A large-scale open-source acoustic simulator for speaker recognition

Marc Ferràs, Srikanth Madikeri, Petr Motlicek, Subhadeep Dey and Hervé Bourlard
Idiap Research Institute, CH-1920 Martigny, Switzerland
{marc.ferras,srikanth.madikeri,petr.motlicek,subhadeep.dey,herve.bourlard}@idiap.ch

***Abstract*—State-of-the-art speaker recognition systems suffer from significant performance loss on degraded speech conditions and acoustic mismatch between enrolment and test phases. Past international evaluation campaigns, such as the NIST Speaker Recognition Evaluation (SRE), have partly addressed these challenges in some evaluation conditions. This work aims at further assessing and compensating for the effect of a wide variety of speech degradation processes on speaker recognition performance. We present an open-source simulator generating degraded telephone, VoIP and interview speech recordings using a comprehensive list of narrow-band, wide-band and audio codecs, together with a database of over 60 hours of environmental noise recordings and over one hundred impulse responses collected from publicly available data. We provide speaker verification results obtained with an i-vector based system using either a clean or degraded PLDA back-end on a NIST SRE subset of data corrupted by the proposed simulator. While error rates increase considerably under degraded speech conditions, large relative EER reductions were observed when using a PLDA model trained with a large number of degraded sessions per speaker.**

***Index Terms*—speaker recognition, degraded speech, codec, noise, robustness, simulation**

I. Introduction

A major challenge for speech technologies is to preserve a satisfactory performance in diverse environmental and recording conditions. For recognition tasks such as speech recognition or speaker recognition, error rates are known to rapidly grow as soon as acoustic mismatch between train/enrolment and test phases is present. Even in matched conditions, error rates are known to be high when speech signals are simply degraded. A number of approaches, at the signal or feature levels [1]–[5] or at the modeling level [6], [7], have been proposed to improve robustness of speaker recognition systems.

Factors such as environmental noise, reverberation, recording equipment or speech coding noise, amongst others, influence the quality of speech signals. While some of these factors may be controllable in some applications, the massive amount of available audio data on public archives such as YouTube, makes addressing robustness an even more challenging problem.

Past international speaker recognition evaluations, e.g. NIST Speaker Recognition Evaluation (SRE) 2005, 2006 and 2008 have highlighted the sensitivity of state-of-the-art systems to acoustic channel mismatch [8]. Researchers have proposed effective *session compensation* and *domain adaptation* techniques such as Joint Factor Analysis (JFA) [9] and Probabilistic Linear Discriminant Analysis (PLDA) [10], to deal with train/test mismatch with in-domain data. However, PLDA models have been shown to be sensitive to in-domain/out-of-domain mismatch between development and evaluation data with close conditions such as landline/cellular telephone speech resulting in a considerable increase in error rates [11].

The presence of environmental noise is also known to be a major source of degradation and it has been partially addressed in the latest NIST SRE 2012 [12]. This evaluation proposed assessing speaker recognition system performance on noisy speech data, corrupted with Heating, Ventilation and Air Conditioning (HVAC) noise as well as artificial crowd noise. Although participants [13], [14] developed systems to face these specific noisy conditions, significant performance loss was observed for noisy trials. In an unconstrained low-SNR noisy scenario, the error rate increase in the presence of noise may be dramatic enough to render a system unreliable.

Some techniques aiming at augmenting the intra-speaker variability of speech data have been explored for speech recognition in recent years. Vocal Tract Length Perturbation [15], augmenting data using random VTL Normalization factors, obtained significant improvements on the TIMIT data. Similar gains were reported for a stochastic feature mapping technique [16] synthesizing content based on speaker adaptation techniques. Speed changes were explored in [17], obtaining gains around 5% relative for several LVCSR tasks.

In this paper, we explore the augmentation of inter-session variability of a data set using an acoustic simulator for speaker recognition tasks. We evaluate a state-of-the-art i-vector speaker recognition system in the presence of both environmental noise, reverberation and telephone, radio and audio codecs. While the data provided in the NIST SRE has traditionally focused on sampling speaker variability, this work uses a database aimed at sampling degradation variability using a large number of simulated degradation processes while keeping a manageable number of speakers. For this purpose, we have developed an open-source simulator [18] using over 60 hours of noise data, twelve speech/audio codecs covering telephony, VoIP and audio applications and over a hundred impulse responses of simulations of answering machine, telephone, playback and loudspeaker devices. Train/test trial lists from the NIST SRE 2010 evaluation campaign are also provided in the same package to make performance comparison possible across systems. We provide experimental results for i-vector systems using clean and degraded speech PLDA backends. The latter are trained using a development database with a large number of degraded speech recordings, following the pooled PLDA approach in [19], proposed for noise compensation. Further PLDA training approaches were explored in [20], with tied and pooled PLDA outperforming per-condition PLDA training in noisy scenarios.

While other simulation initiatives have proposed specific

This work was supported by the European Union under the FP7 Integrated Project SIIP (Speaker Identification Integrated Project). The authors gratefully thank the EU for their financial support and all project partners for a fruitful collaboration.

data sets in the past, e.g. PRISM [21] and QUT [22] data sets, these tend to use a relatively small number of degradation processes, in the tens of noise recordings and few impulse responses. This work goes beyond by publicly releasing a software package addressing large-scale simulation capable of generating thousands of acoustic variations of a single clean speech recording by varying the generation parameters. We believe these are valuable resources for system development and evaluation of speech processing algorithms, speaker recognition in particular.

The paper is organized as follows. Section II describes the proposed acoustic environment simulator. In Section III, the speech database used for system evaluation is presented. Section IV gives details about the speaker recognition systems being evaluated. Experiment details and results are given in Section V and Section VI gives conclusions about this study.

II. Acoustic Simulator

In this work, a simulator is used as an affordable way of generating a number of degraded variants of a given clean speech signal. Three types of non-linear degradation processes are considered, namely additive noise, telephone and audio codecs, and room and device impulse responses. Figure 1 shows a simplified block diagram of the simulator. An input speech signal, recorded in a quiet environment using high-quality equipment, is added a noise recording simulating environmental noise prior to applying reverberation and/or a speech codec setup. This simulation approach to generating data, for either evaluation or training, has been used in the past, e.g. in [21], [23], [24], but at small scale. In this work, we aim at obtaining a large number of degradation processes, made by combining the above blocks, to thoroughly sample the acoustic variability for the same speech sounds of a recording. This is expected to prevent machine learning paradigms to overfit while improving their generalization ability. Clean, noise and impulse response audio is sampled and processed at 16kHz sample rate.

The noise files are randomly chosen from a 60-hour database of 1 to 8 minute long recordings collected from the Freesound online audio archiving site [25]. Around 1400 files[1] were downloaded using the Freesound API and were tagged manually into noise conditions based on the context in which they were apparently recorded. We consider a total of 7 categories, namely outdoors, public, private and transportation settings as well as babble, music and impulsive ambience noises. Table I gives details about the noise database used in this work. These were recorded by users of the Freesound website using all sorts of unknown equipment probably ranging from smartphones to professional equipment. The noise files are added to the clean speech files at levels -15dB lower than the clean signal level.

The package also provides simulation of linear distortion produced by devices and rooms. Over 100 impulse responses, 74 for devices and 54 for rooms, were collected from public sources on the internet under licensing allowing for its redistribution. Impulse responses of smartphones, answering machines, loudspeakers and microphones were included for device simulation whereas small, medium and large room responses were included for simulation of enclosed spaces. For the impulse responses collected online, mainly by private and individual initiative, the exact impulse response estimation methods used are not known. Still, given the amount of time and effort required for the collection of such data, we believe these are priceless resources for system development. A set of 12 speech and audio codecs is also included in the acoustic simulator. These comprise the ITU G.712, P341, IRS and mIRS telephony band-pass filters for narrow-band telephony codecs as well as the following lossy codec groups:

- **Landline** includes mu/A-law companding, following the **ITU G.711** standard for 64 kbps rates. It also includes Adaptive-Differential PCM (ADPCM) coding following the **ITU G.726** standard, allowing for 16, 32, 48 and 64 kbps rates. The latter is used in international trunks of the telephone network and in DECT wireless phones.
- **Cellular** includes two major cellular telephony codecs in Europe[2], namely the Global System for Mobile Communications (GSM) and narrow-band and wide-band Advance Multi-rate (AMR-NB and AMR-WB) codecs. The full-rate specification of GSM (**GSM-FR**) allowing for 13 kbps rate uses linear prediction coding with regular pulse excitation. **AMR-NB** is a multi-rate speech codec using Algebraic Code-Excited Linear Prediction (ACELP) at 4.75-12.2 kbps. **AMR-WB**, following the ITU G.722.2 specification, is the wide-band variant of AMR, coding speech signals up to 7 kHz using bit-rates from 6.6 to 23.8 kbps.
- **Satellite/Radio** includes three codecs that are used in satellite and radio telecommunication systems. The **ITU G.728** is a standard using Low-Delay CELP (LD-CELP) and vector-quantized excitation at 16 kbps. The Continuously Variable Slope Delta (**CVSD**) modulation is a 1-bit sample vocoder using an adaptive quantization step. Bit-rates range from 12 kbps, common for radio and military phones, to 64 kbps, used for wireless headset to mobile phone communication. CVSD has been formerly used for satellite communications as well. **Codec2** is a low bit-rate codec that is patent-free and open source. Codec2 is able to encode speech at 0.7-3.2 kbps via sinusoidal modeling of speech. It is mainly used in radio communications.
- **Voice over IP** includes the ITU G.729 and ITU G.728 standards besides SILK and SILK-WB, former Skype now open-source codecs. **ITU G.729a** is a narrow-band low-complexity codec based on the Code-Excited variant of ACELP (CS-ACELP), operating at 8 kbps. The **ITU G.722** is a wide-band audio codec based on sub-band ADPCM allowing 48, 56 and 64 kbps rates. It is used for

TABLE I
Collected noise database types, along with the number of files, hours, and average length per file.

Noise type	Number of files	Amount (hours)	Length/file (min)
Outdoors	538	23.7	2.5
Public	254	11.7	2.7
Private	135	5.3	2.3
Transportation	191	9.3	2.9
Babble	166	7.35	2.6
Music	103	4.57	2.7
Impulsive	46	1.97	2.6
Total	1433	63.8	2.6

[1]Freesound recordings are made available under Creative Commons License Attributions allowing their use for research purposes.

[2]For copyright reasons, EVRC codecs used in the U.S. are not allowed to be used in the context of this work.

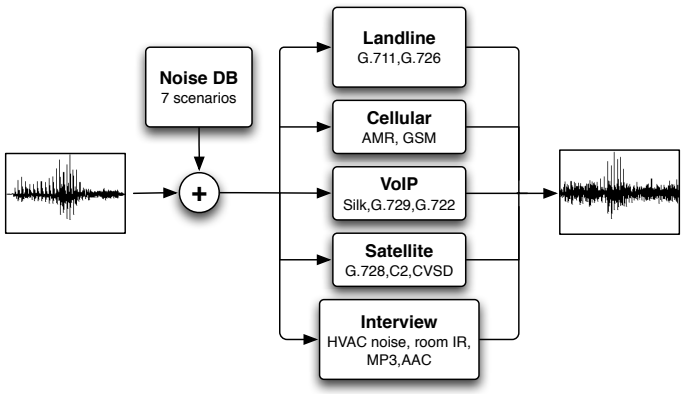

Fig. 1. Simplified block diagram of the environment simulator. A first module adds real noise recordings from a large noise database. The second module further degrades speech using telephone and interview conditions.

voice over IP and radio broadcasters. The **SILK** codec is based on linear prediction and it can adapt its bandwidth from 8 to 24 kHz and quality from 6 to 40 kbps.

- **Audio** includes the Fraunhofer MPEG Layer III (**MP3**) and Advanced Audio Codec (**AAC**) codecs, both being developed for general audio and music. AAC has a more efficient and simpler filter-bank based on the Modified Discrete Cosine Transform (MDCT) and better coding of stationary and transient signals compared to MP3. These codecs are used in many consumer smartphones and audio recorders.

The acoustic simulator is released as a package available online [18]. It includes scripts to download and process noise and impulse response data together with the speech audio codecs. Although many degradation processes can be generated, five conditions were fixed, targeting typical speaker recognition scenarios: **landline**, **cellular**, **satellite**, **voip** and **interview** on clean and 15dB noisy conditions. For Interviews, HVAC noise, small room reverberation and audio codecs were used.

For each of these conditions, codec parameters such as bit-rate, dtx or packet loss as well as noise recordings and impulse responses are randomly sampled. The simulator includes a reliable random number generator to ensure the reproducibility of the degradation processes across different sites.

III. Evaluation

While the motivation in using an environment simulator is to have the ability to generate a large number of degraded speech files from a single one, the amount of generated data becomes rapidly intractable if both speaker and degradation variabilities are densely sampled. In this work, we emphasize on sampling the degradation processes while keeping a reasonable number of speech segments and speakers. This suggests targeting an operating point in the DET curve [26] that is likely to be populated with scores, e.g. the Equal Error Rate (EER) as opposed to the minimum Decision Cost Function (DCF) used in the NIST SRE Evaluations. As an alternative quality measure, Mean Opinion Score - Listening Quality Objective (MOS-LQO) is assessed using the PESQ algorithm [27]. MOS-LQO scores for speech segments with at least 0.5 s leading non-speech and minimum duration of 3.2 s were averaged.

The training and test data involve 361 utterances and speakers and 644 utterances and 361 speakers, respectively, taken from the microphone data of the NIST SRE 2010 data. The simulator was run on each of the train and test data sets using five codec conditions versus a clean and a noisy condition at 15dB SNR, with non-overlapping noise files being used for the development, train and test data sets. Codecs and room impulse responses were available for training the noisy PLDA backend. This experimental setup results in a total of 20 degraded conditions including clean and noisy experiments.

IV. System description

For the speaker verification experiments, we use two Kaldi-based [28] i-vector systems, featuring a standard i-vector extractor, using either clean or degraded PLDA backends.

A. I-vector Frontend

We use 20 Mel-Frequency Cepstral Coefficients (MFCC) computed every 10ms along with delta and double delta features prior to short-term Gaussianization using a 3 s window. The 2048-mixture UBM and the total variability matrix of the i-vector extractor are trained using Fisher English Part I and II data. The dimension of the i-vectors is 400.

B. Clean PLDA Backend

We use LDA to improve the discrimination of speaker i-vectors followed by PLDA scoring. Both LDA and PLDA parameters are trained using the NIST data sets SRE 2004, 2005, 2006, 2008 and 2008 extended and Swithboard Part II and Part III. Details about this system are given in [29].

C. Degraded PLDA Backend

The PLDA hyperparameters of this system are trained using degraded i-vectors from a development data set taken from the NIST SRE 2010 microphone data, from the remaining speakers not included in the train and test sets. This is simulated data involving 13 sessions per speaker in average, for 95 speakers. Each session is degraded using 10 random degradation processes, one per condition as defined in Section II. This makes up a total of around 12,000 i-vectors for PLDA training. Note that, for each session of each speaker, 10 degraded variations from the same exact recording are used. This is expected to help PLDA in separating speaker and session effects, as only variation related to channel and noise is considered. Almost all the possible parameter variations were randomly sampled for the generation of this data set, except SNR which was either clean or 15dB. Sampled parameters include band-pass telephone filters (G.712, P341, IRS, mIRS), codec choice, bit-rates (codec dependant), packet loss probabilities (from 0 to 10%), signal input level (from -26dB to -35dB), noise files and room impulse responses. Uniform distributions were used to sample these parameters.

V. Experiments and results

We performed two sets of experiments involving 10 experiments each. The first set of experiments assessed the performance of an i-vector system using a clean PLDA backend. The second set assess the performance of an i-vector system using a PLDA backend trained on a development data set involving degraded speech data generated using the simulator. These experiments were compared to a baseline system using a clean backend evaluated on clean speech data, downsampled

TABLE II
MOS-LQO scores, ranging from 1 (bad) to 5 (excellent), for each condition in the evaluation data. Rows denote codec conditions and columns denote noise conditions.

	MOS-LQO (1-5)	
Codec Condition	**Clean**	**15dB**
Landline	4.0	3.0
Cellular	3.8	3.0
Satellite	3.1	2.5
VoIP	3.7	2.8
Interview	1.7	1.6

TABLE III
EER (%) for an i-vector speaker recognition system using clean (first and second columns) and degraded (third and foruth columnns) PLDA backends. Rows denote codec conditions and columns denote noise conditions.

	EER (%)			
	Clean PLDA		Degraded PLDA	
Codec Condition	**Clean**	**15dB**	**Clean**	**15dB**
No codec	1.8	-	1.9	-
Landline	4.5	6.0	2.7	4.4
Cellular	6.8	9.6	3.9	5.7
Satellite	19.2	21.1	10.7	12.0
VoIP	3.3	6.2	2.5	4.8
Interview	21.8	25.8	9.5	17.4

from 16 kHz to 8 kHz to match the sample rate of the degraded data. The EER for this system is 1.8%.

Table II gives speech quality assessment scores obtained using the PESQ algorithm. Systems are ranked consistently across clean and noisy conditions, with codec conditions being ranked as landline, cellular, voip, satellite and interview. Satellite shows a 16% decrease of MOS score relative to the worst landline, cellular or voip scores (from 3.7 to 3.1 for clean, 3.0 to 2.5). This condition uses very low-complexity codecs such as CVSD and low bit-rate codecs such as Codec2, using sinusoidal analysis-synthesis, both resulting in a decrease in speech quality that is captured by PESQ. The scores for the interview condition are the worst of all five conditions with scores as low as 1.6. Such low scores are probably due to the time smearing introduced by the room impulse responses, only present in the interview condition. However, it must be noted that reverberant speech is out of the scope of PESQ [30].

The two first columns of Table III shows EERs for speaker recognition systems using a PLDA backend trained using clean data. Error rates mostly follow the trends shown in the speech quality experiments of Table II, with landline, cellular and voip conditions obtaining considerably lower error rates than satellite and interview. In any case, from 1.8% EER for the baseline system, the best performing system achieves 3.3% EER, a 83% relative increase. This is attributed to the mismatch resulting from applying the simulated degratation processes. Although the PLDA backend has been trained using telephone speech from the NIST SRE data sets, large EER increases were found for telephone speech codecs such as landline or cellular (4.5% and 6.8%, respectively, compared to 1.8%). This indicates strong mismatch as well for simulated and real telephone channels. For satellite and interview conditions, error rates rise dramatically, reaching over 20% EER. These results suggest a severe mismatch between development and train/test data, and highlight the sensitivity of speaker recognition systems to it.

For noisy test conditions (second column of Table III), EER become even larger, with the best performing system, landline, achieving 6.0% EER from the 1.8% obtained by the baseline system, more than three times lower error rate. EER increase is large for landline, cellular and voip, the best performing conditions, while satellite and interview conditions achieve less dramatic performance losses of 10% and 18% in relative terms, respectively (19.2 to 21.1 and 21.8 to 25.8).

A set of speaker recognition experiments were conducted to evaluate the potential of using a large development data set with multiple variants of degraded speech recordings. All the session recordings from the remaining 130 speakers from the NIST SRE 2010 data, around 13 sessions per speaker in average, were degraded using each of the 10 conditions used in the simulator, resulting in around 13,000 utterances. The noise files come from a disjoint data set while the same codecs, in its different variants, and impulse responses were used for train, test and development sets. A new PLDA model was trained using these data.

The third and fourth columns of Table III give EER for the system using the degraded PLDA backend. A minor EER increase was found for the baseline condition, from 1.8% to 1.9%. For all degraded conditions, large improvements in EER, six out of ten conditions with over 40% relative EER decrease, were achieved. On one side, this shows that PLDA alone, trained with a large number of sessions per speaker - 12 sessions times 10 degraded versions, is able to considerably improve system performance when test channels have been seen during PLDA training. Therefore, this error reduction can not be solely attributed to data augmentation, but to matching development and train/test channels as well. For noisy experiments, train/test noise was not observed at development time. On the other side, the simulated variants of each session are generated from the same clean recording, thus feeding PLDA training with pure channel/noise variation rather than feeding a conglomerate of aggregate factors such as content or channel from real data. This might facilitate the discrimination of speakers in the i-vector space using PLDA.

VI. Conclusion

An open-source simulator capable of generating a large number of degradation variants from clean speech data was introduced. This simulator uses a large noise recording database covering several ambiences, over a hundred impulse responses for rooms and audio playback devices and telephone speech, voice IP, satellite, radio communication and audio codecs. The simulator has been used to generate ten conditions targeting scenarios of interest for speaker recognition. I-vector systems using clean and degraded data for PLDA backend were evaluated. A minimum relative increase in EER of 80% was found when evaluating an i-vector system with a clean PLDA backend on degraded data, whereas error rates were found to be up to 10 times larger for interview data compared to the baseline condition. Retraining the PLDA backend using simulated development data with around 120 sessions per speaker reduced error rates by 40% relative for six out of the ten conditions. These results serve only as a demonstration of the data augmentation package, and more careful experimentation is required to assess the full potential of data augmentation for speaker recognition, especially under channel mismatched conditions between development, train and test data.

REFERENCES

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," in *IEEE Trans. on Speech and Audio Processing*, 2001.

[2] N. Fakotakis T. Ganchev, I. Potamitis and G. Kokkinaki, "Text-independent speaker verification for real fast-varying noisy environments," *Speech Communication*, p. 281292, 2004.

[3] Q. Wu and L. Zhang, "Auditory sparse representation for robust speaker recognition based on tensor structure," in *EURASIP Journal of Audio Speech and Music Processing*, 2008.

[4] J. Sandberg T. Kinnunen, R. Saeidi and M. Hansson-Sandsten, "What else is new than the hamming window? robust mfccs for speaker recognition via multitapering," in *Proc. INTERSPEECH*, 2010.

[5] S. Thomas S. Ganapathy and H. Hermansky, "Feature extraction using 2-d autoregressive models for speaker recognition," in *Proc. of the IEEE Speaker Odyssey Workshop*, 2012.

[6] L. Burget Y. Lei and N. Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Proc. IEEE ICASSP*, 2013.

[7] N. Scheffer L. Ferrer M. McLaren, Y. Lei, "Application of convolutional neural networks to speaker recognition in noisy conditions," in *Proc. INTERSPEECH*, 2014.

[8] "Nist speaker recognition evaluation results 2005, 2006, 2008 and 2010," http://www.nist.gov/itl/iad/mig/sre.cfm, Accessed November 2015.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2007.

[10] S. Ioffe, "Probabilistic linear discriminant analysis," in *ECCV*, 2006, pp. 531–542.

[11] Daniel Garcia-Romero and Alan McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Proc. IEEE ICASSP*, 2014, pp. 4047–4051.

[12] "Nist speaker recognition evaluation 2012," http://www.nist.gov/itl/iad/mig/sre12.cfm, Accessed March 25th, 2015.

[13] G. Liu N. Shokouhi H. Boril T. Hasan, S. O. Sadjadi and J. H. Hansen, "Crss systems for 2012 nist speaker recognition evaluation," in *Proc. IEEE ICASSP*, 2013.

[14] N. Scheffer Y. Lei M. Graciarena L. Ferrer, M. McLaren and V. Mitra, "A noise-robust system for nist 2012 speaker recognition evaluation," in *Proc. INTERSPEECH*, 2013.

[15] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013.

[16] Xiaodong Cui, Vikas Goel, and Brian Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Proc. IEEE ICASSP*, 2014, pp. 5582–5586.

[17] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015.

[18] "Idiap acoustic simulator," http://github.com/idiap/acoustic-simulator, Accessed November 2015.

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

[20] D. Garcia-Romero and C. Y. Espy-Wilson, "Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. IEEE ICASSP*, 2012.

[21] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, et al., "Promoting robustness for speaker modeling in the community: the prism evaluation set," in *Proceedings of NIST 2011 Workshop*, 2011.

[22] David B Dean, Ahilan Kanagasundaram, Houman Ghaemmaghami, Md Hafizur Rahman, and Sridha Sridharan, "The qut-noise-sre protocol for the evaluation of noisy speaker recognition," 2015.

[23] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition," in *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, France*, 2000.

[24] E. Khoury, L. E. Shaffey, and S. Marcel, "The Idiap Speaker Recognition Evaluation System at NIST SRE 2012," in *NIST Speaker Recognition Conference, Orlando, USA*, 2012.

[25] "Freesound.org," http://freesound.org, Accessed March 25th, 2015.

[26] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proc. EUROSPEECH*, 1997, vol. 4, pp. 1895–1898.

[27] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, 2001, vol. 2, pp. 749–752.

[28] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," 2011.

[29] Petr Motlicek et al., "Employment of subspace gaussian mixture models in speaker recognition," in *To Appear In Proc. of ICASSP 2015*, 2015.

[30] P.862.3, *Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2*, 2007.