Automatic Speaker Recognition System in Adverse Conditions — Implication of Noise and Reverberation on System Performance

Khamis A. Al-Karawi, Ahmed H. Al-Noori, Francis F. Li, and Tim Ritchings

***Abstract*—Speaker recognition has been developed and evolved over the past few decades into a supposedly mature technique. Existing methods typically utilize robust features extracted from clean speech. In real-world applications, especially security and forensics related ones, reliability of recognition becomes crucial, meanwhile limited speech samples and adverse acoustic conditions, most notably noise and reverberation, impose further complications. This paper is presented from a study into the behavior of typical speaker recognition systems in adverse retrieval phases. Following a brief review, a speaker recognition system was implemented using the MSR Identity Toolbox by Microsoft. Validation tests were carried out with clean speech and the speech contaminated by noise and/or reverberation of varying degrees. The image source method was adopted to take into account real acoustic conditions in the spaces. Statistical relationships between recognition accuracy and signal to noise ratios or reverberation times have therefore been established. Results show noise and reverberation can, to different extents, degrade the performance of recognition. Both reverberation time and direct to reverberation ratio can affect recognition accuracy. The findings may be used to estimate the accuracy of speaker recognition and further determine the likelihood a particular speaker.**

***Index Terms*—Clean speech, GMM-UBM, ISM, reverberation, robust speaker recognition, MFCC, MSR toolbox, noise.**

I. Introduction

The performance of speaker recognition can be affected by noise and reverberation and hence degraded. The variation in speech signals is caused by the environments, the speaker themselves and signal acquisition equipment. Robust speaker recognition in real world applications remains a technical challenge. While much of the attention has been paid to channel variability, limited work has been done to address the issue of room acoustics and background noise in far-field of Automatic Speaker Verification (ASV), particularly regarding room reverberation and noise [1]. In far-field applications, it is known that the signal captured by the microphone involves the direct path signal, and a huge number of reflections off the walls, floor, and ceiling. Reverberation causes coloration of the speech signals and temporal spreading, which severely degrades the performance of most automated speech technologies. To address this issue, various techniques have been proposed. Microphone arrays [1], score normalization [2], feature normalization [3] and alternative feature representations [4] have been suggested. The impacts of reverberation on neural network based speaker recognition has been studied in [5]. The effect of reverberation on speaker recognition systems using Gaussian mixture models (GMM), hidden Markov models (HMM) and quantization models (VQ) has been tackled in [1]. Speaker verification using GMM with reverberation has been addressed in [6]. Several methods for Speech enhancement, such as spectral subtraction, have been investigated for noise-robust speaker recognition [7]. In the feature domain, for instance, techniques such as Cepstral Mean Subtraction (CMS) [1], Relative Spectral (RASTA) processing [2], and feature mapping [3] have been utilized to reduce additive and convolutional channel distortions. Previous efforts were typically reistracted to simplistic and specific case reports with too limited information to formulate the relations between the system performance and acoustic conditions, e.g. [1] only reported 3emperical results. Rrecently, the computational auditory scene analysis (CASA) has been engaged to remove noise [8]. In general the community of speaker recognition has concentrated on channel variations in speaker verification. The National Institute of Standards and Technology (NIST) has conducted a series of speaker recognition evaluations (SRE) since 1996. State-of-the-art systems include joint factor analysis [9] and i-vector based techniques [10]. May [11] and Gonzalez-Rodriguez [1] studied the combined impacts utilizing binaural cues and microphone arrays. Krishnamoorthy and Prasanna [12] described the results in noisy and reverberant conditions separately. In this paper, we further investigate the relations between noise and reverberation and the performance of a typical speaker recognition system through a large number of accurately simulated cases. This paper proceeds as follows: Section II, SV background; Section III, the role of noise and reverberation; Section IV, the speaker recognition system; Section V, experiments and results; Section VI, experiment Result Discussion and Section VII, the concluding remarks.

II. SV Background

Speaker recognition is defined as the process of recognising the identity of a person by analysing their speech signals. Speaker recognition can be classified into identification and verification. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the

Manuscript received December 25, 2014; revised February 28, 2015.

The authors are with School of Computing Science and Engineering, University of Salford, UK (e-mail: k.a.yousif@edu.salford.ac.uk).

DOI: 10.7763/IJIEE.2015.V5.571

process of accepting or rejecting the identity claim of a speaker. Speaker recognition methods can also be divided into text-independent and text dependent methods. Text-independence means that the system doesn't use any prior knowledge about the speech contents used in the training. Statistical approach is often used to model speakers as well as to achieve verification [13]. Suppose *X* denotes the collection of feature vectors acquired from the test data, and *k* refers to the claimed speaker identity which has a corresponding reference model C_k. So the verification decision is given by: accept speaker *k*, if $d(C_k, x) < T_k$ or reject speaker *k*, if $d(C_k, x) > T_k$.

where *Tk* is the verification threshold, and *d(Ck, x)* represents some "distance" measure between the test data and the reference model *Ck*. The distance measure is computed by finding the likelihood of the test data being created by the reference model, which is given by:

$$d(c_k, x) = \log(p(x|c_k)) \quad (1)$$

State-of-the-art automatic speaker verification (ASV) systems, today, are based on the extensions to the joint factor analysis framework and constitute the so-called i-vectors, obtained after a total variability feature projection [10].

III. The Role of Noise and Reverberation

Received speech signals by a microphone can be modeled as the convolution between the speech signal and the room impulse response [14], the latter includes direct sound, early reflections and reverberation. When contributions from reflections and reverberation are significant compared to the direct sound, the speech is said to be reverberated. Reverberation time(T60 or RT) is the main parameter used to quantify reverberation [14], [15], which is the time that it takes for the sound pressure in the room to decay by 60dB after the source is switched off. The impact of reverberation can be modeled as the processing of a signal by a linear time invariant system. Taking into account the background noise the received speech signals from a microphone can be written as:

$$y(k) = x(k)h(k) + n(k) \quad (2)$$

where $y(k)$ represent the received speech signals, $x(k)$ is the original speech signal, $h(k)$ denotes the room impulse response, and $n(k)$ is the ambient noise.

IV. The Speaker Recognition System

A. MSR Toolbox

Microsoft Speaker recognition (MSR) identity toolbox [16] is a Matlab toolbox developed by Microsoft Research which includes a collection of Matlab tools and functions to facilitate the development of speaker recognition. It provides flexibility for researchers in developing new front-end and back end techniques, allowing fast prototyping and rapid evaluation of new advancement. The front-end of this toolbox is responsible for transforming the speech signals to acoustic feature. The Cepstral features are most commonly used from this toolbox. The back-end includes the training and testing modules. The training (enrollment) stage is responsible for generating models for each register. In the test stage, the speech segment under testing is scored against all enrolled speaker models to determine if the speaker is the target speaker or just an imposter. The MSR identity toolbox provides two popular tools for speaker modelling the GMM-UBM and i-vector paradigms. In the GMM-UBM framework the universal background model (UBM) represents Gaussian mixture models (GMM) trained in a pool of data from a large number of speakers. The speaker dependent models are then adapted from UBM using maximum a posteriori (MAP) estimation. Fig. 1. Illustrate the structure of the GMM-UBM model. The Gaussian mixture model (GMM)-based speaker identification algorithm is popular due to its good performance [17]. In this paper GMM -UBM is used.

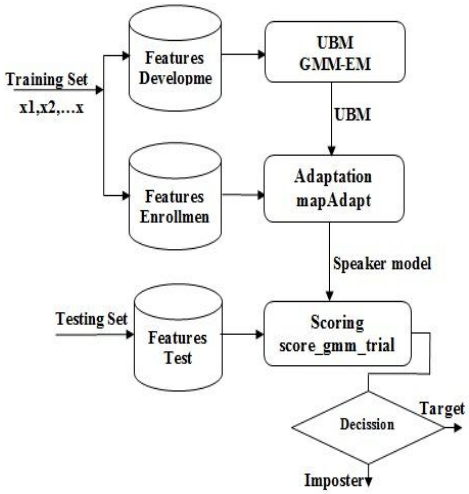

Fig. 1. Block diagram of the GMM-UBM framework [16].

B. Image Source Method

The image-source model (ISM) [18] is a technique used to generate synthetic room impulse responses (RIRs). Once the RIR is available, reverberated speech samples can be simulated by converlution according to (2). The ISM used for typical rooms were utilized to simulate the effects of reverberation, and rectangular rooms were considered. The simulation program takes as its input into four sets of values or dimensions. The first set is the dimensions of the room, length, width and height, the second set is the source location, the third is the receiver location, and the last is 6 reflection coefficients for each surface of the room. The ISM technique has a number of significant benefits compared to other approaches for room acoustics simulation [19], it generates a large number of virtual rooms for the study of the relations between reverberance and speaker recognition in this paper.

As the absorption of sound depends upon frequency it is clear that the reverberation time of an enclosure will also be frequency dependent, it is usual to estimate or mesure reverberation time in octive or third octave bands from 125-4KHz [20]. Fig. 2 shows the process of ISM, in which a box represents the room, the black circle represents the position of the source, and the triangle represents the position of the receiver.

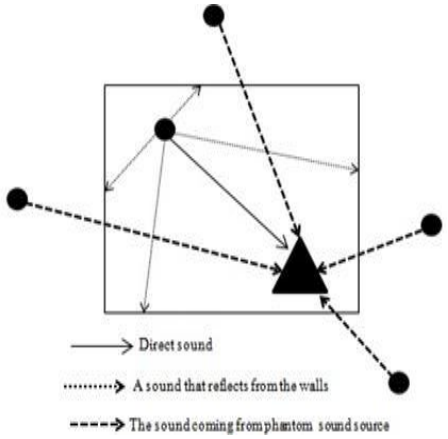

Fig. 2. Image source method process.

C. Simulated Reverberant Speech

Image source method developed in [21] and an implementation in Matlab by Eric A. Lehmann [22] were used in this study to produce reverberated signals from clean signals. The signals in Fig. 3 demonstrate reverberant speech signals produced by the ISM for T60=0.1, 0.5, and 1s. In the clean waveform the sharp points relate to the consonants. When the reverberation increases.

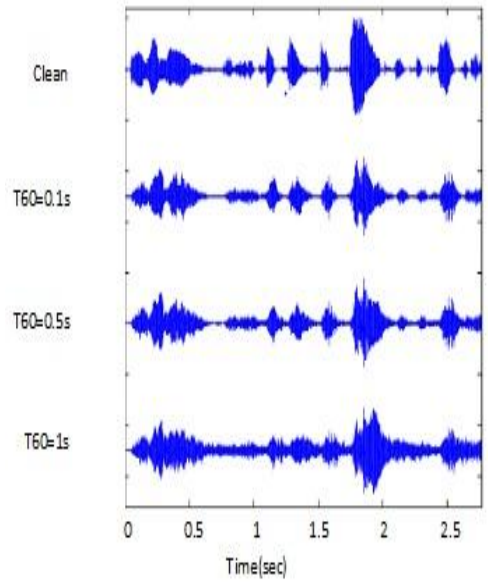

Fig. 3. Waveforms, top to bottom: clean and reverberant speech to T_{60} =0.1 ,0.5 ,1s.

V. Experiments and Results

A. Speech Database

The speech samples were recorded using a Zoom H4 recorder. Speech samples were collected from 19 speakers 11 males and 8 females for both noise and reverberation experiments. Speakers were between 25-40 year of age. Each speaker provided 5 clean speech samples and an additional one was recorded in reverberation condition (in a small reverberation chamber at Salford University), Lengths of speech samples were between 30s and 40s. The utterances for all speakers in reverberation case included the same text (text dependent), while each speaker spoke different text and language in the noise case. The speech is sampled at 16 KHz.

B. Baseline ASV: GMM-UBM

The block diagram of the GMM-UBM baseline is illustrated in Fig. 1. MFCC features, calculated through a set of triangles (Mel) bandpass filters, were utilized. Cepstral coefficients, along with log-energy, delta coefficients were utilized to generate a 39-dimensional feature vector from pre-emphasized speech signal, and then the mean and variance normalized. With the GMM-UBM framework, Gaussian Mixture Model parameters were acquired through the expectation-maximization (EM)algorithm [13]. During enrollment, speaker models were acquired through Maximum a Posteriori (MAP) adaptation. Scoring and the decision was then performed on log-likelihood thresholding.

C. Reverberation Experiment

To quantify the relations between recognition performance and reverberance, two methods for were followed. First, the samples recorded in the Salford university reverberation room. The MSR accuracy as using these samples in the testing phase against clean samples in the learning phase only achieved 10%. Fig. 4 Shows MSR toolbox process with simulation software.

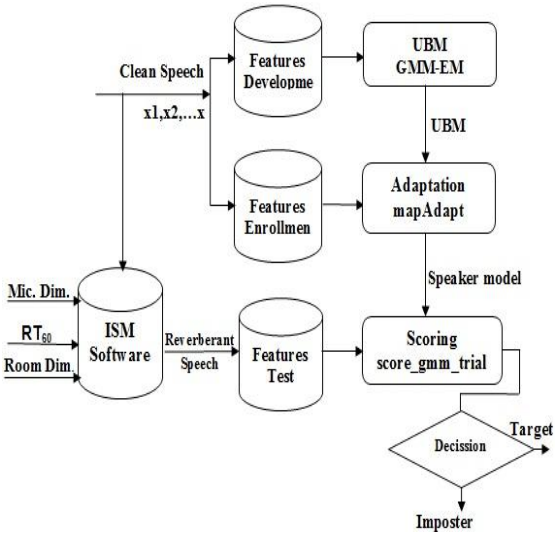

Fig. 4. MSR toolbox process with simulation software.

The other method, as discussed in Section 4.2 was utilized to generate reverberation samples with image source computer simulation. A different room dimensions ($3\times4\times2.5$, $4\times4\times2.5$ and $5\times4\times2.5$ m^3), five different times (0.1, 0.5, 1, 1.5 and 2s) were set with each room. Furthermore, the distance between microphone and source were also variable and three distances were used as shown in Table II.

D. Noise Experiment

Three types of noises were used in the experiments, white, pink and tonal noises. The noisy utterances were obtained by mixing noises with the clean speech with variable ratios according to (3). Fig. 5 illustrates the mixing process of speech and noise

$$Mixingaudio = (S \times NR) + (NS \times 100 - NR) \quad (3)$$

where, *S* the is the speech signal, *NR*, the noise ratio and *NS*, *is the* noise signal. Nine speech samples are mixed with 3 types of noises to produce (297)speech over noise samples

distributed in 11 mixing groups for each type of noise depends on mixing percentage. Table I shows this process.

TABLE I: SPEECH AND NOISE MIXING PERCENTAGE

Speech	100	90	80	70	60	50	40	30	20	10
Noise	0	10	20	30	40	50	60	70	80	90

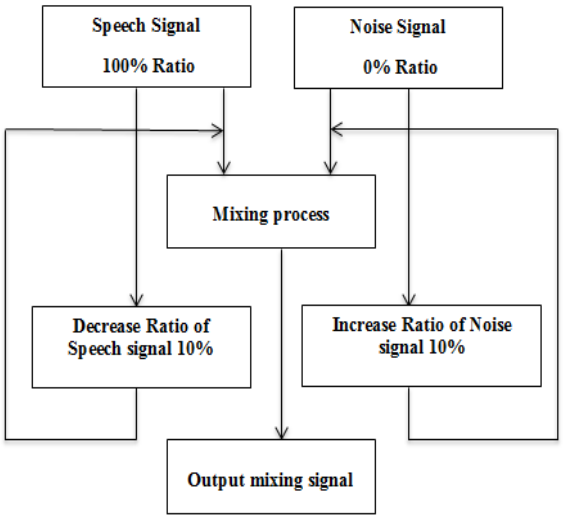

Fig. 5. Mixing of speech and noise with different ratio.

VI. EXPERIMENT RESULT DISCUSSION

A. Noise Experiment Result Discussion

The aim of this experiment is to characterize accuracy of the MSR in different speech to noise ratio (SNR). Fig. 6 illustrates the degradation in recognition accuracy when noise increases. The X axis represents the ratio of speech over noise, while the Y axis represents the recognition accuracy of the system. Regarding additive white noise, the system accuracy drops rapidly from 100% in clean speech to 30% when adding 10% noise. Then the accuracy of the system continues to decrease to 10% when increasing the noise to a 70/30 speech over noise ratio. The pink noise degrades system accuracy to approximately 56% when mixing with 10% noise, and the system accuracy continues to decrease to 0 with a 30/70 speech to noise ratio. On the other hand, tonal noise shows little effect compared with two other noises, but it still has a tangible effect on the accuracy of the system, when mixing with a 40/60 speech to noise ratio, recognition accuracy becomes 55%.

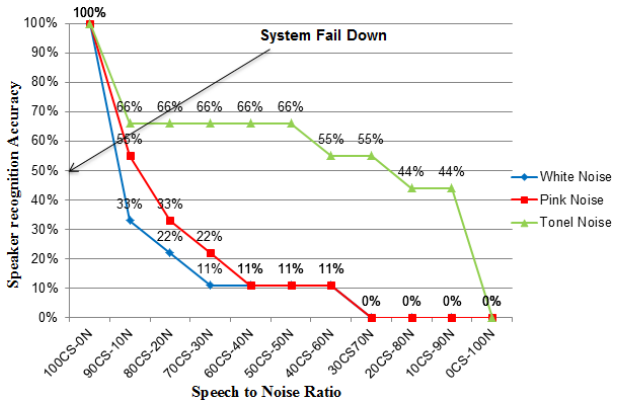

Fig. 6. Performance of speaker recognition in different speech to noise ratio.

B. Reverberation Experiment Result Discussion

Reverberation was added to the speech utilizing the Image source method [21]. Three reverberant impulse responses were utilizing. The specifications of the room used to generate the reverberant impulses are shown in Table II. We first establish a set of baseline results using clean signals, i.e. training and testing with independent but clean speech samples. For the baseline the accuracy of MSR in verification case was 100%. However, the MSR results of reverberation samples which recorded in the described room was only 10%. To further investigate the relations between reverberation and system performance, simulated room impulse responses are used. On the other hand, The direct-to-reverberant energy ratio (DRR) is also a significant parameter. Moreover, there are some applications that depend on DRR such as speech enhancement at a specific distance, source distance estimation, dereverebration and spatial audio coding. Several ways are available for estimating DRR. The Direct to Reverberation Ratio (DRR) is defined as [23]:

$$DDR = 10\log_{10} + \frac{h^2(\delta)}{\sum_{k=0(k\neq\delta)}^{m-1} h^2(k)} dB \quad (4)$$

where, $h(n)$ is the speaker-to-receiver impulse response M, it is length in samples and δ the time-index of the direct path in samples. MSR accuracy results for the simulation of the SV in reverb 1, reverb 2 and reverb 3 using clean training signals and reverberant test signals are shown in Fig. 7.

TABLE II: REVERBERATION SPECIFICATIONS

Specification	Reverberation Model		
	Reverb 1	Reverb 2	Reverb 3
Room dimensions	3× 4×2.5 m	4× 4× 2.5 m	5× 4×2. 5 m
Romm volume	30 m^3	40 m^3	50 m^3
Mic. Position	1	1.5	2 m
Source position	Fixed		
RT_{60}	0.1 , 0.5 , 1 , 1.5 , 2 Second		
Walls reflection coefficients.	0.5 , 0.6 , 0.1 , 0.8		
Ceiling reflection coefficients.	0.9	0.9	0.9
Carpeted floor reflection coefficients.	0.6	0.6	0.6

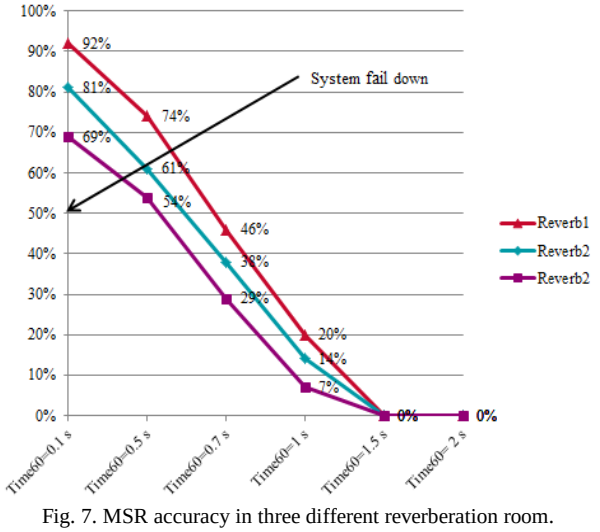

Fig. 7. MSR accuracy in three different reverberation room.

According to the system accuracy data from clean and reverberated samples, a regression model can be obtained

$$DR \approx 100 \times [\frac{RT}{Rvolume} + (1 - RT)] \quad (5)$$

where, *DR* denotes the degradation rate, *RT* denote reverberation time and *Rvoulme* denote room volume.

VII. Conclusion

This paper investigates the relations between the accuracy of speaker recognition and adverse acoustic conditions.

In particular a regression formula has been established to predict the recognition accuracy of a typical speaker recognition system. Validation is left for future work with more room types and speaker recognition engines.

References

[1] J. Gonz ález-Rodr guez, J. Ortega-Garc á, C. Mart ín, and L. Hern ández, "Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proc. Fourth International Conference on Spoken Language*, 1996, pp. 1333-1336.

[2] I. Peer, B. Rafaely, and Y. Zigel, "Reverberation matching for speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4829-4832.

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

[4] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 90-100, 2010.

[5] P. J. Castellano, S. Sradharan, and D. Cole, "Speaker recognition in reverberant enclosures," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 117-120.

[6] I. A. McCowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *Proc. a Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[7] W. Ning, P. C. Ching, Z. Nengheng, and L. Tan, "Robust speaker recognition using denoised vocal source and vocal tract features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 196-205, 2011.

[8] X. Zhao, Y. Shao, and D. Wang, "CASA-based robust speaker identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1608-1616, 2012.

[9] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Montreal,(Report) CRIM-06/08-13, 2005.

[10] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788-798, 2011.

[11] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2016-2030, 2012.

[12] P. Krishnamoorthy and S. M. Prasanna, "Application of combined temporal and spectral processing methods for speaker recognition under noisy, reverberant or multi-speaker environments," *Sadhana*, vol. 34, pp. 729-754, 2009.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[14] H. Kuttruff, *Room Acoustics*, CRC Press, 2009.

[15] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, pp. 409-412, 1965.

[16] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox - A matlab toolbox for speaker recognition research," Microsoft Research, Conversational Systems Research Center (CSRC), 2013.

[17] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, pp. 210-229, 2006.

[18] S. Van Vuuren, "Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch," in *Proc. of Fourth International Conference on Spoken Language*, 1996, pp. 1788-1791.

[19] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1429-1439, 2010.

[20] M. Kahrs and K. Brandenburg, *Applications of Digital Signal Processing to Audio and Acoustics*, vol. 1, Springer, 1998.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small‐room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943-950, 1979.

[22] E. A. Lehmann. (October). [Online]. Available: http://www.eric-lehmann.com/

[23] M. Tonelli, "Blind reverberation cancellation techniques," 2012.

Khamis A. Al-Karawi received a BSc degree in computer science from Baghdad University, Iraq in 1997. He received an Mc.s in computer science from Pune University, India in 2010. He is currently studying towards a PhD at Salford University, United Kindom.

Ahmed H. Al-Noori received a BSc degree in computer science from Al-Nahrain University, Iraq in 2000. He also received an Mc.s in computer science/genetic algorithm-programming from Al-Nahrain University, Iraq in 2003. He is currently doing his PhD at Salford University, United Kindom.

Francis F. Li received a B.Eng. from the East China University of Science and Technology, an MPhil from University of Brighton, and a PhD from the University of Salford, UK. Francis is currently with the School of Computing Science and Engineering, Salford University, where he teaches on BSc and MSc levels, supervises PhDs, and carries out research. His research interests include speech, music and multimedia signal processing; artificial intelligence; soft-computing; architectural acoustics; data communications; bio-medical engineering; and instrumentation. Francis published over 100 research papers and a book. He is an associate editor in chief for SPIJ and an associate technical editor for J. AES.