I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification

R. Saeidi[1] , K. A. Lee[2], T. Kinnunen[3], T. Hasan[4], B. Fauve[5], P. -M. Bousquet[6], E. Khoury[7], P. L. Sordo Martinez[8], J. M. K. Kua[9], C. H. You[2], H. Sun[2], A. Larcher[2], P. Rajan[3], V. Hautamäki[3], C. Hanilci[3], B. Braithwaite[3], R. Gonzales-Hautamäki[3], S. O. Sadjadi[4], G. Liu[4], H. Boril[4], N. Shokouhi[4], D. Matrouf[6], L. El Shafey[7], P. Mowlaee[1], J. Epps[9], T. Thiruvaran[9] D. A. van Leeuwen[1], B. Ma[2], H. Li[2], J. H. L. Hansen[4], J. -F. Bonastre[6], S. Marcel[7], J. Mason[8], E. Ambikairajah[9]

[1]Radboud University Nijmegen, The Netherlands, [2]Institute for Infocomm Research, Singapore, [3]University of Eastern Finland, Finland
[4]CRSS, University of Texas at Dallas, USA, [5]ValidSoft Ltd, London, UK, [6]LIA, University of Avignon, France
[7]Idiap Research Institute, Switzerland, [8]Swansea University, UK, [9]University of New South Wales, Australia

Abstract

I4U is a joint entry of nine research **I**nstitutes and **U**niversities across **4** continents to NIST SRE 2012. It started with a brief discussion during the Odyssey 2012 workshop in Singapore. An online discussion group was soon set up, providing a discussion platform for different issues surrounding NIST SRE'12. Noisy test segments, uneven multi-session training, variable enrollment duration, and the issue of open-set identification were actively discussed leading to various solutions integrated to the I4U submission. The joint submission and several of its 17 subsystems were among top-performing systems. We summarize the lessons learnt from this large-scale effort.

Index Terms: Speaker Verification, NIST SRE 2012, I4U, i-vector

1. Introduction

The I4U submission to National Institute of Standards and Technology (NIST) speaker recognition evaluation 2012 (SRE'12) [1] is a result of active exchange of information between the coalition participants across nine institutions. The name of the institutes and corresponding system identifiers are provided in Table 1. The submitted results are based on the fusion of multiple classifiers. The optimization of the component classifiers and the fusion device were done with development sets jointly designed within the I4U coalition with multiple design iterations, refinement of noise adding protocol and various other details. Different from previous SREs, the task of SRE'12 involves:

Handling noisy test segments: This required speech enhancement algorithms and employing mixed training or parallel model combination techniques.

Imbalanced multi-session training: There are tens of segments available for training some speaker models while only a single segment for some other speakers.

Open-set identification: SRE'12 evaluation protocol allows the use of knowledge of all target speakers in each detection trials which resulted in utilizing compound log-likelihood ratio.

This work was partly supported by Academy of Finland (proj. no 253120 and 132129), Swiss National Science Foundation under the LOBI project, contract no. SNSF-235 and European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 238803.

Table 1: I4U Coalition and assigned system indexes

Site	System index
ValidSoft Ltd (VLD)	Sys1
Swansea University (UWS)	Sys2
University of Avignon (LIA)	Sys3
Radboud University Nijmegen (RUN)	Sys4
University of Texas at Dallas (CRSS)	Sys5–10
University of Eastern Finland (UEF)	Sys11
Institute for Infocomm Research (IIR)	Sys12–16
Idiap Research Institute (IDIAP)	Sys17

This paper is organized as follows: In Section 2, we present the strategies taken to make a development set coping with SRE'12 new conditions. Details of the submitted systems and the component classifiers, together with the strategies to deal with the new challenges listed above are described in Section 3. One of the motivations underlying the I4U coalition is to experiment with the fusion of large numbers of sub-systems. Results for the individual and the fused system are presented in Section 4.

2. Development sets

The development sets were generated to help I4U team members in developing their speaker recognition systems considering the *special* conditions in SRE'12 including multiple segments training for a speaker[1]. All the members of I4U coalition helped in refining the lists with respect to detecting empty or otherwise problematic segments with conflicts in gender and speaker PIN (there are issues with pre-SRE'12 lists like multiple-genders or wrong genders for some speakers). The latest lists from NIST were utilized and speech segments for all 1918 target speakers were fetched from SRE'06, SRE'08 and SRE'10 corpora and corresponding meta-data were extracted. To be able to assess both the recognition systems' generalization and calibration performance, separate development (DEV) and evaluation (EVAL) sets were created. The number of segments, speakers and trials for each set are given in Table 2. In designing these sets, the following criteria were considered:

- Test segments are disjoint for DEV-test and EVAL-test.
- Most of the train segments in DEV-train are added to EVAL-train. The number of train segments in EVAL-train is almost

[1]The lists are available via `http://cls.ru.nl/~saeidi/file_library/I4U.tgz`

Table 2: Number of speakers, speech segments and trials in the development sets.

	Number of speakers				Number of segments				Number of trials			
	DEV		EVAL		DEV		EVAL		DEV		EVAL	
	Train	Test	Train	Test	Train	Test	Train	Test	True	False	True	False
Males	680	868	763	804	16941	19866	29961	21837	14589	13494291	15483	16646148
Females	1039	1243	1155	1102	24693	25980	43119	28548	19863	26973357	20763	32952177

Table 3: Feature extraction setup for the systems in I4U, CMVN: cepstral mean and variance normalization, RFCC: repartitioned frequency cepstral coefficients [2], MHEC: mean Hilbert envelope coefficients [3].

	Features	SAD	Speech enhancement	Features post-processing
Sys1–3	19 LFCCs + Δ + ΔE + first 11 $\Delta\Delta$	Energy-based	Spectral subtraction	CMVN
Sys4	19 MFCCs + E + Δ + $\Delta\Delta$	Energy-based [4]	Wiener filtering[5]	Feature warping [6]
Sys5–7	12 MHEC + logE + Δ + $\Delta\Delta$	Voicing feature [7]	-	RASTALP [8] + CMVN
Sys8–10	12 RFCC + c0 + Δ + $\Delta\Delta$	Statistical SAD [9]	-	Feature warping
Sys11	18 MFCCs + Δ + $\Delta\Delta$	Adaptive SAD [10]	-	RASTA + CMVN
sys12–14	18 MFCCs + logE + Δ + $\Delta\Delta$	Energy-based	Qualcomm-ICSI-OGI [2]	RASTA + CMVN
sys15–16	19 LPCCs + Δ + $\Delta\Delta$ + 12 MFCCs	Energy-based	Qualcomm-ICSI-OGI	RASTA + CMVN
Sys17	19 MFCCs + logE + Δ + $\Delta\Delta$	Energy-based	Qualcomm-ICSI-OGI	CMVN

twice the number of segments in DEV-train. This design choice is made to evaluate the systems performance under the condition that speaker and channel spaces are already trained but the number of enrollment segments for target speaker modeling has increased.

- The segments from train to test have all different LDC-IDs to avoid testing against same session from training.
- Two disjoint sets of speakers from SRE'06 data that do not appear in SRE'12 are added to DEV-test and EVAL-test to form *unknown non-target trials*.
- For those speakers having telephone and microphone data, both types of channels were included in the train set so that systems could benefit from having different channels in training.
- Considering noisy segments inclusion in NIST SRE'12, for every original NIST segment, two noisy versions were generated. Noise adding was carried out using FaNT [3]. We have used ten noise segments for each HVAC (heating, ventilation and air-conditioning) and crowd noise type. Noise signals used to contaminate the speech segments were different from train to test and from DEV to EVAL. Noises are added at two SNR-levels 6dB and 15dB. The mean *measured* SNR-levels were 40dB, 15dB and 10dB for original, 15dB and 6dB segments, respectively[4]. Since there are two noisy versions of each clean segment being utilized in DEV and EVAL sets, the performance of the developed systems are optimized to perform well under noisy condition rather than clean ("not altered") condition.

3. Recognition systems

The systems developed in the I4U coalition were based on state-of-the-art: 1) *i-vector* system [11] with probabilistic linear discriminant analysis (PLDA) [12] modeling, or 2) Gaussian supervector representation and *joint factor analysis* (JFA) [13, 14], or support vector machine (SVM) modeling. All 16 kHz audio data were down-sampled to 8 kHz to match to the existing 8 kHz background data. Energy-based speech activity detection (SAD) was applied to telephone segments, while for interview segments a dual-channel SAD is employed. The automatic speech recognition (ASR) transcripts from NIST for interview segments in SRE'08 and SRE'10 were used to refine the SAD labels. All of the systems are gender-dependent. The components and data usage of sub-systems are presented in Tables 3, 4 and 5 for features, transform and classifier, respectively.

Sys1: Validsoft's i-vector system uses spectral subtraction to enhance energy profile for SAD. Test i-vectors are scored against all target segment i-vectors followed by score averaging.

Sys2: Swansea's i-vectors are normalized with *eigen-factors radial* (EFR) method [15] utilizing total covariance matrix of the background data. LDA-reduced 200-dimensional i-vectors are averaged for each target speaker and used with Mahalanobis scoring.

Sys3: LIA's system uses two fused sub-subsystems. The first uses LDA reduction preceded by iterative i-vector normalization according to the covariance matrix [15] and two-covariance scoring; the second uses PLDA preceded by *spherical nuisance normalization* with within-class covariance matrix [15]. Score is computing as a) the average score of the test i-vector against all target i-vectors and b) an equal-weights combination of these scores according to multiple PLDA subspace dimensions (from 50 to 400 in steps of 50).

Sys 4: RUN's i-vector PLDA system uses dynamic noise suppression within a Wiener filter applied both for speech enhancement and SAD. Noise estimation uses *improved minima controlled recursive averaging* (IMCRA) [5, 19] which averages the previous estimate of the noise power spectra and has proven robust against input SNR and different noise types due to rapid noise tracking. The noise power spectral density estimate is used for decision-directed *a-priori* SNR estimation, which further defines a Wiener filter applied for magnitude enhancement.

Sys5–10: The CRSS's i-vector systems use combinations of two different front-ends and three back-ends [20, 21]. Gaussianized cosine-distance scoring (GCDS) and a discriminative back-end using L2-regularized logistic regression (using LIBLINEAR [18]) are used. The enrollment i-vectors are averaged and then Gaussianized using mean and variance of devset. LDA dimensionality reduction and cosine scoring are used.

Sys11: UEF contributed the overall fusion component for I4U [22, 23] and developed a robust utterance-adaptive SAD [10][5] where 16-component speech and non-speech codebooks are trained from 12 MFCCs including c0. Training labels

[3] `http://dnt.kr.hsnr.de/download.html`

[4] The SNR is measured using *stnr* tool from NIST

[5] SAD available at http://cs.uef.fi/pages/tkinnu/VQVAD/VQVAD.zip

Table 4: Transform details for sub-systems in I4U. Numbers in data columns are standing for corresponding NIST SRE corpus, SW: Switchboard II Phase 2 and 3, Switchboard cellular part 1 and 2, Fis: Fisher, -D: diagonal covariance, -F: full covariance, TV: total variability [11], NAP: nuisance attribute projection [11], ISV: inter session variability [16].

	UBM	UBM data	Transform	Transform data
Sys1	512-D	04	TV 400	04, 05, SW, DEV
Sys2	512-D	04	TV 400	04, 05, 06, SW, Fis
Sys3	512-D	04, 05, Fis	TV 400	04, 05, 06, 08, 10, SW
Sys4	2048-D	04, 05, 06, SW, Fis	TV 400	Same as UBM
Sys5–10	1024-D	Tel only from 04, 05, 06, SW	TV 600	04, 05, 06, SW, DEV
Sys11	1024-D	04, 05, 06 and 08	TV 600	04, 05, 06, SW and Fis
Sys12	512-F	Tel only from 04, 05, 06, SW	TV 600 (400 Tel + 200 mic)	Tel from 04, 05, 06, SW and mic from 05, 06, MIXER5
Sys13,14	1024-D	04	NAP 60	Tel from 04, 05, 06, SW and mic from 05, 06, MIXER5
Sys15	512-F	04	NAP 60	04, 06, 08, 10, 08-followup
Sys16	512-F	04	JFA	06, 08, 10
Sys17	512-D	04	ISV 200	06, 08, 10

Table 5: Classifier details for i-vector based systems in I4U: Lnorm, Length normalization [17], EFR: eigen-factors radial normalization [15], <IV> and <scores>: average over i-vectors or scores in multi-session training.

	Background data for IV processing	IV pre-processing	scoring	#Voice	#Channel	Scoring strategy
Sys1	DEV train	EFR (W)	PLDA	300	50	<scores>
Sys2	DEV train	EFR (C), LDA(300)	Mahanabolis	-	-	<IV>
Sys3	DEV train	i. EFR (C), LDA (50 to 400) ii. EFR (W)	i. 2Cov ii. PLDA	50 to 400	400	<IV>
Sys4	DEV train	LDA(200), centering, WCCN, Lnorm	PLDA	200	50	<IV>
Sys5 and Sys8	04, 05, 06, SW, DEV train	LDA(400), centering, Lnorm	PLDA	400	400	<IV>
Sys6 and Sys9	04, 05, 06, SW, DEV train	LDA(400), Gaussianization, Lnorm	Cosine	-	-	<IV>
Sys7 and Sys 10	04, 05, 06, SW, DEV train	L2-regularized [18]	linear regression	-	-	<IV>
Sys11	04, 06, 08, 10 and SW	-	PLDA	200	0	<IV>
Sys12	04, 06, 08, 10, SW, MIXER5 and DEV train	LDA(400)	PLDA	200	50	<IV> and Snorm

are obtained from reliable frames with the help of aggressive spectral oversubtraction. The recognizer is a standard i-vector PLDA system and, unlike most of the other I4U system, does not use multicondition training.

Sys12 by I2R whitens the first-order sufficient statistics using UBM covariances, which speeds up estimation of the posterior distribution during the total variability matrix (T-matrix) training and i-vector extraction [24]. T-matrix estimation uses two subspaces, T_{tel} and T_{mic}, where T_{tel} is trained from telephone data and T_{mic} from microphone data following decoupled method on [25]. This enables easy control of the dimensionality of the subspaces in $T = [T_{\mathrm{tel}}, T_{\mathrm{mic}}]$ and avoids the problem of data type imbalance encountered when all data are pooled for T-matrix training in one go. For details of the PLDA implementation, refer to [26].

Sys 13 by I2R is a GMM supervector system with KL divergence kernel [27]. Utterance GMM is obtained via MAP adaptation of the UBM means that are concatenated and normalized by the UBM standard deviation and square root of the mixture weights. Nuisance attribute projection (NAP) [28] and tz-norm are applied for channel and score normalization, respectively.

Sys 14 is an *anti-model* variant of **Sys 13**. The use of other target speakers is allowed in SRE'12 which leads to an open-set identification problem. The anti-model approach of [29] is adopted for increased discrimination between target and unseen non-targets. SVM for each target speaker is trained using the supervectors of the other target speakers as the SVM background together with additional data drawn from SRE'04 for the unseen speakers.

Sys 15 is a Bhattacharyya-kernel GMM-SVM system with data-dependent relevance factor [30, 31] and zt-norm. **Sys 16**, in turn, uses joint factor analysis (JFA) implementation for I2R's SRE'10 submission [32]. It is composed of 300 speaker factors, 200 channel factors (100 for telephone, 50 for microphone, 50 for interview), and full rank diagonal matrix. For eigenchannel training, the tel, mic and interview channels were separately trained and concatenated into an eigenchannel matrix. Enrollment and scoring (with zt-norm) are as in **Sys 15**.

Sys 17: IDIAP's system is a single classifier with inter-session variability (ISV) modeling technique of [16]. It is implemented using *Bob*[6], an open-source signal processing and machine learning toolbox. ISV is similar to JFA with linear scoring approximation [33] but with merged eigen-voice and -channel spaces. Scores are normalized using zt-norm.

4. System performance

We analyze and compare system performance on the core task of NIST SRE'12 using the equal error rate (EER) and *primary cost*. The notion of EER is commonly known. What is new in SRE'12 is the use of the so-called primary cost C_{primary}, defined as the average cost at two specific points on the DET curve. At either of these points, the detection cost function (DCF) is defined in normalized form (such that the maximum value is one), as follows

$$C_{\mathrm{Norm}}(\theta) = P_{\mathrm{miss}}(\theta) + \frac{1-P_{\mathrm{tar}}}{P_{\mathrm{tar}}} \times \frac{[P_{\mathrm{fa}}(\theta|\mathrm{known}) + P_{fa}(\theta|\mathrm{unknown})]}{2}.$$

[6]http://idiap.github.com/bob/

Table 6: Analysis of system performance based on equal error rate (EER) and minimum C_{primary} (minC) for $P_{\text{known}} = 0$. NIST SRE'12 common conditions include multi-session in train and specific channel in test; CC1: interview and CC3: added noise interview. Fusion:1) Auto Ridge [22] submitted to SRE'12 as I4U submission 2) Auto Ridge post evaluation 3) FoCal post evaluation.

	Males				Females			
	CC1		CC3		CC1		CC3	
	EER	minC	EER	minC	EER	minC	EER	minC
Sys1	5.55	0.2674	4.22	0.4154	4.26	0.1674	4.07	0.5600
Sys2	5.44	0.2633	4.27	0.4246	4.77	0.1950	4.27	0.5663
Sys3	12.10	0.4998	10.90	0.5579	11.50	0.4363	10.50	0.5498
Sys4	5.75	0.2670	4.83	0.3741	4.86	0.1580	4.09	0.3018
Sys5	4.73	0.2669	4.14	0.3635	4.53	0.1373	3.52	0.3072
Sys6	4.28	0.2168	3.79	0.3053	4.05	0.1118	3.43	0.2420
Sys7	9.71	0.4742	9.32	0.6071	5.81	0.3083	5.18	0.3840
Sys8	4.81	0.3051	4.28	0.3918	4.65	0.1167	3.27	0.3094
Sys9	4.86	0.2374	4.22	0.2894	4.15	0.0948	3.38	0.2346
Sys10	9.84	0.4251	9.61	0.5714	5.56	0.1635	5.13	0.4124
Sys11	13.30	0.5276	9.72	0.6316	12.60	0.3985	7.48	0.5316
Sys12	3.74	0.2765	3.29	0.3322	4.01	0.2290	3.62	0.3877
Sys13	4.77	0.4440	4.50	0.3587	4.36	0.3055	3.35	0.2470
Sys14	5.45	0.3474	5.74	0.3618	4.52	0.1351	3.52	0.1591
Sys15	4.85	0.3347	5.57	0.3751	4.55	0.1708	4.10	0.2188
Sys16	3.78	0.2333	5.56	0.3415	5.17	0.1906	5.31	0.4245
Sys17	9.03	0.4932	7.88	0.4302	8.48	0.4189	5.66	0.3797
Fusion1	3.62	0.2306	3.25	0.3162	3.96	0.1196	2.81	0.2470
Fusion2	3.38	0.2267	3.75	0.3075	4.06	0.0760	2.92	0.2140
Fusion3	3.48	0.2020	2.67	0.2767	3.87	0.0719	2.78	0.2277

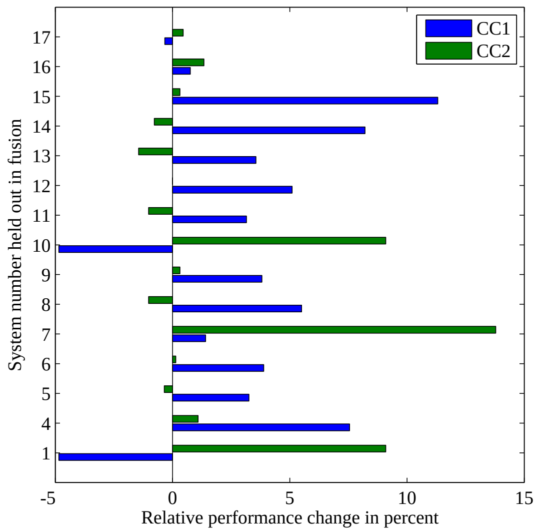
Here, P_{tar} is the *a priori* probability that a trial is a target trial, while $P_{\text{miss}}(\theta)$ and $P_{\text{fa}}(\theta)$ are, respectively, the probability of miss and false alarm at threshold θ. Notice that $P_{\text{fa}}(\theta)$ consists of two components computed separately from the known and unknown non-target trials. Now, let θ_A be the threshold which gives the $C_{\text{Norm}}(\theta_A)$ with $P_{\text{tar}} = 0.01$ and θ_B be the threshold which gives the $C_{\text{Norm}}(\theta_B)$ with $P_{\text{tar}} = 0.001$, the primary detection cost is defined as the average cost between the points on the detection error trade-off (DET) curve, as follows

$$C_{\text{primary}} = \frac{C_{\text{Norm}}(\theta_A)+C_{\text{Norm}}(\theta_B)}{2}$$

Table 6 shows the absolute performance of all 17 systems and their fusion for common conditions 1 and 3 as defined in SRE'12. One obvious point to note here is that, the PLDA i-vector systems give consistently better performance in terms of EER and minimum C_{primary} when the test signal is collected over clean (CC1) and noisy (CC3) interview sessions. It is also obvious that, the GMM-SVM (Sys 13, 14, and 15) and JFA (Sys 16) give equally good performance compared to, and for some instances better than i-vector based systems.

The fusion of large ensemble of recognition systems was by itself a challenging issue, for instance, over-fitting may easily degrade the performance. We followed the recent work in [22, 23] whereby fusion weights are trained using regularization to avoid over fitting. Different regularizers were systematically evaluated, and ridge regression ($L2$-norm regularization) was chosen. Instead of cross-validating the regularization factor λ, we decided to use a simple Bayesian method that allows automatic selection of λ, as described in [35]. This method integrates out λ and the resulting non-convex optimization problem is solved via *majorization-minimization* approach. Convergence was assumed after two iterations. The fusion results are shown in Table 6 with Fusion1-3. Though effective on our DEV set, the ridge-regression regularization (Fusion1 and 2) does not always give improved performance over the single best system,

Figure 1: Analysis of excluding one system at a time in fusion using Focal and employing compound log-likelihood ratio [34] for $P_{\text{known}} = 0.5$. Using the full ensemble of classifiers results in actual C_{primary} of 0.3959 and 0.2836 for first two common conditions (CC1 and CC2) respectively in SRE'12 for the pooled scores of males and females. A positive relative change indicates increased actual C_{primary} by excluding a system in fusion resulting in fusion performance drop. Systems number 2 and 3 are not considered for this analysis.



while the original FoCal[7] fusion (Fusion3) does. One possible insight that we might draw here is that regularization might hamper effective training of fusion parameters when the development data is sufficient. This is a point for future research. The results for Fusion1 are slightly inferior to Fusion2 because of some mis-labeled scores during the evaluation which are corrected for post-evaluation (Fusion2). An analysis of individual systems importance in fusion is provided in Fig. 1. Comparing between interview (CC1) and telephone (CC2) conditions in Fig. 1, the most influensive systems in fusion are not the same across different conditions.

5. Conclusion

This paper provides an overview of fusion of 17 systems submitted to NIST SRE'12 by different sites in I4U coalition. The collaboration of over 30 researchers within the coalition benefited all the sites in preparing robust speaker recognition systems. It is hard to compare the individual subsystems and determine the strengths of each system but in a very general prospective, the systems that utilized more recent features and employ speech enhancement in the front-end were more successful. Averaging the enrollment i-vectors gave about the same performance as averaging the scores of i-vectors. Discriminative training schemes, such as SVMs, using a proper distance kernel on Gaussian supervector representation was found to outperform generative i-vector representation with PLDA classification. The new paradigm shift in NIST SRE'12 is expected to emphasize the discriminative training in modeling and even i-vector representation.

[7] http://niko.brummer.googlepages.com/focal

6. References

[1] NIST speaker recognition evaluation 2012. http://www.nist.gov/itl/iad/mig/sre12.cfm.

[2] H. Boril, P. Fousek, and P. Pollak. Data-driven design of front-end filter bank for Lombard speech recognition. In *Proc. Interspeech 2006 (ICSLP)*, pages 381–384, 2006.

[3] S. O. Sadjadi and J. H. L. Hansen. Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pages 5448 –5451, 2011.

[4] 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, US, 2011.

[5] I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Trans. on Speech and Audio Processing*, 11(5):466 – 475, 2003.

[6] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Odyssey, 2001*.

[7] S. O. Sadjadi and J. H. L. Hansen. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Processing Letters*, pages 197–200, Mar. 2013.

[8] H. Bořil, F. Grézl, and J. H. L. Hansen. Front-end compensation methods for LVCSR under Lombard effect. In *INTERSPEECH 2011*, pages 1257–1260, Florence, Italy, 2011.

[9] J. Sohn and W. Sung. A voice activity detector employing soft decision based noise spectrum adaptation. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1998)*, volume 1, pages 365 –368 vol.1, 1998.

[10] T. Kinnunen and P. Rajan. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.

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

[12] S. J. D. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *11th International Conference on Computer Vision*, pages 1–8, 2007.

[13] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of inter-speaker variability in speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 16(5):980–988, July 2008.

[14] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Speaker and session variability in GMM-based speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 15(4):1448–1460, May 2007.

[15] P. M. Bousquet, A. Larcher, D. Matrouf, J. F. Bonastre, and O. Plchot. Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. In *Odyssey, 2012*, 2012.

[16] R. Wallace, M. McLaren, C. McCool, and S. Marcel. Inter-session variability modelling and joint factor analysis for face authentication. In *International Joint Conference on Biometrics*, 2011.

[17] D. Garcia-Romero and C. Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Proc. Interspeech 2011*, pages 249–252, 2011.

[18] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008.

[19] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. on Speech and Audio Processing*, 9(5):504 –512, 2001.

[20] G. Liu, T. Hasan, H. Bořil, and J.H.L. Hansen. An investigation on back-end for speaker recognition in multi-session enrollment. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.

[21] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, and J.H.L. Hansen. CRSS Systems for 2012 NIST Speaker Recognition Evaluation. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.

[22] V. Hautamäki, K. A. Lee, A. Larcher, T. Kinnunen, B. Ma, and H. Li. Variational Bayes logistic regression as regularized fusion for nist sre 2010. In *Odyssey, 2012*, 2012.

[23] V. Hautamaki, T. Kinnunen, F. Sedlak, K.-A. Lee, B. Ma, and H. Li. Sparse classifier fusion for speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 2013, Accepted for publication.

[24] P. Kenny. A small foot-print i-vector extractor. In *Odyssey, 2012*, pages 1–6, 2012.

[25] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel. An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. In *Odyssey, 2010*, pages 28–33, 2010.

[26] Y. Jiang, K. A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li. PLDA modeling in i-vector and supervector space for speaker verification. In *Proc. Interspeech 2012*, 2012.

[27] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308 – 311, 2006.

[28] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, volume 1, pages 97–100, Toulouse, France, 2006.

[29] P. Matejka, P. Schwarz, L. Burget, and J. Cernocky. Use of anti-models to further improve state-of-the-art PRLM language recognition system. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, volume 1, pages 197–200, 2006.

[30] C. H. You, K. A. Lee, and H. Li. GMM-SVM kernel with a bhattacharyya-based distance for speaker recognition. *IEEE Trans. Audio, Speech and Language Processing*, 18(6):1300 – 1312, 2010.

[31] C. H. You, H. Li, and K. A. Lee. A GMM-supervector approach to language recognition with adaptive relevance factor. In *Proc. 18th European Conf. on Signal Processing (EUSIPCO 2010)*, pages 1993–1997, 2010.

[32] B. Ma, H. Sun, K. A. Lee, C. H. You, D. Zhu, E. Wang, R. Tong, C. L. Huang, C. C. Leung, V. Hautamäki, and H. Li. IIR system description for the 2010 nist speaker recognition evaluation submission. In *Proc. NIST SRE 2010 workshop*, 2011.

[33] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny. Comparison of scoring methods used in speaker recognition with joint factor analysis. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, 2009.

[34] 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. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, 2013.

[35] C. S. Foo, C. B. Do, and A. Y. Ng. A majorization minimization algorithm for (multiple) hyperparameter learning. In *Int. Conf. Mach. Learning*, pages 321–328, 2009.