Cohen *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:33
https://doi.org/10.1186/s13636-021-00219-2

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

An online algorithm for echo cancellation, dereverberation and noise reduction based on a Kalman-EM Method

Nili Cohen, Gershon Hazan, Boaz Schwartz and Sharon Gannot*

Abstract

Many modern smart devices are equipped with a microphone array and a loudspeaker (or are able to connect to one). Acoustic echo cancellation algorithms, specifically their multi-microphone variants, are essential components in such devices. On top of acoustic echos, other commonly encountered interference sources in telecommunication systems are reverberation, which may deteriorate the desired speech quality in acoustic enclosures, specifically if the speaker distance from the array is large, and noise. Although sub-optimal, the common practice in such scenarios is to treat each problem separately. In the current contribution, we address a unified statistical model to simultaneously tackle the three problems. Specifically, we propose a recursive EM (REM) algorithm for solving echo cancellation, dereverberation and noise reduction. The proposed approach is derived in the short-time Fourier transform (STFT) domain, with time-domain filtering approximated by the convolutive transfer function (CTF) model. In the E-step, a Kalman filter is applied to estimate the near-end speaker, based on the noisy and reveberant microphone signals and the echo reference signal. In the M-step, the model parameters, including the acoustic systems, are inferred. Experiments with human speakers were carried out to examine the performance in dynamic scenarios, including a walking speaker and a moving microphone array. The results demonstrate the efficiency of the echo canceller in adverse conditions together with a significant reduction in reverberation and noise. Moreover, the tracking capabilities of the proposed algorithm were shown to outperform baseline methods.

Keywords: Array processing, Acoustic echo cancellation, Dereverberation, Recursive expectation-maximization algorithm, Convolutive transfer function approximation in the STFT domain

1 Introduction

1.1 The echo cancellation problem

Acoustic echo cancellation algorithms are an essential component in many telecommunication systems such as hands-free devices, conference room speakerphones and hearing aids [1–3]. Moreover, in modern devices, such as smart speakers that play loud music, it is mandatory to integrate an acoustic echo cancellation (AEC) algorithm to enable proper functionality of automatic speech recognition (ASR) systems, especially in the task of recognizing a hot-word. Echo control is also common in robot audition applications, to enable proper human-robot interaction. This may impose further complexity to the problem, as the robot may move while capturing the sound from the speakers.

Generally, the role of AEC algorithms is to suppress the interference related to a far-end speaker (a known reference signal) and enhance the desired speech signal, denoted near-end speaker. This task requires an estimate of the acoustic path relating the loudspeaker and the microphone, and is obtained by the application of an adaptive filter [4–8]. Then, the far-end signal is convolved with the estimated echo path to obtain a replica of the echo signal as received by the microphone. An estimate of the desired near-end signal is finally obtained by

*Correspondence: sharon.gannot@biu.ac.il
Faculty of Engineering, Bar Ilan University, 5290002 Ramat-Gan, Israel

subtracting the estimated echo signals from the microphone signal. In [9], an acoustic echo control method is derived, including an echo canceller and a postfilter. The proposed algorithm is based on the Kalman Filter and provides an optimal statistical adaptation framework for the filter coefficients in time-varying and noisy acoustic environments.

1.2 Literature review

Many modern devices are equipped with more than one microphone. The common and most straightforward solution for cancelling the echo signal in the presence of noise is to first independently apply an AEC between the loudspeaker and each of the microphones and then to apply a beamformer. Cascade schemes, implemented in the time-domain, for joint AEC and beamforming are presented in [10, 11], with either AEC preceding the beamformer or vice versa. A frequency-domain implementation addressing the joint noise reduction and multi-microphone echo cancellation is proposed in [12]. The beamformer involves a generalized side-lobe canceller (GSC) structure and the AEC is implemented by applying the block least-mean-square (BLMS) procedure [13]. Another approach, combining a minimum variance distortionless response (MVDR) beamformer and a recursive least-squares (RLS)-based AEC is presented in [14].

A multi-channel echo cancellation is presented in [15], utilizing a low-complexity method. The method relies on a relative transfer function (RTF) scheme for multi-microphone AEC for reducing the overall computational load. Furthermore, it incorporates residual echo reduction into the beamformer design. This method is formulated in the STFT domain using the CTF [16] approximation.

Most studies in the literature assume that the physical distance between the far-end signals and the microphone location is small. It is a reasonable assumption since in many devices the microphones and the loudspeaker are mounted into the same device. However, when the loudspeaker is an external device connected by a cable or wirelessly by Bluetooth, it can be located anywhere in the room. As a result, the received echo signal may include a significant amount of reflections. In such cases, the length of the echo path should take into account the multiple acoustic reflections, implying a long adaptive filter. When the adaptive filter cannot entirely represent the echo path, the AEC output may suffer from a significant residual echo.

A single-microphone approach for jointly suppressing reverberation of the near-end speaker, residual echo and background noise is presented in [17]. A spectral postfilter was developed to efficiently dereverberte the desired speech signal, together with the suppression of the late residual echo and the background noise.

In [18], a two-microphone approach was presented. This algorithm comprises an adaptive filter to eliminate non-coherent signal components such as ambient noise and the reverberation of the near-end speech, in addition to echo cancellation. Another multichannel algorithm that jointly addresses the three problems is presented in [19]. An iterative expectation-maximization (EM technique is used for speech dereverberation, acoustic echo reduction, and noise reduction. The proposed method defines two state-space models, one for the acoustic echo-path and the other for the reverberated near-end speaker. The reverberant speech source model is assumed to follow a noiseless auto-regressive model. Two parameter optimization stages based on the Kalman smoother were applied to each state-space model in the E-step. The joint echo cancellation and dereverberation problem is also discussed in [20] for robot audition. An independent component analysis (ICA) scheme is adopted in order to provide a natural framework for these two problems using a microphone array.

The statistics of acoustic impulse response (AIR) is commonly used in dereverberation algorithms. A single-microphone method for the suppression of late room reverberation based on spectral subtraction is presented in [21]. This concept is extended to the multi-microphone case in [22]. The problem is formulated in the STFT domain while taking into account the contribution of the direct-path in [23].

Yoshioka et al. [24] developed an EM algorithm for derverberation and noise reduction, where the room impulse response (RIR) is modelled as an auto-regressive (AR) process in the STFT domain. An iterative and sequential Kalman expectation-maximization (KEM) scheme for single-microphone speech enhancement in the time-domain was introduced in [25]. This method was extended to a multi-microphone speech dereverberation method in [26], applied in the STFT domain, where the acoustic systems are approximated by the CTF model.

Many modern applications should address cases where the desired speaker, the microphone array and even the interference signal are moving, hence necessitating time-varying online parameter estimate. Unfortunately, the Wiener filter or the Kalman smoother cannot be straightforwardly applied in these cases, as they also utilize future samples. The statistical model of these algorithms should be adjusted to the dynamic scenario.

The REM, which is an efficient scheme for sequential parameter estimation, is particularly suitable for estimating time-varying parameters typical to dynamic scenarios. Titterington [27] formulated an online EM scheme using a stochastic approximation version of the modified gradient recursion. A recursive algorithm is proposed in [28] considering the convergence properties of Titterington's algorithm. The estimates generated by the recursive EM algorithm converged with probability one to a stationary point of the likelihood function. Recursive algorithms based on KEM were presented in [25, 29] using gradient decent algorithm for solving the maximum likelihood (ML) optimization. In [30], recursive EM methods for time-varying parameters were introduced with applications to multiple target tracking. Cappé and Moulines in [31] proposed another online version of the EM algorithm applicable to latent variable models of independent observations. A proof of convergence to a stationary point under certain additional conditions was established in this paper. For dependent observations, a recursive ML method was presented in [32] and is supported by a convergence proof. This method refers to state-space models in which the state process and the observations depend on a finite set of previous observations.

The acoustic path can be treated as stochastic processes under the Bayesian framework. An online EM based dereverberation algorithm is presented [33]. The acoustic paths where represented as random variables with a first-order Markov chain and estimated in the E-step by using the Kalman filter. The speech components were modelled as time-varying parameters and were estimated in the M-step.

An online algorithm for derverberation based on a Kalman expectation-maximization (RKEM) approach is presented in [34], where the acoustic parameters and the clean signal are jointly estimated. We refer to this algorithms as Kalman expectation-maximization for dereverberation (RKEMD). This framework is extended in the current contribution to jointly address echo cancellation, dereverberation and noise reduction problems.

1.3 Main contributions and outline

While most of the studies treat the problems of echo cancellation, dereverberation, and noise reduction separately, only a few propose a combined solution. In this paper, we present an online algorithm for the three problems addressing a unified statistical model using a microphone array. The microphone signal is degraded by an echo signal and an additive noise in highly reverberant environments. The proposed method is applied in the STFT domain using the RKEMD framework and simultaneously addresses all interfering sources. The acoustic systems of the near-end and far-end signals are approximated by the CTF model and the statistical model is represented in a state-space formulation. Using a a doubletalk detector (DTD), our method suspends the adaptation of the acoustic systems coefficients when their relevant signals are inactive, but still enables adaptation during double-talk. It is also capable of tracking time-variations of the acoustic systems. Hence, a feasible solution is provided in realistic dynamic scenarios when the near-end signal is moving, and even when the microphone array itself is moving.

The structure of the manuscript is as follows. In Section 2, the statistical model of the problem is presented. The recursive EM scheme is derived in Section 3. The desired near-end signal is estimated as a byproduct of the E-step of this scheme. In the recursive version, the E-step boils down to a Kalman filter that is applied to the observed signal with the estimated echo signal subtracted. In the M-step, the CTF coefficients and the noise parameters are recursively estimated. It is further shown that the instantaneous speech variance cannot be estimated using the REM procedure and an external estimator is derived instead. Section 4 describes the DTD that facilitates a proper implementation of the echo cancellation stage. An experimental study for different realistic scenario, including the challenging scenario of moving microphone array, was carried out at the Bar-Ilan acoustic lab and is detailed in Section 5. Conclusions are drawn in Section 6.

2 Statistical Model

Let $x[n]$ be the clean near-end signal and $y[n]$ be the far-end signal in the time-domain. The signals are propagating in an acoustic enclosure before being picked up by a J microphone array. The microphone signals are denoted by

$$z_j[n] = x[n] * h_j[n] + y[n] * g_j[n] + v_j[n] \ , \tag{1}$$

where $*$ denotes time-domain convolution and $j \in \mathcal{S}_J = [1, \ldots, J]$ is the microphone index. $h_j[n]$ and $g_j[n]$ are the RIRs relating $x[n]$ and $y[n]$ signals and the jth microphone, respectively. $v_j[n]$ is an additive noise, as received by jth microphone.

The signals $x[n]$ and $y[n]$ are represented in the STFT domain by $x(t,k)$ and $y(t,k)$, respectively, where $t \geq 1$ is the time-frame index and $k \in \mathcal{S}_K = [0, \ldots, K-1]$ is the frequency-bin index. We assume that the clean speech can be modelled as a complex-Gaussian variable, independent across STFT time-frames and frequencies (see [35]), with zero-mean and variance $\phi_x(t,k)$

$$x(t,k) \sim \mathcal{N}_C \left\{0, \phi_x(t,k)\right\} \ , \tag{2}$$

where $\mathcal{N}_C$ denotes a proper complex-Gaussian distribution.

In order to reduce the computational complexity and to facilitate the model analysis, we consider the CTF approximation [16] for the STFT representation of the time-domain RIR. The time-domain model in (1) can be approximated by

$$z_j(t,k) \approx \mathbf{h}_j^\top(k) \cdot \mathbf{x}_t(k) + \mathbf{g}_j^\top(k) \cdot \mathbf{y}_t(k) + v_j(t,k), \tag{3}$$

where the CTF systems are:

$$\begin{aligned} \mathbf{h}_j(k) &= \left[h_{j,L-1}(k), \ \ldots \ , h_{j,0}(k)\right]^\top , \\ \mathbf{g}_j(k) &= \left[g_{j,L-1}(k), \ \ldots \ , g_{j,0}(k)\right]^\top \end{aligned} \tag{4}$$

and the state-vectors of the desired speech signal and the acoustic reference signal are, respectively

$$\begin{aligned} \mathbf{x}_t(k) &= [x(t-L+1,k),\ \ldots\ ,x(t,k)]^\top, \\ \mathbf{y}_t(k) &= \left[y(t-L+1,k),\ \ldots\ ,y(t,k)\right]^\top. \end{aligned} \tag{5}$$

L is the length of CTF systems that depends on the reverberation time.

The noise signal $v_j(t,k)$ is assumed to be a stationary complex-Gaussian spatially uncorrelated random process,

$$v_j(t,k) \sim \mathcal{N}_C\left\{0, \phi_{v_j}(k)\right\} \tag{6}$$

and $E\left\{v_j(t,k)v_i^*(t,k)\right\} = 0$ for $j \neq i$.
For conciseness, the frequency index k will be omitted when no ambiguity arises.

The signal model can be represented in the following state-space form:

$$\begin{aligned} \mathbf{x}_t &= \Phi\mathbf{x}_{t-1} + \mathbf{w}_t, \\ \mathbf{d}_t &\triangleq \mathbf{z}_t - \mathbf{G}\mathbf{y}_t = \mathbf{H}\mathbf{x}_t + \mathbf{v}_t, \end{aligned} \tag{7}$$

where $\mathbf{x}_t$ and $\mathbf{y}_t$ were defined in (5) and $\mathbf{d}_t$ is defined as the observed signal after the subtraction of the echo signal contribution. The state-transition matrix is given by

$$\Phi \equiv \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix},$$

the innovation process is given by

$$\mathbf{w}_t \equiv [0,\ \ldots\ ,\ x(t)]^\top,$$

the measurement vector is given by

$$\mathbf{z}_t \equiv \left[z_1(t),\ \ldots\ ,\ z_J(t)\right]^\top,$$

the observation matrices are

$$\mathbf{H} \equiv \left[\mathbf{h}_1,\ \ldots\ ,\ \mathbf{h}_J\right]^\top\ ,\ \ \mathbf{G} \equiv \left[\mathbf{g}_1,\ \ldots\ ,\ \mathbf{g}_J\right]^\top,$$

with $\mathbf{h}_j$ and $\mathbf{g}_j$ the CTF systems, as defined in (4), and the noise vector is given by

$$\mathbf{v}_t \equiv \left[v_1(t),\ \ldots\ ,v_J(t)\right]^\top.$$

In the algorithm derivation, the following second-order statistics matrices of the innovation and measurement noise signals will also be used:

$$\mathbf{F}_t \equiv E\left\{\mathbf{w}_t\mathbf{w}_t^{\mathrm{H}}\right\} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi_x(t) \end{bmatrix}$$

$$\mathbf{B} \equiv E\left\{\mathbf{v}_t\mathbf{v}_t^{\mathrm{H}}\right\} = \begin{bmatrix} \phi_{v_{j1}} & \cdots & \cdots & 0 \\ 0 & \phi_{v_{j2}} & & \\ 0 & & \ddots & \\ 0 & \cdots & \cdots & \phi_{v_{jJ}} \end{bmatrix},$$

where we assumed that the noise is independent between microphones.

3 Algorithm derivation

The EM algorithm [36] is an iterative-batch procedure that processes the entire dataset in each iteration until convergence to a local maximum of the ML criterion. Hence, it cannot be applied as is to the task of AEC, specifically in time-varying scenarios. We therefore resort to a recursive version of the EM in our algorithm derivation.

3.1 The likelihood function

We start the algorithm derivation by defining the parameter sets and the relevant datasets. As we are interested in causal estimators, the available time-frame indexes for estimating the desired signal at frame t are confined to $\mathcal{S}_t = [1, \ldots, t]$, where $t = 1$ is arbitrarily chosen as the first available time-frame. The EM algorithm is a method for estimating a set of deterministic parameters that maximizes the likelihood criterion. Since the EM works with the notation of *complete-data* it also provides an estimate of the desired signal(s) as a by-product of the estimation procedure.

Let $\mathcal{Z}_t$ be the set of measurements comprising all microphones and all time-frequency (TF) bins:

$$\mathcal{Z}_t = \left\{z_j(\tau,k) : j \in \mathcal{S}_J,\ \tau \in \mathcal{S}_t,\ k \in \mathcal{S}_K\right\}, \tag{8}$$

$\mathcal{Y}_t$ the set of TF bins of the reference signal

$$\mathcal{Y}_t = \left\{y(\tau,k) : \tau \in \mathcal{S}_t,\ k \in \mathcal{S}_K\right\}, \tag{9}$$

and $\mathcal{X}_t$ the unavailable set of TF bins of the desired speech signal

$$\mathcal{X}_t = \{x(\tau,k) : \tau \in \mathcal{S}_t, k \in \mathcal{S}_K\}. \tag{10}$$

Both $\mathcal{Z}_t$ and $\mathcal{Y}_t$ are available, where the set $\mathcal{Z}_t$ describes the available information in microphone signals, and $\mathcal{Y}_t$ the information in far-end signal as transmitted by the local loudspeaker.

The parameter sets of the statistical model presented in Section 2 comprises the following subsets:

$$\begin{aligned} &\boldsymbol{\Theta} \equiv \{\boldsymbol{\Theta}_X, \boldsymbol{\Theta}_H, \boldsymbol{\Theta}_G, \boldsymbol{\Theta}_V\} \\ &\boldsymbol{\Theta}_X \equiv \{\phi_x(t,k)\}_{t,k}, \;\; \boldsymbol{\Theta}_H \equiv \left\{\mathbf{h}_j(k)\right\}_{j,k}, \\ &\boldsymbol{\Theta}_G \equiv \left\{\mathbf{g}_j(k)\right\}_{j,k}, \quad \boldsymbol{\Theta}_V \equiv \left\{\phi_{v_j}(k)\right\}_{j,k} \end{aligned} \tag{11}$$

for all $j \in \mathcal{S}_J$; $t \geq 1$ and $k \in \mathcal{S}_K$.

A note on the time-dependency of the parameters is in place. Two distinct time scales can be defined. While the speech power spectral density (PSD) is rapidly changing from frame to frame, the RIRs relating the desired speech and the echo signal to the microphones, as well as the noise variances, are slowly time-varying. The distinct scales of the time variations imply different types of estimation procedures. While estimating the speech PSD necessitates an external smoothing procedure that maintains the rapid time-variations, estimating the RIRs and the noise variances boils down to recursive aggregation of past statistics. Consequently, slowly time-varying estimated parameters are obtained. In the following sections, estimators for the set of parameters $\boldsymbol{\Theta}$ will be presented in details together with an online estimate of the desired speech signal.

The EM formulation requires the log-likelihood of the *complete-data*. Under the assumed statistical model, it is given by:

$$\begin{aligned} &\log f(\mathcal{X}_t, \mathcal{Z}_t | \mathcal{Y}_t; \boldsymbol{\Theta}) = \\ &\quad \log f(\mathcal{X}_t | \mathcal{Y}_t; \boldsymbol{\Theta}) + \log f(\mathcal{Z}_t | \mathcal{X}_t, \mathcal{Y}_t; \boldsymbol{\Theta}) \stackrel{C}{=} \\ &\quad -\frac{1}{2}\sum_{\tau=1}^{t}\left[\log(\phi_x(\tau)) + \frac{|x(\tau)|^2}{\phi_x(\tau)}\right] \\ &\quad\quad -\frac{1}{2}\sum_{j=1}^{J}\sum_{\tau=1}^{t}\left[\log\phi_{v_j} + \frac{1}{\phi_{v_j}}\left|z_j(\tau) - \mathbf{q}_j^\top \boldsymbol{\mu}_\tau\right|^2\right], \end{aligned} \tag{12}$$

where

$$\mathbf{q}_j \equiv \begin{bmatrix} \mathbf{h}_j^* \\ \mathbf{g}_j^* \end{bmatrix}, \;\; \boldsymbol{\mu}_\tau \equiv \begin{bmatrix} \mathbf{x}_\tau \\ \mathbf{y}_\tau \end{bmatrix} \tag{13}$$

and $\stackrel{C}{=}$ stands for equal up to constants that are independent of $\boldsymbol{\Theta}$. Note that the second and the third lines of (12) are the log-likelihood of the clean speech signal and the log-likelihood of the additive noise, respectively. Both terms are expressed as a summation over the time-frame index $\tau \in \mathcal{S}_t$, as a result of the independence between time-frames of the desired source and the noise signals in the STFT domain. The second term also decomposes to a sum over the J microphones due to the assumed independence of the noise signals across microphones. The likelihood function in (12) is separately calculated for all $k \in \mathcal{S}_K$ due to the independence between frequency bins.

3.2 Recursive EM algorithm

We adopt the online EM formulation presented in [31], in which the auxiliary function is recursively calculated, while the maximization step remains intact. This formulation facilitates online and time-varying estimation of all model parameters.

The auxiliary function at time-frame t is given by a weighted sum of the auxiliary function at the previous time-frames and the innovation of the current measurement:

$$\begin{aligned} &Q\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right) = Q\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t-1)\right) + \\ &\quad \gamma_t \cdot \left\{E\left\{\log\left[f\left(\mathbf{x}_t, \mathbf{z}_t \middle| \mathbf{y}_t; \boldsymbol{\Theta}\right)\right]\middle|\mathcal{Z}_t, \mathcal{Y}_t; \widehat{\boldsymbol{\Theta}}(t)\right\}\right. \\ &\quad\quad\quad \left. - Q\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t-1)\right)\right\}, \end{aligned} \tag{14}$$

where $\widehat{\boldsymbol{\Theta}}(t)$ is the parameter set estimate after measuring the observation $\mathbf{z}_t$ and the far-end echo signal $\mathbf{y}_t$ at time-frame t, and $\gamma_t \in [0, 1)$ is a smoothing parameter, that should decay in time for static scenarios. The maximization is computed over the aggregated auxiliary function (14)

$$\widehat{\boldsymbol{\Theta}}(t+1) = \arg\max_{\boldsymbol{\Theta}} \left\{Q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right]\right\}. \tag{15}$$

Given the measurements and the echo signal, define the expected value of the *instantaneous* complete-data log-likelihood[1]:

$$\rho\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right) \equiv E\left\{\log\left[f\left(\mathbf{x}_t, \mathbf{z}_t\middle|\mathbf{y}_t; \boldsymbol{\Theta}\right)\right]\middle|\mathcal{Z}_t, \mathcal{Y}_t; \widehat{\boldsymbol{\Theta}}(t)\right\}, \tag{16}$$

and substitute the time-varying smoothing parameter with a constant factor $\beta = 1 - \gamma_t$, thus introducing an exponential decay of the contribution of past samples to the calculation, and consequently facilitating recursive estimation of time-varying parameters. Using these definitions, the recursive auxiliary function (14) can be rewritten as

$$\begin{aligned} &Q\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right) = \\ &\quad \beta \cdot Q\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t-1)\right) + (1-\beta)\rho\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right) = \\ &\quad (1-\beta)\sum_{\tau=1}^{t}\beta^{t-\tau}\rho\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(\tau)\right). \end{aligned} \tag{17}$$

[1] In their original contribution, Cappé and Moulines [31] assume independent and identically distributed measurements. This assumption does not hold in our measurement model. We therefore propose to use a slightly different model in which the expectation of the instantaneous complete data is also conditioned on past measurements, namely on $\mathcal{Z}_t, \mathcal{Y}_t$ rather than only on $\mathbf{z}_t, \mathbf{y}_t$. While a proof of such formulation is beyond the scope of this contribution, we note that similar formulations were successfully used in the context of speech processing [25, 34, 37]. To shed more light on underlying mathematical foundations of stochastic approximation, the interested reader is also referred to a comprehensive review on the topic [38].

The complete-data likelihood is independent and identically distributed between time frames. Therefore, we can explicitly write (16) as:

$$\rho\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right) = -\frac{1}{2}\left[\log(\phi_x(t)) + \frac{\widehat{|x(t)|^2}}{\phi_x(t)}\right] - \frac{1}{2}\sum_{j=1}^{J}\left[\log\phi_{v_j} + \frac{1}{\phi_{v_j}}\widehat{\left|z_j(t) - \mathbf{q}_j^\top\boldsymbol{\mu}_t\right|^2}\right]. \quad (18)$$

Finally, the explicit recursive auxiliary function can be calculated by substituting (18) into (17):

$$Q\left(\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right) = -\frac{1-\beta}{2}\sum_{\tau=1}^{t}\beta^{t-\tau}\left[\log\phi_x(\tau) + \frac{\widehat{|x(\tau)|^2}}{\phi_x(\tau)}\right] - \frac{1-\beta}{2}\sum_{\tau=1}^{t}\sum_{j=1}^{J}\beta^{t-\tau}\left[\log\phi_{v_j} + \frac{1}{\phi_{v_j}}\left\{|z_j(\tau)|^2 - 2\Re\left(\mathbf{q}_j{}^\top\hat{\boldsymbol{\mu}}_\tau z_j^*(\tau)\right) + \mathbf{q}_j{}^\top\widehat{\boldsymbol{\mu}_\tau\boldsymbol{\mu}_\tau^{\mathrm{H}}}\mathbf{q}_j{}^*\right\}\right], \quad (19)$$

where

$$\hat{\boldsymbol{\mu}}_\tau \equiv \begin{bmatrix}\hat{\mathbf{x}}_\tau \\ \mathbf{y}_\tau\end{bmatrix}, \widehat{\boldsymbol{\mu}_\tau\boldsymbol{\mu}_\tau^{\mathrm{H}}} \equiv \left[\begin{array}{c|c}\widehat{\mathbf{x}_\tau\mathbf{x}_\tau{}^{\mathrm{H}}} & \hat{\mathbf{x}}_\tau\mathbf{y}_\tau{}^{\mathrm{H}} \\ \hline \mathbf{y}_\tau\hat{\mathbf{x}}_\tau^{\mathrm{H}} & \mathbf{y}_\tau\mathbf{y}_\tau{}^{\mathrm{H}}\end{array}\right]. \quad (20)$$

and the first- and second-order statistics of the near-end speech signal given $\mathcal{Z}_t$ and $\mathcal{Y}_t$ are:

$$\hat{\mathbf{x}}_t \equiv E\left\{\mathbf{x}_t\middle|\mathcal{Z}_t,\mathcal{Y}_t;\widehat{\boldsymbol{\Theta}}(t)\right\} \equiv \hat{\mathbf{x}}_{t|t}, \quad (21a)$$

$$\widehat{\mathbf{x}_t\mathbf{x}_t{}^{\mathrm{H}}} \equiv E\left\{\mathbf{x}_t\mathbf{x}_t{}^{\mathrm{H}}\middle|\mathcal{Z}_t,\mathcal{Y}_t;\widehat{\boldsymbol{\Theta}}(t)\right\} \equiv \hat{\mathbf{x}}_{t|t}\hat{\mathbf{x}}_{t|t}^{\mathrm{H}} + \mathbf{P}_{t|t}, \quad (21b)$$

$$\widehat{|x(t)|^2} \equiv E\left\{|x(t)|^2\middle|\mathcal{Z}_t,\mathcal{Y}_t;\widehat{\boldsymbol{\Theta}}(t)\right\}. \quad (21c)$$

3.2.1 E-Step: Kalman filter

The calculation of the recursive auxiliary function (19) requires the first- and second-order statistics of the clean speech signal (21). These are acquired in the E-step of the recursive procedure by applying the Kalman filter. The Kalman filter, summarized in Algorithm 1, is the optimal *causal* estimator in minimum mean square error (MMSE) sense.

3.2.2 M-Step: Parameter estimation

In the M-step, we update parameters by maximizing the auxiliary function w.r.t. $\boldsymbol{\Theta}$ yielding the subsequent estimate $\widehat{\boldsymbol{\Theta}}(t+1)$,

$$\widehat{\boldsymbol{\Theta}}(t+1) = \arg\max_{\boldsymbol{\Theta}}\left\{Q\left[\boldsymbol{\Theta}\middle|\widehat{\boldsymbol{\Theta}}(t)\right]\right\}, \quad (22)$$

Algorithm 1: The Kalman Filter.

Kalman filter:
for $t \geq 1$ **do**
  Predict:
  $\hat{\mathbf{x}}_{t|t-1} = \Phi\cdot\hat{\mathbf{x}}_{t-1|t-1}$
  $\mathbf{P}_{t|t-1} = \Phi\cdot\mathbf{P}_{t-1|t-1}\cdot\Phi^\top + \mathbf{F}_t$
  Update:
  $\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}^{\mathrm{H}}\left[\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}^{\mathrm{H}} + \mathbf{B}\right]^{-1}$
  $\mathbf{e}_t = \mathbf{d}_t - \mathbf{H}\hat{\mathbf{x}}_{t|t-1}$
  $\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t\cdot\mathbf{e}_t$
  $\mathbf{P}_{t|t} = [\mathbf{I} - \mathbf{K}_t\mathbf{H}]\,\mathbf{P}_{t|t-1}$
end

resulting in the following update rules for the model parameters at the $(t+1)$-th time-frame:

$$\widehat{\mathbf{q}}_j(t+1) = \begin{bmatrix}\widehat{\mathbf{h}}_j^*(t+1) \\ \widehat{\mathbf{g}}_j^*(t+1)\end{bmatrix} = \left[\widehat{\mathbf{R}}_{\mu\mu}^{(t)}\right]^{-1}\widehat{\mathbf{r}}_{\mu z_j}^{(t)}, \quad (23)$$

$$\widehat{\phi}_{v_j}(t+1) = \frac{1}{1-\beta^{t+1}}\left\{r_{z_jz_j}^{(t)} - 2\Re\left[\widehat{\mathbf{q}}_j^\top(t)\widehat{\mathbf{r}}_{\mu z_j}^{(t)}\right] + \widehat{\mathbf{q}}_j^\top(t)\widehat{\mathbf{R}}_{\mu\mu}^{(t)}\widehat{\mathbf{q}}_j^*(t)\right\}, \quad (24)$$

where we define the following aggregated second-order statistics

$$\widehat{\mathbf{R}}_{\mu\mu}^{(t)} \equiv \begin{bmatrix}\widehat{\mathbf{R}}_{xx}^{(t)} & \widehat{\mathbf{R}}_{xy}^{(t)} \\ \widehat{\mathbf{R}}_{xy}^{\mathrm{H}\,(t)} & \mathbf{R}_{yy}^{(t)}\end{bmatrix}, \widehat{\mathbf{r}}_{\mu z_j}^{(t)} \equiv \begin{bmatrix}\widehat{\mathbf{r}}_{xz_j}^{(t)} \\ \mathbf{r}_{yz_j}^{(t)}\end{bmatrix}, \quad (25)$$

with

$$\widehat{\mathbf{R}}_{xx}^{(t)} \equiv (1-\beta)\sum_{\tau=1}^{t}\beta^{t-\tau}\widehat{\mathbf{x}_\tau\mathbf{x}_\tau{}^{\mathrm{H}}} = \beta\cdot\widehat{\mathbf{R}}_{xx}^{(t-1)} + (1-\beta)\widehat{\mathbf{x}_t\mathbf{x}_t{}^{\mathrm{H}}} \quad (26)$$

and similarly

$$\begin{aligned}\mathbf{R}_{yy}^{(t)} &\equiv \beta\cdot\mathbf{R}_{yy}^{(t-1)} + (1-\beta)\mathbf{y}_t\mathbf{y}_t{}^{\mathrm{H}} \\ \widehat{\mathbf{R}}_{xy}^{(t)} &\equiv \beta\cdot\widehat{\mathbf{R}}_{xy}^{(t-1)} + (1-\beta)\hat{\mathbf{x}}_t\mathbf{y}_t{}^{\mathrm{H}} \\ \widehat{\mathbf{r}}_{xz_j}^{(t)} &\equiv \beta\cdot\widehat{\mathbf{r}}_{xz_j}^{(t-1)} + (1-\beta)\hat{\mathbf{x}}_t z_j^*(t) \\ \mathbf{r}_{yz_j}^{(t)} &\equiv \beta\cdot\mathbf{r}_{yz_j}^{(t-1)} + (1-\beta)\mathbf{y}_t z_j^*(t) \\ r_{z_jz_j}^{(t)} &\equiv \beta\cdot r_{z_jz_j}^{(t)} + (1-\beta)|z_j(t)|^2.\end{aligned} \quad (27)$$

Note that (23) is an RLS update rule for estimating both filters and (24) is a recursive estimation of the residual power.

Unlike the estimation procedure of the filters' coefficients, maximizing (22) w.r.t. the speech PSD cannot be applied. In 3.2.3, we explain the reasons for this phenomenon and propose an alternative algorithm for the recursive speech PSD estimation.

3.2.3 Recursive estimation of the speech variance

The speech variance $\phi_x(t)$ is a time-varying parameter, due to the non-stationarity of the speech signal, and hence smoothness over time cannot be assumed, in contrast to the CTF systems $\mathbf{H}$ and $\mathbf{G}$ and the noise variance ϕ_{v_j} that exhibit slower time-variations. In the proposed recursive algorithm, the available observed data refers to the time frames in the interval $\mathcal{S}_t$, thus the derivative of (22) w.r.t. $\phi_x(t+1)$ is zero and does not impose any constraint.

Alternately, we propose to obtain a speech PSD estimator of $\widehat{\phi}_x(t)$, which still maintains some smoothness of the PSD estimates. The spectral amplitude estimator presented in [39] is adapted for this estimation with the necessary changes to incorporate residual echo and reverberation. The optimal speech PSD estimator in the MMSE sense at the jth microphone signal:

$$\widehat{\phi}_{x_j}(t) = \left|\hat{h}_{j,0}(t)\right|^{-2} A_j^2(t)|z_j(t) - \mathbf{g}_j^\top \mathbf{y}_t|^2 \tag{28}$$

$$\approx E\left\{|x_j(t)|^2 \middle| \mathcal{Z}_t, \mathcal{Y}_t; \widehat{\boldsymbol{\Theta}}(t)\right\}, \tag{29}$$

where $A_j(t)$ is a gain function that attenuates the late reverberant component and the noise component. Consequently, $A_j^2(t)|z_j(t) - \mathbf{g}_j^\top \mathbf{y}_t|^2$ represents the variance estimator of the early speech component, $x_j^e(t) = h_{j,0}(t)x(t)$. The gain function is defined as

$$A_j^2(t) = \max\left[\frac{\zeta_{\mathrm{prior},j}(t)}{\zeta_{\mathrm{prior},j}(t)+1}\left(\frac{1+\nu_j(t)}{\zeta_{\mathrm{post},j}(t)}\right), A_{\min}^2\right], \tag{30}$$

where

$$\zeta_{\mathrm{prior},j}(t) \equiv \frac{\widehat{\phi}_{x_j^e}(t)}{\widehat{\phi}_{r_j}(t) + \widehat{\phi}_{v_j}(t)}, \tag{31}$$

$$\zeta_{\mathrm{post},j}(t) \equiv \frac{|z_j(t) - \widehat{\mathbf{g}}_j^\top(t)\mathbf{y}_t|^2}{\widehat{\phi}_{r_j}(t) + \widehat{\phi}_{v_j}(t)}, \tag{32}$$

$$\nu_j(t) = \frac{\zeta_{\mathrm{prior},j}(t)}{1+\zeta_{\mathrm{prior},j}(t)}.$$

and ϕ_{r_j} is the late reverberant spectral variance. Note that $\zeta_{\mathrm{prior},j}$ and $\zeta_{\mathrm{post},j}$ are a priori and a posteriori signal to interference ratio (SIR), respectively. The calculation of (30) is executed for every channel j. The estimation of $\widehat{\phi}_{x_j^e}(t)$ is unobserved and therefore the a priori SIR, $\zeta_{\mathrm{prior},j}(t)$, is estimated by the *decision-directed* estimator proposed by Ephraim and Malah in [40]:

$$\begin{aligned}\bar{\zeta}_{\mathrm{prior},j}(t) =& \alpha_{\mathrm{sir}} A_j^2(t-1)\zeta_{\mathrm{post},j}(t-1) + \\ & [1-\alpha_{\mathrm{sir}}]\max[\zeta_{\mathrm{post},j}(t) - 1, \zeta_{\min}],\end{aligned} \tag{33}$$

where α_{sir} is a smoothing factor and $\zeta_{\min}$ is the minimum SIR that ensures the positiveness of $\zeta_{\mathrm{post},j}(t) - 1$. Note that applying the gain function in (30) on $\zeta_{\mathrm{post},j}(t-1)$ as in (33) represents the a priori SIR resulting from the previous frame process.

For the estimation of late reverberant spectral variance ϕ_{r_j}, the instantaneous power of the reverberation $\widehat{\psi}_{r_j}(t)$ is calculated as in the RKEMD method [34]:

$$\widehat{\psi}_{r_j}(t) = \widehat{\mathbf{h}}_j^\top(t)\boldsymbol{\Phi}\widehat{\mathbf{x}_{t-1}\mathbf{x}_{t-1}^{\mathrm{H}}}\boldsymbol{\Phi}^\top\widehat{\mathbf{h}}_j^*(t), \tag{34}$$

By the definition of $\boldsymbol{\Phi}$, $\hat{h}_{j,0}(t)$ is excluded from (34) and hence only the variance of the late reverberation is taken into account. Then, $\widehat{\phi}_{r_j}(t)$ is estimated by time smoothing using a smoothing parameter $\alpha_r \in [0,1)$:

$$\widehat{\phi}_{r_j} = \alpha_r\widehat{\phi}_{r_j}(t-1) + (1-\alpha_r)\widehat{\psi}_{r_j}(t). \tag{35}$$

The speech PSD $\widehat{\phi}_x(t)$ is finally determined by averaging over all J channels:

$$\widehat{\phi}_x(t) = \frac{1}{J}\sum_{j=1}^{J}\widehat{\phi}_{x_j}(t). \tag{36}$$

It is clear that the presented model in (3) may suffer from *gain ambiguity* in estimating both $\widehat{\phi}_x(t)$ and $\widehat{\mathbf{h}}_j(t)$, attributed to the following equality:

$$\mathbf{h}_j^\top(t,k)\mathbf{x}_t(k) = \left[\nu(t,k)\mathbf{h}_j^\top(t,k)\right]\left[\frac{1}{\nu(k)}\mathbf{x}_t(k)\right], \tag{37}$$

where $\nu(t,k)$ is an arbitrary time- and frequency-dependent gain. To circumvent this problem, we arbitrarily set $|\hat{h}_{j,0}(t,k)| = 1, \forall j$ in (28).

3.3 Alternative M-step 1

Estimating the CTF systems in the M-step (23) boils down to RLS-type update rule. An alternative and commonly used approach for adaptive filtering is the normalized least-mean-square (NLMS) procedure, which is known for its good tracking capabilities, simplicity, and low computational complexity. Conversely, the RLS algorithm is more stable and its convergence rate is faster, at the expense of high computational complexity. The trade-off between fast adaptation and computational complexity should be considered when choosing the appropriate adaptive filtering approach. We develop in the sequel an alternative M-step based on the NLMS procedure.

First, we apply the NLMS procedure for estimating the echo path for each microphone $\mathbf{g}_j, \forall j \in \mathcal{S}_J$ rather than using the estimate resulting from M-step stage in (23). The NLMS update rule is given by:

$$\widehat{\mathbf{g}}_j^{\mathrm{NLMS}}(t+1) = \widehat{\mathbf{g}}_j^{\mathrm{NLMS}}(t) + \lambda\frac{\mathbf{y}_t e_j(t)}{\mathbf{y}_t\mathbf{y}_t^{\mathrm{H}} + \delta_{\mathrm{NLMS}}}, \tag{38}$$

where $\lambda \in (0,2)$ is the step-size, $\delta_{\mathrm{NLMS}} > 0$ is the regularization factor and $e_j(t)$ is the instantaneous estimation error w.r.t. the jth microphone given by:

$$e_j(t) = z_j^*(t) - \mathbf{y}_t^{\mathrm{H}}(\widehat{\mathbf{g}}_j^{\mathrm{NLMS}})^*(t). \tag{39}$$

The update of the other acoustic parameters remains intact and is calculated as described in Section 3.2.2.

Substituting the CTF estimate of the echo path $\widehat{\mathbf{g}}_j$ (23) by $\widehat{\mathbf{g}}_j^{\mathrm{NLMS}}$ leads to a combined structure of NLMS and RKEMD, where the NLMS estimation error of each channel is the input for RKEMD. This new scheme is denoted by NLMS-RKEMD-1.

3.4 Alternative M-step 2

Although the RLS approach in the proposed algorithm is inefficient in means of computational complexity comparing to NLMS, the EM has the advantage of considering the near-end speaker in the echo cancellation model. We therefore introduce another alternative M-step, in which the echo path is estimated using NLMS while still utilizing the benefits offered by the EM formulation. Based on a gradient-descent minimization of the likelihood function, adopted from [41] and [25], we substitute the maximization of $\widehat{\mathbf{g}}_j$ in (23) with:

$$\tilde{\mathbf{g}}_j^{\mathrm{NLMS}}(t+1) = \tilde{\mathbf{g}}_j(t) + \lambda \frac{\partial}{\partial \tilde{\mathbf{g}}_j(t)} Q\left[\boldsymbol{\Theta} \middle| \widehat{\boldsymbol{\Theta}}(t)\right]. \tag{40}$$

Explicitly, carrying out the derivative in (40) (also implementing the normalization operation) yields an adaption rule similar to (38), but with a different error term:

$$\tilde{e}_j(t) = z_j^*(t) - \mathbf{x}_t^{\mathrm{H}} \widehat{\mathbf{h}}_j^*(t) - \mathbf{y}_t^{\mathrm{H}} \left(\tilde{\mathbf{g}}_j^{\mathrm{NLMS}}\right)^*(t)\,. \tag{41}$$

Now, the error signal (41) includes the subtraction of the estimated reverberant near-end signal. We denote this recursive EM variant as NLMS-RKEMD-2.

4 Double talk detector

The statistical model presented in Section 2 assumes a constant activity of the near-end and far-end signals. However, in real scenarios this is not always the case, rendering the statistical modelling inaccurate. To circumvent this intermittency problem, we propose to adopt a DTD to detect the presence of the near-end signal, and to stop the adaptation of the parameters of the CTF model during inactive periods.

We propose to use the normalized cross-correlation method presented in [42], based on the correlation level between the far-end signal and the echo signal, that drops when the near-end signal is active. After some derivation, the decision variable is obtained by:

$$\xi_{t+1} = \frac{\sqrt{\hat{\mathbf{g}}_1^{\mathrm{H}} \mathbf{R}_{yy}^{(t)} \hat{\mathbf{g}}_1}}{\sqrt{\hat{\mathbf{g}}_1^{\mathrm{H}} \mathbf{R}_{yy}^{(t)} \hat{\mathbf{g}}_1 + \widehat{\phi}_x(t)}} \tag{42}$$

where $\hat{\mathbf{g}}_1$ is the CTF estimate at the first microphone. If $\xi_t < \eta$, then a double-talk is detected. Note that ξ_t is calculated using the parameter estimates in previous frame in order to freeze the adaptation in the current frame.

As noted in [43], a *fixed* value of η is not capable of addressing practical scenarios and that an *adaptive* threshold should be used instead:

$$\eta_t = \begin{cases} \eta_{t-1} + \psi_t, & \textit{if } \tilde{\xi}_t > \eta_{t-1} \\ \alpha_d \eta_{t-1} + \\ \quad (1-\alpha_d)(\tilde{\xi}_t - \sqrt{\Sigma_{t-1}}) - \psi_t, & \text{otherwise,} \end{cases} \tag{43}$$

and

$$\Sigma_t = \begin{cases} \Sigma_{t-1}, & \textit{if } \tilde{\xi}_t > \eta_{t-1} \\ \alpha_d \Sigma_{t-1} + (1-\alpha_d)(\tilde{\xi}_t - \eta_t)^2, & \text{otherwise} \end{cases} \tag{44}$$

where $\tilde{\xi}_t$ is minimum ξ_t across the frequency bins in frame t and α_d is a smoothing factor. ψ_t is a small value that was set as $0.002\sqrt{\Sigma_{t-1}}$.

The proposed EM algorithm for echo cancellation, dereverberation an noise reduction, is summarized in Algorithm 2.

Algorithm 2: Kalman-EM algorithm for echo cancellation, dereverberation and noise reduction (RLS version for the M-step).

---

for $t \geq 1$ **do**

1 **DTD:**
 (a) Calculate $\tilde{\xi}_t$ (42), η_t (43) and Σ_t (44).
 (b) Apply the decision rule to detect double-talk.

2 **Speech variance estimation:**
 (a) Calculate $\widehat{\phi}_{r_j}(t)$ (35) and $|z_j(t) - \widehat{\mathbf{g}}_j^{\top}(t)\mathbf{y}_t|^2$.
 (b) Estimate $\widehat{\phi}_x(t)$ (28).

3 **E-step:**
 (a) Echo cancellation:
 Calculate the residual $\mathbf{d}_t = \mathbf{z}_t - \widehat{\mathbf{G}}(t)\mathbf{y}_t$.
 (b) Dereverberation:
 Apply one step of Kalman filtering to the observation $\mathbf{d}_t$ in order to obtain $\hat{\mathbf{x}}_t$ (21a) and $\widehat{\mathbf{x}_t\mathbf{x}_t^{\mathrm{H}}}$ (21b).

4 **M-step:**
 (a) Calculate $\widehat{\mathbf{R}}_{\mu\mu}^{(t)}$ and $\widehat{\mathbf{r}}_{\mu z_j}^{(t)}$ (25)
 (b) Update the acoustic parameters $\widehat{\mathbf{q}}_j(t+1)$ (23) $\widehat{\phi}_{v_j}(t+1)$ (24)

end

5 Performance evaluation

5.1 Setup

The proposed method was evaluated in two dynamic scenarios. The experiments were recorded at the Acoustic Signal Processing Lab, Bar-Ilan University. The room dimensions are $6 \times 6 \times 2.4$ m (length $\times$ width $\times$ height). The reverberation time of the room was set to 650 ms, by adjusting the rooms panels.

The sampling rate was set to 16 kHz, the STFT analysis window is set to a 32 ms Hamming window, with 75% overlap between adjacent time-frames. Avargel et al. [16] define the CTF length L according to the time-domain filter length, the STFT analysis window length and the overlap. The length of the time-domain filter, the RIR in our problem, is determined by the room reverberation time. We set the RIR length to be 650 ms, similar to the reverberation time. Consequently, L was set to 35 frames. Note that setting L to an excessively high value may result in estimation errors and as well as a high computational complexity. Setting L to a lower value than implied by [16], degrades the CTF approximation and can lead to partial dereverberation.

The desired clean speech estimator, $\hat{x}(t)$, was further enhanced by applying a high pass filter to remove frequencies lower than 200 Hz. Finally, the parameters depicted in Table 1 were fixed for all simulations and experiments.

5.2 Experiments using real speakers

For demonstrating the capabilities of our method in realistic cases, we carried out two types of experiments involving human speakers that read out loud sentences and a loudspeaker that plays music. We tested the performance in two scenarios. In Scenario #1, the loudspeaker and the microphone array are static and the subject is moving in the room along a predefined path. In Scenario #2, the loudspeaker and the subject are static and the microphone array is manually moving. Both scenarios are depicted in Fig. 1.

The subjects in the experiments were native English speakers. Two females and three males participated in Scenario #1, and two females and two males in scenario #2. Several recordings of modern music, consisting of musical instruments and a singer, were played throughout the recording session. The SIR in Scenario #1 is set to 5 dB. The average of the measured SIR in Scenario #2 is 4.68 dB.

During the experiments, we tested 2 types of noise. The first is an air-conditioner (AC) noise. The second is a pseudo-diffused babble noise, played from 4 loudspeakers, facing the room walls. In Scenario #1, the reverberated-signal to noise ratio (RSNR) is set to 15 dB. For Scenario #2 the RSNR is time-varying. The average RSNR is 6.62 dB for the AC noise and 9.5 dB for the babble noise.

Table 1 The algorithm parameters

Parameters	Value	Parameters	Value
β	0.99	$A_{\min}$	0.2
α_{sir}	0.2	α_d	0.99
α_r	0.5	λ	1
$\zeta_{\min}$	0.5		

5.3 Baseline methods

We propose to compare the proposed algorithm to a cascade implementation of AEC and a dereverberation algorithm. For the echo cancellation, we applied J instances of a conventional NLMS algorithm to mitigate the echo path relating the far-end signal and each of the microphones. For each frame, the signals at the J outputs of the AECs are further processed by multichannel spectral enhancement (MCSE) algorithm [44]. We denote this approach as NLMS-MCSE. In addition, we present the results of the proposed algorithm considering the alternative M-steps presented in Sections 3.3 and 3.4, NLMS-RKEMD-1 and NLMS-RKEMD-2, respectively. We also refer to the performance of a simple NLMS, without considering any dereverberation approach.

The DTD algorithm that was discussed in Section 4, was also utilized in the implementation of NLMS-based methods. During double talk, the NLMS adaptation is suspended in NLMS-MCSE and NLMS-RKEMD-1. This is in contrast to our method that enables the adaptation of the CTF coefficients also during double talk. Adaptation is only suspended if the relevant signals are inactive.

For the NLMS-MCSE method, $\widehat{\phi_x}(t)$ was substituted by $|\hat{e}_j(t)|^2$ in the detection function (42). In Scenario #2, the echo path is constantly changing during the double talk. Hence, suspending the adaptation during double talk degrades significantly the echo cancellation performance. Ignoring the DTD and allowing adaptation, despite the interfering effect of the near-end speaker to the NLMS convergence, is preferred in this case.

5.4 Speech quality and intelligibility

Two objective measures are used for evaluating the speech quality and intelligibility, namely the log-spectral distortion (LSD) and the short-time objective intelligibility (STOI) [45], respectively.

The LSD between x and $\tilde{z} \in \{z_1, \hat{x}\}$ is calculated for each time frame as:

$$\mathrm{LSD}(t) = \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left[10\log_{10}\left(\frac{\max\{|x(t,k)|, \epsilon(x)\}}{\max\{|\tilde{z}(t,k)|, \epsilon(\tilde{z})\}}\right)\right]^2} \tag{45}$$

(a) Scenario #1: the subject walks from point 1 to point 2, then sits for 2 seconds. Then walks to point 3.

(b) Scenario #2: the microphone array is manually moved from point 1 to point 2.

(c) The microphone array. The leftmost located microphone is Mic.#1.

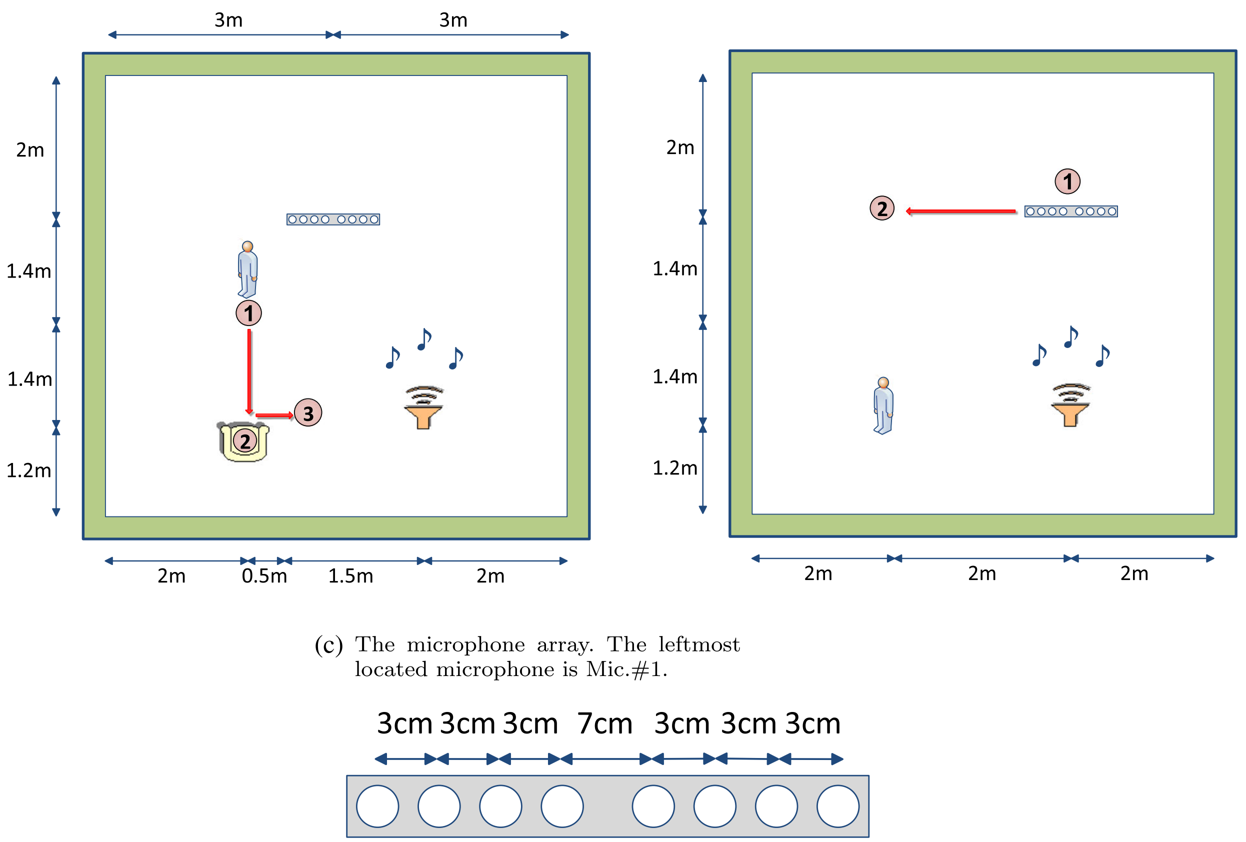

Fig. 1 Experimental setup

where the minimum value is calculated by $\epsilon(C) = 10^{-50/10} \max_{t,k} |C(t,k)|$, which limits the log-spectrum dynamic range of C to about -50dB. The presented value of the LSD is the median value of LSD(t) over all time-frames.

In addition, the dereverberation capabilities of the examined algorithms were evaluated using the SRMR measure [46].

The LSD, SRMR and STOI results for Scenario #1 are presented in Fig. 2. These plots describe the statistics of the measures over 140 experiments, including 5 different speakers, 2 sentences ($20 - 25$ s each), 2 types of noise and 7 songs were played as far-end speaker. The speech quality, intelligibility and dereverberation measures for Scenario #2 are described in Table 2. The table reports the median values of 16 experiments with 4 different speakers, 2 sentences ($20 - 25$ s each), 2 types of noise and 1 song was played as the far-end speaker (different song for every speaker). Whisker plots are not informative enough for this amount of data.

It is evident that the proposed method outperforms the competing algorithms in all measures for both scenarios. The estimated speech of the NLMS-based algorithms in Scenario #2 is severely distorted as compared with Scenario #1. Indeed, the NLMS-MCSE algorithm exhibits comparable performance to the proposed method in scenario #1, but in the more challenging experiment, namely scenario #2, the proposed method significantly outperforms all baseline methods as evident from Table 2. The degradation in NLMS-RKEMD-1 and NLMS-MCSE can be explained by the fact that in Scenario #2, the NLMS keeps updating the echo path during double talk. In contrast, in Scenario #1, the adaptation is suspended. Therefore, the performance gap between the proposed method and its competitors is more pronounced in Scenario #2.

In addition, we observed that the other methods are more sensitive than the proposed method to errors in the DTD. The mis-detection and false-alarm of the DTD lead to severe performance degradation in the NLMS-based methods and consequently results in reduction in speech

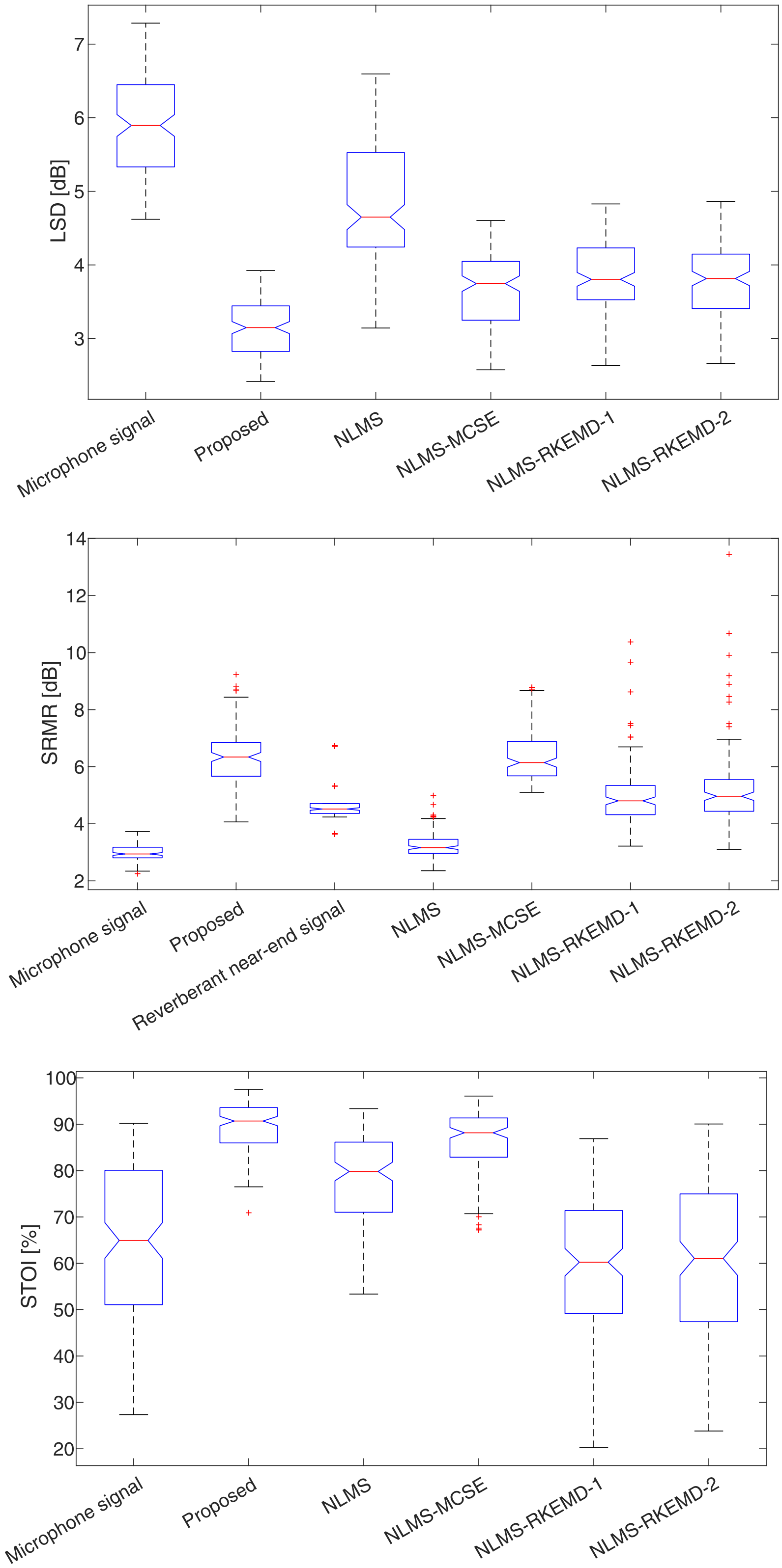

Fig. 2 Speech quality, intelligibility and dereverberation measures for Scenario #1. The microphone signal refers to Mic.1#. NLMS refers to applying NLMS alone with any dereverberation algorithm

Table 2 Results of speech quality, intelligibility and dereverberation measures for Scenario #2. The input signal refers to one of the microphones. NLMS refers to applying NLMS alone with any dereverberation algorithm. The values of LSD, SRMR and SER are in dB and the values of the STOI measure are the percentage of correct words

	LSD	SRMR	STOI	ΔSER
Microphone signal	5.52	3.35	82.15	0
Proposed	**3.54**	**7.40**	**95.57**	**23.16**
NLMS	4.89	3.38	83.64	1.84
NLMS-MCSE	3.74	6.32	90.07	6.03
NLMS-RKEMD-1	3.89	5.28	71.81	11.04
NLMS-RKEMD-2	3.79	5.44	74.49	10.96

quality and intelligibility. It also explains the degradation in NLMS-RKEMD-2. However, our method converges faster even in the presence of these estimation errors and performs better.

We also note that, as expected, the NLMS-RKEMD-2 algorithm outperforms the NLMS-RKEMD-1 algorithm. However, its performance is still inferior to that of the NLMS-MCSE algorithm. In terms of intelligibility, NLMS-RKEMD-1 and NLMS-RKEMD-2 even achieve inferior STOI measures than the microphone signal. However, the speech quality in terms of dereverberation and signal distortion still improved, as evident from a the higher SRMR and lower LSD measures.

5.5 Echo cancellation performance

A common performance measure for evaluating echo cancellation is the ERLE defined for each time-frame as

$$\mathrm{ERLE}(t) = 10\log_{10}\frac{\sum_{k=0}^{K-1}\left(\mathbf{g}_1^\top(k)\mathbf{y}_t(k)\right)^2}{\sum_{k=0}^{K-1}\left(\left(\mathbf{g}_1^\top(k)-\hat{\mathbf{g}}_1^\top(k)\right)\mathbf{y}_t(k)\right)^2}. \quad (46)$$

The ERLE results per frame for Scenario #1 are presented in Fig. 3, depicting the advantage of the proposed method over the competing methods for most frames. Furthermore, we can observe that the ERLE performance is rather stable and insensitive to changes in the far-end signal and to the DTD accuracy.

Note that $\mathbf{g}_1^\top(k)\mathbf{y}_t(k)$ is only available in Scenario #1. In Scenario #2, we cannot separately record the near-end signal and the echo signal and then mix them to generate a test scenario, due to the manual movement of the microphone array, which cannot be exactly repeated. Therefore, for Scenario #2, we propose to use the ratio of the power of the signal when the speech and reference signals are present and the signal power when only the reference signal is active. We refer to this ratio as as signal to echo ratio (SER) and we define it for the input and the output signals:

$$\mathrm{SER}_{\mathrm{input}} = 10\log_{10}\left(\frac{\sum_{n\in\mathcal{N}_a}|z_1[n]|^2}{\sum_{n\in\mathcal{N}_b}|z_1[n]|^2}-1\right) \quad (47a)$$

$$\mathrm{SER}_{\mathrm{output}} = 10\log_{10}\left(\frac{\sum_{n\in\mathcal{N}_a}|\hat{x}[n]|^2}{\sum_{n\in\mathcal{N}_b}|\hat{x}[n]|^2}-1\right). \quad (47b)$$

where

$$\mathcal{N}_a = \{n \in x[n] \text{ is active } \& \; y[n] \text{ is active}\}, \quad (48a)$$

$$\mathcal{N}_b = \{n \in x[n] \text{ is not active } \& \; y[n] \text{ is active }\}. \quad (48b)$$

The improvement between the $\mathrm{SER}_{\mathrm{input}}$ and $\mathrm{SER}_{\mathrm{output}}$ indicates the attenuation in the echo power and is denoted by ΔSER. The length of both $\mathcal{N}_a$ and $\mathcal{N}_a$ is approximately 6 seconds. The median of the measured ΔSER for Scenario #2 is presented in Table 2, also depicting advantage of the proposed method over the competing methods. Recall that the echo path adaptation in NLMS-MCSE and NLMS-RKEMD-1 continues in this scenario even during double talk while the statistical model that is used in these methods is not considering the near-end signal. NLMS-RKEMD-2 echo cancellation performance is worse than our method due to the constantly time-varying echo path and the convergence of the reverberated speech component. Hence, the level of the residual echo is significant and it is reflected in the ΔSER.

5.6 Spectograms assessment

In addition to the quality measures presented in Sections 5.4 and 5.5, we provide the spectograms of one example for Scenario #1 in Fig. 4 and for Scenario #2 in Fig. 5. The spectograms of both scenarios demonstrate the enhancement capabilities and the robustness of the proposed method to double-talk scenarios. Sound examples of both scenarios can be found in the lab website[2].

6 Conclusions

A recursive EM algorithm, based on Kalman filtering, for AEC, dereverberation and noise reduction was presented. The proposed statistical model is addressing the three problems simultaneously. The E-step and M-step are implemented for each STFT time-frame. The E-step is implemented as a Kalman filter. The model parameters are estimated in the M-step. Given the estimate of the acoustic path of the far-end signal, the echo signal at each

[2] www.eng.biu.ac.il/gannot/speech-enhancement/

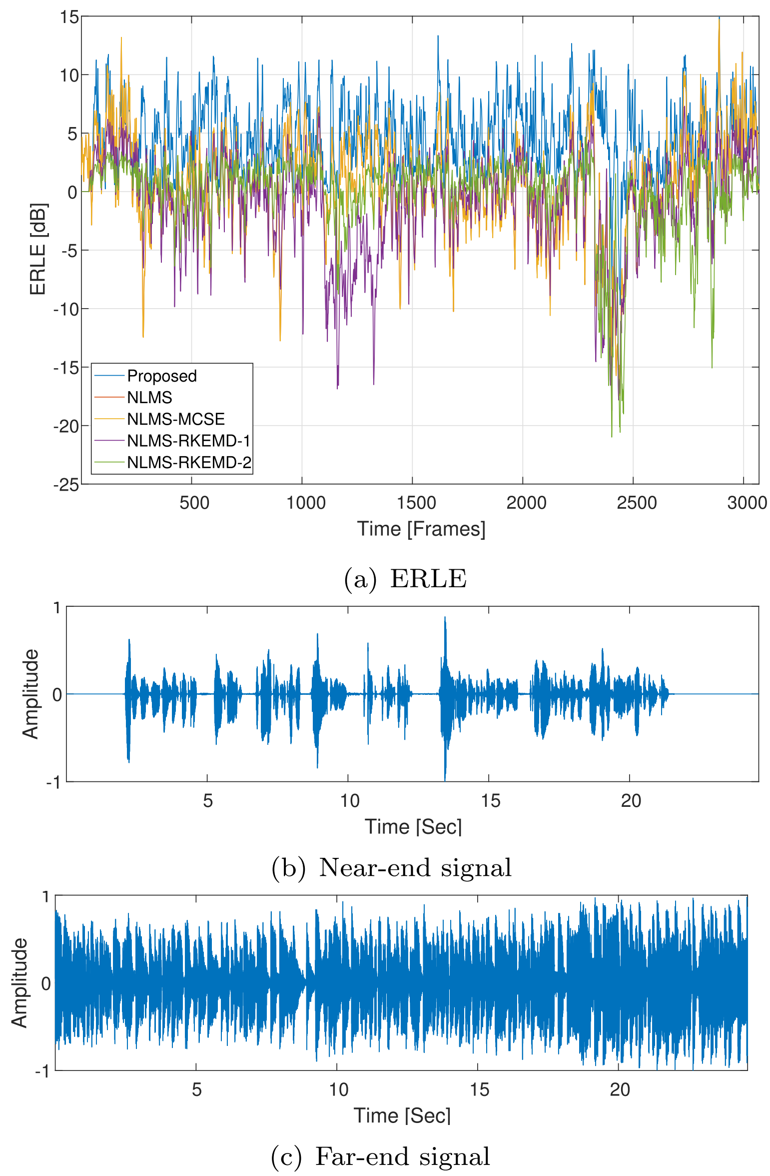

Fig. 3 ERLE per frame for Scenario #1

channel is evaluated. The estimated echo signal is subtracted from the microphone signal and the outcome is further processed by the Kalman filter. The desired speech variance was estimated by adopting a spectral estimation method. The estimated near-end signal is obtained as a byproduct of the E-step. A DTD was utilized in order to suspend the M-step adaptation when the near-end and far-end signal are not active and, consequently, to prevent adaptation errors.

The tracking ability of the algorithm was tested in an experimental study carried out in our lab in very challenging scenarios, including moving speakers and moving microphone array. The algorithm demonstrates convergence capabilities even during double-talk scenarios in time-varying scenarios. Our method is shown to outperform competing methods based on the NLMS algorithm, in terms of intelligibility, speech quality, and echo cancellation performance.

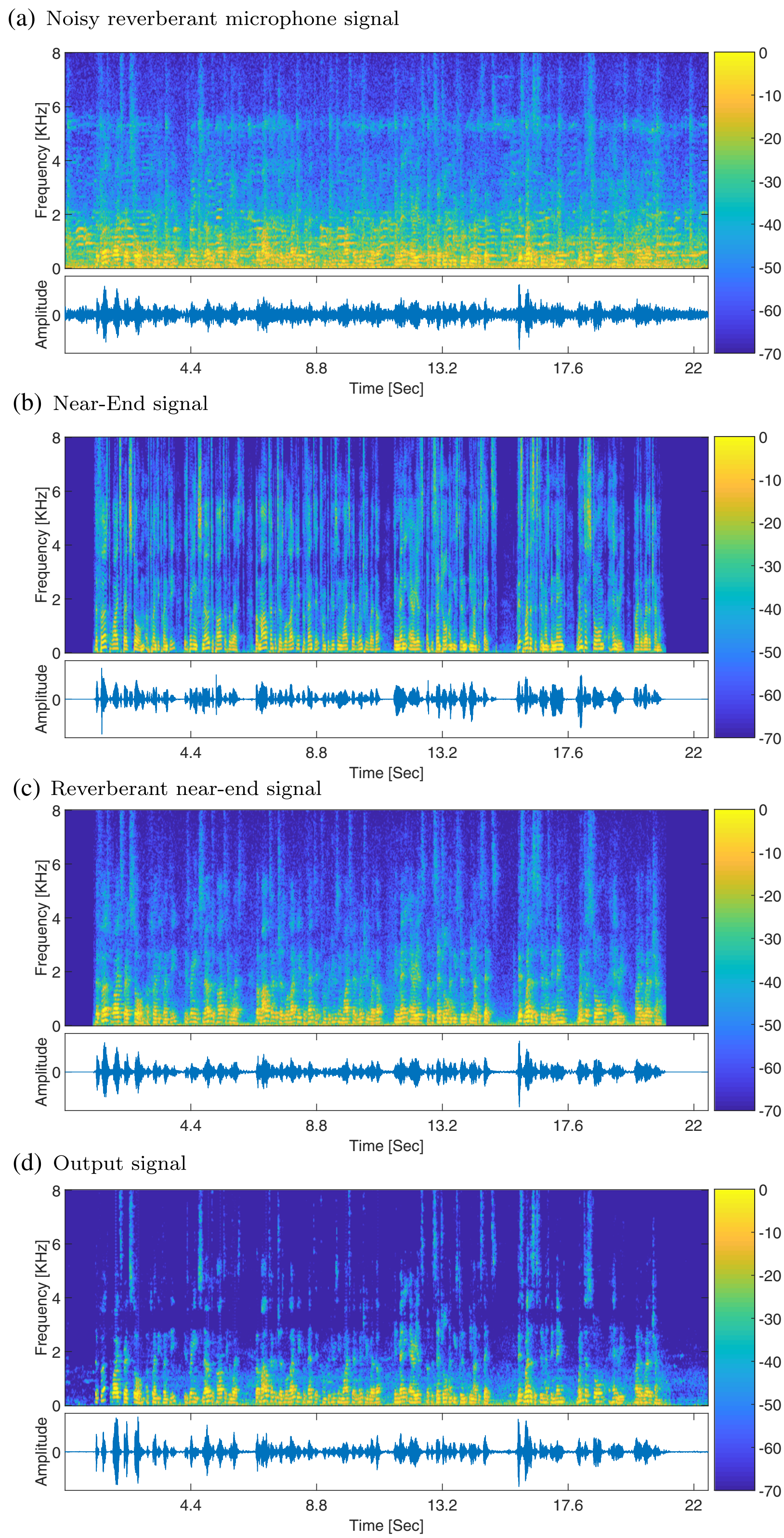

Fig. 4 Spectogrames and waveforms of Scenario #1, $T_{60} = 650$ ms

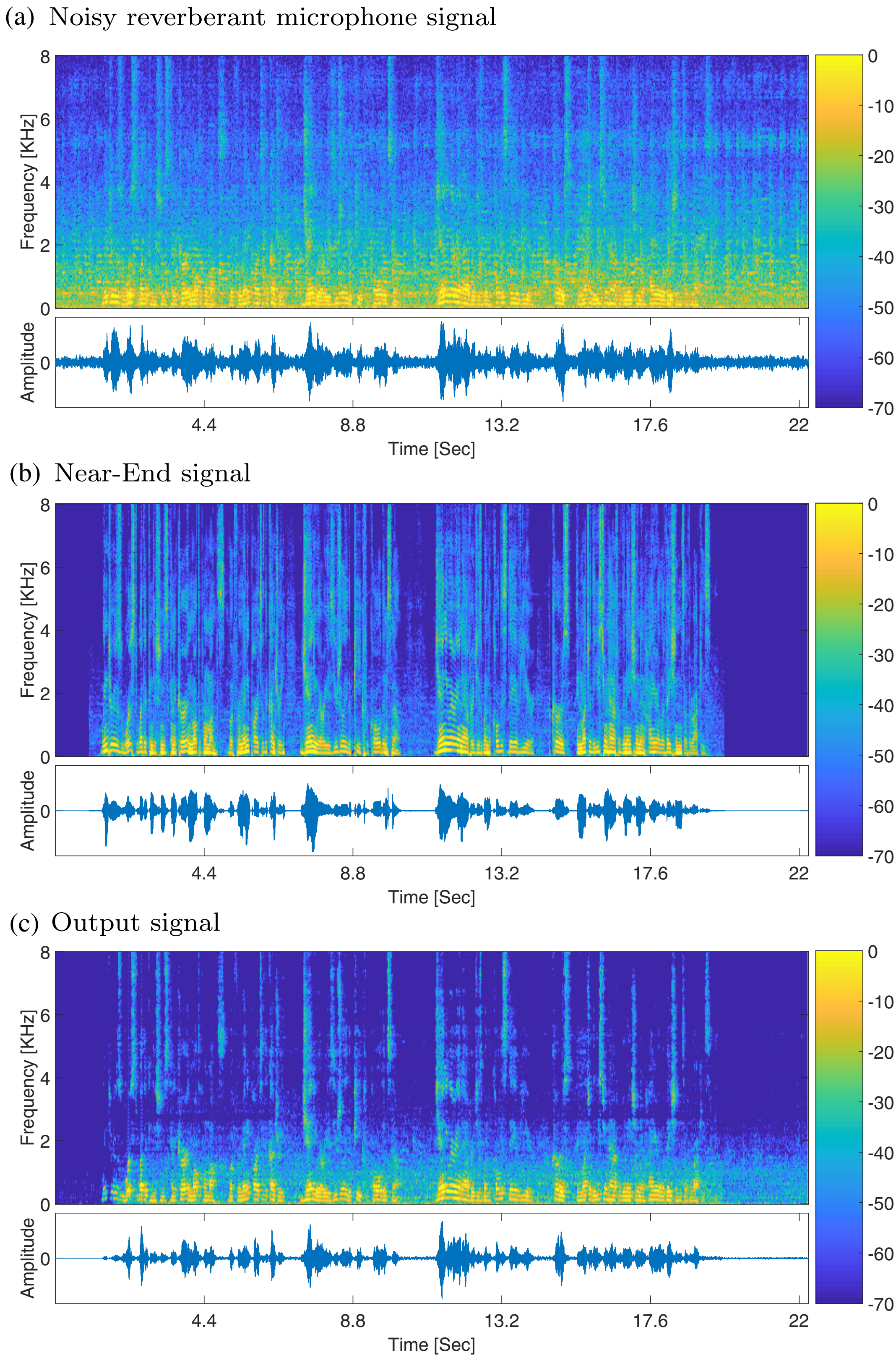

Fig. 5 Spectogrames and waveforms of scenario #2, $T_{60} = 650$ ms. The reverberant near-end signal cannot be extracted in Scenario #2

Abbreviations

ADM: adaptive directional microphone; AIR: acoustic impulse response; AR: auto-regressive; ASR: automatic speech recognition; ATF: acoustic transfer function; BIU: Bar-Ilan University; BSI: blind system identification; BSS: blind source separation; CASA: computational auditory scene analysis; CTF: convolutive transfer function; DLP: delayed linear prediction; DOA: direction of arrival; DRR: direct to reverberant ratio; DSB: delay and sum beamformer; ECM: expectation-conditional maximization; EM: expectation-maximization; EMK: EM-Kalman; E-Step: estimate step; FAU: Friedrich-Alexander University of Erlangen-Nuremberg; FIR: finite impulse response; FIM: Fisher information matrix; GCI: glottal closure instants; GEM: generalized EM; GSC: generalized side-lobe canceller; HOS: high-order statistics; HRTF: head related transfer function; IC: interaural coherence; ICA: independent component analysis; ILD: interaural level difference; ITD: interaural time difference; ITF: interaural transfer function; IS: Itakura-Saito; i.i.d: independent and identically distributed; KEM: Kalman expectation-maximization; KEMD: Kalman expectation-maximization for dereverberation; KEMDS: Kalman expectation-maximization for dereverberation and separation; RKEM: recursive Kalman expectation-maximization; RKEMD: recursive Kalman expectation-maximization for dereverberation; LPC: linear prediction coding; LSD: log-spectral distortion; LSP: line spectral pair; LS: least-squares; RLS: recursive least-squares; NLMS: normalized least-mean-square; LTI: linear time invariant; MA: moving average; MAP: maximum a-posteriori; MCH: multi-channel; MFCC: mel-frequency cepstrum coefficients; MINT: multiple input-output inverse theorem; MKEMD: Multi-Speaker Kalman-EM for dereverberation; ML: maximum likelihood; MLE: maximum likelihood estimation; MMSE: minimum mean square error; MSE: mean square error; MVDR: minimum variance distortionless response; M-Step: maximize step; MWF: multichannel Wiener filter; MWF-N: MWF with partial noise estimate; MTF: multiplicative transfer function; NMF: nonnegative matrix factorization; NPM: normalized projection misalignment; NSRR: normalized signal to reverberant ratio; OM-LSA: optimally-modified log spectral amplitude; PESQ: perceptual evaluation of speech quality; PSD: power spectral density; p.d.f: probability distribution function; WGN: white Gaussian noise; REM: recursive EM; RIR: room impulse response; RSNR: reverberated-signal to noise ratio; RTF: relative transfer function; SDW-MWF: speech distortion weighted multichannel Wiener filter; SE: spectral enhancement; SIR: signal to interference ratio; SNR: signal to noise ratio; SPP: speech presence probability; SRMR: speech to reverberation modulation energy ratio; SRR: signal to reverberant ratio; STFT: short-time Fourier transform; TF: time-frequency; UOL: University of Oldenburg; DUET: degenerate unmixing estimation technique; SDR: signal to distortion ratio; SIR: signal to interference ratio; SAR: signal to artefacts ratio; SRR: signal to reverberant ratio; AEC: acoustic echo cancellation; DTD: doubletalk detector; ERLE: echo return loss enhancement; SER: signal to echo ratio; AC: air-conditioner; STOI: short-time objective intelligibility; MCSE: multichannel spectral enhancement; BLMS: block least-mean-square

Acknowledgements

We would like to thank Mr. Pini Tandeitnik for his professional assistance during the acoustic room setup and the recordings.

Authors' contributions

Model development: NC, BS, GH and SG. Experimental testing: NC, GH and BS. Writing paper: NC, BS, and SG. The authors read and approved the final manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

Availability of data and materials

N/A

Declarations

Consent for publication

All authors agree to the publication in this journal.

Competing interests

The authors declare that they have no competing interests.

Received: 5 April 2021 Accepted: 27 July 2021
Published online: 28 August 2021

References

1. G. Schmidt, in *12th European Signal Processing Conference (EUSIPCO)*. Applications of acoustic echo control - an overview, (2004), pp. 9–16
2. E. Hänsler, G. Schmidt, *Acoustic Echo and Noise Control: a Practical Approach*, vol. 40. (John Wiley & Sons, New-Jersey, 2005)
3. E. Hänsler, G. Schmidt, *Topics in Acoustic Echo and Noise Control: Selected Methods for the Cancellation of Acoustical Echoes, the Reduction of Background Noise, and Speech Processing*. (Springer, Berlin, 2006)
4. A. Gilloire, M. Vetterli, Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation. IEEE Trans. on Signal Process. **40**(8), 1862–1875 (1992)
5. J. Benesty, F. Amand, A. Gilloire, Y. Grenier, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Adaptive filtering algorithms for stereophonic acoustic echo cancellation, vol. 5, (1995), pp. 3099–3102
6. S. L. Gay, in *The 32nd Asilomar Conference on Signals, Systems and Computers*. An efficient, fast converging adaptive filter for network echo cancellation, vol. 1, (1998), pp. 394–398
7. H. Deng, M. Doroslovacki, Proportionate adaptive algorithms for network echo cancellation. IEEE Trans. Signal Process. **54**(5), 1794–1803 (2006)
8. D. L. Duttweiler, Proportionate normalized least-mean-squares adaptation in echo cancelers. IEEE Trans. Speech Audio Process. **8**(5), 508–518 (2000)
9. G. Enzner, P. Vary, Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones. Signal Process. **86**(6), 1140–1156 (2006)
10. W. Kellermann, in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays, vol. 1, (1997), pp. 219–2221
11. W. Kellermann, in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*. Joint design of acoustic echo cancellation and adaptive beamforming for microphone arrays, (1997), pp. 81–84
12. G. Reuven, S. Gannot, I. Cohen, Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller. Speech Commun. **49**(7), 623–635 (2007)
13. J. J. Shynk, Frequency-domain and multirate adaptive filtering. IEEE Signal Process. Mag. **9**(1), 14–37 (1992)
14. A. Cohen, A. Barnov, S. Markovich-Golan, P. Kroon, in *2018 26th European Signal Processing Conference (EUSIPCO)*. Joint beamforming and echo cancellation combining QRD based multichannel AEC and MVDR for reducing noise and non-linear echo, (2018), pp. 6–10
15. M. Luis Valero, E. A. P. Habets, Low-complexity multi-microphone acoustic echo control in the short-time fourier transform domain. IEEE/ACM Trans. Audio Speech Lang Process. **27**(3), 595–609 (2019)
16. Y. Avargel, System identification in the short-time fourier transform domain. PhD thesis, Technion - Israel Institute of Technology (2008). https://israelcohen.com/wp-content/uploads/2018/05/YekutiaelAvargel_PhD_2008.pdf
17. E. A. P. Habets, S. Gannot, I. Cohen, P. Sommen, Joint dereverberation and residual echo suppression of speech signals in noisy environments. IEEE Trans. Audio Speech Lang. Process. **16**(8), 1433–1451 (2008)
18. R. Martin, P. Vary, Combined acoustic echo cancellation, dereverberation and noise reduction: a two microphone approach. Ann. Telecommun. **49**, 429–438 (1994)
19. M. Togami, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Variational Bayes state space model for acoustic echo reduction and dereverberation, (2015), pp. 101–105
20. R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, H. G. Okuno, in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition, (2009), pp. 3677–3680
21. K. Lebart, J. Boucher, P. Denbigh, A new method based on spectral subtraction for speech dereverberation. Acta Acustica Acustica. **87**, 359–366 (2001)
22. E. A. P. Habets, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multi-channel speech dereverberation based on a statistical model of late reverberation, vol. 4, (2005), pp. 173–176
23. E. A. P. Habets, S. Gannot, I. Cohen, Late reverberant spectral variance estimation based on a statistical model. IEEE Signal Process. Lett. **16**(9), 770–773 (2009)

24. T. Yoshioka, T. Nakatani, M. Miyoshi, Integrated speech enhancement method using noise suppression and dereverberation. IEEE Trans. Audio Speech Lang. Process. **17**(2), 231–246 (2009)
25. S. Gannot, D. Burshtein, E. Weinstein, Iterative and sequential Kalman filter-based speech enhancement algorithms. IEEE Trans. Speech Audio Process. **6**(4), 373–385 (1998)
26. B. Schwartz, S. Gannot, E. A. P. Habets, in *European Signal Processing Conference (EUSIPCO)*. Multi-microphone speech dereverberation using expectation-maximization and Kalman smoothing, (Marakech, Morocco, 2013)
27. D. Titterington, Recursive parameter estimation using incomplete data. J. R. Stat. Soc. Ser. B Methodol. **46**(2), 257–267 (1984)
28. P.-J. Chung, J. F. BÕhme, Recursive EM and SAGE-inspired algorithms with application to DOA estimation. IEEE Trans. Signal Process. **53**(8), 2664–2677 (2005)
29. E. Weinstein, A. Oppenheim, M. Feder, J. Buck, Iterative and sequential algorithms for multisensor signal enhancement. IEEE Trans. Signal Process. **42**, 846–859 (1994)
30. L. Frenkel, M. Feder, Recursive expectation-maximization (EM) algorithms for time-varying parameters with applications to multiple target tracking. IEEE Trans. Signal Process. **47**(2), 306–320 (1999)
31. O. Cappé, E. Moulines, On-line expectation-maximization algorithm for latent data models. J. R. Stat. Soc. Ser. B Stat Methodol. **71**(3), 593–613 (2009)
32. B. Schwartz, S. Gannot, E. A. P. Habets, Y. Noam, Recursive maximum likelihood algorithm for dependent observations. IEEE Trans. Signal Process. **67**(5), 1366–1381 (2019)
33. D. Schmid, S. Malik, G. Enzner, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain, (2012), pp. 17–20
34. B. Schwartz, S. Gannot, E. A. P. Habets, Online speech dereverberation using kalman filter and em algorithm. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(2), 394–406 (2015)
35. I. Cohen, Speech enhancement using super-gaussian speech models and noncausal a priori SNR estimation. Speech Commun. **47**, 336–350 (2005)
36. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. **39**(1), 1–22 (1977)
37. E. Weinstein, A. V. Oppenheim, M. Feder, J. R. Buck, Iterative and sequential algorithms for multisensor signal enhancement. IEEE Trans. Signal Process. **42**(4), 846–859 (1994)
38. A. Benveniste, M. Métivier, P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, vol. 22. (Springer, 2012)
39. P. J. Wolfe, S. J. Godsill, in *The 11th IEEE Signal Processing Workshop on Statistical Signal Processing (SSP)*. Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement, (2001), pp. 496–499
40. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoustics Speech Signal Process. **32**(6), 1109–1121 (1984)
41. E. Weinstein, A. V. Oppenheim, M. Feder, *Signal Enhancement Using Single and Multi-sensor Measurements*. (Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 0213 USA, 1990)
42. J. Benesty, D. R. Morgan, J. H. Cho, A new class of doubletalk detectors based on cross-correlation. IEEE Trans. Speech Audio Process. **8**(2), 168–172 (2000)
43. X. Li, R. Horaud, L. Girin, S. Gannot, in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. Voice activity detection based on statistical likelihood ratio with adaptive thresholding, (2016)
44. E. A. P. Habets, in *Speech Dereverberation*. Speech dereverberation using statistical reverberation models (Springer, London, 2010), pp. 57–93
45. C. Sørensen, J. B. Boldt, F. Gran, M. G. Christensen, in *24th European Signal Processing Conference (EUSIPCO)*. Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids, (2016), pp. 1358–1362
46. T. H. Falk, C. Zheng, W. Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio Speech Lang. Process. **18**(7), 1766–1774 (2010)

Publisher's Note