Modulation Spectral Features for Robust Far-Field Speaker Identification

Tiago H. Falk, *Student Member, IEEE*, and Wai-Yip Chan

***Abstract*—In this paper, auditory inspired modulation spectral features are used to improve automatic speaker identification (ASI) performance in the presence of room reverberation. The modulation spectral signal representation is obtained by first filtering the speech signal with a 23-channel gammatone filterbank. An eight-channel modulation filterbank is then applied to the temporal envelope of each gammatone filter output. Features are extracted from modulation frequency bands ranging from 3–15 Hz and are shown to be robust to mismatch between training and testing conditions and to increasing reverberation levels. To demonstrate the gains obtained with the proposed features, experiments are performed with clean speech, artificially generated reverberant speech, and reverberant speech recorded in a meeting room. Simulation results show that a Gaussian mixture model based ASI system, trained on the proposed features, consistently outperforms a baseline system trained on mel-frequency cepstral coefficients. For multimicrophone ASI applications, three multichannel score combination and adaptive channel selection techniques are investigated and shown to further improve ASI performance.**

***Index Terms*—Gaussian mixture model (GMM), modulation spectrum, reverberation, reverberation time, speaker identification.**

I. Introduction

TODAY, the majority of existing automatic speaker identification (ASI) systems use mel-frequency cepstral coefficients (MFCCs) as auditory inspired features and Gaussian mixture models (GMMs) or support vector machines for classification. With burgeoning hands-free communication technologies, however, the performance of such systems is shown to degrade substantially, mostly due to room acoustical effects [1] and to mismatch between training and testing conditions [2]. In order to improve far-field ASI performance, two paradigms have been explored. The first uses compensation schemes to suppress unwanted environment effects (e.g., reverberation) from the *test* speech signal in order to better match the characteristics of clean speech used to train the speaker models. The second approach, viewed as a dual of the compensation methodology, artificially distorts *training* speech data in order to better match the expected characteristics of the distorted test speech signals.

Compensation techniques can operate either at the feature or signal level, or at both the signal and feature levels. Methods that operate at the feature level attempt to reduce environment effects by modifying the extracted features. The most common techniques include cepstral mean substraction (CMS), cepstral mean substraction and variance normalization (CMSVN), and relative spectral (RASTA) filtering [3]. In turn, compensation at the signal level involves performing speech enhancement prior to feature extraction [4], [5]. With far-field ASI, room reverberation acts a major performance degrading factor and speech enhancement consists of reverberation suppression. Dereverberation, however, is a difficult and often ill-conditioned problem, particularly if only a single microphone is available. Moreover, dereverberation may introduce artifacts which can be detrimental to ASI performance. To alleviate the effects of introduced artifacts, combined feature-signal processing has been used. In [6], a microphone array beamformer is used for reverberation suppression and CMS is used to reduce introduced artifacts. Similarly, the work described in [7] uses reverberation suppression in combination with feature warping and CMS for improved far-field ASI performance.

Alternately, the works described in [8]–[11] propose to artificially distort training speech in order to emulate distortions that are expected to be present during testing. Commonly, multiple models are trained per speaker, each obtained with training data distorted by different room acoustical properties. In [8], speaker models are obtained for five different room impulse responses. During testing, a room impulse response classifier is used to determine which speaker model to use. Similarly, in [9] six models are used per speaker to represent unreverberant and five levels of reverberant speech (ranging from low to high). For testing, a "reverberation sensing system" is used to decide which speaker model to use. In [10], [11], it is assumed that some *a priori* information is known about the room in which test signals will be recorded; representative parameters include approximate room size and speaker/microphone positions. Access to such information allows training speech to be distorted with an artificially generated room impulse response which approximates that of the real test environment.

In this paper, an alternate approach to environment-robust speaker identification is presented. In particular, motivated by [12], auditory inspired modulation spectral features are proposed based on extending the work described in [13].

Manuscript received November 26, 2008; revised May 07, 2009. First published May 26, 2009; current version published October 16, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy J. Hazen.

T. H. Falk was with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada. He is now with the Bloorview Kids Rehab, Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON M4G 1R8, Canada (e-mail: tiago.falk@ieee.org).

W.-Y. Chan is with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: geoffrey.chan@queensu.ca).

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/TASL.2009.2023679

The features are shown to be robust to mismatch between training and testing reverberation conditions and insensitive to increasing reverberation levels. Mismatch conditions due to e.g., additive noise or transmission channels are not explored in this paper and are left for future study. The effectiveness of the proposed features is demonstrated with three ASI experiments which use reverberant speech generated with simulated or measured room impulse responses and reverberant speech recorded in a meeting room. Comparisons are carried out with an MFCC-GMM baseline system operating with different compensation methods (e.g., CMSVN, speech enhancement, and a combination of both). Experimental results show that a GMM ASI system based on the proposed features consistently outperforms the baseline. Moreover, three channel selection/combination techniques are explored for multimicrophone applications; experiments show that further improvement in far-field ASI performance can be attained.

The remainder of this paper is organized as follows. Section II describes models of room reverberation as well as introduces methods to artificially generate reverberant speech. Section III presents the proposed modulation spectral features and Section IV the proposed and baseline ASI systems. Section V reports experimental results and Section VI describes three channel selection and combination techniques. Lastly, conclusions are presented in Section VII.

II. Room Reverberation

In this section, models of room reverberation and methods to generate reverberant speech are discussed.

A. Models of Room Reverberation

Speech propagation from a speaker to the microphone in a reverberant room is conventionally modeled as a linear filtering process. The reverberant signal $s(n)$ is modeled as a convolution of the source (clean) speech signal $x(n)$ with the room impulse response $r(n)$

$$s(n) = x(n) * r(n). \tag{1}$$

It is known that under the diffuse sound field assumption, the ensemble average of the squared room impulse response decays exponentially with time [14]

$$\langle r^2(n) \rangle = A \exp(-kn). \tag{2}$$

The angled brackets $\langle \cdot \rangle$ denote the ensemble average, A is a gain term, and k is the damping factor given by

$$k = \frac{(\ln 10^6)}{(F_s \times T_{60})} \tag{3}$$

where F_s is the sampling frequency and T_{60} is the so-called reverberation time, the parameter most widely used to characterize room acoustics. By definition, reverberation time is the time required for the sound energy to decay by 60 dB after the sound source has been turned off [15]. Commonly, the Schroeder integral is used to calculate T_{60} from the room impulse response [16].

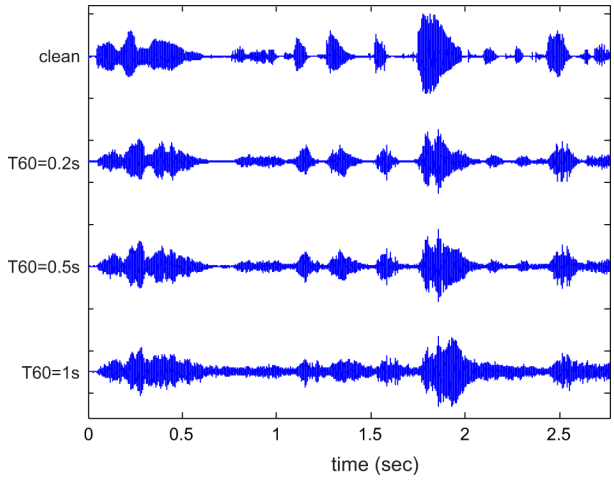

Fig. 1. Waveforms, top to bottom: clean and reverberant speech with $T_{60} = 0.2, 0.5,$ and 1 s.

B. Simulated Reverberant Speech

In our experiments, two tools are used to artificially generate reverberant speech: SIREAC (SImulation of REal ACoustics) [17] and the ITU-T software package described in Recommendation G.191 [18]. The SIREAC tool convolves the source speech signal with artificially generated (office) room impulse responses. The user has the freedom to vary T_{60}, thus simulating office environments of different sizes and dimensions. The waveforms depicted in Fig. 1 exemplify reverberant speech signals produced by the SIREAC simulation tool for $T_{60} = 0.2, 0.5,$ and 1 s. In our experiments, reverberant speech signal levels are normalized to -26 dBov (dB overload) using the ITU-T P.56 voltmeter [19]. In turn, the ITU-T G.191 tool is used to convolve room impulse responses measured from an office environment with clean speech signals. The measured room impulse responses used in our experiments are described in [20] and were collected with a six-channel microphone array and corresponded to $T_{60} \approx 0.5$ s. Microphones were omnidirectional and spaced 5 cm apart in a linear array. The speaker was placed at a 90° angle with respect to the center of the array at a distance of 94 cm.

III. Auditory Inspired Modulation Spectral Features

In this section, a brief description of the proposed modulation spectral features are described (reader is referred to [21] for a more detailed description) and the motivation for their use in far-field speaker identification is presented.

A. Feature Extraction

The proposed modulation spectral features are computed using the signal processing steps depicted in Fig. 2. First, the speech signal $s(n)$ is filtered by a bank of 23 critical-band gammatone filters to emulate the processing performed by the cochlea [22]. Filter center frequencies range from 125 Hz to nearly half the sampling rate (e.g., 3567 Hz for 8-kHz sampling rate). Filter bandwidths are characterized by the

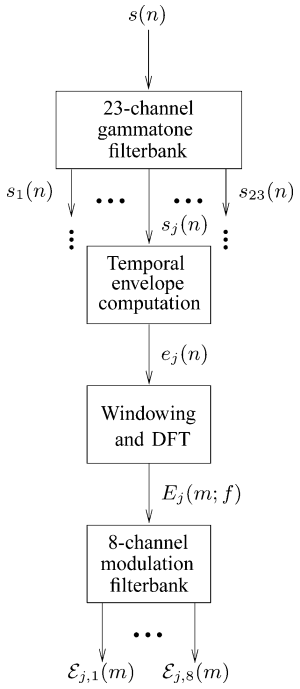

Fig. 2. Block diagram of the signal processing steps involved in the computation of the modulation spectral features.

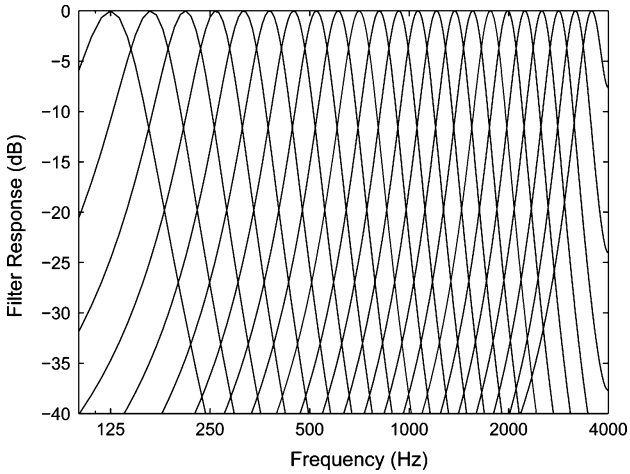

Fig. 3. Filter responses for the 23-channel gammatone filterbank.

equivalent rectangular bandwidth (ERB) [23]. The ERB for filter j, $j = 1, \ldots, 23$, is given by

$$\mathrm{ERB}_j = \frac{f_j}{Q_{\mathrm{ear}}} + B_{\min} \tag{4}$$

where f_j represents the center frequency for the filter and Q_{ear} and $B_{\min}$ are constants set to 9.265 and 24.7, respectively. Fig. 3 illustrates the frequency response of the 23-channel gammatone filterbank used in our experiments.

The output signal of the jth gammatone filter is given by

$$s_j(n) = s(n) * h_j(n) \tag{5}$$

TABLE I
MODULATION FILTER CENTER FREQUENCIES (f_c) AND BANDWIDTHS (BW) EXPRESSED IN HZ

	Modulation Frequency Band Index							
	1	2	3	4	5	6	7	8
f_c	4.0	6.5	10.7	17.6	28.9	47.5	78.1	128.0
BW	1.9	3.4	5.9	9.8	15.9	26.4	43.2	70.8

where $h_j(n)$ is the impulse response of the filter. The temporal envelope of $s_j(n)$ is computed using the Hilbert transform $\mathcal{H}\{\cdot\}$. Temporal envelopes $e_j(n)$ are computed as the magnitude of the complex analytic signal $\tilde{s}_j(n) = s_j(n) + \jmath\mathcal{H}\{s_j(n)\}$. Hence,

$$e_j(n) = \sqrt{s_j(n)^2 + \mathcal{H}\{s_j(n)\}^2}. \tag{6}$$

Temporal envelopes $e_j(n)$ are then multiplied by a 256 ms Hamming window with 32-ms shifts; the windowed envelope for frame m is represented as $e_j(m)$, where the time variable n is dropped for convenience. Frames of 256-ms duration are used in order to obtain appropriate resolution for low-frequency modulation frequencies.

The modulation spectrum for critical band j is obtained by taking the discrete Fourier transform $\mathcal{F}\{\cdot\}$ of the temporal envelope $e_j(m)$

$$E_j(m;f) = |\mathcal{F}(e_j(m))| \tag{7}$$

where f denotes modulation frequency. Modulation frequency bins are grouped into eight bands in order to emulate an auditory-inspired modulation filterbank [24]. The center frequencies and bandwidths of the eight modulation filters used in our experiments are described in Table I; the motivation to discard frequencies below 3 Hz is discussed in Section III-B. Henceforth, the notation $\mathcal{E}_{j,k}(m)$ and $\bar{\mathcal{E}}_{j,k}$ will be used to denote the *per-frame* and *average* (over all frames) modulation energy of the jth critical-band signal grouped by the kth modulation filter. Additionally, the notation $\vec{\mathcal{E}}_k(m)$ and $\vec{\bar{\mathcal{E}}}_k$ will be used to denote the per-frame and average 23-dimensional energy vector for modulation channel k, respectively.

B. Feature Selection for Environment-Robust ASI

Previous research has shown that temporal envelopes of clean unreverberated speech contain dominant frequencies (termed *modulation frequencies*) ranging from 2–16 Hz [25], [26] with spectral peaks at approximately 4 Hz, corresponding to the syllabic rate of spoken speech [27]. With reverberant speech, the room impulse response reverberation tail is often modeled as an exponentially damped Gaussian white noise process [28]. As such, it is expected that reverberant signals attain more Gaussian white-noise like properties with increasing T_{60}. Since temporal envelopes computed using (6) can contain frequencies up to the bandwidth of its originating signal [29], reverberant signals are

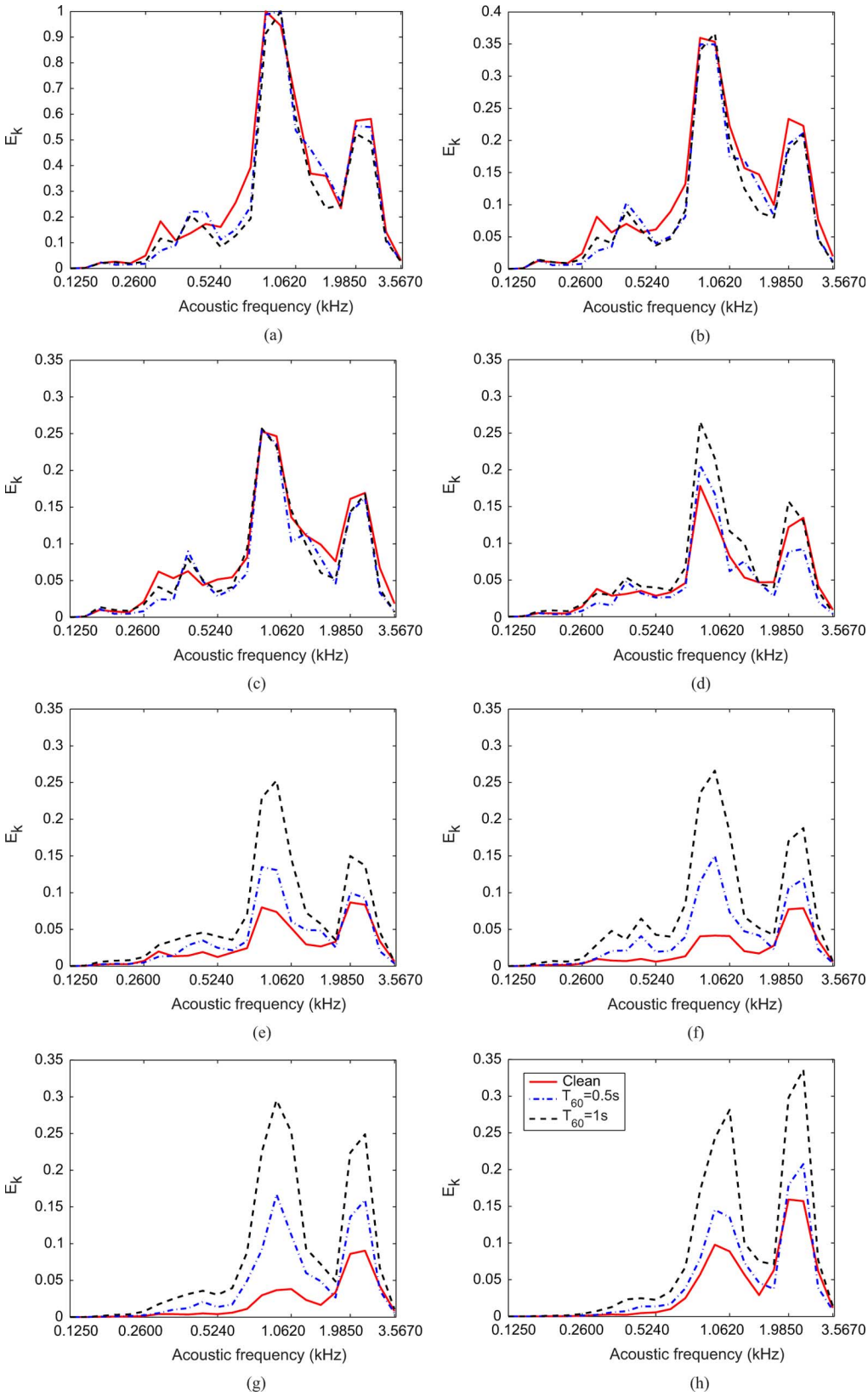

Fig. 4. Plots of $\vec{\mathcal{E}}_k$ versus acoustic frequency for modulation frequency band $k = 1, \ldots, 8$ (subplots (a)–(h) respectively), for clean speech (solid) and reverberant speech with $T_{60} = 0.5$ s (dash-dotted) and $T_{60} = 1$ s (dashed).

expected to contain significant modulation frequency components beyond the 2–16 Hz range of clean speech.

The plots in Fig. 4(a)–4(h) assist in illustrating the effects of T_{60} on $\vec{\mathcal{E}}_k$, for $k = 1, \ldots, 8$, respectively. In the plots, modulation energy values are normalized by the maximum modula-

tion energy obtained for modulation band $k = 1$, corresponding to the syllabic rate of spoken speech. As can be seen from the figure, $\vec{\bar{\mathcal{E}}}_k$ is shown to be robust to increasing T_{60} for modulation bands $k = 1 - 3$, which correspond to modulation frequencies ranging from 3–15 Hz. Moreover, as expected, subplots (d)–(h) show that the modulation energy at higher modulation frequency bands $(k = 4 - 8)$ increases with increasing T_{60}; such bands correspond to modulation frequencies greater than 16 Hz. Hence, in order to devise an environment-robust ASI system, we propose to use $\vec{\mathcal{E}}_k(m),\ k = 1 - 3$, as features. As will be shown in Section VI, information from higher modulation frequency bands $(k = 4 - 8)$ can be used to assist in multichannel score combination.

Moreover, as seen from Table I, modulation frequencies below 3 Hz are not considered. The motivation for discarding such frequencies is twofold. First, reverberation causes temporal smearing since gaps between speech bursts are filled with reverberant energy from adjacent phonemes [26]. Our pilot experiments have shown that temporal smearing causes an increase in modulation energy at low modulation frequencies ranging from DC to approximately 3.2 Hz. Second, with far-field applications, speech is often recorded in noisy environments where (quasi-)stationary noise sources are present (see Section V-D). Our experiments have shown that common office and meeting room noise sources (e.g., fan or air conditioner noise) have dominant modulation frequencies below 3 Hz. Hence, by discarding modulation frequencies below 3 Hz, increased robustness can be attained for far-field ASI. The proposed GMM-based ASI system is described in more detail in Section IV-B.

IV. ASI System Description

In this section, the baseline and proposed systems are described.

A. Baseline System

The widely used GMM based speaker identification system is used as the baseline [30]. A GMM consists of a weighted sum of M component densities

$$p(\mathbf{x}|\mathbf{\Lambda}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{x}) \tag{8}$$

where $\alpha_i,\ i = 1, \ldots, M$ are the mixture weights, with $\alpha_i \geq 0$ and $\sum_{i=1}^{M} \alpha_i = 1$, and $b_i(\mathbf{x})$ are Gaussian densities with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{\Sigma}_i$. The parameter list, $\mathbf{\Lambda} = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M\}$, defines a particular GMM, where $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \mathbf{\Sigma}_i, \alpha_i\}$ are obtained from training data using the expectation-maximization (EM) algorithm [31].

For the baseline, a conventional mel-frequency cepstral coefficient (MFCC)-based system, similar to the ones used in [7], [32], is explored. Feature vectors consist of 12th-order MFCCs appended with 12th-order delta MFCCs. In pilot experiments, it was observed that the inclusion of double-delta coefficients reduced identification accuracy for higher T_{60}. MFCCs are derived from a 26-channel mel-scale filterbank and the zeroth order coefficient (log-energy) is kept to form a 25-dimensional feature vector. Coefficients are computed from 25-ms frames with 10-ms shifts and only informative active speech frames are kept. GMMs with 32 and 64 diagonal components, per speaker, are investigated. Clean speech is used to train the baseline speaker models; different models are obtained for different feature level compensation strategies (e.g., CMS or CMSVN).

ASI is based on the average log-likelihood (LL_s) measure computed for N active speech frames

$$LL_s = \frac{1}{N} \sum_{m=1}^{N} \log(p(\mathbf{x}(m)|\mathbf{\Lambda}_s)) \tag{9}$$

where $\mathbf{x}$ denotes the 25-dimensional MFCC feature vector and $\mathbf{\Lambda}_s$ the GMM parameters obtained for speaker s. Given a group of N_S speakers, the identified speaker $\hat{S}$ is obtained using the following log-likelihood test

$$\hat{S} = \underset{1 \leq s \leq N_S}{\operatorname{argmax}}\, LL_s. \tag{10}$$

B. Proposed System

For the proposed ASI system, one GMM is trained per speaker for each of the first three modulation frequency bands $(k = 1 - 3)$. In our experiments, each model comprises 32 diagonal Gaussian components. Unless stated otherwise, clean speech is used for training of the system and compensation is *not* employed in order to demonstrate the robustness of the proposed system to mismatch between training and testing conditions. Identification is performed based on the average log-likelihood value computed for modulation frequency band k over N' active speech frames

$$LL_{k,s} = \frac{1}{N'} \sum_{m=1}^{N'} \log(p_k(\vec{\mathcal{E}}_k(m)|\mathbf{\Lambda}_{k,s})),\ k = 1 - 3 \tag{11}$$

where p_k represents the GMM for band k and $\mathbf{\Lambda}_{k,s}$ the per-band GMM parameters for speaker s. For the proposed system, the following log-likelihood test is used

$$\hat{S} = \underset{1 \leq s \leq N_S}{\operatorname{argmax}} \max_{1 \leq k \leq 3} LL_{k,s}. \tag{12}$$

V. Experimental Results

In this section, the performance of the proposed and baseline systems is reported for experiments involving reverberant speech generated with simulated and measured room impulse responses as well as recorded reverberant speech.

A. Performance Figures

In the subsections to follow, (percentage) identification accuracy (ACC) is used to quantify system performance. Moreover, two measures are used to quantify the *improvement* in performance attained with the proposed system over the baseline; namely, percentage increase (INC) and percentage error rate reduction (ERR). The measures are given by

$$\mathrm{INC}(\%) = 100 \times \frac{Y - X}{X} \tag{13}$$

$$\mathrm{ERR}(\%) = 100 \times \frac{Y - X}{100 - X} \tag{14}$$

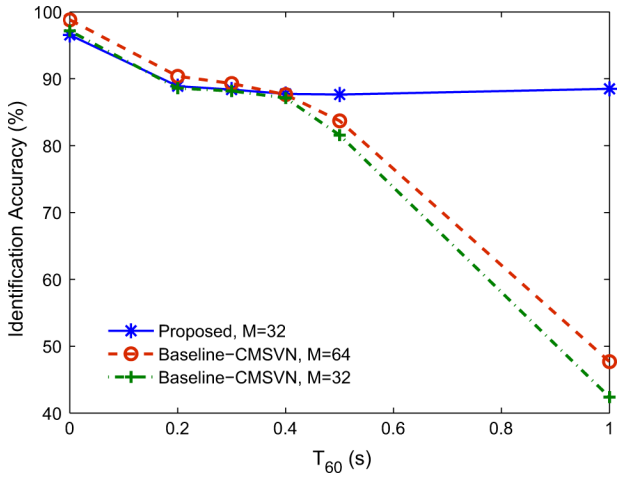

Fig. 5. Identification accuracy versus T_{60} for the proposed method (solid) with $M = 32$, and the baseline with CMSVN compensation for $M = 32$ (dash-dotted) and $M = 64$ (dashed).

where X and Y denote the identification accuracy obtained with the baseline and the proposed system, respectively.

B. Experiment 1: Simulated Room Impulse Responses

For the first experiment, the SIREAC tool is used to create reverberant speech with T_{60} values ranging from 0.2–0.5 s (in 0.1-s increments) and 1 s. The values are chosen as to simulate small, medium, and large office/meeting rooms. In this experiment, reverberant speech is generated by corrupting a subset of the TIMIT database; note that the TIMIT database does not include session variability. TIMIT speech files are downsampled to 8 kHz and utterances from 340 of the 630 speakers are used. Of the ten available utterances per speaker, eight are used to train the speaker models and two are kept for testing. Such separation results in 680 test speech signals for each of the five aforementioned T_{60} conditions, in addition to 680 clean speech signals ($T_{60} = 0$ s). For the baseline CMS, CMSVN, and RASTA filtering are tested as feature level compensation methods.

Plots in Fig. 5 depict identification accuracy versus T_{60} for the proposed system (without compensation) and for the baseline system with CMSVN compensation, as it resulted in superior performance; speaker models for both systems were trained using clean speech. As observed, baseline performance degrades almost linearly for $T_{60} > 0.4$ s. The performance of the proposed system, on the other hand, is shown to be fairly insensitive to increasing T_{60}. Moreover, baseline performance is shown to be slightly superior to that of the proposed system for lower T_{60} values (≤ 0.3 s) for speaker models with $M = 64$. One disadvantage of the proposed system is that longer windows and window shifts are needed relative to the baseline; this amounts to roughly three times less feature vectors available for training. As a consequence, the use of more complex speaker models (larger M) is not feasible with short duration training data, as is the case with the TIMIT database. Strategies, such as maximum *a posteriori* (MAP) adaptation for GMM training, offer room for improvement and are left for a future study.

Table II reports identification accuracy and performance improvements attained with the proposed system over the baseline.

TABLE II
PERFORMANCE COMPARISON OF PROPOSED SYSTEM ($M = 32$) AND THE BASELINE WITH CMSVN COMPENSATION. AVERAGE IMPROVEMENT IS COMPUTED OVER THE FIVE REVERBERATION CONDITIONS

T_{60} (s)	Proposed	Baseline, $M = 32$			Baseline, $M = 64$		
	ACC	ACC	INC	ERR	ACC	INC	ERR
0	96.6	97.2	-0.6	-21.1	98.8	-2.3	-191.5
0.2	88.9	88.6	0.3	2.6	90.4	-1.6	-15.5
0.3	88.4	88.2	0.2	1.7	89.3	-1.0	-8.5
0.4	87.7	87.1	0.7	4.7	87.6	0.1	0.9
0.5	87.6	81.6	7.5	33.0	83.7	4.7	24.1
1	88.5	42.4	108.8	80.0	47.7	85.7	78.0
Average	–	–	23.5	24.4	–	17.6	15.8

As can be seen, the proposed system is shown to attain comparable results with the baseline ($M = 32$) for clean speech (represented by $T_{60} = 0$ s in the table and in Fig. 5); somewhat lower performance is attained relative to the more complex baseline ($M = 64$). Nonetheless, despite the use of less complex speaker models and the absence of feature level compensation, the proposed system is shown to improve over the baseline ($M = 64$) by an average 17.6% INC and 15.8% ERR for $0.2 \leq T_{60} \leq 1$ and by as much as 85.7% INC for $T_{60} = 1$ s.

Moreover, a slight improvement in identification accuracy is observed with the proposed system for $T_{60} = 1$ s relative to 0.3 s $\leq T_{60} \leq 0.5$ s; in contrast, baseline system performance degrades monotonically for $T_{60} > 0.4$ s. Although counter intuitive at first, this improved performance at higher T_{60} values can be explained by insights presented in [33]. As mentioned in Section III-B, reverberation causes low amplitude speech segments and silence intervals to be filled with energy smeared from preceding phonemes [25]. As T_{60} increases, temporal smearing causes the modulation energy at lower modulation frequencies (< 3 Hz) to be amplified [34]. Additionally, as T_{60} increases, due to the Gaussian white-noise like properties of the reverberation tail, modulation energy at increasingly higher modulation frequencies ($\gg 16$ Hz) are also amplified (see Fig. 4). Hence, for large T_{60}, features extracted from modulation frequencies between 3–15 Hz are less affected by room reverberation and result in somewhat improved ASI performance. It is emphasized that this behavior has been observed for T_{60} values up to 2 s [33]. Such highly reverberant scenarios, however, are not included in our experiments due to limitations of the SIREAC tool.

C. Experiment 2: Measured Room Impulse Responses

For the second experiment, the ITU-T G.191 tool is used to convolve the six-channel measured room impulse responses ($T_{60} \approx 0.5$ s) described in Section II-B with clean speech. For this experiment the CHAINS (CHAracterizing INdividual Speakers) clean speech corpus is used [35]. The corpus contains speech files from 36 speakers, recorded in two sessions (two months apart) using different microphones [36]. Each speaker reads four short fables and utters 33 short sentences; the latter

TABLE III
PERFORMANCE COMPARISON OF PROPOSED SYSTEM ($M = 32$) AND THE BASELINE WITH CMSVN COMPENSATION. AVERAGE PERFORMANCE AND IMPROVEMENTS ARE COMPUTED OVER THE SIX-CHANNEL RESULTS

Condition	Proposed	Baseline, $M = 32$			Baseline, $M = 64$		
	ACC	ACC	INC	ERR	ACC	INC	ERR
Clean	96.8	96.7	0.1	3.0	98.6	-1.8	-128.6
Channel 1	84.6	56.9	48.7	64.3	60	41.0	61.5
Channel 2	82.8	55.8	48.4	61.1	58.9	40.6	58.2
Channel 3	83.2	55.6	49.6	62.2	59.1	40.8	58.9
Channel 4	83.4	54.1	54.2	63.8	58.1	43.5	60.4
Channel 5	81.5	56.6	44.0	57.4	60.8	34.0	52.8
Channel 6	80.7	54.3	48.6	57.8	59.1	36.5	52.8
Average	82.7	55.6	48.9	61.1	59.3	39.4	57.4

TABLE IV
BASELINE PERFORMANCE AFTER DEREVERBERATION AND CMSVN. PERFORMANCE IMPROVEMENTS ARE COMPUTED USING THE AVERAGE PROPOSED-SYSTEM PERFORMANCE REPORTED IN TABLE III.

Baseline	DSB			Cepstrum			Subspace		
	ACC	INC	ERR	ACC	INC	ERR	ACC	INC	ERR
$M = 32$	63.0	31.3	53.2	46.0	79.8	68.0	61.9	33.6	54.6
$M = 64$	67.4	22.7	46.9	51.6	60.3	64.3	65.6	26.1	49.7

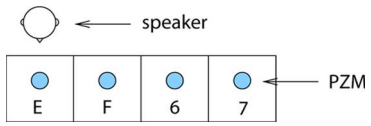

Fig. 6. PZM microphone setup at the ICSI meeting room.

are sentences mostly taken from the TIMIT database to provide a balanced phonetic coverage. The corpus is downsampled to 8 kHz and seven speech files (four fables plus three uttered sentences) are kept for training and the remaining 30 are left for testing. This amounts to a total of 1080 six-channel reverberant and 1080 single-channel clean test speech files.

For the baseline we experiment with three signal level speech enhancement schemes in combination with CMSVN. The selected multichannel dereverberation algorithms are the ones that showed superior performance in the automatic *speech* recognition test described in [20]. The dereverberation algorithms include a delay-and-sum beamformer (DSB), the multichannel cepstrum based algorithm described in [37], and a frequency domain subspace-based algorithm [38]; more detail regarding the speech enhancement algorithms can be found in [20].

Table III shows performance figures attained with the systems for the multichannel reverberant speech signals. In the table, baseline results are reported with CMSVN compensation as it resulted in superior performance. As can be seen, both systems perform comparably for clean speech with the more complex baseline ($M = 64$) attaining slightly improved performance. For reverberant speech, the proposed system is shown to improve over the baseline ($M = 32$) by an average 48.9% INC and 61.1% ERR and the more complex baseline by an average 39.4% INC and 57.4% ERR. Additionally, Table IV shows the performance attained with the baseline once dereverberation (combined with CMSVN) is performed. While delay-and-sum beamforming and subspace-based dereverberation methods are shown to improve baseline performance, cepstrum-based dereverberation decreases identification accuracy. In the table, improvement measures are computed using the average performance of the proposed system reported in Table III ($\mathrm{ACC} = 82.7\%$).

Experiments with the reverberation-suppressed speech signals and the proposed ASI system are also carried out in order to investigate the effects of dereverberation on modulation spectral features. It is observed that dereverberation does not significantly affect system performance. As an example, the proposed system attains 82.8% ACC with reverberant signals processed by the delay-and-sum beamformer. These findings corroborate those reported in [39], where it is shown that multichannel dereverberation algorithms mostly suppress modulation energy content at modulation frequencies greater than 20 Hz, thus causing only subtle changes to the proposed features. Overall, the proposed system, *without* compensation, improves over the baseline ($M = 32$) with delay-and-sum beamforming *and* CMSVN compensation by 31.3% INC and 53.2% ERR; for $M = 64$ the improvements are 22.7% INC and 46.9% ERR.

D. Experiment 3: Reverberant Speech Recordings

The third experiment makes use of the publicly available subset of the International Computer Science Institute (ICSI) Meeting Corpus [40] (the full dataset is licensed by the LDC, Linguist Data Consortium). The corpus contains multichannel noisy and reverberant speech ($T_{60} \approx 0.3$ s) recordings of digit strings read by meeting participants at the beginning and the end of 22 meetings at ICSI. Four omnidirectional pressure zone microphones (PZM) are used and arranged in a staggered line along the center of the conference table, as depicted in Fig. 6. Meetings involved anywhere from three to ten participants (averaging six) with levels of English language fluency ranging from fluent to "hard-to-transcribe." Noise sources include low-level hum of meeting room lights and fans (particularly for microphones numbered 6 and 7), as well as noise from nearby elevators, hallway conversations, and laughter from other meeting participants. Speech files range from 17–35 s and approximately 80% of the available speech data for each speaker was used for training and the remaining 20% was left for testing. Care was exercised to assure that training and test data came from separate meetings.

In this experiment we investigate the effects of training and testing condition mismatch by designing speaker models with data recorded from one of the four microphones and testing with data recorded from the remaining three microphones. During some of the meetings, the microphones also captured intrusive speech-like noises and speech from competing speakers. These noise sources were found to significantly degrade ASI performance of both the baseline and proposed systems, causing

TABLE V
MATCHED AND MISMATCHED ACCURACY AND PERFORMANCE IMPROVEMENTS ATTAINED WITH THE PROPOSED SYSTEM ($M = 32$) OVER THE BASELINE ($M = 64$) FOR THE ICSI MEETING CORPUS

Channel	Matched				Mismatched			
	Baseline	Proposed	INC	ERR	Baseline	Proposed	INC	ERR
E	82.1	95.7	16.6	76.0	78.2	86.9	11.1	39.9
F	81.2	92.8	14.3	61.7	75.1	89.9	19.7	59.4
6	79.6	95.7	20.2	78.9	76.7	85.5	11.5	37.8
7	78.7	91.3	16.0	59.2	75.8	84.1	10.9	34.3
Average	–	–	16.8	68.9	–	–	13.3	42.9

-17% and -6% INC, respectively. As a countermeasure, the noise suppression module of the enhanced variable rate codec [41] is used to reduce nonstationary speech-like noises. Unlike previous experiments, enhanced speech signals are used to test *both* the proposed and the baseline systems. For the baseline system, however, an additional feature level compensation method (CMSVN) is used to further improve performance.

Table V reports identification accuracy for matched and mismatched train-test conditions. Matched conditions indicate that speaker models were trained and tested from signals captured by the same microphone. Mismatched performance is the average over the three remaining train-test combinations. The results reported in the table are for a baseline system with $M = 64$ and the proposed system with $M = 32$. As can be seen, average improvements of 16.8% INC and 68.9% ERR can be attained by using the proposed system in matched conditions. For mismatched conditions, in turn, improvements of 13.3% INC and 42.9% ERR are obtained.

VI. MULTIMICROPHONE CHANNEL SELECTION AND SCORE COMBINATION

Far-field speech applications commonly involve the use of multiple microphones and multichannel speech enhancement techniques to improve ASI performance. Notwithstanding, as shown in Section V-C, enhancement techniques may introduce artifacts that can further degrade ASI performance. It is known that the room impulse responses measured at each microphone change continuously due to, e.g., moving speakers, room temperature, and furniture placement. As a consequence, different microphones capture signals of varied quality or reliability. In [7], multichannel score combination is shown to also improve ASI performance in far-field applications. In this section, conventional channel selection and multichannel score combination techniques are explored in Sections VI-A–VI-B, respectively, as alternative methods to improve ASI performance. A novel objective quality based score combination technique is also proposed in Section VI-C.

A. Maximum Log-Likelihood Based Channel Selection

The goal in adaptive channel selection is to select the microphone which is deemed to capture the most "reliable" speech signal. In this section, the most reliable signal is assumed to be the one that results in the maximum log-likelihood value. For the baseline, let the average log-likelihood value, computed *per microphone*, be denoted by

$$LL_{s,c} = \frac{1}{N}\sum_{m=1}^{N} \log(p(\mathbf{x}_c(m)|\mathbf{\Lambda}_s)), c = 1, \ldots, C \tag{15}$$

where $\mathbf{x}_c$ denotes the 25-dimensional MFCC feature vector computed from the speech signal captured by microphone c and C denotes the total number of microphones in the array. Hence, using the maximum log-likelihood criterion for selection, ASI is performed using the following rule:

$$\hat{S} = \underset{1\le s\le N_S}{\operatorname{argmax}}\, LL_s^* \tag{16}$$

where

$$LL_s^* = \max_{1\le c\le C}\{LL_{s,c}\}. \tag{17}$$

Similarly, for the proposed system let the per-microphone average log-likelihood value be denoted as

$$LL_{k,s,c} = \frac{1}{N'}\sum_{m=1}^{N'} \log(p_k(\vec{\mathcal{E}}_{k,c}(m)|\mathbf{\Lambda}_{k,s})) \tag{18}$$

where k indexes the modulation frequency band $(k = 1-3)$, c indexes the microphone $(c = 1, \ldots, C)$, and $\vec{\mathcal{E}}_{k,c}(m)$ represents the proposed modulation spectral features computed for the speech signal captured by microphone c. Speaker identification is performed based on the following rule:

$$\hat{S} = \underset{1\le s\le N_S}{\operatorname{argmax}} \max_{1\le k\le 3} LL_{k,s}^* \tag{19}$$

where

$$LL_{k,s}^* = \max_{1\le c\le C}\{LL_{k,s,c}\}. \tag{20}$$

B. Mean-Score Multichannel Combination

With multichannel score combination, log-likelihood values (i.e., "scores") computed from speech signals captured by multiple microphones are judiciously combined and used for ASI. As opposed to adaptive channel selection, where information from only one microphone is used, multichannel score combination uses information from all available microphones. With mean-score multichannel combination, baseline identification is performed using (16) where

$$LL_s^* = \frac{1}{C}\sum_{c=1}^{C} LL_{s,c}. \tag{21}$$

Similarly, for the proposed system identification is performed using (19) where

$$LL_{k,s}^* = \frac{1}{C}\sum_{c=1}^{C} LL_{k,s,c}. \tag{22}$$

TABLE VI
PERFORMANCE COMPARISON OF ADAPTIVE CHANNEL SELECTION AND MULTICHANNEL SCORE COMBINATION TECHNIQUES. COLUMNS LABELED "MAX," "MEAN," AND "WEIGHTED" CORRESPOND TO MAXIMUM LOG-LIKELIHOOD BASED CHANNEL SELECTION, MEAN-, AND WEIGHTED-SCORE-BASED CHANNEL COMBINATION, RESPECTIVELY

System	Max			Mean			Weight		
	ACC	INC	ERR	ACC	INC	ERR	ACC	INC	ERR
Proposed, $M = 32$	85.7	–	–	87.0	–	–	87.2	–	–
Baseline, $M = 32$	57.1	50.1	66.7	63.7	36.6	64.2	63.8	36.7	64.6
Baseline, $M = 64$	59.8	43.3	64.4	68.9	26.3	58.2	69.0	26.4	58.7

TABLE VII
PERFORMANCE COMPARISON OF WEIGHTED-SCORE MULTICHANNEL COMBINATION WITH AVERAGE SINGLE-CHANNEL PERFORMANCE AND WITH MULTICHANNEL SIGNAL-BASED DEREVERBERATION. COMPARISONS ARE FOR BOTH THE PROPOSED SYSTEM AND FOR THE BASELINE WITH $M = 64$

Metric	Proposed, $M = 32$		Baseline, $M = 64$				
	Weight	Average	Weight	Average	DSB	Cepstrum	Subspace
ACC	87.2	82.7	69.0	59.3	67.4	51.6	65.6
INC	–	5.4	–	16.4	2.4	33.7	5.2
ERR	–	26.0	–	23.8	4.9	36.0	9.9

C. Weighted-Score Multichannel Combination

With weighted-score multichannel combination, baseline identification is performed using (16) with

$$LL_s^* = \sum_{c=1}^{C} w_c LL_{s,c} \tag{23}$$

where w_c is the assigned weight for microphone c and $\sum_{c=1}^{C} w_c = 1$. Similarly, for the proposed system identification is performed using (19) with

$$LL_{k,s}^* = \sum_{c=1}^{C} w_c LL_{k,s,c}. \tag{24}$$

The goal with weighted-score channel combination is to assign larger weights to microphones that capture signals of higher quality. Here, the speech-to-reverberation modulation energy ratio (SRMR), shown in [39] to be highly (positively) correlated with the perceived quality of reverberant speech, is used to compute the weights. The SRMR measure, computed per-microphone, is given by

$$\mathrm{SRMR}_c = \frac{\sum_{k=1}^{3}\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,k,c}}{\sum_{k=4}^{8}\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,k,c}} \tag{25}$$

and the weight for microphone c is calculated as

$$w_c = \frac{\mathrm{SRMR}_c}{\sum_{c=1}^{C} \mathrm{SRMR}_c}. \tag{26}$$

D. Experiment Results

To test the adaptive channel selection and multichannel score combination strategies, the six-channel ($C = 6$) database described in Section V-C is used. For the baseline, CMSVN is applied as it resulted in improved performance over not applying channel compensation. Table VI reports performance figures attained with the maximum log-likelihood based channel selection method (column labeled "Max") and the mean- and weighted-score channel combination strategies (columns labeled "Mean" and "Weight," respectively). As observed, for this experiment the conventional mean-score combination technique attained comparable performance with the proposed weighted-score combination method. Careful analysis of the computed weights suggest that the signals captured by the six microphones were of similar quality, an expected result as the microphones were only separated 5 cm apart. Nonetheless, the proposed system with weighted-score channel combination can improve on baseline performance by as much as 36.7% INC and 64.6% ERR for $M = 32$ and by 26.4% INC and 58.7% ERR for $M = 64$.

Table VII, in turn, shows performance comparisons between the proposed weighted-score multichannel combination strategy and the average single-channel performance, as reported in Table III, for both the proposed and baseline systems. For the baseline, additional comparisons are shown between the proposed score combination technique and the three signal-based dereverberation algorithms, as reported in Table IV. As can be seen, the proposed score combination technique outperforms average single-channel identification by 5.4% INC and by 26.0% ERR for the proposed system and by 16.4% INC and 23.8% ERR for the baseline. Additionally, score combination is shown to improve baseline performance over all three dereverberation strategies. Improvements of 2.4% INC and 4.9% ERR are observed relative to DSB; relative to cepstrum-based dereverberation, improvements of 33.7% INC and 36.0% ERR are attained.

VII. CONCLUSION

Modulation spectral features are proposed for environment-robust automatic speaker identification. Several experiments conducted with both artificially generated and recorded

reverberant speech serve to demonstrate the effectiveness of the proposed features for far-field speaker identification. Additionally, a novel objective speech quality based score combination technique is proposed. In this paper, the proposed score combination method is shown to only slightly outperform a conventional mean-score combination technique when the microphones are placed in proximity to each other, such that each microphone captures signals of similar quality. It is conjectured that further gains can be attained with the proposed method if microphones are separated further apart; this investigation, however, is left for a future study.

ACKNOWLEDGMENT

The authors would like to thank D. Gelbart for pointers regarding the ICSI Meeting Corpus and Dr. K. Eneman for providing the executables for the dereverberation algorithms and the multichannel room impulse responses.

REFERENCES

[1] Y. Pan and A. Waibel, "The effects of room acoustics on MFCC speech parameter," in *Proc. Int. Conf. Spoken Lang. Process.*, Oct. 2000, pp. 129–132.

[2] P. Castellano, S. Sridharan, and D. Cole, "Speaker recognition in reverberant enclosures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1996, vol. 1, pp. 117–120.

[3] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 7, pp. 1979–1986, Sep. 2007.

[4] Q. Lin, E. Jan, and J. Flanagan, "Microphone arrays and speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 622–629, Oct. 1994.

[5] J. Ortega-Garcia and J. Gonzalez-Rodriquez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, vol. 2, pp. 929–932.

[6] 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. Int. Conf. Spoken Lang. Process.*, 1996.

[7] 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.

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

[9] A. Abu-El-Quran, J. Gammal, R. Goubran, and A. Chan, "Talker identification using reverberation sensing system," in *Proc. IEEE Conf. Sens.*, Oct. 2007, pp. 970–973.

[10] A. Akula and P. de Leon, "Compensation for room reverberation in speaker identification," in *Proc. Eur. Signal Process. Conf.*, Aug. 2008.

[11] P. de Leon and A. Trevizo, "Speaker identification in the presence of room reverberation," in *Proc. IEEE Biometrics Symp.*, Sep. 2007, pp. 1–6.

[12] N. Morgan *et al.*, "Pushing the envelope—aside," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 81–88, Sep. 2005.

[13] T. H. Falk and W.-Y. Chan, "Spectro-temporal features for robust far-field speaker identification," in *Proc. Int. Conf. Spoken Lang. Process.*, Sep. 2008, pp. 634–637.

[14] H. Kuttruff, *Room Acoustics*, 4th ed. : Elsevier, 2000.

[15] W. Sabine, *Collected Papers on Acoustics*. Cambridge, MA: Harvard Univ. Press, 1922.

[16] M. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, no. 3, pp. 409–412, Mar. 1965.

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

[18] ITU-T Rec. G.191, Software Tools for Speech and Audio Coding Standardization, Int. Telecom. Union, 2005.

[19] ITU-T P.56, Ojective Measurement of Active Speech Level, Int. Telecom. Union, 1993.

[20] K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: Experimental validation," *EURASIP J. Audio, Speech, Music Process.*, p. 19, 2007.

[21] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Meas.*, 2009, to be published.

[22] M. Slaney, "An Efficient Implementation of the Patterson–Holdsworth Auditory Filterbank," Apple Computer, Perception Group, 1993.

[23] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1–2, pp. 103–38, 1990.

[24] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I—model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[25] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[26] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2670–2680, May 1994.

[27] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. Int. Conf. Speech Lang. Process.*, Oct. 1996, pp. 2490–2493.

[28] R. Ratnam, D. Jones, B. Wheeler, W. O'Brien, C. Lansing, and A. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Amer.*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[29] Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Lett. Nature*, vol. 416, pp. 87–90, Mar. 2002.

[30] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," in *Speech Commun.*, Aug. 1995, vol. 17, pp. 91–108.

[31] A. Dempster, N. Lair, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[32] J. Ming, T. Hazend, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.

[33] T. H. Falk, H. Yuan, and W.-Y. Chan, "Spectro-temporal processing for blind estimation of reverberation time and single-ended quality measurement of reverberant speech," in *Proc. Int. Conf. Spoken Lang. Process.*, Sep. 2007, pp. 514–517.

[34] A. Kusumoto, T. Arai, T. Kitamura, M. Takahashi, and Y. Murahara, "Modulation enhancement of speech as a preprocessing for reverberant chambers with the hearing-impaired," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. II, pp. 853–856.

[35] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: Characterizing individual speakers," in *Proc. Int. Conf. Speech Comput.*, 2006.

[36] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1097–1111, Aug. 2008.

[37] Q.-G. Liu, B. Champagne, and P. Kabal, "A microphone array processing technique for speech enhancement in a reverberant space," *Speech Commun.*, vol. 18, no. 4, pp. 317–334, 1996.

[38] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," in *Proc. Int. Workshop Acoust. Echo and Noise Control*, Sep. 2001, pp. 47–50.

[39] T. H. Falk and W.-Y. Chan, "A non-intrusive quality measure of dereverberated speech," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sep. 2008.

[40] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2003, vol. I, pp. 364–367.

[41] *Enhanced Variable Rate Codec (EVRC)*, 3GPP2 C.S0014-0, 1999.

Tiago H. Falk (S'00) was born in Recife, Brazil, in September 1979. He received the B.Sc. degree from the Federal University of Pernambuco, Brazil, in 2002, and the M.Sc. (Eng.) and Ph.D. degrees from Queen's University, Kingston, ON, Canada, in 2005 and 2008, respectively, all in electrical engineering.

He is currently a Postdoctoral Fellow at Bloorview Kids Rehab, affiliated with the University of Toronto, Toronto, ON, Canada. His research interests include multimedia quality measurement and enhancement, multimedia coding and communications, biomedical signal processing, rehabilitation engineering, pattern recognition, and communication theory.

Dr. Falk is recipient of several research excellence awards, including the IEEE Kingston Section Ph.D. Research Excellence Award (2008), the Best Student Paper Awards at the International Conference on Acoustics, Speech, and Signal Processing (2005) and the International Workshop on Acoustic Echo and Noise Control (2008), and the Prof. Newton Maia Young Scientist Award (2001). He has also received several prestigious scholarships, most notably the NSERC Postdoctoral Fellowship (2009), the NSERC Canada Graduate Scholarship (2006), and the Harvard-LASPAU Organization of the American States Graduate Scholarship (2003). He is also a member of the International Speech Communication Association and the Brazilian Telecommunications Society.

Wai-Yip Chan received the B.Eng. and M.Eng. degrees from Carleton University, Ottawa, ON, Canada, and the Ph.D. degree from University of California, Santa Barbara, all in electrical engineering.

He is currently with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada. He has held positions with the Communications Research Centre, Bell Northern Research (Nortel), McGill University, and Illinois Institute of Technology. His research interests are in multimedia signal processing and communications. He is an associate editor of the *EURASIP Journal on Audio, Speech, and Music Processing*.

Dr. Chan is a member of the IEEE Signal Processing Society Speech and Language Technical Committee. He has helped organize IEEE sponsored conferences on speech coding, image processing, and communications. He received a CAREER Award from the U.S. National Science Foundation.