CRSS SYSTEMS FOR 2012 NIST SPEAKER RECOGNITION EVALUATION

Taufiq Hasan, Seyed Omid Sadjadi, Gang Liu, Navid Shokouhi, Hynek Bořil, John H.L. Hansen[⋆]

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas (UTD), Richardson, TX 75080-3021, USA

ABSTRACT

This paper describes the systems developed by the Center for Robust Speech Systems (CRSS), for the 2012 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE). Given that the emphasis of SRE'12 is on noisy and short duration test conditions, our system development focused on: (i) novel robust acoustic features, (ii) new feature normalization schemes, (iii) various back-end strategies utilizing multi-session and multi-condition training, and (iv) quality measure based system fusion. Noisy and short duration training/test conditions are artificially generated and effectively utilized. Active speech duration and signal-to-noise-ratio (SNR) estimates are successfully employed as quality measures for system calibration and fusion. Overall system performance was very successful for the given test conditions.

Index Terms— Feature normalization, NIST SRE, robust features, speaker verification, quality measure fusion

1. INTRODUCTION

Consistent with previous NIST evaluations [1, 2], the core task in the NIST SRE-2012 [3] is speaker verification. However, several new challenges are introduced this year requiring a paradigm shift in the system development process. In addition to channel/session variability, test segments with real and artificially added noise as well as segments of varying durations are present in the evaluation conditions. Also, the speaker pins of the target speakers for SRE'12 were released in advance and the evaluation rules allow using all audio recordings from these speakers found in the previous SRE data releases. This enables the system developer to effectively perform multi-session and multi-condition training/enrollment, and also utilize the target speaker's knowledge to optimize the back-end classifier. Moreover, a new cost function is introduced, which is the average of two cost functions at two different operating points. This new cost function poses a significant challenge in the score calibration step. In the following sections, we describe our system development procedure for the NIST SRE-2012 evaluation, and provide details on individual parts of the i-vector [4] based systems and its variants.

2. PREPARATION OF THE DEVELOPMENT SYSTEM

The system development tasks are prepared in collaboration with I4U [5]. The speech segments for all 1918 SRE'12 target speakers are first obtained from SRE'06–10 corpora. Two sets of speaker verification tasks are prepared, namely *Dev* and *Eval*, so that the generalization capability and calibration performance of the systems can be evaluated. The following aspects are considered: (a) the test utterances used in these two tasks are non-overlapping; (b) the *Dev-Test* and *Dev-Train* utterances are also included in *Eval-Train* so that the latter contains the maximum amount of enrollment data per speaker; (c) training and test segments have different Linguistic Data Consortium (LDC) labels (LDC-ID) ensuring a channel mismatch; (d) held out speakers' data from SRE'06 are included in tests to serve as unknown non-target speakers; (e) both telephone and interview speech is used for enrollment if available; (e) for each speech file in training and test sets, two different artificially noised versions (i.e., based on either SNR or noise type) are generated; (f) the 100 speakers data released during evaluation are included in *Eval-Train*; (g) test utterances are cropped randomly to have an active speech duration between 20 s-160 s. In this process, all SRE'12 target speaker data are divided into three disjoint sets where set-1 includes *Dev-Train*, set-2 include *Dev-Test* and *Eval-Train*, and set-3 includes *Eval-Test*. We used the target speakers data in *Dev-Train* for hyper-parameter estimation and discriminative training.

2.1. Noisy file Generation

We collect 10 Heating, Ventilation, and Air Conditioning (HVAC) type noise files from [6] and generated 10 crowd noise files by summing 500–800 NIST SRE utterances from both male and female speakers. The noise files are separated into three disjoint sets with set-1 and set-2 having 6 files each, and set-3 has 8 files (noise types are balanced in each set). We employ our in-house tools to generate the noisy files with the psophometric weighting (ITU-T Recommendation O.41) method as suggested by NIST. The active speech level is measured according to the ITU-T Recommendation P.56. These files are used for speaker enrollment, hyper-parameter [4] and back-end training. For degrading the test files (in *Dev-Test* and *Eval-Test*), we adopted FaNT toolkit with G-712 weighting, to be consistent with I4U. For each training and test file in *Dev* and *Eval*, 6 dB and 15 dB noisy versions are generated. The noise file (and type) is selected randomly from the corresponding set to which the utterance belongs.

2.2. Short Duration Segments

We truncate the test files to have active speech durations of 20 s to 160 s [3] with a 20 s increment. The Voice Activity Detection (VAD) method VAD-2 (Sec. 3.1.2) is used to find the speech segments. The duration values are assigned to the test files randomly. If the assigned duration is larger than the total active speech duration, the full utterance is used. These mixed duration files are used as test segments in the Dev/Eval task. Fig 1 (a)-(d) shows the active speech duration distributions computed for *Eval-Test* and SRE'12 core test recordings using both VADs utilized. Since VAD-2 is used to truncate the *Eval-Test* segments, the histogram shows impulses at the truncation locations. We expect that the use of mixed duration in the development should bring the test conditions closer to that of the actual SRE'12

[⋆]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.

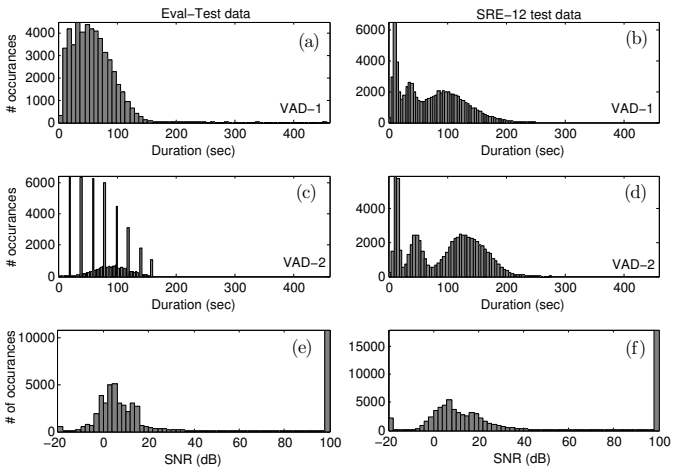

Fig. 1. Histogram plots of active speech duration and SNR values obtained from *Eval-Test* and SRE'12 core test segments. Panels (a–d) show duration, and (e–f) show SNR distributions. Durations in (a–b) and (c–d) are obtained from VAD-1 and VAD-2, respectively.

evaluation and also benefit the fusion and calibration training.

3. SYSTEM COMPONENTS

3.1. Voice Activity Detection (VAD)

3.1.1. VAD Algorithm-1 (VAD-1)

In this algorithm [7], to remove silence and low energy speech segments, a two stage voice activity detection (VAD) is performed. In the first stage, which is used before feature extraction, a soft VAD based on perceptual spectral flux and several voicing measures is utilized to remove the non-speech segments. This strategy saves computations, since in this manner, features are only extracted from speech segments. In the second stage, which is applied after feature extraction, an energy based method is employed to drop low-energy speech frames as well as any residual non-speech frames from the soft VAD in the first stage. These low energy frames are easily affected by noise and channel variabilities, and do not carry much speaker-dependent information.

3.1.2. VAD Algorithm-2 (VAD-2)

The main algorithm used in this VAD closely follows [8]. VAD is performed on both interviewee (A) and interviewer (B) channel, and speech segments detected in channel B is removed from channel A. Since the channel B is usually corrupted by a noise floor to mask the interviewee speech, spectral subtraction [9] is always performed before VAD on channel B. For channel A, first the SNR is estimated using a 2-mixture GMM trained on segment energy. If the SNR is less than 18 dB, spectral subtraction is performed before VAD.

3.2. Acoustic Features

Before feature extraction, all waveforms are first down-sampled to 8 kHz, and blocked into 25 ms frames with a 10 ms skip-rate. All our features use 12 cepstral coefficients and log-energy/C_0, appended with the first and second order time derivatives, thus providing 39 dimensional feature vectors. Individual features are described below.

3.2.1. Mean Hilbert Envelope Coefficients (MHEC)

MHEC features have been shown to be an effective alternative to the conventional MFCCs for robust SID under reverberant and noisy mismatched conditions [10, 11]. A block diagram illustrating the procedure for extracting the MHECs is depicted in Fig. 2.

First, the pre-emphasized speech signal $s(t)$ is decomposed into 24 bands through a 24-channel Gammatone filter-bank covering the frequency range of 300–3400 Hz. Next, the Hilbert envelope $e_s(t, j)$ is calculated and smoothed using a low-pass filter with a cut-off frequency of 20 Hz. In the next stage, the low-pass filtered $e_{sn}(t, j)$ is blocked into frames of 25 ms duration with a skip rate of 10 ms. To estimate the temporal envelope amplitude in frame l, the sample mean $S(l, j)$ is computed. Note that $S(l, j)$ is a measure of the spectral energy at the center frequency of the j^{th} channel, and therefore provides a short-term spectral representation of the speech signal $s(t)$. The next two stages (i.e., log compression, DCT, delta calculation) are commonly used in the extraction of conventional cepstral features such as MFCCs. Here, only the first 12 coefficients (excluding C_0) are retained after DCT and appended with the log-energy for each frame. The final output is a matrix of 39-dimensional cepstral features, entitled the mean Hilbert envelope coefficients (MHEC). The MHEC features are further processed through cepstral mean and variance normalization (CMVN). It is worth noting here that MHECs are extracted from the audio signals pre-processed with VAD-1.

3.2.2. PMVDR Front-End

The power spectrum estimation method used in the extraction of MFCC features is not robust to noise and channel degradations, resulting in large variations in estimated parameters. To alleviate this, a noise robust perceptual spectrum estimation technique with minimum variance was proposed in [12]. The acoustic features extracted using the perceptual Minimum Variance Distortionless Response (MVDR) spectrum have been shown to outperform the conventional MFCCs under noisy conditions for ASR [12] as well as speaker recognition applications [13]. In our system, the PMVDR features are extracted from audio files pre-processed with VAD-1. The PMVDR features are post-processed with CMVN.

3.2.3. Rectangular Filter-Bank Cepstral Coefficients (RFCC)

The RFCC front-end is inspired by perceptual linear prediction (PLP) cepstral features [14]. The original Bark frequency trapezoid filters are replaced by a bank of 24 uniform non-overlapping rectangular filters distributed over a linear frequency scale. The block scheme of the RFCC front-end is shown in Fig. 3. RFCC was initially proposed for robust ASR in noisy/Lombard speech conditions (*20Bands-LPC*) [15]. The tools and a recipe for RFCC extraction are available at [16]. RFCCs are extracted using an open source feature extraction and enhancement tool CTUCopy [17] and normalized using conventional feature Gaussianization [18].

3.2.4. MFCC-QCN-RASTA$_{LP}$

This front-end uses the conventional MFCC features extracted with HTK. Number of channels in the mel filter-bank is 24, and only the first 12 cepstral coefficients along with the log energy are retained and appended with delta and double delta coefficients. This feature stream is processed by Quantile Cepstral Normalization (QCN) [15] with percentile 1 and RASTA$_{LP}$ [19].

3.3. Feature Normalizations

3.3.1. Quantile-Based Cepstral Normalization (QCN)

Similar to cepstral mean-variance normalization (CMVN), QCN [15] aims at minimizing the mismatch between distributions of training and test samples. Unlike CMVN, QCN does not make any assumptions about the distribution properties, and instead performs an alignment of the sample dynamic ranges estimated from distribution quantiles. In our previous studies, QCN provided superior performance gains in ASR under noise and Lombard effect [15] and reverberation [20] compared to other popular normalizations.

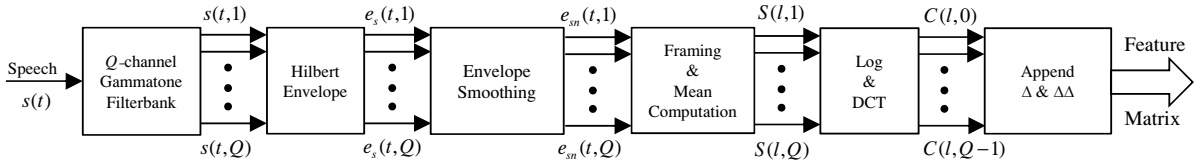

Fig. 2. Block diagram of the MHEC feature extraction framework. The symbols represent the output signals at each stage.

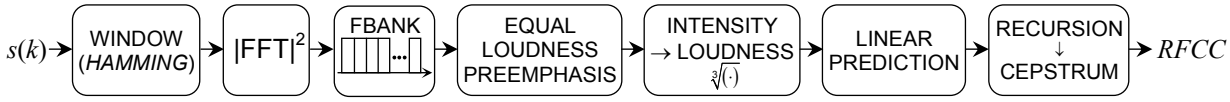

Fig. 3. Block diagram of the Rectangular frequency cepstral coefficient (RFCC) feature extraction scheme.

3.3.2. RASTA$_{LP}$

Temporal filtering is known to reduce the effects of noise and reverberation on speech systems. Recently proposed RASTA$_{LP}$ [19] is a low-pass filter that approximates the low-pass component of the popular RASTA filter [21]. Due to the low order of the RASTA$_{LP}$ filter, the adverse transient effects seen in original RASTA are significantly reduced. In addition, RASTA$_{LP}$ bypasses the mean subtraction functionality of RASTA and can be conveniently combined with distribution normalizations of choice. In our previous ASR studies, RASTA$_{LP}$ considerably outperformed RASTA in noisy, Lombard effect, and reverberated conditions [22, 20].

3.4. UBM Training

Gender dependent 1024-mixture UBMs with diagonal-covariance matrices are trained on telephone utterances selected from the Switchboard-II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the SRE'04-06 enrollment data. Initial four iterations per mixture are gradually increased to 15 for higher order mixtures. For front-end and VAD-2 tuning, we employed data sub-sampling for fast UBM training [23, 24] to perform a large number of experiments. Data sub-sampling is not used for the final evaluation.

3.5. I-vector Extractor Training

For training the i-vector extractor, the UBM training dataset and additional SRE'12 target speakers' data are used (both clean and noisy versions). Here, 600-dimensional i-vectors are extracted using 5 EM iterations. The i-vectors are first mean normalized and then length normalized using radial Gaussianization [25].

3.6. Back-end Classifiers

All the back-end classifiers used in this work utilize noisy and short duration utterances from *Dev-Train* along with the UBM data for training and/or impostor modeling. More details about our back-ends can be found in [26].

3.6.1. I-vector averaged PLDA (PLDA-1)

This is a standard PLDA back-end [27, 28, 29]. I-vector dimension is reduced to 400 using LDA first, then centering and Radial Gaussianization [25] is performed. A diagonal covariance noise based PLDA model with 400 eigenvoice dimensions is used. I-vectors are averaged across multiple segments for speaker enrollment.

3.6.2. Gaussianized Cosine-Distance Scoring (GCDS)

The i-vectors of multiple sessions of the same enroll speakers are first averaged, then Gaussianized with the mean and variance of the development set. LDA is performed for dimensionality reduction, and then cosine distance metric is employed for scoring. Finally, the scores are Gaussianized for each test utterance by the mean and variance of the scores obtained across all the enrollment utterances.

3.6.3. L2-Regularized Linear Regression (L2LR)

In this back-end, an L2-regularized logistic regression is applied using the LIBLINEAR package [30]. One-versus-the-rest approach is used for classifier training. Also, i-vector averaging for speaker enrollment and LDA is applied.

3.6.4. UBS-SVM Anti-Model (UBS-SVM)

The framework is based on SVM anti-modeling as presented in [31]. In this technique, instead of searching for the optimal number of background speakers, a universal background dataset is derived so as to embed impostor speaker knowledge in a balanced way. A cosine kernel is used in the UBS-SVM backend as described in [31].

3.6.5. Score-Averaged PLDA (PLDA-2)

In this back-end, instead of averaging the i-vectors coming from the same enrollment speaker, each test i-vector is first scored against each speaker's individual i-vectors using the PLDA model (as in PLDA-1). Next, the log-likelihoods obtained from the i-vectors of the enrollment speaker is averaged for each test i-vector.

3.7. Score Fusion and Calibration

System calibration and fusion is performed using the Bosaris toolkit [32] utilizing side information/quality measures. The linear logistic regression fusion algorithm is used. The best results were obtained using the active speech duration measured using VAD-1 as quality measures [7]. For the speaker model quality measure, the average active speech duration of all enrollment sessions is used. For test segments, the active speech duration of the corresponding utterance is used. The duration value is directly used as the quality measure function (QMF). An estimate of the SNR computed using the WADA algorithm [33], is also obtained as a secondary side information. Histogram plots of SNR estimates obtained from *Eval-Test* and SRE'12 core test utterances are shown in Fig 1 (e) and (f), respectively. The fusion and calibration is trained on the *Dev* and tested on *Eval*, except for some extended and supplemental submissions, which were trained on *Eval* and blindly applied on the SRE'12 trials. This is done in anticipation of a larger number of trials in those tasks.

4. CRSS SUBMISSION RESULTS

Total 15 sub-systems are constructed by various front-end and back-end combination as summarized in Table 1. The %Equal Error Rate (EER), minC_{primary} and C_{primary} cost functions obtained from these systems in the *Dev* and *Eval* tasks for both genders are also shown. In Table 2, the performance gain obtained by using active speech duration and SNR as quality measure is summarized. These results demonstrate the effectiveness of fusing the CRSS front-end and back-end combinations, resulting in relative improvements in the order of

Table 1. CRSS-UTD Sub-system Results Using Mixed Duration Train/Test

#	Feature/VAD/Norm	Back-end	Male						Female					
			Dev			Eval			Dev			Eval		
			%EER	C_{primary}	$\min C_{\text{primary}}$	%EER	$\min C_{\text{primary}}$	C_{primary}	%EER	$\min C_{\text{primary}}$	C_{primary}	%EER	$\min C_{\text{primary}}$	C_{primary}
1	MHEC-VAD1-CMVN*	PLDA-1	1.358	0.160	0.203	1.934	0.199	0.265	2.496	0.241	0.240	2.448	0.260	0.294
2		GCDS	1.845	0.194	0.378	1.460	0.180	0.359	2.156	0.283	0.473	1.663	0.238	0.440
3		L2LR	2.761	0.248	1.377	2.206	0.223	1.163	3.884	0.338	1.366	2.590	0.255	1.032
4		UBS-SVM	1.905	0.173	0.381	1.460	0.161	0.356	2.293	0.261	0.477	1.719	0.215	0.440
5		PLDA-2	1.139	0.161	0.399	1.382	0.198	0.508	2.163	0.248	0.420	1.709	0.257	0.497
6	RFCC-VAD2-Warp	PLDA-1	1.325	0.183	0.193	1.883	0.218	0.243	2.353	0.249	0.235	2.283	0.259	0.267
7		GCDS	1.690	0.204	0.350	1.354	0.190	0.328	2.066	0.259	0.427	1.379	0.218	0.394
8		L2LR	2.365	0.238	1.274	1.810	0.207	1.288	3.286	0.317	1.367	1.955	0.229	1.189
9		UBS-SVM	1.753	0.188	0.360	1.423	0.188	0.336	2.200	0.239	0.430	1.442	0.200	0.399
10		PLDA-2	0.990	0.180	0.347	1.271	0.216	0.441	1.815	0.249	0.370	1.383	0.254	0.430
11	MFCC-VAD2-QCN-RASTA$_{LP}$	GCDS	1.684	0.210	0.396	1.422	0.190	0.376	2.225	0.296	0.496	1.869	0.249	0.466
12		UBS-SVM	1.749	0.193	0.395	1.428	0.173	0.372	2.424	0.265	0.499	1.929	0.229	0.469
13		PLDA-2	1.048	0.180	0.367	1.221	0.207	0.479	2.132	0.263	0.407	1.777	0.275	0.487
14	PMVDR-VAD1-CMVN	GCDS	1.842	0.199	0.380	1.394	0.179	0.356	2.076	0.258	0.449	1.480	0.216	0.416
15		PLDA-2	1.162	0.183	0.388	1.307	0.206	0.488	2.012	0.250	0.399	1.523	0.248	0.458

Table 2. Fusion and Calibration Performance on Mixed Duration EVAL set using Side Information

#	Systems Fused	Fusion Method	Side Information	Compound LLR	Male			Female		
					%EER	$\min C_{\text{primary}}$	C_{primary}	%EER	$\min C_{\text{primary}}$	C_{primary}
1	2,3,5,7,8,10	Linear	None	No	0.82	0.0884	0.0947	0.85	0.1147	0.1182
2	2,3,5,7,8,10	Linear	None	Yes	0.67	0.0866	0.0888	0.66	0.1076	0.1096
3	2,3,5,7,8,10	Linear+quality	SNR,Duration	No	0.75	0.0881	0.0936	0.69	0.1104	0.1148
4	2,3,5,7,8,10	Linear+quality	SNR,Duration	Yes	**0.64**	**0.0864**	**0.0886**	**0.59**	**0.1040**	**0.1043**

Table 3. Performance of selected CRSS submissions in SRE'12 core, extended and supplemental tasks

Submission name/task	Systems Fused	Fusion Method	Side Info	$\min C_{\text{primary}}$(min cost)					C_{primary} (act cost)				
				c-1	c-2	c-3	c-4	c-5	c-1	c-2	c-3	c-4	c-5
CRSS_core_01_primary	1–4,6–9	Linear+quality	Duration	0.305	0.273	0.227	0.250	0.285	0.506	0.289	1.122	0.271	0.311
CRSS_core_02_alternate	{2,3,5,7,8,10},{11,13}†	Linear+quality	Duration	**0.209**	**0.180**	0.203	**0.184**	**0.180**	**0.226**	0.274	**0.433**	**0.261**	0.312
CRSS_core_03_alternate	2,3,5,7,8,10,11,13	Linear+quality	SNR,Duration	0.241	0.214	**0.201**	0.215	0.212	0.379	0.261	0.725	0.282	0.289
CRSS_core_04_alternate	2,3,5,7,8,10,11,13	Linear+quality	Duration	0.241	0.208	0.202	0.210	0.203	0.369	**0.252**	0.701	0.280	**0.277**
CRSS_core_05_alternate	2,3,5,7,8,10,11,13	Linear	None	0.270	0.226	0.246	0.215	0.253	0.282	0.378	0.597	0.317	0.430
CRSS_01_ext_primary	{2,5,7,10},{11,13}†	Linear+quality	SNR, Duration	0.162	0.228	0.107	0.285	0.259	0.192	0.368	0.113	0.391	0.423
CRSS_02_ext_alternate	2,5,7,10,11,13	Linear+quality	SNR, Duration	0.163	0.214	0.106	0.261	**0.241**	0.231	0.416	0.122	0.429	0.470
CRSS_04_ext_alternate	{2,5,7,10},{11,13}†^{E}	Linear+quality	SNR, Duration	0.168	0.215	0.106	0.246	0.247	0.199	0.333	0.109	0.363	0.388
CRSS_05_ext_alternate	2,5,7,10,11,13^E	Linear+quality	SNR, Duration	**0.159**	**0.210**	**0.098**	**0.244**	0.242	**0.173**	**0.311**	**0.100**	**0.340**	**0.365**
CRSS_01_sup_primary	2–5,7–15	Linear+quality	Duration	-	-	0.129	-	0.172	-	-	0.139	-	0.179
CRSS_02_sup_alternate	2–5,7–15^E	Linear+quality	Duration	-	-	0.127	-	0.169	-	-	0.136	-	0.177

* Results from this front-end (VAD-1), are sub-optimal compared to the front-ends using VAD-2, since the test files are cropped using VAD-2 for the mixed duration tests (see Fig. 1).
† The systems in braces were first linearly fused using equal weights. Next, these fused scores were again fused using quality measures.
^{E} Indicates that the fusion and calibration is trained on the *Eval* task instead of *Dev*. These are "blind" submissions since the fusion/calibration performance is not tested.

$50-60\%$ with respect to all three performance metrics. We also report the performance measures after applying the compound log-likelihood ratio (LLR) transformation [34], though only C_{primary} is a valid metric for compound LLRs.

Selected CRSS submissions for the 2012 NIST SRE core-core, core-extended and core-supplemental tasks are summarized in Table 3 along with the NIST reported performance measures. Results are shown in five common conditions defined as [3]: train on multiple segments and test on: 1) clean interview speech, 2) clean phone call speech, 3) artificially noised interview speech, 4) artificially noised phone call speech, and 5) phone call speech collected in a noisy environment. Consistent with our *Dev-Eval* experiments, we observe significant performance gains when quality measures are used. Training the fusion on the larger *Eval* set also provided some benefit in the extended submissions. In general, the best results are obtained when systems are first linearly fused using equal weights, and then the resulting scores are again fused using quality measures. It may be noted that only the "primary" core-core submission is developed using full duration train/test utterances alone, resulting in a sub-optimal performance. It is interesting to note that even with short utterances included in PLDA, side information is still useful for calibration [35].

5. CONCLUSIONS

We have described the CRSS site speaker recognition system submitted to the 2012 NIST SRE. The systems developed were a fusion of i-vector based sub-systems using four different front-ends and five different back-ends. To address the noisy conditions in SRE'12, artificially noised data were used for speaker enrollment, total variability model training and PLDA. To deal with the mixed duration test utterances, short speech segments were included in PLDA training and the development test trials. Duration and SNR values were used as quality measures. A significant performance gain was achieved using the presented strategies in preparation for the SRE'12 evaluation.

6. ACKNOWLEDGEMENTS

We would like to thank all the members of I4U for their discussion and support on the development task preparation, especially Rahim Saeidi, David A. van Leeuwen and Kong-Aik Lee; CRSS members Ali Ziaei, Keith W. Godin and Abhinav Misra for their support; and finally, Niko Brummer for his suggestions on fusion and calibration.

7. REFERENCES

[1] "The NIST year 2008 speaker recognition evaluation plan," 2008. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/2008/sre08_evalplan_release4.pdf

[2] "The NIST year 2010 speaker recognition evaluation plan," 2010. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/2010/NIST_SRE10_evalplan.r6.pdf

[3] "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

[4] 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.

[5] H. Li *et. al*, "The I4U system in NIST 2008 speaker recognition evaluation," in *Proc. IEEE ICASSP*, April 2009, pp. 4201–4204.

[6] [Online]. Available: www.freesound.org

[7] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised Speech Activity Detection using Voicing Measures and Perceptual Spectral Flux," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 197–200, Dec. 2012.

[8] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[9] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[10] S. O. Sadjadi and J. H. L. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. InterSpeech*, Makuhari, Japan, Sept. 2010, pp. 2138–2141.

[11] ——, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 5448–5451.

[12] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, pp. 142–152, February 2008.

[13] A. D. Lawson, P. Vabishchevich, M. C. Huggins, P. A. Ardis, B. Battles, and A. R. Stauffer, "Survey and evaluation of acoustic features for speaker recognition," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 5444–5447.

[14] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[15] H. Bořil and J. H. L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Trans. Audio Speech Lang. Process.*, pp. 1379–1393, Sep. 2010.

[16] "Hynek Bořil' Homepage." [Online]. Available: http://www.utdallas.edu/~hynek/tools.html

[17] P. Fousek, "CTUCopy – universal speech enhancer and feature extractor," 2007. [Online]. Available: http://noel.feld.cvut.cz/speechlab/

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

[19] H. Bořil and J. H. L. Hansen, "UT-scope: Towards LVCSR under Lombard effect induced by varying types and levels of noisy background," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 4472 – 4475.

[20] S. O. Sadjadi, H. Bořil, and J. H. L. Hansen, "A comparison of front-end compensation strategies for robust LVCSR under room reverberation and increased vocal effort," in *Proc. IEEE ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4701–4704.

[21] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, 2002.

[22] H. Bořil, F. Grézl, and J. H. L. Hansen, "Front-end compensation methods for LVCSR under Lombard effect," in *Proc. InterSpeech*, Florence, Italy, Aug. 2011, pp. 1257–1260.

[23] T. Hasan, Y. Lei, A. Chandrasekaran, and J. H. L. Hansen, "A novel feature sub-sampling method for efficient universal background model training in speaker verification," in *Proc. IEEE ICASSP*, Dallas, TX, March 2010, pp. 4494 – 4497.

[24] T. Hasan and J. H. L. Hansen, "A study on universal background model training in speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, pp. 1890–1899, Sep. 2011.

[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] G. Liu, T. Hasan, H. Bořil, and J. H. Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *Proc. IEEE ICASSP*, Vancouver, Canada, May. 2013.

[27] 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.

[28] 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.

[29] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. IEEE ICASSP*. IEEE, 2012, pp. 4253–4256.

[30] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

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

[32] N. Brummer and E. de Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," in *NIST SRE Analysis Workshop*, Atlanta, GA, Dec. 2011.

[33] C. Kim and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," *Proc. InterSpeech*, pp. 2598–2601, 2008.

[34] N. Brummer, "SRE'12 - BOSARIS Toolkit." [Online]. Available: https://sites.google.com/site/bosaristoolkit/sre12

[35] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration Mismatch Compensation for I-vector based Speaker Recognition Systems," in *Proc. IEEE ICASSP*, Vancouver, Canada, May. 2013.