Improving the Performance of Far-Field Speaker Verification Using Multi-Condition Training: The Case of GMM-UBM and i-vector Systems

Anderson R. Avila[1,2]*, Milton Sarria-Paja*[1]*, Francisco J. Fraga*[2]*,*
Douglas O'Shaughnessy[1]*, and Tiago H. Falk*[1]

[1]INRS-EMT, University of Quebec, Montreal, Quebec, Canada
[2]Universidade Federal do ABC (UFABC), Santo André, São Paulo, Brazil

Abstract

While considerable work has been done to characterize the detrimental effects of channel variability on automatic speaker verification (ASV) performance, little attention has been paid to the effects of room reverberation. This paper investigates the effects of room acoustics on the performance of two far-field ASV systems: GMM-UBM (Gaussian mixture model - universal background model) and i-vector. We show that ASV performance is severely affected by reverberation, particularly for i-vector based systems. Three multi-condition training methods are then investigated to mitigate such detrimental effects. The first uses matched train/test speaker models based on estimated reverberation time (RT) values. The second utilizes two-condition training where clean and reverberant models are used. Lastly, a four-condition training setup is proposed where models for clean, mild, moderate, and severe reverberation levels are used. Experimental results show the first and third multi-condition training methods providing significant gains in performance relative to the baseline, with the latter being more suitable for practical resource-constrained far-field applications.
Index Terms: Automatic speaker verification, GMM-UBM, i-vector, far-field, reverberation time.

1. Introduction

Channel variability and/or train-test mismatch have been regarded as serious detrimental factors for automatic speech and speaker recognition technologies. To overcome these limitations, several approaches have been proposed. In the feature domain, for example, techniques such as cepstral mean subtraction (CMS) [1], relative spectral (RASTA) processing [2], and feature mapping [3] have been used to minimize additive and convolutional channel distortions. In the scoring domain, in turn, normalization techniques such as Hnorm and Tnorm have been developed, to name a few [4]. Recently, joint factor analysis has been used to separate inter-speaker and intersession variability from augmented feature spaces (e.g., Gaussian mixture model supervectors) [5]. State-of-the-art automatic speaker verification (ASV) systems, today, are based on extensions to the joint factor analysis framework and constitute the so-called i-vectors, obtained after a total variability feature projection (e.g., [6, 7]).

While a lot of attention has been given to channel variability, limited work has been done to address the issue of room acoustics in far-field ASV, particularly regarding room reverberation. It is known that in far-field speech applications, the signal captured by the microphone is comprised of the direct path signal, plus numerous reflections off the walls, floor, and ceiling. Reverberation causes colouration of the speech signal, plus temporal smearing, which severely degrades the performance of several automated speech technologies [8, 9], particularly ASV. To overcome these detrimental effects, different approaches have been proposed. As examples for speaker verification and identification, microphone arrays [10], score normalization [11], feature normalization [8], and alternative feature representations [12, 13, 14] have been proposed. By far the most popular method of combating room reverberation, however, has been multi-condition training, where speaker models are developed for different reverberation levels and the best model is found during verification via a reverberation time (RT) estimator [11, 15, 16, 17]. Moreover, in these previous studies, synthetic room impulse responses (RIRs) and traditional features (mel-frequency cepstral coefficients, MFCC) have been commonly explored (e.g., [16]). To the best of the authors' knowledge, only one recent study has investigated the effects of reverberation on i-vectors corrupted by low reverberation levels using synthetic RIRs [18]. As such, the effects of real recorded RIRs, as well as higher reverberation levels, are still unknown.

Having this said, the goal of this paper is three-fold. First, we explore the gains achieved with three multi-condition training paradigms to combat the effects of reverberation on ASV performance. More specifically, we investigate the widely-used method of train-test reverberation level matching, as well as two alternate methods: i) the use of clean and "global" reverberant speech models, and ii) the use of clean, low, medium, and high reverberation level speaker models. In both cases, an in-house RT regressor/classifier is used [19]. Second, we test these multi-condition strategies using the ubiquitous MFCC features, as well as the more recent i-vector features. Lastly, we explore the performance of the above-mentioned feature and multi-condition training combinations using RIRs recorded in a varechoic chamber with a wide range of RT between $0.39 - 2$ s.

The remainder of this paper is organized as follows. Section 2 gives an overview on the two tested ASV paradigms, namely GMM-UBM and i-vector, as well as presents the three multi-condition training setups. Section 3 describes the experimental setup and Section 4 presents the experimental results and a discussion. Lastly, conclusions are presented in Section 5.

2. Baseline and multi-condition training

Typical text-independent speaker verification systems are comprised of a front-end for feature extraction, a modelling method for speaker enrolment (e.g., universal background model, UBM) and a final decision process (e.g., likelihood scoring in GMM-UBM systems). In the sections to follow, two baseline systems are described – GMM-UBM and i-vector – as well as three multi-condition training paradigms to combat the effects of room reverberation on ASV performance.

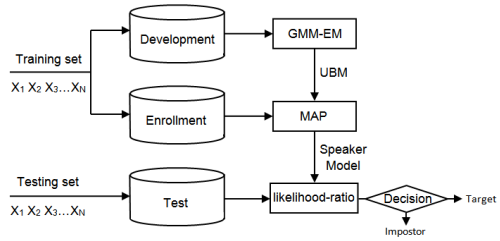


Figure 1: Block diagram of the GMM-UBM framework.

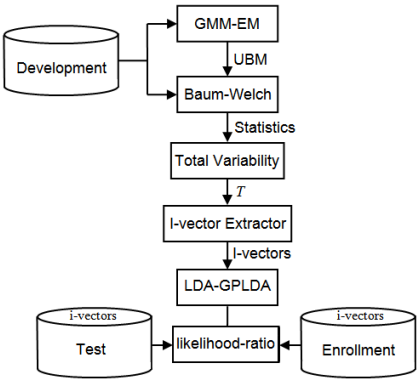


Figure 2: Block diagram of the i-vector framework.

2.1. Baseline ASV: GMM-UBM

The block diagram of the GMM-UBM baseline is depicted by Fig. 1. MFCC features, computed via a set of 24 triangular (mel) bandpass filters, were used. More specifically, 19 cepstral coefficients, along with log-energy, delta and double-delta coefficients were used to generate a 60-dimensional feature vector. With the GMM-UBM framework, Gaussian mixture model (GMM) parameters were obtained via the well-known expectation-maximization (EM) algorithm [20]. In our experiments, a 64-component diagonal covariance matrices GMM was used; we found this value to strike a good balance between model complexity and system performance on our dataset. During enrolment, speaker models are obtained via Maximum a Posteriori (MAP) adaptation. Scoring and decision are then performed based on log-likelihood thresholding.

2.2. Baseline ASV: i-vector

The block diagram of the i-vector based baseline system is depicted by Fig. 2. I-vectors are obtained from a joint factor analysis (JFA) framework [5, 21] where the means of the speaker-dependent MAP-adapted GMMs (trained on the above-mentioned 60-dimensional features) are combined into the so-called supervector *M*. The basic assumption is that the supervector conveys speaker dependent, speaker independent, channel dependent, and residual components. Each component can then be represented by a low-dimensional set of factors, which operate along the principal dimensions (also known as eigendimensions) of the corresponding component. Mathematically, this is represented as $M = m + Vy + Ux + Dz$, where m is the speaker and channel-independent supervector, V the speaker eigenvoice matrix, D the diagonal residual matrix, U is the eigen-session matrix, and y, z, x correspond to the low-dimensional eigenvoice, speaker-specific eigen-residual, and eigen-channel factors, respectively. More recently, the above framework was modified to include a total variability space projection, i.e., $M = m + Tw$, where m remains the mean supervector extracted from the universal background model, T corresponds to a rectangular low-rank matrix and w is a random vector with normal distribution, i.e., the so-called i-vectors [7].

Within the i-vector framework, the decision process consists of computing the similarity between the target speaker factors and the test speaker factors. To achieve this goal, several post-processing steps have been proposed, including linear discriminant analysis (LDA) to maximize the between-class variance and minimize the within-class variance, as well as Gaussian probabilistic linear discriminant analysis (PLDA) [22]. Decisions are then made based on log-likelihood thresholding of PLDA hyperparameters. In our experiments, we used 50 total factors in the total variability matrix, obtained via UBM and Baum-Welch statistics, as suggested in [7, 23]. LDA is then used to reduced the dimensionality to 35. In the dataset used in the experiments herein, these values showed to be optimal.

2.3. Multi-condition training

2.3.1. Train/test RT matching

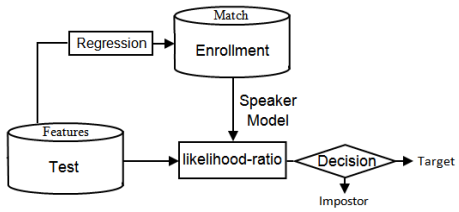
The first multi-condition training scheme represents what has been typically proposed in the literature, i.e., reverberation time matching between training and testing conditions. This setup is depicted by Fig. 3a and typically involves an RT regressor to estimate the RT of the test file. During training, UBM models are obtained under different reverberation time conditions. The enrolment consisted of the same reverberation time conditions used previously for modelling. During verification, the models that more closely match the RT of the test data are used. Such an approach, despite its popularity, has some disadvantages. First, it requires storage of several speaker models, thus may place a burden on resource-constrained ASV applications. Second, RT estimation may be sensitive to additive ambient noise, thus generating erroneous RT estimates in practical everyday settings. Lastly, train-test RT matching may result in overly optimistic performances, considering that a perfect reverberant matching very difficult in real world applications. In our experiments, the RT estimator described in [19] was used, as it has been shown to be robust to high reverberation levels with and without ambient noise [24].

2.3.2. Two-condition training: clean and reverberant models

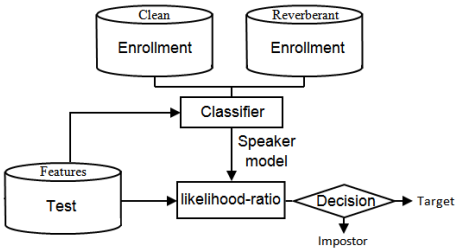
To overcome some of the issues mentioned above of train/test RT matching (particularly that of model storage and generalizability to everyday scenarios), a two-condition training scheme is tested where speaker models are obtained for clean speech and for reverberant speech (see Fig. 3b). Unlike the RT-matched case, here a global model of reverberant speech is used encompassing training data corrupted by varying RT values. In our experiments, we use RT values in the range of 300-2000 ms to generate the "global" reverberant speaker models and also for enrolment. A support vector machine classifier, as proposed in [19], is used to classify between clean and reverberant speech.

2.3.3. Four-condition training: clean, low-, medium-, and high-reverberation speaker models

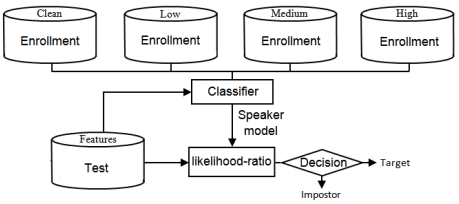
The third multi-condition training setup tested involves an intermediate between the two previously-mentioned schemes. More specifically, a four-condition training scheme is used where speaker models of clean speech and speech corrupted by low



(a)



(b)



(c)

Figure 3: Block diagram of (a) train/test RT-matched, (b) two-, and (c) four-condition multi-condition training paradigms.

(RT$\leq$ 0.7 s), medium (0.7 $<$RT$<$ 1s), and high levels (RT$\geq$ 1 s) of reverberation are used (see Fig. 3c). As before, the enrolment process considered the same RTs used for training and a support vector machine classifier is used to detect which class the test speech file belongs to.

3. Experimental setup

3.1. Speech database

In our experiments, the CHAINS corpus was used [25, 26]. The database comprises recordings of 36 subjects, including male and female, with different English accents. Six different speaking styles are provided (e.g., normal, whisper, fast) but we used only the clean normal speaking style data recorded in a professional studio. Each speaker read 37 prepared texts. The first four files (read paragraphs) were used for training and resulted in roughly 120 seconds of speech material per speaker. The remaining 33 speech files were used for testing and provided approximately 60 seconds of speech material per speaker. Data was originally sampled at 16 kHz with 16-bit resolution, but downsampled to 8 kHz and energy normalized to -26 dBov prior to feature extraction using the ITU-T P.56 voltmeter [27].

3.2. Recorded RIR and reverberant material generation

Room impulse responses (RIRs) were obtained in a professional varechoic chamber (length 7 m, width 9 m x height 3.5 m) available at the École de Technologie Supérieure in Montreal, Quebec, Canada. Measurements were conducted using four powered loudspeakers positioned in the lower corners of the room.

(a)

(b)

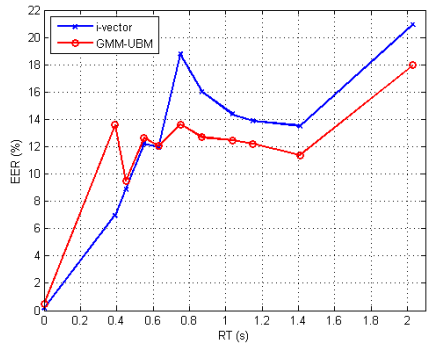
Figure 4: Varechoic reverberation chamber (a) without and (b) absorbers to control reverberation time values

RIRs were recorded using the chirp method with a microphone placed in the centre of the room. Absorbers were placed around the reverberation chamber to result in ten different RT values: 0.39, 0.45, 0.55, 0.63, 0.75, 0.87, 1.04, 1.15, 1.41, and 2.03 seconds. To generate the reverberant speech files used in our experiments, the clean speech files from the CHAINS corpus were convolved with the recorded room impulse responses. Noticed that the same RIRs were used for both training and testing. For our experiments, the MSR identity toolbox was used [28].

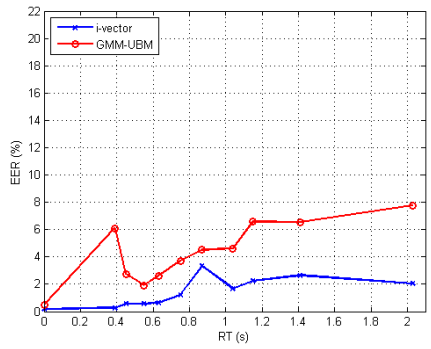
4. Experimental Results and Discussion

Figures 5 (a-d) depict the equal error rates (EER) obtained with the baseline, train/test RT-matched, two-condition, and four-condition training setups, respectively. As can be seen with the baseline setup (Fig. 5a), the GMM-UBM outperforms the i-vector system across the majority of the investigated RT values. As expected, significant performance degradation is observed with an increase in RT. Table 1 presents statistics of the obtained EERs for the four tested scenarios. For clean conditions, an EER of 0.5% was seen. This increased to 18% at RT$=$ 2 s for the GMM-UBM and 21% for the i-vector based systems.

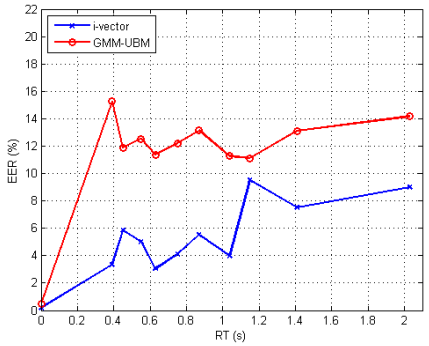
In the train/test RT-matched scenario (Fig. 5b), an inverse relationship was found and the i-vector outperformed the GMM-UBM ASV system across all tested RTs. In the i-vector case, the EER remained below 1% up to an RT$=$ 0.75 s and increased to 2% at RT$=$ 2 s. This amounts to a significant relative reduction of up to 90% in EER. For the GMM-UBM case, on the other hand, EER stayed below 6% up to an RT$=$ 1 s and increased to 8% at the highest RT value. Despite the poor performance relative to the i-vector based system, the RT-matched setup significantly improved performance and a relative reduction of 56% could be seen at this highest RT value. With this setup, the RT estimator from [19] achieved a correlation of 0.96 with true RT and an error of 12 ms. Despite these promising results, an RT-matched system is not very practical, due to e.g.,



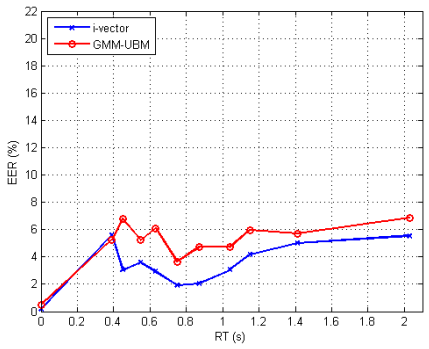
(a) baseline GMM-UBM and i-vector ASV



(b) train/test RT-matched



(c) two-condition training scenario



(d) four-condition training scenario

Figure 5: Equal error rate (EER) performance comparison.

	GMM-UBM				i-vector			
Setup	**min**	**max**	$\overline{x}$	σ^2	**min**	**max**	$\overline{x}$	σ^2
Baseline	0.5	17.96	10.80	17.99	0.18	20.93	11.55	32.54
RT-matched	0.5	7.77	4.04	5.13	0.18	3.33	1.36	1.12
2-condition	0.5	15.27	10.63	14.95	0.18	9.50	4.83	7.47
4-condition	0.5	6.88	4.71	3.18	0.18	5.61	3.17	2.75

Table 1: GMM-UBM and i-vector performance for baseline, train/test RT-matched, two-, and four-condition training setups.

storage and time-varying RT limitations; in such cases, the two- and four-condition training schemes are more appropriate.

For the two-condition training setup (Fig. 5c), the gains relative to the baseline were not as substantial as with the RT-matched system. For the GMM-UBM, EER had a significant drop in performance even for low RT values. As an example, from clean to RT$=$ 0.39 s, the EER went from 0.5% to 15% (97% relative increase). Performance, however, did remain stable across all tested RT values and ranged between 11 and 15%. For the i-vector system, the EER increased almost linearly with RT. At RT$=$ 2 s, an EER of 9% was achieved, thus still significantly better than the baseline (57% relative reduction).

Lastly, for the four-condition training setup (Fig. 5d) it was observed that the gap between GMM-UBM and i-vector performances was small, with the i-vector slightly outperforming the GMM-UBM system. In both cases, results were fairly stable and varied between 2-7% across all tested RT values. In both cases, the performance was best between 0.7$\leq$RT$\leq$1 s, suggesting the mid-reverberation models captured speaker characteristics more reliably. Overall, the four-condition training setup provided the most practical solution to hands-free far-field ASV, as it provided reliable results across a wide range of RT values (EER$<$7% for GMM-UBM; EER$<$6% for i-vector) without the burden of multi-model storage and computational complexity. In these experiments involving only reverberation, the RT classification stage was shown to be accurate and with a small footprint; additional studies are needed to test RT-level classification performance in the presence of additive ambient noise.

5. Conclusion

This paper has investigated the effects of reverberation on two automatic speaker verification systems: one based on the GMM-UBM paradigm and another on the burgeoning i-vector features. In a baseline setup, the i-vector based systems were shown to be most sensitive to the detrimental reverberation effects. Three multi-condition training setups were then investigated: train/test reverberation time (RT)-matched, two-condition (clean and reverberant), and four-condition (clean, low-, medium-, and high-reverberation levels) training. While the RT-matched scenario resulted in the best overall performance, with the i-vector outperforming the GMM-UBM system, the four-condition training setup resulted in the most practical solution for resource constrained applications, achieving reliable performances across a wide range of RT values.

6. Acknowledgements

The authors acknowledge funds from the Centre for Advanced Systems and Technologies in Communications (SYTACOM) and the Emerging Leaders in the Americas Program (ELAP).

7. References

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans.on Speech and Audio Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[3] D.A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2003, vol. 2, pp. 53–56.

[4] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D.A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.

[5] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep. CRIM-06/08-13, CRIM, 2005.

[6] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proc. INTERSPEECH*, 2009, vol. 9, pp. 1559–1562.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[8] S. Ganapathy, J. Pelecanos, and M.K. Omar, "Feature normalization for speaker verification in room reverberation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2011, pp. 4836–4839.

[9] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2023–2032, 2007.

[10] J. González-Rodríguez, J. Ortega-García, C. Martín, and L. Hernández, "Increasing robustness in gmm speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proc Intl Conference on Spoken Language*. IEEE, 1996, vol. 3, pp. 1333–1336.

[11] I. Peer, B. Rafaely, and Y. Zigel, "Reverberation matching for speaker recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2008, pp. 4829–4832.

[12] S.O. Sadjadi and J.H. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. IEEE Intl Conference on Acoustics, Speech and Signal Processing*. IEEE, 2011, pp. 5448–5451.

[13] T.H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 90–100, 2010.

[14] T.H. Falk and W.-Y. Chan, "Spectro-temporal features for robust far-field speaker identification," in *Proc. INTERSPEECH*, 2008, pp. 634–637.

[15] J.S. Gammal and R.A. Goubran, "Combating reverberation in speaker verification," in *Proc. of the IEEE Instrumentation and Measurement Technology Conference*. IEEE, 2005, vol. 1, pp. 687–690.

[16] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Trans Audio Speech and Language Processing*, vol. 22, no. 4, pp. 836–845, April 2014.

[17] L. Couvreur and C. Couvreur, "Blind model selection for automatic speech recognition in reverberant environments," in *Real World Speech Processing*, pp. 115–129. Springer, 2004.

[18] D. Garcia-Romero, X. Zhou, 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 Int. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2012, pp. 4257–4260.

[19] T.H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, 2010.

[20] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[21] N. Dehak, R. Dehak, Kenny P., and P. Dumouchel, "Comparison between factor analysis and GMM support vector machines for speaker verification," in *Proc. IEEE-Odyssey of the Speaker and Language Recognition Workshop*, 2008.

[22] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th Int. Conf. on Computer Vision*, 2007, pp. 1–8.

[23] P. Kenny, "A small foot-print i-vector extractor," in *Proc. IEEE-Odyssey of the Speaker and Language Recognition Workshop*, 2012, pp. 1–6.

[24] N.D. Gaubitch, H.W. Loellmann, M. Jeub, T.H. Falk, P.A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Proc Intl Workshop on Acoustic Signal Enhancement*. VDE, 2012, pp. 1–4.

[25] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains speech corpus: Characterizing individual speakers," in *Proc. Int. Conf. on Speech and Computer (SPECOM)*, 2006, pp. 1–6.

[26] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1097–1111, 2008.

[27] ITU-T, "ITU-T recommendation P.56: Objective measurement of active speech level," Tech. Rep., International Telecommunication Union, Geneva, S., 2011.

[28] S.O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox - a matlab toolbox for speaker recognition research," Tech. Rep. MSR-TR-2013-133, Microsoft Research, Conversational Systems Research Center (CSRC), 2013.