Odyssey 2016
21-24 Jun 2016, Bilbao, Spain

Incorporating uncertainty as a Quality Measure in I-Vector Based Language Recognition

Amir Hossein Poorjam [1], *Rahim Saeidi* [2], *Tomi Kinnunen* [1], *Ville Hautamäki* [1]

[1] Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Finland
[2] Department of Signal Processing and Acoustics, Aalto University, Finland

[1]{amir,tkinnu,villeh}@cs.uef.fi
[2] rahim.saeidi@aalto.fi

Abstract

State-of-the-art language recognition systems involve modeling utterances with the i-vectors. However, the uncertainty of the i-vector extraction process represented by the i-vector posterior covariance is affected by various factors such as channel mismatch, background noise, incomplete transformations and duration variability. In this paper, we propose a new quality measure based on the i-vector posterior covariance and incorporate it into the recognition process to improve the recognition accuracy. The experimental results with LRE15 database and various duration conditions show a 2.9% relative improvement in terms of average performance cost as a result of incorporating the proposed quality measure in language recognition systems.

1. Introduction

Language recognition [1] is the task of recognizing the spoken language in a speech utterance. It has applications in multilingual translation systems [2], automatic speech recognition [3], targeted advertising, forensics and biometric authentication [4].

Originally introduced for speaker verification, the so-called *i-vector* [5] provides a low-dimensional representation of a speech signal. Based on factor analysis of Gaussian mixture model (GMM) mean supervectors [5], i-vector captures variability in the GMM supervector. An estimated i-vector is a *maximum-a-posteriori* (MAP) point estimate of a latent variable given the acoustic features [5], providing access to both posterior mean and posterior covariance. I-vectors have become the *de facto* standard in speaker verification and have received attention in other tasks as well, including language recognition [6], accent recognition [7, 8] and speaker profiling [9].

Studies have shown the effectiveness of incorporating *quality measures* of speech utterances into the recognition process. These include, for instance, signal-to-noise ratio (SNR), duration, F_0 deviations [10] and signal entropy. They can be incorporated in different stages of the recognition process. In [11], duration and SNR are considered as the quality factors and included in the calibration transformation. Authors in [10] incorporate SNR, F_0 deviations and the *ITU P.563* objective speech quality assessment [12] into the score computation and score fusion phases. In [13], duration information is exploited in the i-vector post-processing functions to improve the performance of speaker recognition systems.

In recent studies, instead of using the above mentioned quality factors, uncertainty in utterance modeling is incorporated as a quality measure of an utterance into the recognition process. In [14], uncertainty in acoustic feature extraction, induced by noise and short duration is incorporated into speaker verification. In [15], uncertainty in acoustic features is propagated both into i-vector extractor and the back-end classifier.

The i-vector posterior covariance matrix, conveying potentially useful information about the uncertainty of the i-vector extraction process [16] induced by channel mismatch, background noise and short utterances, is seldom utilized for reasons of simplicity and computation. Because of this simplifying assumption, i-vectors extracted from short or noisy utterances are equally contributed in the model as those extracted from noise free and long utterances. Recently, uncertainty in i-vectors is modeled through a *probabilistic discriminant analysis* (PLDA) model [17]. Specifically, using the *uncertainty decoding* (UD) technique [16], the i-vector uncertainty is first propagated through the post-processing functions and then integrated to the PLDA model. In [18], i-vector uncertainty is modeled with PLDA using three sets of utterances of different duration (namely, 3s, 10s and 30s). The results suggest that modeling i-vector uncertainty improves the performance of language recognition systems for short segments without compromising accuracy for long utterances. In [19], the authors extended the UD method in speaker recognition by applying a so-called modified imputation technique [20] in conjunction with uncertainty decoding to modify both the input and the model.

In this paper, we employ the i-vector posterior covariance to propose a new i-vector quality measure. The proposed quality measure, defined as the inverse of the trace of the i-vector posterior covariance, is a real-valued scalar that ranges between zero and infinity. Since the i-vector posterior covariance is computed through the i-vector extraction process, calculating the proposed quality measure does not require additional processing steps. In this study, the proposed quality measure is incorporated into the language recognition at the i-vector level which provides a computational advantage over UD technique by eliminating several matrix calculations during propagating uncertainty through i-vector post-processing steps and integrating it to the PLDA model.

2. I-vector based language recognition

In this section, we describe the main components of an i-vector/PLDA-based language recognition system.

2.1. The i-vector framework

An i-vector is a low-dimensional feature vector for representing utterances of arbitrary duration. We assume that each utterance possesses a speaker- and channel-dependent GMM mean supervector, **M**, in the form [5]:

doi: 10.21437/Odyssey.2016-11

$$\mathbf{M} = \boldsymbol{\mu} + \mathbf{T}\boldsymbol{\phi}\,, \tag{1}$$

where $\boldsymbol{\mu}$ is the universal background model (UBM) mean supervector and $\mathbf{T}$ is the total variability matrix. The i-vector $\boldsymbol{\phi}$ is a low-rank latent variable with standard normal prior distribution.

The posterior distribution of $\boldsymbol{\phi}$ is Gaussian with the following mean $\boldsymbol{\phi}_\mu$ and covariance matrices $\boldsymbol{\phi}_\Sigma$ [21]:

$$\boldsymbol{\phi}_\Sigma = \left(\mathbf{I} + \sum_{c=1}^{C} N_c \mathbf{T}_c^\top \boldsymbol{\Sigma}_c^{-1} \mathbf{T}_c\right)^{-1} \tag{2}$$

$$\boldsymbol{\phi}_\mu = \boldsymbol{\phi}_\Sigma \mathbf{T}^\top \boldsymbol{\Sigma}^{-1} \mathbf{f} \tag{3}$$

where $^\top$ represents a transpose, $\mathbf{I}$ is an identity matrix, $\boldsymbol{\Sigma}_c$ is the covariance of the c^{th} Gaussian, $\boldsymbol{\Sigma}$ is a block-diagonal matrix with $\boldsymbol{\Sigma}_c$s as its entries, $\mathbf{T}_c$ is the sub-matrix of $\mathbf{T}$ corresponding to the c^{th} mixture component, $\mathbf{T} = [\mathbf{T}_1^\top,...,\mathbf{T}_C^\top]^\top$, $N_c = \sum_t \gamma_{c,t}$ is the zero-order statistics estimated for the c^{th} Gaussian component of the UBM. Finally, $\mathbf{f} = [\mathbf{f}_1^\top,...,\mathbf{f}_C^\top]^\top$ where $\mathbf{f}_c$ is the first-order statistics estimated on the c^{th} Gaussian component as:

$$\mathbf{f}_c = \sum_t (\gamma_{c,t} \boldsymbol{o}_t) - N_c \boldsymbol{\mu}_c \tag{4}$$

where $\boldsymbol{o}_t$ is the acoustic vector at time t and $\gamma_{c,t}$ is the occupation count for the c^{th} mixture component and the t^{th} frame. An efficient procedure for training $\mathbf{T}$ and for MAP adaptation of the i-vectors can be found in [21].

2.2. PLDA model

Generative and discriminative models are two general approaches for language recognition based on i-vectors. Although the reported results using discriminative methods such as multiclass logistic regression and support vector machines are comparable to those of using generative models [22] such as Gaussian and probabilistic linear discriminant analysis (PLDA) models, the generative models provide an appropriate framework to benefit from the uncertainty in the i-vector extraction process through the posterior covariance matrix of the i-vector [23].

PLDA [24], originally studied in image processing, has been very successful in speaker and language recognition. In the Gaussian PLDA framework (also known as simplified PLDA [25]), i-vector generation is modeled as:

$$\boldsymbol{\phi}_{ij} = \mathbf{m} + \mathbf{V}\mathbf{y}_i + \boldsymbol{\epsilon}_{ij}\,, \tag{5}$$

where $\boldsymbol{\phi}_{ij}$ denotes the i-vector of the j^{th} utterance in the i^{th} language and $\mathbf{m}$ is the global mean of training i-vectors. $\mathbf{V}$ is a matrix whose columns span the subspace for the latent variable $\mathbf{y}$ which (in the case of language recognition) represents the language. $\mathbf{y}$ is a vector of latent factors with standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Here, $\mathbf{m}$+$\mathbf{V}\mathbf{y}_i$ is a language-dependent term and $\boldsymbol{\epsilon}_{ij} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$ is an utterance-dependent Gaussian residual term representing the variability not captured through the latent variable $\mathbf{y}$.

Given two i-vectors $\boldsymbol{\phi}_{\text{e}}$ and $\boldsymbol{\phi}_{\text{t}}$, the enrollment (average of i-vectors of each language class) and the test i-vectors, the verification score in the PLDA framework can be computed as:

$$\begin{aligned} s &= \frac{P(\boldsymbol{\phi}_{\text{e}}, \boldsymbol{\phi}_{\text{t}}|H_{\text{s}})}{P(\boldsymbol{\phi}_{\text{e}}, \boldsymbol{\phi}_{\text{t}}|H_{\text{d}})} \\ &= \frac{\displaystyle\int_{\mathbf{y}} P(\boldsymbol{\phi}_{\text{e}}|\mathcal{M}) P(\boldsymbol{\phi}_{\text{t}}|\mathcal{M}) P(\mathbf{y}) d\mathbf{y}}{\displaystyle\int_{\mathbf{y}} P(\boldsymbol{\phi}_{\text{e}}|\mathcal{M}) P(\mathbf{y}) d\mathbf{y} \int_{\mathbf{y}} P(\boldsymbol{\phi}_{\text{t}}|\mathcal{M}) P(\mathbf{y}) d\mathbf{y}} \end{aligned} \tag{6}$$

where $\mathcal{M}$ is the trained PLDA model, H_{s} stands for the same-language hypothesis and implies that both i-vectors ($\boldsymbol{\phi}_{\text{e}}$ and $\boldsymbol{\phi}_{\text{t}}$) originate from the same language, and H_{d} is the different-language hypothesis, indicating that i-vectors originate from different languages. Given the Gaussian assumption and assuming that i-vectors are centered with respect to their global mean, the above log-likelihood ratio can be computed in closed form [26]:

$$\begin{aligned} s_{\text{PLDA}} = &\log \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\phi}_{\text{e}} \\ \boldsymbol{\phi}_{\text{t}}\end{bmatrix}; \begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda} + \mathbf{V}\mathbf{V}^\top & \mathbf{V}\mathbf{V}^\top \\ \mathbf{V}\mathbf{V}^\top & \boldsymbol{\Lambda} + \mathbf{V}\mathbf{V}^\top\end{bmatrix}\right) \\ &- \log \mathcal{N}(\boldsymbol{\phi}_{\text{e}}; \mathbf{0}, \boldsymbol{\Lambda} + \mathbf{V}\mathbf{V}^\top) \\ &- \log \mathcal{N}(\boldsymbol{\phi}_{\text{t}}; \mathbf{0}, \boldsymbol{\Lambda} + \mathbf{V}\mathbf{V}^\top) \end{aligned} \tag{7}$$

3. Proposed i-vector quality measure

Utterances with higher quality provide more reliable information for the model than those with lower quality. Thus, the contribution of each utterance in the recognition process should be controlled by a quality measure [27]. In addition to the typical quality measures such as signal-to-noise ratio (SNR), duration and F_0 deviations, recently, uncertainty of the utterance modeling has been incorporated into the recognition process [16].

Uncertainty of an i-vector, represented by its posterior covariance, is mainly affected by segment duration. The posterior covariance matrix of an i-vector extracted from a short utterance possesses larger entries compared to that of extracted from a long utterance. Taking advantage of this property, in this paper, we propose a new quality measure for an utterance at the i-vector level as:

$$Q(\boldsymbol{\phi}_\Sigma) = \frac{1}{\text{tr}(\boldsymbol{\phi}_\Sigma)}\,, \tag{8}$$

where $\text{tr}(\bullet)$ is the *trace* operator,

$$\text{tr}(\boldsymbol{\phi}_\Sigma) = \sum_{i=1}^{n} \boldsymbol{\phi}_{\Sigma_{ii}} = \sum_i \lambda_i\,, \tag{9}$$

where $\boldsymbol{\phi}_{\Sigma_{ii}}$ denotes the i^{th} entry in the main diagonal of $\boldsymbol{\phi}_\Sigma$, n is the dimension of $\boldsymbol{\phi}_\Sigma$ and λ_i is the i^{th} eigenvalue of $\boldsymbol{\phi}_\Sigma$. The *trace* operator maps the i-vector posterior covariance matrix to a single real number which represents sum of variances for individual dimensions of i-vector. The inversion operation gives the sense of quality to this number.

Fig.1 indicates a high correlation (0.98) between the proposed quality measure and utterance duration in the LRE15 database (described in Section 5.1). However, since the i-vector posterior covariance is also influenced by other factors such as background noise, channel type, incomplete transformations and the acoustic content of the utterance [19, 23], we expect that the proposed quality measure captures more information about the quality of an utterance than its duration.

Quality measures can be incorporated into the recognition process at different stages. In this paper, we study the inclusion

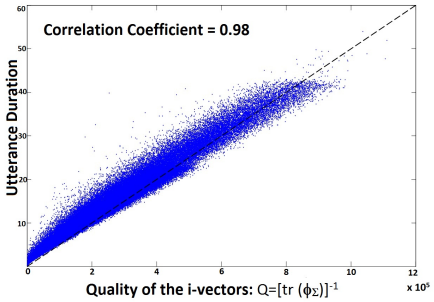

Figure 1: The correlation between the proposed quality measure and duration of the i-vectors of the data set.

of the proposed quality measure into the recognition process at the i-vector level. To this aim, the computed quality measure of the i-vector is normalized to have the same range as the other elements of the i-vectors and then, appended as an additional feature to the i-vector. The normalized quality measure is calculated as:

$$\hat{q} = \frac{(\boldsymbol{\phi}_{\mu,\max} - \boldsymbol{\phi}_{\mu,\min}) \times (q - q_{\min})}{q_{\max} - q_{\min}} + \boldsymbol{\phi}_{\mu,\min} \qquad (10)$$

where $\boldsymbol{\phi}_{\mu,\min}$ and $\boldsymbol{\phi}_{\mu,\max}$ are respectively the minimum entry and the maximum entry in all i-vectors of development set and $q_{\min}$ and $q_{\max}$ are the minimum and the maximum values of the quality measures calculated for all training data. These parameters are then used to normalized the quality measures of both the train and the test i-vectors.

4. Uncertainty Handling

4.1. Uncertainty propagation through the i-vector post-processing functions

Typical post-processing stages after i-vector extraction include linear discriminant analysis (LDA), whitening and length normalization. Usually, one first applies LDA to enhance the class separability and to reduce the dimensionality of the i-vectors. Whitening and length normalization are then used prior to PLDA training. Here, our post-processing steps include LDA followed by mean removal, whitening and length-normalization, as shown in Fig. 2. Transformation functions, often used in combination, should in principle be applied both on the i-vector posterior means and posterior covariances.

Centering i-vectors around the global mean of all training i-vectors, $\mathbf{m}$, followed by applying LDA and whitening results in the transformed i-vector posterior mean and covariance as:

$$\tilde{\boldsymbol{\phi}}_\mu = \mathbf{WV}(\boldsymbol{\phi}_\mu - \mathbf{m}) \qquad (11)$$

$$\tilde{\boldsymbol{\phi}}_\Sigma = \mathbf{WV}\boldsymbol{\phi}_\Sigma\mathbf{V}^\top\mathbf{W}^\top \qquad (12)$$

where $\mathbf{V}$ and $\mathbf{W}$ are LDA [28] and whitening [29] transformation matrices, respectively. These equations indicate that LDA and whitening are linear transforms. In contrast, length normalization [30] that maps i-vectors on the unit sphere by $\bar{\boldsymbol{\phi}} = \boldsymbol{\phi} / \parallel \boldsymbol{\phi} \parallel$ does not satisfy the Gaussian distribution and hence, the Gaussian assumption in PLDA is no longer applicable. To address this issue, one can either use a non-Gaussian assumption for the PLDA model such as the heavy-tailed PLDA model [25] or make the transformation linear using first-order Taylor series expansion around the i-vector posterior mean [23]. Applying a simplified version of first-order Taylor expansion around the i-vector posterior mean [23] results in the length normalized i-vector posterior mean and covariance as:

$$\bar{\boldsymbol{\phi}}_\mu = \frac{\mathbf{WV}(\boldsymbol{\phi}_\mu - \mathbf{m})}{\parallel \mathbf{WV}(\boldsymbol{\phi}_\mu - \mathbf{m}) \parallel} \qquad (13)$$

$$\bar{\boldsymbol{\phi}}_\Sigma = \frac{\mathbf{WV}\boldsymbol{\phi}_\Sigma\mathbf{V}^\top\mathbf{W}^\top}{\parallel \mathbf{WV}(\boldsymbol{\phi}_\mu - \mathbf{m}) \parallel^2} \qquad (14)$$

4.2. Incorporating the uncertainty into the PLDA scoring

In principle, the uncertainty in the i-vectors can be incorporated both to PLDA training and evaluation [23]. In [19], the authors made a simplifying assumption about the uncertainties of the i-vectors in training phase, namely, that the uncertainty of the training i-vectors is small because of the availability of long utterances and the number of the i-vectors per language is sufficient for reliable PLDA training. Based on this assumption, the uncertainty in training PLDA can be safely discarded and applied only on the PLDA scoring stage.

In order to take the uncertainty of the test i-vector (represented by its posterior covariance matrix $\boldsymbol{\phi}_\Sigma$) into account, the conventoinal PLDA scoring represented in Eq. (7) is modified as [19]:

$$\begin{aligned} &s^{\mathrm{UD}}_{\mathrm{PLDA}} = \\ &\log\mathcal{N}\left(\begin{bmatrix}\boldsymbol{\phi}_{\mathrm{e}}\\ \boldsymbol{\phi}_{\mathrm{t}}\end{bmatrix}; \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Lambda}+\boldsymbol{YY}^\top & \boldsymbol{YY}^\top\\ \boldsymbol{YY}^\top & \boldsymbol{\Lambda}+\boldsymbol{YY}^\top+\boldsymbol{\phi}_\Sigma\end{bmatrix}\right) \\ &-\log\mathcal{N}(\boldsymbol{\phi}_{\mathrm{e}}; \mathbf{0}, \boldsymbol{\Lambda}+\boldsymbol{YY}^\top) \\ &-\log\mathcal{N}(\boldsymbol{\phi}_{\mathrm{t}}; \mathbf{0}, \boldsymbol{\Lambda}+\boldsymbol{YY}^\top+\boldsymbol{\phi}_\Sigma) \end{aligned} \qquad (15)$$

5. Experimental setup

5.1. Database

The National Institute of Standard and Technology (NIST) has held biannual language recognition evaluations (LRE) in the past decade. With each LRE, a large corpus of telephone bandwidth broadcast radio conversations is released. Conversations typically last up to 60 seconds and originate from a large number of speakers with known language labels.

We adopt the most recent LRE15 corpus [31] for our experiments. It includes 18 target languages grouped into five language clusters as presented in Table 1. Utterances of less than 1 second long were excluded, leading to a dataset of 205839 utterances. It was further divided into three disjoint sets including 102606, 51860 and 51373 utterances for training, development and testing, respectively. The duration histograms of language clusters of the test set are illustrated in Fig 3.

To study the impact of duration on our recognizers, we built 6 test sets of different duration conditions. In the first condition, all features of utterances less than 3 seconds long were used, whereas for utterances of more than 3 seconds, we only used first 3 seconds of their features. Therefore, in the first set, there are only utterances of up to 3 seconds long. This procedure was repeated with 5s, 10s, 20s and 30s conditions. Finally, in the 6^{th} set, test utterances were not truncated and we used all features of the test segments. These six sets/conditions are labeled as $\boldsymbol{\mathcal{S}}_{3\mathrm{s}}$, $\boldsymbol{\mathcal{S}}_{5\mathrm{s}}$, $\boldsymbol{\mathcal{S}}_{10\mathrm{s}}$, $\boldsymbol{\mathcal{S}}_{20\mathrm{s}}$, $\boldsymbol{\mathcal{S}}_{30\mathrm{s}}$ and $\boldsymbol{\mathcal{S}}_{\mathrm{fulls}}$ in the rest the paper. The number of utterances are the same for all six test sets.

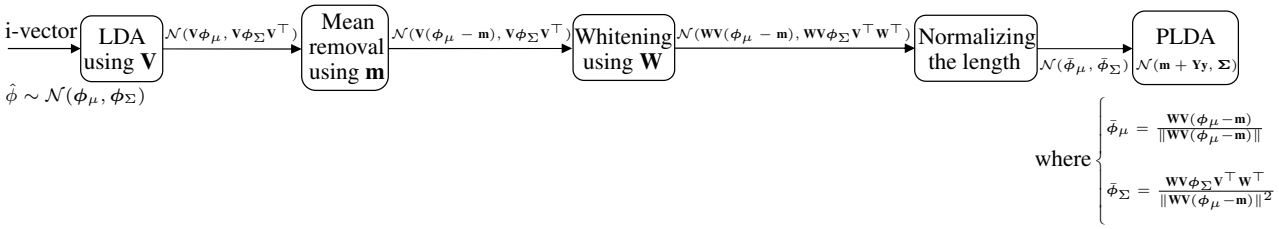

Figure 2: Uncertainty propagation through the i-vector post-processing steps.

Table 1: *Target languages and language clusters of the LRE15 database.*

Cluster	Target Languages
Arabic	Egyptian, Iraqi, Levantine, Maghrebi, Modern Standard
Chinese	Cantonese, Mandarin, Min, Wu
English	British, General American, Indian
Slavic	Polish, Russian
Iberian	Caribbean Spanish, European Spanish, Latin American Spanish, Brazilian Portuguese

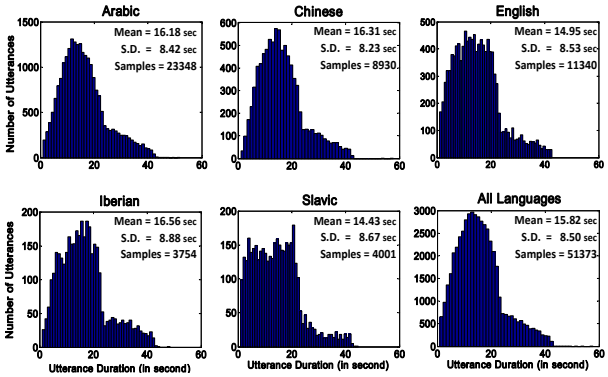

Figure 3: Duration histograms of language clusters of the test set.

5.2. Feature extraction and utterance modeling

For each utterance in the dataset, 7 MFCCs [32] and 49 shifted-delta-cepstral (SDC) features [33] have been extracted. To extract SDC-MFCC features, a 20ms hamming window shifted by 10ms is used. The SDC parameters (N-d-P-k) are configured as 7-1-3-7. These two feature sets are then concatenated into 56-dimensional SDC-MFCC feature vectors. After feature extraction, a standard energy-based speech activity detection (SAD) is applied to remove frames detected as silence or noise. A frame is decided to be speech if 70% of a frame content is deemed speech by the energy-based SAD. Finally, global cepstral mean and variance normalization [34] is applied to the features to suppress linear channel effects. The variable-duration feature vector sequences are then transformed into 400-dimensional i-vectors based on GMMs with 1024 mixture components trained on features form all training utterances.

5.3. I-vector pre-processing

The configuration of language recognition system in this study is illustrated in Fig. 4. Prior to PLDA evaluation, linear discriminant analysis (LDA) is applied on the i-vectors. Since there are 18 language classes in the database, the dimensionality of the resulting i-vectors after LDA is 17. After mean removal, i-vectors are whitened and length-normalized as described in Section 4.1.

5.4. Calibration and detection scores

The obtained scores from the PLDA back-end are calibrated using a multiclass logistic regression [35] trained on the scores from development data. By denoting the PLDA log-likelihood of trial t given the language l by $\log p(t|l)$, the calibrated language log-likelihood is:

$$\log \hat{p}(t|l) = \alpha \log p(t|l) + \beta_l \,, \tag{16}$$

where α is a weighting coefficient and β_l is a language-dependent translation vector of dimension equal to the number of language classes. We perform calibration using the FoCal Multiclass toolkit [36].

Finally, by normalizing each language likelihood with respect to the other language likelihoods, the calibrated scores are transformed to log-likelihood ratios (LLRs). After this transformation, a decision about the estimated languages can be made by comparing the LLRs to a threshold of 0.

5.5. Performance metrics

To evaluate the effectiveness of the proposed method, equal-error-rate (EER) and the average cost performance (C_{avg}) are used. EER is a value on the receiver operating characteristic (ROC) curve where the probabilities of both false acceptance (P_{FA}) and false rejection (P_{miss}) are equal. C_{avg}, in turn, is computed as:

$$\begin{aligned} C_{\text{avg}} = & \frac{C_{\text{miss}} \times P_{\text{tar}}}{R} \times \sum_{L_{\text{t}}} P_{\text{miss}}(u_{\text{t}}) \\ & + \frac{C_{\text{FA}} \times (1 - P_{\text{tar}})}{R(R-1)} \times \sum_{u_{\text{t}}} \sum_{u_{\text{n}}} P_{\text{FA}}(u_{\text{t}}, u_{\text{n}}) \end{aligned} \tag{17}$$

where u_{t} and u_{n} are, respectively, the target and non-target languages and R is the number of languages. $C_{\text{miss}} = C_{\text{FA}} = 1$ and $P_{\text{tar}} = 0.5$ are the application model parameters defined for LRE15 [31]. Since C_{avg} has the sense of cost, the lower value represents a better recognizer.

6. Results and discussion

6.1. Baseline System

In this study, we take the Gaussian PLDA [30] as the baseline for our experiments. The results of the PLDA system for different language clusters and under different duration conditions

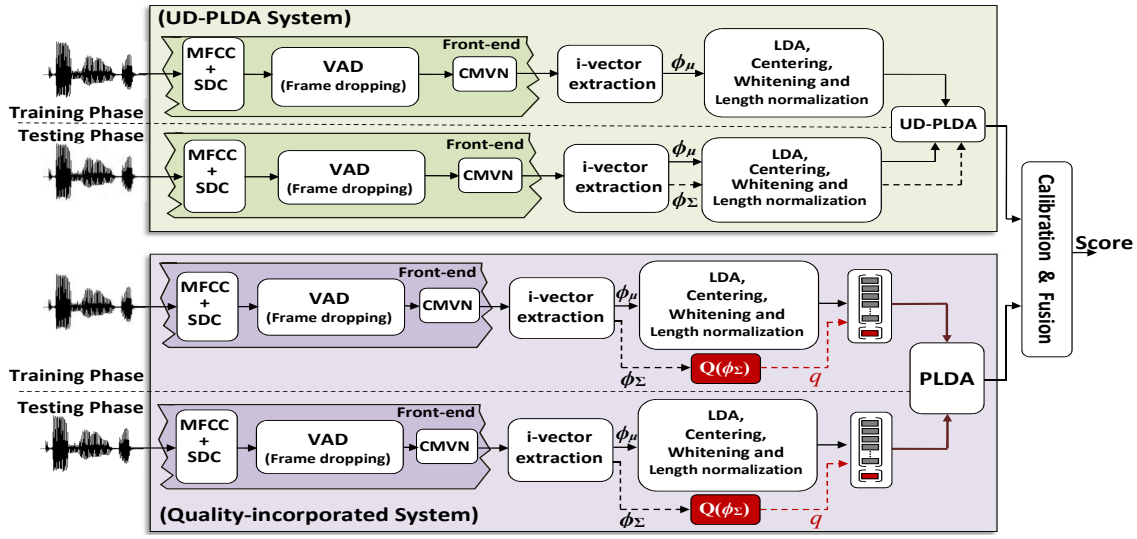

Figure 4: The block diagram of the proposed language recognition system in training and testing phases. UD stands for uncertainty decoding and $Q(\phi_\Sigma)$ is defined in Eq. (8).

are presented in Table 2. As expected, recognition accuracy dramatically degrades when the model is evaluated using short-duration utterances.

6.2. Using uncertainty decoding

Table 2 presents the results of using uncertainty decoding (UD) technique to account for the uncertainty of i-vectors (labeled as UD-PLDA in the table). We find that exploiting the uncertainty of the i-vector extraction process in the PLDA framework enhances the recognition accuracy in most cases.

6.3. Augmenting i-vectors

We consider the quality of the i-vector as an additional feature and append it to the i-vector after post-processing steps performed. The results of the quality-incorporated system (QI-PLDA) under different test conditions are presented in Table 2. Comparing the results of the QI-PLDA and the conventional PLDA reveals that providing quality information of the extracted i-vectors as an additional feature improves the performance. The results indicate that QI-PLDA system performs better than UD-PLDA system in terms of C_{avg} and EER in almost all language clusters and under various test conditions. This has also a computational advantage over UD-PLDA since the QI-PLDA avoids doing several matrix calculations to account for i-vector uncertainty.

Both the histograms of the normalized quality measure (Fig. 5) and the correlation between the i-vector quality measure and utterance duration (Fig. 1) indicate that the proposed quality measure can reflect the duration variability in data as the main source of uncertainty in i-vectors. The histograms of $Q(\phi_\Sigma)$ in Fig. 5 also reveals that increasing the duration of an utterance does not necessarily result in increasing the respective quality measure. Apart from the normalization effect inside $Q(\phi_\Sigma)$, the number of utterances with low $Q(\phi_\Sigma)$ increases by moving towards the full duration i-vectors. Such behavior of $Q(\phi_\Sigma)$ suggests that in some cases, having longer utterance duration produces larger i-vector uncertainty.

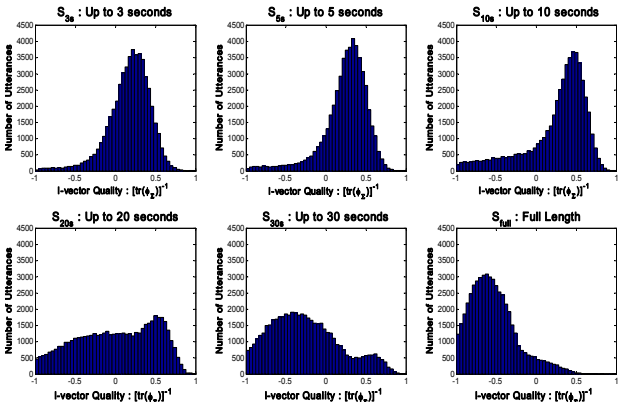

Figure 5: Histograms of the normalized i-vector quality measure for various duration conditions.

6.4. Fusion of UD-PLDA and quality incorporated systems

Following the improvements in previous experiments, we also evaluate a fusion of UD-PLDA and QI-PLDA systems at score level. To perform score fusion, scores of the development set computed by UD-PLDA and QI-PLDA systems along with the corresponding language labels are used to train a logistic regression. Then, the fused scores are obtained using the trained logistic regression. In this study, score fusion is performed using the FoCal Multiclass toolkit [36].

We find that fusion of UD-PLDA and QI-PLDA enhance the recognition accuracy under various test conditions compared to the conventional PLDA. Whereas, the improvement as a result of applying the fusion scheme compared to the UD-PLDA system is observed only for long-duration utterances (i.e. for $\boldsymbol{\mathcal{S}}_{10\text{s}}$, $\boldsymbol{\mathcal{S}}_{20\text{s}}$, $\boldsymbol{\mathcal{S}}_{30\text{s}}$ and $\boldsymbol{\mathcal{S}}_{\text{full}}$ sets).

6.5. Duration as a feature

Instead of augmenting i-vectors with estimated quality $Q(\phi_\Sigma)$, it is informative to use normalized utterance duration as a mea-

Table 2: *Language recognition accuracy in terms of C_{avg} and EER for different systems under six duration conditions.*

System	C_{avg}						EER					
	$\boldsymbol{\mathcal{S}}_{3\text{s}}$	$\boldsymbol{\mathcal{S}}_{5\text{s}}$	$\boldsymbol{\mathcal{S}}_{10\text{s}}$	$\boldsymbol{\mathcal{S}}_{20\text{s}}$	$\boldsymbol{\mathcal{S}}_{30\text{s}}$	$\boldsymbol{\mathcal{S}}_{\text{full}}$	$\boldsymbol{\mathcal{S}}_{3\text{s}}$	$\boldsymbol{\mathcal{S}}_{5\text{s}}$	$\boldsymbol{\mathcal{S}}_{10\text{s}}$	$\boldsymbol{\mathcal{S}}_{20\text{s}}$	$\boldsymbol{\mathcal{S}}_{30\text{s}}$	$\boldsymbol{\mathcal{S}}_{\text{full}}$
	Arabic											
Baseline (PLDA)	12.75	8.47	5.29	4.33	4.20	4.20	12.83	8.51	5.44	4.50	4.37	4.37
UD-PLDA	12.72	**8.41**	5.28	4.30	4.17	4.17	12.82	8.45	5.43	4.45	4.33	4.33
QI-PLDA	12.78	8.42	**5.21**	**4.18**	**4.11**	**4.05**	12.93	**8.44**	**5.26**	**4.26**	**4.18**	**4.12**
QI-PLDA & UD-PLDA	**12.71**	8.42	5.28	4.28	4.16	4.16	**12.81**	8.45	5.43	4.45	4.32	4.32
	Chinese											
Baseline (PLDA)	15.26	9.60	4.71	3.63	3.59	3.58	15.67	9.86	5.10	3.74	3.70	3.70
UD-PLDA	15.21	9.59	4.81	3.68	3.63	3.60	**15.62**	9.86	5.12	3.81	3.65	3.64
QI-PLDA	**15.19**	**9.45**	**4.69**	**3.49**	**3.47**	**3.43**	15.63	**9.48**	**4.92**	**3.56**	**3.55**	**3.50**
QI-PLDA & UD-PLDA	15.24	9.61	4.76	3.66	3.63	3.59	15.66	9.93	5.02	3.76	3.73	3.64
	English											
Baseline (PLDA)	7.73	5.00	3.17	2.32	2.33	2.32	16.08	5.04	3.23	2.45	2.44	2.47
UD-PLDA	7.84	4.94	3.07	2.22	2.19	2.19	16.22	4.97	3.21	2.42	2.32	2.31
QI-PLDA	**7.61**	**4.66**	3.10	2.21	**2.11**	**2.01**	**15.95**	**4.70**	3.19	**2.25**	**2.13**	**2.08**
QI-PLDA & UD-PLDA	7.86	4.92	**3.00**	**2.15**	2.16	2.17	16.19	4.94	**3.09**	2.30	2.30	2.30
	Slavic											
Baseline (PLDA)	20.60	14.30	9.38	7.62	7.29	7.29	20.96	14.36	9.50	7.72	7.61	7.64
UD-PLDA	20.85	14.41	9.25	7.43	7.29	7.23	21.23	14.47	9.40	7.54	7.45	7.43
QI-PLDA	**20.34**	**14.04**	**8.96**	7.39	**7.16**	**7.02**	**20.73**	**14.10**	**9.02**	7.52	7.41	**7.40**
QI-PLDA & UD-PLDA	20.93	14.51	9.20	**7.37**	7.25	7.22	21.24	14.59	9.28	**7.49**	**7.39**	7.43
	Iberian											
Baseline (PLDA)	13.59	9.23	7.24	6.73	6.58	6.68	14.01	9.66	7.36	**6.78**	6.93	6.75
UD-PLDA	13.25	9.30	7.44	6.81	6.69	6.79	**13.65**	9.64	7.60	7.05	7.09	6.97
QI-PLDA	13.54	**9.13**	**7.18**	**6.71**	**6.19**	**6.24**	13.73	**9.60**	**7.26**	6.86	**6.28**	**6.55**
QI-PLDA & UD-PLDA	**13.23**	9.32	7.33	6.81	6.69	6.78	**13.65**	9.75	7.51	7.10	7.09	6.97

sured quality in the augmentation process. Duration is measured by applying an energy-based SAD to the utterances and converting its output to the time scale (in second). Similar to the QI-PLDA, in duration-incorporated system (DI-PLDA), duration is first normalized to have the same range as the other entries of the i-vectors.

For comparison purposes, we present the results of QI-PLDA and DI-PLDA systems under various test conditions by taking average of C_{avg} over different language clusters in Table 3. Results suggest that the proposed quality measure provides more information about the quality of an utterance and consequently, is considered as a better quality measure for language recognition systems than duration. The obtained relative improvements by incorporating $Q(\phi_\Sigma)$ into the recognition process compared to the baseline system in $\boldsymbol{\mathcal{S}}_{3\text{s}}$, $\boldsymbol{\mathcal{S}}_{5\text{s}}$, $\boldsymbol{\mathcal{S}}_{10\text{s}}$, $\boldsymbol{\mathcal{S}}_{20\text{s}}$, $\boldsymbol{\mathcal{S}}_{30\text{s}}$ and $\boldsymbol{\mathcal{S}}_{\text{full}}$ conditions are 0.71%, 1.93%, 2.35%, 2.84%, 4.17% and 5.41%, respectively.

Table 3: *Comparison between baseline, quality-incorporated (QI-PLDA) and duration-incorporated (DI-PLDA) systems under six duration conditions. The results are presented by taking average of C_{avg} over different language clusters.*

System	avg (C_{avg})					
	$\boldsymbol{\mathcal{S}}_{3\text{s}}$	$\boldsymbol{\mathcal{S}}_{5\text{s}}$	$\boldsymbol{\mathcal{S}}_{10\text{s}}$	$\boldsymbol{\mathcal{S}}_{20\text{s}}$	$\boldsymbol{\mathcal{S}}_{30\text{s}}$	$\boldsymbol{\mathcal{S}}_{\text{full}}$
Baseline	13.99	9.32	5.96	4.93	4.80	4.81
DI-PLDA	13.90	9.20	5.86	4.81	4.61	4.57
QI-PLDA	**13.89**	**9.14**	**5.82**	**4.79**	**4.60**	**4.55**

7. Conclusions

In this paper, a new i-vector quality measure based on the i-vector posterior covariance has been proposed. This quality measure was incorporated into the recognition process at the i-vector level to improve the recognition accuracy. Moreover, a fusion of uncertainty decoding and quality incorporated systems at score level was investigated. The average relative improvement in C_{avg} as a result of incorporating the proposed quality measure into the recognition process (QI-PLDA) compared to the baseline system (conventional PLDA) is 2.9%.

The proposed quality measure can reflect the duration variability in data as the main source of uncertainty in i-vectors since it has a high correlation (0.98) with utterance duration. However, the behavior of this quality measure in different duration conditions indicates that in some cases, a larger quality measure is not necessarily produced by a longer utterance. Thus, more robust i-vector quality measures are required to be proposed in future work.

8. References

[1] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: From fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[2] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, 2000.

[3] F. Biadsy, *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*, Ph.D. thesis, 2011.

[4] P. L. Patrick, "Language analysis for determination of origin: Objective evidence for refugee status determination," in *The Oxford Handbook of Language and Law*, chapter 38, pp. 533–546. Oxford University Press, 2012.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[6] N. Dehak, P. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction.," in *INTERSPEECH*, 2011, pp. 857–860.

[7] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C. Lee, "i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 29–41, 2016.

[8] M. H. Bahari, R. Saeidi, H. Van hamme, and D. Van Leeuwen, "Accent recognition using i-vector, Gaussian Mean Supervector and Gaussian posterior probability supervector for spontaneous telephone speech," in *ICASSP*, 2013, pp. 7344–7348.

[9] A. H. Poorjam, M. H. Bahari, and H. Van hamme, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *Int. Conf. on Computer and Knowledge Engineering*, 2014, pp. 7–12.

[10] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Using quality measures for multilevel speaker recognition," *Computer Speech and Language*, vol. 20, pp. 192–209, 2006.

[11] M. Mandasari, R. Saeidi, M. McLaren, and 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*, vol. 21, no. 11, pp. 2425–2438, 2013.

[12] ITU-T Recommendation P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," 2004.

[13] B. Vesnicer, J. Zganec-Gros, S. Dobrisek, and V. Struc, "Incorporating Duration Information into I-Vector-Based Speaker-Recognition Systems," in *Odyssey*, 2014, pp. 241–248.

[14] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating Auditory Feature Uncertainties in Robust Speaker Identification," in *ICASSP*, 2007, vol. 4, pp. IV–277–IV–280.

[15] C. Yu, G. Liu, S. Hahm, and J. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *ICASSP*, 2014, pp. 4017–4021.

[16] P. Kenny, T. Stafylakis, P. Oullete, J. A. Md., and P. Dumouchel, "PLDA for Speaker Verification with Utterances of Arbitrary Duration," in *ICASSP*, 2013, pp. 7649–7653.

[17] T. Stafylakis, P. Kenny, P. Ouellet, and J. Perez, "Text-dependent speaker recognition using PLDA with uncertainty propagation," in *INTERSPEECH*, 2013.

[18] S. Cumani, R. Fer, and O. Plchot, "Exploiting i-vector posterior covariances for short-duration language recognition," in *INTERSPEECH*, 2015, pp. 1002–1006.

[19] R. Saeidi and P. Alku, "Accounting For Uncertainty of i-vectors in Speaker Recognition Using Uncertainty Propagation and Modified Imputation," in *INTERSPEECH*, 2015, pp. 3546–3550.

[20] B. Raj, M. L. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4 SPEC. ISS., pp. 275–296, 2004.

[21] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[22] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in iVectors space," in *INTERSPEECH*, 2011, pp. 861–864.

[23] S. Cumani, O. Plchot, and P. Laface, "On the use of ivector posterior distributions in Probabilistic Linear Discriminant Analysis," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 846–857, 2014.

[24] S. J.D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE 11th Int. Conf. on Computer Vision*, 2007, pp. 1–8.

[25] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," *Odyssey*, pp. 14–24, 2010.

[26] S. J. D. Prince, *Computer Vision: Models, Learning, and Inference*, 2012.

[27] R. Saeidi, R. Astudillo, and D. Kolossa, "Uncertain LDA: Including observation uncertainties in discriminative transforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8828, no. c, pp. 1–1, 2015.

[28] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[29] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *Statistics*, p. 12, dec 2015.

[30] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, 2011, pp. 249–252.

[31] Http://www.nist.gov/itl/iad/mig/lre15.cfm, "NIST 2015 Language Recognition Evaluation," .

[32] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[33] P. A. Torres-Carrasquillo, E. Singer, M. Kohler, R. Greene, D. A. Reynolds, and J. R. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features.," in *INTERSPEECH*, 2002.

[34] V. Prasad and S. Umesh, "Improved cepstral mean and variance normalization using Bayesian framework," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 156–161.

[35] N. Brummer and D. Van Leeuwen, "On calibration of language recognition scores," in *Odyssey*, 2006, pp. 1–8.

[36] N. Brümmer, "FoCal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition Scores: Tutorial and user manual," Tech. Rep., 2007.