify trained speaker models to minimize the effects of various acoustic factors, for example D-Norm which is based on the Kullback-Leibler (KL) distance between the claimed speaker model and imposter model [20]. Finally, score normalization is the transformation of speaker verification output scores to enhance the effectiveness of the detection threshold by aligning the score distribution speaker models. For example Z-Norm attempts to align between-speaker differences of imposter scores distributions [7] and T-norm is another popular score normalization methods which can be considered as the enhanced version of Z-Norm. T-norm speaker models are scored in parallel with the target speaker model [7].As the adapted universal background model (UBM) provides fast scoring so T-norm is efficient in an adapted UBM system [8].

Till date, most of the speaker verification system operates only in a single-language environment. Multilingual speaker recognition and language identification are key to the development of spoken dialogue systems that can function in multilingual environments [9]. For a highly multilingual country like India, the effect of multiple languages on state-of-art speaker verification system needs to be investigated. Most of the publicly available databases for speaker verification research are developed in western context, which is not suitable for evaluating the performance of the system in Indian context. Further, the linguistic scenario of North-East India is different from the rest of India. This is the region where two major linguistic families- Indo-European and Tibeto-Burman meet together and speak each others' language fluently.

To evaluate the speaker verification system in multi-lingual environment, a multi-lingual 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 [10][11].The rest of the paper is organized as follows: Section–2 describes the details of the speaker verification database. Section–3 details the architecture of speaker verification system. The baseline system is briefly explained in Section-4, the experiment and result obtained are described in Section-5.

II. SPEAKER VERIFICATION CORPUS

In this section we used the recently collected Arunachali Language Speech Database (ALS-DB) [10][11][22].

To study the impact of language variability as well as sensor variability on speaker verification task, ALS-DB is a multilingual and multichannel speech database. 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 (i) Device 1: Table mounted microphone, (ii) Device 2: Headset microphone, (iii) Device 3: Laptop microphone and (iv) Device 4: Portable Voice Recorder.

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 100 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 in each language 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. In this experiment we only concentrate on the speech data of Device 2: headset microphone with for the Galo linguistic group of speakers who can speak three languages namely Local (Galo language), Hindi and English clearly.

III. ARCHITECTURE OF SPEAKER VERIFICATION SYSTEM

The entire Speaker Verification is to verify correctly the identity of the claim speaker whether he/she accept or reject from the given speech segment. For verification, it can be considered as a classification problem which is based on hypothesis testing. As we know there are two types of error can be seen in this SV system. First one is known as false rejection which means that the testing speech utterance of the true speaker is rejected incorrectly, another one is known as false acceptance means that incorrectly accepted the speech utterance of impostor.

In this case, we can consider two hypothesizes:

H_0 : X , speech segment is from the hypothesis speaker S,
H_1 : X is not from the hypothesis speaker S.

The optimum test to decide between these two hypotheses is a likelihood ration (LR) [20] test that given by

$$\Lambda(X) = \frac{p(X|H_0)}{p(X|H_1)} = \frac{p(X|S_{TAR})}{p(X|S_{UBM})} \begin{cases} > \emptyset, accept\ H_0 \\ < \emptyset, accept H_1 \end{cases} \quad (1)$$

Here $p(X|H_0)$ is the probability density function for the hypothesis H_0 evaluated for the observed speech segment X also known as the likelihood of hypothesis H_0 and similarly $p(X|H_1)$ for the likelihood of hypothesis H_1 . Also S_{TAR} and S_{UBM} represent model for the hypothesis target speaker and imposter (universal background model or anti model) respectively. The decision threshold for accepting or rejecting H_0 is $\emptyset$. Furthermore, the equation (1) can be expressed in log likelihood ratio as follows:

$$log\Lambda(X) = log\ p(X|S_{TAR}) - log\ p(X|S_{UBM}) \quad (2)$$

Here the observations of the segment X is statistically independent, therefore in we consider the input X speech segment as T short-time feature vectors then X can be expressed as $X=\{x_1, x_2, x_3, \ldots\ldots x_T\}$. Then we have the log likelihoods of the observed sequences to the hypothesis target speaker model and the imposter (UBM) model as follows:

$$log\ p(X|S_{TAR}) = \frac{1}{T}\sum_{t=1}^{T} log\, p(x_t|S_{TAR}) \quad (3)$$

$$log\ p(X|S_{UBM}) = \frac{1}{T}\sum_{t=1}^{T} log\, p(x_t|S_{UBM}) \quad (4)$$

III.I.1 FRONT-END PROCESSING AND FEATURE EXTRACTION

The frame size and frame rate is set to 20ms and 10ms respectively. Thirteen-dimensional Mel-frequency cepstral coefficients (MFCC) are extracted from silence removal with VAD as well as bandlimited data first. The channel effect is compensated by transforming the MFCCs with feature warping as well as CMS and CVN feature normalization techniques. The other 13 delta and 13 delta delta coefficients are calculated based on the warped MFCCs are appended to form a 39-dimensional spectral feature vectors. The zeroth cepstral coefficients (the DC level of the log-spectral energies) are not used in the feature vector. Another high level feature, a nine dimensional prosodic features vector consist of 1st, 2nd and 3rd formant frequencies (F1,F2 and F3), pitch , short time energy and its first and second order derivatives (Δpitch, Δenergy, $\Delta\Delta$pitch and $\Delta\Delta$energy) also added with MFCC features. So, now we have a total 48 dimensional feature vectors which makes it more robustness features to minimize noise, session-variability, language-variability affects.

III.I.2 FEATURE LEVEL NORMALIZATION

Normalizations at the stage of feature extraction are implemented to reduce the effect of the noise, speech signal distortion as well as the channel distortion. State-of-the –art speaker recognition system have used several approached in order to enhance the performance in feature level scores. In the log-spectral and cepstral domains, convolutive channel noise becomes additive [13]. The cepstral mean substraction (CMS) [14] is a blind deconvolution that comprises the substraction of the utterance mean of the cepstral coefficients from each feature which become zero-mean and the effect of the channel is reduced. In the similar way, the variance normalization (CVN) is also applied. Hence, the new features will fit a zero mean and variance one distribution. Another well- known feature normalization is RASTA (Relative Spectras).While CMS focus on the stationary convolution of the noise due to the channel, RASTA reduces the effect of the varying channel; which removes low and high modulation frequencies [15]. The three of them are the most commonly used feature normalization techniques in the SV system. In this experiment we have used both CMS and CVN feature normalization techniques.

III.I.3 GMM-UBM AS SPEAKER MODELING

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].

Normally, SV systems use mel-frequency cepstral coefficients (MFCCs) as a feature vector and the speaker model λ_s is parameterized by the set $\{w_i, \mu_i, \Sigma_i\}$ where w_i are the weights, μ_i are the mean vectors, and Σ_i are the covariance matrices. In the testing stage, feature vectors **X** are extracted from a test signal. A log-likelihood ratio $\Lambda(\mathbf{X})$ is computed by scoring the test feature vectors against the claimant model and the UBM.

$$\Lambda(\mathbf{X}) = log\ p(X|S_{TAR}) - log\ p(X|S_{UBM}) \quad (5)$$

The claimant speaker is accepted if $\Lambda(\mathbf{X}) \geq \theta$ or else rejected. The important problem in SV is to find a decision threshold θ for the decision making [16]. The uncertainty in θ is mainly due to score variability between the trials.

III.I.4 MODEL LEVEL NORMALIZATION

he purpose of score normalization is to alleviate the variability caused by numerous reasons, and currently, most normalization approaches are achieved by rescaling the impostor score distribution of each speaker to a normal distribution (zero mean and unit variance)[17]. In the model level normalization D-norm was one of the most popular

normalization technique that proposed by Ben et al. in 2002 [20] that mainly deals with the problem of pseudo-impostor data availability by generating the data using UBM model [19]. D-Norm doesn't need any additional speech data or external speaker population in addition to UBM model for its implementation in reality which is its advantage [17][21].

In D-Norm, A Monte Carlo-based approach is utilized to get a set of speaker and impostor data using speaker and UBM models respectively. The Normalized score is given by

$$\check{S}_{\lambda}(X) = \frac{S_{\lambda}(X)}{KL2(\lambda,\lambda')} \quad (6)$$

Here, $KL2(\lambda,\lambda')$ is the estimation of the summarized Kullback-Leibler distance between the speaker model (λ) and UBM models (λ'). And $S_{\lambda}(X)$ is the λ speaker score for the utterance X.

III.I.5 SCORE LEVEL NORMALIZATION

In score normalization, the final score of the SV system is normalized relative to a set of other speaker models termed as cohort. The application of score normalization techniques has become important in GMM based speaker verification system for reducing the effects of the lots of sources of statistical variability with log likelihood ratio scores [16]. Score normalization techniques have been mainly derived from the study of Li and Porter [18]. The main purpose of score normalization is to transform scores from different speakers into a similar range so that a common speaker independent verification threshold can be used [18]. As we know that in SV system the score variability comes from various sources. First, the probable mismatch between enrollment data which is used for training speaker models and the data that is used for testing is one of the main problems in SV system. Secondly, the nature and properties of the enrollment data can vary between the speakers, the phonetic content, the duration, the environmental noises as well as the quality of the speaker model training. Other two main factors intra-speaker and inter-speaker variability also affects in the performance in SV system. On the other hand some environment condition changes in transmission channel, recording devices or acoustical environment may also considered as a potential factor affecting the reliability of decision boundaries. To overcome above problems score normalization techniques have been introduced to cope with score variability and to make speaker-independent decision threshold tuning easier [19].

The basic of the normalization techniques is to center the imposter score distribution by applying on each score generated by the SV system. The general formula to compute score normalization for speech signal X and speaker model λ is given as follows.

$$\check{s}_{\lambda}(X) = \frac{S_{\lambda}(X) - \mu_{\lambda}}{\sigma_{\lambda}} \quad (7)$$

Where $\check{s}_{\lambda}(X)$ is the normalized score and $S_{\lambda}(X)$ is final score and μ_{λ} and σ_{λ} are normalized parameters known as estimated mean and standard deviation of the imposter score distribution. Imposter distribution represents the largest part of the score distribution variance.

There are different types of normalization techniques can be seen in speaker recognition system. These are Z-norm, H-norm, T-norm, HT-norm, C-norm etc. The zero normalization (Z-norm) had been more used in SV in the middle of nineties. The advantage of Z-norm is that the estimation of the normalized parameters can be performed offline during speaker model training [19].The handset normalization (H-norm) deals with handset or channel mismatch between training and testing. H-norm normalized parameters are estimated by testing each speaker model against handset dependent speech signals that produced by imposters [19].

The test-normalization (T-norm) can be performed online during testing. In T-norm, during testing, the incoming speech signal is classically compared with claimed speaker model as well as with a set of imposter models to estimate imposter score distribution and normalized parameters consecutively [19].

IV. BASELINE SYSTEM OF SPEAKER VERIFICATION 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 48-dimensional combined of acoustic and prosodic features vector was used in this experiment. The coefficients were extracted from a speech sampled at 16 KHz with 16 bits/sample resolution. A pre-emphasis filter $H(z)=1-0.97z^{-1}$ has been applied before framing. 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 24 triangular filters spaced on Mel-scale and constrained into a frequency band of 300-3400 Hz.

Cepstral Mean Subtraction (CMS) has been applied on all features to reduce the effect of channel mismatch. In this approach we also apply Cepstral Variance Normalization (CVN) which forces the feature vectors to follow a zero mean and a 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 pooling the both male and female

models of total 1024 Gaussian components and finding the average of all these models [7]. 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.

Here, we apply Z-Norm, T-Norm techniques as score normalization and D-norm as model level normalization technique to improve the performance of SV system because of its mismatching phonetic contents in training and testing environments of multilingual SV system. In this T-normalization technique normalized parameters mean and standard deviation are estimated from the imposter score distribution from the same database ALS-DB. In Z-Norm the impostor model also being constructed from the same database. The detection error trade-off (DTE) curve has been plotted using log likelihood ratio between the claimed model and the UBM and the equal error rate (EER) obtained from the DTE curve has been used as a measure for the performance of the speaker verification system. Another measurement MinDCF values has also been evaluated.

V. 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 (VAD) is used to identify and discard the silence frames prior to feature extraction. Only data from the headset microphone (Device 2) has been considered in the present study. All the four available sessions were considered for the experiments. Each speaker model was trained using data from first two sessions. The test sequences were extracted from the next two sessions of Device 2 with language mismatching condition (not same with training language). The training set consists of speech data of length 120 seconds per speaker of total 150 speakers of 90 males and 60 female of the same linguistic group. 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 [9].

The resultant performances of the baseline system for the multilingual speaker verification system has been given in the figure 1, figure 2, figure 3 and figure 4 in the below for both language matching and mismatching conditions.

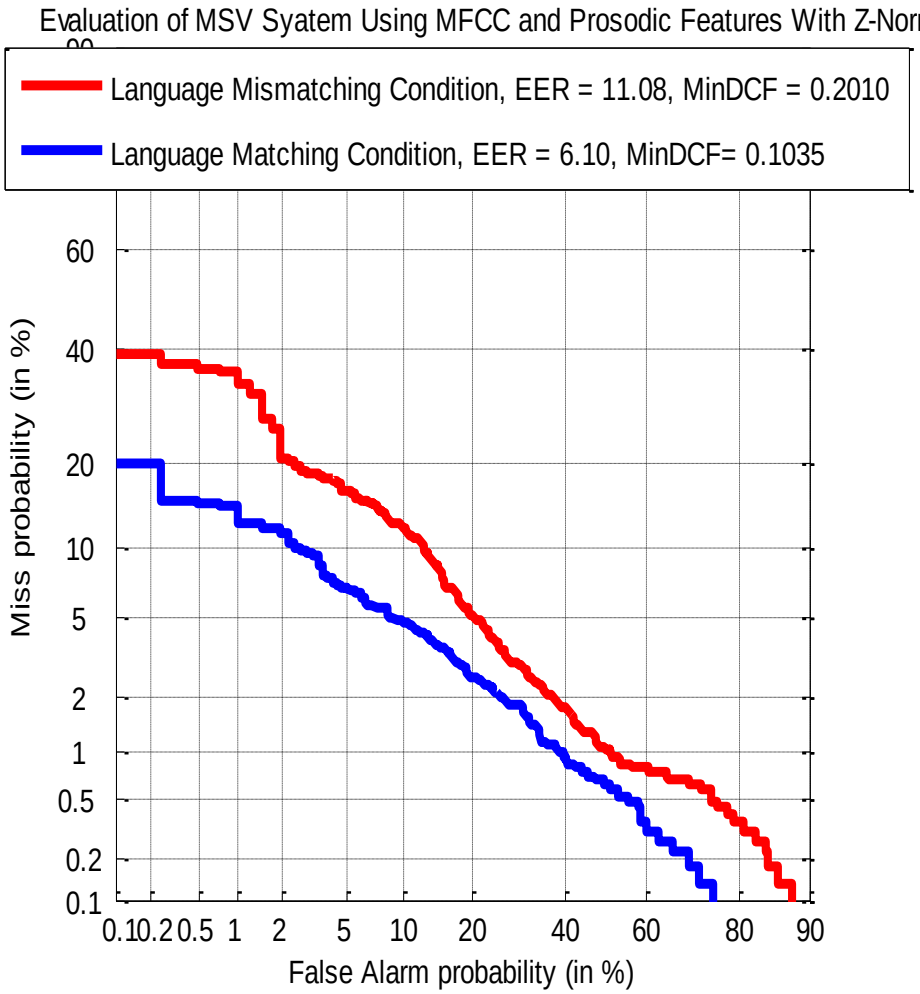

Fig.1: DET curves for Multilingual Speaker Verification System with Z-Norm for both Language matching and mismatching conditions.

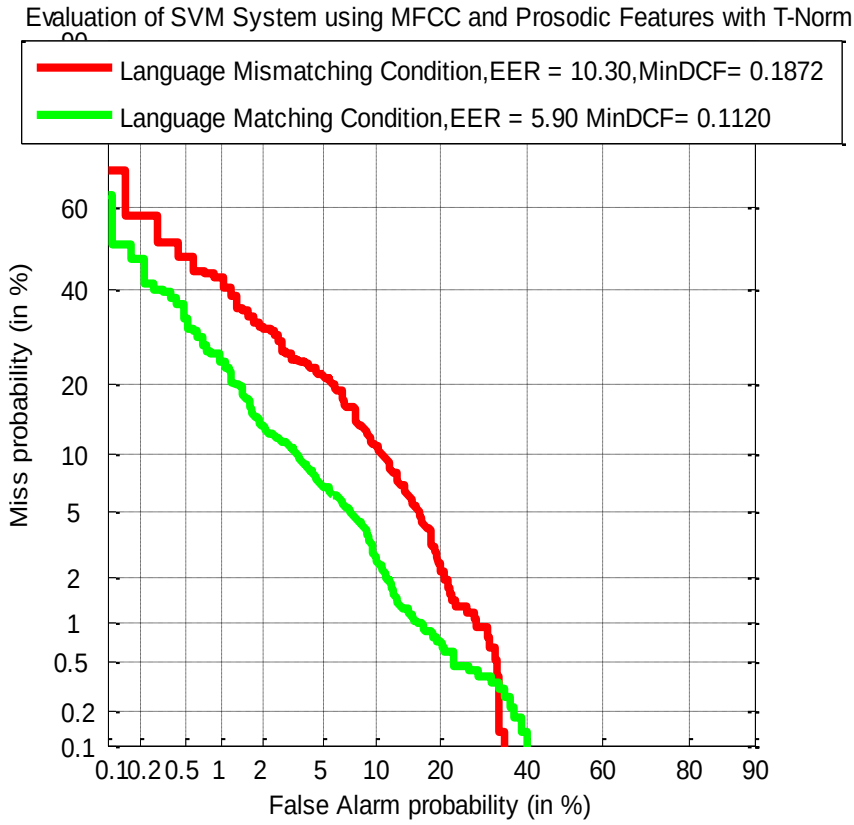

Fig.2: DET curves for Multilingual Speaker Verification System with T-Norm for both Language matching and mismatching conditions.

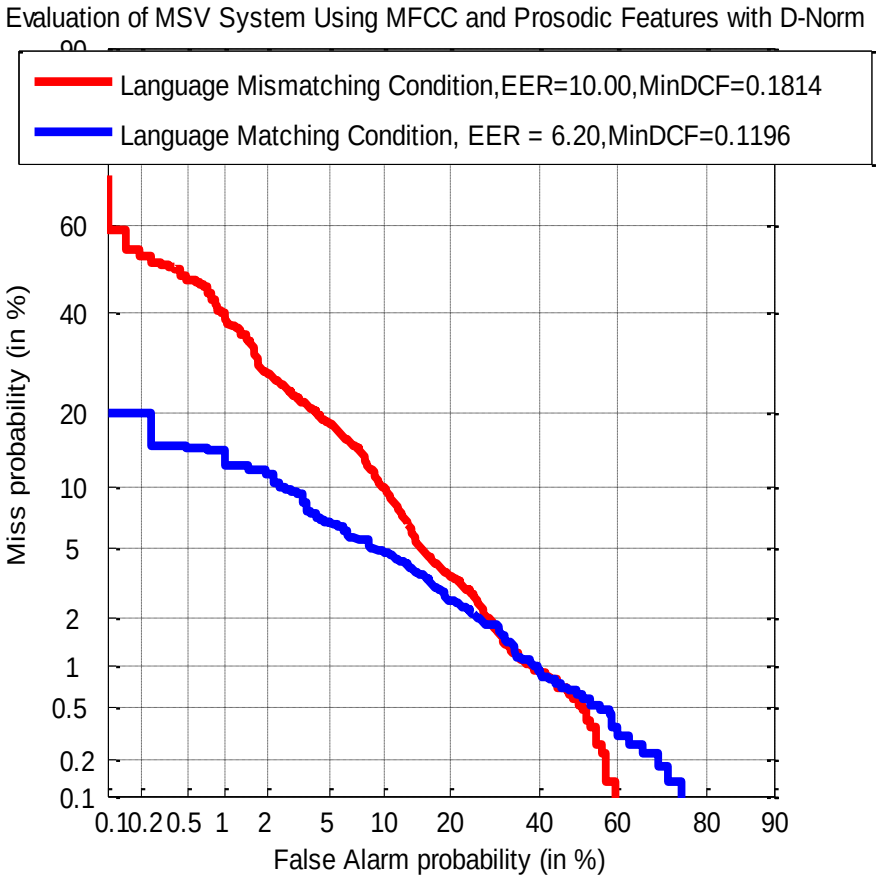

Fig.3: DET curves for Multilingual Speaker Verification System with D-Norm for both Language matching and mismatching conditions.

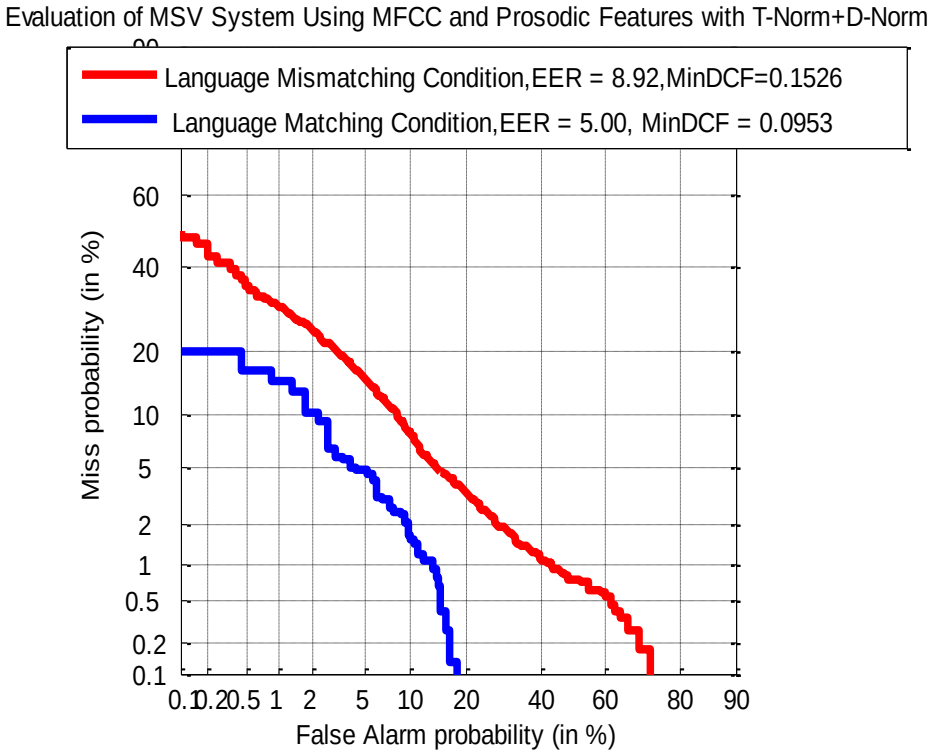

Fig. 4: DET curves for Multilingual Speaker Verification System with D-Norm for both Language matching and mismatching conditions.

Table 1 shows the performance of MSV system in terms of EER values as well as Minimum DCF values.

Table 1: The EER and MinDCF values of Multilingual Speaker Verification System.

Baseline System	**Language Conditions**	**EER%**	**MinDCF**
GMM-UBM + Z-Norm	Matching	6.10	0.1035
	Mismatching	11.08	0.2010
GMM-UBM + T-Norm	Matching	5.90	0.1120
	Mismatching	10.30	0.1872
GMM-UBM + D-Norm	Matching	6.20	0.1196
	Mismatching	10.00	0.1814
GMM-UBM + T-Norm + D-Norm	Matching	5.00	0.0953
	Mismatching	8.92	0.1526

VI. CONCLUSIONS

From the experimental point of view it has been observed that the performance of the multilingual speaker verification system has been improved by applying CMS and CVN in feature level and Z-Norm and T-Norm in score level as well as finally D-Norm in model level normalization. In this case the performance of the MSV system has been evaluated in terms of Equal Error Rates (EER). The performance of the baseline system has been found in EER values of 11.08%,10.30%,10.00%, and 8.90% for the GMM-UBM+Z-Norm, GMM-UBM+ T-Norm, GMM-UBM+D-Norm and GMM-UBM + T-Norm + D_Norm respectively for language mismatching conditions. It has been observed that for the language mismatching condition D-Norm shows better performance than T-Norm and Z-Norm. Combining T-Norm and D-Norm the performance of MSV system improved by approximately 2.00% of its recognition rate. Similarly, for language matching conditions T-Norm shows better performance than that of Z-Norm and D-Norm. The performance of the MSV system enhanced up to 95.00 % accuracy of recognition rate, while applying the combined T-Norm with D-Norm.

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]. J.P. Campbell, Jr, "Speaker recognition: a tutorial", *Proceedings of the IEEE*, 85(9) 1997, Vol.85, pp. 1437–1462,.

[2]. D.A. Reynolds,"Automatic Speaker Recognition: Current Approaches and Future Trends", MIT Lincoln Laboratory, 244 wood St. Lexinton, MA 02140,USA, 2002.
[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]. Ville Hautamaki, Tomi Kinnunen ,Ismo Karkkainen, ,Saastamoinen, Juhani , Tuononen Marko & Pasi Franti, "Maximum a Posteriori Adaptation of the Centroid Model for Speaker Verification," *IEEE signal Processing letters*, vol.15. 2008.
[5]. D.A.Reynolds, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech Communications*, vol. 17, pp. 91-108, 1995.
[6]. W.Campbell, D.Sturim, and D. A. Reynolds," Support vector machines using GMM supervectors for speaker verification", *IEEE Signal Processing Letters* 13,5 pp. 308-311,2006.
[7]. D. A. Reynolds and D.E. Sturim, " Speaker Adaptive Cohort Selection for Tnorm in text-independent speaker verification." MIT Lincoln Laboratory, Lexington, MA USA.
[8]. D.A Reynolds," Comparison of Background Normalization Methods for Text-Independent Speaker Verification." *In Proceeding of EUROSPEECH* '1997, Rhodes, Greece, pp. 963-966.
[9]. NIST2003 Evaluation plan, http://www.itl.nist.gov/iad/mig/tests/sre/2003/2003-spkrec-evalplan-v2.2.
[10]. Utpal Bhattacharjee and Kshirod Sarmah, "A Multilingual Speech Database for Speaker Recognition", *Proc. IEEE, ISPCC*, March 2012.
[11]. Utpal Bhattacharjee and Kshirod 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.
[12]. Xiaojia.Z, , S. Yang., and W, De Liang,"Robust speaker identification using a CASA front-end", *Acoustics, Speech and Signal Processing (ICASSP)*, 2011 *IEEE International Conference* on, pp.5468-5471, 2011.
[13]. Tomi Kinnunen and Haizhou Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors," *Speech Communication* 52(1): pp.12-40, 2010.
[14]. S.Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Transactions on Acoustic, Speech and Signal Processing* 29,2 pp. 254-272, April 1981.
[15]. P.G.Perera, Lopez Leibny, A.Roberto and J.N.Flores," Speaker Verification in Different Database Scenarios," Computation y Sistemas Vol.15 No.1, pp 17-26, 2011.
[16]. R. Auckenthaler, M.Carey, and H.Lloyd-Thomas, "Score normalization for test-independent speaker verification system," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.
[17]. Dong.Yuan, LU.Liang , ZHAO Xian-Yu, and ZHAO Jian, " Studies on Model Distance Normalization Approach in Text-independent Speaker Verification", *ACTA AUTOMATICA SINICA*, vol. 35, No.5, 2009.
[18]. K. P.Li , and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '88)*, 1988. vol. 1, pp. 595–598, New York, NY, USA.
[19]. F. Bimbot, et al, "A tutorial on text-independent speaker verification", *EURASIP Journal on Applied Signal Processing*, 4, 430–451, 2004.
[20]. M.Ben, R. Blouet, and F. Bimbot, "A Monte-Carlo method for score normalization in Automatic speaker verification using Kullback-Leibler distances." *Proceedings of IEEE ICASSP '02*, 2002 vol. 1, pp. 689–692,.
[21]. Utpal Bhattacharjee and Kshirod Sarmah, "Speaker Modeling Distance Normalization technique in Multilingual Speaker Verification", *International Journal of Electrical and Electronics Engineering Research (IJEEER)*,Vol. 3, Issue 2,pp. 319-326, June 2013.
[22]. Utpal Bhattacharjee and Kshirod Sarmah, "GMM-UBM Based Speaker Verification in Multilingual Environments", *International Journal of Computer Science Issues (IJCSI)*, Vol. 9, Issue 6, No 2, pp. 373-380, Nov.2012.
[23]. D.A. Reynolds,"Channel robust speaker verification via feature mapping", *in Proceedings of the International Conference on Acoustic Speech and Signal Processing*, 2003.vol.2, pp. 53-56.

