Joint Dereverberation and Residual Echo Suppression of Speech Signals in Noisy Environments

Emanuël A. P. Habets, *Member, IEEE*, Sharon Gannot, *Senior Member, IEEE*, Israel Cohen, *Senior Member, IEEE*, and Piet C. W. Sommen

***Abstract*—Hands-free devices are often used in a noisy and reverberant environment. Therefore, the received microphone signal does not only contain the desired near-end speech signal but also interferences such as room reverberation that is caused by the near-end source, background noise and a far-end echo signal that results from the acoustic coupling between the loudspeaker and the microphone. These interferences degrade the fidelity and intelligibility of near-end speech. In the last two decades, postfilters have been developed that can be used in conjunction with a single microphone acoustic echo canceller to enhance the near-end speech. In previous works, spectral enhancement techniques have been used to suppress residual echo and background noise for single microphone acoustic echo cancellers. However, dereverberation of the near-end speech was not addressed in this context. Recently, practically feasible spectral enhancement techniques to suppress reverberation have emerged. In this paper, we derive a novel spectral variance estimator for the late reverberation of the near-end speech. Residual echo will be present at the output of the acoustic echo canceller when the acoustic echo path cannot be completely modeled by the adaptive filter. A spectral variance estimator for the so-called late residual echo that results from the deficient length of the adaptive filter is derived. Both estimators are based on a statistical reverberation model. The model parameters depend on the reverberation time of the room, which can be obtained using the estimated acoustic echo path. A novel postfilter is developed which suppresses late reverberation of the near-end speech, residual echo and background noise, and maintains a constant residual background noise level. Experimental results demonstrate the beneficial use of the developed system for reducing reverberation, residual echo, and background noise.**

Manuscript received August 12, 2007; revised April 25, 2008. Current version published October 17, 2008This work was supported by Technology Foundation STW, applied science division of NWO and the Technology Programme of the Ministry of Economic Affairs, and by the Israel Science Foundation under Grant 1085/05. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sen M. Kuo.

E. A. P. Habets is with the School of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel, and also with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: habetse@eng.biu.ac.il).

S. Gannot is with the School of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel (e-mail: gannot@eng.biu.ac.il).

I. Cohen is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: icohen@ee.technion.ac.il).

P. C. W. Sommen is with the Signal Processing Systems Group, Department of Electrical Engineering, Technische Universiteit Eindhoven, 5600 MB Eindhoven, The Netherlands (e-mail: p.c.w.sommen@tue.nl).

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

Digital Object Identifier 10.1109/TASL.2008.2002071

***Index Terms*—Acoustic echo cancellation (AEC), dereverberation, residual echo suppression.**

I. Introduction

CONVENTIONAL and mobile telephones are often used in a noisy and reverberant environment. When such a device is used in hands-free mode the distance between the desired speaker (commonly called near-end speaker) and the microphone is usually larger than the distance encountered in handset mode. Therefore, the received microphone signal is degraded by the acoustic echo of the far-end speaker, room reverberation and background noise. This signal degradation may lead to total unintelligibility of the near-end speaker. Acoustic echo cancellation is the most important and well-known technique to cancel the acoustic echo [1]. This technique enables one to conveniently use a hands-free device while maintaining high user satisfaction in terms of low speech distortion, high speech intelligibility, and acoustic echo attenuation. The acoustic echo cancellation problem is usually solved by using an adaptive filter in parallel to the acoustic echo path [1]–[4]. The adaptive filter is used to generate a signal that is a replica of the acoustic echo signal. An estimate of the near-end speech signal is then obtained by subtracting the estimated acoustic echo signal, i.e., the output of the adaptive filter, from the microphone signal. Sophisticated control mechanisms have been proposed for fast and robust adaptation of the adaptive filter coefficients in realistic acoustic environments [4], [5]. In practice, there is always residual echo, i.e., echo that is not suppressed by the echo cancellation system. The residual echo results from 1) the deficient length of the adaptive filter, 2) the mismatch between the true and the estimated echo path, and 3) nonlinear signal components.

It is widely accepted that echo cancellers alone do not provide sufficient echo attenuation [3]–[6]. Turbin *et al.* compared three postfiltering techniques to reduce the residual echo and concluded that the spectral subtraction technique, which is commonly used for noise suppression, was the most efficient [7]. In a reverberant environment, there can be a large amount of so-called late residual echo due the deficient length of the adaptive filter. In [6], Enzner proposed a recursive estimator for the short-term power spectral density (PSD) of the late residual echo signal using an estimate of the reverberation time of the room. The reverberation time was estimated directly from the estimated echo path. The late residual echo was suppressed by a spectral enhancement technique using the estimated short-term PSD of the late residual echo signal.

In some applications, like hands-free terminal devices, noise reduction becomes necessary due to the relatively large distance between the microphone and the speaker. The first attempts to develop a combined echo and noise reduction system can be attributed to Grenier *et al.* [8], [9] and to Yasukawa [10]. Both employ more than one microphone. A survey of these systems can be found in [4] and [11]. Beaugeant *et al.* [12] used a single Wiener filter to simultaneously suppress the echo and noise. In addition, psychoacoustic properties were considered in order to improve the quality of the near-end speech signal. They concluded that such an approach is only suitable if the noise power is sufficiently low. In [13], Gustafsson *et al.* proposed two postfilters for residual echo and noise reduction. The first postfilter was based on the log spectral amplitude estimator [14] and was extended to attenuate multiple interferences. The second postfilter was psychoacoustically motivated.

When the hands-free device is used in a noisy reverberant environment, the acoustic path becomes longer and the microphone signal contains reflections of the near-end speech signal as well as noise. Martin and Vary proposed a system for joint acoustic echo cancellation, dereverberation, and noise reduction using two microphones [15]. A similar system was developed by Dörbecker and Ernst in [16]. In both papers, dereverberation was performed by exploiting the coherence between the two microphones as proposed by Allen *et al.* in [17]. Bloom [18] found that this dereverberation approach had no statistically significant effect on intelligibility, even though the measured average reverberation time and the perceived reverberation time were considerably reduced by the processing. It should however be noted that most hands-free devices are equipped with a single microphone.

A single-microphone approach for dereverberation is the application of complex cepstral filtering of the received signal [19]. Bees *et al.* [20] demonstrated that this technique is not useful to dereverberate continues reverberant speech due to so-called segmentation errors. They proposed a novel segmentation and weighting technique to improve the accuracy of the cepstrum. Cepstral averaging then allows to identify the acoustic impulse response (AIR). Yegnanarayana and Murthy [21] proposed another single microphone dereverberation technique in which a time-varying weighting function was applied to the linear prediction (LP) residual signal. The weighing function depends on the signal-to-reverberation ratio (SRR) of the reverberant speech signal and was calculated using the characteristics of the reverberant speech in different SRR regions. Unfortunately, these techniques are not accurate enough in a practical situation and do not fit in the framework of the postfilter which is commonly formulated in the frequency domain. Recently, practically feasible single microphone speech dereverberation techniques have emerged. Lebart proposed a single microphone dereverberation method based on spectral subtraction of the spectral variance of the late reverberant signal [22]. The late reverberant spectral variance is estimated using a statistical model of the AIR. This method was extended to multiple microphones by Habets [23]. Recently, Wen *et al.* presented results obtained from a listening test using the algorithm developed by Habets [24]. These results showed that the algorithm in [23] can significantly increase the subjective speech quality. The methods in [22] and [23] do not require an estimate of the AIR. However, they do require an estimate of the reverberation time of the room which might be difficult to estimate blindly. Furthermore, both methods do not consider any interferences and implicitly assume that the source–receiver distance is larger than the so-called critical distance, which is the distance at which the direct path energy is equal to the energy of all reflections. When the source–receiver distance is smaller than the critical distance the contribution of the direct path results in overestimation of the late reverberant spectral variance. Since this is the case in many hands-free applications, the latter problems need to be addressed.

In this paper, we develop a postfilter which follows the traditional single microphone acoustic echo canceller (AEC). The developed postfilter jointly suppresses reverberation of the near-end speaker, residual echo, and background noise. In Section II, the problem is formulated. The near-end speech signal is estimated using an optimally-modified log spectral amplitude (OM-LSA) estimator which requires an estimate of the spectral variance of each interference. This estimator is briefly discussed in Section III. In addition, we discuss the estimation of the *a priori* signal-to-interference ratio (SIR), which is necessary for the OM-LSA estimator. The late residual echo and the late reverberation spectral variance estimators require an estimate of the reverberation time. A major advantage of the hands-free scenario is that due to the existence of the echo an estimate of the reverberation time can be obtained from the estimated acoustic echo path. In Section IV, we derive a spectral variance estimator for the late residual echo using the same statistical model of the AIR that is used in the derivation of the late reverberant spectral variance estimator. In Section V, the estimation of the late reverberant spectral variance in presence of additional interferences and direct path is investigated. An outline of the algorithm and discussions are presented in Section VI. Experimental results that demonstrate the beneficial use of the developed postfilter are presented in Section VII.

II. Problem Formulation

An AEC with postfilter and a loudspeaker enclosure microphone (LEM) system are depicted in Fig. 1.

The microphone signal is denoted by $y(n)$ and consists of a reverberant speech component $z(n)$, an acoustic echo $d(n)$, and a noise component $v(n)$, where n denotes the discrete time index.

The reverberant speech component $z(n)$ results from the convolution of the AIR, denoted by $\mathbf{a}(n)$, and the anechoic near-end speech signal $s(n)$.

In this paper, we assume that the coupling between the loudspeaker and the microphone can be described by a linear system that can be modeled by a finite-impulse response. The acoustic echo signal $d(n)$ is then given by

$$d(n) = \sum_{j=0}^{N_{\mathrm{h}}-1} h_j(n)\, x(n-j) \tag{1}$$

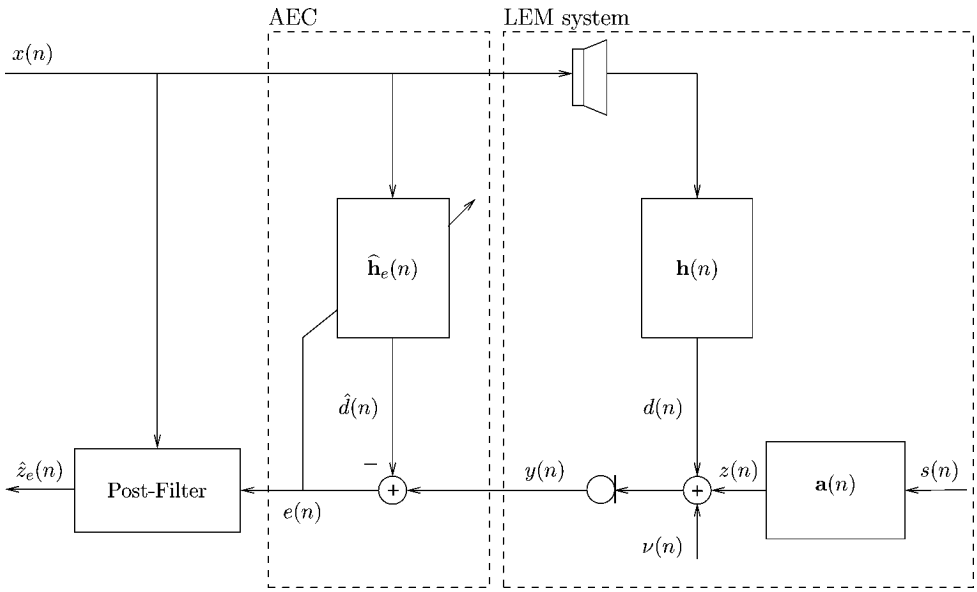

Fig. 1. Acoustic echo canceller with postfilter.

where $h_j(n)$ denotes the jth coefficient of the acoustic echo path at time n, $N_{\rm h}$ is the length of the acoustic echo path, and $x(n)$ denotes the far-end speech signal.

In a reverberant room, the length of the acoustic echo path is approximately given by $f_s\,T_{60}$, where f_s denotes the sampling frequency in Hz, and T_{60} denotes the reverberation time in seconds [2]. At a sampling frequency of 8 kHz, the length of the acoustic echo path in an office with a reverberation time of 0.5 s would be approximately 4000 coefficients. Due to practical reasons, e.g., computational complexity and required convergence time, the length of the adaptive filter, denoted by $N_{\rm e}$, is smaller than N_h. The tail part of the acoustic echo path has a very specific structure. In Section IV, it is shown that this structure can be exploited to estimate the spectral variance of the late residual echo which is related to the part of the acoustic echo path that is not modeled by the adaptive filter. As an example, we use a standard normalized least mean square (NLMS) algorithm to estimate part of the acoustic echo path $\mathbf{h}$. The update equation for the NLMS algorithm is given by

$$\hat{\mathbf{h}}_{\rm e}(n+1) = \hat{\mathbf{h}}_{\rm e}(n) + \mu\,\frac{\mathbf{x}(n)\,e(n)}{\mathbf{x}^T(n)\,\mathbf{x}(n) + \delta_{\rm NLMS}} \tag{2}$$

where $\hat{\mathbf{h}}_{\rm e}(n) = [\hat{h}_{{\rm e},0}(n), \hat{h}_{{\rm e},1}(n), \ldots, \hat{h}_{{\rm e},N_{\rm e}-1}(n)]^T$ is the estimated impulse response vector, μ $(0 < \mu < 2)$ denotes the step-size, $\delta_{\rm NLMS}$ $(\delta_{\rm NLMS} > 0)$ the regularization factor, and $\mathbf{x}(n) = [x(n), \ldots, x(n-N_{\rm e}+1)]^T$ denotes the far-end speech signal state-vector. It should be noted that other, more advanced, algorithms can be used, e.g., recursive least squares (RLS) or affine projection (AP); see, for example, [4] and the references therein. Since $\mathbf{h}_{\rm e}(n)$ is sparse, one might use the improved proportionate NLMS (IPNLMS) algorithm proposed by Benesty and Gay [25]. These advanced techniques are beyond the scope of this paper which focuses on the postfilter.

The estimated echo signal can be calculated using

$$\hat{d}(n) = \sum_{j=0}^{N_{\rm e}-1} \hat{h}_{{\rm e},j}(n)\,x(n-j). \tag{3}$$

The residual echo signal can now be defined as

$$e_r(n) \triangleq d(n) - \hat{d}(n). \tag{4}$$

In general, the residual echo signal $e_r(n)$ is not zero because of the deficient length of the adaptive filter, the system mismatch and nonlinear signal components that cannot be modeled by the linear adaptive filter. While many residual echo suppressions [5], [7] focus on the residual echo that results from the system mismatch, we focus on the late residual echo that results from a deficient length adaptive filter.

Double-talk occurs during periods when the far-end speaker and the near-end speaker are talking simultaneously and can seriously affect the convergence and tracking ability of the adaptive filter. Double-talk detectors and optimal step-size control methods have been presented to alleviate this problem [4], [5], [26], [27]. These methods are out of the scope of this paper. In this paper, we adapt the filter in those periods where only the far-end speech signal is active. These periods have been chosen by using an energy detector that was applied to the near-end speech signal.

The ultimate goal is to obtain an estimate of the anechoic speech signal $s(n)$. While the AEC estimates and subtracts the far-end echo signal a postfilter is used to suppress the residual echo and background noise. The postfilter is usually designed to estimate the reverberant speech signal $z(n)$ or the noisy reverberant speech signal $z(n) + v(n)$. The reverberant speech signal $z(n)$ can be divided into two components: 1) the early speech component $z_{\rm e}(n)$, which consists of a direct sound and early reverberation that is caused by early reflections, and 2) the late reverberant speech component $z_{\rm r}(n)$, which consists of late reverberation that is caused by the reflections that arrive after the early reflections, i.e., late reflections. Independent research [24], [28], [29] has shown that the speech quality and intelligibly are most affected by late reverberation. In addition, it has been shown that the first reflections that arrive shortly after the direct

path usually contribute to speech intelligibility. Therefore, we focus on the estimation of the early speech component $z_{\mathrm{e}}(n)$.

The observed microphone signal $y(n)$ can be written as

$$\begin{aligned} y(n) &= z(n) + d(n) + v(n) \\ &= z_{\mathrm{e}}(n) + z_{\mathrm{r}}(n) + d(n) + v(n). \end{aligned} \tag{5}$$

Using (4) and (5) the error signal $e(n)$ can be written as

$$\begin{aligned} e(n) &= y(n) - \hat{d}(n) \\ &= z_{\mathrm{e}}(n) + z_{\mathrm{r}}(n) + e_r(n) + v(n). \end{aligned} \tag{6}$$

Using the short-time fourier transform (STFT), we have in the time–frequency domain

$$E(l,k) = Z_{\mathrm{e}}(l,k) + Z_{\mathrm{r}}(l,k) + E_{\mathrm{r}}(l,k) + V(l,k) \tag{7}$$

where k represents the frequency bin and l the time frame. In the next section, we show how the spectral component $Z_{\mathrm{e}}(l,k)$ can be estimated.

III. Generalized Postfilter

In this section, the postfilter is developed that is used to jointly suppress late reverberation, residual echo, and background noise. When residual echo and noise are suppressed, Gustafsson *et al.* [30] and Jeannès *et al.* [11] concluded that the best result is obtained by suppressing both interferences together after the AEC. The main advantage of this approach is that the residual echo and noise suppression does not suffer from the existence of a strong acoustic echo component. Furthermore, the AEC does not suffer from the time-varying noise suppression. A disadvantage is that the input signal of the AEC has a low signal-to-noise ratio (SNR). To overcome this problem, algorithms have been proposed where, besides the joint suppression, a noise-reduced signal is used to adapt the echo canceller [31].

Here, a modified version of the OM-LSA estimator [32] is used to obtain an estimate of the spectral component $Z_{\mathrm{e}}(l,k)$. Given two hypotheses, $\mathrm{H}_0(l,k)$ and $\mathrm{H}_1(l,k)$, which indicate, early speech absence and early speech presence, respectively, we have

$$\begin{aligned} &\mathrm{H}_0(l,k) : E(l,k) = Z_{\mathrm{r}}(l,k) + E_{\mathrm{r}}(l,k) + V(l,k), \\ &\mathrm{H}_1(l,k) : E(l,k) = Z_{\mathrm{e}}(l,k) + Z_{\mathrm{r}}(l,k) + E_{\mathrm{r}}(l,k) + V(l,k). \end{aligned}$$

Let us define the spectral variance of the early speech component, the late reverberant speech component, the residual echo signal, and the background noise, as $\lambda_{z_{\mathrm{e}}}$, $\lambda_{z_{\mathrm{r}}}$, λ_{e_r}, and λ_v, respectively. The *a posteriori* SIR is then defined as

$$\gamma(l,k) = \frac{|E(l,k)|^2}{\lambda_{z_{\mathrm{r}}}(l,k) + \lambda_{e_r}(l,k) + \lambda_v(l,k)} \tag{8}$$

and the *a priori* SIR is defined as

$$\xi(l,k) = \frac{\lambda_{z_{\mathrm{e}}}(l,k)}{\lambda_{z_{\mathrm{r}}}(l,k) + \lambda_{e_r}(l,k) + \lambda_v(l,k)}. \tag{9}$$

The spectral variance $\lambda_v(l,k)$ of the background noise $v(n)$ can be estimated directly from the error signal $e(n)$, e.g., by using the method proposed by Martin in [33] or by using the improved minima controlled recursive averaging (IMCRA) algorithm proposed by Cohen [34]. The latter method was used in our experimental study. The spectral variance estimators for $\lambda_{e_r}(l,k)$ and $\lambda_{z_r}(l,k)$ are derived in Sections IV and V, respectively. The *a priori* SIR cannot be calculated directly since the spectral variance $\lambda_{z_{\mathrm{e}}}(l,k)$ is unobservable. Different estimators can be used to estimate the *a priori* SIR, e.g., the decision direct estimator developed by Ephraim and Malah [35] or the recursive causal or noncausal estimators developed by Cohen [36]. In the sequel, the decision directed estimator is used for the estimation of the *a priori* SIR. The decision directed-based estimator is given by [35]

$$\begin{aligned} \hat{\xi}(l,k) = \max\Bigg\{ &\eta \frac{|\hat{Z}_{\mathrm{e}}(l-1,k)|^2}{\lambda(l-1,k)} \\ &+ (1-\eta)\,\max\{\psi(l,k),0\}, \xi_{\min}\Bigg\} \end{aligned} \tag{10}$$

where $\psi(l,k) = \gamma(l,k) - 1$ is the *instantaneous* SIR

$$\lambda(l,k) \triangleq \lambda_{z_{\mathrm{r}}}(l,k) + \lambda_{e_r}(l,k) + \lambda_v(l,k) \tag{11}$$

and $\xi_{\min}$ is a lower-bound on the *a priori* SIR that helps to reduce the amount of musical noise. The weighting factor $\eta (0 \le \eta \le 1)$ controls the tradeoff between the amount of noise reduction and transient distortion introduced into the signal. The weighting factor is commonly chosen close to one, e.g., $\eta = 0.98$. A larger value of η results in a greater reduction of musical noise, but at the expense of attenuated speech onsets and audible modifications of transient components. Although (10) can be used to calculate the total *a priori* SIR, it does not allow to make different tradeoffs for each interference. One can gain more control over the estimation of the *a priori* SIR by estimating it separately for each interference. More information regarding this and combining the separate *a priori* SIRs can be found in Appendix A.

When the early speech component $z_{\mathrm{e}}(n)$ is assumed to be active, i.e., $\mathrm{H}_1(l,k)$ is assumed to be true, the log spectral amplitude (LSA) gain function is used. Under the assumption that $z_{\mathrm{e}}(n)$ and the interference signals are mutually uncorrelated, the LSA gain function is given by [14]

$$G_{\mathrm{H}_1}(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)} \exp\left(\frac{1}{2}\int_{\zeta(l,k)}^{\infty} \frac{e^{-t}}{t}\,\mathrm{d}t\right) \tag{12}$$

where

$$\zeta(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)}\,\gamma(l,k). \tag{13}$$

When the early speech component $z_{\mathrm{e}}(n)$ is assumed to be inactive, i.e., $\mathrm{H}_0(l,k)$ is assumed to be true, a lower-bound $G_{\mathrm{H}_0}(l,k)$ is applied. In many cases, the lower-bound $G_{\mathrm{H}_0}(l,k) = G_{\min}$ is used, where $G_{\min}$ specifies the maximum amount of interference reduction. To avoid speech distortions $G_{\min}$ is usually set between -12 and -18 dB. However, in practice the residual echo and late reverberation needs to be

reduced more than 12–18 dB. Due to the constant lower-bound the residual echo will still be audible in some time–frequency frames [32]. Therefore, $G_{\mathrm{H}_0}(l,k)$ should be chosen such that the residual echo and the late reverberation is suppressed down to residual background noise floor given by $G_{\min}V(l,k)$. When $G_{\mathrm{H}_0}(l,k)$ is applied to those time-frequency frames where hypothesis $\mathrm{H}_0(l,k)$ is assumed to be true, we obtain

$$\hat{Z}_{\mathrm{e}}(l,k) = G_{\mathrm{H}_0}(l,k)\,(Z_{\mathrm{r}}(l,k) + E_{\mathrm{r}}(l,k) + V(l,k)). \quad (14)$$

The desired solution for $\hat{Z}_{\mathrm{e}}(l,k)$ is

$$\hat{Z}_{\mathrm{e}}(l,k) = G_{\min}V(l,k). \quad (15)$$

The least squares solution for $G_{\mathrm{H}_0}(l,k)$ is obtained by minimizing

$$\mathcal{E}\{|G_{\mathrm{H}_0}(l,k)\,(Z_{\mathrm{r}}(l,k)+E_{\mathrm{r}}(l,k)+V(l,k)) - G_{\min}V(l,k)|^2\}.$$

Assuming that all interferences are mutually uncorrelated, we obtain

$$G_{\mathrm{H}_0}(l,k) = G_{\min}\frac{\hat{\lambda}_v(l,k)}{\hat{\lambda}_{z_{\mathrm{r}}}(l,k) + \hat{\lambda}_{e_{\mathrm{r}}}(l,k) + \hat{\lambda}_v(l,k)}. \quad (16)$$

The results of an informal listening test showed that the obtained residual interference was more pleasant than the residual interference that was obtained using $G_{\mathrm{H}_0}(l,k) = G_{\min}$.

The OM-LSA spectral gain function, which minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence probability denoted by $p(l,k)$ [37]. Hence, the modified OM-LSA gain function is given by

$$G_{\mathrm{OM-LSA}}(l,k) = \{G_{\mathrm{H}_1}(l,k)\}^{p(l,k)}\,\{G_{\mathrm{H}_0}(l,k)\}^{1-p(l,k)}. \quad (17)$$

The speech presence probability $p(l,k)$ was efficiently estimated using the method proposed by Cohen in [37].

The spectral speech component $Z_{\mathrm{e}}(l,k)$ of the early speech component can now be estimated by applying the OM-LSA spectral gain function to each spectral component $E(l,k)$, i.e.,

$$\hat{Z}_{\mathrm{e}}(l,k) = G_{\mathrm{OM-LSA}}(l,k)\,E(l,k). \quad (18)$$

The early speech component $\hat{z}_{\mathrm{e}}(n)$ can then be obtained using the inverse STFT and the weighted overlap-add method [38].

IV. Late Residual Echo Spectral Variance Estimation

In Fig. 2, a typical AIR and its energy decay curve (EDC) are depicted. The EDC is obtained by backward integration of the squared AIR [39] and is normalized with respect to the total energy of the AIR. In Fig. 2, we can see that the tail of the AIR exhibits an exponential decay and that the tail of the EDC exhibits a linear decay.

Enzner [6] proposed a recursive estimator for the short-term PSD of the late residual echo which is related to $\mathbf{h}_{\mathrm{r}}(n) = [h_{N_{\mathrm{e}}}(n), \ldots, h_{N_{\mathrm{h}}-1}(n)]^T$. The recursive estimator exploits the fact that the exponential decay rate of the AIR is directly related to the reverberation time of the room, which can

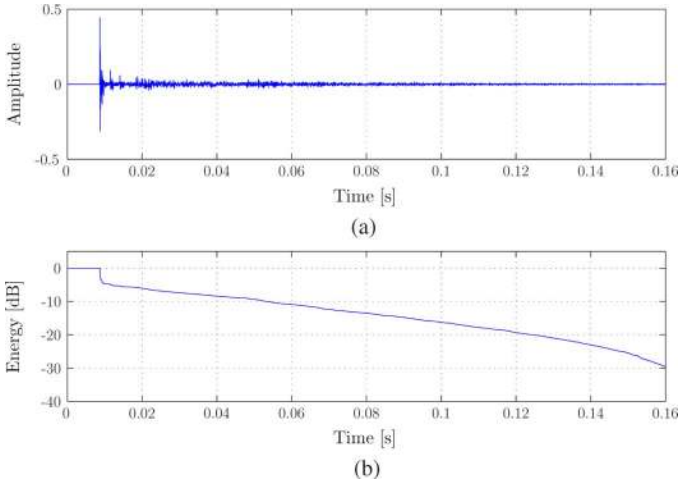

Fig. 2. Typical acoustic impulse response and related energy decay curve. (a) Typical acoustic impulse response. (b) Normalized energy decay curve of (a).

be estimated using the estimated echo path $\hat{\mathbf{h}}_{\mathrm{e}}$. Additionally, the recursive estimator requires a second parameter that specifies the initial power of the late residual echo.

In this section, an essentially equivalent recursive estimator is derived, starting in the time-domain rather than directly in the frequency-domain as in [6]. Enzner applied a direct fit to the log-envelope of the estimated echo path to estimate the required parameters, viz, the reverberation time and the initial power of the late residual echo, which are both assumed to be frequency independent. It should, however, be noted that these parameters are usually frequency dependent [40]. Furthermore, in many applications, the distance between the loudspeaker and the microphone is small, which results in a strong direct echo. The presence of a strong direct echo results in an erroneous estimate of both the reverberation time and the initial power (cf. [41]). Therefore, we propose to apply a linear curve fit to part of the EDC, which exhibits a smoother decay ramp. Details regarding the estimation of the reverberation time $T_{60}(k)$ and the initial power can be found in Appendices B and C, respectively.

Using a statistical reverberation model and the estimated reverberation time the spectral variance of the late residual echo can be estimated. In the sequel, we assume that $N_{\mathrm{h}} = \infty$. The late residual echo $e_{\mathrm{r}}(n)$ can then be expressed as

$$e_{\mathrm{r}}(n) = \sum_{j=0}^{\infty} h_{\mathrm{r},j}(n)\,x_{\mathrm{r}}(n-j) \quad (19)$$

where $x_{\mathrm{r}}(n) = x(n-N_{\mathrm{e}})$.

The spectral variance of $e_{\mathrm{r}}(n)$ is defined as

$$\lambda_{e_{\mathrm{r}}}(l,k) \triangleq \mathcal{E}\{|E_{\mathrm{r}}(l,k)|^2\}. \quad (20)$$

In the STFT domain, we can express $E_{\mathrm{r}}(l,k)$ as [42]

$$E_{\mathrm{r}}(l,k) = \sum_{k'=0}^{N_{\mathrm{DFT}}-1}\sum_{i=0}^{\infty} H_{\mathrm{r},i}(l,k,k')\,X\left(l-i-\frac{N_{\mathrm{e}}}{R},k'\right) \quad (21)$$

where R denotes the number of samples between two successive STFT frames, N_{DFT} denotes the length of the discrete fourier

transform (DFT), $H_{\mathrm{r},i}(l,k,k')$ may be interpreted as the response to an impulse $\delta(l-i,k-k')$ in the time–frequency domain (note that the impulse response is translation varying in the time- and frequency-axis), and i denotes the coefficient index. Note that N_e should be chosen such that N_e/R is an integer value.

Polack proposed a statistical reverberation model where the AIR is described as one realization of a nonstationary process [43]. The model is given by $h(n) = b(n)\,\mathrm{e}^{-\rho n/f_\mathrm{s}}\,\forall\, n \geq 0$, where $b(n)$ is a white Gaussian noise with zero mean, and ρ denotes the decay rate which is related to the reverberation time T_{60} of the room. Using this model, it can be shown that

$$\mathcal{E}\{H_{\mathrm{r},i}(l,k,k')\,H_{\mathrm{r},i+\tau}(l,k,k')\} = 0 \quad \forall \tau \neq 0. \tag{22}$$

Using statistical room acoustics, it can be shown that correlation between different frequencies drops rapidly with increasing $|k-k'|$ [44]. Therefore, the correlation between the cross-bands $k \neq k'$ can be neglected, i.e.,

$$\mathcal{E}\{H_{\mathrm{r},i}(l,k,k')\,H_{\mathrm{r},i}(l,k,k')\} = 0 \quad \forall k \neq k'. \tag{23}$$

Using (20)–(23), we can express $\lambda_{e_\mathrm{r}}(l,k)$ as

$$\begin{aligned}\lambda_{e_\mathrm{r}}(l,k) &= \mathcal{E}\left\{\sum_{k'=0}^{K}\sum_{i=0}^{\infty}|H_{\mathrm{r},i}(l,k,k')|^2 \times \left|X\left(l-i-\frac{N_\mathrm{e}}{R},k'\right)\right|^2\right\} \\ &= \sum_{i=0}^{\infty}\mathcal{E}\{|H_{\mathrm{r},i}(l,k)|^2\} \times \mathcal{E}\left\{\left|X\left(l-i-\frac{N_\mathrm{e}}{R},k\right)\right|^2\right\}\end{aligned} \tag{24}$$

where $H_{\mathrm{r},i}(l,k) \triangleq H_{\mathrm{r},i}(l,k,k)$.

Using Polack's statistical reverberation model, the energy envelope of $H_{\mathrm{r},i}(l,k)$ can be expressed as

$$\mathcal{E}\{|H_{\mathrm{r},i}(l,k)|^2\} = c(l-i,k)\,\alpha^i(k) \tag{25}$$

where $c(l-i,k)$ denotes the initial power of the late residual echo in the kth subband at time $(l-i)R$, $\alpha(k) = e^{-2\rho(k)R/f_\mathrm{s}}$ $(0 \leq \alpha(k) < 1)$, and $\delta(k)$ denotes the frequency dependent decay rate. The decay rate $\rho(k)$ is related to the frequency dependent reverberation time $T_{60}(k)$ through

$$\rho(k) \triangleq \frac{3\ln(10)}{T_{60}(k)}. \tag{26}$$

Using (25) and the fact that $\lambda_x(l,k) = \mathcal{E}\{|X(l,k)|^2\}$, we can rewrite (24) as

$$\lambda_{e_\mathrm{r}}(l,k) = \sum_{i=0}^{\infty} c(l-i,k)\,\alpha^i(k)\,\lambda_x\left(l-i-\frac{N_\mathrm{e}}{R},k\right). \tag{27}$$

By using $i' = l-i$ and extracting the last term of the summation in (27), we can derive a recursive expression for $\lambda_{e_\mathrm{r}}(l,k)$ such that only the spectral variance $\lambda_x(l-N_\mathrm{e}/R,k)$ is required, i.e.,

$$\begin{aligned}\lambda_{e_\mathrm{r}}(l,k) &= \sum_{i'=-\infty}^{l} c(i',k)\,\alpha^{l-i'}(k)\,\lambda_x\left(i'-\frac{N_\mathrm{e}}{R},k\right) \\ &= \sum_{i'=-\infty}^{l-1} c(i',k)\,\alpha^{l-i'}(k)\,\lambda_x\left(i'-\frac{N_\mathrm{e}}{R},k\right) \\ &\quad + c(l,k)\,\lambda_x\left(l-\frac{N_\mathrm{e}}{R},k\right) \\ &= \alpha(k)\lambda_{e_\mathrm{r}}(l-1,k) + c(l,k)\,\lambda_x\left(l-\frac{N_\mathrm{e}}{R},k\right).\end{aligned} \tag{28}$$

Given an estimate of the reverberation time $\mathrm{T}_{60}(k)$ (see Appendix B), an estimate of the exponential decay rate $\rho(k)$ is obtained using (26). Using the initial power $\tilde{c}(l,k)$ (see Appendix C), we can now estimate $\lambda_{e_\mathrm{r}}(l,k)$ using

$$\hat{\lambda}_{e_\mathrm{r}}(l,k) = \mathrm{e}^{-2\hat{\rho}(k)\frac{R}{f_\mathrm{s}}}\hat{\lambda}_{e_\mathrm{r}}(l-1,k) + \tilde{c}(l,k)\,\hat{\lambda}_x\left(l-\frac{N_\mathrm{e}}{R},k\right) \tag{29}$$

where $\hat{\lambda}_x(l,k)$ can be calculated using

$$\hat{\lambda}_x(l,k) = \eta_x\,\hat{\lambda}_x(l-1,k) + (1-\eta_x)\,|X(l,k)|^2 \tag{30}$$

where $\eta_x\,(0 \leq \eta_x < 1)$ denotes the smoothing parameter. In general, a value $\eta_x = \exp(-R/(f_s\,12\,\mathrm{ms}))$ yields good results.

V. Late Reverberant Spectral Variance Estimation

In this section, we develop an estimator for the late reverberant spectral variance of the near-end speech signal $z(n)$.

In [22], it was shown that, using Polack's statistical room impulse response model [43], the spectral variance of the late reverberant signal can be estimated directly from the spectral variance of the reverberant signal using

$$\hat{\lambda}_{z_\mathrm{r}}(l,k) = \alpha^{\frac{N_\mathrm{r}}{R}}(k)\hat{\lambda}_z\left(l-\frac{N_\mathrm{r}}{R},k\right). \tag{31}$$

The parameter N_r (in samples) controls the time instance (measured with respect to the arrival time of the direct sound) where the late reverberation starts and is chosen such that N_r/R is an integer value. In general, N_r is chosen between 20 and 60 ms. While 20–35 ms yields good results when the SRR is larger than 0 dB, a value larger than 35 ms is preferred when the SRR is smaller than 0 dB.

In [22] and [23], it was implicitly assumed that the energy of the direct path was small compared to the reverberant energy. However, in many practical situations, the source is close to the microphone, and the contribution of the spectral variance that is related to the direct path is larger than the spectral variance that is related to all reflections. When the contribution of the direct path is ignored, the late reverberant spectral variance will be overestimated. Since this overestimation results in a distortion

of the early speech component, we need to compensate for the spectral variance that related to the direct path.

In Section V-A, it is shown how an estimate of the spectral variance of the reverberant spectral component $Z(l,k)$ can be obtained which is required to calculate (31). In Section V-B, a method is developed to compensate for the spectral variance contribution that is related to the direct path.

A. Reverberant Spectral Variance Estimation

The spectral variance of the reverberant spectral component $Z(l,k)$, i.e., $\lambda_z(l,k)$, is estimated by minimizing

$$\mathcal{E}\{(|Z(l,k)|^2 - |\hat{Z}(l,k)|^2)^2\} \tag{32}$$

where $\hat{Z}(l,k) = G_{\mathrm{SP}}(l,k)E(l,k)$.

As shown in [45] this leads to the following spectral gain function:

$$G_{\mathrm{SP}}(l,k) = \sqrt{\frac{\xi_{\mathrm{SP}}(l,k)}{1+\xi_{\mathrm{SP}}(l,k)}\left(\frac{1}{\gamma_{\mathrm{SP}}(l,k)} + \frac{\xi_{\mathrm{SP}}(l,k)}{1+\xi_{\mathrm{SP}}(l,k)}\right)} \tag{33}$$

where

$$\xi_{\mathrm{SP}}(l,k) = \frac{\lambda_z(l,k)}{\lambda_{e_r}(l,k) + \lambda_v(l,k)} \tag{34}$$

and

$$\gamma_{\mathrm{SP}}(l,k) = \frac{|E(l,k)|^2}{\lambda_{e_r}(l,k) + \lambda_v(l,k)} \tag{35}$$

denote the *a priori* and *a posteriori* SIRs, respectively. The *a priori* SIR is estimated using the decision directed method. An estimate of the spectral variance of the reverberant speech signal $z(n)$ is then obtained by

$$\hat{\lambda}_z(l,k) = \eta_z\,\hat{\lambda}_z(l,k) + (1-\eta_z)(G_{\mathrm{SP}}(l,k))^2|E(l,k)|^2 \tag{36}$$

where η_z $(0 \le \eta_z < 1)$ denotes the smoothing parameter. In general, a value $\eta_z = \exp(-R/(f_s\,80\ \mathrm{ms}))$ yields good results.

B. Direct Path Compensation

The energy envelope of the AIR of the system between $s(n)$ and $y(n)$ can be modeled using the exponential decay rate of the AIR, and the energy of the direct path and the energy of all reflections in the kth subband, denoted by $Q_{\mathrm{d}}(k)$ and $Q_{\mathrm{r}}(k)$, respectively. For the kth subband we then obtain in the z-transform domain

$$\tilde{A}_k(\mathrm{z}) = Q_{\mathrm{d}}(k) + Q_{\mathrm{r}}(k)\,\tilde{R}_k(\mathrm{z}) \tag{37}$$

where $\tilde{R}_k(\mathrm{z})$ denotes the normalized energy envelope of the reverberant part of the AIR, which starts at $l=1$, i.e.,

$$\tilde{R}_k(\mathrm{z}) = \frac{1-\alpha(k)}{\alpha(k)}\sum_{l=1}^{\infty}(\alpha(k))^l\,\mathrm{z}^{-l}. \tag{38}$$

Note that $\sum_{l=1}^{\infty}(\alpha(k))^l$ equals $\alpha(k)/(1-\alpha(k))$. By expanding the series in (38), we obtain

$$\tilde{R}_k(\mathrm{z}) = \frac{1-\alpha(k)}{\alpha(k)}\,\frac{\alpha(k)\mathrm{z}^{-1}}{1-\alpha(k)\,\mathrm{z}^{-1}}. \tag{39}$$

To eliminate the contribution of the energy of the direct path in $\hat{\lambda}_z(l,k)$, we apply the following filter to $\hat{\lambda}_z(l,k)$:

$$F_k(\mathrm{z}) = \frac{Q_{\mathrm{r}}(k)\,\tilde{R}_k(\mathrm{z})}{Q_{\mathrm{d}}(k) + Q_{\mathrm{r}}(k)\,\tilde{R}_k(\mathrm{z})}. \tag{40}$$

We now define $\kappa(k)$, which is inversely proportional to the direct to reverberation ratio (DRR) in the kth subband, as

$$\kappa(k) \triangleq \frac{1-\alpha(k)}{\alpha(k)}\,\frac{Q_{\mathrm{r}}(k)}{Q_{\mathrm{d}}(k)}. \tag{41}$$

In this paper, it is assumed that $\kappa(k)$ is known *a priori*. In practice, $\kappa(k)$ could be estimated online, by minimizing $\mathcal{E}\{(|Z(l,k)|^2 - \lambda'_z(l,k))^2\}$ during the so-called free-decay of the reverberation in the room. Recently, an adaptive estimation technique was proposed in [46].

Using the normalized energy envelope $\tilde{R}_k(\mathrm{z})$, as defined in (39), (40), and (41), we obtain

$$F_k(\mathrm{z}) = \frac{\alpha(k)\,\kappa(k)\,\mathrm{z}^{-1}}{1-\alpha(k)\,(1-\kappa(k))\,\mathrm{z}^{-1}}. \tag{42}$$

Using the difference equation related to the filter in (42), we obtain an estimate of the reverberant spectral variance with compensation of the direct path energy, i.e.,

$$\hat{\lambda}'_z(l,k) = \alpha(k)\,(1-\kappa(k))\,\hat{\lambda}'_z(l-1,k) + \alpha(k)\,\kappa(k)\,\hat{\lambda}_z(l-1,k). \tag{43}$$

To ensure the stability of the filter $|\alpha(k)(1-\kappa(k))| < 1$. Furthermore, from a physical point of view it is important that only the source can increase the reverberant energy in the room, i.e., the contribution of $\hat{\lambda}'_z(l-1,k)$ to $\hat{\lambda}'_z(l,k)$ should always be smaller than, or equal to, $\alpha(k)$. Therefore, we require that $0 < \kappa(k) \le 1$.

If $Q_{\mathrm{d}}(k) \gg Q_{\mathrm{r}}(k)$, i.e., $\kappa(k)$ is small, $\hat{\lambda}'_z(l,k)$ mainly depends on $\alpha(k)\hat{\lambda}'_z(l-1,k)$. If $Q_{\mathrm{d}}(k) \ll Q_{\mathrm{r}}(k)$, we reach the upper-bound of $\kappa(k)$, i.e., $\kappa(k)=1$, and $\hat{\lambda}'_z(l,k)$ is equal to

$$\hat{\lambda}'_z(l,k) = \alpha(k)\,\hat{\lambda}_z(l-1,k). \tag{44}$$

The late reverberant spectral variance $\hat{\lambda}_{z_{\mathrm{r}}}(l,k)$ with direct path compensation (DPC) can now be obtained by using $\hat{\lambda}'_z(l,k)$, i.e.,

$$\hat{\lambda}_{z_{\mathrm{r}}}(l,k) = \alpha^{\frac{N_{\mathrm{r}}}{R}-1}(k)\,\hat{\lambda}'_z\left(l - \frac{N_{\mathrm{r}}}{R} + 1, k\right). \tag{45}$$

By substituting (44) in (45), we obtain the estimator (31) that was proposed in [22].

VI. Algorithm Outline and Discussion

In the previous sections, a novel postfilter that is used for the joint suppression of residual echo, late reverberation, and back-

ground noise was developed. This postfilter is used in conjunction with a standard AEC. The steps of a complete algorithm, that includes the estimation of the echo path, the estimation of the spectral variance of the interferences and the OM-LSA gain function, are summarized in Algorithm 1.

Algorithm 1 Summary of the developed algorithm.

1) **Acoustic Echo Cancellation**: Update the adaptive filter $\hat{h}_{\mathrm{e}}(n)$ using (2) and calculate $\hat{d}(n)$ using (3).
2) **Estimate Reverberation Time**: Estimate $T_{60}(k)$ as described in Appendix B.
3) **STFT**: Calculate the STFT of $e(n) = y(n) - \hat{d}(n)$ and $x(n)$.
4) **Estimate Background Noise**: Estimate $\lambda_v(l,k)$ using [34].
5) **Estimate Late Residual Echo Spectral Variance**: Calculate $\tilde{c}(l,k)$ using (57) and $\hat{\lambda}_{e_{\mathrm{r}}}(l,k)$ using (29).
6) **Estimate Late Reverberant Spectral Variance**: Calculate $G_{\mathrm{SP}}(l,k)$ using (33)–(35). Estimate $\lambda_z(l,k)$ using (36), and calculate $\hat{\lambda}_{z_{\mathrm{r}}}(l,k)$ using (43) and (45).
7) **Postfilter**:
   a) Calculate the *a posteriori* using (8) and *a priori* SIR using (51)–(54).
   b) Calculate the speech presence probability $p(l,k)$ [37].
   c) Calculate the gain function $G_{\mathrm{OM-LSA}}(l,k)$ using (16) and (17).
   d) Calculate $\hat{Z}_{\mathrm{e}}(l,k)$ using (18).
8) **Inverse STFT**: Calculate the output $\hat{z}_{\mathrm{e}}(n)$ by applying the inverse STFT to $\hat{Z}_{\mathrm{e}}(l,k)$.

In this paper, we used a standard NLMS algorithm to update the adaptive filter. Due to the choice of N_{e} ($N_{\mathrm{e}} < N_{\mathrm{h}}$), the length of the adaptive filter is deficient. When the far-end signal $x(n)$ is not spectrally white, the filter coefficients are biased [47], [48]. However, the filter coefficients, that are mostly affected, are in the tail region. Accordingly, this problem can be partially solved by slightly increasing the value of N_{e} and calculating the output using the original N_{e} coefficients of the filter. Alternatively, one could use a, possibly adaptive, prewhitening filter [2], or another adaptive algorithm like AP or RLS.

An estimate of the reverberation time is required for the late residual echo spectral variance and late reverberant spectral variance estimation. In some applications, e.g., conference systems, this parameter may be determined using a calibration step. In this paper, we proposed a method to estimate the reverberation time online using the estimated filter $\hat{\mathbf{h}}_{\mathrm{e}}$, assuming that the convergence of the filter $\hat{\mathbf{h}}_{\mathrm{e}}$ is sufficient. Instantaneous divergence of the filter coefficients, e.g., due to false double-talk detection or echo path changes, do not significantly influence the estimated reverberation time $\hat{T}_{60}$ because it is updated slowly. In the case when the filter coefficients cannot convergence, for example due to background noise, the estimated reverberation time will be inaccurate. Overestimation of the reverberation time results in an overestimation of the spectral variance of the late residual echo $\lambda_{e_{\mathrm{r}}}(l,k)$ and the late reverberation $\lambda_{z_{\mathrm{r}}}(l,k)$. During double-talk periods, this introduces some distortion of the early speech component. Informal listening tests indicated that estimations errors $>10\%$ resulted in audible distortions of the early speech component. When only the far-end speech signal is active the overestimation of $\lambda_{e_{\mathrm{r}}}(l,k)$ does not introduce any problems since the suppression is limited by the residual background noise level. Underestimation of the reverberation time results in an underestimation of the spectral variances. Although the underestimation reduces the performance of the system in terms of late residual echo and reverberation suppression, it does not introduce any distortion of the early speech component.

Postfilters that are capable of handling both the residual echo and background noise are often implemented in the STFT domain. In general, they require two STFT and one inverse STFT, which is equal to the number of STFTs used in the proposed solution. The computational complexity of the proposed solution is comparable to former solutions since the estimation of the reverberation time and the late reverberant spectral variance only requires a few operations. The computational complexity of the AEC can be reduced by using an efficient implementation of the AEC in the frequency domain (cf. [49]), rather than in the time-domain.

VII. Experimental Results

In this section, we present experimental results that demonstrate the beneficial use of the developed spectral variance estimators and postfilter.[1] In the subsequent subsections, we evaluate the ability of the postfilter to suppress background noise and nonstationary interferences, i.e., late residual echo and late reverberation. First, the performance of the late residual echo spectral variance estimator and its robustness with respect to changes in the tail of the acoustic echo path is evaluated. Second, the dereverberation performance of the near-end speech is evaluated in the presence of background noise. We compare the dereverberation performance obtained with, and without, DPC that was developed in Section V-B. Finally, we evaluate the performance of the entire system when all interferences are present, i.e., during double-talk.

The experimental setup is depicted in Fig. 3. The room dimensions were 5 m $\times$ 4 m $\times$ 3 m (length $\times$ width $\times$ height). The distance between the near-end speaker and the microphone (r_s) was 0.5 m, the distance between the loudspeaker and microphone (r_l) was 0.25 m. All AIRs were generated using Allen and Berkley's image method [50], [51]. The wall absorption coefficients were chosen such that the reverberation time is approximately 500 ms. The microphone signal $y(n)$ was generated using (5). The analysis window $w(n)$ of the STFT was a 256-point Hamming window, i.e., $N_{\mathrm{w}} = 256$, and the overlap between two successive frames was set to 75%, i.e., $R = 0.25\,N_{\mathrm{w}}$. The remaining parameter settings are shown in Table I. The additive noise $v(n)$ was speech-like noise, taken from the NOISEX-92 database [52].

A. Residual Echo Suppression

The echo cancellation performance, and more specifically the improvement due to the postfilter, was evaluated using the echo

[1]The results are available for listening at the following web page: http://home.tiscali.nl/ehabets/publications/tassp08/tassp08.html

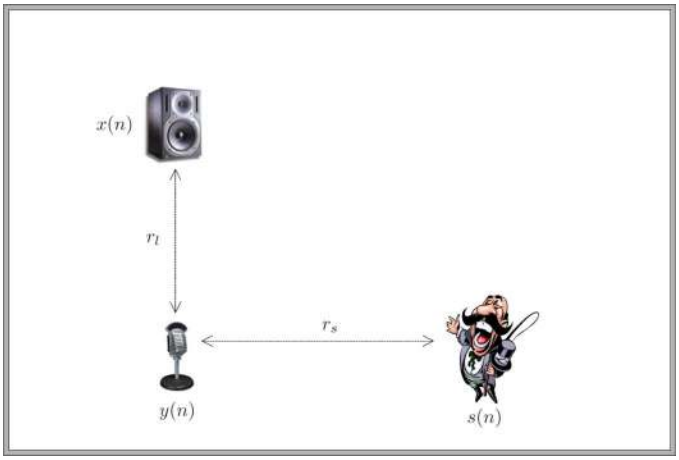

Fig. 3. Experimental setup.

TABLE I
PARAMETERS USED FOR THESE EXPERIMENTS

$f_s = 8000$ Hz	$N_e = 0.128\, f_s$	$N_r = 0.024\, f_s$
$G_{\min}^{dB} = 18$ dB	$\beta^{dB} = 9$ dB	w = 3
$\mu = 0.35$	$\eta_x = 0.5$	$\eta_z = 0.9$

return loss enhancement (ERLE). This experiment was conducted without noise, and the postfilter was configured such that no reverberation was reduced, i.e., $\lambda_{z_r}(l,k) = 0$. The ERLE achieved by the adaptive filter was calculated using

$$\mathrm{ERLE}(l) = 10\log_{10}\left(\frac{\sum_{n=lR'}^{lR'+L'-1} d^2(n)}{\sum_{n=lR'}^{lR'+L'-1}(d(n)-\hat{d}(n))^2}\right)\ \mathrm{dB} \tag{46}$$

where $L' = 0.032\, f_s$ is the frame length and $R' = L'/4$ is the frame rate. To evaluate the total echo suppression, i.e., with postfilter, we calculated the ERLE using (46) and replaced $(d(n)-\hat{d}(n))$ by the residual echo at the output of the postfilter which is given by $(\hat{z}_e(n) - z(n))$. Note that by subtracting near-end speech signal $z(n)$ from the output of the postfilter $\hat{z}_e(n)$, we avoid the bias in the ERLE that is caused by $z(n)$. The final normalized misalignment of the adaptive filter was -24 dB (SNR $= 25$ dB). It should be noted that the developed postfilter only suppresses the residual echo that results from the deficient length of the adaptive filter. Hence, the residual echo that results from the system mismatch of the adaptive filter cannot be compensated by the developed postfilter. The microphone signal $y(n)$, the error signal $e(n)$, and the ERLE with and without postfilter are shown in Fig. 4. We can see that the ERLE is significantly increased when the postfilter is used. A significant reduction of the residual echo was observed when subjectively comparing the error signal and the processed signal. A small amount of residual echo was still audible in the processed signal. However, in the presence of background noise (as discussed in Section VII-C), the residual echo in the processed signal is masked by the residual noise.

We evaluate the robustness of the developed late residual echo suppressor with respect to changes in the tail of the acoustic echo path when the far-end speech signal was active. Let us assume that the AEC is working perfectly at all times, i.e., the $\hat{h}_e(n) = h_e(n)$. We compared three systems: 1) the perfect AEC, 2) the perfect AEC followed by an adaptive filter of length 1024 which compensate for the late residual echo, and 3) the perfect AEC followed by the developed postfilter. It should be noted that the total length of the filter that is used to cancel the echo in system 2 is still shorter than the acoustic echo path. The output of system 2 is denoted by $e'(n)$. At 4 s, the acoustic echo path was changed by changing the position of the loudspeaker in the x-y plane. Here, the loudspeaker position was rotated by 30°, the microphone position was the center of the rotation. The time at which the position changes is marked with a dash-dotted line. The microphone signal $y(n)$, the error signal $e(n)$ of the standard AEC, the signal $e'(n)$ and $\hat{z}_e(n)$, and the ERLEs are shown in Fig. 5. From the results, we can see that the ERLEs of $e'(n)$ and $\hat{z}_e(n)$ are improved compared to the ERLE of $e(n)$. When listening to the output signals, an increase in late residual echo was noticed when using the adaptive filter (system 2), no increase was noticed when using the developed late residual echo estimator and the postfilter (system 3). Since the late residual echo estimator is mainly based on the exponential decaying envelope of the AIR, which does not change over time, the postfilter does not require any convergence time and it does not suffer from the change in the tail of the acoustic echo path. Furthermore, during double-talk, the adaptive filter might not be able to converge due to the low echo to near-end speech-plus-noise ratio of the microphone signal $y(n)$. In the latter case, the developed late residual echo suppressor would still be able to obtain an accurate estimate of the late residual echo.

B. Dereverberation

The dereverberation performance was evaluated using the segmental SIR and the log spectral distance LSD. The parameter $\kappa(k)$ was obtained from the AIR of the system relating $s(n)$ and $z(n)$. An estimate of the reverberation time $\hat{T}_{60}(k)$ was obtained using the procedure described in Appendix B. After convergence of the adaptive filter $\hat{T}_{60}$ was 493 ms. The parameter N_r was set to $0.024\, f_s$.

The instantaneous SIR of the lth frame is defined as

$$\mathrm{SIR}(l) = 10\log_{10}\left(\frac{\sum_{n=lR'}^{lR'+L'-1} z_e^2(n)}{\sum_{n=lR'}^{lR'+L'-1}(z_e(n)-v(n))^2}\right)\ \mathrm{dB} \tag{47}$$

where $v \in \{y, \hat{z}_e\}$. The segmental SIR is defined as the average instantaneous SIR over the set of frames where the near-end speech is active.

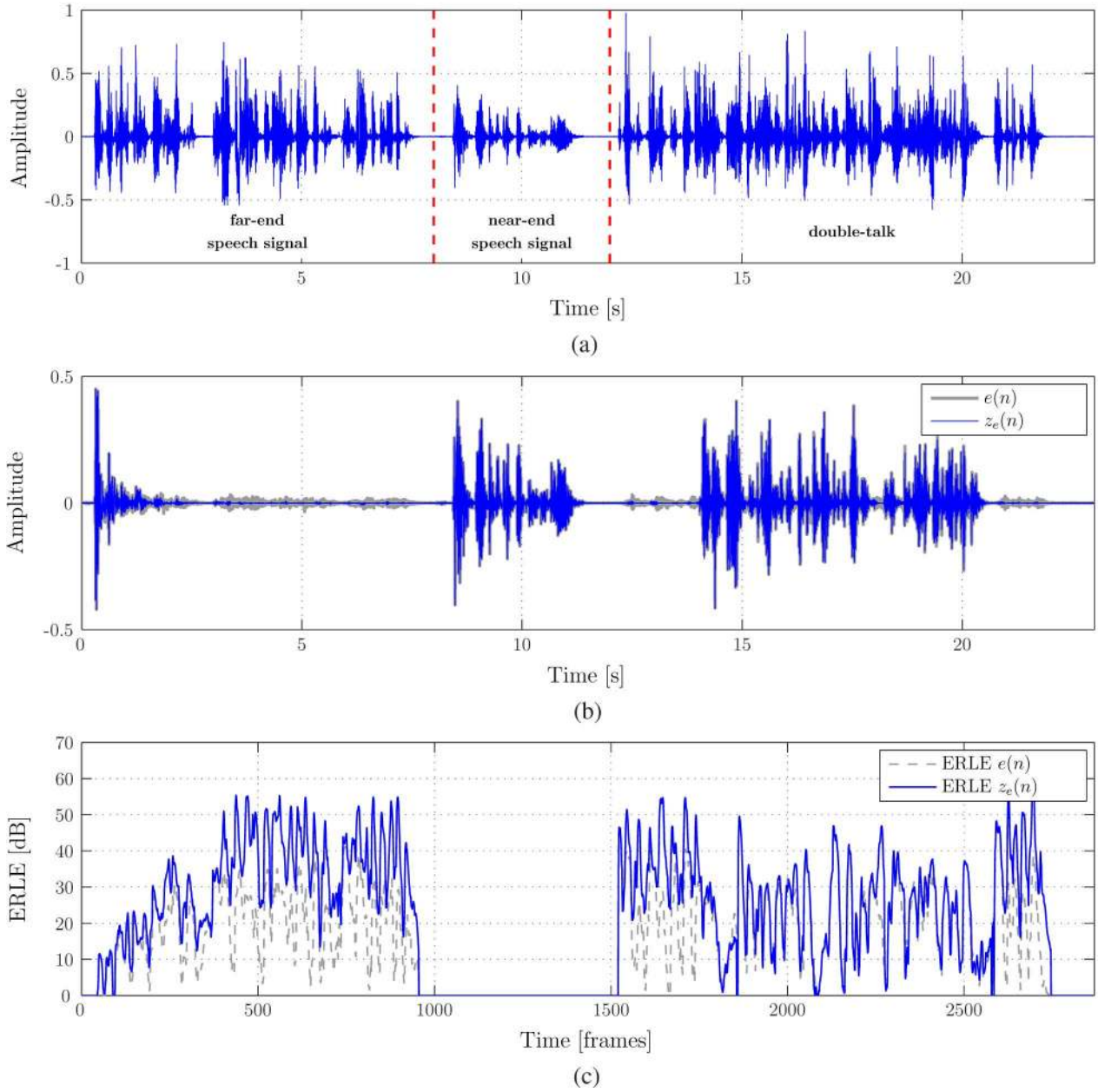

Fig. 4. Echo suppression performance. (a) Microphone signal $y(n)$. (b) Error signal $e(n)$ and the estimated signal $\hat{z}_{\mathrm{e}}(n)$. (c) Echo return loss enhancement of $e(n)$ and $\hat{z}_{\mathrm{e}}(n)$.

The LSD between $z_{\mathrm{e}}(n)$ and the dereverberated signal is used as a measure of distortion. The distance in the lth frame is calculated using

$$\mathrm{LSD}(l) = \frac{1}{K}\sum_{k=0}^{K-1}\left|10\log_{10}\left(\frac{\mathcal{C}\{|Z_{\mathrm{e}}(l,k)|^2\}}{\mathcal{C}\{|\Upsilon(l,k)|^2\}}\right)\right| \ \mathrm{dB} \qquad (48)$$

where $\Upsilon \in \{Y, \hat{Z}_{\mathrm{e}}\}$, K denotes the number of frequency bins, and $\mathcal{C}\{|A(l,k)|^2\} \triangleq \max\{|A(l,k)|^2, \epsilon\}$ denotes a clipping operator which confines the log-spectrum dynamic range to about 50 dB, i.e., $\epsilon = 10^{-50/10}\max_{l,k}\{|A(l,k)|^2\}$. Finally, the LSD is defined as the average distance over all frames.

The dereverberation performance was tested using different segmental SNRs. The segmental SNR value is determined by averaging the instantaneous SNR of those frames where the near-end speech is active. Since the nonstationary interferences, such as the late residual echo and reverberation, are suppressed down to the residual background noise level the postfilter will always include the noise suppression. To show the improvement related to the dereverberation process, we evaluated the segmental SIR and LSD measures for the unprocessed signal, the processed signal [noise suppression (NS) only], the processed signal without DPC [noise and reverberation suppression (NS+RS)], and the processed signal with DPC (NS+RS+DPC). It should be noted that the late reverberant spectral variance estimator without DPC is similar to the method in [22]. The results,

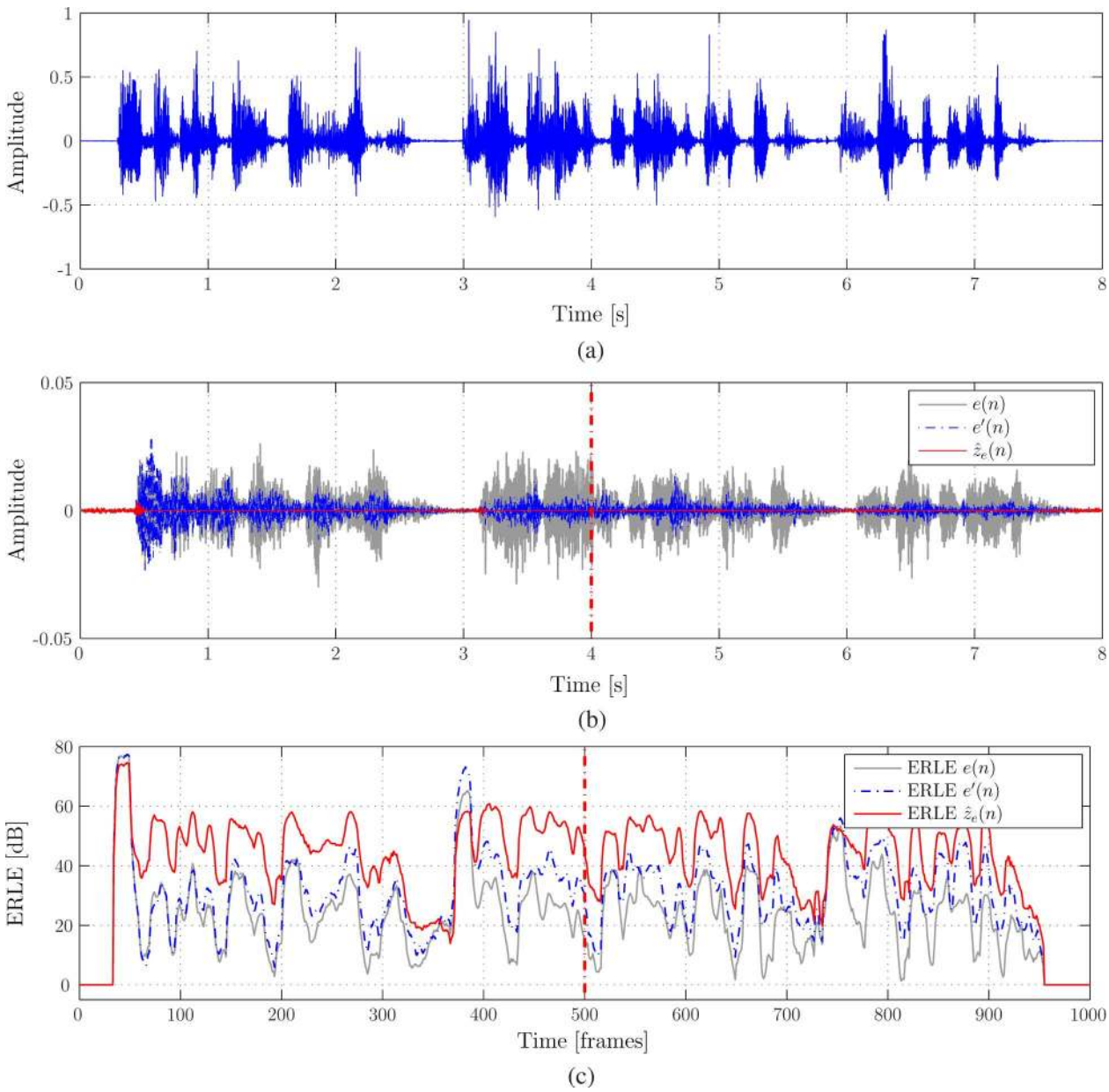

Fig. 5. Echo suppression performance with respect to echo path changes. (a) Microphone signal $y(n)$. (b) Error signals $e(n)$ and $e'(n)$, and the estimated signal $\hat{z}_{\mathrm{e}}(n)$. (c) Echo return loss enhancement of $e(n)$, $e'(n)$, and $\hat{z}_{\mathrm{e}}(n)$.

TABLE II
SEGMENTAL SIR AND LSD FOR DIFFERENT SEGMENTAL SIGNAL-TO-NOISE RATIOS

	segmental SNR = 5 dB		segmental SNR = 10 dB		segmental SNR = 25 dB	
	segSIR	LSD	segSIR	LSD	segSIR	LSD
Unprocessed	-3.28 dB	8.21 dB	0.21 dB	5.77 dB	4.74 dB	2.66 dB
Post-Filter (NS)	2.70 dB	3.54 dB	4.15 dB	2.83 dB	5.31 dB	2.40 dB
Post-Filter (NS+RS)	2.47 dB	4.02 dB	4.48 dB	3.26 dB	6.94 dB	2.45 dB
Post-Filter (NS+RS+DPC)	3.57 dB	3.41 dB	5.38 dB	2.62 dB	7.93 dB	1.71 dB

presented in Table II, show that compared to the unprocessed signal, the segmental SIR and LSD are improved in all cases. It can be seen that the DPC increases the segmental SIR and reduces the LSD, while the reverberation suppression without DPC distorts the signal. When the background noise is suppressed the late reverberation of the near-end speech becomes more pronounced. The results of an informal listening test indicated that the near-end signal that was processed without DPC sounds unnatural as it contains rapid amplitude variations, while the signal that was processed with DPC sounds natural.

The instantaneous SIR and LSD results obtained with a segmental SNR of 25 dB together with the anechoic, reverberant and processed signals are presented in Fig. 6. Since the SNR is relatively high, the instantaneous SIR mainly relates to the amount of reverberation, such that the SIR improvement is related to the reverberation suppression. The instantaneous SIR

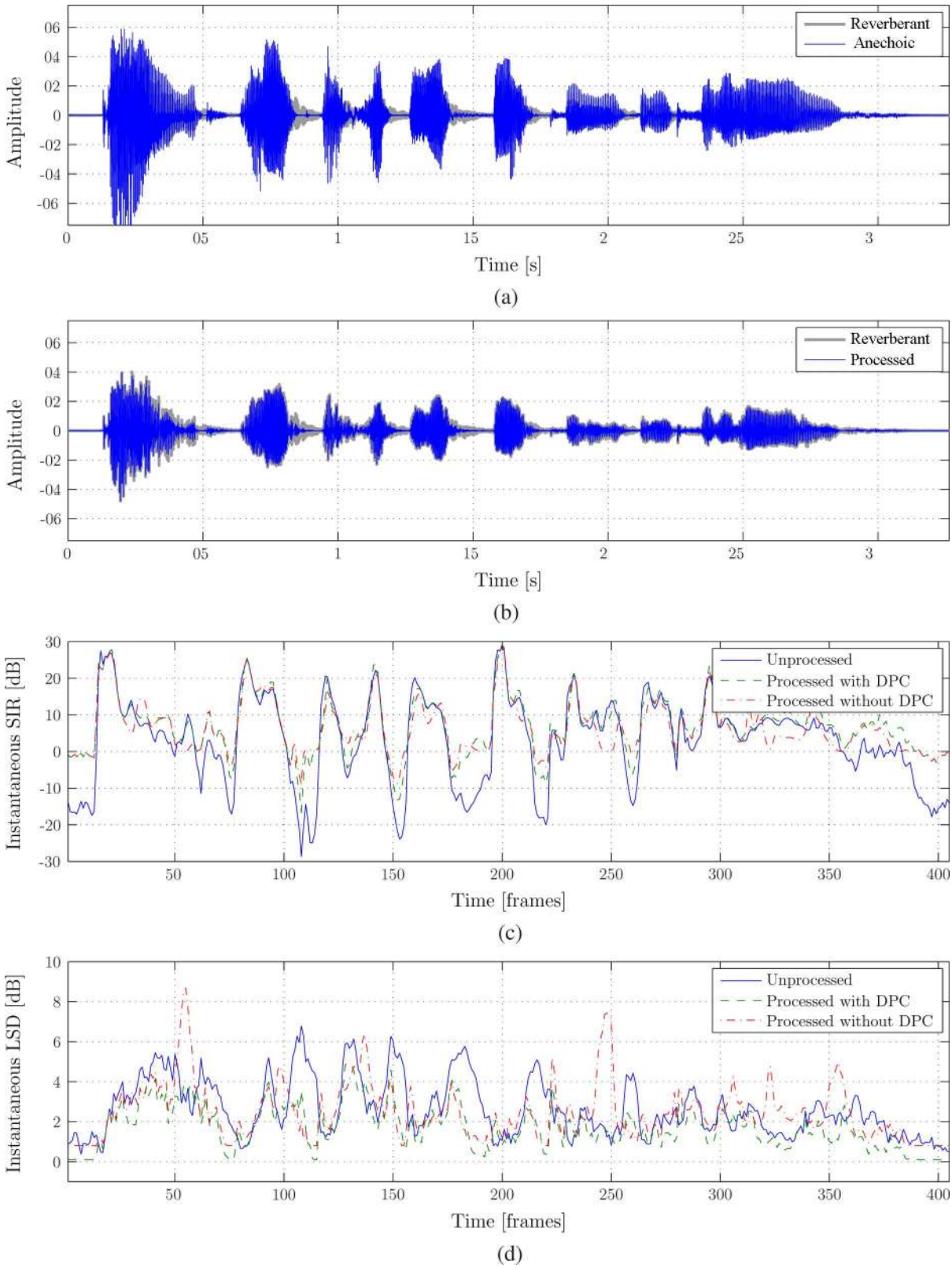

Fig. 6. Dereverberation performance of the system during near-end speech period ($T_{60} \approx 0.5$ s). (a) Reverberant and anechoic near-end speech signal. (b) Reverberant near-end speech signal and estimated early speech component. (c) Instantaneous SIR of the unprocessed and processed (with and without direct path compensation) near-end speech signal. (d) LSD of the unprocessed and processed (with and without direct path compensation) near-end speech signal.

and LSD are, respectively, increased and decreased, especially in those areas where the SIR of the unprocessed signal is low. During speech onsets, some speech distortion may occur due to using the decision directed approach for the *a priori* SIR estimation [36]. We can also see that the processed signal without DPC introduces some spectral distortions, i.e., for some frames

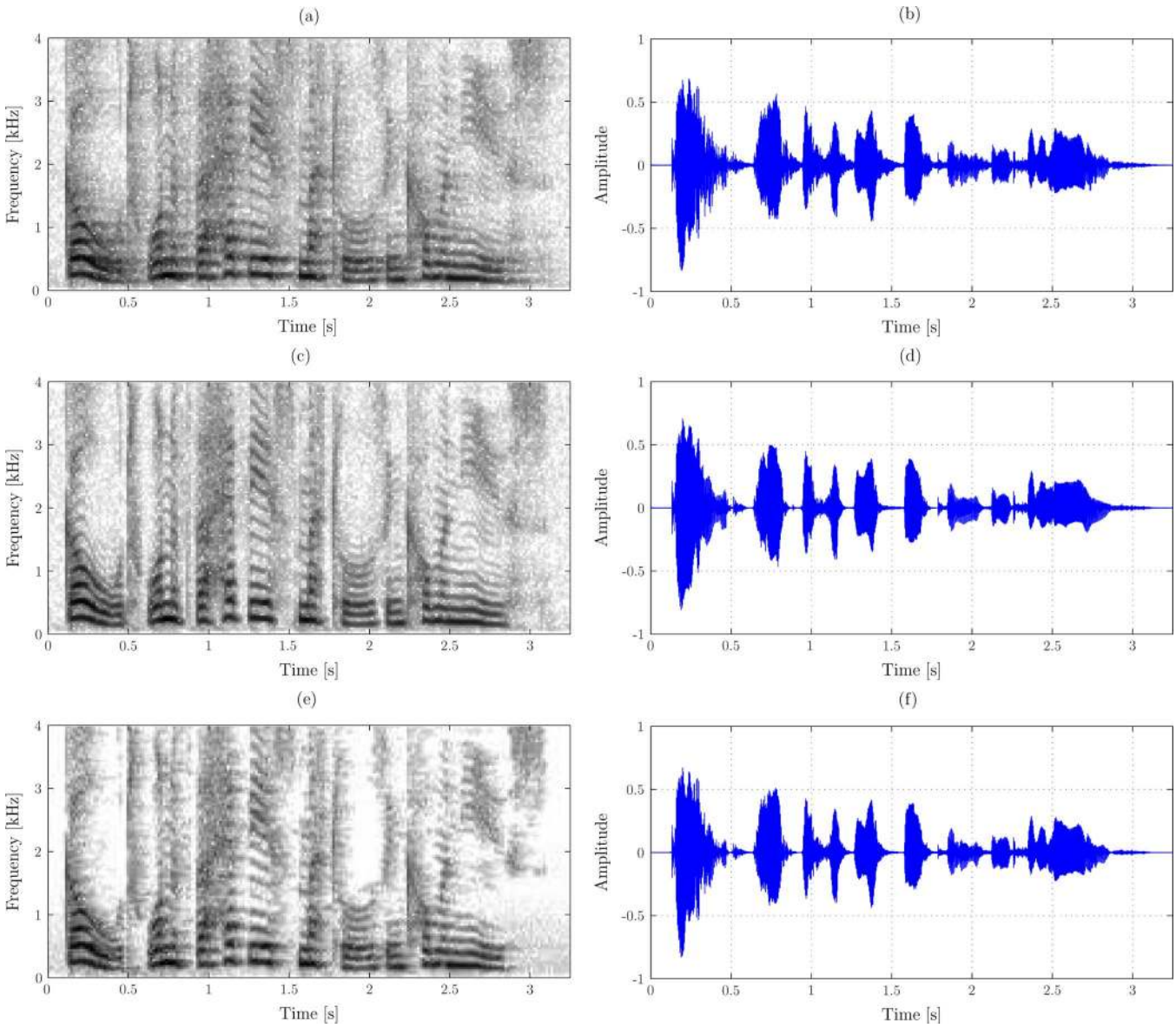

Fig. 7. Spectrogram and waveform of (a), (b) the reverberant near-end speech signal $z(n)$, (c), (d) the early speech component $z_{\mathrm{e}}(n)$, and (e), (f) the estimated early speech component $\hat{z}_{\mathrm{e}}(n)$ (segmental SNR $=$ 25 dB, $T_{60} \approx 0.5$ s).

the LSD is higher than the LSD of the unprocessed signal, while the processed signal with DPC does not introduce such distortions. In general, these distortions occur during spectral transitions in the time–frequency domain. While the distortions are often masked by subsequent phonemes they are clearly audible at the onset and offset of the full-band speech signal. These distortions can best be described as an abrupt increase or decrease of the sound level.

The spectrograms and waveforms of the near-end speech signal $z(n)$, the early speech component $z_{\mathrm{e}}(n)$, and the estimated early speech component $\hat{z}_{\mathrm{e}}(n)$ are shown in Fig. 7. From these plots, it can be seen (for example, at 0.5 s) that the smearing in time due to the reverberation has been reduced significantly.

In Section V-B, we have developed a novel spectral estimator for the late reverberant signal component $z_{\mathrm{r}}(n)$. The estimator requires an additional parameter $\kappa(k)$ which is inversely dependent on the DRR. In the present work, it is assumed that $\kappa(k)$ is *a priori* known. However, in practice, $\kappa(k)$ needs to be estimated online. In this paragraph, we evaluate the robustness with respect to errors in $\kappa(k)$ by introducing an error of $\pm 10\%$. The segmental SIR and LSD using the perturbated values of $\kappa(k)$ are shown in Table III. From this experiment, we can see that

TABLE III
SEGMENTAL SIR AND LSD, SEGMENTAL SNR $=$ 25 dB, AND $\hat{\kappa}(k) = \{\kappa(k), 0.9 \cdot \kappa(k), 1.1 \cdot \kappa(k)\}$

$\hat{\kappa}(k)$	Processed	
	segSIR	LSD
$\kappa(k)$	7.93 dB	1.71 dB
$0.9 \cdot \kappa(k)$	7.87 dB	1.72 dB
$1.1 \cdot \kappa(k)$	7.92 dB	1.71 dB

TABLE IV
SEGMENTAL SIR AND LSD FOR DIFFERENT SEGMENTAL SIGNAL TO NOISE RATIOS DURING DOUBLE-TALK

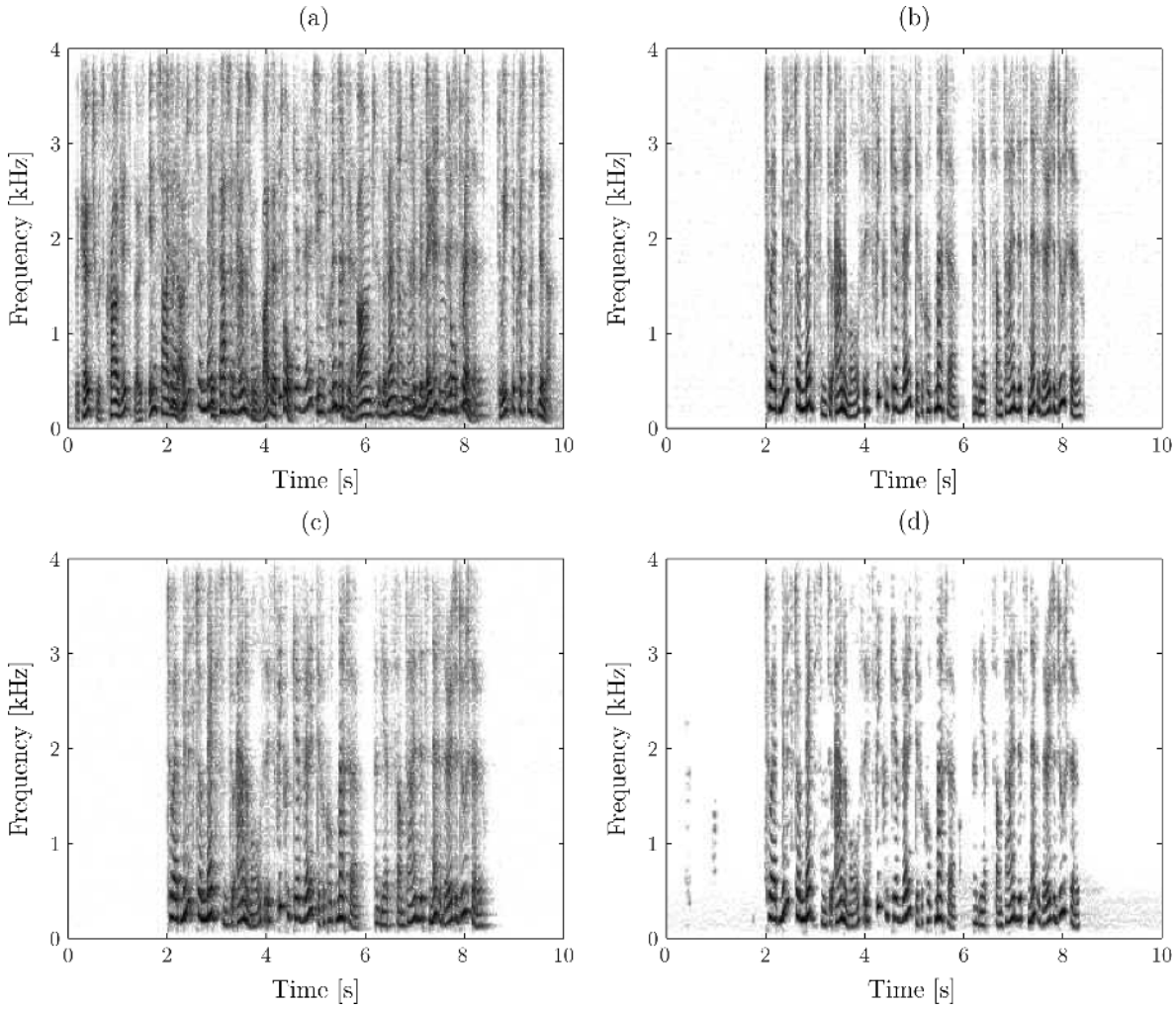

Fig. 8. Spectrograms of (a) the microphone signal $y(n)$, (b) the early speech component $z_{\mathrm{e}}(n)$, (c) the reverberant near-end speech signal $z(n)$, and (d) the estimated early speech component $\hat{z}_{\mathrm{e}}(n)$, during double-talk (segmental SNR $= 25$ dB, $T_{60} \approx 0.5$ s).

the performance of the proposed algorithm is not very sensitive to errors in the parameter $\kappa(k)$. Furthermore, when an estimator for $\kappa(k)$ is developed it is sufficient to obtain a "rough" estimate of $\kappa(k)$.

C. Joint Suppression Performance

We now evaluate the performance of the entire system during double-talk. The performance is evaluated using the segmental SIR and the LSD at three different segmental SNR values. To be able to show that the suppression of each additional interference results in an improvement of the performance we also show the intermediate results. Since all non-stationary interferences, i.e., the late residual echo and reverberation, are reduced down to the residual background noise level, the background noise is suppressed first. We evaluated the performance using i) the AEC, ii) the AEC and postfilter (noise suppression), iii) the AEC and postfilter (noise and residual echo suppression), and iv) the AEC and postfilter (noise, residual echo, and reverberation suppression). The are presented in Table IV. These results show a significant improvement in terms of SIR and LSD. An improvement of the far-end echo to near-end speech ratio is observed when listening to the signal after the AEC (system i). However, reverberant sounding residual echo can clearly be noticed. When the background noise is suppressed (system ii) the residual echo and reverberation of the near-end speech becomes more pronounced. After suppression of the late residual echo (system iii) almost no echo is observed. When in addition the late reverberation is suppressed (system iv) it sounds like the near-end speaker has moved closer to the microphone. Informal listening tests using normal hearing subjects showed a significant improvement of the speech quality when comparing the output of system ii and system iv.

The spectrograms of the microphone signal $y(n)$, the early speech component $z_{\mathrm{e}}(n)$, and the estimated signal $\hat{z}_{\mathrm{e}}(n)$ for a segmental SNR of 25 dB and 5 dB, are shown in Figs. 8 and 9,

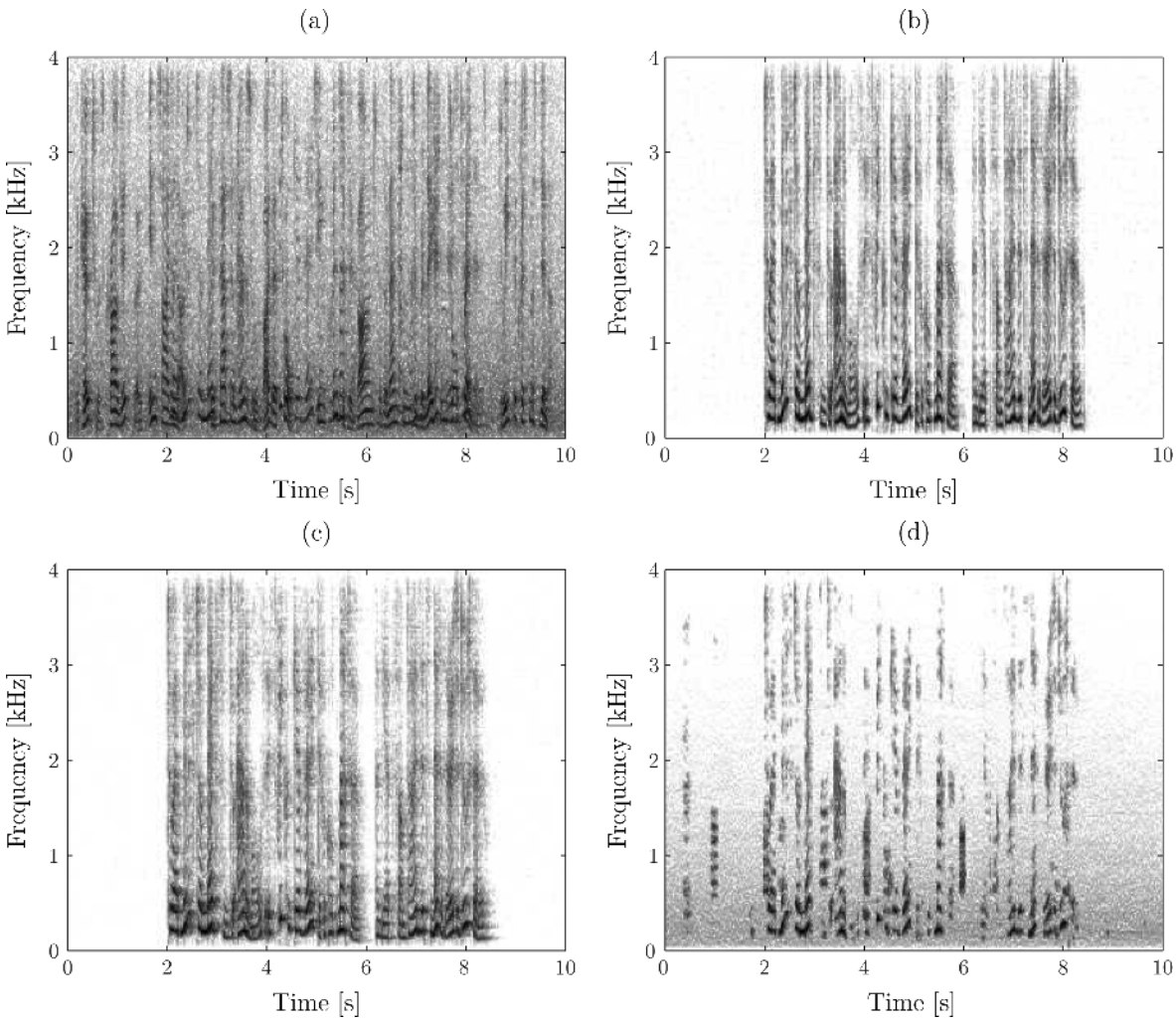

Fig. 9. Spectrograms of (a) the microphone signal $y(n)$, (b) the early speech component $z_e(n)$, (c) the reverberant near-end speech signal $z(n)$, and (d) the estimated early speech component $\hat{z}_e(n)$, during double-talk (segmental SNR $= 5$ dB, $T_{60} \approx 0.5$ s).

respectively. The spectrograms demonstrate how well the interferences are suppressed during double-talk.

VIII. Conclusion

We have developed a novel postfilter for an AEC which is designed to efficiently reduce reverberation of the near-end speech signal, late residual echo and background noise. Spectral variance estimators for the late residual echo and late reverberation have been derived using a statistical model of the AIR that depends on the reverberation time of the room. Because blind estimation of the reverberation time is very difficult, a major advantage of the hands-free scenario is that due to the existence of the echo an estimate of the reverberation time can be obtained from the estimated acoustic echo path. Finally, the near-end speech is estimated based on a modified OM-LSA estimator. The modification ensures a stationary residual background noise level of the output. Experimental results demonstrate the performance of the developed postfilter and its robustness to small changes in the tail of the acoustic echo path. During single- and double-talk periods a significant amount of interference is suppressed with little speech distortion.

The statistical model of the AIR does not take the energy contribution of the direct path into account. Hence, a late reverberant spectral variance estimator, which is based on this model, results in an overestimated spectral variance. This phenomenon is pronounced when the source-microphone distance is smaller than the critical distance and results in spectral distortions of the desired speech signal. Therefore, we derived an estimator that compensates for the contribution of the direct path energy. The compensation requires one additional (possibly frequency dependent) parameter that is related to the DRR of the AIR. We demonstrated that the proposed estimator is not very sensitive to estimation errors of this parameter. Future research will focus on the blind estimation of this parameter.

When multi-microphones are available rather than a single-microphone the spatial diversity of the signal can be used to increase the suppression of reverberation and other interferences. Extending the postfilter to the case where a microphone array is

available, rather than a single microphone, is a topic for future research.

APPENDIX A
A Priori SIR ESTIMATOR

Rather than using one *a priori* SIR it is possible to calculate one value for each interference. By doing this, one gains control over i) the trade-off between the interference reduction and the distortion of the desired signal, and ii) the *a priori* SIR estimation approach of each interference. Note that in some cases it might be desirable to reduce one of the interferences at the cost of larger speech distortion, while other interferences are reduced less to avoid distortion. Gustafsson *et al.* also used separate *a priori* SIRs in [13], [30] for two interferences, i.e., background noise and residual echo. In this section we show how the Decision Directed approach can be used to estimate the individual *a priori* SIRs, and we propose a slightly different way of combining them. It should be noted that each *a priori* SIR could be estimated using a different approach, e.g., the Decision Directed *a priori* SIR estimator proposed by Ephraim and Malah in [35] or the non-causal *a priori* SIR estimator proposed by Cohen in [36]. In this work we have used the Decision Directed *a priori* SIR estimator.

The *a priori* SIR in (9) can be written as

$$\frac{1}{\xi(l,k)} = \frac{1}{\xi_{z_\mathrm{r}}(l,k)} + \frac{1}{\xi_{e_r}(l,k)} + \frac{1}{\xi_v(l,k)} \tag{49}$$

with

$$\xi_\vartheta(l,k) = \frac{\lambda_{z_\mathrm{e}}(l,k)}{\lambda_\vartheta(l,k)} \tag{50}$$

where $\vartheta \in \{z_\mathrm{r}, e_r, v\}$.

Let us assume that there always is a certain amount of background noise. When the power of the near-end speech is very low and the power of the late reverberant and/or residual echo is very low, the *a priori* SIR $\xi_{z_\mathrm{r}}(l,k)$ and/or $\xi_{e_r}(l,k)$ may be unreliable since $\lambda_{z_\mathrm{e}}(l,k)$ and $\lambda_{z_\mathrm{r}}(l,k)$ and/or $\lambda_{e_r}(l,k)$ are close to zero. Due to this the *a priori* SIR $\xi(l,k)$ may be unreliable. Because the LSA gain function as well as the speech presence probability $p(l,k)$ depend on $\xi(l,k)$, an inaccurate estimate can decrease the performance of the postfilter. We propose to calculate $\xi(l,k)$ using only the most important and reliable *a priori* SIRs as follows:[2]

$$\xi(l,k) = \begin{cases} \xi_v, & \text{if } 10\log_{10}\left(\frac{\lambda_v}{\lambda_{z_\mathrm{r}}+\lambda_{e_r}}\right) > \beta^{\mathrm{dB}} \\ \xi', & \text{otherwise} \end{cases} \tag{51}$$

and

$$\xi'(l,k) = \begin{cases} \frac{\xi_{e_r}\,\xi_v}{\xi_{e_r}+\xi_v}, & \text{if } 10\log_{10}\left(\frac{\lambda_{e_r}}{\lambda_{z_\mathrm{r}}}\right) > \beta^{\mathrm{dB}} \\ \frac{\xi_{z_\mathrm{r}}\,\xi_v}{\xi_{z_\mathrm{r}}+\xi_v}, & \text{if } 10\log_{10}\left(\frac{\lambda_{z_\mathrm{r}}}{\lambda_{e_r}}\right) > \beta^{\mathrm{dB}} \\ \frac{\xi_{z_\mathrm{r}}\,\xi_v\,\xi_{e_r}}{\xi_v\,\xi_{e_r}+\xi_{z_\mathrm{r}}\,\xi_{e_r}+\xi_{z_\mathrm{r}}\,\xi_v}, & \text{otherwise} \end{cases} \tag{52}$$

where the threshold β^{dB} specifies the level difference in dB. When the noise level is β^{dB} higher than the level of residual echo and late reverberation (in dB), the total *a priori* SIR, $\xi(l,k)$, will be equal to $\xi_v(l,k)$. Otherwise $\xi(l,k)$ will be calculated depending on the level difference between $\lambda_{z_\mathrm{r}}(l,k)$ and $\lambda_{e_r}(l,k)$ using (52): When the level of residual echo is β^{dB} larger than the level of late reverberation, $\xi(l,k)$ will depend on both $\xi_v(l,k)$ and $\xi_{e_r}(l,k)$. When the opposite is true, $\xi(l,k)$ will depend on both $\xi_v(l,k)$ and $\xi_{z_\mathrm{r}}(l,k)$. In any other case $\xi_v(l,k)$ will be calculated using all *a priori* SIRs.

To estimate $\xi_\vartheta(l,k)$ we use the following expression

$$\hat{\xi}_\vartheta(l,k) = \max\left\{\eta_\vartheta \frac{|\hat{Z}_\mathrm{e}(l-1,k)|^2}{\lambda_\vartheta(l-1,k)} + (1-\eta_\vartheta)\,\max\{\psi_\vartheta(l,k),0\}, \xi_{\min,\vartheta}\right\} \tag{53}$$

where

$$\psi_\vartheta(l,k) = \frac{\lambda(l,k)}{\lambda_\vartheta(l,k)}\,\psi(l,k) = \frac{|E(l,k)|^2 - \lambda(l,k)}{\lambda_\vartheta(l,k)} \tag{54}$$

and $\xi_{\min,\vartheta}$ is the lower-bound on the *a priori* SIR $\xi_\vartheta(l,k)$.

APPENDIX B
ESTIMATION OF THE REVERBERATION TIME

The reverberation time can be estimated directly from the EDC of $\hat{\mathbf{h}}_\mathrm{e}(n)$. It should be noted that the last EDC values are not useful due to the finite length of $\hat{\mathbf{h}}_\mathrm{e}(n)$ and due to the final misalignment of the adaptive filter coefficients. Therefore, we use only a dynamic range of 20 dB[3] to determine the slope of the EDC. The estimated reverberation time is then updated using an adaptive scheme.

In general, the reverberation time T_{60} is frequency dependent due to frequency dependent reflection coefficients of walls and other objects and the frequency dependent absorption coefficient of air [40]. Instead of applying the above procedure to $\hat{\mathbf{h}}_\mathrm{e}(n)$, we can apply the above procedure to band-pass filtered versions of $\hat{\mathbf{h}}_\mathrm{e}(n)$, denoted by $\hat{\mathbf{h}}_\mathrm{e}(n,b)$ for $b = \{1,\ldots,B\}$, where b denotes the sub-band index and B denotes the number of band-pass filters. We used 1-octave band filters to acquire the reverberation time $T_{60}(n,k_b)$, where k_b denotes the center frequency of band-pass filter b. The obtained values are then interpolated and extrapolated to obtain an estimate of $T_{60}(n,k)$ for each frequency bin k and time n. A detailed description for estimating $T_{60}(n,k_b)$ given $\hat{\mathbf{h}}_\mathrm{e}(n,b)$ can be found in Alg. 2.

[2]The time and frequency indices at the right-hand side have been omitted.

[3]It might be necessary to decrease the dynamic range when N_e is small or the reverberation time is long.

Algorithm 2 Estimation of the reverberation time $T_{60}(n, k_b)$ given a band-pass filtered echo path impulse response $\hat{\mathbf{h}}_{\mathrm{e}}(n, b)$.

1) Calculate the Energy Decay Curve of $\hat{\mathbf{h}}_{\mathrm{e}}(n,b)$, where b denotes the sub-band index, using

$$\mathrm{EDC}(n,m,b) = 20\log_{10}\left(\sum_{j=m}^{N_{\mathrm{e}}-1}(\hat{h}_{\mathrm{e},j}(n,b))^2\right) \quad \text{for} \quad 0 \le m \le N_{\mathrm{e}}-1.$$

2) A straight line is fitted through a selected part of the EDC values using a least squares approach. The line at time n is described by $p(n,b)+q(n,b)\,m$, where $p(n,b)$ and $q(n,b)$ denotes the offset and the regression coefficient of the line, respectively. The regression coefficient $q(n,b)$ is obtained by minimizing the following cost function:

$$J(p(n,b),q(n,b)) = \sum_{m=m_s(b)}^{m_e(b)}(\mathrm{EDC}(n,m,b)-(p(n,b)+q(n,b)\,m))^2$$

where $m_s(b)$ $(0 \le m_s < N_{\mathrm{e}}-1)$ and $m_e(b)$ $(m_s < m_e \le N_{\mathrm{e}}-1)$ denote the start-time and end-time of EDC values that are used, respectively. A good choice for $m_s(b)$ and $m_e(b)$ is given by

$$m_s(b) = \arg\min_m \left|\frac{\mathrm{EDC}(n,m,b)}{\mathrm{EDC}(n,0,b)}+5\right|$$

and

$$m_e(b) = \arg\min_m \left|\frac{\mathrm{EDC}(n,m,b)}{\mathrm{EDC}(n,0,b)}+25\right|$$

respectively.

3) The reverberation time for frequency bin k_b, where k_b denotes the center frequency of the b^{th} band-pass filter, can now be calculated using

$$\hat{T}_{60}(n,k_b) = \hat{T}_{60}(n-1,k_b) + \mu_{\mathrm{T}_{60}}\left(\frac{60}{q(n,b)f_{\mathrm{s}}} - \hat{T}_{60}(n-1,k_b)\right)$$

where $\mu_{\mathrm{T}_{60}}$ denotes the adaptation step-size.

To reduce the complexity of the estimator the reverberation time can be estimated at regular time intervals, i.e., for $n = u\,R_{\mathrm{T}_{60}}$, where $u \in \mathbb{N}$ and $R_{\mathrm{T}_{60}}$ denotes the estimation rate of the reverberation time.

APPENDIX C
ESTIMATION OF THE INITIAL POWER

The initial power $c(l,k)$ can be calculated using the following expression

$$c(l,k) = \left|\sum_{j=0}^{N_{\mathrm{w}}-1}\hat{h}_{\mathrm{r},j}(lR)\,e^{-\iota\frac{2\pi k}{N_{\mathrm{DFT}}}j}\right|^2 \quad \text{for} \quad k=\{0,\ldots,N_{\mathrm{DFT}}-1\} \quad (55)$$

where $\iota=\sqrt{-1}$ and N_{w} is the length of the analysis window. Since $\hat{\mathbf{h}}_{\mathrm{r}}(n)$ is not available, we use the last N_{w} coefficients of $\hat{\mathbf{h}}_{\mathrm{e}}(n)$ and extrapolate the energy using the estimated decay. We then obtain an estimate of $c(l,k)$ by

$$\hat{c}(l,k) = \alpha^{\frac{N_{\mathrm{w}}}{R}}(k)\left|\sum_{j=0}^{N_{\mathrm{w}}-1}\hat{h}_{\mathrm{e},N_{\mathrm{e}}-N_{\mathrm{w}}+j}(lR)\,e^{-\iota\frac{2\pi k}{N_{\mathrm{DFT}}}j}\right|^2 \quad \text{for} \quad k=\{0,\ldots,N_{\mathrm{DFT}}-1\}. \quad (56)$$

The estimated initial power might contain some spectral zeros, which can easily be removed by smoothing $\hat{c}(l,k)$ along the frequency axis using

$$\tilde{c}(l,k) = \sum_{i=-w}^{w} b_i\hat{c}(l,k+i) \quad (57)$$

where b is a normalized window function $(\sum_{i=-w}^{w} b_i = 1)$ that determines the frequency smoothing.

In this work we have calculated $\tilde{c}(l,k)$ for every frame l. However, in many cases it can be assumed that the acoustic echo path is slowly time-varying. Therefore, $\tilde{c}(l,k)$ does not have to be calculated for every frame l. By calculating $\tilde{c}(l,k)$ at a lower frame rate the computational complexity of the late residual echo estimator can be reduced.

ACKNOWLEDGMENT

The authors like to thank the anonymous reviewers for their constructive comments which helped to improve the presentation of this paper.

REFERENCES

[1] G. Schmidt, "Applications of acoustic echo control: An overview," in *Proc. Eur. Signal Process. Conf. (EUSIPCO'04)*, Vienna, Austria, 2004, pp. 9–16.

[2] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control—An application of very-high-order adaptive filters," *IEEE Signal Process. Mag.*, vol. 16, no. 4, pp. 42–69, Jul. 1999.

[3] E. Hänsler, "The hands-free telephone problem: An annotated bibliography," *Signal Process.*, vol. 27, no. 3, pp. 259–271, 1992.

[4] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. New York: Wiley, Jun. 2004.

[5] V. Myllylä, "Residual echo filter for enhanced acoustic echo control," *Signal Process.*, vol. 86, no. 6, pp. 1193–1205, Jun. 2006.

[6] G. Enzner, "A model-based optimum filtering approach to acoustic echo control: Theory and practice," Ph.D. dissertation, RWTH Aachen Univ., Aachen, Germany, Apr. 2006, Wissenschaftsverlag Mainz, ISBN 3-86130-648-4.

[7] V. Turbin, A. Gilloire, and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '97)*, 1997, vol. 1, pp. 307–310.

[8] Y. Grenier, M. Xu, J. Prado, and D. Liebenguth, "Real-time implementation of an acoustic antenna for audio-conference," in *Proc. Int. Workshop Acoust. Echo Control*, Berlin, Sep. 1989, CD-ROM.

[9] M. Xu and Y. Grenier, "Acoustic echo cancellation by adaptive antenna," in *Proc. Int. Workshop Acoust. Echo Control*, Berlin, Sep. 1989, CD-ROM.

[10] H. Yasukawa, "An acoustic echo canceller with sub-band noise cancelling," *IEICE Trans. Fundamentals Electron., Commun., Comput. Sci.*, vol. E75-A, no. 11, pp. 1516–1523, 1992.

[11] R. Le, B. Jeannes, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: A survey," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 808–820, Nov. 2001.

[12] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals," *Signal Process.*, vol. 64, no. 1, pp. 33–47, Jan. 1998.
[13] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Process.*, vol. 10, no. 5, pp. 245–256, Jul. 2002.
[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[15] R. Martin and P. Vary, "Combined acoustic echo cancellation, dereverberation and noise reduction: A two microphone approach," *Proc. Annales des Telecomm.*, vol. 49, no. 7–8, pp. 429–438, Jul. –Aug. 1994.
[16] M. Dörbecker and S. Ernst, "Combination of two-channel spectral subtraction and adaptive wiener post-filtering for noise reduction and dereverberation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO 1996)*, Triest, Italy, 1996, pp. 995–998.
[17] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, no. 4, pp. 912–915, 1977.
[18] P. Bloom and G. Cain, "Evaluation of two input speech dereverberation techniques," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, (ICASSP'82)*, 1982, vol. 1, pp. 164–167.
[19] A. Oppenheim, R. Schafer, and J. T. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, no. 8, pp. 1264–1291, Aug. 1968.
[20] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'91)*, 1991, vol. 2, pp. 977–980.
[21] 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.
[22] K. Lebart and J. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.
[23] E. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. 30th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'05)*, Philadelphia, PA, Mar. 2005, pp. 173–176.
[24] J. Wen, N. Gaubitch, E. Habets, T. Myatt, and P. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. 10th Int. Workshop Acoust. Echo and Noise Control (IWAENC'06)*, Paris, France, Sep. 2006, pp. 1–4.
[25] J. Benesty and S. L. Gay, "An improved PNLMS algorithm," in *Proc. 27th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, 2002, pp. 1881–1884.
[26] E. Hänsler and G. Schmidt, "Hands-free telephones—Joint control of echo cancellation and postfiltering," *Signal Process.*, vol. 80, pp. 2295–2305, 2000.
[27] T. Gänsler and J. Benesty, "The fast normalized cross-correlation double-talk detector," *Signal Process.*, vol. 86, pp. 1124–1139, Jun. 2006.
[28] F. Aigner and M. Strutt, "On a physiological effect of several sources of sound on the ear and its consequences in architectural acoustics," *J. Acoust. Soc. Amer.*, vol. 6, no. 3, pp. 155–159, 1935.
[29] J. Allen, "Effects of small room reverberation on subjective preference," *J. Acoust. Soc. Amer.*, vol. 71, pp. S1–S5, 1982.
[30] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Process.*, vol. 64, no. 1, pp. 21–32, Jan. 1998.
[31] G. Faucon and R. L. B. Jeannès, "Joint system for acoustic echo cancellation and noise reduction," in *EuroSpeech*, Madrid, Spain, Sep. 1995, pp. 1525–1528.
[32] E. Habets, I. Cohen, and S. Gannot, "MMSE log spectral amplitude estimator for multiple interferences," in *Proc. 10th Int. Workshop Acoust. Echo Noise Control (IWAENC 2006)*, Paris, France, Sep. 2006, pp. 1–4.
[33] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[34] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.
[35] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[36] I. Cohen, "Relaxed statistical model for speech enhancement and *a priori* SNR estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sep. 2005.
[37] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.
[38] R. Crochiere and L. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
[39] M. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *J. Acoust. Soc. Amer.*, vol. 66, no. 2, pp. 497–500, 1979.
[40] H. Kuttruff, *Room Acoustics*, 4th ed. London, U.K.: Spon Press, 2000.
[41] M. Karjalainen, P. Antsalo, A. Mkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc.*, vol. 11, pp. 867–878, 2002.
[42] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.
[43] J. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Univ. du Maine, La Mans, France, 1988.
[44] M. Schroeder, "Frequency correlation functions of frequency responses in rooms," *J. Acoust. Soc. Amer.*, vol. 34, no. 12, pp. 1819–1823, 1962.
[45] A. Accardi and R. Cox, "A modular approach to speech enhancement with an application to speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '99)*, 1999, vol. 1, pp. 201–204.
[46] A. Abramson, E. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation using garch modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, Las Vegas, NV, 2008.
[47] D. Schobben and P. Sommen, "On the performance of too short adaptive fir filters," in *Proc. CSSP-97, 8th Annu. ProRISC/IEEE Workshop Circuits, Syst. Signal Process.*, J. Veen, Ed., Utrecht, The Netherlands, 1997, pp. 545–549, STW, Technology Foundation, ISBN 90-73461-12-X.
[48] K. Mayyas, "Performance analysis of the deficient length LMS adaptive algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 53, no. 8, pp. 2727–2734, 2005.
[49] J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14–37, Jan. 1992.
[50] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.
[51] P. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, no. 5, pp. 1527–1529, Nov. 1986.
[52] A. Varga and H. 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.*, vol. 12, pp. 247–251, Jul. 1993.

Emanuël A. P. Habets (S'02–M'07) received the B.Sc. degree in electrical engineering from the Hogeschool Limburg, Limburg, The Netherlands, in 1999 and the M.Sc. and Ph.D. degrees in electrical engineering from the Technische Universiteit Eindhoven, Eindhoven, The Netherlands, in 2002 and 2007, respectively.

In February 2006, he was A Guest Researcher at Bar-Ilan University, Ramat-Gan, Israel. Since March 2007, he has been a Postdoctoral Researcher at the Technion—Israel Institute of Technology and at the Bar-Ilan University. His research interests include statistical signal processing and speech enhancement using either single or multiple microphones with applications in acoustic communication systems. His main research interest is speech dereverberation.

Dr. Habets was a member of the organization committee of the 9th International Workshop on Acoustic Echo and Noise Control (IWAENC), Eindhoven, The Netherlands, 2005.

Sharon Gannot (S'93–M'01–SM'06) received the B.Sc. degree (summa cum laude) from the Technion—Israel Institute of Technology, Haifa, Israel, in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Tel-Aviv, Israel, in 1995 and 2000, respectively, all in electrical engineering.

From 1986 to 1993, he was head of a research and development section, in an R&D center of the Israeli Defense Forces. In 2001, he held a Postdoctoral Position in the Department of Electrical Engineering (SISTA), K. U. Leuven, Leuven, Belgium. From 2002 to 2003, he held a research and teaching position at the Signal and Image Processing Lab (SIPL), Faculty of Electrical Engineering, Technion—Israel Institute of Technology. Currently, he is a Senior Lecturer in the School of Engineering, Bar-Ilan University, Ramat-Gan, Israel. His research interests include parameter estimation, statistical signal processing, and speech processing using either single- or multimicrophone arrays. He is an Associate Editor of the *EURASIP Journal Applied signal Processing*, an Editor of a special issue on Advances in Multimicrophone Speech Processing of the same journal, a Guest Editor of *ELSEVIER Speech Communication* journal and a reviewer of many IEEE journals.

Israel Cohen (M'01–SM'03) received the B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees in electrical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 1990, 1993, and 1998, respectively.

From 1990 to 1998, he was a Research Scientist at RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate at the Computer Science Department, Yale University, New Haven, CT. In 2001, he joined the Electrical Engineering Department, the Technion, where he is currently an Associate Professor. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification and adaptive filtering. He is a Guest Editor of a special issue of the *EURASIP Journal on Applied Signal Processing* on advances in multimicrophone speech processing and a special issue of the *EURASIP Speech Communication Journal* on speech enhancement. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2007).

Dr. Cohen received the Technion Excellent Lecturer Award in 2005. He serves as Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS.

Piet C. W. Sommen received the Ingenieur degree in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 1981 and the Ph.D. degree from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 1992.

From 1981 to 1989, he was with Philips Research Laboratories, Eindhoven, and since 1989, with the faculty of Electrical Engineering, Eindhoven University of Technology, where he is currently an Associate Professor. He is involved in internal and external courses, all dealing with different basic and advanced signal processing topics. His main field of research is in adaptive array signal processing, with applications in acoustic communication systems.

Dr. Sommen has been a member of the faculty board as a Research Dean (1993–1995), member of ProRISC board (1993–2002), Vice President of IEEE Benelux Signal Processing Chapter (1998–2002), Officer of the Administrative board of EURASIP (1998–2003). EURASIP Newsletter Editor (1998–2003), Editor of the *Journal of Applied Signal Processing* (2002–2004), Editor of a special issue on Signal Processing for Acoustic Communications (2003), reviewer of the MEDEA+ project (2002–2004) and cochair of International Workshop on Acoustic Echo and Noise Control (2005).