The QUT-NOISE-SRE Protocol for the Evaluation of Noisy Speaker Recognition

David Dean, Ahilan Kanagasundaram, Houman Ghaemmaghami, Md Hafizur Rahman, Sridha Sridharan

Speech and Audio Laboratory, Queensland University of Technology, Brisbane, QLD, Australia

ddean@ieee.org, {a.kanagasundaram, houman.ghaemmaghami, m20.rahman, s.sridharan}@qut.edu.au

Abstract

The QUT-NOISE-SRE protocol is designed to mix the large QUT-NOISE database, consisting of over 10 hours of background noise, collected across 10 unique locations covering 5 common noise scenarios, with commonly used speaker recognition datasets such as Switchboard, Mixer and the speaker recognition evaluation (SRE) datasets provided by NIST. By allowing common, clean, speech corpora to be mixed with a wide variety of noise conditions, environmental reverberant responses, and signal-to-noise ratios, this protocol provides a solid basis for the development, evaluation and benchmarking of robust speaker recognition algorithms, and is freely available to download alongside the QUT-NOISE database. In this work, we use the QUT-NOISE-SRE protocol to evaluate a state-of-the-art PLDA i-vector speaker recognition system, demonstrating the importance of designing voice-activity-detection front-ends specifically for speaker recognition, rather than aiming for perfect coherence with the true speech/non-speech boundaries.

Index Terms: noisy speaker verification, speech databases, evaluation protocols

1. Introduction

While research in the field of speaker recognition has been ongoing for decades, the greatest cause of errors still remains the same: the issue of mismatch caused by intersession variability. The term intersession variability encompasses a number of phenomena contributing to this mismatch, including different recording devices, speech coding, transmisison channels, and variability introduced by the speaker (such as linguistic content). One particular area of intersession variability that has recently received renewed interest is the effect of noise and reverberation in the recording environment.

The recent development of the PRISM evaluation set [1], and the inclusion of background noise in the evaluation of speaker recognition systems in the 2012 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) [2] provoked a significant increase in the investigation of modern i-vector speaker recognition techniques in the presence of (often simulated) environmental noise [3, 4, 5, 6, 7].

However, one of the main unsolved issues in evaluating speaker recognition algorithms is the lack of a suitable large corpus of noisy speech available covering many speakers in a variety of noisy environments, and with a wide range of noise levels. In order to begin to approach the volume required to properly evaluate speaker recognition systems, most approaches have mixed existing clean speech databases with relatively short background noise data collected separately at the required noise level [3, 6, 7]. However, while the large speech corpora available to researchers through this approach allow a wide variety of speakers to be evaluated for VAD, the limited conditions, and short recordings–typically less than 5 minutes, of existing popular noise datasets such as NOISEX-92 [8], AURORA-2 [9] or freesound.org [1], have limited the ability to adequately test speaker recognition algorithms in a wide range of background noise conditions, especially when the length of the speech recordings exceed the noise.

Having noticed an earlier, similar, shortcoming in the voice-activity-detection literature, We previously [10] collected the comparatively large QUT-NOISE corpus, consisting of 20 recordings of at least 30 minutes of background noise and reverberant responses in a wide variety of locations covering common noise scenarios. This data was then combined with the clean-speech TIMIT [11] database, to create the QUT-NOISE-TIMIT database of over 600 hours of noisy recordings, and was demonstrated as a useful resource for the evaluation of voice-activity detection in noisy conditions.

In this paper, we will take the QUT-NOISE data originally collected for the construction of the QUT-NOISE-TIMIT database, and design the freely available QUT-NOISE-SRE protocol[1] that will allow for the comprehensive evaluation and benchmarking of modern speaker recognition approaches across a wide variety of background noise scenarios, noise levels and reverberation conditions. The paper concludes with a short evaluation of the front-end effect of voice activity detection on state-of-the-art speaker recognition systems.

2. The QUT-NOISE background noise corpus

This section will briefly reintroduce the QUT-NOISE background noise corpus from [10], which will be used to provide the background noise for the construction of the mixed-speech QUT-NOISE-SRE protocol outlined later in this paper.

2.1. Scenarios

In order to provide a simulation of noisy speech in a wide variety of typical background noise conditions, the corpus consists of 20 noise sessions of at least 30 minutes duration each. Two separate noise recordings, separated by at least one day in all but the CAR scenario, were conducted in 10 separate locations over 5 commonly encountered background noise scenarios.

CAFE: The two locations of the CAFE scenario were a typical outdoor cafe environment (CAFE-CAFE) and

This research was funded by the Australian Research Council (ARC) Linkage Grant No: LP130100110.

[1] Visit https://qut.edu.au/research/saivt to download the database and protocol information.

a typical indoor shopping centre food-court (CAFE-FOODCOURTB). These recordings are typified by medium to high levels of background speech babble, and kitchen noises from the cafe environment.

HOME: The two locations for the HOME scenario were recorded in a kitchen (HOME-KITCHEN) and living-room (HOME-LIVINGB) during typical home activities. The kitchen recordings consist of sections of relative silence interrupted occasionally by typical kitchen noises. The living room recordings consist of children singing, talking and playing alongside (public domain) television or music noise.

STREET: The two locations for the STREET scenario were at the roadside near typical inner-city (STREET-CITY) and outer-city (STREET-KG) traffic-light controlled intersections. Both recordings largely consist of road traffic noise, with the inner-city recordings also having significant pedestrian traffic as well as bird noise from a nearby park, while the outer-city recordings mostly consisting of cycles of traffic noise as the traffic lights changed.

CAR: As only one car was available for the CAR scenario, in lieu of two separate locations, the scenario was divided into driving with the windows down (CAR-WINDOWNB) or with the windows up (CAR-WINUPB). Because the car used was only available for a short time, all recordings were conducted on a single day. For both 'locations' the first session was recorded as highway driving, and the second was recorded based upon driving in city and suburban areas. All recordings are characterised by road (and wind for CAR-WINDOWNB) noise and typical car-interior noises (such as indicator, key or luggage-movement noise) but with no radio or speech noise.

REVERB: The two locations for the REVERB scenario were an enclosed indoor pool (REVERB-POOL) and an partially enclosed carpark (REVERB-CARPARK). Both locations were chosen as environments that were expected to produce a large reverberant response. In addition to the large levels of reverberation, the pool environment is characterised by splashing and running water noises, while the carpark environment is characterised by nearby road noise and occasional carpark vehicular noise.

2.2. Recording setup

As was outlined in the original publication [10], each of the 20 noise sessions were recorded with a prosumer-quality Zoom H2 set to record raw stereo WAV output with a sampling rate of 48 kHz, and 16 bits per sample.

In order to calculate the room response in the reverberant CAR and REVERB scenarios, 10-second-long frequency sweeps were played with a high-quality KRK RP5 studio monitor positioned several metres away from the microphone. Each reverberant session contained 12 frequency sweeps, with 6 before the main 30+ minute recording session and 6 after. Based on the work of Farina [12], the recorded sweeps could be averaged and deconvolved with a clean sweep to estimate the environment's reverberant response, for later use in simulating the reverberation of the clean speech in the target environment.

Each of the noise sessions collected was manually labeled with the boundaries of the main 30+ minute recording session, as well as the locations of each individual frequency sweep in the reverberant sessions. In addition, the locations of any bad portions of data (such as microphone failure) were labeled to allow them to be avoided.

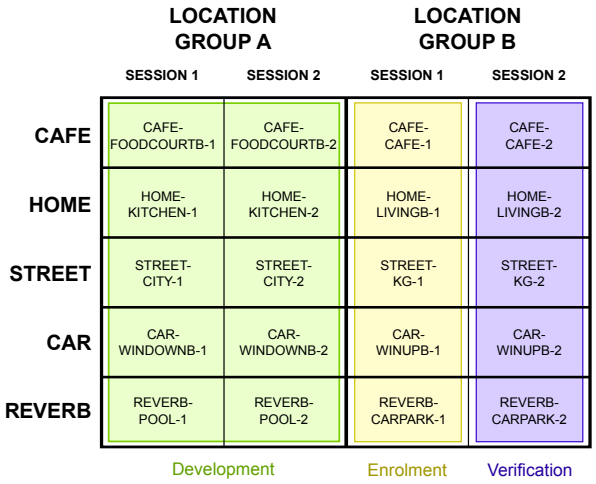

Figure 1: An overview of the noise sequences available in the QUT-NOISE database [10], with an example of a possible partitioning for development, enrolment and verification of a speaker recognition system.

3. The QUT-NOISE-SRE mixed speech evalutation protocol

This section will outline the QUT-NOISE-SRE protocol for the construction of a mixed speech corpus by mixing background noise sessions chosen from the QUT-NOISE corpus outlined previously with clean speech chosen from typical speaker recognition database, such as Switchboard [13], Mixer [14] and/or the various recent NIST SRE corpora [15, 16, 2].

3.1. QUT-NOISE-SRE partitioning

An overview of the speech sequences available in the QUT-NOISE database is shown in Figure 1, alongside an example configuration dividing the noise recordings into separate partitions for development, enrollment and verification.

While not by any means the only possible partitioning of the QUT-NOISE database, this particular partitioning allows for the development of background models, including UBM, total-variability and PLDA training, in noise conditions that are similar to, but not in the same location, to the noise used during the evaluation (enrollment and verification) of those trained models. Additionally the separation of the second location into two temporally separated sessions allows for the enrollment and verification to be performed in the same environment, but not against the exact same recordings. This partitioning approach could also easily be permutated to allow for at least four folds with the enrollment and verification noises swapped, and with the development and evaluation locations swapped.

Within the partitioning scheme outlined here, further permutations are also possible with the ability to study a variety of signal-to-noise ratios (SNRs), the effect of cross-scenario enrollment and verification, and whether development is performed across all scenarios (and/or noise levels), or done with knowledge of the target scenario and/or noise level.

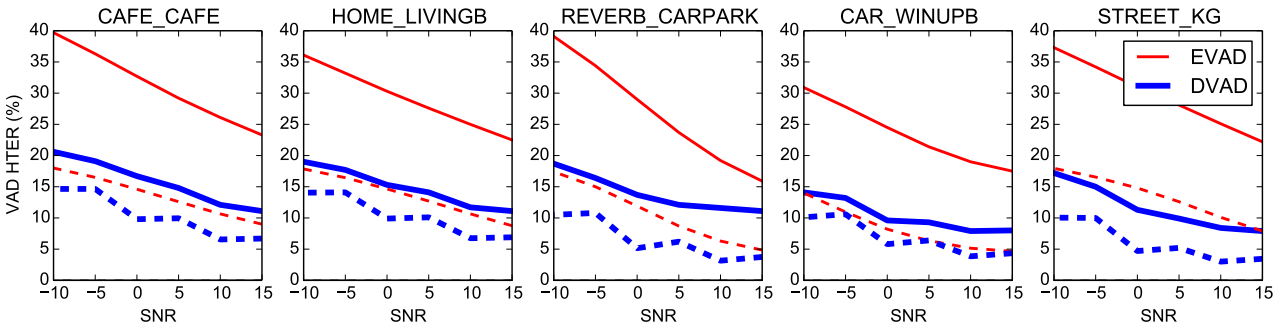

Figure 2: HTER performance of the energy-based thresholding (EVAD) and dissimilarity clustering (DVAD) VAD approaches on the different noise scenarios of the QUT-NOISE-SRE-corrupted NIST 2008 *short2-short3* dataset. The contribution of the FAR (or $\frac{\text{FAR}}{2}$) to the HTER is shown as a matching dashed line for both techniques. Oracle VAD is not shown, as it serves as the reference labeling and is always zero.

3.2. Construction of noisy sequences

3.2.1. Background noise and reverberation

Given a particular chosen noise session from the QUT-NOISE corpus, for each specified SNR, the QUT-NOISE-SRE protocol randomly selects a specified number of clean speech sequences (up to all sequences) from a chosen speech corpora. For each clean sequence selected, a random segment of noise of the same length as the complete clean-speech sequence (including all silence) is chosen from the labeled main portion of the recorded noise session, excluding the first five minutes[2], and restricted in such a way as to avoid any portion of the recorded noise session labeled as bad data.

Once the location of the background audio scene was chosen, the left-hand channel was taken and low-pass filtered and down-sampled from the original 48 kHz to match the sample rate of the clean speech files.

The reverberant response of the room, if the chosen noise session was in the CAR or REVERB scenarios, was also recorded and used through convolution with the chosen clean speech sequences to simulate the reverberant environmental response.

3.2.2. Combining speech with background noise

As many speech files provided for training and evaluation of speaker recognition algorithms have significant periods of silence in them, it was important that the desired SNR was produced only for sections where the speaker was speaking, otherwise the effect of silence would artificially raise the true short-term active-speech SNR compared to that desired. Accordingly, before the clean speech could be combined with the background noise at the desired SNR, the active portions of the clean speech had to first be identified through the use of simple VAD based on a ITU-T P.56 active level calculation, as implemented in the VOICEBOX MATLAB toolkit [17]. The average level of the active-speech portions only was then used to scale the entire clean speech sequence to have an ITU-T standard (P830) reference signal level of -26 dBov, and then the average background noise energy was scaled in relation to this reference speech level according to the desired SNR. This approach ensured that all created noisy speech sequences had a well defined speech signal level.

Once the background noise levels and speech sequences had been scaled appropriately, the final noisy speech sequences were obtained by a sample-by-sample summing of the speech sequence and background noise. While this approach has resulted in some clipping (but only to the noise, not the speech) at low SNR levels due to high noise energy, it was deemed more important to maintain a consistent reference energy level in similar SNR sequences.

[2]The first five minutes were excluded to allow for the possibility of training specific noise models on noise not used in the final noisy speech files in future research.

4. Speaker recognition experiments

In order to demonstrate the utility of the QUT-NOISE-SRE protocol outlined in the previous section, a short examination of the effect of VAD on speaker recognition performance will be conducted across a range of SNRs in all the noise scenarios available.

4.1. QUT-NOISE-SRE configuration

For the speaker recognition experiments conducted in this section we used a similar configuration to that outlined in Figure 1, but to simplify the experiments in this paper, development remained on clean speech data taken from Switchboard II and the NIST 2004, 2005 and 2006 SRE corpora. Only the evaluation experiments were corrupted using the Location Group B of the QUT-NOISE-SRE protocol, using the NIST 2008 *short2-short3* evaluation data as the basis for the noisy speech sequences, with the *short2* enrolment data corrupted with noise from Location Group B, Session 1 and the short3 verification data corrupted with noise from Location Group B, Session 2. Enrolment and verification were always matched by noise scenario and SNR, with six SNR levels chosen: -10, -5, 0, 5, 10 and 15 dB. Only the telephone-speech portions of the NIST 2008 evaluation sets were used for these experiments, with each channel treated separately when choosing the particular portion of noise to be added.

4.2. Voice activity detection

Before speaker recognition can be conducted, a voice activity detection (VAD) process must be run to first determine which portions of the speech sequence are actually speech. For the speaker recognitions evaluations we will perform in this paper, we investigated three approaches to VAD:

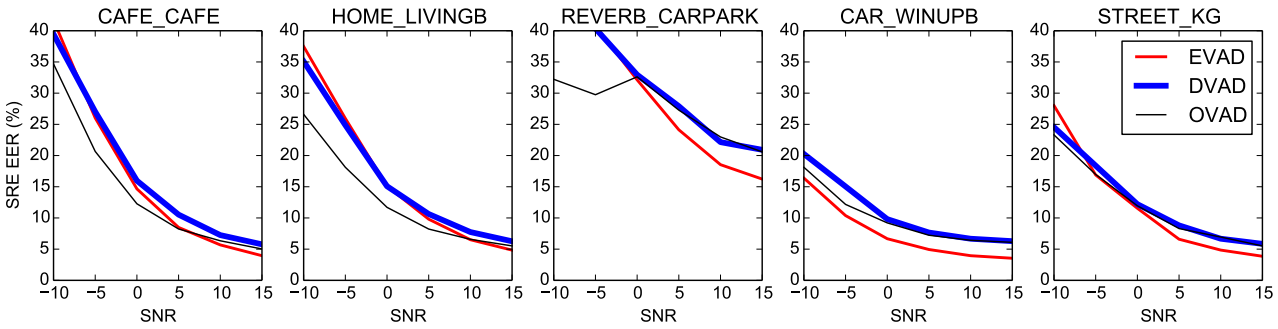

Figure 3: EER speaker-recognition performance after EVAD, DVAD and OVAD on the different noise scenarios of the QUT-NOISE-SRE-corrupted NIST 2008 *short2-short3* dataset. Enrolment and verification utterances were corrupted from different noise sessions, but with matched noise-scenarios, and SNRs.

Energy-based thresholding (EVAD), where speech is determined to be active if the frame-energy is above an exponentially-weighted moving average of all preceeding frames. This approach is the one we typically use for clean speech.

Dissimilarity clustering (DVAD) [18], where speech and non-speech GMMs are trained on Location Group A (but using the QUT-NOISE-TIMIT database [10] rather than QUT-NOISE-SRE corrupted development data), and used to calculate a dissimilarity measure for use in complete-linkage clustering between two final clusters: speech and non-speech.

Oracle VAD (OVAD), where the active speech labels calculated using the ITU-T P.56 active level VAD on the clean speech during construction of the noisy speech are kept and used as the VAD labels on the constructed noisy speech.

The performance of the EVAD and DVAD techniques, measured using the half-total error rate (HTER), in comparison the OVAD ground truth is shown in Figure 2. The contribution of the false-alarm rate (FAR) and miss-rate (MR) to the HTER is also shown as a dashed line, below which is $\frac{\text{FAR}}{2}$ and above which is $\frac{\text{MR}}{2}$. From this, it can be seen that while DVAD easily outperforms the simple EVAD approach across all noise scenarios and SNRs, the contribution of false alarms, or incorrectly-detected speech, to DVAD is proportionally much higher than in EVAD.

4.3. Speaker recognition

Our speaker recognition experiments were conducted with a Gaussian probabilistic linear discriminant analysis (GPLDA) approach based upon linear discriminant analysis (LDA) transformed, length-normalised, i-vectors, similar to our existing work [19] and other state-of-the-art speaker recognition systems in this space.

In our system we used 13 dimensional feature-warped mel-frequency cepstral coefficients (MFCCs) with appended delta coefficients. Two gender-dependent universal background model Gaussian mixture models (UBM-GMMs) with 512 components were used to map the MFCC features into a higher dimensional space, from which 500-dimensional i-vector were extracted and, prior to GPLDA modeling, LDA was used to reduce i-vectors dimension to 150 dimensions and WCCN was used to compensate intra-speaker variance. Length normalization i-vectors using i-vector centering and whitening were estimated to model the GPLDA parameters. Scoring was conducted using the batch likelihood ratio, followed by symmetric score normalisation (S-norm).

The speaker recognition performance across the 5 QUT-NOISE-SRE scenarios at a range of SNRs, and using the VAD techniques outlined in the previous section, is shown in Figure 3. Interestingly, it can be seen that while the dissimilarity-clustering DVAD had much better performance in VAD than the simple energy-thresholding EVAD, the EVAD technique often is comparable or better, particularly in the cleaner conditions. Even the perfect OVAD performance, when used for speaker recognition, is often bettered by EVAD, suggesting that, when using VAD as a front-end for speaker recognition, having perfect adhesion to the real speech boundaries is not important, as long as some useful speech is found. Further study should be conducted into how important false-alarms and misses are for VAD as a front-end to speaker recognition.

Across the noisy scenarios, it appears that all but the REVERB scenario perform similarly in cleaner conditions (SNRs > 0 dB), but that CAFE and HOME scenarios degraded further in noisier conditions, due largely to the more prominence of speech in the noisy recordings in these scenarios. The REVERB scenario is an outlier, but it is not clear at this stage if this is due to the high reverberation or the type of background noise present. Further study should investigate the creation of noisy speech in this scenario, but with the reverberation not applied to investigate this further.

5. Conclusion

Within this paper, we have outlined the development of the freely-available QUT-NOISE-SRE noisy speaker recognition protocol based upon the QUT-NOISE database [10]. We have also demonstrated the use of the QUT-NOISE-SRE protocol for the evaluation of a state-of-the-art PLDA i-vector speaker recognition system, demonstrating the importance of designing specific speaker-recognition focused VAD techniques.

We believe that this protocol provides a solid basis for the development, evaluation and benchmarking of robust speaker recognition algorithms that can operate across a wide variety of noise scenarios. To encourage further research in this area, we have made the QUT-NOISE-SRE protocol freely available available alongside the QUT-NOISE database at our website[3].

[3] https://qut.edu.au/research/saivt

6. References

[1] Luciana Ferrer, Harry Bratt, Lukas Burget, Honza Cernocky, Ondrej Glembek, Martin Graciarena, Aaron Lawson, Yun Lei, Pavel Matejka, Olda Plchot, et al., "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proceedings of NIST 2011 Workshop*, 2011.

[2] "The NIST year 2012 speaker recognition evaluation plan," Tech. Rep., National Institute of Standards and Technology, 2012.

[3] M.I. Mandasari, M. McLaren, and D. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *ICASSP 2012*. IEEE, 2012, pp. 4249–4252.

[4] Yun Lei, Lukas Burget, Luciana Ferrer, Martin Graciarena, and Nicolas Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4253–4256.

[5] Daniel Garcia-Romero, Xinhui Zhou, and Carol Y Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4257–4260.

[6] Luciana Ferrer, Mitchell McLaren, Nicolas Scheffer, Yun Lei, Martin Graciarena, and Vikramjit Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation.," in *INTERSPEECH*, 2013, pp. 1981–1985.

[7] Chengzhu Yu, Gang Liu, Seongjun Hahm, and John H.L. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *ICASSP*, Florence, Italy, 2014.

[8] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[9] D. Pearce and H.G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Sixth International Conference on Spoken Language Processing*. Citeseer, 2000.

[10] David B. Dean, S. Sridharan, Robert J. Vogt, and Michael W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Interspeech 2010*, Japan, September 2010, pp. 3110–3113.

[11] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[12] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *108th Audio Engineering Society Convention*. 2000, Citeseer.

[13] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, Mar 1992, vol. 1, pp. 517–520 vol.1.

[14] Christopher Cieri Joseph, Joseph P. Campbell, Hirotaka Nakasone, David Miller, and Kevin Walker, "The Mixer corpus of multilingual, multichannel speaker recognition data," in *Proc. 4 th International Conference on Language Resources and Evaluation*, 2004, pp. 26–28.

[15] "The NIST year 2008 speaker recognition evaluation plan," Tech. Rep., NIST, 2008.

[16] "The NIST year 2010 speaker recognition evaluation plan," Tech. Rep., NIST, 2010.

[17] Mike Brookes, "VOICEBOX: Speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[18] Houman Ghaemmaghami, David Dean, Shahram Kalantari, Clinton Fookes, and Sridha Sridharan, "Complete linkage clustering for voice activity detection in audio and visual speech," in *Interspeech 2015*, Dresden, Germany, 2015.

[19] Ahilan Kanagasundaram, Robert J Vogt, David B Dean, and Sridha Sridharan, "PLDA based speaker recognition on short utterances," in *The Speaker and Language Recognition Workshop (Odyssey 2012)*. ISCA, 2012.