DURATION MISMATCH COMPENSATION FOR I-VECTOR BASED SPEAKER RECOGNITION SYSTEMS

Taufiq Hasan[1]*, Rahim Saeidi*[2]*, John H. L. Hansen*[1] *and David A. van Leeuwen*[2]

[1]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, USA*
[2]Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands†
{taufiq.hasan,john.hansen}@utdallas.edu, {d.vanleeuwen, r.saeidi}@let.ru.nl

ABSTRACT

Speaker recognition systems trained on long duration utterances are known to perform significantly worse when short test segments are encountered. To address this mismatch, we analyze the effect of duration variability on phoneme distributions of speech utterances and i-vector length. We demonstrate that, as utterance duration is decreased, number of detected unique phonemes and i-vector length approaches zero in a logarithmic and non-linear fashion, respectively. Assuming duration variability as an additive noise in the i-vector space, we propose three different strategies for its compensation: i) multi-duration training in Probabilistic Linear Discriminant Analysis (PLDA) model, ii) score calibration using log duration as a Quality Measure Function (QMF), and iii) multi-duration PLDA training with synthesized short duration i-vectors. Experiments are designed based on the 2012 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) protocol with varying test utterance duration. Experimental results demonstrate the effectiveness of the proposed schemes on short duration test conditions, especially with the QMF calibration approach.

Index Terms— Speaker verification, short utterance, quality measure fusion (QMF), i-vector

1. INTRODUCTION

Mismatch in utterance duration is a long-standing problem in speaker recognition. In real world applications, it is very common that sufficient data is available for speaker enrollment/training, but the test utterances are very short during recognition. When speaker models are trained using sufficient amount of data, a variety of phonemes are captured from the speaker enabling the model to better represent the speaker's acoustic space while enhancing its discriminating ability. This has been especially true since Gaussian Mixture Models (GMM) were introduced for modeling the acoustic space for speaker recognition [1]. The subsequent research endeavors focused more on modeling the speaker dependent GMM mean super-vectors using various factor analysis methods, leading to techniques such as Eigenvoice [2, 3], Eigenchannel, Joint Factor Analysis (JFA) [4], and i-vector [5] based speaker recognition systems. Most of these systems were specifically designed to handle channel mismatch. The state-of-the-art systems generally utilize the i-vectors as utterance dependent features and use a Probabilistic Linear Discriminant Analysis (PLDA) [6, 7] model for classification, which are still prone to performance degradation when short duration utterances are encountered in test [8]. Recently, short duration test utterance conditions are re-introduced in the NIST SRE 2012 [9] within common test conditions, leading the research community to further concentrate on this problem.

A considerable amount of study has been done related to utterance duration mismatch for speaker recognition. Among the earlier works, in [10], this problem was addressed for vector quantization (VQ) based speaker recognition systems in a small dataset. Short duration utterances in both training and test for in-set/out-of-set speaker recognition was considered in [11, 12] using a GMM-Universal Background Model (UBM) based framework. More recent works considered short duration mismatch in GMM-SVM [13], JFA [14, 15] and i-vector system framework [8, 15, 16]. In [15], mismatch in train and test duration was compensated in the total variability (TV) training phase, demonstrating that short utterances in the hyper-parameter training provides the optimal performance when short test utterances are encountered. In [17], special attention is paid to the effect of duration on calibration.

In this paper, we systematically analyze the effect of duration on speech utterances and compensate for the mismatch using three different approaches. Firstly, by introducing short utterances in the PLDA training so that the model may learn the duration variability in the i-vector space [8]. Secondly, assuming that duration variation causes a linear shift in the speaker recognition scores, we propose a compensation strategy through score calibration. Thirdly, we propose a method of artificially generating i-vectors [18] of variable duration utterances and use them for PLDA training.

2. EFFECT OF DURATION ON SPEECH UTTERANCES

2.1. Analysis of Phoneme Distribution

To analyze the effect of truncating speech utterances on the phoneme space, we collect 19167 telephone recordings of both genders from the SRE'04,05 and 06 corpora and perform English phoneme recognition [19]. After removing the silence portions detected from the phoneme transcripts, we calculate the number of unique phonemes detected from the full utterances and also their truncated versions having durations of 2, 5, 10, 20 and 40 seconds. The average number of unique phonemes detected for different durations are plotted in Fig. 2 in a logarithmic scale, showing that this quantity reduces exponentially with duration. This partially explains why short duration utterances cause speaker recognition performances to degrade in an exponential manner [8]. Histograms of phonemes detected from a single utterance in its full duration, and short versions are plotted in Fig. 1, showing the "acoustic holes"/missing phonemes [12] observed when utterances are truncated. The effect of number of sam-

†This work is partly funded by European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 238803.

*This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

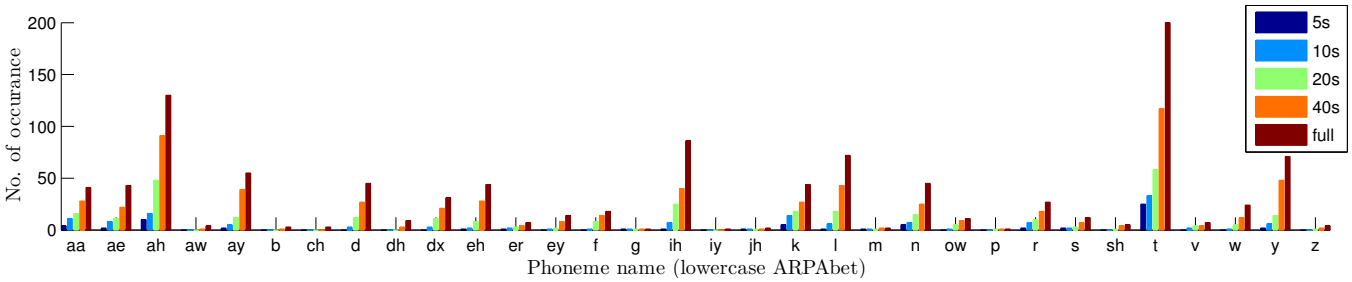

Fig. 1. Histogram of phonemes detected from an utterances in five different truncated conditions: 5 s, 10 s, 20 s, 40 s and full duration.

ples of unique phonemes and presence of unique phonemes in train and test phase of speaker recognition requires a deeper investigation.

2.2. Analysis of I-vector Length and Variance

Since i-vectors are Maximum A-posteriori (MAP) point estimates, when the utterance duration approaches zero, the i-vector approaches the zero vector. Thus, the i-vector length should gradually reduce to zero as the utterance is truncated. In this paper, an i-vector extracted from a full duration and truncated speech utterance will be referred to as a "*full* i-vector" and "*truncated* i-vector," respectively.

To observe the effect of duration on i-vector length, we extract i-vectors from the female DEV-train utterances (see Sec 3). For each speech segment, i-vectors are extracted from the full segment and four truncated versions of duration 5 s, 10 s, 20 s and 40 s. Let $\mathbf{w}_s \triangleq \mathbf{w}_{sM}$ denote a $D \times 1$ *full* i-vector extracted from utterance s, and $\mathbf{w}_{si}$ denote a *truncated* i-vector extracted from the same utterance. Here, $i \in [1, M]$ is an integer defining several fixed durations in the set $\mathcal{D} = \{d_i\}$, where M is the total number of duration values considered. In our case, $\mathcal{D} = \{5, 10, 20, 40, full\}$ and $M = 5$. The average length of truncated i-vectors for duration d_i is computed by:

$$\bar{L}_i = \frac{1}{N} \sum_{s=1}^{N} ||\mathbf{w}_{si}||^2. \tag{1}$$

Here, N is the total number of speech segments in the dataset in consideration. The average diagonal covariance is computed across all the i-th duration truncated i-vectors from all segments as:

$$\bar{\sigma}_i = \frac{1}{ND} \sum_{s=1}^{N} \text{Tr}\left((\mathbf{w}_{si} - \mathbf{w}_s)(\mathbf{w}_{si} - \mathbf{w}_s)^T\right) \tag{2}$$

Here $\text{Tr}(\cdot)$ denotes the trace operation. The values of $\bar{L}_i$ and $\bar{\sigma}_i$ for each fixed duration d_i is summarized in Table 1. From this table, it is observed that as duration is reduced, i-vector length approaches zero while the average deviation from full i-vectors is increased.

Table 1. Analysis of length and average variance of truncated i-vectors obtained from different segment durations

Measure	Duration d_i (seconds)			
	5	10	20	40
$\bar{\sigma}_i$	4.10	3.034	1.919	0.935
$\bar{L}_i$	602.7	1038.4	1567.7	2072.4

2.3. Additive Noise Model for Duration Variability

Following the observations, it is natural to model the duration variability in the i-vectors using an additive noise. In essence, if the i-vector extracted from a full duration utterance is considered "clean,"

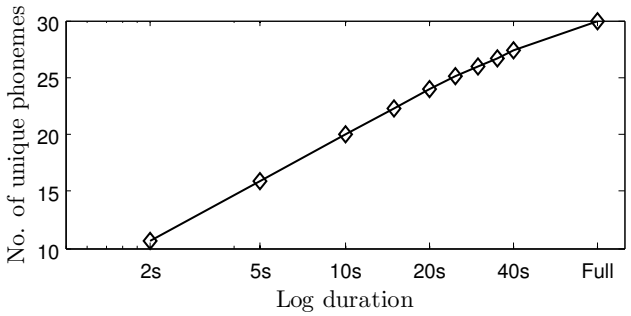

Fig. 2. Number of unique phonemes detected from varying duration utterances. The figure is obtained by averaging over 19167 utterances from NIST SRE'04,05,06 gender-mixed telephone data.

the i-vectors extracted from a truncated version of the utterance can be considered "noisy." We assume that a truncated i-vector can be represented by:

$$\mathbf{w}_{si} = \mathbf{w}_s + \epsilon_i \tag{3}$$

where $\epsilon_i \sim \mathcal{N}(0, \mathbf{\Sigma})$ and $\mathbf{\Sigma}$, represent noise, and within segment covariance matrix due to duration variability, respectively.

3. EXPERIMENTAL FRAMEWORK

The experimental framework is very similar to what we have used for the NIST SRE 2012 submission, considering the *special* conditions in SRE'12 including multiple segments training for a speaker. The latest lists from NIST are utilized and speech segments for all 1918 speakers are fetched from SRE'06, SRE'08 and SRE'10 corpora and corresponding meta-data are extracted. To be able to assess the recognition systems' generalization capability as well as calibration performance, separate development (DEV) and evaluation (EVAL) sets are prepared. Only segments having a duration greater than 40 sec are considered. Number of segments, speakers and trials for each set are given in table 2. In making these sets, the followings are considered:

1. Test segments are different for DEV-test and EVAL-test.
2. Most of the train segments in DEV-train are added to EVAL-train. The number of train segments in EVAL-train is almost twice the number of segments in DEV-train. This is done to evaluate the systems performance under the condition that speaker and channel space are already trained but number of enrollment segments for target speaker modeling has increased.
3. The segments from train and test always had different LDC-IDs so that the trials always contain a session mismatch.
4. Two disjoint sets of speakers (segments) from SRE'06 data that did not appear in SRE'12 are added to DEV-test and EVAL-test to serve as the unknown non-targets.

Table 2. Number of speakers and speech segments in the DEV and EVAL data sets. The number of known/unknown (k/u) non-target trials are also shown for the trial statistics.

Gender	Number of speakers				Number of segments				Number of trials			
	DEV		EVAL		DEV		EVAL		DEV		EVAL	
	Train	Test	Train	Test	Train	Test	Train	Test	Target	Non-target (k/u)	Target	Non-target (k/u)
Male	680	828	723	803	5475	6501	9730	7078	4801	3259879/1156000	4997	3607834/1504563
Female	1039	1182	1095	1101	8067	8460	14061	9333	6506	6753228/2030206	6770	7406380/2806485

5. For the speakers that telephone and microphone data were available, both types of channels are included in the training to make the system more robust to channel variability.

To systematically introduce the duration variability, we prepared separate versions of these tasks when the test utterances are short. For each test utterance in DEV-test and EVAL-test, four truncated segments are created having durations of 5, 10, 20 and 40 seconds. Thus, including the full utterance, there were five different versions of each segment. Using these segments, we prepared five experimental conditions with full duration training and variable duration test. It should be noted that short utterances are generated after speech activity detection (SAD) and feature extraction, retaining the exact number of feature frames required for a specified duration.

4. BASELINE SYSTEM

The speaker recognition system consists of a standard i-vector [5] configuration with PLDA modeling.

4.1. Front-end Processing

In the feature extraction we use 19 MFCC's plus log energy computed over a window of 30 ms with a skip-rate of 10 ms. After appending delta and acceleration, feature warping [20] was applied. Speech activity detection is based on Gaussian modeling of the frame energy which closely followed [21]. Since we have expected to deal with noise segments in SRE'12, a Wiener filtering based speech enhancement module was incorporated in the system front-end which applied in signal domain for both feature extraction and speech activity detection. In this way, the speech signal is first enhanced by applying the Wiener filter on the amplitude spectrum of the frames and the noise spectrum is estimated by *improved minima controlled recursive averaging* (IMCRA) approach [22, 23]. This method works based on averaging the previous estimation of the noise power spectra followed up by smoothing over consecutive frames using a forgetting factor.

4.2. UBM and TV Space Training

For each gender, a 2048-component diagonal UBM is trained, using segments from NIST SRE'04–06, Switchboard cellular phase 1 and 2, and Fisher English corpora. The Baum-Welch statistics up to second order are computed using the UBMs, which is later used to train the i-vector extractor matrix T. The matrix T with a rank of 400 has been trained using the statistics from the same utterances used for training the UBM. For each relevant utterance, a 400 dimension i-vector is extracted using the sufficient statistics and the matrix T.

4.3. I-vector Conditioning and PLDA modeling

To enhance the class separability, Linear Discriminant Analysis (LDA) is employed, reducing the i-vector dimension to 200. Finally, i-vectors are centered, whitened [24] and length-normalized [25]. The speaker and session dependent i-vector distribution is modeled using PLDA. All of the utterances in DEV-train are used for training the speaker and channel space with 200 and 50 dimensions, respectively. Each enrollment speaker is modeled by an average i-vector computed over all of the enrollment i-vectors corresponding to that speaker. No truncation is performed on enrollment data. The output scores of PLDA, which is already in the form of log-likelihood-ratio (LLR), is then converted to *calibrated LLR* using a linear calibration transformation. In the last step, the calibrated LLR are mapped to *compound LLR* described in [26].

5. PROPOSED METHOD

5.1. Training with Truncated I-vectors

In this method, i-vectors extracted from truncated DEV-train utterances are included in the PLDA training. The long files are cropped to smaller parts in a overlapping fashion. For each utterance in the dataset, 4 truncated versions are created by cropping the acoustic features and i-vectors are extracted. In this way, the number of i-vectors used in PLDA is increased by 5 times compared to the baseline setup, but the same speech data/information content is utilized.

5.2. Score Domain Compensation

We propose a QMF based calibration method for duration mismatch compensation applied on the raw recognizer scores. The calibration we use for scores $s(x, y)$ for a trial involving training speech segments x and test segment y is

$$\lambda(x, y) = w_0 + w_1 s(x, y) + Q(x, y, w_2, \ldots) \tag{4}$$

where $Q(x, y, w_2, \ldots)$ is a QMF depending on certain quality measures of the speech samples. The quality measures we used in this research are the duration of the target train segments and test segment (after SAD), $d(x)$ and $d(y)$, respectively. Inspired by Fig. 2, we used the two-parameter function

$$Q = w_2 \log d(x)/d_0 + w_3 \log d(y)/d_0. \tag{5}$$

Here d_0 is an arbitrary duration constant to keep the dimensions proper. The total number of parameters was 4 which we found by minimizing the multi-class cross entropy H_{mc} [27] over the development set. H_{mc} is defined in terms of the posterior probability of the true class, by

$$H_{\text{mc}} = \sum_{i=0}^{N} \frac{\pi_i}{N_i} \sum_{j=1}^{N_i} -\log P(i \mid x, y_j). \tag{6}$$

Here i indexes the N target speakers, using $i = 0$ for an unknown speaker, and j runs over all N_i test segments for which i is the the speaker. For the priors π_i, we were inspired by the SRE'12 core conditions, setting $\pi_0 = \frac{1}{2}$ and $\pi_{i>0} = \frac{1}{2N}$. The posterior in (6) is computed using

$$P(i \mid x, y_j) = \frac{\pi_i e^{\lambda_i(x, y_j)}}{\pi_0 + \sum_{k=1}^{N} \pi_k e^{\lambda_k(x, y_j)}}. \tag{7}$$

Table 3. Performance comparison of the proposed schemes in DEV and EVAL tasks for male trials.

PLDA	Calibration	Actual DCF'12 (C_{primary}) measures for different test durations									
		DEV					EVAL				
		5 s	10 s	20 s	40 s	full	5 s	10 s	20 s	40 s	full
Full	None	0.9971	0.8508	0.5407	0.2479	0.1046	1.012	0.8531	0.5032	0.2324	0.1208
	QMF	0.6898	0.3547	0.164	0.0791	0.0471	0.6552	0.3388	0.1977	0.1366	0.0953
Full+truncated	None	0.8273	0.5258	0.2642	0.1318	0.0836	0.8537	0.5449	0.2717	0.1594	0.1201
	QMF	**0.4981**	**0.2571**	0.1419	**0.0856**	**0.0620**	**0.4244**	**0.2163**	**0.1334**	**0.0988**	**0.0747**
Full+synthesized	None	0.7969	0.4562	0.2188	0.1230	0.0861	0.8260	0.4803	0.2455	0.1595	0.1213
	QMF	0.5638	0.2757	**0.1405**	0.0899	0.0673	0.4909	0.2424	0.1412	0.1051	0.0825

Table 4. Performance comparison of the proposed schemes in DEV and EVAL tasks for female trials.

PLDA	Calibration	Actual DCF'12 (C_{primary}) measures for different test durations									
		DEV					EVAL				
		5s	10s	20s	40s	full	5s	10s	20s	40s	full
Full	None	0.9831	0.8301	0.5552	0.3068	0.1611	1.0395	0.8706	0.5440	0.2703	0.1636
	QMF	0.7184	0.4280	0.2189	0.1256	0.0977	0.7191	0.3973	0.2149	0.1434	0.1220
Full+truncated	None	0.8234	0.5710	0.3238	0.1845	0.1293	0.8587	0.5898	0.3093	0.1789	0.1502
	QMF	**0.5515**	**0.3299**	**0.1897**	**0.1269**	**0.1059**	**0.4807**	**0.2685**	**0.1579**	**0.1110**	**0.1048**
Full+synthesized	None	0.7930	0.5033	0.2754	0.1730	0.1348	0.7249	0.3973	0.2138	0.1424	0.1209
	QMF	0.6638	0.3725	0.2088	0.1405	0.1208	0.5969	0.3087	0.1681	0.1212	0.1152

Note that we use the notation $\lambda_i(x, y_j)$ to indicate the simple likelihood ratio for test segment y_j with speaker i in the target hypothesis using all available training material x. We used a standard general numerical optimizer `nlm` from the R software package for finding the calibration parameters.

5.3. Synthesized I-vectors using Intra-segment Covariance

In this section, we propose to add artificially generated *truncated* i-vectors in PLDA modeling. This is carried out by utilizing the distributions learned from the global covariance structure of i-vectors extracted from DEV-train utterances with 4 truncated versions. Using the additive noise model proposed in (3), we can synthesize truncated i-vectors by adding Gaussian random vectors having an intra-segment covariance matrix $\mathbf{\Sigma}$ to the full duration i-vector $\mathbf{w}_s$. The procedure of obtaining the inter segment covariance matrix $\mathbf{\Sigma}$ and generation of synthesized i-vectors is provided in Alg. 1 and 2, respectively. We note that the synthesized i-vectors $\hat{\mathbf{w}}_{sj}$ do not represent a specific duration, rather they follow a global distribution of truncated i-vectors extracted from fixed duration utterances.

Initialize: $\mathbf{\Sigma} \leftarrow \mathbf{0}_{D\times D}$;
for *segment* $s \leftarrow 1$ **to** N **do**
  Compute mean: $\bar{\mathbf{w}}_s \leftarrow 1/M \sum_{i=1}^{M} \mathbf{w}_{si}$
  Update scatter: $\mathbf{\Sigma} \leftarrow \mathbf{\Sigma} + \sum_{i=1}^{M} (\mathbf{w}_{si} - \bar{\mathbf{w}}_s)(\mathbf{w}_{si} - \bar{\mathbf{w}}_s)^T$
end
Force symmetry: $\mathbf{\Sigma} \leftarrow (\mathbf{\Sigma}^T + \mathbf{\Sigma})/2$
Normalize: $\mathbf{\Sigma} \leftarrow \frac{1}{N}\mathbf{\Sigma}$

Algorithm 1: Estimation of i-vector scatter matrix

6. RESULTS

The results are summarized in Table 3 and 4 for male and female trials, respectively. We use the NIST SRE'12 detection cost function (DCF), C_{primary}, using $P_{\text{known}} = 0.5$ for evaluating the systems. From the results, we observe the general trend of performance degradation as test utterance duration is reduced [8, 15]. As expected,

Compute Cholesky factorization: $\mathbf{\Sigma} = \mathbf{R}\mathbf{R}^T$
for *Segment* $s \leftarrow 1$ **to** N **do**
  Extract full i-vector $\mathbf{w}_s$
  for *Index* $j \leftarrow 1$ **to** M **do**
    Random vector: $\mathbf{v} \leftarrow randn_{D\times 1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Synthesized i-vector: $\hat{\mathbf{w}}_{sj} \leftarrow \mathbf{w}_s + \mathbf{v}\mathbf{R}$
  end
end

Algorithm 2: Computation of synthetic i-vectors

adding truncated i-vectors significantly improve the system performance, as in [8]. It is worth noting that when truncated i-vectors are used in PLDA, the underlying training dataset or the acoustic feature vectors are essentially the same as when full i-vectors are used. The performance improvement is attained simply through data reorganization (see Sec 5.1). Similar improvements are also observed when the synthesized i-vectors are added in PLDA training. This verifies that our additive noise model in (3) of duration degradation is valid. The most significant gain, however, is achieved through the proposed QMF calibration. The best performance is achieved by using truncated i-vectors in PLDA and with QMF calibration. If QMF is not applied on the scores, the synthesized i-vectors can provide benefit compared to the case when only 4 truncated versions are available to train PLDA. Thus, the application of QMF to model the scores' behavior with respect to the duration variability is more beneficial copared to modeling the duration effect in the i-vector space.

7. CONCLUSIONS

In this paper, we have addressed the problem of duration mismatch in an i-vector based speaker recognition system. Three different approaches are presented to compensate for the mismatch: using multi-duration PLDA training, score domain compensation using quality measure function, and synthetically generated short duration i-vectors in PLDA training. The proposed methods demonstrate encouraging performance improvements on short duration test conditions when compared to the baseline system trained on full duration utterances.

8. REFERENCES

[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19 – 41, 2000.

[2] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.

[3] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in Eigenvoice space," *IEEE Trans. Audio Speech Lang. Process.*, vol. 8, no. 6, pp. 695–707, 2000.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus Eigenchannels in speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 99, pp. 788 – 798, May 2010.

[6] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 4828 – 4831.

[7] J. Villalba and N. Brummer, "Towards fully Bayesian speaker recognition: Integrating out the between-speaker covariance," in *Proc. InterSpeech*, Florence, Italy, Oct. 2011, pp. 505 – 508.

[8] A. Sarkar, D. Matrouf, P. Bousquet, and J. Bonastre, "Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification," in *Proc. InterSpeech*, Portland, OR, September 2012.

[9] "The NIST year 2012 speaker recognition evaluation plan," 2012. [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf

[10] K. Li and E. Wrench Jr, "An approach to text-independent speaker recognition with short utterances," in *Proc. IEEE ICASSP*, vol. 8, 1983, pp. 555–558.

[11] P. Angkititrakul and J. Hansen, "Discriminative in-set/out-of-set speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 2, pp. 498–508, 2007.

[12] J.-W. Suh and J. H. Hansen, "Acoustic hole filling for sparse enrollment data using a cohort universal corpus for speaker recognition," *The Journal of the Acoustical Society of America*, vol. 131, p. 1515, 2012.

[13] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short duration SVM-and GMM-based speaker verification," in *Proc. Odyssey*, Stellenbosch, South Africa, 2008.

[14] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Proc. InterSpeech*, Brisbane, Australia, 2008, pp. 853–856.

[15] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-vector based speaker recognition on short utterances," in *Proc. InterSpeech*, Florence, Italy, 2011, pp. 2341–2344.

[16] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," in *Proc. IEEE ICASSP*, Vancouver, Canada, May. 2013.

[17] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "Evaluation of i-vector speaker recognition systems for forensic application," in *Proc. Insterspeech*, Florence, Italy, August 2011, pp. 21–24.

[18] W. Rao and M. Mak, "Utterance partitioning with acoustic vector resampling for i-vector based speaker verification," in *Proc. Odyssey*, Singapore, June 2012.

[19] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. IEEE ICASSP*, vol. 1, Toulouse, France, May 2006, pp. 325–328.

[20] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey*, Crete, Greece, 2001, pp. 213–218.

[21] M. McLaren and D. A. van Leeuwen, "A simple and effective speech activity detection algorithm for telephone and microphone speech," in *Proc. NIST SRE 2011 workshop*, Atlanta, GA, 2011.

[22] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[23] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul 2001.

[24] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. InterSpeech*, Pittsburgh, Pennsylvania, 2006.

[25] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-Vector length normalization in speaker recognition systems," in *Proc. InterSpeech*, Florence, Italy, Oct. 2011, pp. 249 – 252.

[26] D. A. van Leeuwen and R. Saeidi, "Knowing the non-target speakers: the effect of the i-vector population for PLDA training in speaker recognition," in *Proc. IEEE ICASSP*, 2013.

[27] L. J. Rodríguez-Fuentes, N. Brümmer, M. Penagarikano, A. Varona, M. Diez, and G. Bordel. (2012, Nov.) The Albayzin 2012 language recognition evaluation plan. [Online]. Available: http://www.ehu.es/~ljrf/tmp/Albayzin_LRE12_EvalPlan_v1.2.pdf