2016

Performance comparison of speaker recognition systems in presence of duration variability

Poddar, A

Institute of Electrical and Electronics Engineers (IEEE)

conferenceObject
info:eu-repo/semantics/acceptedVersion

http://dx.doi.org/10.1109/INDICON.2015.7443464

Performance Comparison of Speaker Recognition Systems in Presence of Duration Variability

Arnab Poddar*, Md Sahidullah†, Goutam Saha‡
*‡ Dept of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India
†Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Joensuu, Finland
Email: *arnabpoddar@iitkgp.ac.in, †sahid@cs.uef.fi, ‡gsaha@ece.iitkgp.ernet.in

***Abstract*—Performance of speaker recognition system is highly dependent on the amount of speech data used in training and testing. In this paper, we compare the performance of two different speaker recognition systems in presence of utterance duration variability. The first system is based on state-of-the-art total variability (also known as i-vector system), whereas the other one is classical speaker recognition system based on Gaussian mixture model with universal background model (GMM-UBM). We have conducted extensive experiments for different cases of length mismatch on two NIST corpora: NIST SRE 2008 and NIST SRE 2010. Our study reveals that the relative improvement of total variability based system gradually drops with the reduction in test utterance length. We also observe that if the speakers are enrolled with sufficient amount of training data, GMM-UBM system outperforms i-vector system for very short test utterances.**

Keywords*—*Duration Variability, Gaussian Mixture Model-Universal Background Model (GMM-UBM), Gaussian PLDA (GPLDA), i-vector, NIST SRE, Short Utterance, Speaker Recognition.

I. Introduction

Speech signal conveys information regarding the physiological aspects of a speaker because it is affected by the unique shape and size of vocal tract, mouth, nasal cavity, etc [1], [2]. It also carries information related to the behavioral aspects of a speaker like accent and involuntary transforms of acoustic parameters. Therefore, voice samples can be used as a biometric in real-life application. Speaker recognition is the process of automatically recognizing the speakers from their voice samples. Its potential applications include telephone banking system, system access control, providing forensic evidence, call centers and many more [1], [2]. Speaker recognition task can be sub-divided into two major tasks: speaker identification (SI) [3] and speaker verification (SV). Speaker identification is to find the identity of the speaker from a speech utterance. On the other hand, speaker verification refers to the authentication of a claimed identity of a person from his/her speech data. SV system can be broadly categorized as text-dependent (TD) [4] and text-independent (TI) modes depending on the speech content in training and test phase [1], [2]. The TD-SV requires the same set of text to be spoken during training as well as testing. In the case of TI-SV, it does not have any restriction over train and test data.

A TI speaker recognition system includes three fundamental modules [1], [2]: a feature extraction unit, which represents the speech signal in a compact manner, a modeling block to characterize those features using statistical approaches, and lastly, a classification scheme to classify the unknown utterance. Mel frequency cepstral coefficients (MFCCs), perceptual linear prediction (PLP), etc. are commonly used as speech features for speaker recognition [5], [6]. For classification, various modeling techniques such as vector quantization (VQ) [7], dynamic time warping (DTW) [8], Gaussian mixture model (GMM) [9] were used. During the last two decades in speaker recognition research, most of the notable developments in classifier-level are based on the GMM concept [10], [11], [12]. It also found applications in various field like speech language recognition [13], voice conversion [14], detection of spoofing attacks [15] in SR systems etc. Subsequently, joint factor analysis (JFA) based approach is introduced which successfully integrates session variability compensation techniques [16], [17]. Here, the concatenated means of adapted GMM (known as GMM supervector) are decomposed into speaker and session dependent component using factor analysis technique. Speaker factors are compared for training and test segments after subtracting the session related factors. Inspired by the earlier use of JFA, Dehak *et al.* proposed *total-variability* based approach for reducing the dimensionality of GMM-supervector [18]. Here, unlike JFA, a single space called *total variability* space is used to represent the GMM supervector corresponding to a speech utterance. This low-dimensional representation of high-dimensional supervector is known as identity vectors or i-vectors. The state-of-the-art speaker recognition system uses i-vector based system with Gaussian probabilistic linear discriminant analysis (GPLDA) based scoring where the i-vectors are further decomposed into speaker and channel subspace to efficiently handle intersession variability [19], [20]. Though i-vector based speaker recognition systems are shown to give best recognition accuracy in latest NIST SREs [18], [19], [21], they require huge computational resources as well as massive amount of development data for estimating its parameters and hyper-parameters. For this reason, GMM-UBM systems are still popular and widely used, particularly when suitable amount development data is inadequate [10], [5], [14].

The performance of a speaker recognition system is severely degraded with the reduction in amount of speech data during train and test phase [21], [22], [23], [24], [25]. State-of-the-art i-vector system gives considerable recognition accuracy when more than two minutes of speech data are available in both phases [19], [21]. But practically, real-time systems may not facilitate the luxury on amount of speech data. When designing a practical speech-based authentication system, the requirement in training segment duration can be fulfilled by enrolling the speaker with adequate amount of speech data. However, this is impractical to maintain during the verification

978–1–4673–6540–6/15/$31.00

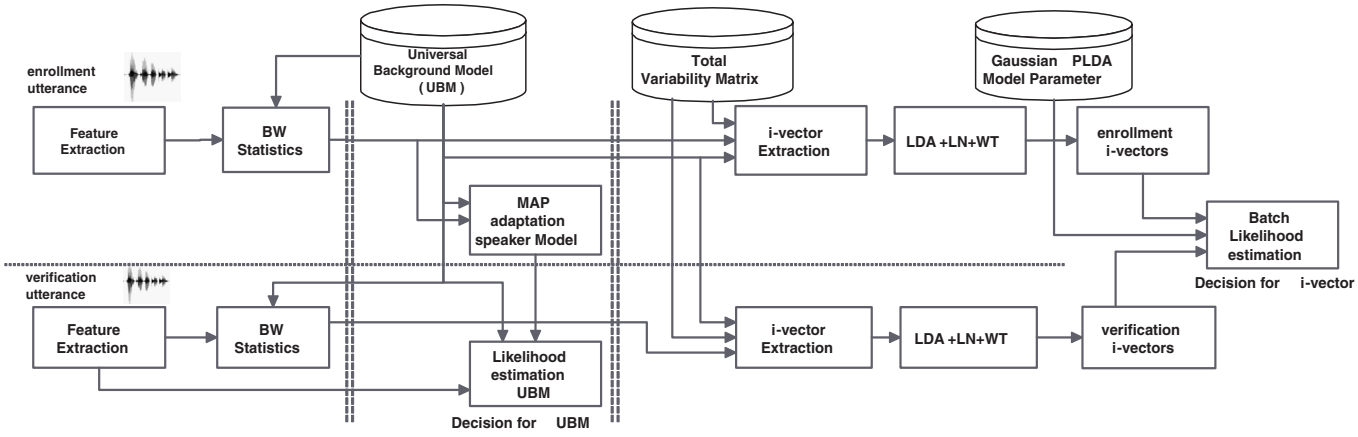

Figure 1: Block Diagram showing how the scores of GMM-UBM and i-vector system are calculated from enrollment and verification utterance.

phase. The test speech duration should be as low as possible so that decision regarding acceptance or rejection can be made in real-time. Another issue associated with this state-of-the-art i-vector system is that it requires various computationally expensive processes for getting the recognition results [26]. On the other hand, the GMM-UBM system gives the decision in relatively short time by just computing the likelihood ratio directly using the cepstral feature. GMM-UBM systems perform well for short test segments [27], [28]. To the best of our knowledge, a systematic comparison of these two systems including the effect of duration variability is not available in present literatures even though it has practical significance. In this work, we explore the impact of duration variability on both GMM-UBM and TV systems using the same benchmark data and performance evaluation metrics. Exhaustive experiments are carried out by varying both the length of training and test utterance. Our experimental results reveal that though TV system is performing better than GMM-UBM in many conditions, but the classical approach is still better than the state-of-the-art technique for condition very similar to practical requirements i.e. when speakers are enrolled with sufficient amount of speech data and tested with short segments.

The rest of the paper is organized as follows. Section II briefly describes GMM-UBM based system. In section III, we have discussed about i-vector GPLDA system. In section IV, the set-up arranged to conduct the experiments is described. Experimental results are presented in section V. Finally, the paper is concluded in VI.

II. GMM-UBM BASED SV SYSTEM

The task of a typical SV system is to discriminate between target and imposter speakers based on two hypothesis, i.e, whether the verification utterance belongs to the target speaker or not. The block diagram of a typical SV system is shown in Fig 1 which shows both TV and GMM-UBM framework. In GMM-UBM, prior to enrollment phase, a single speaker independent universal background model (UBM) is created by using a large development data [10], [14]. The UBM represented as $\boldsymbol{\lambda}_{UBM} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{C}$ where C is the total number of Gaussian mixture components, w_i is the weight or prior of i^{th} mixture component, $\boldsymbol{\mu}_i$ is the mean and co-variance matrix is given by $\boldsymbol{\Sigma}_i$. Parameter w_i satisfies the constrain $\sum_{i=1}^{C} w_i = 1$.

A group of S speakers is represented by their corresponding model as $\{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_S\}$. In the GMM-UBM system, we derive the target speaker model by adapting the GMM-UBM parameters. The model parameters are adapted by maximum a posteriori (MAP) method. Initially, sufficient statistics N_i and $\mathbf{E}_i$ from a hypothesised speaker's utterance with T frames $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, are calculated as,

$$N_i = \sum_{t=1}^{T} Pr(i|\mathbf{x}_t) \text{ and } \mathbf{E}_i(\mathbf{X}) = \frac{1}{N_i} \sum_{t=1}^{T} Pr(i|\mathbf{x}_t)\mathbf{x}_t$$

where probability distribution of component density conditioned on speech data $Pr(i|\mathbf{x}_t)$ is given by

$$Pr(i|\mathbf{x}_t) = \frac{w_i p_i(\mathbf{x}_t)}{\sum_{j=1}^{C} w_j p_j(\mathbf{x}_t)} \tag{1}$$

Each component density is a d-variate Gaussian function of the form

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\} \tag{2}$$

Finally, the sufficient statistics from training data are used to adapt GMM-UBM parameters to obtain adapted parameters $\hat{w}_i$, $\hat{\boldsymbol{\mu}}_i$ for target speakers.

In the testing phase, average log-likelihood ratio $\Lambda(\mathbf{X})$ is determined using test feature vector $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ against both target model and the background model.

$$\Lambda_{UBM}(\mathbf{X}^{test}) = log\ p(\mathbf{X}^{test}|\boldsymbol{\lambda}_{target}) - log\ p(\mathbf{X}^{test}|\boldsymbol{\lambda}_{UBM}) \tag{3}$$

where $log\ p(\mathbf{X}^{test}|\boldsymbol{\lambda}) = \frac{1}{T}\sum_{t=1}^{T} log\ p(\mathbf{x}_t^{test}|\boldsymbol{\lambda})$ Finally, in decision logic block, an algorithm is applied to decide whether the claimant speaker will be accepted or rejected by the SV system. Popularly a decision threshold θ is used for decision, like if $\Lambda_{UBM}(\mathbf{X}) \geq \theta$ then the claim will be accepted, else rejected.

III. I-VECTOR BASED SV SYSTEM

i-vector is considered as the state-of-the-art in SV research. The basic block diagram of i-vector based SV system system is shown in Fig. 1. The i-vector represents the GMM supervector by a single variability space which reduces high dimensional GMM supervector into lower dimensional total variability space [18]. In TV space, GMM supervector, i.e, the concatenated means of GMM mixture components, is rewritten as

$$\mathbf{M} = \mathbf{m} + \mathbf{\Phi}\mathbf{y} \tag{4}$$

where $\mathbf{\Phi}$ is a low-rank total variability matrix and $\mathbf{y}$ is represented as i-vector, $\mathbf{m}$ is the speaker and channel independent supervector (taken to be UBM supervector) and $\boldsymbol{M}$ is the speaker-and channel-dependent GMM supervector.

A UBM model consisting of C Gaussian components can be represented by the parameter set $\boldsymbol{\nu} = \{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots, \boldsymbol{\nu}_C\}$, where i^{th} mixture component is characterised by $\boldsymbol{\nu}_i = \{w_i, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\}$. Then, for single utterance $\mathbf{X}$ with feature sequence $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, the zeroth and first order centered sufficient statistics N_i and $\mathbf{F}_i$ respectively are calculated as follows $N_i = \sum_{t=1}^{T} Pr(i|\mathbf{x}_t)$ $\mathbf{F}_i = \frac{1}{N_i}\sum_{t=1}^{T} Pr(i|\mathbf{x}_t, \boldsymbol{\lambda}_i)(\mathbf{x}_t - \boldsymbol{\mu}_i)$. These N_i and $\mathbf{F}_i$ are used to obtain the i-vectors $\mathbf{y}$.

The prior distribution of i-vectors $p(\mathbf{y})$, is assumed to be $\mathcal{N}(0, I)$ and posterior distribution of $\mathbf{F}$, conditioned on the i-vector $\mathbf{y}$ is hypothesised to be $p(\mathbf{F}|\mathbf{y}) = \mathcal{N}(\mathbf{\Phi}\mathbf{y}, \mathbf{N}^{-1}\mathbf{\Sigma})$. The MAP estimate of $\mathbf{y}$ conditioned on $\mathbf{F}$ is given by

$$E(\mathbf{y}|\mathbf{F}) = (\mathbf{I} + \mathbf{\Phi}^{\top}\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{\Phi})^{-\mathbf{1}}\mathbf{\Phi}^{\top}\mathbf{\Sigma}^{-1}\mathbf{N}\mathbf{F} \tag{5}$$

the mean of the posterior distribution of $\mathbf{y}$ conditioned on $\mathbf{F}$ is adopted as the i-vector of an utterance.

A. Gaussian Probabilistic Linear Discriminate Analysis (GPLDA)

A recent attempt to model speaker and channel variability in i-vector space is accomplished through Probabilistic LDA (PLDA) modelling approach. In this paper, we concentrate on a simplified variant of PLDA, named as Gaussian PLDA [19]. Here, the inter-speaker variability is modelled by a full covariance residual term. The generative model for s^{th} speaker and j^{th} recording of new i-vector variability projected space is given by

$$\mathbf{y}_{s,j} = \boldsymbol{\eta} + \mathbf{\Psi}\mathbf{z}_s + \boldsymbol{\epsilon}_{s,j} \tag{6}$$

where, $\boldsymbol{\eta}$ is the mean of the development i-vectors, $\mathbf{\Psi}$ is eigen-voice subspace and $\mathbf{z}$ is a vectors of latent factors, which is assumed to have prior distribution $\mathcal{N}(0, \mathbf{I})$. The residual term $\boldsymbol{\epsilon}$ represents the variability not captured by the latent variables. This regenerative model approach of i-vector space representation has been applied successfully with significant improvement in speaker recognition research [19].

B. Likelihood Computation

GPLDA based i-vector system score calculation uses batch likelihood ratio [19]. For a projected enrollment and verification i-vector $\mathbf{z}_{target}$ and $\mathbf{z}_{test}$ respectively, the batch likelihood ratio $\Lambda_{GPLDA}(\mathbf{z}_{target}, \mathbf{z}_{test})$ can be calculated as follows

Table I: Database set-up description of NIST 2008 and NIST 2010.

Database	Verification Task	No of speaker model	No of test trials
NIST 2008	short2-10sec	1270	3958
NIST 2008	short2-short3	1270	6615
NIST 2010	core-10sec	1203	11990
NIST 2010	core-core	1203	14060

$$\Lambda_{GPLDA}(\mathbf{z}_{target}, \mathbf{z}_{test}) = log\ \frac{p(\mathbf{z}_{target}, \mathbf{z}_{test}|H_1)}{p(\mathbf{z}_{target}|H_0)\ p(\mathbf{z}_{test}|H_0)} \tag{7}$$

where H_1: The i-vectors belong to the same speaker.
H_0: The i-vectors belong to different speaker.

IV. EXPERIMENTAL SET-UP

Both GMM-UBM and i-vector based systems use mel frequency cepstral coefficient (MFCC) with 20 ms frame size and 10 ms frame shift as in [5]. Hamming window is applied in MFCC extraction process [29]. The non-speech frames are dropped using energy based voice activity detector (VAD) [30] and at the end cepstral mean and variance normalisation (CMVN) is applied on coefficients [5]. 19 dimensional MFCC with appended delta and double delta coefficients (57 dimensional) are used throughout the experiments. Gender dependent UBM of 512 mixture components are trained with 10 iterations of EM algorithm. We have used NIST 2004 and NIST 2005 corpora as development data to generate UBM, GPLDA and LDA model parameters. Total variability subspace of dimension 400 is implemented for i-vector. LDA on i-vector space is used to reduce the dimension to 200, and speaker variability subspace i.e, eigen-voice space is applied to further reduce the i-vector dimension to 150.

A. Experiments and Corpora

The performance of two popular speaker modelling methods were evaluated on NIST SRE 2008 [31] and NIST SRE 2010 [32] corpora. We have used NIST 2008 *short2-short-3*, *short2-10 sec* and NIST 2010 *core-core*, *core-10 sec* speaker recognition task for evaluation. Later, we have used utterances which are utterance truncated versions of NIST 2008 Short2-Short3 task for experiments in varying utterance duration condition. Truncation of speech utterances is done in 2 sec (200 active frames), 5 sec (500 active frames), 10 sec (1000 active frames), 20 sec (2000 active frames) and 40 sec (4000 active frames) duration. For truncation of utterances, the prior 500 active speech frames are discarded to avoid phoneme dependency which refers to, capturing phonetically similar data. Only male speakers of *english* trials from NIST 2008 and *telephone-telephone* trials of from NIST SRE 2010 are used in the following experiments. The experiments are preformed on the male data subset of the corpora, taken from both NIST databases. The database description is summarised in Table I.

B. Performance metrics

For both NIST 2008 and NIST 2010, the performance was evaluated using *equal error rate (EER)* and *detection cost*

Table II: Results for comparison of i-vector-GPLDA(TV) based system vs GMM-UBM based system on 2008 NIST SRE short2-short3, short2-10sec conditions.

Verification Task	Training duration	Testing duration	EER [%] (TV)	EER [%] (UBM)	RI_{TV}^{EER}[%]	DCF × 100 (TV)	DCF × 100 (UBM)	RI_{TV}^{DCF}[%]
short2-short3	Full	Full	**3.48**	12.30	72	**2.16**	5.28	59.04
short2-10 sec	Full	10 sec	**11.49**	18.46	38	**4.62**	6.30	26.67
short2-10 sec	10 sec	10 sec	**16.81**	20.76	19	**6.45**	8.05	19.88

Table III: Results for comparison of i-vector-GPLDA(TV) based system vs GMM-UBM based system on 2010 NIST SRE core-core, core-10sec conditions.

Verification Task	Training duration	Testing duration	EER [%] (TV)	EER [%] (UBM)	RI_{TV}^{EER}[%]	DCF × 100 (TV)	DCF × 100 (UBM)	RI_{TV}^{DCF}[%]
core-core	Full	Full	**4.53**	14.16	68	**2.33**	4.88	52.25
core-10 sec	Full	10 sec	**11.41**	18.64	39	**5.60**	6.84	18.13
core-10 sec	10 sec	10 sec	**19.54**	23.82	18	**7.85**	8.39	6.44

function (DCF). EER is the point on detection error tradeoff (DET) plot, where probability of false acceptance and probability of false rejection are equal. The DCF is computed by creating a cost function assigning some unequal weight on false alarm and false rejection followed by computation of threshold where cost function is minimum. The cost function is computed as

$$C_{Det} = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1 - P_{Target}), \quad (8)$$

DCF is calculated using the parameter value $C_{Miss} = 10$, $C_{FalseAlarm} = 1$ and $P_{target} = 0.01$ for both databases NIST 2008 and NIST 2010 [31], [32]. A measurement of relative improvement of EER and DCF rate of i-vector system over GMM-UBM is calculated as

$$RI_{TV}^{EER} = \frac{(EER_{TV} - EER_{UBM})}{EER_{UBM}} \times 100\% \quad (9)$$

$$RI_{TV}^{DCF} = \frac{(DCF_{TV} - DCF_{UBM})}{DCF_{UBM}} \times 100\% \quad (10)$$

V. Results and Discussion

We have chosen NIST 2008 [31] and NIST 2010 [32] database to conduct the experiments as they have wide intersession and channel variability. In addition to that, they also provide a large number of speaker trials as described in Table I. We have used two different databases to show the consistency of the indications emerging from the results over databases. Table II and Table III show the results on NIST SRE 2008 and NIST SRE 2010 respectively, depicting the behavior and comparison of two systems for the different utterance duration. For the third case in Table II and Table III, training utterance is truncated to 1000 frames. The first 500 voiced frames are dropped to avoid phonetic similarity and also to ensure text independence of speech segments. This procedure is maintained for truncation in other experiments as well, of which the results are given in Table IV and V. The general trend shown by Table II and Table III is, as the utterance length decreases, significant degradation of performance occurs in both i-vector and GMM-UBM system. The Fig. 2(a) and Fig. 4(a) show that the performance of both system degrades for reduction in duration of utterance. We also observe that the realtive improvement of TV system over GMM-UBM system degrades for shorter test segments. As these trends of results are found in 4 different SV system evaluation plan, it establishes consistency of trends of results over databases.

The results of the experiments reported in Table IV and Table V exhibits a deeper comparative study. In Table IV and V, it depicts that the relative improvement RI_{TV}^{EER} and RI_{TV}^{DCF} decreases monotonically with the reduction in utterance duration. If we go on increase the no of active frames in utterance irrespective of training and testing, the relative performance improvement of i-vector based system exhibited higher rate of over GMM-UBM system. In Fig. 2(a) and Fig. 4(a), the red lines representing i-vector performance, showed steeper curves with respect to the blue lines representing GMM-UBM system. This indicates the fact that the relative improvement of i-vector system over GMM-UBM system increases with utterance length. Figure 2(b), Fig. 4(b), Table IV and Table V supports this observation for both *Full-duration training - truncated duration testing* and *truncated duration training - truncated duration testing* condition. In Fig. 3 and Fig. 5 detection error tradeoff (DET) plots of i-vector and UBM based system are shown, which shows a deeper comparative performance study on decision threshold. DET curves are given for both *Full-duration training - truncated duration testing* and *truncated duration training - truncated duration testing* in Fig. 3 and Fig. 5 respectively.

The overall result shows some relevant observations. The results from all the tables shows that i-vector based system worked significantly better than GMM-UBM for longer utterances. The results from Table II show upto 72% relative improvement of i-vector based system over GMM-UBM based system. For real-time application, SV system with very short duration with minimum complexity is desired. Both the systems' performance fall on durations as small as 2 sec, 5 sec etc. From Table IV and Table V, we observe that in case of very short duration utterances specially in *Full duration training-2 sec testing* and *2 sec training-2 sec testing*, GMM-UBM based system showed better performance over i-vector based system.

Table IV: NIST 2008 short 2 short 3 Results on truncated Training and Testing.

Training duration	Testing duration	EER [%] (TV)	EER [%] (UBM)	RI_{TV}^{EER}[%]	DCF × 100 (TV)	DCF × 100 (UBM)	RI_{TV}^{DCF}[%]
2 sec	2 sec	**35.37**	37.04	4.51	9.81	**9.59**	-2.24
5 sec	5 sec	**23.23**	26.19	11.3	8.65	**8.57**	-0.93
10 sec	10 sec	**13.47**	16.62	18.95	**6.35**	7.00	9.29
20 sec	20 sec	**7.51**	13.43	44.04	**4.11**	6.21	33.82
40 sec	40 sec	**4.92**	12.04	59.14	**2.85**	5.69	49.91
Full	Full	**3.48**	12.30	71.71	**2.16**	5.28	59.09

Table V: NIST 2008 short-2 short 3 Results on Full length Training and truncated Testing.

Training duration	Testing duration	EER [%] (TV)	EER [%] (UBM)	RI_{TV}^{EER}[%]	DCF × 100 (TV)	DCF × 100 (UBM)	RI_{TV}^{DCF}[%]
Full	2 sec	22.09	**20.50**	-7.76	7.79	**7.57**	-2.91
Full	5 sec	**11.61**	15.44	24.81	**5.43**	6.22	12.70
Full	10 sec	**8.33**	14.35	41.95	**4.14**	5.81	28.74
Full	20 sec	**6.21**	13.43	53.76	**3.22**	5.55	41.90
Full	40 sec	**4.55**	12.92	64.78	**2.73**	5.52	50.54
Full	Full	**3.48**	12.30	71.71	**2.16**	5.28	59.09

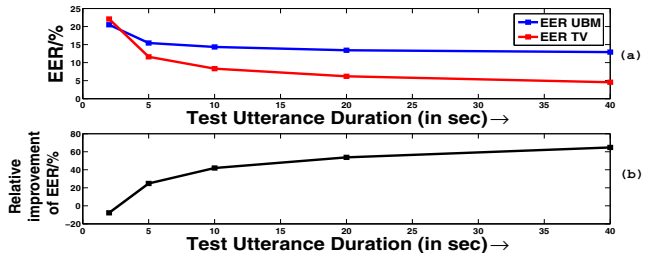

Figure 2: (a) Plot of EER of i-vector system and UBM system. (b) Relative improvement of EER for full length training-truncated testing condition in NIST 2008, short2 short3 corpora.

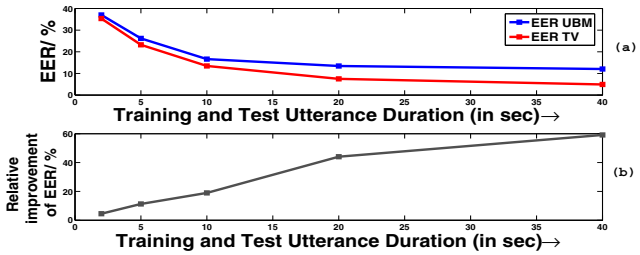

Figure 4: (a) Plot of EER of i-vector system and UBM system. (b) Relative improvement of EER for truncated training-truncated testing condition in NIST 2008, short2 short3 corpora.

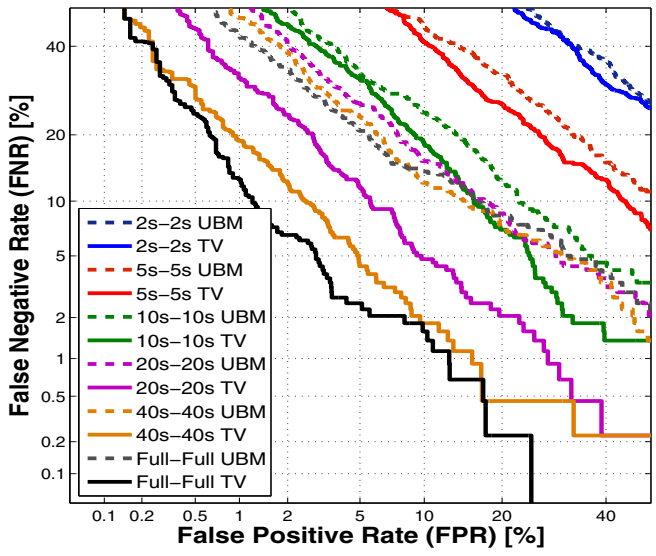

Figure 3: DET plot of i-vector (TV) system and GMM-UBM system for Full utterance duration training-truncated testing condition in NIST 2008, short2 short3 corpora.

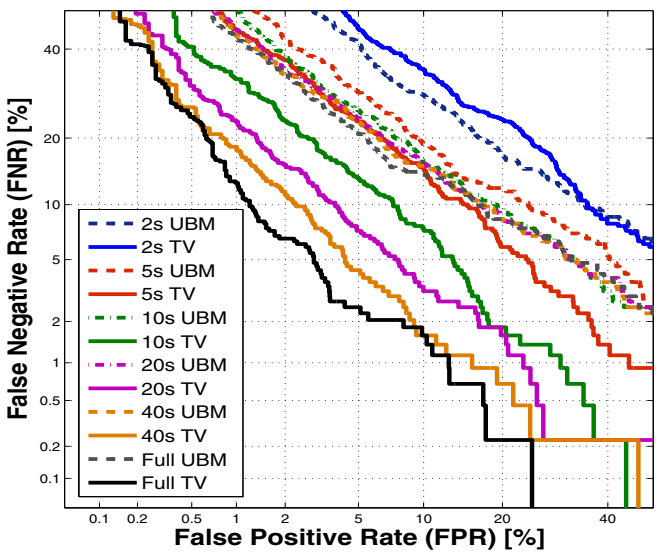

Figure 5: DET plot of i-vector (TV) system and UBM system for truncated training - truncated testing condition in NIST 2008, short2 short3 corpora.

VI. Conclusion

The primary concern of state-of-the-art SV system is the modelling part of it. A critical comparison of factor analysis based state-of-the-art modelling method and a background model based straight-forward modelling method is presented in this work. The present study gives an indication of merits and demerits of i-vector based and GMM-UBM based system for different train-test condition. It is found that modelling speakers in total variability subspace framework exhibits a significant relative performance improvement upto 72% on long duration utterances. But, small utterance duration is desirable for a real-time SV system. Both GMM-UBM and i-vector based systems degrade severely when test utterance length falls below 10 sec. This characteristics limits the utility of SV system in real life scenario. The relative improvement measures of i-vector based system over GMM-UBM based system are also found to get reduced significantly in utterance length below 10 sec. Hence performance of the classical straight-forward GMM-UBM system is more close to i-vector based system in short duration utterance. Moreover, in case of very short utterances like 2 sec, the GMM-UBM based system has performed better over i-vector based system.

Acknowledgements

This work is partially supported by Indian Space Research Organization (ISRO), Government Of India. The work of the second author is also partially supported by the Academy of Finland (projects 253120 and 283256).

References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication* 52, no. 1 (2010): 12-40.

[2] J. Campbell, J. P., "Speaker recognition: A tutorial," *Proceedings of the IEEE* 85, no. 9 (1997): 1437-1462.

[3] S. Chakroborty and G. Saha, "Improved text-independent speaker identification using fused MFCC and IMFCC feature sets based on Gaussian filter," *International Journal of Signal Processing* 5, no. 1 (2009): 11-19.

[4] M. Hbert, "Text-dependent speaker recognition," *In Springer handbook of speech processing*, pp. 743-762. Springer Berlin Heidelberg, 2008.

[5] M. Sahidullah, and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition" *Speech Communication* 54, no. 4 (2012): 543-565.

[6] S. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech", *Speech Communication*, vol. 48, no. 10, pp. 1243 1261

[7] F. K. Soong, A. E. Rosenberg, B. Juang and L. R. Rabiner, "Report: A vector quantization approach to speaker recognition," *AT & T technical journal* 66, no. 2 (1987): 14-26.

[8] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech Signal Process*, 1981 vol. 29 no. 2, 254272

[9] D. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on* 3, no. 1 (1995): 72-83.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing* 10, no. 1 (2000): 19-41.

[11] W. M. Campbell, D. E. Sturim and D. A. Reynolds "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE 13*, no. 5 (2006): 308-311.

[12] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I-I. IEEE, 2006.

[13] D. Sengupta and G. Saha, "Study on Similarity among Indian Languages Using Language Verification Framework," In *Advances in Artificial Intelligence* , vol. 2015, Article ID 325703, 24 pages, 2015. doi: 10.1155/2015/325703

[14] M. Pal and G. Saha, "On robustness of speech based biometric systems against voice conversion attack", *Applied Soft Computing* 30 (2015): 214-228

[15] D. Paul, M. Pal and G. Saha, " Novel Speech Features for Improved Detection of Spoofing Attacks", In *India Conference (INDICON), 2015 Annual IEEE*, (Accepted)

[16] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on* 15, no. 4 (2007): 1435-1447.

[17] P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on* 16, no. 5 (2008): 980-988.

[18] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on* 19, no. 4 (2011): 788-798.

[19] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," In *Odyssey*, p. 14. 2010.

[20] A. Kanagasundaram, R. J. Vogt, D. B. Dean and S. Sridharan, "PLDA based speaker recognition on short utterances," In *The Speaker and Language Recognition Workshop (Odyssey 2012)*. ISCA, 2012.

[21] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason. "I-vector based speaker recognition on short utterances," In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pp. 2341-2344. International Speech Communication Association (ISCA), 2011.

[22] A. K. Sarkar, D. Matrouf, P. M. Bousquet and J. Bonastre, "Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification," In *INTERSPEECH. 2012*.

[23] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Communication 59*, (2014): 69-82.

[24] V. Hautamki, Y. Cheng, P. Rajan, and C.H. Lee, "Minimax i-vector extractor for short duration speaker verification," In *INTERSPEECH*, pp. 3708-3712. 2013.

[25] R. Travadi, M. V. Segbroeck and S. Narayanan, "Modified-prior i-vector estimation for language identification of short duration utterances," *In Proc. Interspeech. 2014*.

[26] O. Glembek, L. Burget, P. Matjka, M. Karafit and P. Kenny, "Simplification and optimization of i-vector extraction," In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 4516-4519. IEEE, 2011.

[27] A. Larcher, JF. Bonastre, and J. SD. Mason, "Short utterance-based video aided speaker recognition," In *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pp. 897-901. IEEE, 2008.

[28] T. Kinnunen, J. Saastamoinen, V. Hautamki, M. Vinni and P. Frnti, "Comparative evaluation of maximum a posteriori vector quantization and Gaussian mixture models in speaker verification," *Pattern Recognition Letters 30*, no. 4 (2009): 341-347.

[29] M. Sahidullah and G. Saha, "A novel windowing technique for efficient computation of MFCC for speaker recognition," *Signal Processing Letters, IEEE 20*, no. 2 (2013): 149-152.

[30] M. Sahidullah, and G. Saha, "Comparison of speech activity detection techniques for speaker recognition," *arXiv preprint arXiv:1210.0297 (2012)*.

[31] "The NIST year 2008 speaker recognition evaluation plan", *tech.rep., NIST, April 2008*

[32] "The NIST year 2010 speaker recognition evaluation plan," *tech.rep., NIST, 2010*.