Promoting robustness for speaker modeling in the community: the PRISM evaluation set

Luciana Ferrer*, Harry Bratt*, Lukas Burget*, Honza Cernocky†, Ondrej Glembek†, Martin Graciarena*, Aaron Lawson*, Yun Lei*, Pavel Matejka†, Olda Plchot†, Nicolas Scheffer*
*Speech Technology and Research Laboratory
SRI International, Menlo Park, California, USA
Email: lferrer,harry,burget,martin,aaron,yunlei,scheffer@speech.sri.com
†Brno University of Technology
Czech Republic
Email: cernocky,glembek,matejkap,iplchot@fit.vutbr.cz

***Abstract*—We introduce a new database for evaluation of speaker recognition systems. This database involves types of variability already seen in NIST speaker recognition evaluations (SREs) like language, channel, speech style and vocal effort, and new types not yet available on any standard database like severe noise, and reverberation. The database is created using data from NIST SREs from 2004 to 2010. We present results of a state-of-the-art system on the different subset of this database. The database will be publicly available, and this work aims at encouraging other sites to adopt it and improve it.**

I. Introduction

New challenges face the speaker recognition community at every NIST SRE evaluation. The last few years have seen a dramatic increase in the number of trials and the amount of data to be processed, the introduction of a new type of speech style (interview) and a wide variety of microphones, as well as the introduction of speech recorded with different vocal effort levels. Despite these challenges, speaker recognition accuracy hits a record high evaluation after evaluation, thus questioning the need for further research on the problem.

This work aims at introducing (or re-introducing) challenging variabilities in a speaker evaluation set in order to drive the research toward the creation of systems that can better handle the problems faced in realistic conditions and improving their robustness.

The PRISM (Promoting Robustness in Speaker Modeling) evaluation set is a very large speaker recognition set based on NIST SRE data released from 2004 to 2010, where the scope is extended to additional types of variabilities namely, noise and reverberation. In addition, it includes variabilities already seen in one or more NIST SREs namely, language, channel type, speech style and vocal effort level.

The language condition leverages data from multiple Mixer corpora [1] to assess speaker recognition performance under multiple languages, including same-language and cross-language trials. The reverb and noise conditions are created from a clean data set that is artificially degraded at different signal-to-noise ratio (SNR) levels, using different real noises, and different reverb delays and room types. These simulated sets are carefully crafted so that audio files and tools used to simulate these degradations are all openly available and at no cost. The other conditions use data from SRE08 and SRE10 to address the effect of channel type, speech style and vocal effort level.

The PRISM set comprises three main pieces of information:

- Definition of multiple trial sets including the different types of variability.
- A recipe to create the simulated degraded data.
- A unified list of labels for all previously released Mixer data, Switchboard data and Fisher data, with standardized naming conventions as well as normalized names for the available metadata. This is the key information that was used to build the trial definition for the different trial sets, and is of crucial need for the researchers to build their training sets.

The PRISM set will be open to the community, and this work aims at encouraging other sites to adopt it. Results and analysis from the authors will be shown as reference on this new and innovative evaluation set.

II. Baseline System

To give the reader a reference to compare against, we present results for our baseline system on some of the conditions included in the database. The system used for this purpose is a state-of-the-art Mel frequency cepstral coefficient (MFCC) iVector system with probabilistic linear discriminant analysis (PLDA) modeling [2], [3], [4]. Nineteen cepstral coefficients along with the energy with appended deltas and double deltas are used. The background model is a gender-dependent 2048-component diagonal covariance Gaussian mixture model. iVectors of dimension 600 are extracted using a total variability subspace model. This dimension is further reduced using linear discriminant analysis (LDA) to 250. Then, length normalization is used to gaussianize the distribution of the iVectors. Finally, the distribution of the normalized iVectors is modeled and a score for each trial is obtained using full-rank probabilistic LDA (PLDA). The training data used for the background models, iVector extractor, LDA and PLDA is described in Section III-C.

Results are shown in terms of equal error rate (EER) and decision cost function (DCF) as recently defined by NIST for the core condition of 2010 SRE [5].

III. PRISM Evaluation Set

The PRISM evaluation set is created using data from all NIST SREs beginning with the year 2004 (that is, SREs 2004, 2005, 2006, 2008 and 2010). Fisher [6] and Switchboard [7] data is also included in the database, although used only for training purposes, not to create evaluation trials.

The evaluation set is divided into different subsets designed to test the effect of different kinds of variability: language, noise, reverberation, speech style and channel, and vocal effort. In addition, the SRE10 conditions for 1-side and 8-side training are included as separate sets (**sre10** and **sre108s**) for ease of comparison with previous results.

Only waveforms from the SRE databases of lengths that were included in the core conditions in the corresponding evaluations are used to create trials. That is, waveforms created by NIST to test 10 or 30 second conditions are not included in the trial definitions, although they are present in the metadata files.

A. Unified Metadata Files

To create the PRISM evaluation set and its accompanying training set, we had to standardize the information available on all the different NIST SRE databases, and Fisher and Switchboard corpora. Most available pieces of information in each database were recorded in a set of metadata files with a single unified format. For each available waveform corresponding to a recorded channel, these files list the following information:

- **Database id:** Original database under which the waveform was released (e.g., Fisher 1, SRE 2004).
- **Speaker pin:** A unique id for the speaker present in the channel. Summed channel waveforms are not considered in the PRISM evaluation set. Hence, a single speaker pin corresponds to each waveform even though other speakers might be hearable in the waveform through cross-channel effects.
- **Session name:** An id for the recording session. This is the name used for the session in the original database. Note that the same session id might have been used by NIST in different databases to refer to totally different recordings. The only way to tell whether two recordings are the same is by using the session id below along with the channel, channel type, and speech type.
- **Channel:** The recording channel: A or B for stereo waveforms, or X for one-channel waveforms.
- **Session id:** The original filename from where the waveform was extracted by NIST for the SRE. In some SRE databases, original waveforms from a certain recording session were used to generate various waveforms of different lengths. Also, in some SRE databases, several recordings using different microphones were made during the same session. This information is essential to avoid creating trials that test and train on different parts or microphones from the same recording session. This information is available only for some databases. When not available, we assume that a single waveform was extracted from each session.
- **Gender:** Female or male
- **Year of birth:** Year of birth of speaker in recording, if available.
- **Year of recording:** Year in which recording was made, if available.
- **Age:** Age of speaker at time of recording, if available. Age, year of birth and year of recording are redundant. If two of them are available, the third one can be trivially calculated. Nevertheless, some databases only contain year of birth. For this reason, we decided to keep the three fields in our metadata files.
- **Speech type:** Either telephone conversation or interview.
- **Channel type:** Either a telephone channel or a microphone id as defined in the original database. The same microphone number might correspond to different devices in different original databases. Hence, the database id is appended to the microphone id to avoid confusion.
- **Nominal length:** For SRE data, this information is derived from the task in which the waveform was used and it can be 10 seconds, 300 seconds (5 minutes), and so on.
- **Language:** Language spoken in the recording. This field is normalized across databases to use the same three-letter name for each language.
- **Native language:** Native language for the speaker in the recording.
- **Vocal effort:** Vocal effort level as prompted during the recording (this is not necessarily the actual vocal effort perceived when listening to the signal). This information is available only for a subset of SRE10 data.

B. Evaluation Subsets

We explain in detail the different subsets in the PRISM evaluation set, starting with a description of the clean corpus used as a basis to create the noise and reverb sets.

1) **Clean Corpus:** The noise and reverb sets are created by adding real noise (i.e., recorded noise samples) and reverberation to data extracted from the SRE10 and SRE08 corpora. Only clean microphone data is selected from those corpora. Specifically, microphone 2 (lavalier microphones) waveforms are chosen from both interview and telephone conversations. Only SRE08 data is used for training, while SRE10 data and a small portion of SRE08 data is used to create trials (enrollment and test). The clean trials are created as the Cartesian product of the sessions selected for this purpose (except for same-session trials, which are discarded). That is, all possible target and impostor samples are created for the selected list of clean sessions.

2) **Noise Set:** We selected 15 cocktail noise samples from the free sound repository Freesound.org [8]. These noise samples were collected in bars, cafeterias, offices, and airports. We inspected the samples to remove single-speaker foreground

speech sounds and artifacts (e.g., clicks). The noise samples vary in duration from 1 to 13 minutes and are labeled 1 to 15. We added these 15 noise samples to the full waveforms from the clean corpus described above at 20, 15, and 8 dB SNRs, using the publicly available tool called FaNT [9].

Different noises are added to training, enrollment, and test samples. This avoids the highly optimistic matched case in which the same type of noise is observed when training the systems as in enrollment or test samples, or even in just enrollment and test. Hence, noises are separated into three disjoint sets: enrollment noises corresponding to noise samples 1 through 4, test noises corresponding to samples 5 through 8, and training noises corresponding to noise samples 9 through 15. A randomly chosen sample from each of these groups is selected to be added to signals in each corresponding group.

The noise trials are created following the clean trial definition, where the clean enrollment sample has been degraded by one of the enrollment noises (at a certain SNR level) and the clean test sample has been degraded by one of the test noises (at a possibly different SNR level). Different conditions are created by matching different SNR levels for enrollment and test. Table I shows the number of target and impostor trials in all evaluation conditions.

TABLE I

Number of target and impostor samples for each condition in the noise and reverb sets. The conditions with matched RT are 0.3 vs. 0.3, 0.5 vs. 0.5, 0.7 vs. 0.7 and clean vs. clean. The conditions with matched SNR are 8 dB vs. 8 dB, 15 dB vs. 15 dB, 20 dB vs. 20 dB and clean vs. clean. All matched conditions within the noise and reverb sets have the same number of target and impostor trials as indicated in the table. The conditions with mismatched RT or SNR are created by matching data with RT or SNR of X for enrollment and Y for test and conversely. The all vs. all condition is created by combining all of these conditions within each of the two sets (noise and reverb).

Eval. condition	# tgt	# imp
sets with matched RT or SNR	2450	592,508
sets with mismatched RT or SNR	4900	1,185,016
all vs. all	39,200	9,480,128

Figure 1 and Table II show the results on the different conditions within the noise set for our baseline system. We can see a clear trend of degradation as the SNR decreases from clean data to 20, 15 and 8 dB. In particular, the EER degrades around 9 times from the clean vs clean condition to the 8dB vs 8dB condition. More results on the noise set for multiple systems using different features and showing the effect of adding noisy data to the training data for PLDA can be found in [10].

3) **Reverb Set:** Reverberation is added to the clean signals at different reverberation times (RT) of 0.3, 0.5 and 0.7. Initially, a set of candidate rooms were generated using the rir tool [11], which allows for the modeling of a room impulse response for parameters of room size, microphone and speaker location, wall, floor and ceiling reflection coefficients, speed of sound, and so on. Our rooms were modeled so as to cover common configurations of size, reflectivity, and source and microphone locations and only those configurations resulting in RTs close to 0.3, 0.5 or 0.7 were used. In total, twelve rooms were modeled for train (four for each RT), three for test, and three for enrollment (one for each RT in each case). As for the noise set, all reverberation conditions were exclusive to a single set (train, test, enroll). The fconv tool [11] was then

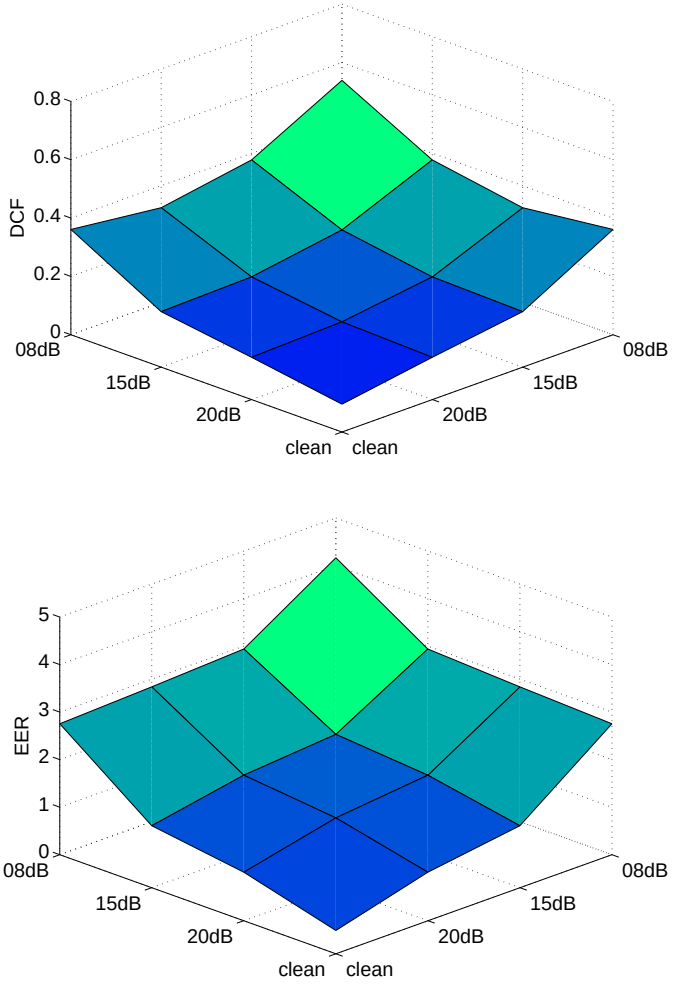

Fig. 1. DCF and EER of the baseline system on the different noise conditions listed in Table II. The x-axis and y-axis correspond to the SNR level in each session involved in the trials (enrollment and test).

TABLE II

Performance of the baseline system on different noise conditions.

Eval. condition	DCF	EER
08 dB vs. 08 dB	0.539	4.16
08 dB vs. 15 dB	0.377	2.94
08 dB vs. 20 dB	0.325	2.84
08 dB vs. clean	0.361	2.76
15 dB vs. 15 dB	0.248	1.84
15 dB vs. 20 dB	0.198	1.67
15 dB vs. clean	0.190	1.30
20 dB vs. 20 dB	0.155	1.47
20 dB vs. clean	0.145	1.02
clean vs. clean	0.095	0.49
all vs. all	0.282	2.18

used to generate the reverberated signals by convolving the room impulse responses with the audio files.

The reverb trials are created, as the noise ones, following the clean trial definition, where the clean enrollment sample has been degraded by one of the enrollment reverb types and the clean test sample has been degraded by one of the test reverb types. As for the noise set, different conditions are created by matching different RTs for enrollment and test. Table I shows the number of target and impostor trials in all evaluation conditions. The clean vs. clean set is identical to the one listed under the noise set.

Figure 2 and Table III show the results in the different conditions within the reverb set for our baseline system. Unlike what we see in Figure 1, the results for the reverb conditions are not as consistently correlated with the RT values of the enrollment and test samples. We expected this to be the case, since it is well known that RT is not as good a predictor of the speech signal quality for reverberated signals as SNR is for noisy signals [12], [13]. In fact, many factors affect the perceived quality of a reverberated signals, with RT value being only one of them. This partly explains the lack of a monotonic relationship between the RT values and the performance measures. Another factor that can explain this behavior is that the reverb dataset only contains one reverb type for each RT value in enrollment and test. This is something we plan to remediate in the near future by adding a wider variety of rooms with those RT values to this set. Nevertheless, we still see a clear degradation of around 5 times in EER from the clean vs. clean condition to the worst reverb condition.

TABLE III
Performance of the baseline system on different reverberation conditions.

Eval. condition	DCF	EER
RT 0.7 vs. RT 0.7	0.285	2.00
RT 0.7 vs. RT 0.5	0.421	2.45
RT 0.7 vs. RT 0.3	0.274	1.75
RT 0.7 vs. clean	0.229	1.22
RT 0.5 vs. RT 0.5	0.357	1.912
RT 0.5 vs. RT 0.3	0.378	1.96
RT 0.5 vs. clean	0.220	1.18
RT 0.3 vs. RT 0.3	0.220	1.31
RT 0.3 vs. clean	0.192	0.96
clean vs. clean	0.095	0.49
all vs. all	0.339	1.71

4) **Language Set:** The language set is created using telephone data from SRE05, SRE06, SRE08 and SRE10 databases. Five hundred speakers for which there is at least one session in some language other than English are selected randomly from SRE05, SRE06 and SRE08, leaving the rest of them for training purposes. An additional set of three hundred randomly chosen speakers that appear only in English conversations are selected from SRE10. Trials are then created as the Cartesian product of all sessions from these eight hundred speakers.

As in the case of the noise and reverb sets, different conditions are defined within the language set to assess the

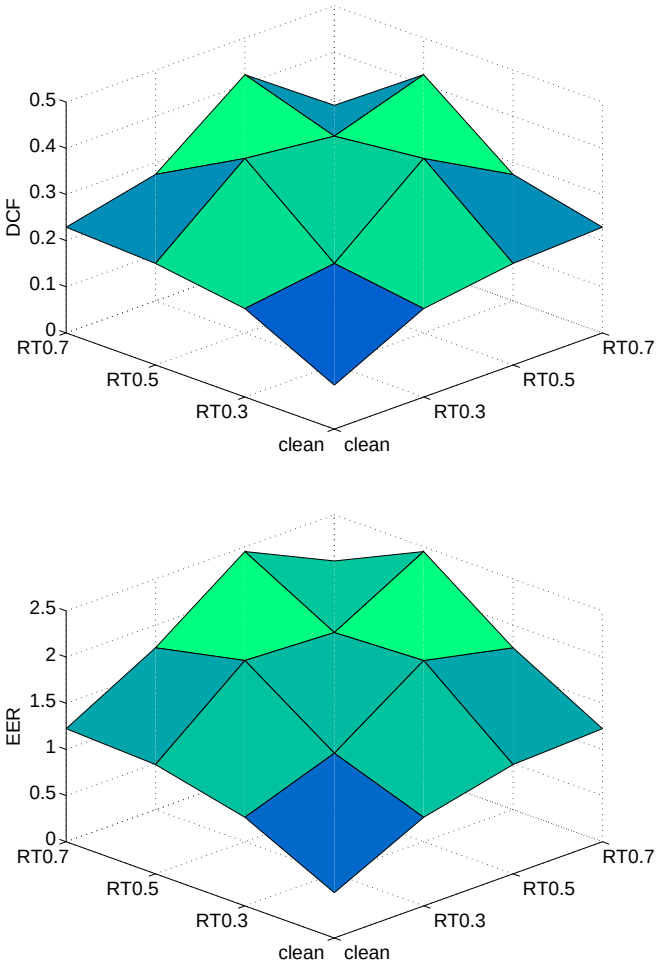

Fig. 2. DCF and EER of the baseline system on the different reverberation conditions listed in Table III. The x-axis and y-axis correspond to the RT in each session involved in the trials.

effect of language in the speaker recognition performance. Table IV shows the number of target and impostor samples and the performance on our baseline system for three of these conditions. Other smaller conditions are defined per language in the database. Results for these conditions are not shown in this paper since the number of trials for them is not large enough to result in a reliable performance estimation.

TABLE IV
Number of target and impostor samples and baseline performance for some conditions in the language set. The last condition is created by first downsampling the English vs. English trials to match the number of available Non-English vs Non-English trials. The languages included in the last two conditions are Arabic, Russian, Chinesse and Thai, the languages for which the per-language conditions have the most trials.

Eval. condition	# tgt	# imp	DCF	EER
English vs. English	28974	8463915	0.251	1.74
Non-English vs. English	5757	4344974	0.439	2.41
Lang X vs. Lang X	9894	428563	0.890	3.17

We can see that performance on English versus English

trials is clearly better than on mismatched trials or on matched trials that include languages other than English. The degradation on mismatched trials can be explained by the fact that detecting target trials must be harder (have lower scores) when enrollment is done on a language different from that found in testing. Impostor trials, on the other hand, will be easier (also have lower scores) for mismatched data but, apparently, this effect is less marked than the shift in the same direction of the target distribution. Finally, the fact that performance on matched trials that include languages other than English is worse than on matched English-only trials might because much less training data is available for languages other than English in our training set.

5) **Vocal Effort Set:** The vocal effort set is created using telephone data from 380 speakers in SRE10 who participated in at least one high or low vocal effort conversation. Sixty-six speakers with these characteristics were held out for training purposes. These speakers are not used in any of the other sets either, except in the SRE10 ones that replicate the definition used by NIST for the evaluation. In this paper, though, we do not use these held-out SRE10 speakers for training the system to ensure that performance shown on SRE10 sets is not overly optimistic.

The conditions created within this set are similar to those defined in SRE10 though they include all possible impostor samples using the sessions from the selected speakers and include a new set that merges all three vocal effort conditions into one. Table V shows the number of target and impostor samples for the four conditions defined within the vocal effort set along with the results on these conditions for our baseline system.

Results show that testing on high vocal effort while enrolling on normal, causes a big degradation in system performance. This, of course, affects the performance when all conditions are merged (last line in the table). A much smaller degradation is observed (and only on DCF) when testing on low vocal effort signals. This might be because low vocal effort (as elicited in this collection) does not affect the quality of the speech, but mostly only the volume. On the other hand, high vocal effort results in a marked effect in speech quality, clearly affecting the spectrum of the speech [14].

TABLE V

Number of target and impostor samples and baseline performance for the conditions in the vocal effort set. The last condition is created by first downsampling the normal vs normal and normal vs high trials to match the number of available normal vs low trials.

Eval. condition	# tgt	# imp	DCF	EER
normal vs. normal	28015	5022527	0.245	2.08
normal vs. low	3579	650034	0.370	1.87
normal vs. high	3933	752505	0.715	3.94
normal vs. all	10814	1950109	0.461	3.13

6) **Speech Style and Channel Set:** This set is created using sessions from all SRE10 speakers that are not held out for training (as described earlier) and 70 SRE08 speakers who participated in telephone conversations recorded over alternate microphones. The SRE08 speakers are added since SRE10 data corresponding to telephone conversations over alternate microphones includes only two microphones, while SRE08 data includes six different microphones. The trials are created as a Cartesian product of all the sessions from these speakers, avoiding trials that involve two waveforms coming from the same recording session but different microphones or different lengths.

Many different per-microphone conditions are defined over this set. Nevertheless, these conditions have relatively few trials, hence, here we present results only on conditions that merge different microphones. Table VI shows the number of target and impostor samples and performance for the baseline system for these conditions. The first two blocks aim at evaluating the effect of the channel. When both enrollment and test samples are interviews, the effect of having trials with matched or mismatched microphones is tested. As expected, the performance for the mismatched case is worse than for the matched case. When these two conditions are merged, after first downsampling the mistmached condition to match the number of trials in the matched one, the performance degrades even more. This happens because matched trials have generally higher scores than mismatched ones (for both target and impostor distributions). When merging the two conditions this shift (miscalibration) results in further degradation of performance.

TABLE VI

Number of target and impostor samples, and baseline performance for some conditions in the speech style and channel set. In this table, int stands for interview speech style, tel for telephone speech style, phn for telephone channel and mic for (alternate) microphone channel. Note that line 3 in the table is a subset of line 8 where trials have been downsampled to keep the balance between matched and mismatched microphone trials from lines 1 and 2. Also, lines 6 and 9 are identical and are repeated for ease of comparison within each block.

Eval. condition					
Speech style	Channel	# tgt	# imp	DCF	EER
int vs int	matched mic	7281	1903536	0.279	1.76
	mismatched mic	48490	11766137	0.402	2.20
	mic vs mic	14576	3807101	0.436	2.48
tel vs tel	phn vs phn	31670	6601291	0.251	2.194
	phn vs mic	17030	3927234	0.436	2.402
	mic vs mic	2468	580881	0.370	2.553
	all vs all	7361	1744500	0.430	2.880
int vs int	mic vs mic	55771	13669673	0.471	2.363
tel vs tel		2468	580881	0.370	2.553
int vs tel		16536	5663159	0.391	2.074
all vs all		7343	1744612	0.420	2.329

A similar effect can be seen when both enrollment and test samples are telephone conversations: the matched telephone channel performance is better than the mismatched one, in which one of the samples in the trial is a microphone recording. When both sides are microphone recordings, the performance is comparable to that for telephone channel versus microphone channel. This is reasonable, considering that a telephone channel can be simply seen as just another microphone. Finally, a balanced merge of all these three conditions results, as in the case of int vs int, in further

degradation of performance due to the shift (miscalibration) between the distributions for the individual conditions.

The final block in Table VI gives us an idea of the effect of speech style on system performance. Clearly, the effect is much smaller than that of channel variability, given that all four conditions in the table give relatively similar performance. From this, we conclude that the mismatch between interview and telephone conversations is not a big challenge to our system. This might simply be due to the fact that these interviews are very conversational in nature and elicit a similar kind of speech as in telephone conversations.

7) **SRE10 Sets:** These sets are added to the PRISM evaluation database for ease of comparison with older results on SRE10 conditions. Many SRE10 conditions, though, are represented in some other set within the database, and, in most cases, the conditions within those other sets include more trials than the original ones used in SRE10 (even though, in our sets, some speakers are held out for training purposes). This is true because the extended sets in SRE10 did not necessarily include all possible impostor and target samples, while our sets do. Furthermore, in some cases, our sets include data from other corpora apart from SRE10, increasing the size of the available list of sessions to create the trials.

Note that, while all other sets exclude any signals from the 66 held-out speakers (as explained in Section III-B5), this set does not. This is to ensure that the trials are identical to those released by NIST.

Table VII shows the number of target and impostor samples and baseline performance on the SRE10 conditions included in the PRISM evaluation set. Compared with performance levels demonstrated during SRE10, these results show that the baseline system for which all other results presented in this paper are shown is clearly a state-of-the-art system.

C. Training Data

For the background models, i-Vector extractor, LDA and PLDA training data was extracted from Fisher 1 and 2, Switchboard phase 2 and 3 and Switchboard cellphone phases 1 and 2, along with all Mixer speakers not used for any of the sets described above. This includes the 66 held out speakers from SRE10 (see Section III-B5), and 965, 980, 485 and 310 speakers from SRE08, SRE06, SRE05 and SRE04, respectively. A total of 13,916 speakers are available in Fisher data and 1,991 in Switchboard data.

Simulated noisy and reverberated signals were also added to the training set starting from a set of held-out lavalier mic data from SRE08. As mentioned above, the noises and reverberation parameters used on the training data were different from those used on enrollment and test data.

The complete list of sessions that can be used for training is available as part of the release of the PRISM database.

For our baseline system, for which results were presented in the previous sections, background models are trained using only Mixer data, while the i-Vector extractor is trained using every training session available from all databases. LDA and PLDA models are trained using all training data corresponding to speakers who participated in at least six sessions. This restriction discarded most of the Fisher speakers. Noisy and reverberated data are used only in the LDA/PLDA stage. As mentioned before, the 66 held-out SRE10 speakers are not used for training purposes in the experiments for this paper to allow for a fair assessment of performance on the SRE10 set.

IV. Conclusion

We presented the PRISM database for evaluation and training of speaker recognition systems. The data used to create this database comes from NIST SRE corpora from 2004 to 2010. Fisher and Switchboard corpora are also used, but for training purposes only. The database includes types of variability that have not been present in any standard large-scale evaluation database for speaker recognition: severe noise and reverberation. Furthermore, the database includes other types of variability that are found in different NIST speaker recognition evaluations: language, speech style, channel and vocal effort. In these cases, the proposed database expands the number of trials for these conditions, in many cases, including speakers from across SRE databases. The database defines meaningful conditions that can be used to assess the effect of the different types of variability, keeping all other factors as constant as possible.

We presented results on the defined conditions for a state-of-the-art MFCC system based on iVector/PLDA modeling. Results usually coincide with intuition as to which conditions should be harder than others, indicating that the database is a reasonable testbed for new methods designed to compensate for the different types of variabilities found in this data.

The PRISM database will be publicly available. We encourage the community to adopt it and perhaps debug, enlarge, or improve it in any way.

Acknowledgment

This work was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U. S. Government.

References

[1] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: The mixer 3, 4 and 5 corpora," in *Proc. Interspeech*, Antwerp, Aug. 2007.

[2] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 788–798, may 2011.

[3] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," June 2010, Keynote presentation.

[4] P. Matejka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," .

[5] "NIST SRE10 evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf.

[6] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *International Conference On Language Resources And Evaluation*, 2004.

TABLE VII

Number of target and impostor samples, and baseline performance for the conditions in the SRE10 set. The closest condition from the other PRISM sets is also indicated, where **vel** *refers to the vocal effort set, and* **sch** *refers to the speech style and channel set.*

Eval. condition		# tgt	# imp	DCF	EER	Matching condition in other sets
1-side training	1: int vs. int, matched mic	4304	795995	0.138	0.63	sch set, int vs. int, matched mic
	2: int vs. int, mismatched mic	15084	2789534	0.227	1.01	sch set, int vs. int, mismatched mic
	3: int vs. tel, mic vs. phn	3989	637850	0.304	1.88	
	4: int vs. tel, mic vs. mic	3637	756775	0.182	0.83	sch set, int vs. tel, mic vs. mic
	5: tel vs. tel, phn vs. phn	7169	408950	0.444	1.79	sch set, tel vs. tel, phn vs. phn
	6: tel vs. tel, normal vs. high vel	4137	461438	0.639	3.29	vel set, normal vs. high
	7: mic vs. mic, normal vs. high vel	359	82551	0.541	1.98	
	8: tel vs. tel, normal vs. low vel	3821	404848	0.344	1.29	vel set, normal vs. low
	9: mic vs. mic, normal vs. low vel	290	70500	0.146	0.69	
8-side training	5: tel vs. tel, phn vs. phn	442	687447	0.179	0.90	
	6: tel vs. tel, normal vs. high vel	236	52499	0.284	1.76	
	8: tel vs. tel, normal vs. low vel	223	46692	0.075	0.11	

[7] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," .

[8] "Freedsound," http://www.freesound.org.

[9] G. Hirsch, "Fant," http://dnt.kr.hs-niederrhein.de/download.html.

[10] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. ICASSP*, Kyoto, Mar. 2012.

[11] S. G. McGovern, "A model for room acoustics," http://www.2pi.us/rir.html.

[12] M. Wolf and C. Nadeu, "On the potential of channel selection for recognition of reverberated speech with multiple microphones," in *Proc. Interspeech*, Makuhari, Japan, Sept. 2010.

[13] K. Lebart and J. M. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acustica*, vol. 87, pp. 359–366, 2001.

[14] R. Goldenberg, A. Cohen, and I. Shallom, "The lombard effect's influence on automatic speaker verification systems and methods for its compensation," in *ITRE '06. International Conference on Information Technology: Research and Education*, 2006.