Quality Measures for Speaker Verification with Short Utterances

Arnab Poddar, Md Sahidullah, Goutam Saha

Quality Measures for Speaker Verification with Short Utterances

Arnab Poddar[a,*], Md Sahidullah[b], Goutam Saha[a]

[a] *Department of Electronics & Electrical Communication Engineering, Indian Institute of Technology, India-721302, Kharagpur, India*
[b] *MULTISPEECH Team, Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France*

Abstract

The performances of the *automatic speaker verification* (ASV) systems degrade due to the reduction in amount of speech used for enrollment and verification. Combining multiple systems based on different features and classifiers considerably reduces speaker verification error rate with short utterances. This work attempts to incorporate supplementary information during the system combination process. We use quality of the estimated model parameters as a supplementary information. We introduce a class of novel quality measures formulated using the zero-order sufficient statistics used during the i-vector extraction process. We have used the proposed quality measures as side information for combining ASV systems based on *Gaussian mixture model-universal background model* (GMM-UBM) and i-vector. Considerable improvement is found in performance metrics by the proposed system on NIST SRE corpora in short duration conditions. We have observed improvement over state-of-the-art i-vector system.

Keywords: Duration Variability, Gaussian Mixture Model (GMM), Identity Vector (i-vector), Posterior Probability, Quality Measure, Short Utterances, Speaker Verification, System Fusion, Total Variability, Universal Background Model (UBM), Voice Authentication,

1. Introduction

The *automatic speaker verification* (ASV) technology uses the characteristics of human voice for the detection of individuals [1, 2]. The technology provides a low cost biometric solution suitable for real-world applications such as in banking [3], finance [4], and forensics [5]. Similar to other traditional pattern recognition applications, an ASV system includes three fundamental modules [1, 6]: an *acoustic feature extraction unit* that extracts relevant information from the speech signal in a compact manner, a *modeling block* to represent those features and a *scoring and decision* scheme to distinguish between genuine speakers and impostors. The state-of-the-art ASV system uses *i-vector* technology that represents a speech utterance with a single vector of fixed length either using *Gaussian mixture model-universal background model* (GMM-UBM) [7] or *deep neural network* (DNN) technology [8]. More recently, deep neural network (DNN) based embeddings are used for speaker recognition [9]. First, a DNN trained in a supervised manner to classify different speakers with known labels. Then, the trained DNN is employed to find a fixed-dimensional representation, known as *x-vectors* [9], corresponding to a variable length speech utterance.

Despite of these recent technological advancements, the mismatch issues are still a major concern for its real-world applications [10]. The performance of ASV system considerably degrades in presence of mismatch due to *intra-speaker variability* caused by the variations in speech duration [10, 11], background noise [12], vocal effort [13], spoken languages [14], emotion [15], channels [16], room reverberation [17], etc. In this paper, we focus on one of the most important mismatch factor, speech duration, the amount of speech data used in enrollment and verification.

*Corresponding author
Email addresses: arnabpoddar@iitkgp.ac.in (Arnab Poddar), md.sahidullah@inria.fr (Md Sahidullah), gsaha@ece.iitkgp.ernet.in (Goutam Saha)

1.1. Short utterance in speaker recognition

State-of-the-art ASV systems exhibit satisfactory performance with adequately long (2 minutes) speech data. However, reduction in amount of speech drastically degrades the ASV performance [10, 12, 18, 19, 20]. The requirement of sufficiently long speech for training or testing, especially in presence of large intersession variability has limited the potential of widespread real-world implementations. An ASV system, in real world, is naturally constrained on the amount of speech data. Though this requirement can be fulfilled in training in some special cases, it is not always possible to maintain the same in verification for end-user convenience. In forensics applications, it is less likely to get sufficient data even for enrollment also [10, 19]. Therefore, getting reliable performance for short duration speech is one of the most important requirement in ASV application.

The performance of ASV systems are notably degraded with the reduction of amount of speech due to the lack of information provided in short utterance condition [19, 18, 21, 22]. In [7], it is reported that the i-vector based ASV systems are less sensitive to limited duration utterances than *support vector machine* (SVM) and JFA. The performance still deteriorates considerably with limited duration utterance as reported in [20, 18]. The duration variability problem is handled by extracting the duration pattern from the automatic speech recognition prior to modeling and scoring process in [23]. In [24], the short duration problem is approached, demonstrating the potential of fusion between GMM-UBM and SVM based systems using logistic regression. The work in [25] attempted to model the duration variability as noise and also by a synthetic process. The work in [26] has attempted to model variability caused by short duration segments in i-vector domain. In [27, 28], i-vector based ASV system is calibrated for short duration using duration based quality measures. The work in [29] attempted to improve short utterance speaker recognition by modeling speech unit classes.

The latest DNN-based speaker embedding approaches have shown promising results for speaker recognition with short utterances [9, 30]. Another recent work demonstrates that DNN-based i-vector mapping is useful for speaker recognition with short utterances [31]. Even though the DNN-based methods give good recognition accuracy, they require massive amount of training data, careful selection of network architecture and related tuning parameters. In this current work, we aim at improving the speaker recognition performance by efficiently combining two popular ASV systems based on GMM-UBM and i-vector representation which require lesser number of tuning parameters and amount of training data compared to the DNN-based methods. Moreover, the GMM-UBM and i-vector method are suitable with limited computational resources.

1.2. Quality measure for duration-invariant speaker recognition

The research dealing with the effect of duration in speaker recognition have concentrated mostly on the consequences of classification performance, expressed in terms of *equal error rate* (EER) and *minimum detection cost function* (DCF) assuming the speaker model parameters are estimated satisfactorily. However, the speaker models are affected due to duration variability in short duration. The idea of quality metric was successfully applied in biometric authentication systems [32, 33]. The quality metrics were employed to improve the efficiency of the multi-modal biometric systems [34, 35, 36]. The work in [37] was motivated by a need to test claims that quality measures are predictive of matching performance. They also evaluated it by quantifying the association between estimated quality metric values and observed matching results.

The quality metrics are also successfully used in speech based bio-metric systems [38, 39]. The work in [39] studied a frame-level quality measure, obtaining encouraging results. However, the work in [38] showed a conventional user-independent multilevel SVM-based score fusion, adapted for the inclusion of quality information in the fusion process. The work in [40] focused on quality measure based system fusion, giving the emphasis on noisy and short duration test conditions using NIST 2012 database. The commonly used ASV systems, such as i-vector and GMM-UBM, do not include the information about the quality of estimated speaker models and information of duration variability. The work documented in [41], analyzed several quality measures for speaker verification from the point of view of their utility in an authentication task by selecting several quality measures derived from classic indicator like ITU P.563 estimator of subjective quality, signal to noise ratio and kurtosis of linear predictive coefficients. Moreover, the work [41] proposed a novel quality measure derived from what we have called universal background model likelihood (UBML).

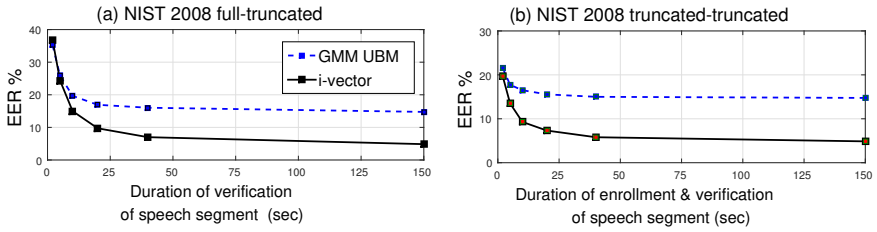

Figure 1: Plot of speaker recognition performance on (a) NIST 2008 corpus (short2-short3) with truncated enrollment and truncated verification. (b) NIST 2008 short2-short3 corpus with full duration enrollment and truncated verification.

The work in [42], analyzed the factors that negatively impact the biometric quality and also depict a review of overall framework for the challenges of biometric quality.

The work in [28] used duration of speech segments to formulate the quality metrics and subsequently utilized the same for the calibration of recognition scores. However, the duration based quality metrics may not improve performance where the duration is fixed for either enrollment or verification or both. These duration based quality measures ignored the information of quality of speaker-model estimation. The quality of speaker-model parameters are not only dependent on duration, noise but also on phonetic distribution, intelligibility of speech etc. However, to develop a solution by targeting the basic building blocks of an ASV system, we attempted to incorporate the information of duration variability which degrades the quality of speaker-models. The concept of quality may be defined as degree of goodness of an element [39, 38], which, in our case, is the speaker-models. We treat BW statistics not only as the source of speaker information but also as a source of quality of estimated speaker models.

The Baum-Welch (BW) statistics, which represent the speech features in the intermediate step of i-vector extraction process, is affected by the duration variability. Consequently, the variability gets propagated in the subsequent representation, i.e, i-vector. We hypothesize that BW statistics can help to extract the quality of speaker-model parameters. We demonstrate through graphical analysis that the utterance duration is associated with the dissimilarity measures between intermediate statistics and background model parameters. We propose to use this measure as a quality measure. In this work, we propose to formulate this quality measure from the BW statistics and universal background model (UBM) parameters.

The proposed quality measures can be infused as additional information in the ASV technique to improve the system performance. The quality measures can be incorporated in potentially four possible stages of ASV system: feature extraction, speaker-model training, score computation and fusion of scores [38]. The use of quality measures in score fusion stage is most straightforward and has been successfully applied in speech, finger-print, face based multimodal person authentication systems [43, 44]. In this paper, we incorporate the proposed quality measures in score fusion stage to improve the performance of speaker recognition system in various duration conditions. In short duration, the linear score fusion strategy showed efficient performance with GMM and SVM based classification framework [24]. However, the i-vector based system (with GPLDA based channel compensation) was reported to perform more efficiently over JFA and GMM-SVM based framework in short utterance conditions [7]. Here, we show a comparative performance study of i-vector and GMM-UBM on NIST corpora (Fig. 1). We observe that though i-vector system performs better than GMM-UBM for long duration speech, the GMM-UBM system still shows comparable or even better performance for short duration conditions [45, 46]. This observation inspire to fuse i-vector and GMM-UBM to develop a more accurate and reliable solution for practical application of ASV systems. We have incorporated the estimated quality measures while blending the GMM-UBM and i-vector based ASV system. Incorporation of quality measures not only showed considerable improvement in performance but also consistency in various duration conditions. The proposed systems showed more relative improvement in short duration conditions which is more relevant for practical requirement. A preliminary version of this work was presented in [45]. In this work, we conduct extensive analysis and experiments.

The rest of the paper is organized as follows. The theoretical aspects of classical GMM-UBM and i-vector GPLDA system are discussed in Section 2. Analysis on intermediate subsystems under different duration

variability condition is presented in Section 3. Section 4 describes the proposed quality measures and quality aided fusion based system. Details of experimental setup are provided in Section 5. The comparison of performance metrics of baseline GMM-UBM and i-vector GPLDA based ASV system and results on proposed quality aided fusion system are reported in Section 6. Finally, conclusion is drawn in Section 7.

2. Automatic Speaker Recognition System

Speaker recognition system, based on Gaussian mixture model, has emerged as the most widely used fundamental approach with the introduction of universal background model [47]. Subsequently, GMM supervector based SVM [48], and JFA [49] were introduced in ASV technology. Recent state-of-the-art speaker recognition concentrates on compact representation of GMM supervectors, named as i-vectors [7]. This work considers ASV system based on subspace modeling of i-vectors using PLDA [50]. This section presents a brief explanation of GMM-UBM, i-vector and PLDA.

2.1. GMM-UBM based ASV system

In GMM-UBM, prior to enrollment phase, a single speaker independent universal background model is created by using a large development data [47]. The UBM is represented as $\boldsymbol{\lambda}_{\text{UBM}} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{C}$ 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 $\boldsymbol{\Sigma}_i$ represents the co-variance matrix. 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. First, sufficient statistics N_i and $\mathbf{E}_i$ from a hypothesised speaker's utterance with T speech frames $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, are calculated as,

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

where posterior probability of i-th component $Pr(i|\mathbf{x}_t)$, is conditioned on speech data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$.

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

$$\Lambda_{\text{UBM}}(\mathbf{X}^{\text{test}}) = \log\ p(\mathbf{X}^{\text{test}}|\boldsymbol{\lambda}_{\text{target}}) - \log\ p(\mathbf{X}^{\text{test}}|\boldsymbol{\lambda}_{\text{UBM}}) \tag{2}$$

Finally, a decision logic is applied to decide whether the claimant speaker will be accepted or rejected. A decision threshold is used for decision, like if $\Lambda_{\text{UBM}}(\mathbf{X}^{\text{test}})$ exceeds a predefined threshold then the claim will be accepted, else rejected.

2.2. i-vector based ASV system

The i-vector represents the GMM supervector in a *total variability* space which reduces dimension of GMM supervector [7]. In i-vector space, the GMM supervector, i.e, the concatenated means of GMM mixture components, is represented as $\mathbf{M} = \mathbf{m} + \boldsymbol{\Phi}\mathbf{y}$, where $\boldsymbol{\Phi}$ is a low-rank total variability matrix and $\mathbf{y}$ represents i-vector, $\mathbf{m}$ is the speaker and channel independent supervector (taken to be UBM supervector) and $\mathbf{M}$ is the speaker and channel dependent GMM supervector.

Zeroth and first order BW statistics N_i and $\mathbf{E}_i$ respectively are used to obtain the i-vector $\mathbf{y}$. The prior distribution of i-vectors $p(\mathbf{y})$ is assumed to be $\mathcal{N}(0, I)$ and posterior distribution of $\mathbf{E}$, conditioned on the i-vector $\mathbf{y}$ is hypothesized to be $p(\mathbf{F}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\Phi}\mathbf{y}, \mathbf{N}^{-1}\boldsymbol{\Sigma})$. The MAP estimate of $\mathbf{y}$ conditioned on $\mathbf{E}$ is given by

$$\mathbb{E}(\mathbf{y}|\mathbf{E}) = (\mathbf{I} + \boldsymbol{\Phi}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{N}\boldsymbol{\Phi})^{-\mathbf{1}}\boldsymbol{\Phi}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{N}(\mathbf{E} - \bar{\mathbf{m}}), \tag{3}$$

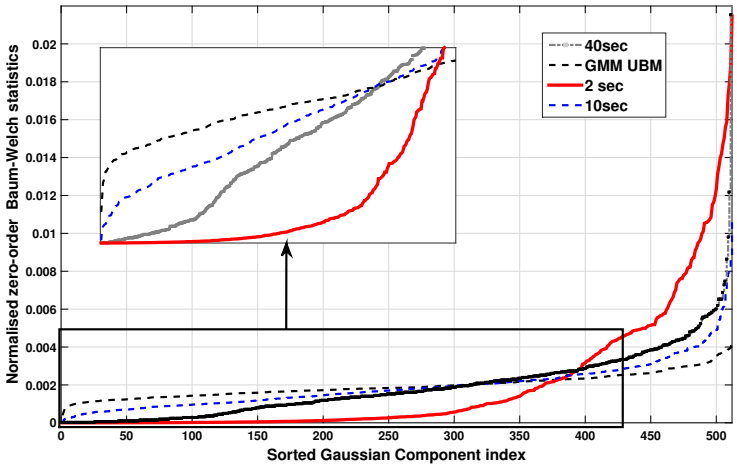

Figure 2: Sorted distribution of normalized zeroth-order BaumWelch statistics ($\tilde{N}_i$) corresponding to various duration conditions (2 sec, 10 sec, 40 sec). Prior weights of GMM-UBM components (w_i) are also shown in sorted ascending order.

where $\mathbb{E}(\mathbf{y}|\mathbf{E})$, the expected value of the posterior distribution of $\mathbf{y}$ conditioned on $\mathbf{E}$ is considered as the *i-vector* representation of a speech utterance. Here $\mathbf{I}$ refers to Identity matrix, the term $\mathbf{\Phi}$ refers to total variability matrix, estimated from the development data. The symbol $\mathbf{\Sigma}$ refers the co-variance matrix adopted from UBM. The term $\bar{\mathbf{m}}$ represents the concatenated mean of the UBM components.

2.3. 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) modeling approach. In this paper, we have used a simplified variant of PLDA, named as Gaussian PLDA [50]. The inter-speaker variability is modeled by a full co-variance residual term. The generative model for s-th speaker and j-th recording of new i-vector variability projected space is given by

$$\mathbf{y}_{s,j} = \boldsymbol{\eta} + \mathbf{\Psi}\mathbf{z}_s + \boldsymbol{\epsilon}_{s,j}, \tag{4}$$

where $\boldsymbol{\eta}$ is the mean of the development i-vectors, $\mathbf{\Psi}$ is eigen-voice subspace and $\mathbf{z}$ is a vector of latent factors. The residual term $\boldsymbol{\epsilon}$ represents the variability not captured by the latent variables. This generative model approach of i-vector space representation has been applied successfully with considerable improvement in recognition accuracy [50].

2.4. Likelihood Computation

score calculation of GPLDA based i-vector system uses likelihood ratio [50]. For a projected enrollment and verification i-vector $\mathbf{z}_{\text{target}}$ and $\mathbf{z}_{\text{test}}$ respectively, the likelihood ratio $\Lambda_{\text{GPLDA}}(\mathbf{z}_{\text{target}}, \mathbf{z}_{\text{test}})$ can be calculated as follows:

$$\mathbf{\Lambda}_{\text{GPLDA}}(\mathbf{z}_{\text{target}}, \mathbf{z}_{\text{test}}) = \log \frac{p(\mathbf{z}_{\text{target}}, \mathbf{z}_{\text{test}}|H_1)}{p(\mathbf{z}_{\text{target}}|H_0)\ p(\mathbf{z}_{\text{test}}|H_0)} \tag{5}$$

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

3. Analysis on BW statistics extraction procedure

Previous studies dealing with duration variability concentrated on the final performance metrics measured in terms of EER, DCF, etc [18]. Some studies focused the variability in i-vector space [26]. However in this work, we present a study on how duration variability affects the intermediate steps of ASV system. BW statistics represent the total information from the speech and are transformed into i-vectors for decision making. Since, in most of modern ASV systems, BW statistics is an indispensable and important step, we initially conduct analysis on the BW sufficient statistics to investigate the characteristics and effect of the short duration.

We have the relationship for zeroth order BW statistics as, $N_i = \sum_{t=1}^{T} Pr(i|\mathbf{x}_t)$, summing over all Gaussian mixture components C we obtain,

$$\sum_{i=1}^{C} N_i = \sum_{i=1}^{C} \sum_{t=1}^{T} Pr(i|\mathbf{x}_t) = T \tag{6}$$

Normalizing zeroth order statistics for a single Gaussian mixture component i we get

$$\tilde{N}_i = \frac{1}{T} \sum_{t=1}^{T} Pr(i|\mathbf{x}_t) \quad \text{and} \quad \sum_{i=1}^{C} \tilde{N}_i = 1 \tag{7}$$

Hence normalized zeroth order BW statistics (NBS) has the same property as weights of the GMM-UBM, i.e., $\sum_{i=1}^{C} w_i = 1$. Moreover, $\tilde{N}_i$ can be regarded as the mixture weight indicator of the Gaussian component i for a particular speech segment. It is a standard statistical hypothesis that BW statistics are better estimated with sufficiently large speech data which capture all kinds of variability with meaningful proportion. Hence, it is expected that the higher value of T with phonetically rich speech segment would lead to better quality of estimation of $\tilde{N}_i$. On the other hand, the intermediate statistics may be expected to be updated more sparsely for reduced speech data or degraded quality of speech. On this core note, the characteristics of $\tilde{N}_i$ are investigated. Systematic studies are presented separately on single utterance and multiple utterances from multiple speakers.

3.1. Baum-Welch statistics for short-utterance

Initially, a telephone utterance from NIST SRE 2008 short-2 enrollment corpus of male speakers is taken for analysis on NBS of different duration conditions, e.g., 2 sec, 10 sec and full length (1.5 - 2.5 mins). In Fig. 2, $\tilde{N}$ is plotted in ascending order with respect to Gaussian mixture components for different duration conditions. We observe from Fig. 2 that the gradient is steeper for short duration compared to the other longer duration conditions, whereas that of the UBM showed more flat nature. This observation refers to the fact that only a few number of Gaussian components are associated with most of the speech frames in limited duration conditions. A large number of Gaussian components do not associates adequate speaker-specific frames which finally affects the quality of model estimation. This effect reduces as the duration of the utterance increases. As GMM-UBM is estimated from sufficiently large pool of speech data, most of the Gaussian components are occupied by adequate speech frames. Thus, it shows more uniform nature in Fig. 2. We hypothesize that the introduction of greater variability in zeroth order statistics especially for short duration condition indicates the lower quality of estimation of model parameters (NBS). This indicates the quality of estimation of NBS is degraded in short duration condition. We treat NBS not only as a source of speaker information, it can help measure the quality of estimated model of speakers.

3.2. Analysis of NBS in multiple speech utterances

In order to interpret the NBS, we have done an analysis on entire male part of NIST SRE 2008 consisting. The effect of speech duration on the NBS is presented in Fig. 3 by showing mean (Fig. 3(a)) and standard deviation(Fig. 3(b)) of NBS per Gaussian components for three duration conditions (2 sec, 10 sec and full

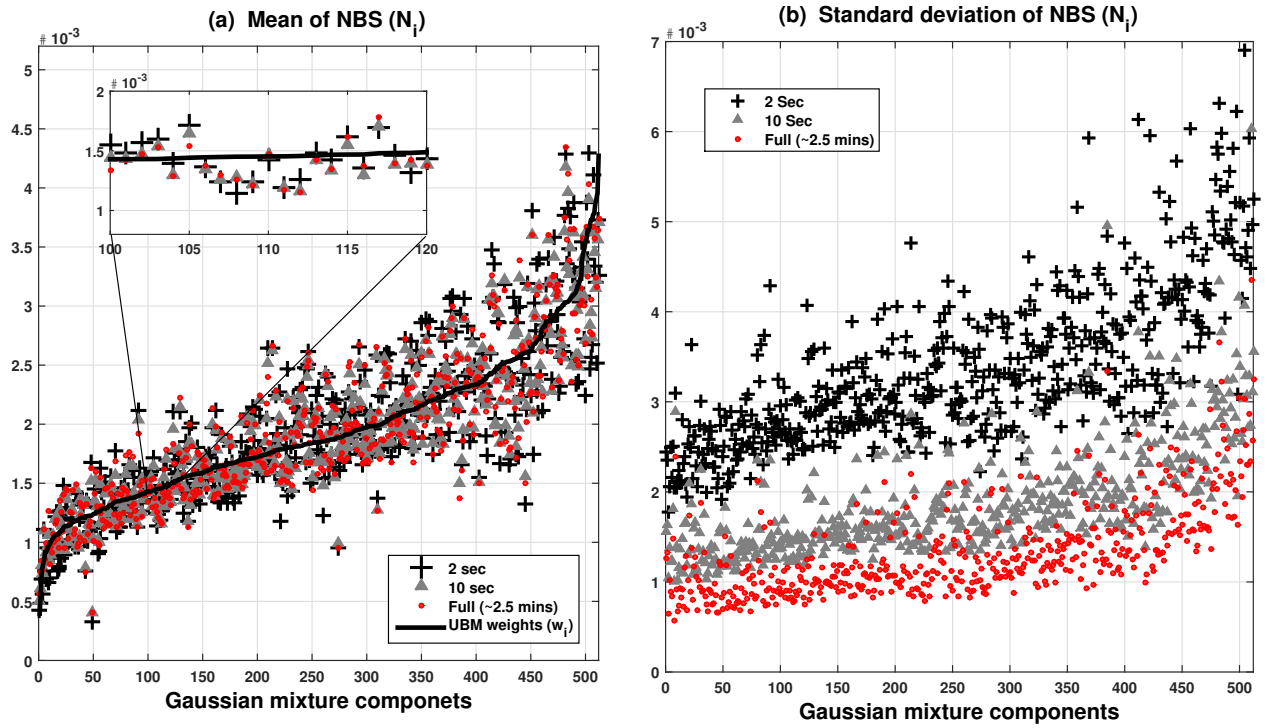

Figure 3: (a)Mean and (b)standard deviation of normalized Baum-Welch statistics (NBS) for each of the 512 Gaussian components computed on three duration conditions (2 Sec, 10 Sec and full). Figure(a) includes the weights of GMM-UBM for corresponding Gaussian. We have used all the male speakers of NIST 2008 (short2) corpus for this analysis.

length). We have also shown the weights of GMM-UBM of corresponding mixture components (w_i) are presented in Fig. 3(a).

We observe in Fig. 3(a) that the means of $\tilde{N}$ for different duration condition follows the value of UBM weight of corresponding Gaussian mixture component (w_i). We also notice that for different duration conditions, the means of different duration condition for a particular Gaussian component remains nearly similar. This is observed in almost all Gaussian components shown in Fig. 3(a). These observations on means of $\tilde{N}$ distribution and weights of corresponding Gaussian mixture component inspired us to use GMM-UBM weights (w_i) as reference to measure the variability in $\tilde{N}$.

The short segments in Fig. 3(b) also show greater standard deviation referring greater variability introduced in NBS. We observe gradual increment in the variability with a reduction in speech segment length. As the variability in NBS is due to the duration, we hypothesize that the duration-related quality can be reflected in the NBS.

4. Quality measures for speech segments

The observations in previous section illustrate that the variability in BW statistics is in someway associated with the duration of speech. The work in [51] also attempted to model the sparsity and variability to compensate the performance of ASV in short duration. The work in [52] introduced an uncertainty measure computed from the the i-vector posterior parameter to compensate the duration variability effect. In our current work, we propose to apply dissimilarity metrics between NBS ($\tilde{N}_i$) and prior of corresponding Gaussian component of UBM model (w_i) to measure the impact of duration on speaker representation. Subsequently, this is incorporated as supporting information in fusion of ASV system.

Table 1: Mathematical expression for quality measures formulated using normalized Baum-Welch statistics and UBM weights.

Quality Measure	Short form
$Q_1(\tilde{N}_s) = \sum_{i=1}^{C}(\tilde{N}_{i,s} log \frac{\tilde{N}_{i,s}}{w_{i,\text{ubm}}})$	kl-1
$Q_2(\tilde{N}_s) = \sum_{i=1}^{C}(w_{i,\text{ubm}} log \frac{w_{i,\text{ubm}}}{\tilde{N}_{i,s}})$	kl-2
$Q_3(\tilde{N}_s) = \frac{1}{2}\sum_{i=1}^{C}(\tilde{N}_{i,s} log \frac{\tilde{N}_{i,s}}{w_{i,\text{ubm}}} + w_{i,\text{ubm}} log \frac{w_{i,\text{ubm}}}{\tilde{N}_{i,s}})$	kl-avg
$Q_4(\tilde{N}_s) = \sum_{i=1}^{C} \lvert\tilde{N}_{i,s} - w_{i,\text{ubm}}\rvert$	ℓ_1 norm
$Q_5(\tilde{N}_s) = \sqrt{(\sum_{i=1}^{C} \lvert\tilde{N}_{i,s} - w_{i,\text{ubm}}\rvert^2}$	ℓ_2 norm
$Q_6(\tilde{N}_s) = \sqrt{(\sum_{i=1}^{C} \tilde{N}_{i,s} w_{i,\text{ubm}})}$	bh

4.1. Quality Measure Modeling

The trends observed in Section 3, are exploited to empirically model the quality of speech segments. In this paper, six types of dissimilarity measures are adopted to model the duration variability which degrades quality of speaker model estimation. The mathematical expressions to model the quality Q of a segment s are presented in Table 1. The adopted quality measures differ from the way of measuring dissimilarity of $\tilde{N}_i$ from the weights of UBM (w_i), which is treated as reference. Quality measure operators Q_1, Q_2 and Q_3 models the Kullback-Liebler divergence between NBS ($\tilde{N}_{i,s}$) and weights of Gaussian mixture components of UBM $w_{\text{i,ubm}}$. These metrics attempts to capture the divergence of the distribution of NBS and UBM weights. Furthermore, we have also used other metrics to measure the dissimilarity. However, ℓ_1-norm and ℓ_2-norm are applied in quality measure operator $Q_4 and Q_5$ respectively. The Bhattacharyya distance is used in to measure the overlap of the population samples of NBS and weights of UBM.

4.2. Statistical analysis of Quality Measures

An analysis for statistical relevance of the modeling of quality measures is illustrated in this section. Segments from NIST 2008 short2 enrollment corpus with 1270 male speakers are used for assessment. The distribution mean of dissimilarity measures $\bar{Q}_j$ of all 1270 segments and its truncated versions (2 sec, 5 sec, 10 sec, 20 sec, full) are presented in Table 2. The mathematical expression for the mean of quality measures of type $\bar{Q}_j$ estimated from NIST 2008 short-2 corpus consisting of H utterances is given by

$$\bar{Q}_j = \frac{1}{H}\sum_{s=1}^{H} Q_j(\tilde{N}_s) \qquad j = 1, 2, \ldots 6 \tag{8}$$

Separate analysis are presented in Table 2 for six types of dissimilarity measures ($Q_j, \quad j = 1, 2, \ldots 6$) . The distribution of quality measure of Q_4 for truncated version of 2 sec, 10 sec and full duration are plotted in Fig.4. Results indicate that $\bar{Q}_j$ is decreased as the duration of speech segment rises as shown in Table 2 and Fig. 4. Hence, 2 sec segment shows highest value for $\bar{Q}_j$ and full duration segment shows the lowest. In Table 2, we observe gradual decrements in distribution mean of dissimilarity measure $\bar{Q}_j$ with the increment of speech duration. The decrements are consistent for almost all kind dissimilarity measures presented in Table 1. These observations leads to use the proposed dissimilarity measure as the overall quality of speaker model.

4.3. Fusion of ASV systems with quality measures

Quality measures can be incorporated in different stages of a speaker verification system like model training [53], computation of scores [28, 27], scores fusion [54], etc. The existing literature show that

Table 2: Mean of proposed quality measures of all segments in NIST 2008 short2 enrollment corpus and that of its truncated versions for all six types quality measures.

duration	$\bar{Q}_1$ kl-1	$\bar{Q}_2$ kl-2	$\bar{Q}_3$ kl-avg	$\bar{Q}_4$ ℓ_1 norm	$\bar{Q}_5$ ℓ_2 norm	$\bar{Q}_6$ bh
2 sec	2.250	1.051	1.652	1.144	0.084	0.536
5 sec	1.314	0.679	0.997	0.899	0.061	0.399
10 sec	0.963	0.521	0.749	0.742	0.055	0.327
20 sec	0.782	0.437	0.608	0.654	0.050	0.273
Full	0.309	0.226	0.268	0.499	0.031	0.178

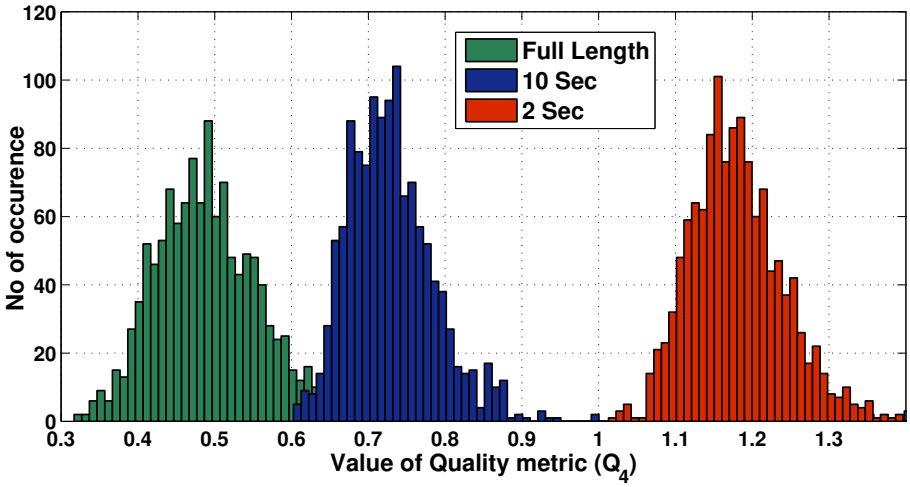

Figure 4: Distribution of quality measure Q_4 for *2 sec, 10 sec, and full* segment, estimated from *NIST 2008 short2 enrollment* corpus.

the quality information can be used in many applications, especially for score fusion in biometric person recognition system [43, 44, 55, 36]. In our work also, we incorporate the quality measures in score fusion step. In ASV, fusion-based approaches have found very much suitable for improving recognition accuracy [40, 56]. Though i-vector and GMM-UBM based ASV systems uses different approaches, we show in Section 6 that GMM-UBM exhibit competitive or even better performance over i-vector in short utterance conditions. We fuse i-vector and GMM-UBM to exploit the information captured simultaneously by two systems. The fusion parameters are trained using logistic regression objective using the BOSARIS toolkit [57]. Separate NIST corpora, SRE 2008 [58] and SRE 2010 [59], are used for training and evaluation of the fusion parameters respectively.

We confine our work to score level fusion with fusion function f which combines two base classifier scores, Λ_{UBM} and Λ_{GPLDA}, into a single score,

$$\mathbf{\Lambda} = \{\Lambda_{UBM}, \Lambda_{GPLDA}\}^\top. \tag{9}$$

The decision is made by comparing the fused score to a threshold. An annotated development set $\mathcal{D} = \{(\mathbf{\Lambda}_i, c_i), i = 1, 2, \ldots, N_{dev}\}$ with $c_i \in \{+1, -1\}$ representing corresponding speech frame from target speaker ($c_i \in \{+1\}$) or imposter ($c_i \in \{-1\}$) is used to train the fusion parameters.

The general model of linear fusion of the two systems is represented by:

$$f_{\text{lin}}(\mathbf{\Lambda}) = \alpha^\top \mathbf{\Lambda} + \bar{\theta}, \tag{10}$$

where α is the fusion weight and θ is the bias. These fusion parameters are estimated by logistic regression on the development scores.

After incorporating the quality measures for enrollment ($Q(\tilde{N}_{\text{enrol}})$) and test ($Q(\tilde{N}_{\text{test}})$) utterances, the general model of fusion is given by:

$$f_Q(\mathbf{\Lambda}) = \alpha_Q^\top \mathbf{\Lambda} + \theta_Q + \beta \times Q(\tilde{N}_{\text{enrol}})Q(\tilde{N}_{\text{test}}), \tag{11}$$

where α_Q, θ_Q, and β are the parameters for quality measure fusion. Note that quality-based fusion is also a type of linear fusion where the quality parameters are incorporated as additional similar scores. We have also conducted a separate experiment where we have incorporated the proposed quality metric in single i-vector system for duration-based score calibration [28]. The general model for adding quality measure in single i-vector based system is given by,

$$f_C(\mathbf{\Lambda}) = \alpha_C \Lambda_{GPLDA} + \theta_C + \beta_C \times Q(\tilde{N}_{\text{enrol}})Q(\tilde{N}_{\text{test}}) \tag{12}$$

where the parameters α_C, θ_C and β_C are the parameters for calibration and they are estimated on the development set with *known* labels and applied evaluation set with *unknown* labels. In our experiments, we have observed that the i-vector based scores and GMM-UBM based scores can be fused linearly which yields encouraging performance improvement [46, 45, 60]. Hence, we attempt to add quality information in linear fusion of i-vector and GMM-UBM system. However, while defining fusion model, we have incorporated the quality information in score fusion using trial-by-trial manner. We have attempted to model the quality of a trial by simply multiplying the quality measures, separately estimated from the BW statistics of enrolment (Q_{enrol}) and test (Q_{test}) utterance. Thus, we obtain the overall trial quality as. $Q_{train} \times Q_{test}$. We have used six types of quality measures to improve the performance of ASV system in various duration condition in Table 1.

For a quality fusion function $f_Q(\mathbf{\Lambda})$ with parameters (α, β, θ), the development data $\mathcal{D}$ and an empirical cost function $\hat{C}((\alpha, \beta, \theta), \mathcal{D})$ are given, the optimal fusion function is obtained by

$$(\alpha^{\text{dev}}, \beta^{\text{dev}}, \theta^{\text{dev}}) = argmin_{\alpha,\beta,\theta} \hat{C}(\{\alpha, \beta, \theta\}, \mathcal{D}), \tag{13}$$

where *decision cost function*, C is defined as,

$$C_{\text{det}}(\zeta) = C_{\text{miss}} P_{\text{miss}}(\theta) P_{\text{tar}} + C_{\text{fa}} P_{\text{fa}}(\zeta)(1 - P_{\text{tar}}) \tag{14}$$

where ζ is the threshold, P_{tar} is the prior probability of the target speaker, C_{miss} is the cost of the miss and C_{fa} is the cost of the false alarm.

A schematic diagram for overall ASV system with linear as well as quality-measure based fusion is shown in Fig. 5.

5. Experimental Setup

Both GMM-UBM and i-vector based systems use same mel frequency cepstral coefficients (MFCCs)[61] as front-end acoustic features. We extract MFCCs using frame size of 20 ms and frame shift of 10 ms as in [62]. The Hamming window is used in MFCC extraction process [63]. The non-speech frames are dropped using energy-based speech activity detector (SAD) [64]. Finally, we performan cepstral mean and variance normalization (CMVN) to remove the convolutive channel effect[62]. 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 SRE 2004, 2005, 2006 and Switchboard II corpora as development data to estimate UBM, LDA and GPLDA parameters. Total variability subspace of dimension 400 is chosen for i-vector extractor. We perform LDA on i-vectors to improve the speaker discriminatibility and the dimensions are reduced to 200. Finally, GPLDA with 150 eigen-voice space is used for scoring. We estimated the GPLDA parameters with random initialization and 20 iterations of EM algorithm.

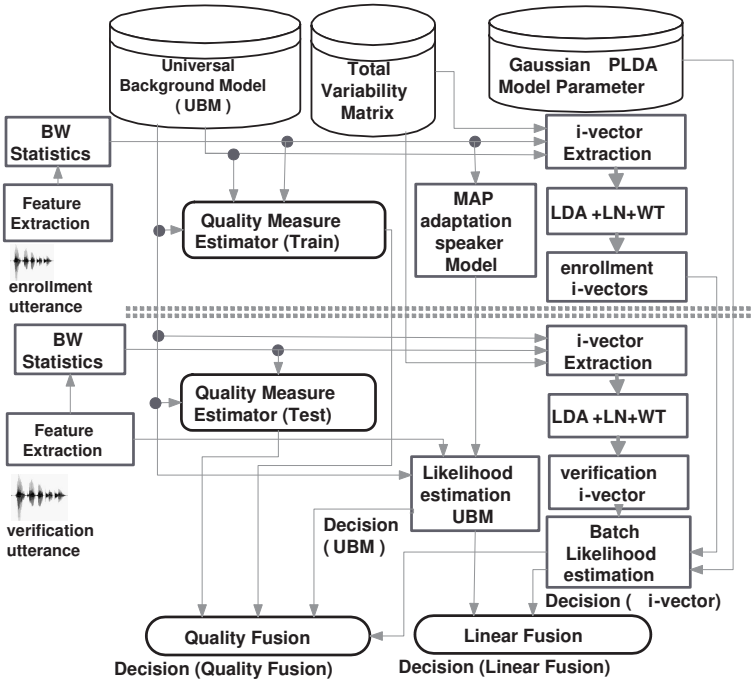

Figure 5: Schematic block diagram for proposed speaker verification system showing decisions obtained with both linear fusion and quality-based fusion.

Table 3: Summary of speech corpora used in the experiments.

Specifications	NIST SRE 2008	NIST SRE 2010
#target model	482	489
#test segments	858	351
#genuine trials	874	353
#imposter trials	11637	13707

5.1. Experiments and Corpora

The performance of two major speaker modelling methods and proposed methods were evaluated on NIST SRE 2008 [58] and NIST SRE 2010 [59] corpus. NIST 2008 *short2-short-3* and NIST 2010 *core-core* speaker recognition protocol is used as development and evaluation data respectively. Utterance truncated versions of both databases are used for experiments in varying utterance duration condition. Only *telephone-telephone* trials of male speakers from NIST SRE 2008 and NIST SRE 2010 are used in the following experiments. The summary of the databases used in the experiments are shown in Table 3.

5.2. Utterance duration and truncation procedure

In core condition of NIST SRE corpora, the duration of speech segments are long (2.5 min of speech). In order to conduct experiments in short duration conditions, truncation of speech utterances is done in 2 sec (200 active frames), 5 sec (500 active frames), 10 sec (1000 active frames) and 20 sec (2000 active frames) duration. For truncation of utterances, the prior 500 active speech frames are discarded at the feature level after VAD to avoid phonetic similarity in initial greetings of telephonic conversations which introduces text dependence.

The original utterances from the NIST SRE corpus without any truncation is referred to as *full* condition in this paper. From the *full* condition features, and features generated from truncated segments, six test sets with different duration conditions in both model and verification segments collection are designed. Fourteen

Table 4: Speaker verification performance on NIST 2008 using GMM-UBM (UBM) and i-vector (TV) based system. The results are shown in terms of EER (in %) and DCF (× 100).

Train-Test duration	EER (UBM)	EER (TV)	$RI_{\text{TV}}^{\text{EER}}$ [%]	DCF (UBM)	DCF (TV)	$RI_{\text{TV}}^{\text{DCF}}$ [%]
Truncated training - Truncated testing						
2s-2s	**35.24**	36.84	-4.54	**9.69**	9.93	-2.47
5s-5s	25.25	**24.37**	3.49	8.89	**8.65**	2.69
10s-10s	19.67	**14.98**	23.84	8.23	**6.58**	20.04
20s-20s	16.93	**9.72**	42.58	8.19	**4.62**	43.58
Full training - Truncated testing						
Full-2s	21.56	**19.67**	8.76	**7.75**	7.91	-2.06
Full-5s	17.73	**13.50**	23.85	7.32	**5.99**	18.16
Full-10s	16.66	**8.78**	46.93	6.75	**4.50**	33.33
Full-20s	15.52	**7.32**	52.83	6.58	**3.63**	41.90
Full-Full	14.75	**4.86**	67.05	6.23	**2.70**	59.09

trial conditions are arranged by combining different duration train-test segments and <*full*>-<*duration of test segment*> for both NIST SRE 2008 and 2010. We use the notation '<*duration of model segment*>-<*duration of test segment*> condition,' in which duration is measured in seconds or *full*.

5.3. Performance Evaluation Metrics

We have evaluated the performance using EER and DCF as performance evaluation metric. The EER is the point on *detection error trade-off* (DET) plot where the probability of false acceptance and probability of false rejection are equal. The DCF is computed by creating a cost function assigning separate weights on false alarm and false rejection followed by computation of threshold where cost function is minimum. The cost function is computed as

$$C_{\text{Det}} = C_{\text{Miss}} \times P_{\text{Miss|Target}} \times P_{\text{Target}} + C_{\text{FalseAlarm}} \times P_{\text{FalseAlarm|NonTarget}} \times (1 - P_{\text{Target}}) \tag{15}$$

The DCF is calculated using the parameter value $C_{\text{Miss}} = 10$, $C_{\text{FalseAlarm}} = 1$ and $P_{\text{target}} = 0.01$ for both databases NIST 2008 and NIST 2010 [58, 59]. We also report relative improvement for parameter p of system s_1 over system s_2, calculated as

$RI_{\text{s}_1}^{\text{p}} = \frac{(p_{\text{s}_1} - p_{\text{s}_2})}{p_{\text{s}_2}} \times 100\%$

6. Experiments Results and Discussion

6.1. Baseline Performance

Initially, we have investigated the performance of state-of-the-art i-vector and classical GMM-UBM based ASV system under various duration condition. The experiments are executed on male subset of both NIST 2008 *short2-short3* corpus. Comparison of performance is accomplished in eleven different duration conditions separately.

The results of the experiments reported in Table 4. Fig.6 exhibits a systematic comparative study. Table 4 depicts the relative performance improvement of i-vector based system over GMM-UBM based system i.e. $RI_{\text{TV}}^{\text{EER}}$ and $RI_{\text{TV}}^{\text{DCF}}$ decreases monotonically with the reduction in utterance duration. Table 4 also shows that i-vector based system worked better than GMM-UBM for longer utterances but both the system performance falls on duration as small as *2 sec, 5 sec* etc. We also observed that in case very similar to real-time requirements i.e., very short duration utterances, specially in *full duration training - 2*

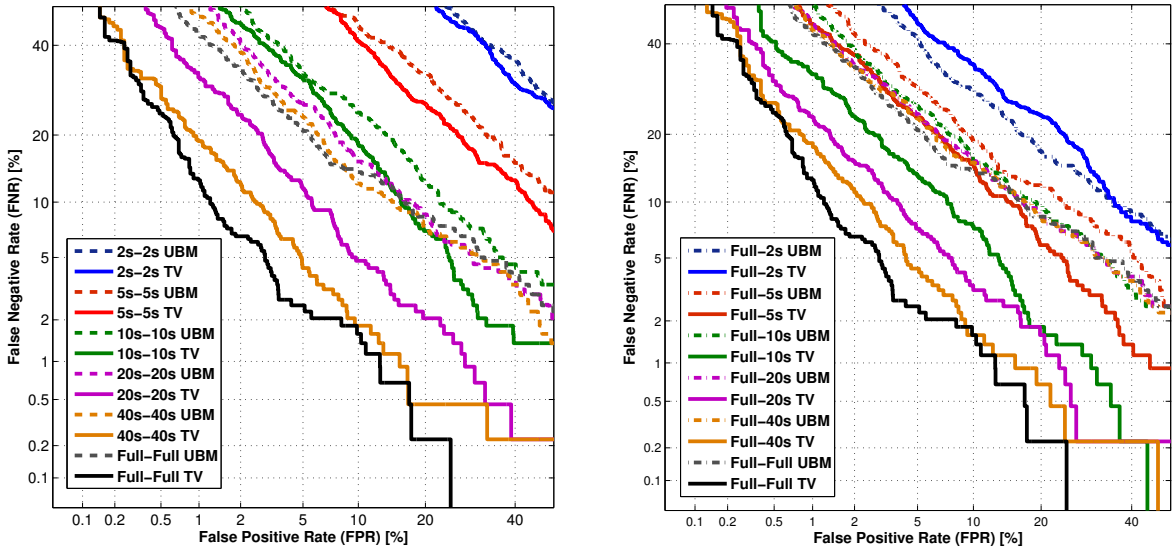

Figure 6: Speaker verification DET plot of i-vector (TV) system and UBM system in short2-short3 sub-condition of NIST SRE 2008 corpus.

sec testing and *2 sec training - 2 sec testing*, GMM-UBM based system showed comparable or even better performance over i-vector based system.

The above observations invoke to fuse GMM-UBM and i-vector system which should combine the advantage of both the systems. This is also provides opportunity to include the proposed quality metric as an additional information. Fusion parameters are estimated on NIST 2008 *short2-short3* corpus and validated the same on NIST 2010 *core-core* task for the sake of generality. Quality measure are incorporated in the ASV system to support the fusion device. Table 5, 6, 7 and 8 represents the results of both linear fusion and quality measure based fusion. Table 5 and 7 represents the results in terms of EER whereas Table 6 and 8 represents in terms of DCF values on development data and evaluation data respectively. In all these tables, we have shown the results for 15 different duration conditions including the *Full-Full* scenario.

6.2. Results of proposed ASV technique

It may be observed from Tables 5, 6, 7, 8, that the fusion gives higher improvement in shorter utterances. Irrespective of enrollment and verification, as the utterance length of speech segments becomes longer, the relative difference between performance of GMM-UBM and i-vector system increased considerably. The fusion based system evolved to be more effective in cases where performance of GMM-UBM and i-vector based system is more comparable i.e., in short utterance cases. The relative boost of the linear fusion based system over i-vector GPLDA based baseline ASV system showed high values up-to $10\% - 15\%$ in cases like *full-2 sec, full-5 sec, 20 sec-5 sec, 2 sec-2 sec*, etc. This advocates more potential of the fusion based system in real-world scenario.

Quality measures of speech signals are used to support the fusion ASV system further. Inclusion of proposed quality metrics derived from BW statistics in the proposed system showed significant improvement in performance. Here quality measures are proposed in such a way that it requires almost negligible additional computation and no additional parameters estimation. Performance measures of fusion method using six types of quality measure are shown in Table 5, 6, 7 and 8. DET curve showing speaker verification performance of the i-vector baseline system, GMM-UBM based system and proposed quality fusion based system is presented in Fig. 7.

Table 5: Results in EER of fusion of GMM-UBM and i-vector based system on NIST 2008 telephone-telephone development corpora

Train-Test duration	i-vec GPLDA	linear fusion	Q_{tv}	Q_1 kl-1	Q_2 kl-2	Q_3 kl-avg	Q_4 ℓ_1	Q_5 ℓ_2	Q_6 bh
Truncated training - Truncated testing									
2s-2s	36.84	32.95	33.98	31.92	**31.80**	31.75	31.92	31.88	**31.80**
5s-2s	30.48	30.37	28.33	**27.23**	27.45	27.57	27.45	28.48	28.72
5s-5s	24.37	23.11	23.17	22.99	21.62	**21.22**	21.25	21.70	21.62
10s-2s	26.10	25.40	23.91	23.34	23.01	23.79	**22.88**	23.21	23.11
10s-5s	19.56	17.50	16.84	17.16	16.93	17.16	16.84	16.82	**16.73**
10s-10s	14.98	14.53	13.95	12.92	13.72	14.30	13.84	**13.15**	14.07
20s-2s	24.20	21.73	22.05	20.59	20.48	20.59	**20.18**	20.48	20.36
20s-5s	17.29	14.94	15.95	14.94	14.94	15.07	14.94	15.23	**14.66**
20s-10s	12.35	11.67	11.66	11.62	11.59	11.56	**11.55**	11.67	11.63
20s-20s	9.72	9.83	9.83	**9.64**	9.71	9.83	9.91	9.80	9.77
Full training - Truncated testing									
Full-2s	19.67	19.10	21.51	**16.47**	16.70	16.47	16.81	17.45	16.76
Full-5s	13.50	12.70	13.79	11.78	11.67	11.70	11.89	**11.83**	12.12
Full-10s	9.29	9.72	14.73	**9.15**	9.24	9.27	9.09	9.26	9.49
Full-20s	7.32	7.30	9.14	7.32	7.36	7.31	**7.26**	7.39	7.35
Full training - Full testing									
Full-Full	4.86	4.91	7.29	5.20	5.20	5.14	5.06	5.16	5.14

Table 6: Results in DCF of fusion of GMM-UBM and i-vector based system on NIST 2008 telephone-telephone development corpora

Train-Test duration	i-vec GPLDA	linear fusion	Q_{tv}	Q_1 kl-1	Q_2 kl-2	Q_3 kl-avg	Q_4 ℓ_1	Q_5 ℓ_2	Q_6 bh
Truncated training - Truncated testing									
2s-2s	9.93	9.53	9.64	**9.52**	**9.50**	9.53	**9.50**	9.56	9.54
5s-2s	9.75	9.67	9.73	8.85	**8.87**	8.87	8.92	9.13	8.89
5s-5s	8.65	8.27	8.33	8.19	8.16	8.13	**8.19**	8.24	8.24
10s-2s	9.23	9.05	8.54	8.27	8.28	**8.27**	8.30	8.36	8.32
10s-5s	7.80	7.34	7.30	7.15	**7.12**	7.15	**7.13**	7.29	7.29
10s-10s	6.58	6.40	6.32	**5.93**	6.04	6.04	5.86	6.03	5.99
20s-2s	8.77	8.20	8.14	7.79	7.73	7.76	**7.79**	7.79	7.80
20s-5s	6.98	6.65	6.71	6.39	6.37	**6.34**	6.34	6.47	6.46
20s-10s	5.48	5.42	5.28	5.36	5.24	5.26	**5.22**	5.33	5.30
20s-20s	4.62	4.59	4.61	4.55	4.52	4.53	**4.51**	4.52	**4.51**
Full training - Truncated testing									
Full-2s	7.75	7.53	7.09	7.01	7.02	**6.99**	7.01	7.09	7.09
Full-5s	5.99	5.95	6.15	5.44	**5.36**	5.42	5.19	4.93	5.57
Full-10s	4.50	4.59	6.92	4.27	4.26	**4.25**	4.23	4.36	4.32
Full-20s	3.63	3.60	4.46	**3.61**	3.65	3.65	**3.61**	3.62	3.62
Full training - Full testing									
Full-Full	2.70	2.73	3.61	2.63	2.63	2.63	2.66	2.64	**2.61**

Table 7: Results in EER of fusion of GMM-UBM and i-vector based system on NIST 2010 telephone-telephone evaluation corpora

Train-Test duration	i-vec GPLDA	linear fusion	Q_{tv}	Q_1 kl-1	Q_2 kl-2	Q_3 kl-avg	Q_4 ℓ_1	Q_5 ℓ_2	Q_6 bh
Truncated training - Truncated testing									
2s-2s	37.67	33.52	34.59	33.71	33.42	33.66	32.91	33.28	**32.84**
5s-2s	32.01	31.44	29.13	27.44	**27.53**	**27.47**	27.76	28.13	27.97
5s-5s	25.95	23.93	24.17	**21.62**	22.09	21.81	21.92	22.37	22.37
10s-2s	28.61	27.76	24.95	24.92	**24.65**	24.65	24.54	24.92	24.92
10s-5s	20.52	18.13	17.01	17.56	16.93	16.84	16.84	**16.82**	**16.99**
10s-10s	14.44	14.16	13.88	13.59	**13.31**	13.59	13.59	13.99	13.94
20s-2s	24.44	21.81	21.65	21.24	21.74	21.55	**21.12**	21.51	21.50
20s-5s	16.94	14.44	15.45	14.44	14.44	14.57	14.44	14.73	**14.16**
20s-10s	11.28	10.19	10.08	10.07	**9.91**	10.02	10.33	10.19	10.04
20s-20s	8.78	8.03	9.87	8.49	8.47	8.44	8.49	**8.21**	8.30
Full training - Truncated testing									
Full-2s	21.81	19.18	18.41	**17.96**	18.01	17.99	18.13	18.53	18.13
Full-5s	12.46	11.89	13.79	11.61	11.84	11.70	11.89	**11.44**	11.99
Full-10s	7.36	7.97	13.31	7.36	7.64	7.41	7.88	**7.08**	7.59
Full-20s	5.38	5.5	7.24	**5.38**	5.44	**5.38**	5.58	5.59	5.59
Full training - Full testing									
Full-Full	2.90	2.95	5.26	3.11	3.11	3.11	3.39	3.11	3.11

The performance of ASV system with proposed quality metrics improved irrespective of development and evaluation corpora and utterance duration condition as well. It showed up-to 12% relative improvement over the linear fusion (GMM-UBM+i-vector) based ASV system in conditions like *full-10s, 10s-2s, 5s-2s* etc. These conditions are more close to desirable real-time requirements of ASV systems which encourages to find implementations of proposed system. Indication of similar improvements both in development and evaluation corpus, shown in Tables 5, 6, 7 and 8 respectively, authenticates generality of the proposed system. Consistent improvement of accuracy of the ASV system in various duration and databases established relevance of the proposed quality measures based on intermediate statistics. The system is more suitable when the duration speech utterances are limited, especially when it is trained with long enrollment data and tested with very short duration of speech.

6.3. Comparison with uncertainty based quality metric

A comparison of performance of the proposed quality metrics and an i-vector uncertainty based metric Q_{tv} as in [52], is accommodated in the aforementioned Tables for different duration condition. The quality measure reflects the duration variability in data as the main source of uncertainty in i-vectors since it has a high correlation with utterance duration. The posterior distribution of i-vector $\mathbf{y}$ is Gaussian with the following covariance matrix[65]

$$\mathbf{y}_\Sigma = (\mathbf{I} + \mathbf{\Phi}^\top \mathbf{\Sigma}^{-1} \mathbf{N} \mathbf{\Phi})^{-\mathbf{1}} \tag{16}$$

Here the quality measure $Q_{tv}(\mathbf{y}_\Sigma)$ is calculated as

$$Q_{tv}(\mathbf{y}_\Sigma) = \frac{1}{\text{trace}(\mathbf{y}_\Sigma)} \tag{17}$$

Table 8: Results in DCF of fusion of GMM-UBM and i-vector based system on NIST 2010 telephone-telephone evaluation corpora

Train-Test duration	i-vec GPLDA	linear fusion	Q_{tv}	Q_1 kl-1	Q_2 kl-2	Q_3 kl-avg	Q_4 ℓ_1	Q_5 ℓ_2	Q_6 bh
Truncated training - Truncated testing									
2s-2s	9.98	9.77	9.86	**9.74**	9.77	**9.74**	9.76	9.77	9.77
5s-2s	9.74	9.72	9.31	9.16	9.14	**9.14**	9.11	9.15	9.14
5s-5s	9.01	8.10	8.68	**8.01**	8.05	8.03	8.05	8.07	8.07
10s-2s	9.50	9.32	8.66	8.59	8.54	**8.50**	8.56	8.48	8.48
10s-5s	7.67	7.17	7.14	7.20	7.21	7.14	7.20	**7.04**	**7.04**
10s-10s	6.52	6.29	6.18	5.97	6.17	6.05	6.05	**5.93**	6.19
20s-2s	9.04	8.05	8.14	7.89	7.89	**7.84**	7.97	7.94	7.88
20s-5s	6.98	6.57	6.91	6.34	6.34	6.33	**6.28**	6.34	6.35
20s-10s	5.47	4.93	4.94	4.92	4.90	4.90	4.92	**4.89**	4.92
20s-20s	4.11	3.85	4.89	4.03	3.96	3.94	**3.88**	3.94	3.97
Full training - Truncated testing									
Full-2s	8.52	7.55	7.48	7.63	7.65	7.71	7.72	**7.61**	**7.61**
Full-5s	5.47	4.92	6.15	5.14	5.13	5.14	5.19	**4.93**	5.01
Full-10s	3.69	3.67	6.00	3.65	3.62	**3.52**	3.58	3.57	3.67
Full-20s	2.70	2.71	3.61	2.74	2.75	2.73	**2.71**	2.72	**2.71**
Full training - Full testing									
Full-Full	2.01	2.00	2.67	1.96	1.96	1.97	**1.95**	1.95	1.95

Here, in the experiments, truncation of speech utterances are done in fixed durations. Hence, the duration based quality metric, as used for calibration of ASV scores in [28], becomes non-functional. The recognition performance with duration based quality metric are compared later in the experiments where the enrollment/verification segment duration are randomized.

6.4. Results of quality metrics with only i-vector based system

The proposed quality metrics can also be applied for calibration of the stand-alone classifier based ASV system. The recognition scores of i-vector based system can be calibrated using the proposed quality metrics computed from the speech utterances of the corresponding trials. We have conducted separate experiments to observe the performance of the quality metrics in single classifier based system. We have incorporated the quality metric with the recognition score of i-vector based ASV system using the Eq. 12. The performance on NIST 2008 is reported in Table 9 and 10 (in %EER and DCF $\times$ 100 respectively). Whereas, the results on NIST 2010 is presented in Table 11 and 12 (in %EER and DCF $\times$ 100 respectively). The performance metrics as depicted in Tables 9, 10, 11, 12, indicate that the quality metrics showed some improvement over the stand-alone i-vector based ASV system. The results shows marginal improvements with the long and very short duration like Full, 20sec, 2 sec etc. However, the experiments yielded better result in durations like 5 sec, 10 sec etc.

6.5. Experiments and Results in Mixed duration

The evaluation of proposed quality metrics are further extended to more challenging variable duration conditions. Both the enrollment and verification segments are truncated with random duration within a range of *2 sec - 20 sec*. Here the results of the proposed methods are compared with duration-based existing

Table 9: Results in EER of i-vector with proposed quality metrics on NIST 2008 telephone-telephone evaluation corpora

Train - Test Duration	i-vec GPLDA	Q_1	Q_2	Q_2	Q_4	Q_5	Q_6
Truncated Training - Truncated Testing							
2s - 2s	36.84	36.76	36.84	**36.38**	36.84	36.84	36.84
5s - 2s	30.48	30.66	30.42	30.49	30.54	30.53	**30.47**
5s - 5s	24.37	**24.29**	24.48	24.37	24.48	24.37	24.37
10s - 2s	26.10	26.58	26.27	26.25	26.24	**26.10**	26.19
10s - 5s	19.56	19.56	19.01	19.31	**18.99**	19.22	19.20
10s - 10s	14.98	14.64	14.75	**14.64**	14.87	14.98	14.87
20s - 2s	24.20	23.94	23.90	23.91	**23.82**	23.91	23.91
20s - 5s	17.29	16.47	16.49	16.41	16.36	**16.31**	16.59
20s - 10s	12.35	**12.14**	12.31	12.12	12.35	12.24	12.24
20s - 20s	9.72	**9.63**	9.83	9.67	9.72	9.72	9.72
Full Training - Truncated Testing							
Full - 2s	19.67	**19.44**	21.04	19.45	19.85	19.60	19.54
Full - 5s	13.50	**13.17**	13.25	13.32	13.23	13.50	13.38
Full - 10s	9.29	9.49	9.54	9.49	9.45	**9.28**	9.54
Full - 20s	7.32	7.38	**7.30**	7.33	7.32	7.32	7.35

Table 10: Results in DCF of i-vector with proposed quality metrics on NIST 2008 telephone-telephone evaluation corpora

Train - Test Duration	i-vector GPLDA	Q_1	Q_2	Q_2	Q_4	Q_5	Q_6
Truncated Training - Truncated Testing							
2s - 2s	9.93	9.92	9.93	9.91	9.93	9.93	9.93
5s - 2s	9.75	9.74	9.76	9.74	9.76	9.75	9.71
5s - 5s	8.65	8.72	8.70	8.71	8.67	8.65	8.64
10s - 2s	9.23	9.34	9.25	9.27	9.24	9.23	9.22
10s - 5s	7.80	7.84	7.77	7.88	7.79	7.78	7.76
10s - 10s	6.58	6.46	6.42	6.44	6.42	6.57	6.45
20s - 2s	8.77	8.89	8.86	8.96	8.83	8.77	8.78
20s - 5s	6.98	7.01	6.95	6.95	6.94	6.95	6.99
20s - 10s	5.48	5.54	5.47	5.52	5.50	5.48	5.51
20s - 20s	4.61	4.64	4.61	4.61	4.61	4.62	4.60
Full Training - Truncated Testing							
Full - 2s	7.75	7.88	7.89	7.87	7.91	7.91	7.87
Full - 5s	5.99	5.93	5.92	5.91	5.91	5.99	5.94
Full - 10s	4.50	4.38	4.35	4.34	4.34	4.50	4.42
Full - 20s	3.63	3.60	3.54	3.62	3.51	3.63	**3.62**

quality metrics as used for calibration in [28]. We have used three types of duration based quality metrics $(Q_{dur1}, Q_{dur2}, Q_{dur3})$, as shown in Table 13. The quality measures $Q_{dur1}, Q_{dur2}, Q_{dur3}$ denote function

Table 11: Results in EER of i-vector with proposed quality metrics on NIST 2010 telephone-telephone evaluation corpora

Train - Test Duration	i-vec GPLDA	Q_1	Q_2	Q_2	Q_4	Q_5	Q_6
Truncated Training - Truncated Testing							
2s - 2s	37.67	37.67	37.67	**38.41**	37.67	37.67	37.67
5s - 2s	32.01	32.17	32.35	32.57	32.57	32.01	**32.01**
5s - 5s	25.95	25.77	25.77	26.01	**25.54**	25.91	26.00
10s - 2s	28.61	28.32	28.51	28.32	28.32	**28.54**	28.39
10s - 5s	20.52	20.39	20.46	20.24	**20.24**	20.58	20.67
10s - 10s	14.44	14.34	14.16	**14.16**	14.16	14.44	14.32
20s - 2s	24.44	25.49	24.57	24.36	**24.39**	24.44	24.07
20s - 5s	16.94	16.43	16.02	16.37	16.14	**16.43**	15.86
20s - 10s	11.28	**11.22**	11.32	11.30	11.20	11.29	11.22
20s - 20s	8.21	**8.49**	8.49	8.50	8.49	8.78	8.49
Full Training - Truncated Testing							
Full - 2s	21.81	**21.24**	21.26	21.04	21.52	21.81	21.81
Full - 5s	12.46	**12.18**	12.33	12.18	12.46	12.46	12.62
Full - 10s	7.36	7.41	7.41	7.36	7.37	**7.36**	7.42
Full - 20s	5.38	5.38	**5.38**	5.38	5.38	5.38	5.38

Table 12: Results in DCF of i-vector with proposed quality metrics on NIST 2010 telephone-telephone evaluation corpora

Train - Test Duration	i-vector GPLDA	Q_1	Q_2	Q_2	Q_4	Q_5	Q_6
Truncated Training - Truncated Testing							
2s - 2s	9.98	9.99	9.99	9.99	9.99	**9.98**	9.99
5s - 2s	9.74	9.73	**9.72**	9.74	9..75	9.74	9.71
5s - 5s	9.01	**8.98**	9.02	8.99	9.02	9.01	9.01
10s - 2s	9.50	9.72	9.52	9.53	9.51	9.51	9.52
10s - 5s	7.67	7.89	7.73	7.93	7.86	7.65	7.73
10s - 10s	6.52	6.59	6.54	6.57	6.53	6.52	6.60
20s - 2s	9.04	**9.02**	**9.02**	9.17	**9.02**	9.04	9.03
20s - 5s	6.98	7.07	7.01	7.04	6.99	**6.97**	7.03
20s - 10s	5.47	**5.43**	5.44	5.46	5.46	5.46	5.44
20s - 20s	4.11	4.12	4.12	4.11	**4.10**	4.11	4.12
Full Training - Truncated Testing							
Full - 2s	8.52	8.58	8.62	8.63	8.57	8.53	**8.41**
Full - 5s	5.47	5.37	5.39	5.44	5.51	5.47	**5.32**
Full - 10s	3.69	3.76	3.82	3.80	3.86	3.70	3.75
Full - 20s	2.71	2.72	2.73	**2.70**	2.71	**2.70**	2.73

related to the duration of model segment, d_m and the duration of test segment, d_t. The value of d_c is kept at 20 sec for the experiments. The fusion parameters (α, β, θ) as shown in equation 11 are estimated from the

Table 13: Mathematical expressions of duration based quality measures as proposed in [28]. Here, d_m and d_t denote the duration of enrolment and test segment where d_c and k are additional parameters.

Quality Measure	Additional parameter
$Q_{dur1}(\tilde{N}_s) = k\lvert \log \frac{d_m}{d_t} \rvert$	k
$Q_{dur2}(\tilde{N}_s) = k \log^2 \frac{d_m}{d_t}$	k
$Q_{dur3}(\tilde{N}_s) = k \log \frac{d_m}{d_c} \log \frac{d_c}{d_t}$	k, d_c

Table 14: Comparison of performance of i-vector and GMM-UBM and their linear fusion based ASV system in NIST 2008 telephone-telephone evaluation corpora with Randomized short duration varying between *(2 sec-20 sec)*.

Perf. Metric	i-vec GPLDA	GMM-UBM	Linear Fusion i-vec+GMM-UBM
NIST SRE 2008			
EER	17.41	27.68	17.24
DCF	6.83	8.67	6.71
NIST SRE 2010			
EER	17.93	27.64	17.28
DCF	7.09	8.40	6.94

development set using NIST SRE 2008 and applied for the evaluation with NIST SRE 2010. Table 15 and 14 present the results on this randomized duration condition for both NIST 2008 and NIST 2010 corpora. In Table 14, we have compared the three baseline systems namely i-vector (GPLDA), GMM-UBM and linear fusion of the two in randomized duration condition for both train and test utterances. However, in Table 15, the performance of the existing duration based and uncertainty based quality metrics are compared with proposed quality metrics, taking linear fusion of i-vector and GMM-UBM system as baseline. The results show a consistent improvement for the proposed methods. Further, the proposed methods outperform the duration-based and uncertainty based quality measures in both databases.

7. Conclusion and Future Scopes

In this work, we have introduced new quality measures for improving the speaker recognition performance under short duration conditions. We derive the quality measures using Baum-Welch sufficient statistics which are used for computing i-vector representation. We demonstrate that the dissimilarity between the normalized zero-order Baum-Welch statistics and the weights of universal background model (UBM) is associated with the speech duration. We formulate the quality measures based on the normalized zero-order Baum-Welch statistics and UBM weights. This quality measure estimation method does not require additional parameter estimation as they directly derived from already estimated parameters. The proposed quality measures of speech are incorporated as side information in fusion-based ASV system with i-vector and *Gaussian mixture model-Universal background model* (GMM-UBM) system as two subsystems. The score fusion with proposed quality measures substantially enhanced the ASV performance, especially for short utterance. We observed up to 12.63% relative improvement over linear fusion based ASV system. The performance is also considerably better than the performance obtained with existing speech duration and uncertainty based quality measures. Even though we have observed considerable improvement with the distance-based proposed quality measures, we do not observe any clear indication on which quality measure

Table 15: Comparison of performance of proposed Quality metrics in NIST 2008 telephone-telephone evaluation corpora with randomized short duration varying between *(2 sec-20 sec)*.

	Baseline	Existing Quality Metrics				Proposed Quality Metrics					
	Linear Fusion	Duration Based			Uncertainty Based	BW Statistics Based					
Perf. Metric	i-vec + GMM-UBM	Q_{dur1}	Q_{dur2}	Q_{dur3}	Q_{tv}	Q_1 kl-1	Q_2 kl-2	Q_3 kl-avg	Q_4 ℓ_1	Q_5 ℓ_2	Q_6 bh
NIST SRE 2008											
EER	17.24	17.03	17.28	17.04	24.02	16.57	16.47	16.59	**16.36**	17.04	16.80
DCF	6.71	6.59	6.73	6.59	8.59	**6.36**	6.45	6.40	6.39	6.56	6.51
NIST SRE 2010											
EER	17.28	16.14	15.86	**15.58**	25.77	**15.58**	**15.58**	**15.58**	**15.58**	15.83	15.93
DCF	6.94	6.77	6.95	6.52	8.81	6.52	6.49	**6.46**	6.50	6.53	6.59

distance function is most appropriate. This also opens up the possibility of further optimizing the distance measures for quality estimation.

In this work, we have considered GMM-based i-vectors. As an extension of this study, similar investigations on quality measure fusion can be made with the latest DNN-ASR-based i-vector system. The proposed method can also be explored for ASV system fusions including x-vector system as a subsystem. Moreover, the distance-based quality measures implicitly represent the acoustic variations in the speech utterance with respect to the mean of UBM or the distribution of acoustic space. Therefore, it would be interesting to explore the general use case of the proposed quality measures where acoustic variability needs to be computed.

8. Acknowledgment

This work is partially supported by Indian Space Research Organization (ISRO), Government Of India. The work of Md Sahidullah is supported by Region Grand Est, France. The authors would like to express their sincere thanks to the anonymous reviewers and the editors for their comments and suggestions, which greatly improved the work in quality and content. We further thank Dr. Tomi Kinnunen (University of Eastern Finland) for his valuable comments on an the earlier version of this work. Finally, the authors would also like to acknowledge Dr. Monisankha Pal (Signal Analysis and Interpretation Laboratory, University of Southern California), Shefali Waldekar (ABSP lab, Dept. of E & ECE, IIT Kharagpur), and Dipannita Podder (VIP Lab, Dept. of CSE, IIT Kharagpur) for their helpful suggestions in different stages of this work.

References

[1] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: From features to supervectors, Speech Communication 52 (1) (2010) 12–40.

[2] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. Bonastre, D. Matrouf, Forensic speaker recognition, IEEE Signal Processing Magazine 26 (2) (2009) 95–103.

[3] ICICI bank introduces voice recognition for biometric authentication, `https://www.icicibank.com/aboutus/article.page?identifier=news-icici-bank-introduces-voice-recognition-for-biometric-authentication-20152505124050634`, accessed: 2019-01-25.

[4] Death of the password?, `https://www.barclayscorporate.com/insight-and-research/fraud-smart-centre/biometrics.html`, accessed: 2019-01-25.

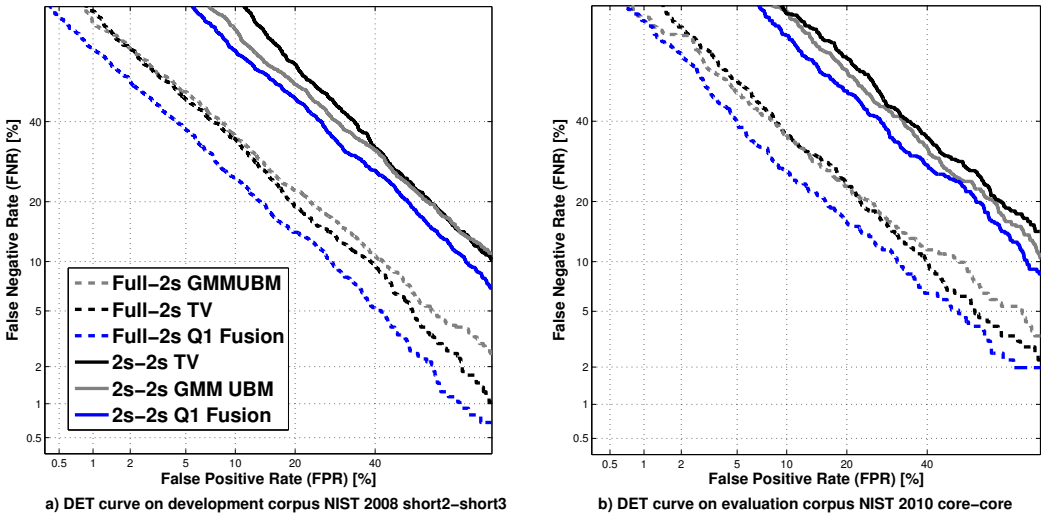

Figure 7: DET plot of *Full duration training-2 sec testing* and *2 sec training-2 sec testing 2 sec* on NIST 2008 and NIST 2010 corpora

[5] Interpols new software will recognize criminals by their voices, `https://spectrum.ieee.org/tech-talk/consumer-electronics/audiovideo/interpols-new-automated-platform-will-recognize-criminals-by-their-voice`, accessed: 2019-01-25.

[6] J. P. Campbell Jr, Speaker recognition: A tutorial, Proceedings of the IEEE 85 (9) (1997) 1437–1462.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 19 (4) (2011) 788–798.

[8] P. Matějka, O. Glembek, O. Novotnỳ, O. Plchot, F. Grézl, L. Burget, J. H. Cernockỳ, Analysis of dnn approaches to speaker identification, in: Proc. ICASSP, IEEE, 2016, pp. 5100–5104.

[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN embeddings for speaker recognition, in: Proc. ICASSP, 2018.

[10] A. Poddar, M. Sahidullah, G. Saha, Speaker verification with short utterances: a review of challenges, trends and opportunities, IET Biometrics 7 (3) (2018) 91–101.

[11] Y. A. Solewicz, G. Jardine, T. Becker, S. Gfroerer, Estimated intra-speaker variability boundaries in forensic speaker recognition casework, Proceedings of Biometric Technologies in Forensic Science (BTFS)(Nijmegen) (2013) 31–33.

[12] J. Ming, T. J. Hazen, J. R. Glass, D. A. Reynolds, Robust speaker recognition in noisy conditions, IEEE Transactions on Audio, Speech, and Language Processing 15 (5) (2007) 1711–1723.

[13] R. Saeidi, P. Alku, T. Bäckström, Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch, IEEE/ACM Transactions on Audio, Speech and Language Processing 24 (1) (2016) 42–53.

[14] M. McLaren, L. Ferrer, A. Lawson, Exploring the role of phonetic bottleneck features for speaker and language recognition, in: Proc. ICASSP, IEEE, 2016, pp. 5575–5579.

[15] S. Parthasarathy, C. Zhang, J. H. Hansen, C. Busso, A study of speaker verification performance with expressive speech, in: Proc. ICASSP, IEEE, 2017, pp. 5540–5544.

[16] D. Wang, Y. Zou, J. Liu, Y. Huang, A robust DBN-vector based speaker verification system under channel mismatch conditions, in: 2016 IEEE International Conference on Digital Signal Processing (DSP), IEEE, 2016, pp. 94–98.

[17] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, T. Kinnunen, Time-varying autoregressions for speaker verification in reverberant conditions, in: Proc. INTERSPEECH, 2017, pp. 1512–1516.

[18] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, M. W. Mason, I-vector based speaker recognition on short utterances, in: Proc. INTERSPEECH, International Speech Communication Association (ISCA), 2011, pp. 2341–2344.

[19] M. I. Mandasari, M. McLaren, D. A. van Leeuwen, Evaluation of i-vector speaker recognition systems for forensic application, in: Proc. INTERSPEECH, 2011, pp. 21–24.

[20] A. Kanagasundaram, R. J. Vogt, D. B. Dean, S. Sridharan, PLDA based speaker recognition on short utterances, in: Proc. Odyssey: The Speaker and Language Recognition Workshop, ISCA, 2012.

[21] A. K. Sarkar, D. Matrouf, P.-M. Bousquet, J.-F. Bonastre, Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification, in: Proc. INTERSPEECH, 2012.

[22] B. G. Fauve, N. W. Evans, N. Pearson, J.-F. Bonastre, J. S. Mason, Influence of task duration in text-independent speaker verification, in: Proc. INTERSPEECH, 2007, pp. 794–797.

[23] L. Ferrer, H. Bratt, V. R. Gadde, S. S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, A. Venkataraman, Modeling duration patterns for speaker recognition, in: Proc. EUROSPEECH, 2003.

[24] B. Fauve, N. Evans, J. Mason, Improving the performance of text-independent short duration SVM-and GMM-based speaker verification, in: Proc. Odyssey: The Speaker and Language Recognition Workshop, 2008, p. 18.
[25] T. Hasan, R. Saeidi, J. H. Hansen, D. van Leeuwen, Duration mismatch compensation for i-vector based speaker recognition systems, in: Proc. ICASSP, IEEE, 2013, pp. 7663–7667.
[26] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, D. Ramos, Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques, Speech Communication 59 (2014) 69–82.
[27] M. I. Mandasari, R. Saeidi, D. A. van Leeuwen, Quality measures based calibration with duration and noise dependency for speaker recognition, Speech Communication 72 (2015) 126–137.
[28] M. I. Mandasari, R. Saeidi, M. McLaren, D. A. van Leeuwen, Quality measure functions for calibration of speaker recognition systems in various duration conditions, IEEE Transactions on Audio, Speech, and Language Processing 21 (11) (2013) 2425–2438.
[29] L. Li, D. Wang, C. Zhang, T. F. Zheng, Improving short utterance speaker recognition by modeling speech unit classes, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (6) (2016) 1129–1139.
[30] C. Zhang, K. Koishida, J. Hansen, Text-independent speaker verification based on triplet convolutional neural network embeddings, IEEE/ACM Transactions on Audio, Speech and Language Processing 26 (9) (2018) 1633–1644.
[31] J. Guo, N. Xu, K. Qian, Y. Shi, K. Xu, Y. Wu, A. Alwan, Deep neural network based i-vector mapping for speaker verification using short utterances, Speech Communication 105 (2018) 92–102.
[32] N. Poh, J. Kittler, T. Bourlai, Quality-based score normalization with device qualitative information for multimodal biometric fusion, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 40 (3) (2010) 539–554.
[33] N. Poh, S. Bengio, Improving fusion with margin-derived confidence in biometric authentication tasks, in: International Conference on Audio-and Video-Based Biometric Person Authentication, Springer, 2005, pp. 474–483.
[34] N. Poh, T. Bourlai, J. Kittler, A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms, Pattern Recognition 43 (3) (2010) 1094–1105.
[35] N. Poh, J. Kittler, A unified framework for biometric expert fusion incorporating quality measures, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (1) (2012) 3–18.
[36] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, J. Bigun, Discriminative multimodal biometric authentication based on quality measures, Pattern recognition 38 (5) (2005) 777–779.
[37] P. Grother, E. Tabassi, Performance of biometric quality measures, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (4) (2007) 531–543.
[38] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, J. Ortega-Garcia, Using quality measures for multilevel speaker recognition, Computer Speech & Language 20 (2-3) (2006) 192–209.
[39] D. Garcia-Romero, J. Fiérrez-Aguilar, J. Gonzalez-Rodriguez, J. Ortega-Garcia, On the use of quality measures for text-independent speaker recognition, in: Proc. Odyssey: The Speaker and Language Recognition Workshop, 2004.
[40] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, J. H. Hansen, CRSS systems for 2012 NIST speaker recognition evaluation, in: Proc. ICASSP, 2013, pp. 6783–6787.
[41] A. Harriero, D. Ramos, J. Gonzalez-Rodriguez, J. Fierrez, Analysis of the utility of classical and novel speech quality measures for speaker verification, in: International Conference on Biometrics, Springer, 2009, pp. 434–442.
[42] F. Alonso-Fernandez, J. Fierrez, J. Ortega-Garcia, Quality measures in biometric systems, IEEE Security & Privacy 10 (6) (2012) 52–62.
[43] C. C. Chibelushi, F. Deravi, J. S. Mason, A review of speech-based bimodal recognition, IEEE Transactions on Multimedia 4 (1) (2002) 23–37.
[44] J. Kittler, N. Poh, O. Fatukasi, K. Messer, K. Kryszczuk, J. Richiardi, A. Drygajlo, Quality dependent fusion of intramodal and multimodal biometric experts, in: Defense and Security Symposium, International Society for Optics and Photonics, 2007, pp. 653903–653903.
[45] A. Poddar, M. Sahidullah, G. Saha, Novel quality metric for duration variability compensation in speaker verification, in: Proc. Ninth International Conference on Advances in Pattern Recognition (ICAPR-2017), 2017.
[46] A. Poddar, M. Sahidullah, G. Saha, Improved i-vector extraction technique for speaker verification with short utterances, International Journal of Speech Technology 21 (3) (2018) 473–488.
[47] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (1) (2000) 19–41.
[48] W. M. Campbell, D. E. Sturim, D. A. Reynolds, A. Solomonoff, SVM based speaker verification using a GMM supervector kernel and NAP variability compensation, in: Proc. ICASSP, Vol. 1, IEEE, 2006, pp. I–I.
[49] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, Joint factor analysis versus eigenchannels in speaker recognition, IEEE Transactions on Audio, Speech, and Language Processing 15 (4) (2007) 1435–1447.
[50] P. Kenny, Bayesian speaker verification with heavy-tailed priors, in: Proc. Odyssey: The Speaker and Language Recognition Workshop, 2010, p. 14.
[51] W. Li, T. Fu, H. You, J. Zhu, N. Chen, Feature sparsity analysis for i-vector based speaker verification, Speech Communication 80 (2016) 60–70.
[52] A. H. Poorjam, R. Saeidi, T. Kinnunen, V. Hautamäki, Incorporating uncertainty as a quality measure in i-vector based language recognition, in: Proc. Odyssey: The Speaker and Language Recognition Workshop, ISCA, 2016, pp. 74–80.
[53] L. Ferrer, M. K. Sönmez, S. S. Kajarekar, Class-dependent score combination for speaker recognition, in: Proc. INTERSPEECH, 2005, pp. 2173–2176.
[54] G. R. Doddington, M. A. Przybocki, A. F. Martin, D. A. Reynolds, The NIST speaker recognition evaluation–overview,

methodology, systems, results, perspective, Speech Communication 31 (2) (2000) 225–254.
[55] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, Multimodal biometric authentication using quality signals in mobile communications, in: Proc. 12th International Conference on Image Analysis and Processing, 2003.
[56] V. Hautamaki, T. Kinnunen, F. Sedlák, K. A. Lee, B. Ma, H. Li, Sparse classifier fusion for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 21 (8) (2013) 1622–1631.
[57] BOSARIS Toolkit [software package], available at: https://sites.google.com/site/bosaristoolkit.
[58] The NIST year 2008 speaker recognition evaluation plan, tech.rep., NIST.
[59] The NIST year 2010 speaker recognition evaluation plan, tech.rep., NIST.
[60] A. Poddar, M. Sahidullah, G. Saha, Performance comparison of speaker recognition systems in presence of duration variability, in: Proc. 2015 Annual IEEE India Conference (INDICON), IEEE, 2015, pp. 1–6.
[61] S. B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech and Signal Processing 28 (4) (1980) 357–366.
[62] M. Sahidullah, G. Saha, Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition, Speech Communication 54 (4) (2012) 543–565.
[63] M. Sahidullah, G. Saha, A novel windowing technique for efficient computation of MFCC for speaker recognition, Signal Processing Letters 20 (2) (2013) 149–152.
[64] M. Sahidullah, G. Saha, Comparison of speech activity detection techniques for speaker recognition, arXiv preprint arXiv:1210.0297.
[65] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, A study of interspeaker variability in speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 16 (5) (2008) 980–988.

Arnab Poddar received his MS (by research) degree in the area of speech processing and machine learning from the Department of Electronics & Electrical Communication Engineering, Indian Institute Technology Kharagpur in 2018. He has worked in the research project entitled as *Reduction of False acceptance an rejection in non-cooperative automatic speaker recognition system*, funded by Indian Space Research Organization (ISRO). Prior to that he has worked as a research project person in the project *Development of Optical Character Recognition system on printed Indian Languages* in Computer Vision and Pattern Recognition (CVPR) Unit, Indian Statistical Institute (ISI). He is currently pursuing Ph.D. at Indian Institute of Technology Kharagpur in area of machine learning and computer vision. His research interests include speech & audio signal processing, image processing, and machine learning.

Md Sahidullah received his Ph.D. degree in the area of speech processing from the Department of Electronics & Electrical Communication Engineering, Indian Institute Technology Kharagpur in 2015. Prior to that he obtained the Bachelors of Engineering degree in Electronics and Communication Engineering from Vidyasagar University in 2004 and the Masters of Engineering degree in Computer Science and Engineering (with specialization in Embedded System) from West Bengal University of Technology in 2006. In 2007-2008, he was with Cognizant Technology Solutions India PVT Limited. In 2014-2017, he was a post-doctoral researcher with the School of Computing, University of Eastern Finland. In January 2018, he joined MULTISPEECH team, Inria, France as a post-doctoral researcher where he currently holds a starting research position. His research interest includes robust speaker recognition, voice activity detection and spoofing countermeasures. He is also a co-organizer of two *Automatic Speaker Verification Spoofing and Countermeasures Challenges*: ASVspoof 2017 and ASVspoof 2019.

Goutam Saha received his B.Tech. and Ph.D. degrees from the Department of Electronics & Electrical Communication Engineering, Indian Institute of Technology (IIT) Kharagpur, India in 1990 and 2000, respectively. In between, he served industry for about four years and obtained a five year fellowship from Council of Scientific Industrial Research, India. In 2002, he joined IIT Kharagpur as a faculty member where he is currently serving as a Professor. His research interests include analysis of audio and bio signals.