Berkun and Cohen *EURASIP Journal on Advances in Signal Processing*
DOI 10.1186/s13634-015-0233-y

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Microphone array power ratio for quality assessment of reverberated speech

Reuven Berkun* and Israel Cohen

Abstract

Speech signals in enclosed environments are often distorted by reverberation and noise. In speech communication systems with several randomly distributed microphones, involving a dynamic speaker and unknown source location, it is of great interest to monitor the perceived quality at each microphone and select the signal with the best quality. Most of existing approaches for quality estimation require prior information or a clean reference signal, which is unfortunately seldom available. In this paper, a practical non-intrusive method for quality assessment of reverberated speech signals is proposed. Using a statistical model of the reverberation process, we examine the energies as measured by unidirectional elements in a microphone array. By measuring the power ratio, we obtain a measure for the amount of reverberation in the received acoustic signals. This measure is then utilized to derive a blind estimation of the direct-to-reverberation energy ratio in the room. The proposed approach attains a simple, reliable, and robust quality measure, shown here through persuasive simulation results.

Keywords: Microphone arrays; Reverberation; Direct-to-reverberation ratio; Quality diagnosis

1 Introduction

Speech signals in closed-space environments are often distorted by reverberation and noise. In a speech communication application with several distributed microphones, it is often desired to quantify the amount of reverberation of the perceived signal at each sensor, in order to select the channel with the highest quality or with the least reverberation.

Many prior studies dealt with the problem of measuring the amount of reverberation and assessing the quality of degraded acoustic signals. The most common methods are based on quantifying the system characteristics, herein termed system-based. The most well-known measure is the direct-to-reverberation energy ratio (DRR), which estimates the reverberation using the room impulse response (RIR) [1, 2]. Another popular approach is based on comparing the distorted signal to a clean reference version [2–5]. Unfortunately, neither an estimate or knowledge of the room characteristics nor a clean reference is normally available, especially in real-time systems. Moreover, some of these methods obtain low correlation with subjective quality tests [6] and thus cannot be used as reliable reverberation measures. Recently, various methods have been proposed to estimate the reverberation ratio and its properties given the distorted signal alone [7–11].

Direct approaches for measuring the reverberation are based on the signal power or the signal-to-noise evaluation [11]. However, such approaches are suitable only when the power of the noise or the late reverberation is uniform, which is not always true. Many popular methods model the coherence of the direct sound and the reverberation, and estimate the signal-to-diffuse ratio of the signal. Jeub et al. [10] measured the complex spatial coherence between a pair of microphones but restricted the arrival of the direct sound at the broadside direction of the array. In [12], Thiergart, Del Galdo, and Habets estimated the signal-to-diffuse ratio based on the coherence, by using omnidirectional microphones and without direction-of-arrival assumptions. However, when using omnidirectional microphones, the signals are highly correlated at low frequencies, resulting in a high estimation variance. Later works already segregated the diffuse and direct part by using beamforming or directional microphones, and obtained a more robust estimation of the signal-to-diffuse ratio [9, 13]. Yet, they were tested only by an artificial simulation of diffuse and coherent noise fields

*Correspondence: berkun@tx.technion.ac.il
Department of Electrical Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel

and not under real scenarios of reverberant speech signals. Falk, Zheng, and Chan [7] quantified the coloration and reverberation based on an analysis in the modulation spectral domain. Their proposed quality measure was tested with speech signals and was reported to outperform several standard quality and intelligibility measurement algorithms. Goetze et al. [14] compared several measures using subjective listening tests for the assessment of dereverberation algorithms. They showed that most of the signal-based objective measures fail to judge different reverberation distortions, where only one signal-based measure showed high correlation with the subjective rating. They also argued that measures that are based on the impulse response (when available), like the *Clarity* measure (C50) [1], showed much higher correlation with the tests.

In this paper, we address the problem of estimating the quality and the reverberation level of distorted reverberant signals, using the microphone signals alone. Using a directional microphone array, we utilize the directivity pattern of the array elements to segregate the reverberation contribution from the direct signal. We measure the ratio between the energies of the unidirectional sensors and derive an objective signal-based measure for the reverberation quantity. Additionally, we expand this method and derive a reliable blind DRR estimator. Our proposed approach attains a reliable measure with high correlation to various reverberation parameters and outperforms state-of-the-art methods for quality estimation.

This paper is organized as follows. In Section 2, we define the problem. In Sections 3 and 4, we describe the general and the directional-array signal model, respectively. Next, in Section 5, we present our proposed reverberation quantity measure, the directional power ratio, together with a blind estimate for the channel DRR parameter. Simulation and real speech performance results are presented in Section 6. Finally, conclusions are given in Section 7.

2 Problem formulation

We consider a single source of an anechoic speech signal $s(t)$, which convolves with a causal time-invariant room impulse response (RIR) $h(t)$. Then, the measured signal is given by

$$z(t) = \int_{-\infty}^{t} s(\tau)h(t-\tau)d\tau + v(t), \tag{1}$$

where $v(t)$ denotes ambient additive noise, which is assumed to be null at this part of the discussion. The reverberation related to the RIR is divided into two segments [15], $h_d(t)$ and $h_r(t)$, such that

$$h(t) = \begin{cases} h_d(t), & \text{for } 0 \le t < T_r \\ h_r(t), & \text{for } t \ge T_r \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $h_d(t)$ represents the direct path propagation from the source to the microphone, plus some early reflections of the acoustic wave. These reflections usually arrive up to 50 ms after the direct signal and thus are not considered as reverberation. The late part $h_r(t)$ represents later high-order reflections, which are perceived as reverberation. These reflections are incoherent with the direct sound and constitute the main factor for temporal smearing and quality degradation in reverberant rooms. The parameter T_r defines the segmentation of the RIR (where $t = 0$ denotes the arrival time of the direct signal), so that $h_d(t)$ consists of the direct part and some early reflections, while $h_r(t)$ is composed of the late reverberant part.

With (2) we can write the reverberant signal (1) as:

$$\begin{aligned} z(t) &= \int_{t-T_r}^{t} s(\tau)h_d(t-\tau)d\tau + \int_{-\infty}^{t-T_r} s(\tau)h_r(t-\tau)d\tau \\ &= z_d(t) + z_r(t). \end{aligned} \tag{3}$$

The DRR is probably the most well-known objective and unambiguous measure for quantifying the amount of reverberation in rooms, which can be used as well for estimating the perceived signal quality. It is defined as [1, 5]

$$\mathrm{DRR} = \frac{E_d}{E_r} = \frac{\int_0^{T_d} h^2(\tau)d\tau}{\int_{T_d}^{\infty} h^2(\tau)d\tau}, \tag{4}$$

where E_d and E_r are the energies of the direct and reverberated part, respectively, and T_d is the arrival time of the direct sound to the microphone. For measured responses that undergo sampling, T_d is usually chosen to be 8–16 ms larger than the approximate arrival time [15], for higher precision.

Accordingly, our objective is to obtain an estimate for the reverberation amount or alternately for the perceived speech quality, based on the received signals alone (without a priori information of the RIR), and to blindly define an objective criterion for the direct-to-reverberation ratio.

3 Reverberation signal model

Due to the high complexity and low resilience in creating an exact model of the reverberant RIR, it is often described by means of statistical room acoustics (SRA) [15–18]. Originally introduced by Polack [16], and later generalized by Habets [15] (who added the direct-path contribution), the RIR is modeled as a stochastic process, of zero-mean white Gaussian noise, modulated by

an exponentially decaying variance envelope. Accordingly, the direct part can be expressed by

$$h_d(t) = \begin{cases} b_d(t)e^{-\delta t}, & \text{for0} \le t < T_r \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $b_d(t)$ is a white Gaussian noise process, with zero mean and variance of σ_d^2. The decay rate δ is given by [16]

$$\delta = \frac{3\ln 10}{T_{60}}, \tag{6}$$

where T_{60} denotes the reverberation decay time to -60 dB. The late reverberant part is modeled by

$$h_r(t) = \begin{cases} b_r(t)e^{-\delta t}, & \text{for } t \ge T_r \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $b_r(t)$ is a white Gaussian noise process, with zero mean and variance of σ_r^2. The direct and the late parts are uncorrelated, i.e., $\mathbb{E}\{b_d(t)b_r(t+\tau)\} = 0,\ \forall\tau$.

The energy of the RIR would be therefore

$$\mathbb{E}_h\left\{h^2(t)\right\} = \begin{cases} \sigma_d^2 e^{-2\delta t}, & \text{for } 0 \le t < T_r \\ \sigma_r^2 e^{-2\delta t}, & \text{for } t \ge T_r \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $\mathbb{E}_h\{\cdot\}$ denotes expectation over the stochastic process h.

The measured signal energy is obtained by calculating the autocorrelation of $z(t)$ [18]. Relying on the statistical independency of $s(t)$ and $h(t)$, and based on the segmentation described in (3), we get

$$\begin{aligned} \mathbb{E}_z\left\{z^2(t)\right\} &= \int_{t-T_r}^{t}\int_{t-T_r}^{t} \mathbb{E}_s\left\{s(\tau)s(\tau')\right\} \\ &\qquad \mathbb{E}_h\left\{h_d(t-\tau)h_d(t-\tau')\right\} d\tau d\tau' \\ &\quad + \int_{-\infty}^{t-T_r}\int_{-\infty}^{t-T_r} \mathbb{E}_s\left\{s(\tau)s(\tau')\right\} \\ &\qquad \mathbb{E}_h\left\{h_r(t-\tau)h_r(t-\tau')\right\} d\tau d\tau' \\ &= e^{-2\delta t}\int_{t-T_r}^{t} \mathbb{E}_s\left\{s^2(\tau)\right\}\sigma_d^2 e^{2\delta\tau} d\tau \\ &\quad + e^{-2\delta t}\int_{-\infty}^{t-T_r} \mathbb{E}_s\left\{s^2(\tau)\right\}\sigma_r^2 e^{2\delta\tau} d\tau. \end{aligned} \tag{9}$$

Note that the second transition is justified since

$$\begin{aligned} &\mathbb{E}_h\left\{h_\alpha(t-\tau)h_\alpha(t-\tau')\right\} \\ &= \sigma_\alpha^2 e^{-2\delta t}e^{\delta(\tau+\tau')}\delta(\tau-\tau'), \qquad \text{for } \alpha = d, r. \end{aligned} \tag{10}$$

The speech signal is considered stationary over short periods of time, particulary with respect to the reverberation time T_{60}. Accordingly, it is assumed stationary during the measurement period, so that the source autocorrelation can be excluded from the integral in (9), yielding

$$\begin{aligned} \mathbb{E}_z\left\{z^2(t)\right\} &= \lambda_s(t)e^{-2\delta t}\int_{t-T_r}^{t}\sigma_d^2 e^{2\delta\tau} d\tau \\ &\quad + \lambda_s(t)e^{-2\delta t}\int_{-\infty}^{t-T_r}\sigma_r^2 e^{2\delta\tau} d\tau \\ &= \lambda_s(t)\frac{1}{2\delta}\left[\sigma_d^2\left(1-e^{-2\delta T_r}\right) + \sigma_r^2 e^{-2\delta T_r}\right], \end{aligned} \tag{11}$$

where $\lambda_s(t) = \mathbb{E}_s\{s^2(t)\}$ denotes the speech energy at time t, i.e., the current variance of the stochastic quasi-stationary speech process.

Finally, similar to the RIR energy representation, if we choose $T_r = T_d$, based on the generalized statistical model, we can easily deduce the direct and late part energies and express (4) as

$$\mathrm{DRR}_{T_r=T_d} = \frac{E_d}{E_r} = \frac{\sigma_d^2}{\sigma_r^2}\cdot\left(e^{2\delta T_d} - 1\right). \tag{12}$$

4 Directional array response

In this section, we expand the RIR model from Section 3 and examine the response for perception by a unidirectional microphone array.

Let us assume that the reverberant signal impinges on a unidirectional microphone array, rather than a single omnidirectional microphone. Such an array can be composed of several directional microphone elements or alternately by applying beamforming techniques with a few closely spaced omnidirectional microphones [19, 20]. The overall source-to-microphone response can be described as a convolution of the RIR (2) with the response of the corresponding directional microphone. The acoustic response of a directional microphone (or beamformer) is time-invariant and is defined only by the frequency and angle of the arriving signal.

Suppose we have a microphone array with few directional elements, each directed at perpendicular direction. Then, the microphone directed toward the source (denoted with superscript $^{\mathrm{dir}}$), will perceive the direct signal plus the reverberation part. On the other hand, the element directed at the opposite direction (denoted with $^{\mathrm{opp}}$ superscript) will not perceive the direct-path signal, since it arrives mainly from the speaker direction. It will sense the reverberation alone, which is modeled as diffuse noise and hence propagates in all directions incoherently and is being perceived similarly by both elements. An example of such a configuration is illustrated in Fig. 1.

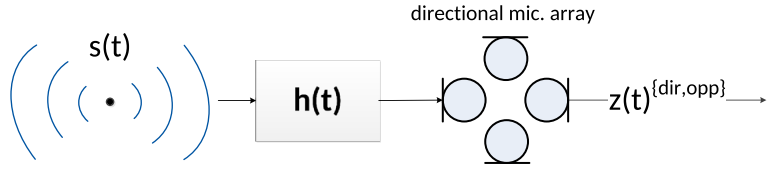

Fig. 1 Directional array configuration. An example of a directional microphone array configuration for reverberant signal propagation acquisition

Let us denote by θ the angle of incidence of the direct signal, and the microphone directional gain at the angle θ by $g^{\text{dir}}(\theta)$. Then, we can express the energy measured by the direct microphone as

$$\mathbb{E}_z\left\{\left[z^{\text{dir}}(t)\right]^2\right\} = \left[g^{\text{dir}}(\theta)\right]^2 \cdot \lambda_s(t)\frac{1}{2\delta}\cdot\sigma_d^2\left(1-e^{-2\delta T_r}\right) + \frac{1}{|\Omega|}\int_\Omega\left[g^{\text{dir}}(\theta')\right]^2 d\theta' \cdot \lambda_s(t)\frac{1}{2\delta}\cdot\sigma_r^2 e^{-2\delta T_r}, \tag{13}$$

where the angular integration is performed across the directional microphone lobe, measuring the uniform power of the diffuse reverberant part. Respectively, the energy measured by the opposite microphone would be

$$\mathbb{E}_z\left\{\left[z^{\text{opp}}(t)\right]^2\right\} = \frac{1}{|\Omega|}\int_\Omega\left[g^{\text{opp}}(\theta')\right]^2 d\theta' \cdot \lambda_s(t)\frac{1}{2\delta}\cdot\sigma_r^2 e^{-2\delta T_r}, \tag{14}$$

where $g^{\text{opp}}(\theta)$ is the opposite microphone angular gain at the angle θ. In light of the aforementioned discussion, we will next derive our proposed approach for measuring the reverberation amount and the reverberant speech quality.

5 Directional power ratio

In the introduced configuration of the directional microphone array, the direct microphone receives both direct-path and late reverberant-part signal, while the opposite microphone receives the reverberant part alone. Therefore, it seems natural to examine the energy ratio of the two:

$$\frac{\mathbb{E}_z\left\{\left[z^{\text{dir}}(t)\right]^2\right\}}{\mathbb{E}_z\left\{\left[z^{\text{opp}}(t)\right]^2\right\}} = \frac{\left[g^{\text{dir}}(\theta)\right]^2 \cdot \lambda_s(t)\frac{1}{2\delta}\cdot\sigma_d^2\left(1-e^{-2\delta T_r}\right) + \frac{1}{|\Omega|}\int_\Omega\left[g^{\text{dir}}(\theta')\right]^2 d\theta' \cdot \lambda_s(t)\frac{1}{2\delta}\cdot\sigma_r^2 e^{-2\delta T_r}}{\frac{1}{|\Omega|}\int_\Omega\left[g^{\text{opp}}(\theta')\right]^2 d\theta' \cdot \lambda_s(t)\frac{1}{2\delta}\cdot\sigma_r^2 e^{-2\delta T_r}}. \tag{15}$$

Let us assume that the directional elements are calibrated, with equal and known gains. If we define

$$\bar{g}^2 = \frac{1}{|\Omega|}\int_\Omega\left[g^{\text{dir}}(\theta')\right]^2 d\theta' = \frac{1}{|\Omega|}\int_\Omega\left[g^{\text{opp}}(\theta')\right]^2 d\theta', \tag{16}$$

then the power ratio (15) is given by

$$\frac{\mathbb{E}_z\left\{\left[z^{\text{dir}}(t)\right]^2\right\}}{\mathbb{E}_z\left\{\left[z^{\text{opp}}(t)\right]^2\right\}} = \frac{\left[g^{\text{dir}}(\theta)\right]^2}{\bar{g}^2}\cdot\left[\frac{\sigma_d^2}{\sigma_r^2}(e^{2\delta T_r}-1)\right]+1 = \frac{\left[g^{\text{dir}}(\theta)\right]^2}{\bar{g}^2}\cdot\text{DRR}+1, \tag{17}$$

where the second transition is immediately inferred from the definition of the DRR (12).

The derived expression for the power ratio may be used as a practical procedure to measure the amount of reverberation and estimate the quality of reverberated signals. In practice, we replace the ensemble averaging $\mathbb{E}_z\{\cdot\}$ with temporal smoothing such as integration over time. Consequently, the measured power ratio (PR) is given by

$$\text{PR}(t) = \frac{P^{\text{dir}}(t)}{P^{\text{opp}}(t)} = \frac{\int_{t-T}^{t}\left[z^{\text{dir}}(\tau)\right]^2 d\tau}{\int_{t-T}^{t}\left[z^{\text{opp}}(\tau)\right]^2 d\tau} = \frac{\left[g^{\text{dir}}(\theta)\right]^2}{\bar{g}^2}\cdot\text{DRR}(t)+1. \tag{18}$$

where the integration should be performed over short intervals of time, in which the speech signal is considered quasi-stationary. Usually the stationarity time-span T is around 20–40 ms [21]. $P^{\{\text{dir,opp}\}}(t)$ denotes the current integrated power as sensed by the direct and opposite microphones, respectively, and $\text{DRR}(t)$ denotes the DRR

at time instance t, with the current speaker and microphone positions. An example of measured power ratio for a reverberated speech signal is given in Fig. 2.

From (18), it is natural to propose a blind estimate for the DRR by

$$\begin{aligned}\text{PR-DRR}(t) &= \frac{\bar{g}^2}{\left[g^{\text{dir}}(\theta)\right]^2}\cdot\left(\frac{P^{\text{dir}}(t)}{P^{\text{opp}}(t)}-1\right)\\ &= \frac{\bar{g}^2}{\left[g^{\text{dir}}(\theta)\right]^2}\cdot\frac{P^{\text{dir}}(t)-P^{\text{opp}}(t)}{P^{\text{opp}}(t)}.\end{aligned} \tag{19}$$

Many popular approaches [8, 10, 22] refer to the DRR and the statistical model as frequency-dependent, due to the frequency dependency of the reflection coefficients and the air absorption coefficient, resulting in a frequency-dependent T_{60} and decay rate δ. Nevertheless, we adopt a frequency-independent model, mainly for simplicity reasons. A frequency-dependent measure achieved similar simulation results to the frequency-independent model, which we would describe next.

6 Experimental results

6.1 In-front simulations

In this section, we evaluate the performance of our proposed method to assess the reverberated signal quality and blindly estimate the DRR, by controlled artificial Matlab simulations. The measure was tested using reverberant human speech, generated by convolving anechoic speech signals with various RIRs. The RIRs were generated by Matlab implementation [23] of the image method [24]. The anechoic speech signal database was composed of 60 male and 60 female speakers, from the TIMIT database [25] (with a sampling rate of $f_s = 8$ kHz). Two types of tests were performed: varying source-microphone distance with fixed reverberation time T_{60}, and varying T_{60} with fixed source-microphone distance. In order to obtain consistent results, we repeated each experiment (for a given distance and T_{60}) by varying the position of the receiver and the source, keeping the source-receiver distance and the reverberation time fixed. We then spatially averaged each set of same-distance and same-T_{60} configuration, to evaluate the ensemble average in a better way [26] and to average over disparities caused by position

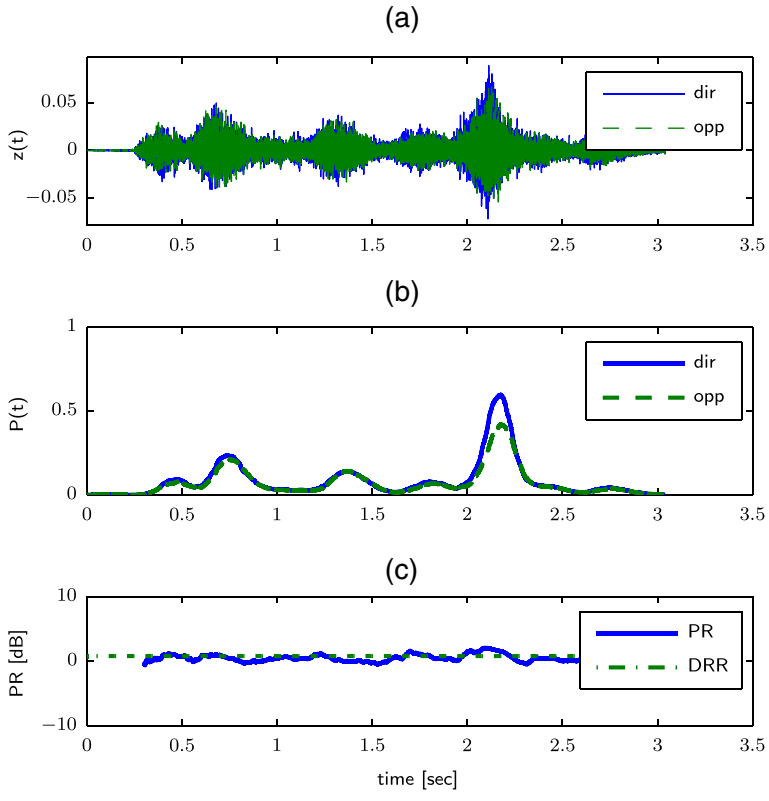

Fig. 2 Measuring the power ratio of a speech signal. Illustration of the reverberation measure for a speech signal, with source-microphone distance = 2 m, $T_{60} = 1$ s. **a** The measured signal $z(t)$ (1) at the direct (*solid line*) and opposite (*dashed line*) microphones. **b** The power $P(t)$ as measured by the direct (*solid line*) and opposite (*dashed line*) microphones. **c** The measured power ratio PR(t) (18) (*solid line*) and $\frac{\left[g^{\text{dir}}(\theta)\right]^2}{\bar{g}^2}\cdot\text{DRR}+1$ (*dashed-dotted line*) as a reference

or local-related effects. We simulated a room of size $5 \times 6 \times 4$ m (length×width×height) [18], with different source and receiver positions, as detailed in Fig. 3. In the simulation, the power ratio integration time T [Eq. (18)] was set to 32 ms. We used four calibrated directional microphones of cardioid directivity, with the microphone array mounted exactly in front of the source. Accordingly, $g^{\mathrm{dir}}(0)$ was set to 1, and $\bar{g}^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}[\frac{1+\cos(\theta')}{2}]^2\, d\theta' = \frac{3}{8}$. The direct microphone was set as the sensor that measured the maximum power. The opposite microphone was set as the sensor in front of it (in 180° angle). Alternately, it can be chosen as the sensor with the minimum power (or by applying localization algorithms).

First, for the quality estimation test, we compared our proposed directional power ratio measure (18) with state-of-the-art quality measures: the speech-to-reverberation modulation energy ratio (SRMR) [7] (using [27]), and the envelope-variance channel selection measure (EV) [11] (uniformly weighted, implemented in Matlab). We computed the Pearson correlation of these approaches with the objective *Clarity* measure C50 [1], which was found to be the most correlative system-based measure with regard to subjective hearing tests [14]. In addition, we calculated correlations to the intrusive quality standard algorithm ITU-T P.563 [28] and the non-intrusive quality algorithm ITU-T P.862 (PESQ) [29] (as done in [7]). Each configuration was first tested with white noise input (of constant temporal variance) [30], and then with reverberated speech signals, where here we calculate the average over all of the speech signals. The correlation results are summarized in Table 1. It details the correlation results of the varying-distance test (with fixed T_{60} and increasing source-microphone distance from 0.25 to 3 m) and the varying-T_{60} test (with fixed distance and increasing T_{60} from 0.1 to 2 s). Note that for the white noise input, the PESQ and the P.563 tests were not performed (they operate only above minimum speech activity level).

The obtained results indicate that the proposed signal-based power ratio approach is highly correlative with the objective system-based C50 quality measure. They also show a relatively high correlation with the PESQ and P.563 scores. Note that the SRMR and the EV obtained even higher correlation with these scores. This can be explained by the fact that they are based on a gammatone [7] and mel-scale [11] subband filtering, like the bark-scale used in the PESQ and P.563. Moreover, note that the standard scores PESQ and P.563 were not developed to measure quality under reverberation conditions and that they attained poor results with respect to subjective tests [14]. However, since they are sensitive to other perceptually important distortions, we use them as additional measures.

Next, we evaluated our proposed blind DRR estimator (19). Its temporal mean was compared with Jeub

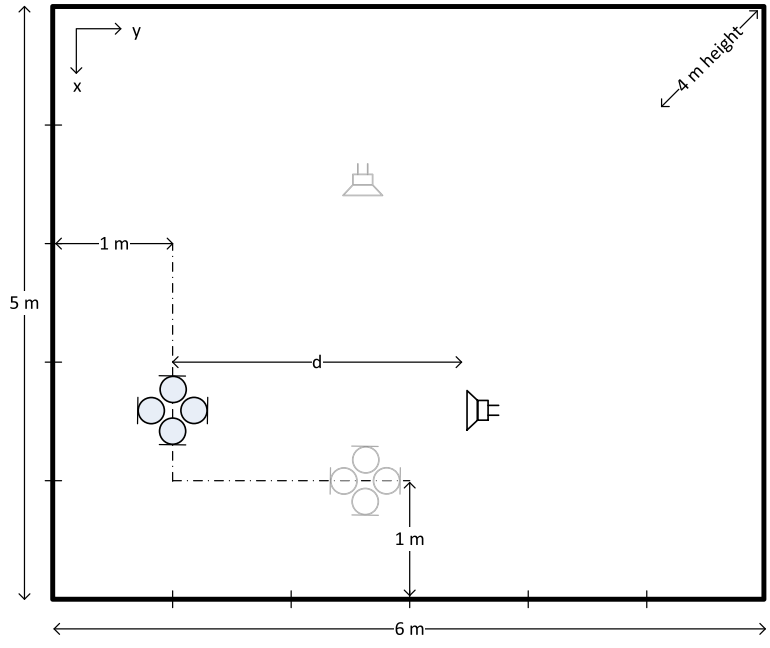

Fig. 3 Simulated room scheme. Illustration of the simulated room of size $5 \times 6 \times 4$ m. For applying spatial averaging, the receiver position was uniformly distributed (with 16 equally spaced locations) along the *L-shaped dashed-dotted line*, keeping the source-microphone distance constant at every configuration. The *source and receiver gray icons* demonstrate such a different location. In addition, the source and receiver positions were swapped to sample more different locations

Table 1 Performance comparison - correlation between the (temporal mean) proposed power ratio (PR) (18), SRMR, and EV values, with Clarity (C50), PESQ, and P.563 algorithms

Input type		White noise	Speech signals		
		Correlation ref.	Correlation ref.		
Test type	Algorithm	C50	C50	PESQ	P. 563
$T_{60} = 0.3$ s,	PR	0.999	0.999	0.911	0.712
vs.	SRMR	−0.27	0.845	0.973	0.934
increasing distance	EV	−0.66	0.931	0.994	0.875
$T_{60} = 0.6$ s,	PR	0.999	0.998	0.970	0.843
vs.	SRMR	0.454	0.967	0.991	0.921
increasing distance	EV	−0.54	0.982	0.991	0.885
Distance = 0.5 m,	PR	0.944	0.951	0.899	0.562
vs.	SRMR	0.392	0.640	0.991	0.873
increasing T_{60}	EV	0.235	0.614	0.984	0.912
Distance = 2 m,	PR	0.973	0.969	0.918	0.674
vs.	SRMR	0.787	0.808	0.998	0.892
increasing T_{60}	EV	−0.33	0.700	0.987	0.958

Table 2 Performance comparison - correlation between the (temporal mean) proposed power ratio-based DRR estimator (PR-DRR) (19) and Jeub et al. CDR-based DRR estimator, with the true DRR measure

Input type		White noise	Speech signals
		Correlation ref.	Correlation ref.
Test type	Algorithm	DRR	DRR
$T_{60} = 0.3$ s,	PR-DRR	0.999	0.999
vs. increasing distance	CDR	0.995	0.992
$T_{60} = 1$ s,	PR-DRR	0.999	0.999
vs. increasing distance	CDR	0.964	0.972
Distance = 0.5 m,	PR-DRR	0.994	0.996
vs. increasing T_{60}	CDR	0.984	0.978
Distance = 2 m,	PR-DRR	0.999	0.999
vs. increasing T_{60}	CDR	0.852	0.913

et al.'s coherent-to-diffuse-based (CDR) blind DRR estimator [10] (using a Matlab code available online). The Pearson correlation coefficient was computed with the DRR measure (4) (with T_d larger than the arrival time by 12 ms). Similarly, first we tested the performance with a white noise input and then with the same 120 reverberant speech sources. The same varying-distance and varying-T_{60} experiments were performed. The corresponding results are summarized in Table 2. In addition, a demonstration of the proposed measure performance vs. distance is illustrated in Fig. 4. It can be inferred that the proposed measure shows high correlation to the theory, with a reliable blind DRR estimation of almost 100 % correlation. As expected, it is inversely proportional to the source-microphone distance and to the reverberation time as well.

Finally, we would like to qualitatively analyze the proposed method performance under noise. In this case, if we add noise $\nu(t) \neq 0$ to the measured signal (1), we obtain a multiplicative bias factor β_{noise} in the proposed DRR estimator (19), such that: $\beta_{\text{noise}} \propto [1 + 2\delta e^{2\delta T_r} \cdot \text{SNR}^{-1}]^{-1}$. Then, we expect that in high signal-to-noise ratio (SNR), the bias would be negligible, whereas in low SNR, the performance would be affected and biased. This type of behavior was observed in a simulation performed over the same 120 speech signals, with a fixed source-microphone distance, a fixed reverberation time, and increasing levels of SNRs from 5 to 25 dB, using additive babble noise [30]. The rest of the parameters were similar to the previous simulations. An example of such simulation is given in Fig. 5, where we measured the absolute difference (AD) between the DRR estimator and the true DRR (in dB), vs. SNR levels [we defined $\text{AD}(x) = 10\log_{10}(x) - 10\log_{10}(\text{DRR})$]. As a reference, it was compared to the AD of Jeub et al.'s CDR-based DRR measure and the true DRR. It seems that even though the proposed approach is sensitive to noise, it still manages to estimate correctly the DRR level based on the signals alone. For very low SNR scenarios, one can first remove the noise by applying speech enhancement methods (e.g., [31, 32]) in a pre-processing stage or estimate the noise variance [33] and use it in the measurement.

6.2 Off main-lobe simulations

In this part, we repeated the experiments above, but instead of varying the array positions in the room and holding the source exactly in front of the receiver, we changed the source-receiver angle. This would give us more interesting and realistic results, since usually the source is not located exactly in front of the microphone, but there is a slight offset from the directional microphone main-lobe axis. In order to achieve a smaller source-receiver angle (such that the direct microphone gain would be higher), one can use more directional elements at every array (or create it using beamforming techniques). However, clearly, this would increase the complexity and the cost of the system. At this part of the simulations, the receiver was positioned in $(x, y, z) = (1, 2, 1)$ m in the room, and the source position was uniformly changed

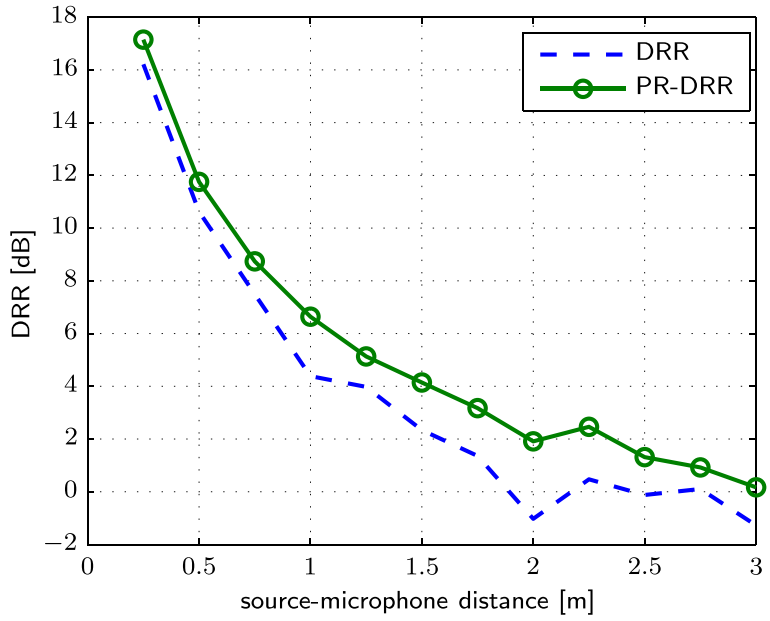

Fig. 4 The proposed DRR estimate vs. distance. Example of the proposed (temporal mean) DRR estimate, PR-DRR (19) [dB] (*solid circled line*), and the true DRR [dB] (*dashed line*), versus source-microphone distance, with fixed $T_{60} = 0.3$ s

along an arch of a fixed radius (for a given distance), creating a $-30°$ to $+30°$ source-receiver angle.

For the varying distance experiment, the range of the source-microphone distance was between 0.25 and 3 m, and for the varying reverberation time, the range of T_{60} was between 0.1 and 1.4 s. Additionally, we repeated the same simulations with swapped source-receiver positions, for a bigger sample space of the experiment.

First, we repeated the quality estimation test and compared the proposed directional power ratio measure (18)

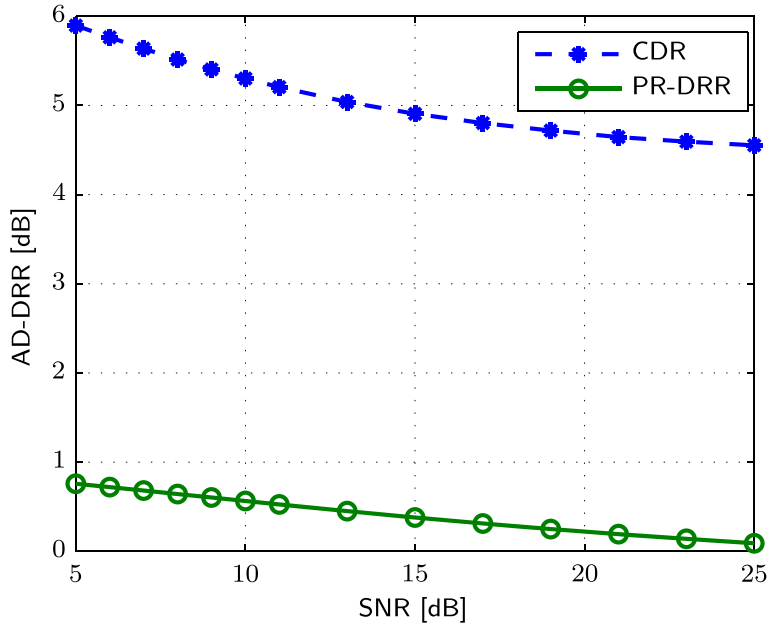

Fig. 5 Performance of the DRR estimate vs. SNR. Example of the AD of the proposed (temporal mean) DRR estimate (19) [dB] (*solid circled line*) compared to the AD of Jeub et al.'s CDR-based (temporal mean) DRR estimate [dB] (*dashed asterisk line*), versus SNR level [dB], with fixed $T_{60} = 0.3$ s and fixed source-microphone distance = 0.5 m.

with the aforementioned reference measures. The corresponding correlation results are detailed in Table 3. If we examine the results (Table 3), we can conclude that the practical scenario (off main-lobe) performance is quite similar to the optimal receiver-in-front case (Table 1).

Next, we repeated the performance comparison of our blind DRR estimator (19) with the CDR blind DRR estimator [10]. Here, too, we performed the same tests with the same speech segments, where instead of varying the receiver position in the room, we changed the source-receiver angle. The correlation results with the true DRR reference are given in Table 4. In addition, an illustration of the proposed DRR measure vs. T_{60} example is shown in Fig. 6. Both the illustration and the correlation results indicate that for the off main-lobe scenario, we obtained promising performance results as well. Moreover, the computed correlation coefficients were as high as the in-front simulations (Table 2), offering a reliable and practical measure.

6.3 Recorded speech experiment

In order to examine our proposed approach in a real environment, we performed speech recordings in a lecture hall of size $15 \times 10 \times 6$ m, using six microphone clusters (with a 3-m spacing between adjacent clusters), each composed of four unidirectional microphone units and each facing 90° apart. For the purpose of analyzing the performance of our proposed measure, we placed the microphone clusters on a line along the hall. The speaker in the experiment moved along the line, advancing from the first array toward the sixth. We divided the speech recordings such that every time the speaker was in front of one array or in between two arrays, a separate speech segment was defined. Then, for every active speech segment, we measured the power ratio (18) and calculated its temporal mean separately. Since we could not restore the reference DRR (or C50) precisely, we chose to demonstrate here a qualitative analysis of the results.

In Fig. 7, we illustrate the power ratio measure at every segment, of all six microphone arrays, vs. the source position. We would expect that the closer the source is to the receiver, the higher the power ratio measure we obtain. Meaning, we expect to have a roughly monotonic increase in the power ratio as the source gets closer to the measuring array, and then a monotonic decrease - as the source moves away from the specific array. Examining Fig. 7, we see that this type of behavior is remarkably noticeable, such that we obtained local maxima exactly in front of the measuring arrays. In addition, we infer that the obtained measure is inversely proportional to the source-microphone distance, as expected. In conclusion, the proposed approach provides a reliable measure, even for non-intuitive scenarios where the reverberation is not proportional to the distance from the source.

Table 3 Off main-lobe performance comparison - correlation between the (temporal mean) proposed power ratio (PR) (18), SRMR, and EV values, with Clarity (C50), PESQ, and P.563 algorithms

Input type		White noise	Speech signals		
		Correlation ref.	Correlation ref.		
Test type	Algorithm	C50	C50	PESQ	P. 563
$T_{60} = 0.3$ s,	PR	0.999	0.998	0.889	0.703
vs.	SRMR	−0.66	0.860	0.984	0.934
increasing distance	EV	−0.65	0.932	0.993	0.871
$T_{60} = 0.6$ s,	PR	0.999	0.999	0.954	0.821
vs.	SRMR	0.399	0.958	0.992	0.938
increasing distance	EV	−0.50	0.981	0.989	0.891
Distance = 0.5 m,	PR	0.946	0.948	0.926	0.573
vs.	SRMR	0.340	0.654	0.986	0.867
increasing T_{60}	EV	0.369	0.640	0.980	0.893
Distance = 2 m,	PR	0.966	0.965	0.947	0.696
vs.	SRMR	0.714	0.809	0.998	0.903
increasing T_{60}	EV	−0.11	0.713	0.976	0.959

The receiver is mounted in $(x, y, z) = (1, 2, 1)$ m, and the source-receiver angle is uniformly distributed from −30° to +30°

Table 4 Off main-lobe performance comparison - correlation between the proposed PR-DRR measure (19) for DRR estimation, and CDR-based DRR estimator [10], with the true DRR measure

Input type		White noise	Speech signals
		Correlation ref.	Correlation ref.
Test type	Algorithm	DRR	DRR
$T_{60} = 0.3$ s,	PR-DRR	0.999	0.999
vs. increasing distance	CDR	0.999	0.998
$T_{60} = 1$ s,	PR-DRR	0.999	0.999
vs. increasing distance	CDR	0.952	0.934
Distance = 0.5 m,	PR-DRR	0.992	0.993
vs. increasing T_{60}	CDR	0.995	0.996
Distance = 2 m,	PR-DRR	0.999	0.999
vs. increasing T_{60}	CDR	0.838	0.745

The microphone array position is distributed in the room creating a source-receiver angle uniformly distributed from −30° to +30°. The receiver position is $(x, y, z) = (1, 2, 1)$ m

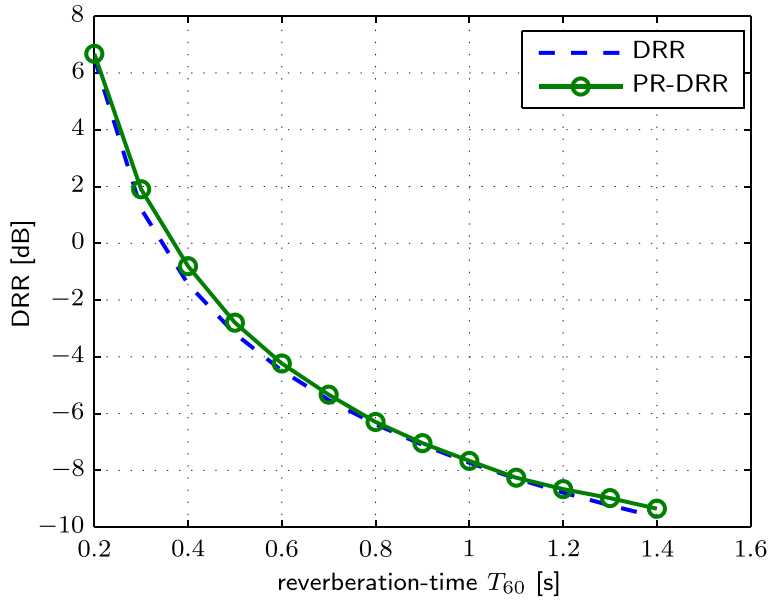

Fig. 6 The proposed DRR estimate vs. T_{60}. The proposed PR-DRR (19) [dB] (*solid circled line*) DRR estimator (temporally averaged), with the true DRR [dB] (*dashed line*), versus the reverberation time T_{60}, for off-main lobe experiment (source-receiver angle is varied from −30° to +30°, and the source-microphone distance is fixed at 2 m)

7 Conclusions

We have proposed a new approach to measure the reverberation ratio for assessment of the acoustic signal quality and mainly for a blind estimation of the direct-to-reverberation ratio of speech signals. Based on a statistical model, we have developed a model for reverberated speech in directional microphones. Supported by this, we measured the power ratio between two opposite unidirectional sensors and segregated the diffuse field influence from the direct signal. This directional-power-ratio

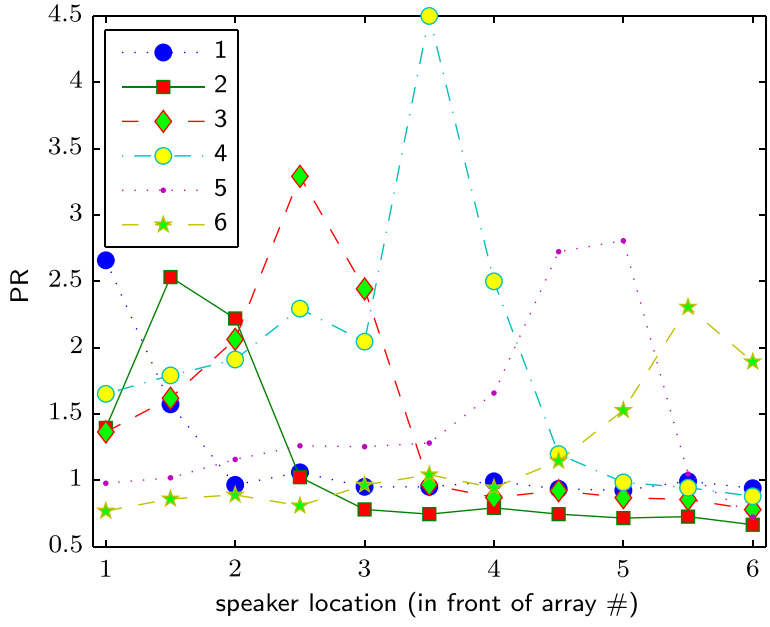

Fig. 7 Recorded speech PR measure vs. source location. The measured (temporally averaged) PR (18) of all microphone arrays (1–6) vs. the source position in the hall. The *x*-axis describes where the source is located relative to the arrays (near an array or between two adjacent arrays). The microphone arrays are located lengthwise along a hall of size 15 × 10 × 6 m, with a 3-m spacing between adjacent arrays

measure was shown to properly estimate the ratio between the direct speech and the reverberation amount, yielding a well-founded signal-based quality measure and a blind DRR estimator. It was compared to various state-of-the-art quality measurement algorithms and DRR measures, and provided reliable results which are highly correlated to the system-based DRR measure. Finally, we tested its performance with some real speech input and managed to show that it can be used as a reliable and robust speech quality measure.

Future work will concentrate on analysis of the optimal directional microphone beampattern and its influence, optimizing and adapting the temporal smoothing to the voice activity level, and combination with de-noising algorithms for integration in real-time quality monitoring systems with distributed microphone arrays.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The authors thank Dr. Baruch Berdugo from Phoenix Audio Technologies for his appreciated assistance with the real-data recordings, and the anonymous reviewers for their constructive comments and useful suggestions.
This research was supported by the Israel Science Foundation (grant no. 1130/11).

Received: 7 January 2015 Accepted: 20 May 2015
Published online: 18 June 2015

References

1. H Kuttruff, *Room Acoustics*. (Taylor & Fanncis Press, New York, USA, 2009)
2. PA Naylor, EAP Habets, in *Speech Dereverberation*, ed. by PA Naylor, EAP Habets, JY-C Wen, and ND Gaubitch. Models, measurement and evaluation (Springer London, UK, 2010)
3. SR Quackenbush, TP Barnwell, MA Clements, *Objective Measures of Speech Quality*. (Yale University Press, Prentice Hall Englewood Cliffs, NJ, 1988)
4. S Wang, A Sekey, A Gersho, An objective measure for predicting subjective quality of speech coders. Selected Areas Commun. IEEE J. **10**(5), 819–829 (1992)
5. PA Naylor, ND Gaubitch, EAP Habets, Signal-based performance evaluation of dereverberation algorithms. J. Electr. Comput. Eng. **2010**, 1–5 (2010)
6. I-T Recommendation. *P. 800: Methods for Subjective Determination of Transmission Quality* (International Telecommunication Union, Geneva, 1996)
7. TH Falk, C Zheng, W-Y Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. Audio Speech Lang. Process. IEEE Trans. **18**(7), 1766–1774 (2010)
8. EAP Habets, S Gannot, I Cohen, Late reverberant spectral variance estimation based on a statistical model. Signal Process. Lett. IEEE. **16**(9), 770–773 (2009)
9. O Thiergart, T Ascherl, EAP Habets, in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference On*. Power-based signal-to-diffuse ratio estimation using noisy directional microphones (Florence Italy, 4–9, May 2014), pp. 7440–7444
10. M Jeub, C Nelke, C Beaugeant, P Vary, in *19th European Signal Processing Conference (EUSIPCO 2011)*. Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals (Barcelona, Spain, Aug. 29–Sep. 2, 2011), pp. 1347–1351
11. M Wolf, C Nadeu, Channel selection measures for multi-microphone speech recognition. Speech Comm. **57**, 170–180 (2014)
12. O Thiergart, G Del Galdo, EAP Habets, in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference On*. Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones (Kyoto, Japan, 25–30 March, 2012), pp. 309–312
13. Y Hioka, K Furuya, K Niwa, Y Haneda, *et al*, in *Acoustic Signal Enhancement, Proceedings of IWAENC 2012, International Workshop On* (VDE, 2012). Estimation of direct-to-reverberation energy ratio based on isotropic and homogeneous propagation model (Aachen, Germany, 4–6 Sept., 2012), pp. 1–4
14. S Goetze, E Albertin, M Kallinger, A Mertins, K-D Kammeyer, in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference On*. Quality assessment for listening-room compensation algorithms (Dallas, TX, 14–19 March, 2010), pp. 2450–2453
15. EAP Habets, Single- and multi-microphone speech dereverberation using spectral enhancement. PhD thesis, Technische Universiteit Eindhoven (2007)
16. J-D Polack, La transmission de l'energie sonore dans les salles. PhD thesis, Universite du Maine, Le Mans, France (1988)
17. K Lebart, J-M Boucher, P Denbigh, A new method based on spectral subtraction for speech dereverberation. Acta Acustica united with Acustica. **87**(3), 359–366 (2001)
18. EAP Habets, in *Speech Dereverberation*, ed. by PA Naylor, ND Gaubitch. Speech dereverberation using statistical reverberation models (Springer London, UK, 2010)
19. S Gannot, I Cohen, in *Springer Handbook of Speech Processing*, ed. by J Benesty, MM Sondhi, and Y Huang. Adaptive beamforming and postfiltering (Springer Berlin, Germany, 2008), pp. 945–978
20. M Brandstein, D Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. (Springer, Berlin, Germany, 2001)
21. JR Deller Jr, JG Proakis, JH Hansen, Discrete-Time Processing of Speech Signals. (Macmillan Pub. Co, New York, 1993)
22. EAP Habets, S Gannot, I Cohen, in *Proc. IEEE Convention of Electrical & Electronics Engineers in Israel (IEEEI)*. Speech dereverberation using backward estimation of the late reverberant spectral variance (Eilat, Israel, 3–5 Dec., 2008), pp. 384–388
23. EAP Habets, Room Impulse Response (RIR) Generator. http://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator. Accessed Dec. 2014
24. JB Allen, DA Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
25. JS Garofolo, *TIMIT: Acoustic-Phonetic Continuous Speech Corpus*. (Linguistic Data Consortium, Philadelphia, 1993)
26. J-M Jot, L Cerveau, O Warusfel, in *Audio Engineering Society Convention 103*. Analysis and synthesis of room reverberation based on a statistical time-frequency model (Audio Engineering Society, 1997)
27. TH Falk, C Zheng, W-Y Chan, SRMR (speech-to-reveberation modulation energy ratio) Matlab Toolbox. http://musaelab.ca/pdfs/SRMRtoolbox.zip. Accessed in Dec. 2014
28. I-T Recommendation, *P. 563: Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*. (International Telecommunication Union, Geneva, 2004)
29. I-T Recommendation, *P. 862, Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. vol, 23*. (International Telecommunication Union, Geneva, 2001)
30. A Varga, HJM Steeneken, Assessment for automatic speech recognition: Ii. noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**(3), 247–251 (1993)
31. I Cohen, B Berdugo, Speech enhancement for non-stationary noise environments. Signal Process. **81**(11), 2403–2418 (2001)
32. I Cohen, S Gannot, in *Springer Handbook of Speech Processing*, ed. by J Benesty, MM Sondhi, and Y Huang. Spectral enhancement methods (Springer Berlin, Germany, 2008), pp. 873–902
33. I Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. Speech Audio Process. IEEE Trans. **11**(5), 466–475 (2003)