Temporal Dynamics for Blind Measurement of Room Acoustical Parameters

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

***Abstract*—In this paper, short- and long-term temporal dynamic information is investigated for the blind measurement of room acoustical parameters. In particular, estimators of room reverberation time (T_{60}) and direct-to-reverberant energy ratio (DRR) are proposed. Short-term temporal dynamic information is obtained from differential (delta) cepstral coefficients. The statistics computed from the zeroth-order delta cepstral sequence serve as input features to a support vector T_{60} estimator. Long-term temporal dynamic cues, on the other hand, are obtained from an auditory spectrotemporal representation of speech commonly referred to as *modulation spectrum*. A measure termed as reverberation-to-speech modulation energy ratio, which is computed per modulation frequency band, is proposed and serves as input to T_{60} and DRR estimators. Experiments show that the proposed estimators outperform a baseline system in scenarios involving reverberant speech with and without the presence of acoustic background noise. Experiments also suggest that estimators of subjective perception of spectral coloration, reverberant tail effect, and overall speech quality can be obtained with an adaptive speech-to-reverberation modulation energy ratio measure.**

***Index Terms*—Delta cepstrum, direct-to-reverberation ratio, modulation spectrum, reverberation time, temporal dynamics.**

I. Introduction

When speech is produced in an enclosed environment, the acoustic signal follows multiple paths from source to receiver. Such reflections may arrive with delays ranging from a few milliseconds to a few seconds, depending on the room geometry and the sound absorption properties. Early reflections, which are on the order of a few tens of milliseconds, modify the signal short-time spectrum, causing a change in signal timbre; such an effect is termed spectral coloration [1], [2]. Delays greater than 50 ms (termed late reflections), on the other hand, are perceived as distinct copies of the direct path signal and cause temporal coloration distortions. The exponential decay of late reflections results in temporal smearing, which, in turn, decreases the perceived speech quality and intelligibility.

As hands-free communication technologies advance, reverberation has become a burden, in particular for applications with far-field microphones. A strategy to improve system performance is to estimate the room acoustical parameters and use signal processing techniques that are most appropriate for the environment (e.g., see [3] and [4]). Traditionally, the time-domain room impulse response (IR) or room geometry and wall absorption properties are used to measure the room acoustical parameters. Offline measurement of the room IRs, however, is a laborious task. In addition, the IR varies with acoustic source positioning, room temperature, and placement of room furnishings. As a consequence, the room acoustical parameters obtained from room IR measurements are not feasible for real-time signal processing applications. To this end, blind signal-based measurement, where the room acoustical parameters are obtained from the reverberant speech signal, has been the focus of more recent research. Special emphasis has been given to the blind estimation of the reverberation time (T_{60}) parameter (see Section II-B).

In the past, a handful of blind T_{60} estimators have been proposed. In [5], the diffuse tail of the reverberation is modeled as exponentially damped Gaussian white noise. A maximum-likelihood (ML) estimate of the time constant of the decay is used to characterize T_{60}. With ML-based approaches, it is common to assume that the source signal abruptly stops and has long pauses between speech segments; such requirements are needed to attain reliable estimates. As expected, the performance of ML-based methods is compromised for noise-corrupted reverberant speech. Notwithstanding, the work described in [6] proposes a "generalized" ML procedure that loosens the aforementioned assumptions and allows for blind T_{60} estimation under noisy environments.

Alternately, the work described in [7] shows that reverberation corrupts the harmonic structure of voiced speech segments. Hence, a measure of pitch "strength" (or periodicity) is used to blindly estimate T_{60}. The estimator, however, is shown to be sensitive to speaker gender. Additionally, the kurtosis of linear prediction (LP) residuals is used in [8] for blind T_{60} characterization. The idea is that for clean voiced speech segments, LP residuals have strong peaks corresponding to glottal pulses. The peaks become smeared in time as reverberation increases, thus reducing the LP residual kurtosis to that of a Gaussian distribution. LP residual-based methods have also been successfully used in the past for noise and reverberation suppression [9]–[11].

In this paper, we investigate the use of temporal dynamic information for the blind measurement of room acoustical parameters. Short-term dynamic information is obtained from commonly used differential (delta) cepstral coefficients. The statistics computed from the zeroth-order delta cepstral sequence are shown to provide useful cues for blind T_{60} estimation. Moreover, long-term dynamic information is obtained by means of spectral analysis of the temporal envelopes of speech,

Manuscript received July 15, 2008; revised December 22, 2008. Current version published March 20, 2010. The Associate Editor coordinating the review process for this paper was Dr. Jesús Ureña.

The authors are with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: tiago.falk@ece.queensu.ca; 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/TIM.2009.2024697

which is a process commonly termed as *modulation spectrum* processing. Here, the work described in [12] is extended to allow for the blind measurement of several room acoustical parameters, including measures of subjective perception of spectral coloration, reverberant tail effect, and overall speech quality. Experiments show that the proposed estimators outperform a baseline system in scenarios involving reverberant speech with and without the presence of acoustic background noise.

The remainder of this paper is organized as follows: Section II describes models, characterization, and simulation of room reverberation. Section III provides motivation and a description of the features obtained from temporal dynamic information. The experimental results are presented in Section IV. An objective assessment of the perceived reverberation effects is discussed in Section V, and conclusions are given in Section VI.

II. Room Reverberation

In this section, models of room reverberation are presented. The parameters commonly used to characterize reverberation are presented, as well as methods to generate reverberant speech.

A. Models of Room Reverberation

Conventionally, the propagation from source to microphone in a reverberant enclosure is modeled as a linear filtering process. The reverberant signal $s(n)$ is modeled as a convolution of the anechoic source speech signal $x(n)$ with the room IR $r(n)$ as

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

If additive background noise $N(n)$ is present, then (1) becomes

$$s(n) = x(n) * r(n) + N(n). \tag{2}$$

It is known that under the diffuse sound field assumption, the ensemble average of the squared room IR exponentially decays with time [13] as

$$\langle r^2(n) \rangle = A \exp(-kn). \tag{3}$$

The angled brackets $\langle \cdot \rangle$ denote the ensemble average, A is a gain term, and k is the damping factor given by [13]

$$k = \log 10^6 / (F_s \times T_{60}) \tag{4}$$

where F_s is the sampling frequency, and T_{60} is the so-called reverberation time, as described in Section II-B. The plot in Fig. 1 illustrates the exponential decay of a room IR generated via the image method [14] with $T_{60} = 0.5$ s and $F_s = 8$ kHz. The dashed curve in the figure illustrates the exponential decay given by (3) with $A = 0.0045$.

B. Characterization of Room Reverberation

Reverberation time (T_{60}) is the parameter most widely used to characterize room acoustics. By definition, it is the time required for the sound energy to decay by 60 dB after the sound source has been turned off [15]. Commonly, the so-called Schroeder integral is used to calculate T_{60} from the room IR [16]. Other parameters that characterize room acoustics and are obtained from the room IR include the early decay time (interval required for the energy to decay by 10 dB), the speech clarity index (energy ratio between the 50-ms early reflections and the remaining late reflections) [17], and the direct-to-reverberant energy ratio (DRR). The DRR, which is expressed in decibels, is the energy ratio between the direct sound and the room reverberation and is given by

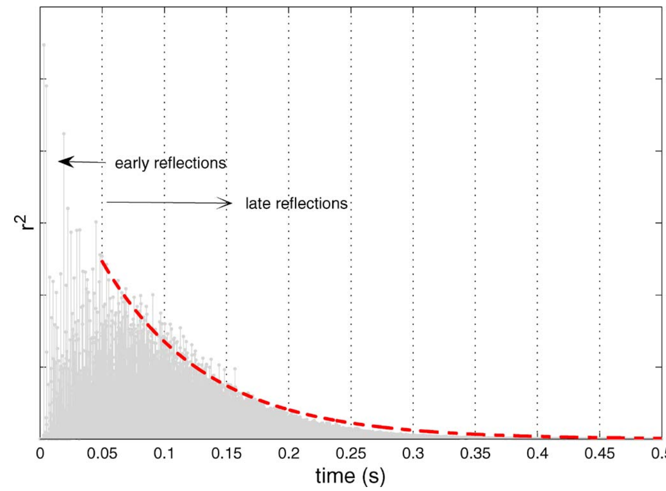

Fig. 1. Exponential decay of the late reflections of a room with $T_{60} = 0.5$ s.

$$\mathrm{DRR} = 10 \log_{10} \left(\frac{\sum_{n=0}^{n_d} r^2(n)}{\sum_{n=n_d+1}^{\infty} r^2(n)} \right) \tag{5}$$

where $n_d F_s$ is the direct sound arrival time.

Moreover, the spectral content of the room IR can provide information regarding spectral coloration. In [18] and [19], the second-order moment of the room frequency response is proposed as a measure of spectral coloration. Additionally, subjective listening tests may be used to characterize the perceived quality of speech signals produced in reverberant enclosures. In [20], subjective listening tests are used to characterize the perception of timbre. Recently, listening tests have been used to characterize the subjective perception of coloration, reverberation decay tail effects, and overall quality for reverberant and reverberation-suppressed speech [21]. The test follows the guidelines described in the International Telecommunications Union (ITU-T) Recommendation P.835 [22].

C. Simulation of Reverberant Speech

Two tools are used to generate reverberant speech: 1) SImulation of REal ACoustics (SIREAC) [23] and 2) the ITU-T software package described in Recommendation G.191 [24]. Anechoic speech from eight speakers (half male, half female) is used throughout our experiments. A total of 256 utterances (averaging 6 s each) are spoken per speaker; half of the utterances are in English and the other half in French. Each of the speech samples consists of two sentences separated by an

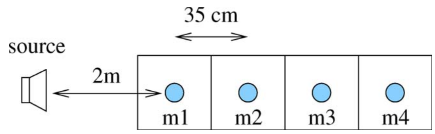

Fig. 2. Microphone array setup at the Bell Labs varechoic chamber.

TABLE I
ROOM ACOUSTICAL PARAMETERS FOR REAL ROOM IRs

Room	$T_{60}(s)$	DRR (dB)
VC-100%-m1	0.3	1
VC-100%-m2	0.3	1
VC-100%-m3	0.3	-1
VC-100%-m4	0.3	-1
VC-43%-m1	0.5	0
VC-43%-m2	0.5	-3
VC-43%-m3	0.5	-2
VC-43%-m4	0.5	-5
VC-0%-m1	0.9	-5
VC-0%-m2	0.9	-7
VC-0%-m3	0.9	-7
VC-0%-m4	0.9	-9
Office	0.6	-4
Meeting	0.9	-7
Lavatory	1.3	-9
Cafeteria	1.5	-14

approximately 800-ms pause; all signals are stored with 8-kHz sampling rate and 16-bit precision. SIREAC is used to artificially generate reverberant speech with T_{60} values between 0.2 and 1 (with 0.1-s increments), 1.5, and 2 s. The level of the reverberant speech signal is normalized to -26 dB overload (dBov) using the ITU-T P.56 voltmeter [25].

The ITU-T G.191 tool is used to convolve the room IRs collected from real environments with the anechoic speech signals. The real room IRs are stored with 8-kHz sampling rate and include those collected with a four-channel linear microphone array (as depicted in Fig. 2) at the Bell Labs varechoic chamber[1] with 100%, 43%, and 0% panels open [26] and those collected with a single microphone in a large cafeteria, a medium-sized meeting room, a small lavatory, and a medium-sized office [3]. As with the simulated data, the reverberant speech signals are normalized to -26 dBov. Table I reports the parameters T_{60} and DRR, which are computed from the room IRs, for the aforementioned environments. In the table, varechoic chamber data are represented as "VC-%-mi," where "%" represents the percentage of open reflective panels, and "mi" represents the microphone number in the microphone array (see Fig. 2).

III. TEMPORAL DYNAMICS AND PROPOSED ESTIMATORS

In this section, a description of the features used to capture short- and long-term temporal dynamics is given; the proposed T_{60} and DRR estimators are also described.

[1]The Bell Labs varechoic chamber is a rectangular room with 368 independently actuated surfaces in the walls, ceiling, and floor. T_{60} is controlled by the percentage of open panels.

A. Short-Term Temporal Dynamics

Short-term energy dynamics is used for the blind measurement of T_{60}. In this paper, the zeroth-order mel-frequency cepstral coefficient is proposed as a measure of the short-term log-spectral energy, and the zeroth-order delta coefficient is proposed as a measure of the log-energy rate of change [27]. Since such coefficients are commonly extracted by speech and speaker recognition systems, blind T_{60} estimation can be used to improve the recognition performance while requiring negligible computational overhead. Let $c_0(m)$ denote the zeroth-order cepstral coefficient for frame m. $\Delta c_0(m)$ represents the zeroth-order delta coefficient and is computed as [28]

$$\Delta c_0(m) = \sum_{l=-L}^{L} l c_0(m+l) \tag{6}$$

where the normalization factor $\sum_{l=-L}^{L} l^2$ is omitted as it does not affect the results; in our simulations, $L = 5$ is used.

Fig. 3(a) depicts (from top to bottom) the waveform c_0 and Δc_0 sequences for a clean speech signal, respectively. As observed, speech onsets induce positive "peaks" in the Δc_0 sequence; analogously, speech offsets induce negative peaks. Fig. 3(b) and (c) illustrates the effects of increasing T_{60} on the speech offset regions (e.g., between 1.75 and 2.25 s); the plots correspond to $T_{60} = 0.4$ and 1 s, respectively. As can be seen, as T_{60} increases, c_0 decays at a slower rate, which, in turn, decreases the log-energy rate of change. Moreover, due to temporal smearing, the intervals between phonemes are filled with reverberant energy (e.g., between 0.5 and 1.75 s), thus also decreasing the log-energy rate of change.

To capture such reverberation tail effects, sample statistics are computed from $N \Delta c_0$ samples (x_i). In particular, standard deviation (σ_Δ), skewness $(\mathcal{S}_\Delta)$, kurtosis $(\mathcal{K}_\Delta)$, and median absolute deviation $(\mathcal{D}_\Delta)$ are computed according to

$$\sigma_\Delta = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{7}$$

$$\mathcal{S}_\Delta = \frac{\sqrt{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{3/2}} \tag{8}$$

$$\mathcal{K}_\Delta = \frac{N\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{2}} - 3 \tag{9}$$

$$\mathcal{D}_\Delta = \text{median}_i\left(|x_i - \text{median}_j(x_j)|\right) \tag{10}$$

where $\bar{x}$ indicates the sample average of x_i.

The aforementioned inverse relationship between T_{60} and the log-energy rate of change can be observed in the σ_Δ versus T_{60} plots depicted in Fig. 4 (solid curve). Moreover, since reverberation tail effects are more pronounced in speech offset intervals, it is expected that, with an increase in T_{60}, fewer negative peaks will occur in the Δc_0 sequence. A direct consequence of this effect is the increase in positive skewness

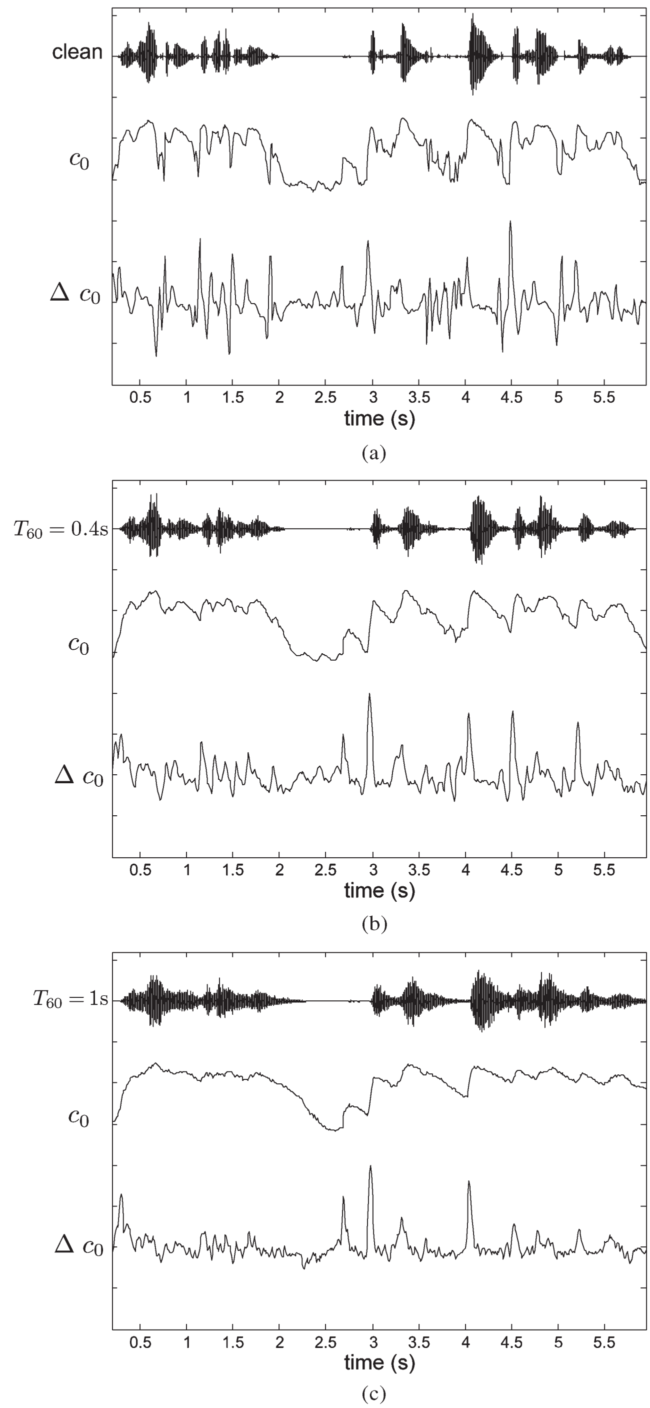

Fig. 3. (From top to bottom) Waveform, c_0, and Δc_0, for (a) clean speech and reverberant speech with (b) $T_{60} = 0.4$ s and (c) $T_{60} = 1$ s.

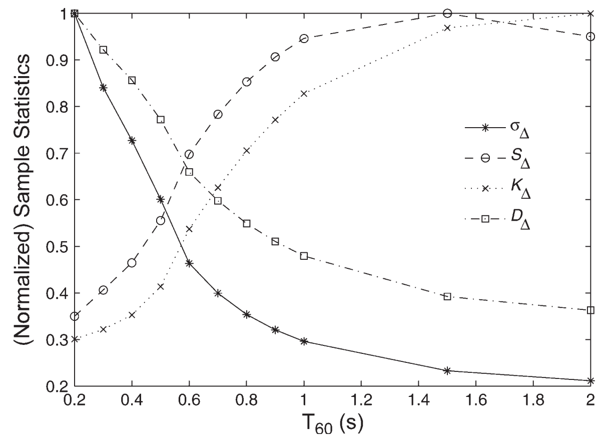

Fig. 4. Plots of (normalized) sample statistics versus T_{60}. Data points represent average statistics for simulated reverberant speech signals.

$\mathcal{S}_\Delta$, as illustrated in Fig. 4 (dashed curve). Note that, with our speech data, both speech offsets and onsets are severely affected by the reverberation tail for very large reverberation times, hence the decrease in $\mathcal{S}_\Delta$ for $T_{60} = 2$ s. Additionally, it is observed that an increase in T_{60} will result in a shift of the variance to large deviations, rendering the Δc_0 distribution with a heavier tail. Hence, an increase in $\mathcal{K}_\Delta$ is observed, as illustrated in Fig. 4 (dotted curve). Finally, $\mathcal{D}_\Delta$ (dash-dot curve) is used as it provides increased robustness (relative to σ_Δ) to extreme Δc_0 deviations around the mean, which is an effect commonly observed in multiple-sentence speech signals with an intersentence duration that is longer than T_{60}.

Due to the nonlinear relationship between T_{60} and Δc_0 sample statistics, we propose to use machine learning algorithms to blindly estimate the room acoustical parameters. In our experiments, a support vector regressor (SVR) is used to estimate T_{60}; the reader is referred to [29] for a comprehensive description of support vector regression. Further performance improvements may be attained with alternate machine learning paradigms, such as neural networks [30] or relevance vector machines [31]; such investigation, however, is left for future study. The input to the SVR is a 4-D vector consisting of $\mathbf{u}_s = [\sigma_\Delta, \mathcal{S}_\Delta, \mathcal{K}_\Delta, \mathcal{D}_\Delta]$. As will be shown in Section IV-D, a simple adaptation procedure can be used to improve the estimation performance in the presence of acoustic background noise.

B. Long-Term Temporal Dynamics

To capture the long-term temporal dynamics of the reverberant speech signal, we propose to use an auditory spectrotemporal representation of speech, which is commonly referred to as *modulation spectrum*. The modulation spectrum characterizes the frequency content (or rate of change) of the long-term speech temporal envelopes. In our experiments, the spectrotemporal signal representation is obtained using the signal processing steps depicted in Fig. 5.

First, the speech signal $s(n)$ is filtered by a bank of critical-band filters. In our simulations, a critical-band gammatone filterbank, with 23 filters, is used to emulate the processing performed by the cochlea [32]. Filter center frequencies range from 125 Hz to nearly half the sampling rate (e.g., 3567 Hz for 8-kHz sampling rate). The filter bandwidths are characterized by the equivalent rectangular bandwidth (ERB) [33]. The ERB for filter j, $j = 1, \ldots, 23$, is given by

$$\mathrm{ERB}_j = \frac{f_j}{Q_{\mathrm{ear}}} + B_{\min} \tag{11}$$

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. The plot in Fig. 6 illustrates the frequency response of the 23-channel gammatone filterbank used in our experiments.

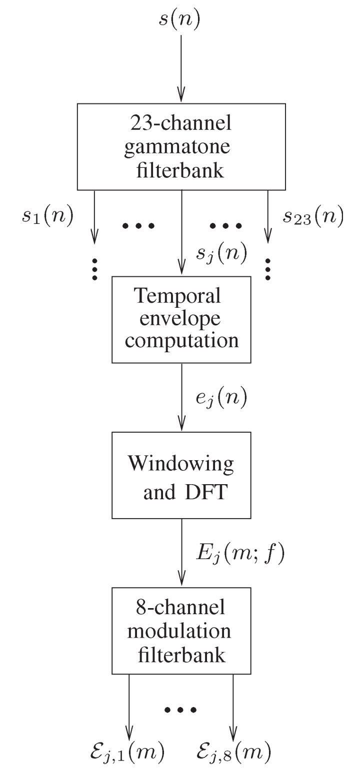

Fig. 5. Block diagram of the signal processing steps involved in the computation of the spectrotemporal signal representation.

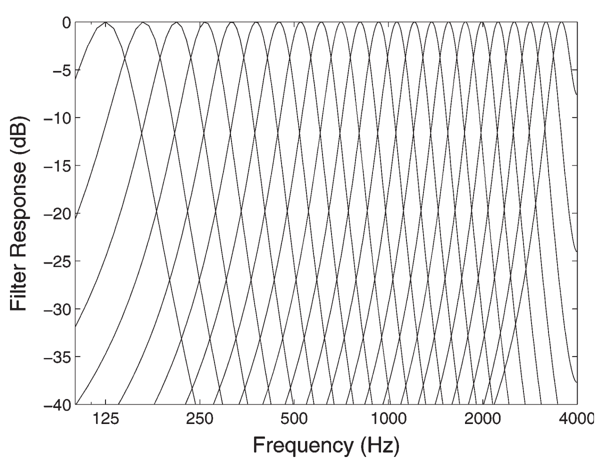

Fig. 6. Filter responses for the 23-channel gammatone filterbank.

The output signal of the jth channel is given by

$$s_j(n) = s(n) * h_j(n) \tag{12}$$

where $h_j(n)$ is the IR of the jth critical-band filter. Temporal dynamics information is obtained from the temporal envelope of $s_j(n)$. In our experiments, the Hilbert transform $\mathcal{H}\{\cdot\}$ is used to obtain the temporal envelopes $e_j(n)$. The temporal envelope (also called Hilbert envelope) is computed as the magnitude of the complex analytic signal $\tilde{s}_j(n) = s_j(n) + \jmath\mathcal{H}\{s_j(n)\}$. Hence

$$e_j(n) = \sqrt{s_j(n)^2 + \mathcal{H}\left\{s_j(n)\right\}^2}. \tag{13}$$

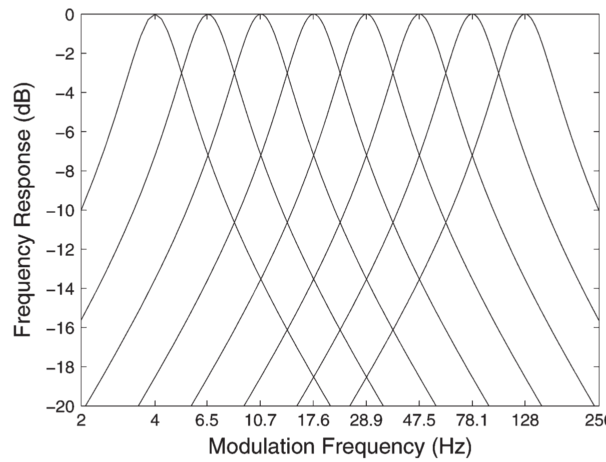

Fig. 7. Filter responses for the eight-channel modulation filterbank.

The 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. Here, 256-ms frames are used to obtain long-term temporal dynamics information and appropriate resolution for low-frequency modulation frequencies (e.g., around 4 Hz).

The modulation spectrum for critical-band j is obtained by taking the discrete Fourier transform $\mathcal{F}\{\cdot\}$ of the temporal envelope $e_j(m)$, i.e., $E_j(m; f) = |\mathcal{F}(e_j(m))|$, where f denotes the modulation frequency. Modulation frequency bins are grouped into K-bands to emulate an auditory-inspired modulation filterbank [34]. The kth modulation band energy for frame m is denoted as $\mathcal{E}_{j,k}(m)$, $k = 1, \ldots, K$. In the experiments described in Section IV, $K = 8$ is used as it resulted in superior performance. For the experiments described in Section V, on the other hand, optimal values for K are chosen on a per-signal basis. Fig. 7 depicts the frequency response of the eight-channel modulation filterbank used in our experiments. The filters are second-order bandpass with quality factor $Q = 2$, as suggested in [34].

The modulation energy $\mathcal{E}_{j,k}(m)$ is then averaged over all active speech frames to obtain

$$\bar{\mathcal{E}}_{j,k} = \frac{1}{N_{\text{act}}} \sum_{i=1}^{N_{\text{act}}} \mathcal{E}_{j,k}^{\text{act}}(i) \tag{14}$$

where N_{act} denotes the number of active speech frames, and $\mathcal{E}_{j,k}^{\text{act}}(i)$ is the modulation energy of such frames; the voice activity detection (VAD) algorithms used in our experiments are described in Section IV-E. The $\bar{\mathcal{E}}_{j,k}$ notation will be used throughout the remainder of this paper to indicate the active speech modulation energy of the jth critical-band signal grouped by the kth modulation filter. A representative illustration of $\bar{\mathcal{E}}_{j,k}$ for a clean speech signal is depicted in Fig. 9(a). Moreover, the notation $\vec{\bar{\mathcal{E}}}_k$ will be used to denote the 23-D energy vector for modulation channel k.

For clean (unreverberated) speech, it is known that the Hilbert temporal envelopes contain dominant frequencies ranging from 2 to 20 Hz [35], [36] with spectral peaks at approximately 4 Hz, which corresponds to the syllabic rate of spoken speech [37]. With reverberant speech, the diffuse IR reverberant tail is often modeled as an exponentially damped Gaussian white noise process [5]. As such, it is expected that reverberant signals attain more Gaussian white-noise-like properties with increasing T_{60}. Since the Hilbert envelope can contain frequencies (which are also termed *modulation frequencies*) up to the bandwidth of its originating signal [38], the reverberant signals are expected to contain significant modulation frequency components beyond the 2- to 20-Hz range of syllabic modulation frequencies. The plots in Fig. 8 assist in illustrating the effects of T_{60} on the temporal envelopes. Subplot (a) depicts $e_j(n)$ and the positive portion of $s_j(n)$ $(s_j^+(n))$ for a 256-ms frame of clean speech. Subplots (b) and (c), in turn, depict the corresponding signals for reverberant speech with $T_{60} = 0.4$ and 1 s, respectively. The plots in the figure are for $j = 14$, which corresponds to a filter center frequency of 1.2 kHz.

Fig. 9 depicts the active speech modulation energy $\bar{\mathcal{E}}_{j,k}$ for the speech signals used to produce Fig. 8. In the plots, the modulation energy values are normalized by the maximum energy obtained over all modulation frequency bands. Fig. 9(a) depicts the normalized modulation energy for a clean speech signal. As observed, the most significant modulation frequency components lie below 20 Hz. The plots in Fig. 9(b) and (c), in turn, depict $\bar{\mathcal{E}}_{j,k}$ for the corresponding reverberant speech signals with $T_{60} = 0.4$ and 1 s, respectively. An increased modulation energy at higher modulation frequency bands is observed in these two plots. Additionally, more pronounced reverberation effects are observed for modulation frequencies greater than 20 Hz (i.e., $k = 5{-}8$).

It can also be observed from Fig. 9 that an increase in T_{60} has negligible effect on $\vec{\bar{\mathcal{E}}}_1$, which corresponds to the 4-Hz modulation frequency attributed to the syllabic rate of speech. This insight is used to develop a reverberation-to-speech modulation energy ratio (RSMR) measure computed per modulation frequency channel k and given by

$$\mathrm{RSMR}_k = \frac{\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,k}}{\sum_{j=1}^{23} \bar{\mathcal{E}}_{j,1}}. \tag{15}$$

To illustrate the nonlinear effects of T_{60} on RSMR, the plots in Fig. 10 depict RSMR_k versus T_{60} for $k = 5{-}8$. The data points reflect the average RSMR for the simulated reverberant speech signals described in Section II-C.

As expected, more pronounced effects are observed for $k = 8$ with an increase in T_{60}. In pilot experiments, we have observed that estimators based only on RSMR_8 attain reliable performance for simulated data, but a slightly lower performance is attained for reverberant speech generated from recorded room IRs. To design estimators that are robust to unseen (real) conditions, an SVR is used to estimate T_{60}. The 4-D vector input to the SVR is given by $\mathbf{u}_l = [\mathrm{RSMR}_5, \mathrm{RSMR}_6, \mathrm{RSMR}_7, \mathrm{RSMR}_8]$.

Moreover, as aforementioned, reverberation tail effects can be quantified from $\vec{\bar{\mathcal{E}}}_k$, $k = 5{-}8$. Speech information, on the

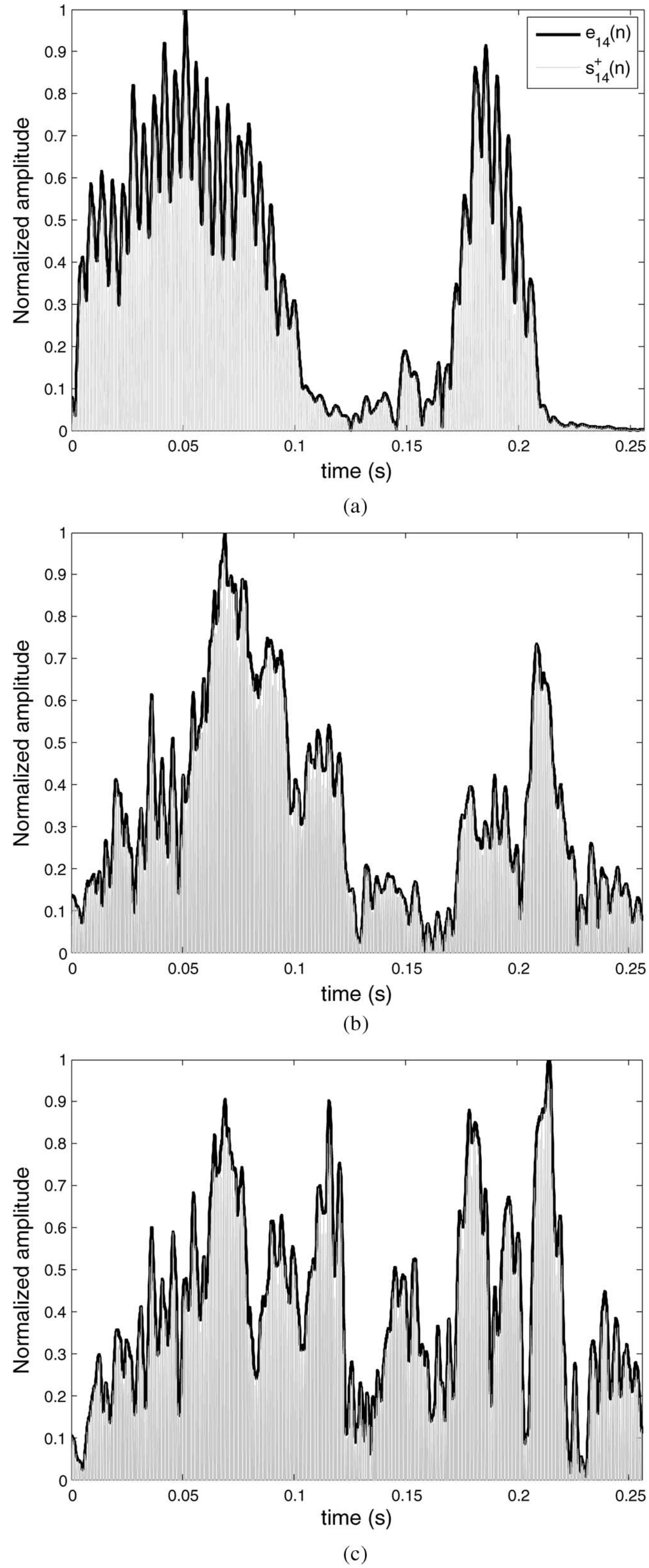

Fig. 8. Temporal envelope $e_j(n)$ and positive portion of the gammatone-filtered signal $s_j^+(n)$ for (a) clean speech and reverberant speech with (b) $T_{60} = 0.4$ s and (c) $T_{60} = 1$ s. The plots are for $j = 14$ corresponding to $f_{14} = 1.2$ kHz.

other hand, can be obtained from $\vec{\bar{\mathcal{E}}}_1$. This insight is used to compute an overall RSMR (ORSMR) measure, which is shown to be highly correlated with DRR. The measure ORSMR is

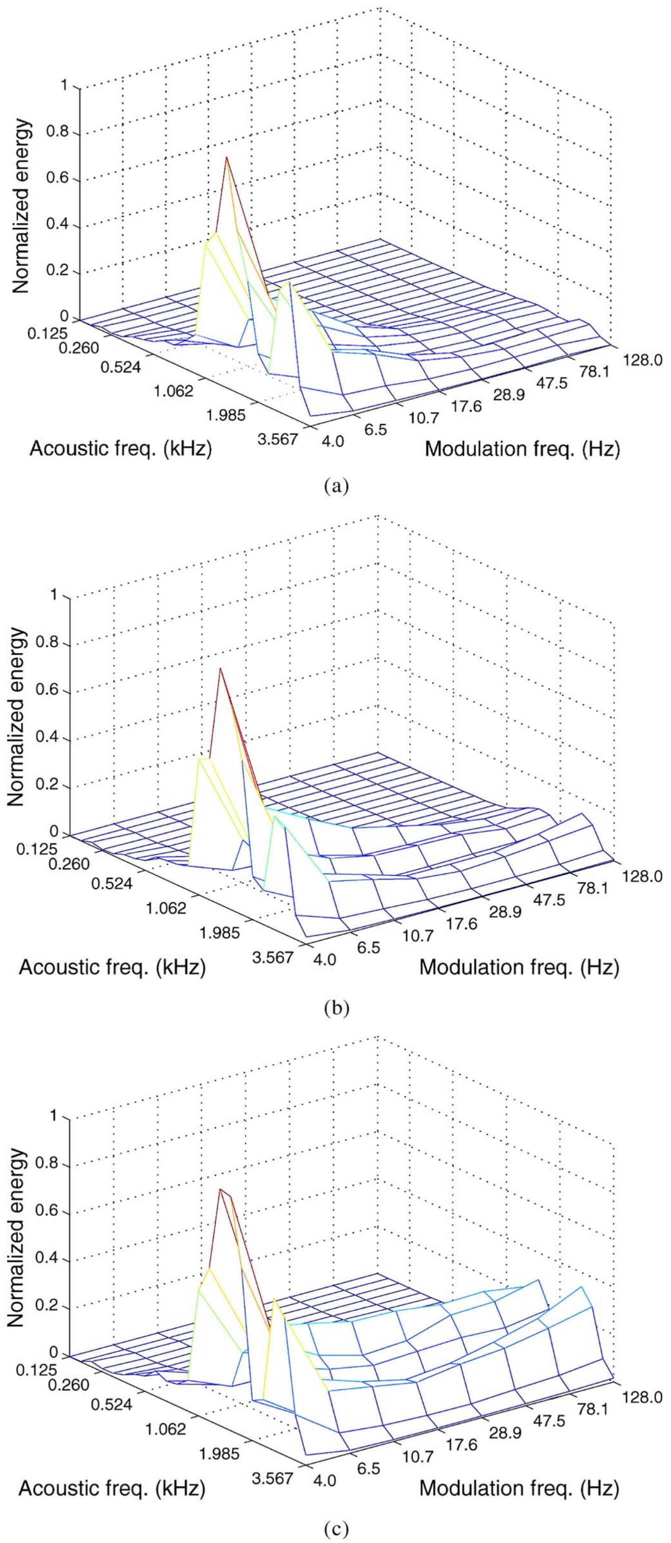

Fig. 9. $\bar{\mathcal{E}}_{j,k}$ for (a) clean speech and reverberant speech with (b) $T_{60} = 0.4$ s and (c) $T_{60} = 1$ s.

given by

$$\text{ORSMR} = \frac{\sum_{k=5}^{8}\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,k}}{\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,1}} = \sum_{i=5}^{8}\text{RSMR}_i. \tag{16}$$

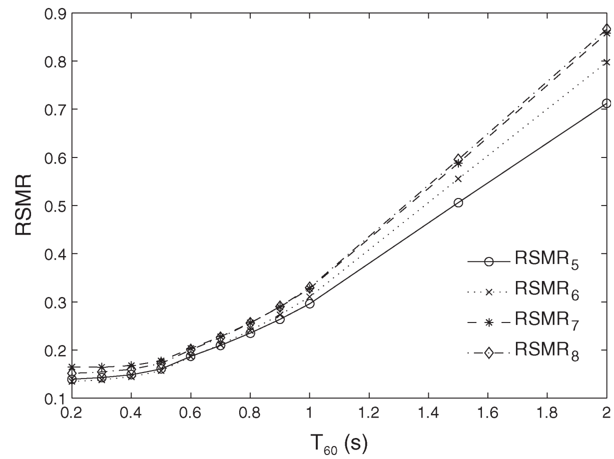

Fig. 10. Plots of RSMR_k versus T_{60} for $k = 5{-}8$.

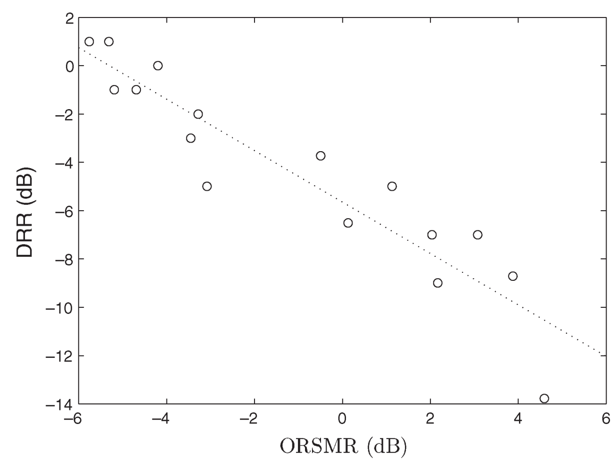

Fig. 11. Plot of DRR versus ORSMR; the latter is given by (16).

The plot in Fig. 11 illustrates a linear regression relationship between ORSMR (expressed in decibels) and DRR. The data points represent DRR values described in Table I and average ORSMR values obtained from English reverberant speech signals generated with recorded room IRs. Hence, the following DRR estimator ($\widehat{\text{DRR}}$) is proposed:

$$\widehat{\text{DRR}} = -5.6467 - 1.0644 \times \text{ORSMR (in decibels)} \tag{17}$$

where ORSMR and $\widehat{\text{DRR}}$ are expressed in decibels. In the sequel, the proposed estimators are tested on simulated and recorded reverberant speech.

IV. Experiments

In this section, experimental setup, performance figures, baseline estimator, and two experiments are described. The first experiment tests the performance of the proposed estimators in reverberant enclosures, and the second experiment tests the performance of the proposed estimators in environments corrupted by reverberation and acoustic background noise.

A. Experimental Setup

Reverberant speech signals generated with the SIREAC tool are used to train the SVRs. Throughout the remainder of this section, the notations SVR_s and SVR_l will be used to distinguish blind T_{60} estimators derived from short- and long-term temporal dynamics, respectively. On our data, SVR with radial basis kernels and parameters optimized via linear search are shown to provide the best estimation performance. The results to follow are all based on using radial basis SVRs. The reader is referred to [29] for a comprehensive review on support vector machines. Additionally, the SIREAC tool is used to generate speech signals degraded by reverberation and acoustic background noise. Reverberant speech is generated with T_{60} ranging from 0.2 to 1 s (with 0.1-s increments) and with babble noise at five SNR levels (5–25 dB with 5-dB increments). As shown in Section IV-D, a simple adaptation process can be used to increase the performance of the proposed T_{60} estimators in the presence of acoustic noise. The "adapted" SVR is termed $\widetilde{\mathrm{SVR}}$ throughout the remainder of this paper.

B. Performance Figures and Baseline Estimator

The correlation (ρ), the mean square error (ϵ), and the median absolute error (γ) are used as estimator figures of merit. The correlation between blindly estimated parameter values (w_i) and parameter measurements obtained from room IR (y_i) is computed via Pearson's formula [39] as

$$\rho = \frac{\sum_{i=1}^{N}(w_i - \bar{w})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(w_i - \bar{w})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{18}$$

where $\bar{w}$ is the average of w_i, and $\bar{y}$ is the average of y_i. The mean square error ϵ is given by

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}(w_i - y_i)^2 \tag{19}$$

and the median absolute error γ is given by

$$\gamma = \mathrm{median}_i\left(|w_i - y_i|\right). \tag{20}$$

In the sequel, error measures are reported in milliseconds for T_{60} estimators and in decibels for DRR estimators.

The performance of the proposed T_{60} estimators is compared with a baseline estimator based on the kurtosis of twelfth-order LP residuals (κ_{LP}) computed over 32-ms frames. In pilot experiments, the LP residual-based method was found to be more robust to background noise when compared with other existing ML-based schemes (e.g., [5], [6]) and, as opposed to pitch-based methods, was found to be insensitive to speaker gender. The plot in Fig. 12 shows the nonlinear relationship between κ_{LP} and T_{60}. As can be seen, the LP residual kurtosis approaches that of a Gaussian distribution with increasing T_{60}. A clean speech, which is represented by $T_{60} = 0$ s in the plot, attains a high κ_{LP}; this is expected as the LP residual for clean speech contains sparse peaks corresponding to glottal pulses. In our experiments, simulated data are used to train the baseline SVR, which is henceforth referred to as SVR_κ. Moreover, to the best of our knowledge, ours is the first blind estimator of DRR; thus, comparisons with a baseline are not carried out for $\widehat{\mathrm{DRR}}$.

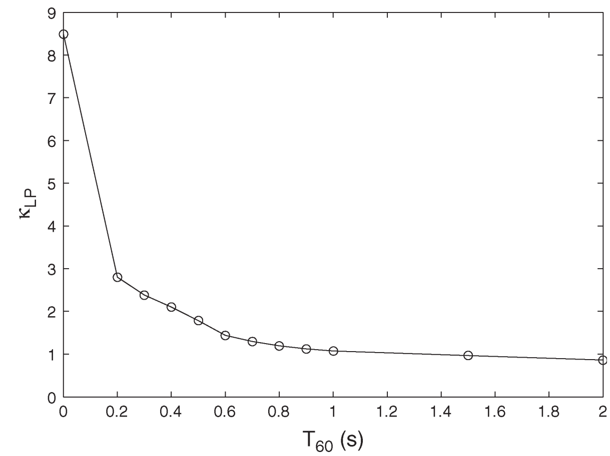

Fig. 12. Plot of κ_{LP} versus T_{60}. The LP residual kurtosis for clean unreverberated speech is represented as $T_{60} = 0$ in the plot.

C. Experiment 1—Reverberation Only

As aforementioned, reverberant signals simulated with the SIREAC tool are used to train T_{60} estimators SVR_s, SVR_l, and SVR_κ. Bilingual reverberant data generated with real single- and multichannel recordings of room IR are regarded as unseen data and are used for testing. Table II reports the performance figures for the proposed estimators as well as for the baseline estimator. Columns labeled "%" indicate the percentage increase in ρ or percentage decrease in ϵ and γ attained with the proposed measures relative to the baseline. As observed, both proposed estimators outperform the baseline method. SVR_s results in superior improvements relative to SVR_l for the data generated with the multichannel room IR. For data generated from the single-channel room IR, both estimators attain similar performance figures, with SVR_l obtaining somewhat lower γ.

Moreover, as aforementioned, English reverberant speech data are used to train the coefficients in (17). Hence, French reverberant speech data are regarded as unseen and used to test the performance of the proposed DRR estimator. Fig. 13 depicts DRR versus the average $\widehat{\mathrm{DRR}}$ for the unseen test set; $\rho = 0.98$, $\epsilon = 1.11$ (dB), and $\gamma = 0.97$ (dB) are attained. The results are encouraging given that no knowledge of the room IR is used for estimation. Additionally, both T_{60} and DRR estimators are found to be *insensitive* to speaker gender.

D. Experiment 2—Reverberation and Background Noise

To test the performance of the proposed estimators in practical scenarios, we use speech corrupted by reverberation and babble (crowd) noise. Table III reports the performance measures for SVR_κ, SVR_s, and SVR_l for various noise levels. As can be seen, both proposed estimators outperform the baseline, with SVR_l showing reduced sensitivity to the noise level. This behavior is expected as babble noise has speechlike

TABLE II
PERFORMANCE COMPARISON OF PROPOSED T_{60} ESTIMATORS FOR SPEECH CORRUPTED BY REVERBERATION

	Multi-Channel Room IR					Single-Channel Room IR				
	SVR_κ	SVR_s	%	SVR_l	%	SVR_κ	SVR_s	%	SVR_l	%
ρ	0.71	0.96	35.2	0.90	26.8	0.82	0.89	8.5	0.86	4.9
ϵ (ms)	30.4	11.7	61.5	29.5	3.0	69.9	30.6	56.2	32.2	53.9
γ (ms)	109.1	80.8	25.9	105.9	2.9	173.5	99.2	42.8	94.7	45.4

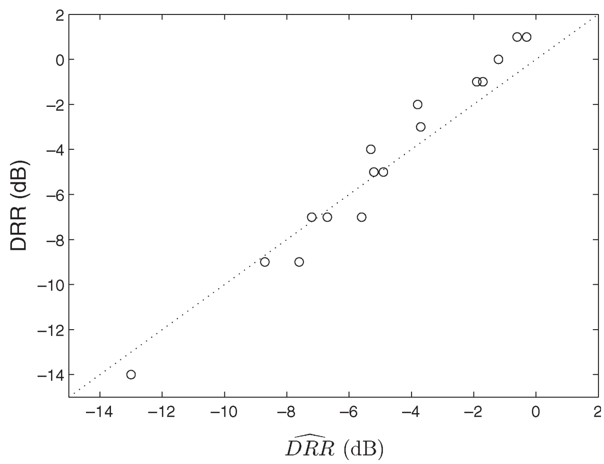

Fig. 13. Plot of DRR versus average $\widehat{\text{DRR}}$ for unseen French test data.

characteristics, thus mostly affecting $\vec{\vec{\mathcal{E}}}_1$. Overall, SVR_s attains average improvements over the baseline of 38.2%, 35.4%, and 23.1% in ρ, ϵ, and γ, respectively, and SVR_l attains average improvements of 22.2%, 81%, and 68.2%.

Despite improved performance over the baseline, high ϵ and γ errors compromise the usability of SVR_s for practical applications. To reduce estimation errors, an "adaptation" process is proposed where the estimated SNR is introduced as an added feature to the support vector estimators. Here, the noise analysis module of the ITU-T P.563 algorithm [40] is used to estimate the SNR. Estimation is performed by calculating the levels of speech and noise sections identified during VAD [41]. In a controlled experiment, the estimated SNR is shown to be highly correlated with the true SNR ($\rho = 0.96$). Table IV reports improvements in ϵ and γ for the adapted T_{60} estimators; as observed, adaptation substantially reduces the estimation errors. Relative to the adapted baseline, $\widetilde{\text{SVR}}_s$ attains average improvements of 37.3% in ϵ and 20.7% in γ. $\widetilde{\text{SVR}}_l$ obtains average improvements of 50.5% and 40.9%, respectively. The improvements in ρ over the nonadapted estimators are considerably lower—on the order of 7%—for all three estimators and, thus, are omitted from the table.

E. Discussion

As can be seen from (15) and (16), the proposed measures are based on summing the per-band modulation energy over 23 acoustic frequency channels. To reduce the algorithmic processing time, the critical-band gammatone filterbank can be omitted, and the per-band modulation energy can be computed over the entire 4-kHz signal bandwidth. On our data, such simplified configuration is capable of reducing the algorithmic processing time by a maximum 40%. It has been observed, however, that the reduced-complexity configuration lowers the measurement performance by as much as 20%, in particular for noise-corrupted environments and for enclosures with low T_{60} (≤ 0.3 s). As a consequence, the reduced-complexity alternative should be considered only if limited resources are available. Moreover, as will be described in Section V, the critical-band gammatone filterbank is useful for the objective assessment of perceived reverberation effects and, thus, has been kept in our experiments.

Additionally, we have experimented with two VAD algorithms. The first is available in the ITU-T G.729 speech codec [42], and the second is available in the adaptive multirate (AMR) wireless speech codec [43]. For the reverberant speech files used in Experiment 1 (Section IV-C), both VAD algorithms attained similar detection performance. On the other hand, for the noise-corrupt speech files used in Experiment 2 (Section IV-D), the AMR VAD attained improved detection performance, as expected. For the purpose of blind room acoustic characterization, however, a similar T_{60} measurement performance is attained with either VAD algorithm, thus signaling the robustness of the proposed measures to VAD errors.

Overall, the use of temporal dynamics information for the blind characterization of room acoustics has several advantages over existing schemes. As estimators of T_{60}, the proposed measures are found to be more robust to background noise and, unlike pitch-based methods, are found to be insensitive to speaker gender. Moreover, the proposed measures, based on long-term temporal dynamics, allow for the blind estimation of DRR and, as will be shown in Section V, also the blind estimation of subjective perception of coloration, reverberation tail effects, and overall quality—which are functionalities *not* available with existing algorithms.

V. TOWARDS OBJECTIVE ASSESSMENT OF PERCEIVED REVERBERATION EFFECTS

Subjective listening tests may be used to characterize the subjective perception of room reverberation effects, such as coloration and temporal smearing, as well as to quantify the perceived quality of reverberant speech. Subjective speech quality assessment, however, is laborious and expensive. For the purpose of real-time quality measurement, an automated *objective* speech quality assessment is required.

In [21], a subjectively scored multichannel acoustic reverberation database (MARDY) is used to test the performance of several objective quality measures, namely, segmental SNR, Bark spectral distortion, cepstral distance, and reverberation decay tail. Such measures are termed "intrusive" as the anechoic

TABLE III
PERFORMANCE COMPARISON OF T_{60} ESTIMATORS FOR SPEECH CORRUPTED BY REVERBERATION AND ACOUSTIC NOISE

	SVR_κ			SVR_s						SVR_l					
SNR (dB)	ρ	ϵ	γ	ρ	%	ϵ	%	γ	%	ρ	%	ϵ	%	γ	%
25	0.67	144.1	331.6	0.94	40.3	45.3	68.6	184.4	44.4	0.76	13.4	45.9	68.1	146.6	55.8
20	0.65	192.6	401.1	0.92	41.5	92.2	52.1	273.8	31.7	0.76	16.9	46.5	75.9	151.5	62.2
15	0.63	274.2	498.7	0.88	39.7	185.2	32.5	403.5	19.1	0.75	19.0	46.9	82.9	153.7	69.2
10	0.60	397.2	612.6	0.81	35.0	331.6	16.5	538.3	12.1	0.75	25.0	46.7	88.2	153.9	74.9
5	0.55	551.9	728.2	0.74	34.5	510.1	7.6	669.4	8.1	0.75	36.4	55.6	89.9	154.1	78.8
Average	–	–	–	–	**38.2**	–	**35.4**	–	**23.1**	–	**22.2**	–	**81.0**	–	**68.2**

TABLE IV
PERFORMANCE COMPARISON OF ADAPTED T_{60} ESTIMATORS FOR SPEECH CORRUPTED BY REVERBERATION AND ACOUSTIC NOISE

	$\widetilde{SVR}_\kappa$		$\widetilde{SVR}_s$				$\widetilde{SVR}_l$			
SNR (dB)	ϵ	γ	ϵ	%	γ	%	ϵ	%	γ	%
25	62.8	203.3	32.8	47.8	107.2	47.3	35.0	44.3	117.7	42.1
20	65.6	201.1	39.9	39.2	144.4	28.2	35.2	46.3	114.9	42.9
15	71.1	203.6	46.2	35.0	168.8	17.2	34.8	51.1	120.9	40.6
10	74.8	202.5	52.4	29.9	190.0	6.2	35.2	52.9	119.8	40.8
5	85.7	204.6	56.1	34.5	195.1	4.6	35.9	58.1	126.5	38.2
Average	–	–	–	**37.3**	–	**20.7**	–	**50.5**	–	**40.9**

source signal is used as a reference in the computation process. A nonintrusive measurement, akin to "blind" measurement, does not require a reference signal and constitutes a more challenging approach. In [21], the intrusive measures are tested as estimators of the subjective perception of coloration (COL), reverberation tail effects (RTE), and overall quality (QUAL). It is reported that most measures attain poor correlation with subjective listening quality scores ($\rho \leq 0.40$), and the reverberation decay tail measure attains the highest correlation ($\rho = 0.62$) with respect to RTE. Such poor performance signals the need for more reliable objective quality measures.

Here, long-term temporal dynamics information is investigated for the *nonintrusive* estimation of perceived reverberation effects. As aforementioned, the modulation frequency content for acoustic frequency band j is upper bounded by the bandwidth of the critical-band filter j. Hence, speech signals with different acoustic frequency content, which are subjected to the same quality-degrading reverberation effects, may result in different modulation spectra. In our experiments, an adaptive measure is found to attain superior performance relative to (16). To devise a measure that positively correlates with the subjective quality, an adaptive speech-to-reverberation modulation energy measure (SRMR) is proposed and given by

$$\mathrm{SRMR} = \frac{\sum_{k=1}^{4}\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,k}}{\sum_{k=5}^{K^*}\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,k}}. \tag{21}$$

The measure is adaptive as the upper summation bound K^* in the denominator is dependent on the test speech signal.

In our simulations, K^* is chosen on a per-signal basis and depends on the bandwidth of the lowest gammatone filter for which 90% of the total modulation energy is accounted for. As an example, the plot in Fig. 14 depicts the percentage of modulation energy present per acoustic frequency band for a speech signal, produced by a male speaker, with a reverberation time of 319 ms. As can be seen, 90% of the total energy is obtained below 447 Hz. The bandwidth of the gammatone filter centered at this frequency is approximately 73 Hz. According to Fig. 7, negligible energy at modulation frequency band $k = 8$ is expected, and $K^* = 7$ is chosen.

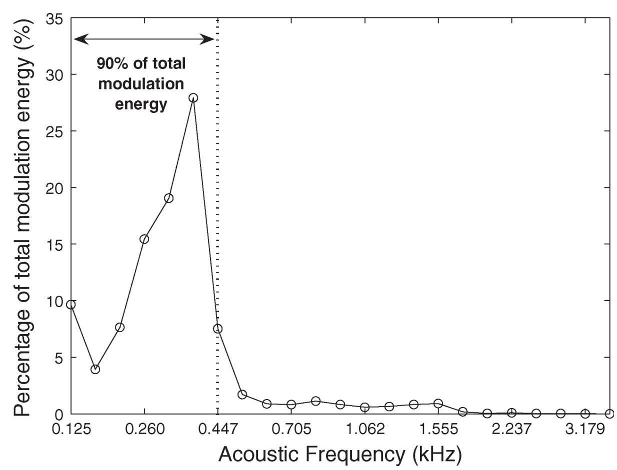

Fig. 14. Percentage of modulation energy, per acoustic frequency band, for a speech signal from a male speaker.

To test the performance of the proposed SRMR measure, a subset of the aforementioned MARDY database is used. The database was developed with room IRs collected with a linear microphone array in an anechoic chamber with reflective panels and absorptive panels installed [21]. The speaker-to-microphone distances varied between 1 and 4 m (1-m increments); the T_{60} values range from 291 to 447 ms.

Reverberant speech was generated with the collected room IRs and anechoic speech from two speakers (one male and one female); additionally, three dereverberation algorithms were used. In our experiments, speech signals are downsampled to 8-kHz sample rate, and only reverberant speech and speech processed by a conventional delay-and-sum beamformer are used.

The proposed SRMR measure is shown to attain $\rho = 0.81$, 0.73, and 0.70 with quality dimensions COL, RTE, and QUAL, respectively. For comparison purposes, objective quality scores obtained from the state-of-the-art nonintrusive ITU-T P.563 standard algorithm [40] attain $\rho = 0.44$, 0.46, and 0.35 with COL, RTE, and QUAL, respectively. Moreover, the state-of-the-art intrusive ITU-T P.862 standard algorithm [44] attains $\rho = 0.69$, 0.81, and 0.73, respectively. As can be seen, the results obtained with the proposed *nonintrusive* measure are comparable with those attained with a state-of-art *intrusive* standard algorithm but with the advantage that a reference speech signal is *not* required.

VI. Conclusion

Temporal dynamics information has been used to design blind estimators of room acoustic parameters, namely, reverberation time and direct-to-reverberation energy ratio. Estimators, based on short- and long-term temporal dynamics information, are shown to outperform a baseline system on reverberant speech data with and without the presence of acoustic background noise. The proposed measures are also shown to reliably estimate the perceived room reverberation effects, such as coloration and reverberation tail effects, in addition to the overall reverberant speech quality.

Acknowledgment

The authors would like to thank D. Gelbart for pointers regarding the recorded room IRs, Dr. L. Couvreur for providing the single-channel IRs, J. Wen for making the MARDY database available, and the anonymous reviewers for their insightful comments.

References

[1] T. Halmrast, "Sound coloration from (very) early reflections," in *Proc. Meeting Acoust. Soc. Amer.*, Jun. 2001. 7 p. [Online]. Available: http://www.akutek.info/Papers/TH_Coloration2001.pdf

[2] P. Rubak, "Coloration in room impulse responses," in *Proc. Joint Baltic-Nordic Acoust. Meet.*, Jun. 2004, pp. 1–14.

[3] L. Couvreur and C. Couvreur, "Blind model selection for automatic speech recognition in reverberant environments," *J. VLSI Signal Process.*, vol. 36, no. 2/3, pp. 189–203, Feb. 2004.

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

[5] 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.

[6] H. Lollmann and P. Vary, "Estimation of the reverberation time in noisy environments," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sep. 2008. 4 p. [Online]. Available: http://www.engr.washington.edu/epp/iwaenc2008/proceedings/contents/papers/9033.p

[7] M. Wu and D. Wang, "A pitch-based method for the estimation of short reverberation time," *Acta Acust./Acustica*, vol. 92, no. 2, pp. 337–339, Mar./Apr. 2006.

[8] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 3701–3704.

[9] N. Gaubitch, D. Ward, and P. Naylor, "Statistical analysis of autoregressive modeling of reverberant speech," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4031–4039, Dec. 2006.

[10] B. Yegnanarayana and P. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[11] E. Habets, N. Gaubitch, and P. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2008, pp. 4577–4580.

[12] 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.

[13] H. Kuttruff, *Room Acoustics*, 4th ed. New York: Elsevier, 2000.

[14] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–951, Apr. 1979.

[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] *Acoustics—Measurement of the Reverberation Time of Rooms With Reference to Other Acoustical Parameters*, 2000. ISO3382.

[18] T. Curtis, "Characterization of room coloration by moments of room spectral response," *J. Acoust. Soc. Amer.*, vol. 58, no. S1, p. S78, Nov. 1975.

[19] J. Jetztz, "Critical distance measurement of rooms from the sound energy spectral envelope," *J. Acoust. Soc. Amer.*, vol. 65, no. 5, pp. 1204–1211, May 1979.

[20] S. Bech, "Timbral aspects of reproduced sound in small rooms," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3539–3549, Jun. 1996.

[21] J. Wen, N. Gaubitch, E. Habets, T. Myatt, and P. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2006, pp. 1–4.

[22] *Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithms*, 2003. ITU-T P.835.

[23] H. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proc. Interspeech*, 2005, pp. 2697–2700.

[24] *ITU-T Software Tool Library*, Aug. 2005. ITU-T Rec. G.191.

[25] *Objective Measurement of Active Speech Level*, 1993. ITU-T P.56.

[26] W. Ward, G. Elko, R. Kubli, and C. McDougald, "The new varechoic chamber at AT&T Bell Labs," in *Proc. Wallace Clement Sabine Centennial Symp.*, 1994, pp. 343–346.

[27] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[28] J. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, pp. 1215–1247, Sep. 1993.

[29] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[30] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[31] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Jun. 2001.

[32] M. Slaney, "An efficient implementation of the Patterson–Holdsworth auditory filterbank," Perception Group, Apple Computer, Inc., Cupertino, CA, Apple Computer Tech. Rep. No. 35, 1993.

[33] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1/2, pp. 103–138, Aug. 1990.

[34] 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, Jun. 1996.

[35] 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.

[36] 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.

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

[38] Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, Mar. 2002.

[39] K. Pearson, "Contribution to the mathematical theory of evolution," *Philos. Trans. Roy. Soc. London A, Math. Phys. Sci.*, vol. 185, pp. 71–110, 1894.

[40] *Single-Ended Method for Objective Speech Quality Assessment in Narrowband Telephony Applications*, 2004. ITU-T P.563.

[41] L. Malfait, J. Berger, and M. Kastner, "P.563—The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.

[42] *A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70*, 1996.

[43] *Adaptive Multi-Rate (AMR) Speech Codec: Voice Activity Detector (VAD), Release 6*, Dec. 2004. 3GPP2 TS 26.094.

[44] *Perceptual Evaluation of Speech Quality: An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, 2001. ITU-T P.862.

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

He is currently with the Bloorview Research Institute, University of Toronto, Toronto, ON, Canada. His research interests include multimedia quality measurement and enhancement, multimedia coding and communications, biomedical signal processing, pattern recognition, and communication theory.

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

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

He has held positions with the Communications Research Centre, Ottawa, Bell Northern Research (Nortel), Ottawa, McGill University, Montreal, QC, Canada, and Illinois Institute of Technology, Chicago, IL. He is currently with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada. His current research interests are multimedia coding and communications, and speech quality measurement and enhancement. He is an Associate Editor of EURASIP *Journal on Audio, Speech, and Music Processing*. He has helped organize conferences on speech coding, image processing, and communications.

Dr. Chan is a Member of the IEEE Signal Processing Society Speech and Language Technical Committee. He was the recipient of a CAREER Award from the U.S. National Science Foundation.