AN INVESTIGATION ON BACK-END FOR SPEAKER RECOGNITION IN MULTI-SESSION ENROLLMENT

*Gang Liu, Taufiq Hasan, Hynek Bořil, John H.L. Hansen**

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas 75083, USA
{Gang.Liu, Taufiq.Hasan, Hynek, John.Hansen}@utdallas.edu

ABSTRACT

This study explores various back-end classifiers for robust speaker recognition in multi-session enrollment, with emphasis on optimal utilization and organization of speaker information present in the development data. Our objective is to construct a highly discriminative back-end framework by fusing several back-ends on an i-vector system framework. It is demonstrated that, by using different information/data configuration and modeling schemes, performance of the fused system can be significantly improved compared to an individual system using a single front-end and back-end. Averaged across both genders, we obtain a relative improvement in EER and minDCF by 56.5% and 49.4%, respectively. Consistent performance gains obtained using the proposed strategy validates its effectiveness. This system is part of the CRSS' NIST SRE 2012 submission system.

Index Terms— Universal Background Support, PLDA, speaker recognition, GCDS, classification algorithms

1. INTRODUCTION

Speaker verification, similar to other recognition/verification tasks, largely depends on the training and development utterances or instances from each speaker/class. The more information/instances available for one speaker/class, the more accurate the modeling can be. In more than a decade of speaker recognition evaluation (SRE) history hosted by the National Institute of Standards and Technology (NIST), a single instance (session) per enrollment/training for each speaker has been the dominating task [1, 2]. Consequently, speaker recognition systems, including universal background modeling-Gaussian Mixture modeling (UBM-GMM) [3], Gaussian Super-vector SVM (G-SVM) [4], and i-vector with probabilistic linear discriminative analysis (PLDA) [5-8] back-end, have become more or less optimized for the speaker detection task for single utterance enrollment. In NIST SRE 2012, multiple session enrollment and noisy test data were introduced for the very first time. These represent significant diversions from past evaluations and thus require a vigorous re-design and optimization of the classification framework making maximal use of the available development and training data.

Multiple session/instances for the enrollment speaker can drastically improve the system since the modeling process can learn more from the variability/nuances of the different utterances of the same speaker. Also, using out-of-set/impostor speakers in the modeling becomes more important as classifiers such as support vector machines (SVM) become more common. However, information from the non-target data can be practically unlimited, and thus the question of how exactly to employ them for maximum benefit is critical. We propose a series of solutions aimed at finding a generic and effective strategy of utilizing development data for optimal speaker classification.

Multi-session enrollment for speaker recognition has been investigated in the past, based on corpora such as TIMIT, YOHO [9], and ROSSI [10]. These datasets contain a limited number of speakers and/or noise variations, which precludes large scale experiments similar to SRE. This study, to the best of our knowledge, together with the efforts of other sites participating in NIST-SRE 2012, is a first attempt in using large noisy multi-session data for speaker recognition. This work is based on the methods proposed in [5-8, 11-12], but the methods are further adapted and modified to handle the multi-session scenarios.

This paper is organized as follows: Sec. 2-5 describe five back-ends utilized in this study and discuss what kind of information and processing they employ. Back-end fusions are shown in Sec. 6. A comprehensive experiment is established and discussed in Sec. 7 and 8, and research findings are summarized in Sec. 9. Relation to prior work is detailed in Sec. 10.

2. GAUSSIANIZED COSINE DISTANCE SCORING (GCDS)

The classical cosine distance scoring (CDS) for i-vector based system is [12]:

$$k(\omega_1, \omega_2) = \frac{(A^t\omega_1)^t(A^t\omega_2)}{\sqrt{(A^t\omega_1)^t(A^t\omega_1)}\sqrt{(A^t\omega_2)^t(A^t\omega_2)}} \quad (1)$$

where A is a projection matrix, which may come from within class covariance normalization (WCCN) or linear discriminative analysis (LDA) projection, 't' indicates the transpose operation, and ω_j denotes the i-vector of the j^{th} speech utterance. The operations are generally performed in a cascade fashion: where the i-vector is first projected through WCCN matrix and then LDA transformation is applied, both of which are estimated from a background data set.

We note that performance of classical LDA-WCCN-CDS methods highly depend on the WCCN projection, which is usually difficult to estimate (especially in noisy and/or channel mismatch conditions). Therefore, we propose to replace the WCCN with background data based Gaussianization, named Gaussianized CDS (GCDS). The algorithm is outlined below (source code is provided online [13].):

Step 1: Average the i-vectors of the j^{th} enrollment speaker;

*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.

Step 2: Calculate the mean and variance of the background data, which are used to Gaussianize the i-vectors from Step1;
Step 3: Apply length normalization on all the data [14];
Step 4: Apply LDA on all the data to reduce the dimensionality;
Step 5: Repeat Step 3;
Step 6: Perform cosine distance scoring;
Step 7: Score normalization (calculate the mean and variance of scores involved in the i^{th} test file, which are then Gaussianized by the derived mean and variance).

3. UBS-SVM ANTI-MODEL (UBSSVM)

This approach is based on Universal Background Support data-selection and SVM (UBSSVM) anti-modeling [11], where a cosine kernel is used. We note that, during the SVM modeling process, only the support vector has an impact on the final classification performance. Also, care needs to be taken in building the SVM with only a few positive examples (enrollment speaker) and a large number of negative examples (imposter speaker) for a specific enrollment. If the positive and negative examples are not properly balanced, the final performance may be compromised. Our usual practice is to begin with an individual SVM model trained by using a fixed imposter background dataset with the assumption that the imposters share a common subspace. This is a natural assumption, but an unbalanced data set may produce an over-fitting hyper-plane for the enrollment speaker. To address this imbalance, an imposter dataset selection was proposed in [11] and adopted here:

Step 1: Average the i-vectors of the j^{th} enrollment speaker;
Step 2: Follow the first 3 steps of [11];
Step 3: Score normalization (same as in Section 3).

This approach utilizes information not only from limited available enrollment data, but also from a large amount of imposter data, which helps the verification decision when a non-target trial is encountered.

4. L2-REGULARIZED LINEAR REGRESSION (L2LR)

L2-regularized logistic regression (L2LR) was applied by means of the LIBLINEAR toolkit (the source code having been slightly modified for parallel computation) [15]. The classifier training method is *one-versus-the-rest*. In contrast to approaches in GCDS and UBSSVM, the i-vectors of each enrollment speaker were not averaged and LDA was not applied. The parameters of this classifier are optimized on the Dev-set by using the hill-climbing method. The scores are again normalized (same as in Sec. 3).

It is need to note that this approach only makes use of enrollment data to build the model for the target speaker, and no 3rd party imposter data is involved (That is, for a specific target speaker, all the other speakers' data are used as imposter data).

5. MULTI-SESSION PLDA (PLDA1, PLDA2)

Traditionally, only one instance is available for each target speaker, where PLDA is the state-of-art back-end. We tested two methods of handling the multi-session case with PLDA.

5.1. Before-scoring Average PLDA (PLDA1)

The i-vectors of the j^{th} enrollment speaker are grouped and averaged before applying PLDA to perform verification. This allows us to use the centroid of multiple instances of each speaker to average out the potential noise and/or channel mismatch.

5.2. Post-scoring Average PLDA (PLDA2)

Each i-vector of the target file is treated as if originated from a different speaker. After applying PLDA, scores of the i^{th} test file against instances of j^{th} enrollment speaker are averaged, and used as the likelihood score of the trial of i^{th} test files coming from the j^{th} speaker. This is equivalent to a majority vote on the decision with the hope that each individual sample/utterance captures some combination of the acoustic-based speaker characteristics and environment distortion. This basically can be understood as multi-condition training and is an echo to the multi-condition preparation for the enrollment files, which is neglected in PLDA-1 in Sec. 5.2.

6. BACK-END FUSION

As noted in the previous sections, each back-end focuses on different data utilization. Some back-ends only use enrollment data and test data (GCDS, L2LR) for modeling, some back-ends use imposter data (UBSSVM), and other back-ends use similar data to learn the data variability and thus compensate for the environment mismatch (PLDA) in potential final trials. Table 1 summarizes the similarities and dissimilarities between the different back-ends.

Table 1. *Comparison of data usage among different back-ends.*

Back-end	Need imposter data	Uses LDA	Feature Avg.	Score Avg.
GCDS	**Y**	**Y**	**Y**	*N*
UBSSVM	**Y**	**Y**	**Y**	*N*
L2LR	*N*	*N*	**Y**	*N*
PLDA1	**Y**	**Y**	**Y**	*N*
PLDA2	**Y**	**Y**	*N*	**Y**

We observe that various back-ends differ in how much speaker information is utilized (whether using imposter data or full dimension) and how instances of each category are utilized (with or without averaging). Different information utilization contributes different discriminative modeling. Therefore, by fusion, we expect that the likelihood score will be reinforced more towards the correct decision, while contradictory decision will become less likely. We perform fusion using the linear-logistic regression algorithm from the BOSARIS toolkit [16].

7. EXPERIMENT SETUP

7.1. System Setup

The speaker recognition task is developed to address the NIST SRE'12 [2] evaluation. The flowchart is illustrated in Fig 1. The difference of this task compared to conventional SRE is outlined as: (1) the test data is corrupted by additive noise, (2) test utterances have different durations (varying from 20s through 160s), (3) multiple samples for enrollment data are available, and (4) all enrollment data is allowed to be used for collective modeling.

7.2. Front-end Processing

We utilized two different acoustic features: (a) Mel frequency Cepstral coefficients (MFCC), MFCCs are normalized by quantile cepstral normalization (QCN) [17] and low-pass RASTA filtering [18], and (b) Rectangular Filter-bank Cepstral Coefficients (RFCC) [17], which are processed through feature warping [19]. All features are 39-dimensional (12 cepstral coefficients+C_0+Δ+$\Delta\Delta$).

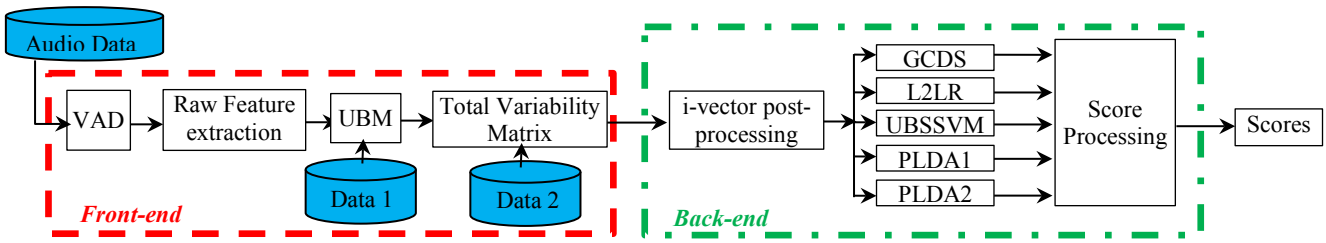

Figure 1: *i-vector based Speaker verification system block diagram. Data 1 and 2 correspond to raw feature data for UBM and Total Variability matrix, respectively. 'Audio data' means all the acoustic data involved for the verification task.*

7.3. I-Vector Extraction

Gender dependent UBMs with 1024 mixtures are trained on telephone utterances (See Table 3 for details). For total variability matrix, T, the UBM training dataset and additional SRE-12 target speakers' data in both clean and noisy versions are used. Five iterations are used for the EM training [20]. The i-vector dimension was set to 600. In the i-vector post-processing module, LDA was applied to reduce i-vector to 400 dimensions and then length normalization was applied (L2LR skipped this step; see Table 1).

7.4. Experimental and Development Dataset

To ensure a comprehensive evaluation, the proposed method will be tested on both male and female cases for two feature front-ends.

7.4.1. Experimental Setup

In preparing the development system, we maintained a close collaboration with the I4U consortium. The 1918 SRE'12 target speakers' utterances are collected from SRE'06-10 and a train-test list pair is prepared for evaluation. The training list included multiple sessions per speaker and the test list included both known and unknown non-target speakers, following SRE'12 protocol (Table 2). The trials are designed based on the three criteria: ***I***) train and test files are always from a different session, ***II***) both telephone and microphone recordings are kept for enrollment, ***III***) for each segment, 6dB and 15dB noisy versions are created. The noise sample is randomly chosen from a pool of HVAC and crowd noise files [20,23]. No same noise sample is used for train and test.

Table 2. *Number of speakers, segments and trials in the eval-set*

Gender	No. speakers		No. segments		No. trials	
	Train	Test	Train	Test	True	False
Male	763	804	29961	21837	15483	16646148
Female	1155	1102	43119	28548	20763	32952177

7.4.2. Development Dataset

This dataset is constructed to assist the development of different back-ends, which involves different corpora. Specifically, the development dataset for PLDA-1 and 2 includes the UBM training dataset and the clean and noisy enroll speaker data. The UBS-SVM development data set includes only UBM data, which has no overlapping speaker data with the enrollment data. For the back-end of GCDS, UBM training data and the enrollment data (clean and noisy) are only used in Step 2 of the GCDS algorithm. L2LR does not need a development set with the aim to avoid potential information distortion coming with development sets. Table 3 summarizes how the corpora are used in different tasks.

8. RESULTS AND ANALYSIS

First, we want to validate the performance of GCDS, which is detailed in Table 4. It is observed that score normalization can significantly enhance the performance in both classical CDS and

Table 3. *Corpora used to estimate the system components. (Note: "X" means that data from this corpus was used)*

Corpora	Switch-board	SRE 04	SRE 05	SRE 06	SRE 08	SRE 10	SRE 12
UBM	X	X	X	X			
T	X	X	X	X	X	X	
Dev Dataset				X	X	X	
Eval Dataset				X	X	X	X
LDA	X	X	X	X			
SVM-imposter	X	X	X	X			

Table 4. *Performance comparison between classical CDS and GCDS. Relative Gain is computed between 5th and 6th column.*

Gender	Back-end	CDS	GCDS	CDS	GCDS	Gain
	Score Norm	N	N	Y	Y	
Male	EER(%)	2.67	1.78	1.73	**1.42**	**+17.9%**
	minDCFx100	29.1	23.8	20.9	**16.3**	**+22.0%**
Female	EER(%)	3.60	2.41	2.29	**1.87**	**+18.3%**
	minDCFx100	37.2	30.5	27.2	**21.3**	**+21.7%**

proposed GCDS. We also observe similar trends in L2LR and UBSSVM, so for all these three back-ends, score normalization is always used. Secondly, we can see that GCDS outperforms CDS significantly, which proves the validity of GCDS.

The performances of the individual back-ends are shown in Fig. 2 using Detection Error Tradeoff (DET) curves. It may be noted that GCDS and PLDA-2 are very competitive in terms of EER. UBSSVM can offer the most competitive minDCF in both male and female cases. L2LR is inferior to the other three top back-ends (i.e., PLDA-2, GCDS. and UBSSVM). It, however, can provide a significant performance gain in fusion, especially for the minDCF measure, which is defined in [2].

From Table 5, we observe that the performances of the two individual front-ends are very similar (dark shaded area in Table 5). After introducing the four other back-ends, the discriminating capability is greatly enhanced for the individual i-vector system (light shaded area in Table 5). Compared with the MFCC+PLDA-1-i-vector system, after the back-end expansion (MFCC + five back-ends i-vector systems), the EER and minDCF is relatively improved by 45.0% and 48.7% for the male condition, and 48.7% and 36.8% in the female condition (the average relative EER and minDCF improvement across gender is 46.9% and 37.2%; and the same comparison will be applied thereafter for all across gender cases for simplicity). This suggests that the proposed multisession enrollment processing in this study have a better discriminating capability learned from the available data, which is realized by using our alternate information modeling approach.

To further verify the merits of different fusion approaches, we also summarize in Fig. 3 the relative gain of different fusion approaches against averaged PLDA-1 based individual front-ends ('Avg.1' in Table 5). It is noted that the front-end based fusion can relatively improve the EER and minDCF by around 16% for both

Table 5. *Performance comparison of individual front-end/back-end systems and different fusion combination systems. 'All-1' indicates a fusion system based on all two front-ends with back-end 1 (PLDA1). 'All-All' indicate the fusion is based on all two front-ends with all five back-ends (total 2 X 5 systems). 'Avg.1' is the average performance of back-end 1-based single front-end (dark shaded area) systems. 'Avg.2' is the average performance of the back-end-based fusion system (light shaded area). 'Gain 1', 'Gain 2', and 'Gain 3' are the relative gains obtained from 'All-1' vs. 'Avg.1', 'Avg.2' vs. 'Avg-1', and 'All-All' vs. 'All-1', respectively. Back-ends 1~5 are the five back-ends, corresponding to: PLDA1, GCDS, L2LR, UBSSVM, and PLDA2, respectively. minDCF is defined in [2].*

Gender	Front-end	MFCC	RFCC	*Avg.1*	*All- 1*	MFCC	RFCC	*Avg.2*	*All-All*	*Gain 1 (%)*	*Gain 2 (%)*	*Gain 3 (%)*
	Back-end	PLDA1	PLDA1			1~5	1~5					
Male	EER(%)	1.80	1.93	*1.87*	*1.56*	0.99	1.03	*1.01*	***0.74***	*16.6*	*46.0*	*52.6*
	minDCF**x**100	18.9	19.9	*19.4*	*16.3*	11.8	10.6	*11.2*	***8.88***	*16.0*	*42.3*	*49.0*
Female	EER(%)	2.69	2.37	*2.53*	*2.12*	1.38	1.03	*1.21*	***0.84***	*16.2*	*52.2*	*60.4*
	minDCF**x**100	25.0	23.8	*24.4*	*21.1*	15.8	12.6	*14.2*	***10.6***	*13.5*	*41.8*	*49.8*

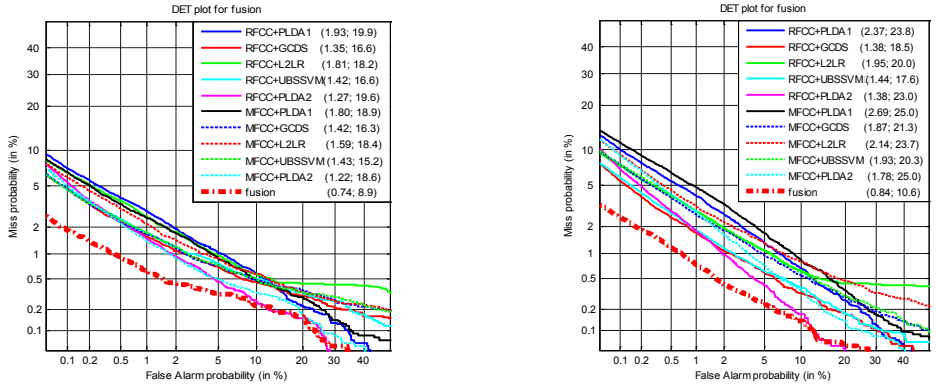

Figure 2: *DET plot for male (Left) and female (Right) condition. Numbers in parentheses are EER (%) and minDCF (x100), respectively.*

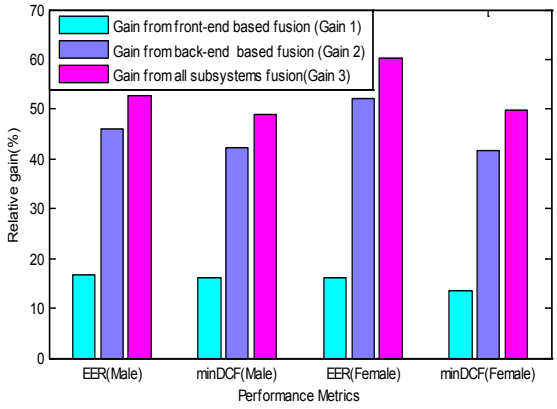

Figure 3: *Relative gain of different fusion combinations vs. averaged individual front-end performance metrics (based on back-end 1: PLDA1). Gain 1~3 come from Table 5.*

genders. However, this benefit is acquired at a high computation cost for i-vector extraction, which involves raw feature extraction, UBM training, total variability matrix and i-vector extraction for each front-end involving thousands of audio files. Compared with the averaged single front-end and back-end case ('Avg.1' in Table 5), fusion based on multiple back-ends but a single front-end performs much better, obtaining a relative gain (Gain 2) of 49% and 42% in EER and minDCF, respectively. Finally, when two front-ends and five back-ends are combined, we have a fusion of ten systems providing a relative gain (Gain 3) of 56.5% and 49.4%, in EER and minDCF, compared to the case 'Avg.1' in Table 5, where extra 7% gain against Gain 2 comes from the different feature, which is very limited due to the similar signal processing procedure behind them.

9. CONCLUSION

An information organization-based back-end fusion framework was proposed to fully explore the patterns present in the speaker ID development data. We considered the difficult real-life scenario of speaker recognition when the test utterances are noisy and of varying duration, similar to what is posed by NIST SRE 2012. For addressing noise and channel mismatch, robust front-end processing is an obvious necessity. In this study, for an i-vector based system, we demonstrated that by properly designing different back-end classifiers and subsequent fusion, a much greater benefit can be achieved compared to the scenario when multiple front-end features but one backend is utilized for the system fusion. Consistent and significant performance gains were obtained from the proposed strategy, which proves the validity of the methods presented. It is also noted that the proposed GCDS performs significantly better than the classical CDS in the current multisession enrollment verification experiment.

10. RELATION TO PRIOR WORK

This work is based on the methods proposed in [5-8, 11-12], which are adapted here to handle the new multi-session SRE-12 scenarios. Specifically, GCDS is based on CDS but far outperforms the latter. PLDA1 is similar with traditional PLDA [6] and is taken as the back-end baseline. PLDA2 is a modification of traditional PLDA scoring, and is also shown to outperform the traditional approach. This works is also aiming at extending the endeavor of [22].

11. ACKNOWLEDGEMENT

The authors would like to thank Yun Lei, for the discussion, the members of the I4U group (esp. Rahim Saeidi), for the development experiment setup, Omid Sadjadi, for noise adding. We also want to thank Navid Shokouhi, Keith W. Godin, Abhinav Misra and Ali Ziaei for the help and participation in NIST SRE-12.

12. REFERENCES

[1] NIST, "The NIST year 1997 - 2010 speaker recognition evaluation plans," [Online] Available: http://www.itl.nist.gov/iad/mig/tests/sre.

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

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

[4] W. Campbell, J. Campbell, D. Reynolds, E. Singer, P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition", *Comput. Speech Lang.*, 20 (2–3) (2006), pp. 210–229.

[5] S. J. D. Prince, J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV, 2007*, pp. 1–8.

[6] N. Brummer, "EM for Probabilistic LDA," Feb. 2010. [Online]. Available: https://sites.google.com/site/nikobrummer

[7] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Proc. Odyssey*, Brno, Czech Republic, 2010.

[8] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification", *in Proc. IEEE ICASSP*, 2011

[9] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1, pp. 91-108, 1995.

[10] G. Liu, Y. Lei, J. H.L. Hansen, "Robust feature front-end for speaker identification," *in Proc. IEEE ICASSP*, Kyoto, Japan, 2012. pp. 4233-4236.

[11] G. Liu, J.W. Suh, J. H.L. Hansen, "A fast speaker verification with universal background support data selection", *in Proc. ICASSP*, Kyoto, Japan, 2012. pp.4793-4796.

[12] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, P. Dumouchel,"Front-end Factor Analysis for Speaker Verification", *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 19, no. 4, pp. 788-798, August 2010.

[13] "GCDS Algorithm source code". [Online]. Available: www.utdallas.edu/~gang.liu/code.htm

[14] D. Garcia-Romero, C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," *in Proc. Interspeech*, Florence, Italy, August 2011, pp. 249-252.

[15] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research 9(2008), 1871-1874. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear.

[16] N. Brummer, E. de Villiers, "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF," *in Proc. NIST SRE Analysis Workshop*, Atlanta, USA, Dec. 2011.

[17] H. Bořil, J. H.L. Hansen, "Unsupervised Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments," *IEEE Trans. on Audio, Speech, and Lang. Process.*, 18(6), (2010). 1379-1393.

[18] H. Bořil, J. H.L. Hansen, "UT-Scope: Towards LVCSR under Lombard Effect Induced by Varying Types and Levels of Noisy Background," *in Proc. IEEE ICASSP*, 4472-4475, Prague, Czech Republic, May 2011.

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

[20] T. Hasan, G. Liu, S. O. Sadjadi, N. Shokouhi, H. Bořil, A. Ziaei, A. Misra, K. W. Godin, J. H.L. Hansen, "UTD-CRSS Systems for 2012 NIST Speaker Recognition Evaluation," *in Proc. NIST SRE Workshop*, Orlando, FL, USA, 2012

[21] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, N. Scheffer, "Towards Noise-Robust Speaker Recognition using Probabilistic Linear Discriminant Analysis", *in Proc. IEEE ICASSP*, Kyoto, Japan, 2012.

[22] N. Scheffer, Y. Lei, L. Ferrer, "Factor analysis back ends for MLLR transforms in speaker recognition", *in Proc. Interspeech*, Florence, Italy, August 2011, pp. 257-260.

[23] T. Hasan, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, J. H. L. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluation", *in Proc. IEEE ICASSP*, Vancouver, Canada, 2013.