Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch

Seyed Omid Sadjadi and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Room reverberation poses various deleterious effects on performance of automatic speech systems. Speaker identification (SID) performance, in particular, degrades rapidly as reverberation time increases. Reverberation causes two forms of spectro-temporal distortions on speech signals: i) self-masking which is due to early reflections and ii) overlap-masking which is due to late reverberation. Overlap-masking effect of reverberation has been shown to have a greater adverse impact on performance of speech systems. Motivated by this fact, this study proposes a blind spectral weighting (BSW) technique for suppressing the reverberation overlap-masking effect on SID systems. The technique is blind in the sense that prior knowledge of neither the anechoic signal nor the room impulse response is required. Performance of the proposed technique is evaluated on speaker verification tasks under simulated and actual reverberant mismatched conditions. Evaluations are conducted in the context of the conventional GMM-UBM as well as the state-of-the-art i-vector based systems. The GMM-UBM experiments are performed using speech material from a new data corpus well suited for speaker verification experiments under actual reverberant mismatched conditions, entitled MultiRoom8. The i-vector experiments are carried out with microphone (interview and phonecall) data from the NIST SRE 2010 extended evaluation set which are digitally convolved with three different measured room impulse responses extracted from the Aachen impulse response (AIR) database. Experimental results prove that incorporating the proposed blind technique into the standard MFCC feature extraction framework yields significant improvement in SID performance under reverberation mismatch.**

***Index Terms*—Mismatch conditions, NIST SRE, overlap-masking effect, reverberation, speaker verification.**

I. Introduction

RECENT advancements in DSP technology have enabled integration of automatic speech systems into a variety of electronic/mobile components of an individual's daily life. Nevertheless, providing robustness to these systems still remains a challenge because of the variety of acoustic mismatch scenarios that may occur between training and test conditions due to background noise, room reverberation, communication channel, accent, language, emotions, vocal effort, etc.

Manuscript received September 13, 2013; revised December 23, 2013; accepted March 06, 2014. Date of publication March 11, 2014; date of current version April 04, 2014. This project was supported in part by the AFRL under Contract FA8750-12-1-0188 and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Thomas Fang Zheng.

The authors are with the Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, Richardson, TX 75080-3021 USA (e-mail: john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2311329

Specifically, performance of automatic speaker identification (SID) systems severely degrades in the presence of room reverberation [1], [2]. Reverberation poses various detrimental effects on spectro-temporal characteristics of speech signals, most notably including temporal smearing, filling dips and gaps in the temporal envelope, increasing the prominence of low-frequency energy, and flattening formant transitions [3]. These effects in turn mask higher frequencies in the speech spectrum and blur spectral details, both of which are useful acoustic cues for speaker identification.

Several compensation techniques for alleviating the adverse impact of room reverberation on SID performance have been reported in the literature, most of which were first developed for automatic speech recognition (ASR) or speech enhancement. The techniques have been applied at different stages of SID systems, i.e., front-end (signal and feature) [4]–[6], modeling [7], [8], and scoring stages [5], [7].

At the signal level, multichannel (e.g., microphone arrays) speech processing techniques have been employed to provide robustness to SID systems in reverberant and/or noisy conditions [4], [9]. However, this is not applicable in cases where only a single-channel signal (e.g., telephone) or prerecorded mono speech data is available. Long-term log-spectral subtraction (LTLSS) [10] is a single-channel homomorphic filtering method that works in a similar manner as cepstral mean subtraction (CMS) [11], except that it is applied at the signal level over a relatively long analysis window (typically longer than 1 second). The use of a long-term analysis window imposes the need for signal reconstruction before feature extraction. A more recent signal level technique to address reverberation is based on non-negative matrix factorization (NMF) [12]. The NMF based approach, which also requires signal reconstruction, is more sophisticated when compared to LTLSS and has been widely adopted for source separation tasks. In a more recent attempt [6], a method based on inversion of the modulation transfer function was proposed for spectral enhancement of reverberant speech. The method was shown to result in improved speaker recognition performance on artificially reverberated speech.

At the feature level, despite its simplicity, CMS [11] has been shown to be helpful, but only for small reverberation times (a.k.a. T_{60}[1]) where the length of analysis windows is comparable to that of the room impulse response (RIR) [2]. Relative spectral (RASTA) processing [13], [14], which is a modulation filtering technique in the log-spectral domain, has also been shown to be successful in removing the short-term

[1] T_{60} is defined as the time period required for the signal power to decay by 60-dB after the sound source is switched off.

effects of linear channel distortion on the speech signal. In [5], it was assumed that reverberation can be modeled as additive noise, and a spectral subtraction method was adopted to suppress the reverberation before applying CMS. A feature warping method was also applied and a significant SID accuracy improvement was obtained over their baseline system. As an alternative feature level solution, several studies have introduced reverberant-robust acoustic feature parameters with varying SID improvement levels compared to the baseline mel-frequency cepstral coefficients (MFCC). In particular, feature parameters obtained from subband Hilbert envelopes have shown promise for SID tasks under reverberant mismatched conditions [15]–[17].

At the model level, assuming that there is access to RIRs and that a rough estimate of T_{60} can be calculated, reverberation classification and acoustic model matching based on reverberant background model (RBM) have been successfully employed [7], [18]. Furthermore, in a recent study [8], multi-style training of probabilistic linear discriminant analysis (PLDA) models was adopted to bring reverberation robustness to an i-vector based speaker recognition system.

Finally, at the scoring level, similar to methods used for channel mismatch conditions, in [5] and [7] a combination of different normalization strategies were used to remove possible biases in the calculated likelihoods.

In this study, the focus is on robust front-end solutions for improved SID under reverberant mismatched conditions. Specifically, we formulate a blind spectral weighting (BSW) technique for mitigating the destructive reverberation effects on SID performance. The weights are computed using a parametric gain function which is based on *a priori* signal-to-interference ratio (SIR) estimate. A smoothed and shifted version of the reverberant power spectrum is used as an approximation for the late reverberation spectral variance. The technique is entirely blind, meaning that prior knowledge of neither the anechoic signal nor the RIR is required.

Performance of the proposed technique in mitigating the adverse reverberation impact on SID is evaluated through speaker verification experiments under simulated and actual reverberant mismatched scenarios. Evaluations are conducted in the context of the conventional GMM-UBM [19] as well as the state-of-the-art i-vector [20] based systems. The GMM-UBM experiments are performed using seven distinct actual reverberant mismatched scenarios from the MultiRoom8 corpus made available by AFRL. The i-vector experiments are carried out with microphone data (interview and phone call) from the NIST SRE 2010 extended evaluation set which are digitally convolved with three different *measured* RIRs, with T_{60} ranging from 0.48 s to 1.15 s, extracted from Aachen impulse response (AIR) database [21], [22]. We employ the proposed spectral weighting solution as a pre-processing step in the standard MFCC feature extraction framework, and evaluate its effectiveness in suppressing the late reverberation effect on SID. For the sake of comparison, we also perform the same experiments with RASTA [13], two other blind reverberation compensation strategies, namely LTLSS, [10], and Gammatone subband based NMF [12], and the recently introduced mean Hilbert envelope coefficients (MHEC) feature [16].

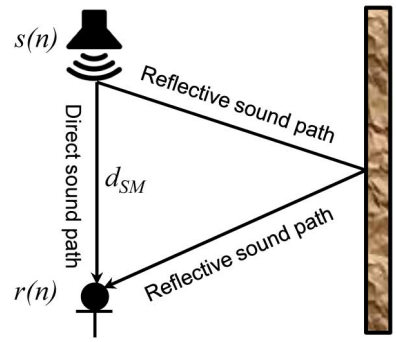


Fig. 1. Direct and reflected sound signal components. d_{SM} denotes the source-to-microphone distance.

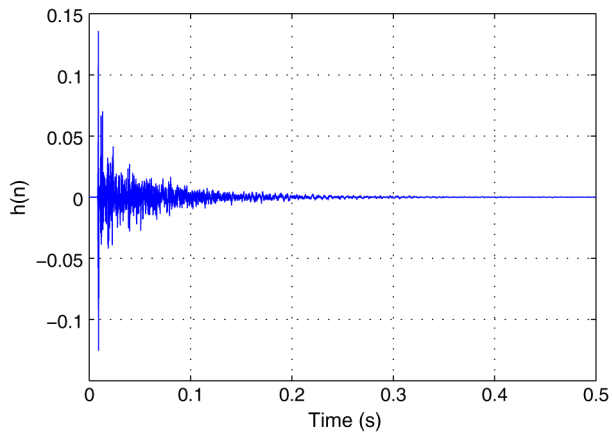


Fig. 2. Measured room impulse response for a $5.0 \times 6.4 \times 2.9m^3$ office room with $T_{60} = 0.48$ s.

II. Room Reverberation

A. Background

In a reverberant enclosure, sound waves arrive at the receiver (e.g., ears or microphone) via a direct path, and via multiple paths and directions after reflecting off walls and objects defining the acoustic enclosure. This is illustrated in Fig. 1 where the sound source $s(n)$ is captured at the microphone as a delayed sum of the direct sound and its reflections from the wall. The reflections arriving within 50-80 ms after the direct sound are called early reflections, which tend to build up to a level louder than the direct sound and cause an internal smearing effect known as the "self-masking effect". Echoes reaching the receiver after early reflections are called late reflections, which tend to smear the direct sound over time and mask succeeding sounds. This phenomenon is commonly referred to as the "overlap-masking effect", and has been shown to be the primary cause of degraded speech identification performance for both human and machine listeners [23]–[26]. The overlap-masking effect can also mask/obscure spectral details and acoustic cues essential for automatic SID, resulting in a major drop in performance [1], [5], [16].

Room reverberation can be completely characterized through the RIR with which it is possible to compute the reverberation time, T_{60}, and the direct-to-reverberant ratio (DRR), both of which are important parameters for understanding the reverberation effects on speech. An example RIR is shown in Fig. 2 for a $5.0 \times 6.4 \times 2.9m^3$ office room with $T_{60} = 0.48$ s. The sharp

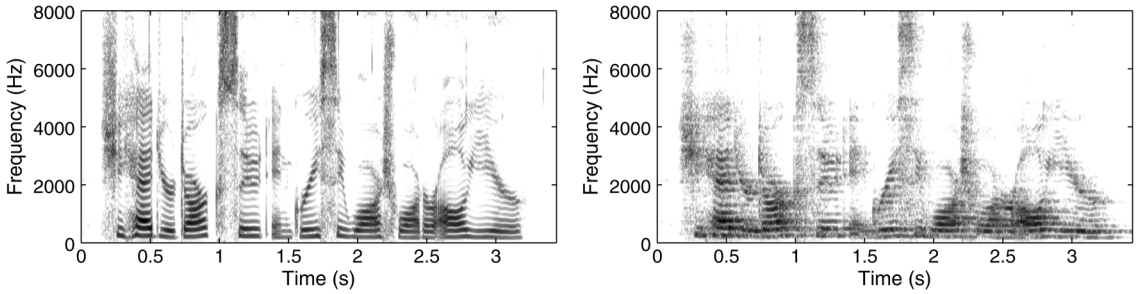


Fig. 3. Sample spectrograms for the sentence "she had your dark suit in greasy wash water all year", in anechoic quiet condition (left), and under reverberant condition with $T_{60} = 0.48$ s (right).

spike at the beginning represents the direct sound power, while the decaying portion of the $h(n)$ determines how long the attenuated reflections persist over time after the sound source is switched off.

As noted earlier, reverberation has various destructive effects on spectro-temporal characteristics of speech signals. These effects are demonstrated in Fig. 3 in which sample spectrograms are shown for the TIMIT sentence "she had your dark suit in greasy wash water all year", in anechoic (left) and reverberant conditions (right). Both self and overlap-masking effects of reverberation are evident in this figure. The self-masking effect blurs spectral details of individual phonemes and results in flattened formant transitions, while the overlap-masking effect smears the high energy phonemes (e.g., vowels) over time and fills envelope gaps which in turn increases the prominence of low-frequency energy in the speech spectrum. Taken together, these effects pose a deleterious impact on performance of automatic SID systems, especially under mismatched conditions.

B. Mathematical Model of Reverberation

In order to develop objective solutions for compensating the reverberation effects on performance of SID, it is imperative to mathematically model reverberation. As discussed earlier, in a reverberant environment, the speech signal received at the microphone is a delayed sum of a direct sound and its reflections from walls and objects in the acoustic enclosure; hence reverberation can be modeled as the convolution of the RIR with the speech signal,

$$r(n) = \sum_{j=0}^{L-1} s(n-j)h(j) = s(n) * h(n), \tag{1}$$

where $r(n)$ and $s(n)$ are the reverberant and anechoic signals, respectively, and $h(n)$ is the RIR. The RIR, $h(n)$, can be partitioned into two components as,

$$h(n) = \begin{cases} 0, & n < 0 \\ h_e(n), & 0 \le n < n_e \\ h_\ell(n), & n_e \le n < L \end{cases} \tag{2}$$

where L is the length of $h(n)$, and n_e is a time window threshold chosen such that $h_e(n)$ consists of the direct path signal and a few early reflections, while $h_\ell(n)$ consists of all the late reflections. The time threshold n_e is commonly set to a value within 50-80 ms range[2]. Late reflections that smear the speech spectra and reduce signal quality, are characterized by T_{60}. These have a long-term effect on speech signals and therefore cannot be effectively compensated for using conventional cepstral mean subtraction (CMS) within the short-term speech analysis framework [2]. On the other hand, early reflections that cause coloration distortion are characterized by the DRR which is dependent on the distance between the sound source and microphone.

Taking (2) into account, (1) can be rewritten as,

$$r(n) = \underbrace{\sum_{j=0}^{n_e-1} s(n-j)h_e(j)}_{r_e(n)} + \underbrace{\sum_{j=n_e}^{L-1} s(n-j)h_\ell(j)}_{r_\ell(n)}, \tag{3}$$

where $r_e(n)$ and $r_\ell(n)$ are referred to as the early and late reverberant speech components, respectively. Our objective is to blindly suppress the late reverberant speech using spectral weighting to improve SID performance in reverberation. This is discussed in the next section.

III. Blind Spectral Weighting (BSW) Algorithm

As discussed earlier, from a signal processing perspective, reverberation can be considered a convolutive/channel distortion. Nevertheless, in the seminal work of [24] it has been shown that the overlap-masking effect can be modeled as an uncorrelated additive interference. Hence, it can be compensated via spectral subtraction, given that an estimate of the late reverberation spectral variance is available. This has inspired several single and multichannel approaches that have considered spectral subtraction for blind late reverberation suppression [9], [28]–[30]. Because a rough estimate of the reverberation time is required to compute the late reverberation spectral variance, performance of these approaches are highly dependent on the accuracy of the T_{60} estimation.

In this study, following the uncorrelated and additive assumption for late reverberation, we introduce a spectral weighting technique to mitigate the reverberation overlap-masking effect on automatic SID performance. The weights are computed using a parametric gain function which is based on the *a priori*

[2]This choice of time boundary between early and late reflections is motivated by the temporal integration property observed with human auditory system in reverberant sound fields [27]. Due to this property, early reflections arriving within the first 50 ms time period after the direct sound are not perceived as separate sounds. Note, however, that this time threshold is dependent on characteristics of the source signal as well. For instance, the time threshold, n_e, for slowly varying music is set to 80 ms (as opposed to 50 ms for speech).

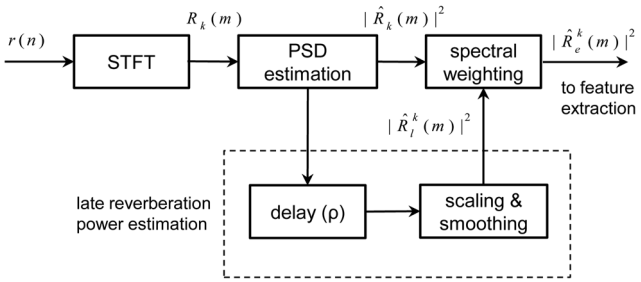


Fig. 4. Block diagram of the proposed spectral weighting technique for suppression of the late reverberation.

signal-to-interference ratio (SIR) estimate. A smoothed and shifted version of the reverberant power spectrum is used as an approximation for the late reverberation spectral variance. The technique is entirely blind, meaning that prior knowledge of neither the anechoic signal nor the RIR is required.

A. Problem Formulation

A block diagram of the proposed spectral weighting solution for late-reverberation suppression is depicted in Fig. 4. The late reverberant speech is suppressed in the short-time Fourier transform (STFT) domain by applying spectral weights as,

$$\hat{R}_e^k(m) = G_k(m) \cdot R_k(m), \tag{4}$$

with m and k being the time frame and frequency-bin indices, respectively. The spectral weights are computed using a parametric gain function defined as,

$$G_k(m) = \left(\frac{\xi_k(m)}{\xi_k(m) + \alpha} \right)^{\beta}, \tag{5}$$

where $\xi_k(m)$ denotes the *a priori* SIR, and α and β are some constant parameters. The *a priori* SIR is defined as,

$$\xi_k(m) = \frac{\lambda_{r_e}^k(m)}{\lambda_{r_\ell}^k(m)}, \tag{6}$$

where $\lambda_{r_e}^k(m) = E[|R_e^k(m)|^2]$ and $\lambda_{r_\ell}^k(m) = E[|R_\ell^k(m)|^2]$ denote spectral variances of the early and late speech components, respectively, both of which are to be estimated.

It is common practice to recursively estimate $\xi_k(m)$ via the decision-directed method [31] as,

$$\hat{\xi}_k(m) = \eta \frac{|\hat{R}_e^k(m) - 1|^2}{\hat{\lambda}_{r_\ell}^k(m-1)} + (1-\eta)\max[\gamma_k(m) - 1, 0], \tag{7}$$

where η $(0 \leq \eta \leq 1)$ is a smoothing constant that controls the trade-off between interference reduction and transient distortion introduced into the signal. The first term $\frac{|\hat{R}_e^k(m-1)|^2}{\hat{\lambda}_{r_\ell}^k(m-1)}$, represents the estimate of $\xi_k(m)$ from the previous time frame, while the second term $\max[\gamma_k(m) - 1, 0]$, is the maximum likelihood (ML) estimator for $\xi_k(m)$ and solely dependent on the current frame. The parameter $\gamma_k(m)$ is called the *a posteriori* SIR and is defined as,

$$\gamma_k(m) = \frac{E[|R_k(m)|^2]}{\lambda_{r_\ell}^k}. \tag{8}$$

The two SIRs are related via $\gamma_k(m) = \xi_k(m) + 1$. The recursive relationship in (7) provides smoothness in the estimate of $\xi_k(m)$ which consequently helps suppress the musical noise distortion. In practice, to further reduce distortions introduced by the spectral weighting, the gain function $G_k(m)$ is lower bounded by a constant gain floor G_f as,

$$G_k(m) = \max\left[G_k(m),\, G_f\right]. \tag{9}$$

The motivation behind employing a gain function in the form of (5), is twofold. First, the parametric Wiener filtering [32] has been successfully applied to a similar problem in the context of noisy speech enhancement[3]. In addition, it can be easily shown that common speech enhancement algorithms such as spectral subtraction and maximum-likelihood methods, are special cases of the parametric Wiener filtering [35]. Second, the two parameters α and β provide more degrees of freedom and control over the late reverberation suppression and speech distortion reduction. It has been shown in [36] that the excessive speech distortion introduced by speech enhancement algorithms, which typically occurs due to inappropriate selection of noise suppression parameters, can result in severe performance degradation for automatic SID systems.

It is worthwhile remarking here that although the parametric Wiener filtering is adopted in this study to suppress the late reverberation, the non-stationary "reverberation noise" cannot be compensated for using traditional speech enhancement techniques that estimate the noise power spectrum from the initial silence and update it during gaps and silence regions within words and sentences. Late reverberation suppression requires a different treatment that involves estimation of the spectral variance of late reverberation, which is addressed in the next section.

B. Late Reverberation Power Estimation

In order to estimate the two SIRs, i.e., $\gamma_k(m)$ and $\xi_k(m)$, an estimate of the late reverberation spectral variance must be available. In [24], a simple statistical model for the RIR was considered and an estimator for $\lambda_{r_\ell}^k(m)$ was derived. The estimator is dependent on T_{60}, which can be estimated directly from reverberant data, albeit at the cost of a more complex algorithm. This approach was further investigated in [28] and [29] to accommodate for estimation and reduction of additive noise. In addition, an ML approach for T_{60} estimation was proposed in [29].

Here, an alternative approach for estimation of late reverberation spectral variance is taken which obviates the need for direct T_{60} estimation. Considering the smearing effects of the late reverberation on speech, the power spectrum of the late speech component can be assumed to be a smoothed and shifted version of the reverberant speech power spectrum. It has also been

[3]While stability issues can occur in this solution [32], constrained iterative speech enhancement with intra- and extra- frame constraints have proven to be effective for noise reduction and front-end enhancement for ASR [33], [34].

proved mathematically in [24] that such assumption is valid. The spectral variance of the late speech component is thus expressed as [37],

$$\hat{\lambda}^k_{r_\ell}(m) = \mu\, w(m-\rho) * |R_k(m)|^2 \tag{10}$$

where the symbol $*$ denotes convolution in the time domain, $w(m)$ is a smoothing function, and $\rho \propto n_e$ is the time threshold between early and late components of the RIR. As noted earlier, n_e is commonly set to a value within 50-80 ms, and is independent of reverberation characteristics. The parameter μ is a scaling factor that specifies the relative strength of the late speech component.

Since RIRs have a decaying exponential shape (see Fig. 2), a right skewed smoothing function with a long tail would be a reasonable choice for $w(m)$. Therefore, as in [37], Rayleigh distribution function is adopted,

$$w(m) = \begin{cases} 0, & m \le -b \\ \frac{m+b}{b^2} \exp\left(\frac{-(m+b)^2}{2b^2}\right), & m > -b \end{cases} \tag{11}$$

where $b \propto n_e$ determines the overall spread of the smoothing function, and is set in accordance with the time threshold between early and late components of the RIR.

IV. Experiments

Performance of the proposed blind spectral weighting technique for suppression of late reverberation is evaluated in the context of speaker verification tasks with GMM-UBM [19] and i-vector based [20] SID systems. We report equal error rates (EER) as performance measures. The proposed technique is integrated into the MFCC feature extraction framework as a pre-processing stage, and performance is compared to that of the baseline system with no pre-processing. Performance comparison is also made with RASTA [13] along with two other blind reverberation suppression techniques, namely LTLSS [10], and Gammatone subband based NMF [12], as well as the MHEC front-end [16].

For GMM-UBM experiments, speech material from the MultiRoom8 corpus are utilized. The MultiRoom8 database, which is made available by AFRL, was designed to capture multi-session audio impacted by environmental contamination, i.e., background noise and room reverberation. It contains 1) a development set with a total of 100 speech recordings which are used to train background models, 2) 7 different enrollment-test conditions representing a range of distinct reverberant and noisy mismatched scenarios, and 3) an enrollment-test condition involving different communication channels which is not considered in this study. All recordings are sampled at 8 kHz. Four different rooms were used for data collection including: small, medium, large, and a conference room. The rooms are labeled as Sm, Med, Lg, and Enroll, respectively. Except for the conference room where recordings were collected using only close-talking microphones (CTM), for each environment, 6 uni- and omni-directional microphones located at a range of distinct distances from the speaker were used for speech capture. Each session was recorded at least 1 week from the previous session for each speaker. In an interview-like scenario, a total of 52 speakers were recorded, although not every speaker is present for every room/microphone configuration. The average length of the recordings is approximately 3 minutes. Here, a 1024-mixture UBM is built on the development set, and individual speaker models are adapted from the UBM using maximum *a posteriori* (MAP) estimation [19] with a relevance factor of 19.0. Table I summarizes room/microphone configurations used to design the seven different enrollment-test conditions available in the MultiRoom8 corpus. The reverberation times, T_{60}, are computed using the well-known method proposed by Schroeder [38]. As evident from the table, among different room/microphone configurations, speech recordings captured via microphone number 4 in the large room should be impacted the most by the overlap-masking effect of reverberation. Therefore, it is expected that speaker verification on trials involving Lg4 should be more challenging compared with other conditions.

TABLE I
Properties of the Room/Microphone Setups in the MultiRoom8 Corpus. d_{SM} Denotes the Source-to-Microphone Distance

Room/Mic	Dim. (m^2)	d_{SM} (m)	T_{60} (s)	DRR (dB)
Sm3	5.3×3.6	1.6	0.45	-2.31
Sm4		3.3	0.55	-3.95
Sm5		0.9	0.35	-1.95
Sm6		1.8	X†	X†
Med3	11.3×3.6	2.9	0.50	-2.80
Med5		0.9	0.30	-0.90
Lg4	14.6×12.9	8.1	1.10	-6.00
Lg5		0.9	0.40	-0.82
Enroll	X†	≤ 0.3	≈ 0	inf

† Information not provided for this setup.

For i-vector based speaker verification experiments, microphone speech data (interview and phone call) from the NIST SRE 2010 extended evaluation set are used (only male speakers are considered in this study), which corresponds to core evaluation conditions 1, 2, and 4 (CC-1, CC-2, and CC-4). These three evaluation conditions share the same 1,108 interview models, however, the test conditions are different. CC-1 comprises 1,108 interview test segments recorded using the same microphone types as the models with 346,857 impostor trials and 1,978 target trials. CC-2 consists of 3,328 interview test segments captured via different microphone types compared to the models with 1,215,586 impostor trials and 6,932 target trials. CC-4 comprises 440 conversational telephone test segments recorded over room microphone channels, with 364,308 impostor trials and 1,886 target trials. To simulate different reverberant conditions, *measured* RIR samples extracted from the AIR database are digitally convolved with the test material. Three RIRs with distinct source-to-microphone distances (d_{SM}) and with T_{60} ranging from 0.48 s to 1.15 s are used. The RIRs were measured in office and lecture rooms as well as an stairway. Further information concerning the RIRs is summarized in Table II. To learn the i-vector extractor [20], a gender dependent 1024-mixture UBM with diagonal covariance matrices is first trained using a total of 9,676 5-minute conversational telephone recordings (English only) from 951

TABLE II
PROPERTIES OF THE RIRS EXTRACTED FROM THE AIR DATABASE FOR EXPERIMENTS. d_{SM} DENOTES THE SOURCE-TO-MICROPHONE DISTANCE

Room Type	Dim. (m^3)	d_{SM} (m)	T_{60} (s)	DRR (dB)
Office	$5.0 \times 6.4 \times 2.9$	3.0	0.48	-0.89
Lecture	$10.8 \times 10.9 \times 3.15$	10.2	0.83	-5.62
Stairway	$5.2 \times 7.0 \times X^{\dagger}$	3.0	1.15	-6.03

† Information not provided for this setup.

male speakers. The data is selected from the NIST SRE 2004, 2005, 2006, as well as the Switchboard 2 (Phase III) and Switchboard Cellular (Part 1 and 2) corpora. These data corpora are available through Linguistic Data Consortium (LDC) [39] or by participating in the SRE evaluations (e.g., see [40]). The zeroth and first order Baum-Welch statistics are then computed for each recording and used to learn a 400-dimensional total variability subspace. After extracting 400-dimensional i-vectors, we use linear discriminant analysis (LDA) to reduce the dimensionality to 200. The dimensionality reduced i-vectors are then mean and length normalized. For scoring, a Gaussian probabilistic LDA (PLDA) model with full covariance residual noise term [41], [42] is learned using the i-vectors extracted from the UBM set as well as from microphone data found in NIST SRE 2005 and 2006 data releases. The Eigenvoice matrix in the PLDA model is full-rank with 200 columns. Note that all the models (i.e., the UBM, i-vector extractor, and PLDA) are trained only with the original non-reverberated data.

To perform the spectral weighting, the reverberant signals are transformed into the STFT domain using Hamming windowed frames of 25 ms duration with a 10 ms skip rate. The *a priori* SIR is estimated using the decision-directed approach (7) with a smoothing factor $\eta = 0.6$. The time threshold between early and late components of the RIRs is set to 50 ms which, considering the 10 ms skip rate, corresponds to 5 frames. The parameters of the parametric gain function (5) are set according to our preliminary experiments in [43], where closed-set SID accuracy on a randomly selected subset of TIMIT corpus was used as the criterion for parameter tuning. It was found that setting $\alpha = 2$ and $\beta = 2.5$, on average yields the best performance across the various reverberant mismatched conditions which were simulated with RIRs from the AIR database. The gain floor parameter G_f is fixed to 0.01 which is equivalent to a maximum attenuation of -20 dB. In contrast to the findings reported in [37], tuning the scaling factor μ, that specifies the relative strength of the late speech component, seems to be very important for SID tasks. Here, μ is set to 0.1, since greater values for this parameter will result in speech distortion that is intolerable for the SID system, which in turn can lead to a great drop in performance [36]. MFCC features are then extracted from the processed speech spectra. Out of 24 filterbank log-energies, the first 13 cepstral coefficients are retained after applying the DCT (including c_0), and the first and second temporal cepstral derivatives are appended to form a 39-dimensional feature vector for each frame. The MFCCs are also extracted from the unprocessed data to serve as the baseline. In order to perform non-speech frame dropping, we use time labels generated with an unsupervised speech activity detector called Combo-SAD [44]. After dropping the non-speech frames, cepstral mean and variance normalization (CMVN) is applied to remove the short-term linear channel effects.

V. RESULTS

Table III presents speaker verification results obtained from the GMM-UBM experiments on different reverberant mismatched conditions available in the MultiRoom8 corpus. The results are reported in terms of EER, with and without the proposed blind spectral weighting algorithm as the pre-processing stage for the MFCC feature extraction. In addition, speaker verification performances are shown with RASTA along with two other blind reverberation suppression techniques. Also shown in Table III (last column) are results obtained with the MHEC front-end. Several observations can be made from the results given in this table. First, speaker verification performance degrades significantly when the mismatch between training (or enrollment) and test conditions is large. For instance, as noted earlier, the range of the reverberation time and the DRR in Lg4 is totally different from the other room/microphone conditions in MultiRoom8 corpus. Therefore, speaker verification performance on trials involving the recordings collected with Lg4 is inferior by far compared to the other setups. Furthermore, as discussed earlier, as the reverberation time increases the overlap-masking effect becomes the dominant cause of speech distortion. The overlap-masking effect has been shown to be the major source of performance degradation of speech system under reverberant conditions [16], [24]. This constitutes another reason for the high error rates seen on Lg4-Med5 condition. The same argument holds for Enroll-Sm4 and Enroll-Sm6 conditions where speakers are enrolled using anechoic speech data collected with CTM in the Enroll room, while tests are performed using reverberant speech recordings.

The second observation from Table III is that RASTA technique is not as effective in suppressing the reverberation effects on speaker verification performance. It is well-established that RASTA is useful in reducing the short-term linear channel effects (i.e., telephone and microphone channels) in the log-spectral (or cepstral) domain. Nevertheless, it is clear from the results in the table that RASTA cannot effectively deal with the long-term reverberation effects for speaker verification tasks.

Third, it is clear that incorporating BSW within the MFCC feature extraction framework consistently results in significant improvements in speaker verification performance. An average absolute improvement of 3.56% is achieved over the baseline system with MFCC features extracted from unprocessed spectra.

Finally, to compare the performance of the proposed BSW technique with other blind reverberation compensation strategies, we perform the same speaker verification experiments using MFCC features extracted from speech data pre-processed with the LTLSS [10], and Gammatone subband NMF [12]. Additionally, experiments are conducted with acoustic features extracted using the MHEC front-end. It is evident from the results in Table III that the proposed technique consistently outperforms the other strategies in suppressing the reverberation effects on SID. The MHEC front-end provides significant

TABLE III
PERFORMANCE OF BLIND REVERBERATION COMPENSATION FRONT-ENDS IN TERMS OF EER (%), OBTAINED FROM SPEAKER VERIFICATION EXPERIMENTS ON MULTIROOM8 CORPUS

Enrollment-Test	EER [%]					
	MFCC	MFCC-RASTA	MFCC-BSW	MFCC-LTLSS	MFCC-NMF	MHEC
Lg5-Sm4	10.53	11.91	**5.47**	11.50	11.50	8.10
Sm4-Lg5	7.90	7.90	**5.26**	7.90	11.34	8.71
Enroll-Sm6	18.61	15.40	13.95	18.61	22.51	**13.91**
Enroll-Sm4	11.44	10.79	**5.95**	9.44	11.63	9.30
Med3-Sm3	10.50	10.32	**7.69**	12.82	10.26	8.17
Lg4-Med5	19.44	16.67	**16.67**	21.27	23.76	19.44
Med5-Sm5	6.61	6.98	**5.13**	10.26	7.69	6.14
Avg.	12.15	11.42	**8.59**	13.11	14.10	10.54

improvements over the baseline MFCCs, however, it cannot achieve the same performance level obtained with the BSW method. Note that the system performance with LTLSS and NMF under reverberant conditions is even worse than the performance with plain MFCCs. This is due to the fact that these methods introduce a great amount of processing artifacts which are intolerable for the SID system (this was confirmed through informal listening experiments). In addition to the superior performance, there is no need for signal reconstruction with the proposed technique, as required with both the LTLSS and NMF strategies.

Results for i-vector based speaker verification experiments conducted with artificially reverberated microphone data (interview and phone call) selected from the extended evaluation sets in NIST SRE 2010 are displayed in Figs. 5 and 6. Verification results for the three different room setups (i.e., office, lecture, and stairway) are shown in Fig. 5 in terms of EER. In general, the same trend in results is observed with the i-vector experiments compared to the GMM-UBM experiments. As the reverberation time increases and the overlap-masking effect becomes more dominant, the verification performance of the baseline system degrades more rapidly. For instance, the EER for the baseline system increases from 4% to 7%, which is an absolute performance degradation of 3%. The results shown here indicate that, in line with the findings in psychoacoustic studies (e.g., see [[23], [26]) and when compared to the self-masking effect, the overlap-masking effect has a greater impact on performance of SID systems. Suppressing this effect can thus alleviate its adverse impact on the performance. Here, BSW technique consistently provides significant gains in performance, especially for rooms with larger reverberation times (i.e., lecture and stairway). The MHEC front-end also provides gains in performance over the baseline system, although the gains are not as significant as those obtained with the proposed method. However, unlike the behavior observed with the GMM-UBM experiments, both RASTA and LTLSS methods only result in degraded performance, while the Gammatone subband based NMF yields moderate performance improvements. The variation in the behavior of these front-ends can be due to the change in the speaker verification paradigm, that is i-vector versus GMM-UBM framework. It was recently shown that [45] improvements/degradations due to front-end processing in the GMM-UBM paradigm may not necessarily translate into improvements/degradations in the i-vector paradigm and vice versa. This variation in behavior can also be attributed to the artificial nature of the task compared to the realistic scenarios in the MultiRoom8 corpus. More specifically, in the artificial scenario (i.e., SRE-2010 task) clean speech signal is digitally convolved with the RIR, while in the realistic scenario (i.e., MultiRoom8) speech signal is recorded using far-filed microphones and thus contains a perceivable level of background noise (e.g., AC noise). Therefore, the NMF based technique, which essentially represents a filtering operation formulated using a least-squares error criterion and is based on the convolutive assumption for the interference, performs as expected only on the artificial task. Nevertheless, the proposed BSW technique is the only reverberation suppression method that performs consistently well across the two paradigms.

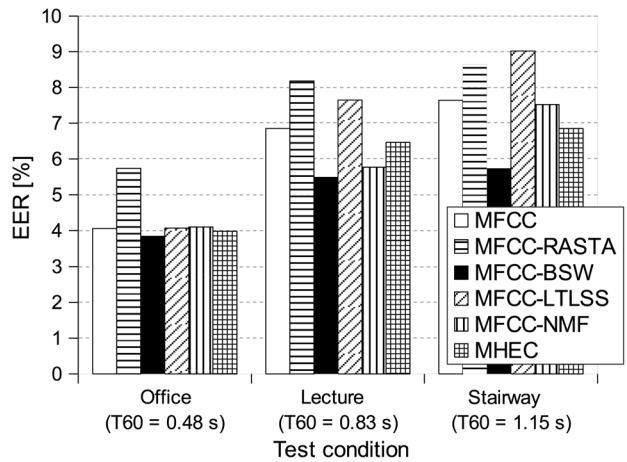


Fig. 5. Performance comparison of different blind reverberation suppression techniques on artificially reverberated microphone data (interview and phone call) from NIST SRE 2010. Results are given in terms of EER for three different test conditions corresponding to reverberation times $T_{60} = 0.48$, 0.83, and 1.15 s.

Fig. 6 presents the results from the same i-vector based experiments for the extended core conditions 1, 2, and 4 in NIST SRE 2010. The results are averaged across the three reverberant conditions (i.e., office, lecture, and stairway) for each test condition. The highest EERs are seen for CC-2, followed by CC-4 and CC-1. This is expected because, in addition to reverberation mismatch, there is also a microphone mismatch in CC-2.

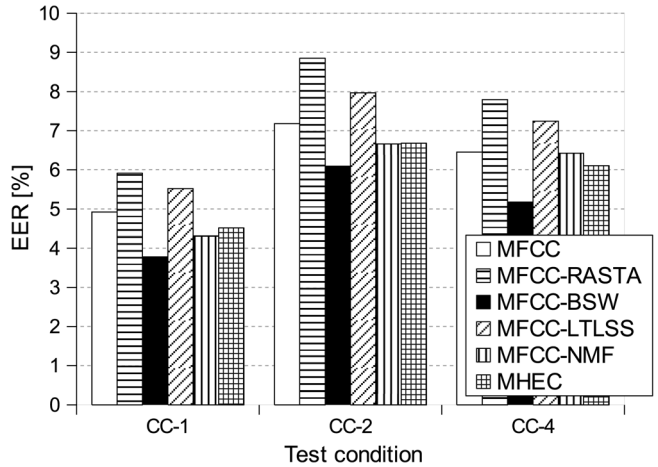


Fig. 6. Performance comparison of different blind reverberation suppression techniques on artificially reverberated microphone (interview and phone call) data from NIST SRE 2010. Results are given in terms of EER for extended core conditions (CC) 1, 2, and 4, averaged across the three reverberant conditions.

The general performance trends remain the same for all the front-ends across the three core evaluation conditions considered, and the BSW technique consistently yields the highest boosts in the performance.

VI. Conclusion

Reverberation overlap-masking effect causes severe performance degradations for both human and machine listeners. In this study we proposed a blind spectral weighting (BSW) technique for alleviating the impact of late reverberation on performance of SID systems. The technique is blind in the sense that prior knowledge of neither the anechoic signal nor the room impulse response is required. In addition, the late reverberation spectral variance was estimated without the direct need for T_{60} estimation. It was confirmed that incorporating the proposed BSW technique as a pre-processing stage in the MFCC feature extraction framework results in significant improvements in automatic SID performance under simulated (NIST SRE 2010) and actual (MultiRoom8) reverberant mismatched conditions. The performance improvements were shown to be consistent across various evaluation conditions in both GMM-UBM and i-vector speaker verification paradigms. Performance comparisons were made with RASTA along with two other blind reverberation suppression techniques, namely LTLSS and Gammatone subband based NMF. It was shown that the BSW technique consistently outperformed the other methods in improving speaker verification performance under reverberation mismatch. We believe that this technique can potentially benefit other automatic speech applications, such as automatic speech recognition (ASR), under the same mismatched conditions.

References

[1] P. Castellano, S. Sridharan, and D. Cole, "Speaker recognition in reverberant enclosures," in *Proc. IEEE ICASSP*, Atlanta, GA, USA, May 1996, vol. I, pp. 117–120.

[2] Y. Pan and A. Waibel, "The effects of room acoustics on MFCC speech parameter," in *Proc. ICSLP*, Beijing, China, Oct. 2000, pp. 129–132.

[3] P. Assmann and A. Summerfield, "The perception of speech under adverse conditions," in *Speech Processing in the Auditory System*, S. Greenberg, W. Ainsworth, A. Popper, and R. Fay, Eds. New York, NY, USA: Springer-Verlag, 2004, pp. 231–308.

[4] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez, "Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proc. ICSLP*, Philadelphia, PA, USA, Oct. 1996, pp. 1333–1336.

[5] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2023–2032, Sep. 2007.

[6] B. J. Borgström and A. McCree, "The linear prediction inverse modulation transfer function (LP-IMTF) filter for spectral enhancement, with applications to speaker recognition," in *Proc. IEEE ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4065–4068.

[7] I. Peer, B. Rafaely, and Y. Zeigel, "Reverberation matching for speaker recognition," in *Proc. IEEE ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 4829–4832.

[8] 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 ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4257–4260.

[9] L. Wang, K. Odani, and A. Kai, "Dereverberation and denoising based on generalized spectral subtraction by multi-channel LMS algorithm using a small-scale microphone array," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–11, 2012.

[10] D. Gelbart and N. Morgan, "Double the trouble: Handling noise and reverberation in far-field automatic speech recognition," in *Proc. ICSLP*, Denver, CO, USA, Sep. 2002, pp. 2185–2188.

[11] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, pp. 1304–1312, Jun. 1974.

[12] K. Kumar, R. Singh, B. Raj, and R. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 4604–4607.

[13] H. Hermansky, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 5, pp. 578–589, Oct. 1994.

[14] B. E. D. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. IEEE ICASSP*, Munich, Germany, Apr. 1997, pp. 1259–1262.

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

[16] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 5448–5451.

[17] S. Ganapathy, J. Pelecanos, and M. Omar, "Feature normalization for speaker verification in room reverberation," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 4836–4839.

[18] J. Gammal and R. Goubran, "Combating reverberation in speaker verification," in *Proc. IEEE Conf. Instrum. Meas. Technol., IMTC'05*, May 2005, pp. 687–690.

[19] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, Jan. 2000.

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

[21] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. IEEE DSP*, Santorini, Greece, Jul. 2009, pp. 1–5.

[22] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant, and P. Vary, "Do we need dereverberation for hand-held telephony?," in *Proc. Int. Congress Acoust., ICA*, Sydney, Australia, Aug. 2010, pp. 1–7.

[23] A. K. Nabelek, T. R. Letowski, and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, vol. 86, pp. 1259–1265, Oct. 1989.

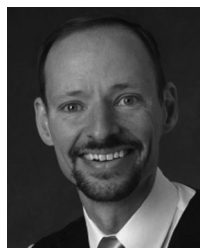
[24] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, pp. 359–366, 2001.

[25] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *J. Acoust. Soc. Am.*, vol. 133, no. 3, pp. 1607–1614, Mar. 2013.

[26] O. Hazrati, S. O. Sadjadi, P. C. Loizou, and J. H. L. Hansen, "Simultaneous suppression of noise and reverberation in cochlear implants using a ratio masking strategy," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. 3759–3765, Nov. 2013.

[27] H. Gözler and M. Kleinschmidt, "Importance of early and late reflections for automatic speech recognition in reverberant environments," in *Proc. Elektronische Sprachsignalverarbeitung*, Karlsruhe, Germany, Sep. 2003, pp. 1–8.
[28] E. A. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. IEEE ICASSP*, Philadelphia, PA, USA, Mar. 2005, vol. 4, pp. 173–176.
[29] H. Lollmann and P. Vary, "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *Proc. IEEE ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3989–3992.
[30] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 4, pp. 534–545, May 2009.
[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[32] J. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proc. IEEE*, Dec. 1979, vol. 67, pp. 1586–12–1604.
[33] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 795–805, Apr. 1991.
[34] J. H. L. Hansen and L. M. Arslan, "Markov model-based phoneme class partitioning for improved constrained iterative speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 98–104, Jan. 1995.
[35] P. C. Loziou, *Speech Enhancement: Theory and Practice*, 2nd Ed. ed. Boca Raton, FL, USA: CRC, 2013, ch. 6.
[36] S. O. Sadjadi and J. H. L. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. INTERSPEECH*, Makuhari, Japan, Sep. 2010, pp. 2138–2141.
[37] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 774–784, May 2006.
[38] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, no. 3, pp. 409–412, 1965.
[39] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: The mixer 3, 4 and 5 corpora," in *Proc. INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 950–953.
[40] "The NIST year 2010speaker recognition evaluation plan," 2010, [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/
[41] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. Comput. Vis., ICCV '07*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.
[42] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. INTERSPEECH*, Florence, Italy, Sep. 2011, pp. 249–252.
[43] S. O. Sadjadi and J. H. L. Hansen, "Blind reverberation mitigation for robust speaker identification," in *Proc. IEEE ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4225–4228.
[44] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 197–200, Mar. 2013.
[45] K. W. Godin, S. O. Sadjadi, and J. H. L. Hansen, "Impact of noise reduction and spectrum estimation on noise robust speaker identification," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 3656–3660.

Seyed Omid Sadjadi received the Ph.D. degree in Electrical Engineering from The University of Texas at Dallas in 2014, and M.S.E.E. and B.S.E.E. degrees from Amirkabir University of Technology (Tehran Polytechnic) in 2008 and 2005, respectively. Currently, he is a Research Staff Member at IBM T. J. Watson Research Center, Yorktown Heights, NY. From 2008 to 2013, he was a graduate research assistant with the Center for Robust Speech Systems (CRSS) at The University of Texas at Dallas, when this research was conducted. His research has been primarily focused on the development of robust front-end processing techniques for speech applications under adverse mismatched conditions. He was a recipient of the IBM Research Travel Grant at IEEE ICASSP-2013, Vancouver, BC, for the paper describing speaker ID systems submitted from CRSS to the NIST SRE 2012. A member of IEEE and ISCA, he has been a reviewer for IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE SIGNAL PROCESSING LETTERS, IEEE TRANSACTIONS ON FORENSICS AND INFORMATION SECURITY, and IEEE TRANSACTIONS ON MULTIMEDIA. He has authored/coauthored around 30 papers in the field of speech processing and language technology. In the summer of 2013, he developed the MSR Identity Toolbox for speaker recognition while he was a Research Intern at Microsoft, Mountain View, CA.



John H. L. Hansen (S'81–M'82–SM'93–F'07) received the Ph.D. and M.S. degrees in Electrical Engineering from Georgia Institute of Technology, Atlanta, Georgia, in 1988 and 1983, and B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, N.J. in 1982. He joined University of Texas at Dallas (UTD), Erik Jonsson School of Engineering and Computer Science in the fall of 2005, where he served as Department Head of Electrical Engineering from (2005–2012), and presently serves as Associate Dean for Research for the Erik Jonsson School of Engineering and Computer Science. He also holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech & Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Dept. of Speech, Language and Hearing Sciences (SLHS), and Professor in the Dept. of Electrical & Computer Engineering, at University of Colorado Boulder (1998–2005), where he co-founded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. In 2007, he was named IEEE Fellow for contributions in Robust Speech Recognition in Stress and Noise, and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee (2005–08; 2010–13; elected and served as TC Chair in 2011–2012, presently serving as Past-TC Chair in 2013), and Educational Technical Committee (2005–08; 2008–10). Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/06), Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–99), Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the *IEEE Signal Processing Magazine* (2001–03). He has also served as guest editor of the Oct. 1994 special issue on Robust Speech Recognition forIEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–03), and is serving as a member of the ISCA (Inter. Speech Communications Association) Advisory Council. In 2010, he was recognized as ISCA Fellow, for contributions on research for speech signals under adverse conditions. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has supervised 66 (33 Ph.D., 33 MS/MA) thesis candidates, was recipient of The 2005 University of Colorado Teacher Recognition Award as voted on by the student body, author/co-author of 505 journal and conference papers and 11 textbooks in the field of speech processing and language technology, coauthor of the textbook Discrete-Time Processing of Speech Signals, (IEEE Press, 2000), co-editor of DSP for In-Vehicle and Mobile Systems (Springer, 2004), Advances for In-Vehicle and Mobile Systems: Challenges for International Standards (Springer, 2006), In-Vehicle Corpus and Signal Processing for Driver Behavior (Springer, 2008), and lead author of the report The Impact of Speech Under Stress on Military Speech Technology, (NATO RTO-TR-10, 2000). He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, Sept. 16–20, 2002; Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX, and Co-Chair and Organizer for IEEE SLT-2014: Spoken Language Technology Workshop, Lake Tahoe, NV.