AUDIO ENHANCING WITH DNN AUTOENCODER FOR SPEAKER RECOGNITION

Oldřich Plchot[1], *Lukáš Burget*[1], *Hagai Aronowitz*[2], *Pavel Matějka*[1]

[1]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic
[2]IBM Research - Haifa

iplchot,burget,matejkap@fit.vutbr.cz, hagaia@il.ibm.com

ABSTRACT

In this paper we present a design of a DNN-based autoencoder for speech enhancement and its use for speaker recognition systems for distant microphones and noisy data. We started with augmenting the Fisher database with artificially noised and reverberated data and trained the autoencoder to map noisy and reverberated speech to its clean version. We use the autoencoder as a preprocessing step in the later stage of modelling in state-of-the-art text-dependent and text-independent speaker recognition systems. We report relative improvements up to 50% for the text-dependent system and up to 48% for the text-independent one. With text-independent system, we present a more detailed analysis on various conditions of NIST SRE 2010 and PRISM suggesting that the proposed preprocessig is a promising and efficient way to build a robust speaker recognition system for distant microphone and noisy data.

Index Terms: speaker recognition, denoising, de-reverberation, neural networks, DNN

1. INTRODUCTION

The last years have seen a great growth in the market with various portable devices that are equipped with microphone to process a speech input in various environments and applications. Such devices include smartphones, tablets, gaming consoles, voice-controlled navigation devices and other voice-controlled systems. The presence of various environmental noises and reverberation in the input speech signal has a significant negative impact on the performance of most applications that deal with speech.

Various techniques for speech and signal processing have been introduced to cope with the distortions caused by noise and reverberation in distant microphone data. One way to tackle this problem at the very beginning is to use multiple microphones that allow for effective use of techniques such as active speaker tracking, active noise cancelling, beamforming and filtering [1]. While using smart microphone arrays is effective, their use is still limited to larger and non-portable devices. For single microphone systems, front-ends utilize signal pre-processing methods such as Wiener filtering, adaptive voice activity detection (VAD), gain control, etc. [2]. In the later stages, various designs of robust features [3] are used in combination with normalization techniques such as cepstral mean and variance normalization or short-time gaussianization [4].

All above referenced techniques work on a signal or acoustic-feature level and can be combined with speech enhancement techniques such as denoising or dereverberation. In the last years, we have seen several application of speech enhancement based on neural networks (NN): for example, in [5], a classical approach of removing the room impulse response is proposed, but the filter is estimated using a NN. NNs have also been used for speech separation [6] instead of popular computational auditory scene analysis (CASA) techniques. NN-based autoencoder for speech enhancement was proposed in [7] with optimization in [8] and finally, reverberant speech recognition with signal enhancement by a deep autoencoder was tested in the Chime Challenge and presented in [9].

In this paper, we investigate the use of a DNN autoencoder as an audio preprocesing front-end for speaker recognition. The autoencoder is trained to learn a mapping from noisy and reverberated speech to clean speech. The frame-by-frame aligned examples for DNN training are artificially created by adding noise and reverberation to the Fisher speech corpus. We have developed and successfully applied this preprocessing scheme for automatic speech recognition (ASR) in our system for the IARPA Automatic Speech recognition In Reverberant Environments (ASpIRE) challenge[1].

Here, we demonstrate that the proposed method increases the performance of two state-of-the-art systems for text-dependent and text-independent speaker recognition. The text-dependent system is based on GMM-NAP framework [11], while text-independent system is based on i-vectors and Probabilistic Linear Discriminant Analysis (PLDA) [12, 13]. By design, both systems already include compensation for the unwanted variability caused by noise or reverberation and serve as good baselines for our experiments. As it was already shown that performing multi-condition training with added noisy and reverberated data helps significantly for both ASR [10, 14] and speaker recognition [15, 16], we also explore this scenario for text-independent system in section 4.

2. AUTOENCODER TRAINING AND DATASET DESIGN

Fisher English database parts 1 and 2 were used for training the autoencoder. It contains over 20,000 telephone conversational sides or approximately 1800 hours of audio.

2.1. Adding noise

Our training data were processed by artificially adding different types of noises from the following two categories: stationary noises

This work was supported by the DARPA RATS Program under Contract No. HR0011-15-C-0038. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The work was also supported by Czech Ministry of Interior project No. VI20152020025 "DRAPAK" and European Union's Horizon 2020 programme under grant agreement No. 645523 BISON. The aouthors also thank Sandro Cumani, Pietro Laface and the Politecnico di Torino for organinzing the TOSREW workshop, where part of this work was conducted.

[1] https://www.innocentive.com/ar/challenge/9933624 [10]

and transient noises. Stationary noises contain 285 samples (4 minutes long) taken from the Freesound library[2] and include recordings of the following categories: real fan, HVAC, street, city, shop, crowd, library, office and workshop. Their character is mainly stationary, with minor portions of transient noises and babbling. Additionally, we use 7 samples (4 minutes long) of artificially generated noises: various spectral modifications of white noise + 50 and 100 Hz hum. Transient noises contain 60 samples (4 minutes long) from Freesound and include recordings of the following categories: dishes, motor, workshop, doors, city, keyboard, library, office. The character is mainly transient, with some minor portion of stationary noises. Additionally we created 25 samples (4 minutes long) of babbling noises by merging speech from 100 random speakers from Fisher database using speech activity detector.

2.2. Reverberation

We generated artificial room impulse responses (IR) using "Room Impulse Response Generator" tool from E. Habets [17]. The tool can model the size of room (3 dimensions), reflectivity of each wall, type of microphone, position of source and microphone, orientation of microphone towards the audio source, and number of bounces (reflections) of the signal. Each room model consists of a pair of IR. One is used to reverberate (convolution with IR) the speech signal and the other is used to reverberate the noise signal. These signals are then mixed into a single recording. Just coordinates of audio sources (speech/noise) differ for each of the IRs in such pair. We randomly set all parameters of the room for each room model.

2.3. Composition of the training set

We used `fant` tool [18] to mix reverberated speech and reverberated noise at given SNR. Speech signal was compensated for the delay caused by the reverberation.

The autoencoder training dataset consists of 1800 hours of clean Fisher data augmented with another three copies of artificially corrupted Fisher data. IRs were generated for rooms where each dimension was limited to the range of $2 - 5$ meters. Noises were added at SNRs ranging from 0 dB to 27 dB. Two noises were always added into each recording: one random stationary noise and one random transient noise.

2.4. Audio enhancement by DNN autoencoder

The role of the autoencoder is to enhance (de-noise and de-reverberate) the speech signal. It is trained on the artificially created parallel clean-noisy Fisher corpora as described in the previous section. The inputs of the NN are 129 dimensional vectors of log magnitude spectrum stacked over 31 frames (e.g. 3999 dimensional vector). The desired outputs are 129 dimensional vectors (again log spectrum) corresponding to the clean version of the central input frame. A standard feed-forward architecture is used: 3999 inputs, 3 hidden layers with 1500 neurons, 129 outputs, tanh nonlinearities in the hidden layers. The NN is initialized in such a way that it (approximately) passes its input to the output and it is trained using conventional stochastic gradient descent to minimize the MSE objective.

We have experimented with different strategies of normalizing NN input and output. To achieve a good performance, utterance level mean and variance normalization is applied to both the NN input and the desired NN output. To synthesize the cleaned-up speech log spectrum, the NN output is de-normalized based on the global mean and variance of clean speech. The cleaned-up log magnitude spectrum is further converted to speech signal by using an overlap-add algorithm. The information about phase is taken from the original noisy spectrum.

3. SPEAKER RECOGNITION SYSTEMS

3.1. Text dependent system

The proposed deep audio enhancement method is evaluated on a common passphrase authentication task. The speaker recognition system is based on the GMM-NAP framework, as i-vector-based systems have shown to be inferior for the common passphrase task when a small-to-medium sized text matched development set is available [11].

In the GMM-NAP framework a UBM is MAP adapted to each session (enrollment, test and development). The resulting session-dependent GMM is transformed into a mean supervector. A linear projection named NAP (Nuisance Attribute Projection) is estimated from the development set and is used to compensate intra-speaker intersession variability in the evaluation data. Scores are computed as a dot product between supervectors. Finally, scores are normalized using ZT-norm. A detailed system description of the system can be found in [19].

Note that all trainable components (UBM, NAP and score normalization parameters) are trained solely on the common passphrase utterances from the development set. In this paper, we report results for two front-ends. The first is based on plain Mel frequency cepstral coefficients (MFCC) and the second is based on a variant of the noise robust PNCC features [20]. For the first system, we take 12 MFCCs together with zero-*th* coefficient and their delta. Both features are transformed by means of feature warping on a 3 s sliding window [4]. We used an energy-based voice activity detection to remove non-speech frames.

3.1.1. Datasets

The data were collected internally at IBM as part of a multi-modal data collection effort described in detail in [21]. Subjects were recorded by a smartphone or a tablet held at arm-length, which degrades the quality of the audio signal significantly.

The data were collected in two separate phases with disjoint sets of subjects. First, the development/tuning data (using iPad 2 and iPhone 4, and then the evaluation dataset (using iPad 2 and iPhone 5) were collected. The development data were collected in a relatively quiet room (denoted by clean). For the evaluation set, two sessions were recorded (per subject and per device) in clean condition, and a third session was recorded (per speaker and per device) in a noisy cafeteria (denoted by noisy). The distribution of signal-to-noise ratios (SNR) in the evaluation set is given in Figure 1.

3.2. Text independent system

Our systems are based on i-vectors [12, 13]. To train i-vector extractors, we always used 2048-component diagonal-covariance Universal Background Model (GMM-UBM) and we set the dimensionality of i-vectors to 600.

Before using the i-vectors, we apply LDA to reduce the dimensionality to 200. Such processed i-vectors are then transformed by global mean normalization and length-normalization [12, 22].

A speaker verification score is produced by comparing two i-vectors corresponding to the segments in the verification trial by

[2] `http://www.freesound.org`

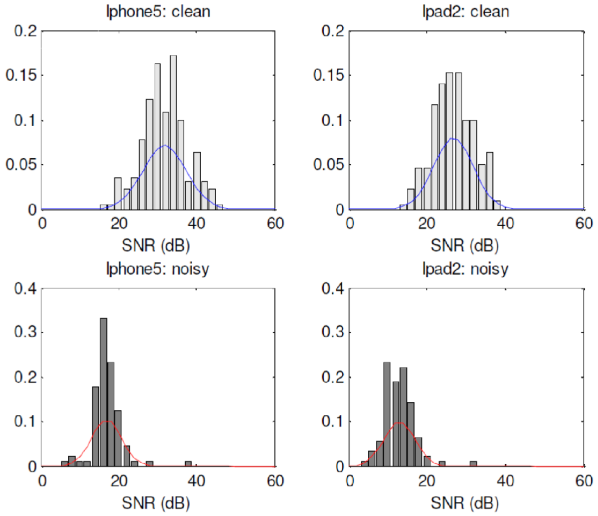

Fig. 1. SNR distribution for the clean and noisy environments on iPhone5 and iPad2.

means of PLDA [13] —a generative model that models i-vector distributions allowing for direct evaluation of the desired log-likelihood ratio verification score.

In our experiments, we used cepstral features, extracted using a 25 ms Hamming window. We used 24 Mel-filter banks and we limited the bandwidth to the 120–3800Hz range. 19 MFCCs together with zero-*th* coefficient were calculated every 10 ms. This 20-dimensional feature vector was subjected to short time mean- and variance-normalization using a 3 s sliding window. Delta and double delta coefficients were then calculated using a five-frame window giving a 60-dimensional feature vector.

After feature extraction, voice activity detection (VAD) was performed by the BUT Czech phoneme recognizer [23], dropping all frames that are labeled as silence or noise. The recognizer was trained on the Czech CTS data, but we have added noise with varying SNR to 30% of the database.

3.2.1. Datasets

We used the PRISM [24] training dataset definition without added noise or reverb to train UBM and i-vector transformation. Two variants of gender independent PLDA were trained: one only on the same training data without added noise and second included also artificially added cocktail noise and reverb. Artificially added noise and reverb segments totaled approximately eleven thousand segments or 10% of total number of segments for PLDA training. The PRISM set comprises the Fisher 1 and 2, Switchboard phase 2 and 3 and Switchboard cellphone phases 1 and 2, along with a set of Mixer speakers. This includes the 66 held out speakers from SRE10 (see Section III-B5 of [24]), 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.

We evaluated our systems on the *female* portions of the following conditions in NIST SRE 2010 [25] and PRISM [24]:

- **tel-tel**: SRE 2010 extended telephone condition involving normal vocal effort conversational telephone speech in enrollment and test (known as condition 5).
- **int-int**: SRE 2010 extended interview condition involving interview speech from different microphones in enrollment and test (known as condition 2).
- **int-mic**: SRE 2010 extended interview-microphone condition involving interview enrollment speech and normal vocal effort conversational telephone test speech recorded over a room microphone channel (known as condition 4).
- **prism,noi**: Clean and artificially created noisy waveforms from both interview and telephone conversations recorded over lavalier microphones. Noise was added with different SNR levels and recordings tested against each other.
- **prism,rev**: Clean and artificially created reverberated waveforms from both interview and telephone conversations recorded over lavalier microphones. Reverberation was added with different RTs and recordings tested against each other.
- **prism,chn**: English telephone speech with normal vocal effort recorded over different microphones from both SRE2008 and 2010 tested against each other.

The recognition performance is evaluated in terms of the equal error rate (EER) and the normalized minimum detection cost functions (DCF) as defined in both the NIST 2010 SRE task ($\mathrm{DCF}^{\mathrm{min}}_{\mathrm{new}}$) and the previous SRE 2005, 2006, 2008 evaluations ($\mathrm{DCF}^{\mathrm{min}}_{\mathrm{old}}$).

4. EXPERIMENTS AND DISCUSSION

4.1. Text dependent speaker recognition

Table 1 reports the performance of the proposed audio enhancing method for the MFCC-based frontend, and Table 2 reports the results for the PNCC-inspired fronted. For the experiments denoted as *enhanced*, only the evaluation data are processed with our autoencoder. The PNCC-inspired frontend proves to be superior for noisy and mixed data. The proposed method improves significantly the clean-clean condition. It is important to note that this condition still contains reverb as the data are recorded over a distant microphone. For noisy data, enhancing improves significantly the MFCC-based system and less significantly the PNCC-inspired system. In this case, the error is probably dominated by noise for which our enhancement is not that effective.

Table 1. Results using MFCC based frontend.

	Condition (dev-eval)	Baseline EER [%]	Enhanced EER [%]	Relative improv. [%]
iPad	clean-clean	1.01	0.60	41
	clean-noisy	8.89	7.50	16
	all-all	7.68	6.16	20
iPhone	clean-clean	1.50	1.45	3
	clean-noisy	4.24	3.45	19
	al–all	3.82	3.55	7

4.2. Text independent speaker recognition

Table 3 reports the performance of our speaker recognition systems under four different training scenarios. First, we study the effect of multi-condition training, when we add noise and reverberation into part of the PLDA training data, against the situation when only

Table 3. Results obtained with text-independent system in four scenarios. The first two blocks correspond to a system trained only with clean data without enhancing and to the same system, but trained with enhancing. The last two blocks correspond to a system trained in a multi-condition fashion (with noise and reverberated data in PLDA) and to the same system but with enhancing.

	PLDA trained on **clean** data						PLDA trained on **multi-condition** data					
	Original data			Enhanced data			Original data			Enhanced data		
Condition	$\mathrm{DCF}_{\mathrm{new}}^{\mathrm{min}}$	$\mathrm{DCF}_{\mathrm{old}}^{\mathrm{min}}$	EER	$\mathrm{DCF}_{\mathrm{new}}^{\mathrm{min}}$	$\mathrm{DCF}_{\mathrm{old}}^{\mathrm{min}}$	EER	$\mathrm{DCF}_{\mathrm{new}}^{\mathrm{min}}$	$\mathrm{DCF}_{\mathrm{old}}^{\mathrm{min}}$	EER	$\mathrm{DCF}_{\mathrm{new}}^{\mathrm{min}}$	$\mathrm{DCF}_{\mathrm{old}}^{\mathrm{min}}$	EER
tel-tel	0.372	0.108	2.07	0.370	0.109	2.18	0.382	0.109	2.14	0.392	0.111	2.24
prism,noi	0.415	0.126	2.94	0.364	0.099	2.28	0.316	0.078	1.88	0.316	0.076	1.73
prism,rev	0.408	0.108	2.07	0.224	0.059	1.37	0.303	0.079	1.62	0.206	0.053	1.28
int-int	0.310	0.077	1.74	0.251	0.064	1.68	0.270	0.071	1.69	0.230	0.062	1.58
int-mic	0.244	0.053	1.09	0.216	0.046	1.04	0.243	0.046	0.90	0.207	0.044	0.89
prism,chn	0.307	0.048	0.79	0.178	0.021	0.47	0.282	0.039	0.59	0.175	0.020	0.40

Table 2. Results using a PNCC inspired frontend.

	Condition (dev-eval)	**Baseline EER [%]**	**Enhanced EER [%]**	**Relative improv. [%]**
iPad	clean-clean	1.07	0.53	50
	clean-noisy	5.13	4.86	5
	all-all	4.66	4.22	9
iPhone	clean-clean	1.91	1.27	34
	clean-noisy	2.82	2.72	4
	all-all	2.98	2.69	10

relatively clean data are available for PLDA training. We can see that multi-condition training is effective for all conditions with the exception of *tel-tel* condition, which probably does not contain reverberation or high levels of noise. It is interesting to observe that multi-condition training helped also in relatively clean *int-int* and *int-mic* conditions.

After establishing the baselines with clean and multi-condition training, we study the effect of the proposed audio enhancement. In the experiments denoted as *enhanced test data* in Table 3, we pass all PLDA training data together with enrollment and test data through our DNN autoencoder and synthesize new audio. We keep the UBM and i-vector extractor trained on original data. With the exception of *tel-tel* condition, we can see improvements everywhere and in both clean and multi-condition scenarios. In both cases we can observe large relative improvements for *prism,rev* (artificially added reverberation) and *prism,chn* (real reverberation) conditions indicating, that our autoencoder is compensating mainly the effect of reverberation. The relative gains from audio enhancing are larger when training on clean data (26% average relative improvement) compared to multi-condition training (18% average relative improvement).

At this point, it is interesting to analyze the results obtained on the *prism,noi* condition containing artificially added additive noise. We can observe that most of the improvement is achieved with multi-condition training and further audio enhancing does not bring additional significant performance boost. This behavior is in line with other noise robust modelling techniques such as extracting noise-compensated i-vectors [15]. In this work, nice improvement was also achieved with PLDA trained only on clean data, but multi-condition training already solved most of the problems and the proposed technique was not very effective under this scenario.

To complete the analysis of presented results, we can also study the effect of enhancing and multi-condition training against enhancing and training on clean data (comparing columns five to seven with columns eleven to thirteen of Table 3). Again, we see improvements with the exception for *tel-tel* condition where we observe up to 6% relative degradation. The average improvement for all conditions except *tel-tel* is 10%. These results suggest that both multi-condition training and audio enhancing can be successfully used simultaneously, especially for data containing reverberation and additive noise.

Finally, after conducting these experiments and being encouraged by a good performance, we enhanced all of our data and retrained the whole system including UBM and i-vector extractor, which were trained only on unprocessed data so far. We observed a small relative improvement or degradation for all conditions. An average relative improvement across all presented conditions was 1.1%. This result suggests, that enhancing is important in the later stage of modelling when we are dealing with an unwanted variability in the data. Our DNN indeed removes some of the variability in the data or normalizes them into a common domain which helps our generative model. It is important to note that this processing also introduces some variability and therefore it is important to always enhance also the PLDA training data. When we trained the PLDA on the original data and enhanced only enrollment and test data, we were observing degradation w.r.t. testing on the original data.

5. CONCLUSIONS

We have presented our approach towards building a robust speaker recognition system. We concentrated on improving the performance on noisy and reverberant data by means of a DNN autoencoder, which is trained to remove both additive noise and reverberation from audio. We showed that our method significantly improves the performance of both state-of-the-art text-dependent and text-independent speaker recognition systems in the domain of distant microphone recordings. We analyzed and discussed the effect of the proposed method both on real-world data as well as on artificially created data. The artificially created data allowed us to measure the effect of enhancing separately for distortions caused by additive noise or reverberation. From these experiments, we conclude that the proposed audio enhancing method compensates well for the distortions caused by reverberation, while distortions caused by additive noise can be very well dealt with by means of multi-condition training.

6. REFERENCES

[1] Kenichi Kumatani, Takayuki Arakawa, Kazumasa Yamamoto, John McDonough, Bhiksha Raj, Rita Singh, and Ivan Tashev, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *APSIPA Annual Summit and Conference*, Hollywood, CA, USA, December 2012.

[2] ETSI, "Speech processing, transmission and quality aspects (STQ)," Tech. Rep. ETSI ES 202 050, European Telecommunications Standards Institute (ETSI), 2007.

[3] Oldřich Plchot, Spyros Matsoukas, Pavel Matějka, Najim Dehak, Jeff Ma, Sandro Cumani, Ondřej Glembek, Hynek Heřmanský, Nima Mesgarani, Mohammad Mehdi Soufifar, Samuel Thomas, Bing Zhang, and Xinhui Zhou, "Developing a speaker identification system for the darpa rats project," in *Proceedings of ICASSP 2013*, Vancouver, CA, 2013.

[4] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop*, Crete, Greece, 2006.

[5] B.D. Dufera and T. Shimamura, "Reverberated speech enhancement using neural networks," in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2009.*, Jan 2009, pp. 441–444.

[6] Tu Yanhui, Du Jun, Xu Yong, Dai Lirong, and Lee Chin-Hui, "Deep neural network based speech separation for robust speech recognition," in *Proceedings of ICSP2014*, 2014, pp. 532–536.

[7] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, Jan. 2014.

[8] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Global variance equalization for improving deep neural network based speech enhancement," in *Proc. IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*, 2014, pp. 71 – 75.

[9] Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders," in *Proc. Reverb Challenge Workshop*, Florence, Italy, 2014.

[10] Martin Karafiát, František Grézl, Lukáš Burget, Igor Szőke, and Jan Černocký, "Three ways to adapt a CTS recognizer to unseen reverberated speech in BUT system for the ASpIRE challenge," in *Proceedings of Interspeech 2015*, 2015.

[11] Hagai Aronowitz and Asaf Rendel, "Domain adaptation for text dependent speaker verification," in *Proc. Interspeech 2014*, Singapore, 2014, pp. 1337–1341.

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," vol. PP, no. 99, pp. 1 –1, 2010.

[13] P. Kenny, "Bayesian speaker verification with heavy–tailed priors," keynote presentation, Proc. of Odyssey 2010, June 2010.

[14] M. Karafiat, K. Vesely, I. Szoke, L. Burget, F. Grezl, M. Hannemann, and J. Cernocky, "BUT ASR system for BABEL surprise evaluation 2014," in *Proceedings of 2014 Spoken Language Technology Workshop*, South Lake Tahoe, Nevada, 2014, pp. 501–506.

[15] David González Martínez, Lukáš Burget, Themos Stafylakis, Yun Lei, Patrick Kenny, and Eduardo LLeida, "Unscented transform for ivector-based noisy speaker recognition," in *Proceedings of ICASSP 2014*, Florencie, IT, 2014.

[16] Yun Lei, Lukáš Burget, Luciana Ferrer, Martin Graciarena, and Nicolas Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proceedings of ICASSP*, Kyoto, JP, 2012.

[17] Emanuël A.P. Habets, "Room impulse response generator," https://www.audiolabs-erlangen.de/content/05-fau/professor/00-habets/05-software/01-rir-generator/rir_generator.pdf.

[18] H.G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proc. Interspeech 2005*, Lisabon, Portugal, 2005.

[19] Hagai Aronowitz, "Score stabilization for speaker recognition trained on a small development set," in *Proc. Interspeech 2015*, Dresden, Germany, 2015.

[20] Chanwoo Kim and Richard M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *Proc. ICASSP 2012*, Kyoto, Japan, 2012, pp. 4101–4104.

[21] Hagai Aronowitz, Min Li, Orith Toledo-Ronen, Sivan Harary, Amir B. Geva, Shay Ben-David, Asaf Rendel, Ron Hoory, Nalini K. Ratha, Sharath Pankanti, and David Nahamoo, "Multi-modal biometrics for mobile authentication," in *Proc. of IEEE International Joint Conference on Biometrics, Clearwater, IJCB 2014*, FL, USA, 2014.

[22] Daniel Garcia-Romero, "Analysis of i-vector length normalization in Gaussian-PLDA speaker recognition systems," 2011.

[23] Pavel Matějka, Lukáš Burget, Petr Schwarz, and Jan Černocký, "Brno university of technology system for NIST 2005 language recognition evaluation," in *Proceedings of Odyssey 2006*, San Juan, PR, 2006.

[24] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, and N. Scheffer, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proceedings of SRE11 analysis workshop*, Atlanta, Dec. 2011.

[25] "National institute of standards and technology," http://www.nist.gov/speech/tests/spk/index.htm.