Joint Beamforming and Echo Cancellation Combining QRD Based Multichannel AEC and MVDR for Reducing Noise and Non-Linear Echo

Alejandro Cohen, Anna Barnov, Shmulik Markovich-Golan and Peter Kroon

Communication and Devices Group, Intel Corporation
Email: {alejandro.cohen,shmulik.markovich-golan,peter.kroon}@intel.com

Abstract—**The problems of echo and noise contaminating a desired talker signal in a communication or an entertainment device are considered. In the following, we propose a combined method comprising a linear echo-canceller followed by a weighted minimum variance distortionless response (MVDR) beamformer designed to reduce noise and echo residues. For the echo-canceller stage we use a fast-converging multichannel QR decomposition (QRD)-recursive least squares (RLS) method. For the beamformer stage, we adopt and modify our recently proposed method of a fast-tracking QRD based MVDR beamformer [1]. We model that the residual echo is dominated by non-linearly distorted components which undergo the same echo paths as the non-distorted component. Thereby, the MVDR beamformer is designed to minimize a weighted sum of the powers of the noise and of the non-linear echo while maintaining the desired talker undistorted. The computational and memory complexities of the proposed algorithm are sufficiently low, making it appropriate for implementation in mobile devices. The performance of the proposed method is tested using real recordings from two commercial devices, a mobile-phone and a smart-speaker.**

I. Introduction

Noise and echo are fundamental problems in speech enhancement applications, e.g., voice communication, speech recognition etc. Treating them is especially important for distant talker scenarios, where the desired talker component is weak and the corresponding signal-to-noise ratio (SNR) and signal-to-echo ratio (SER) are low. Modern devices usually include a microphone array which enables introducing spatial filtering, also referred to as beamforming, for enhancing the desired talker component. Mobile devices with small form factor usually drive the loudspeaker close to its compression point for increased efficiency at the expense of introducing non-linear distortions to the emitted signal. In this case, the performance attained by a linear echo-canceller is limited and more complicated non-linear echo cancellation methods, e.g. [2]–[4], are needed.

Numerous methods address the problem of a joint echo cancellation and noise reduction solution [5]. Two common approaches for combining echo cancellation and beamforming exist: either beamformer followed by echo-canceller or the other way around. In the first approach, beamformer followed by echo-canceller, a lower complexity is attained, since it requires a single channel echo-canceller. Furthermore, this approach allows for improved echo cancellation since the noise at the echo-canceller input is reduced by the beamformer. However, for consistent performance, the echo path estimator should track variations in both the acoustic echo path and in the beamformer design. In addition designing the beamformer in this case is intricate since its inputs contain also echo. In the second approach the echo-canceller design is decoupled from the subsequent beamformer. However, this decoupling comes at the expense of additional computations which are required for applying a multichannel echo-canceller. More advanced approaches try to alleviate some of the problem listed above, see [5]–[7]. Other approaches treat the echo as a spatial interference that can be reduced by properly designing the beamformer, see [8]–[10]. However, the latter approaches involve increased computations since the beamformer and echo-canceller are jointly optimized using a single multi-dimensional cost function. Due to this coupling of the echo-canceller design and the beamformer design, changes in the echo path, noise field or desired talker position will shift the optimal point and require update. Rapid changes will lead in the best case to increased computations and in worse cases to performance degradation.

In the current contribution we address the problem of joint echo cancellation and beamforming. The problem is formulated in Sec. II. We propose to combine reference based echo cancellation with beamforming based echo reduction and adopt the echo-canceller followed by beamformer approach, see Sec. IV. For the echo-canceller stage we propose to use the fast-converging multichannel QR decomposition (QRD)-recursive least squares (RLS) method. For the beamformer stage, we adopt and modify our recently proposed method of a fast-tracking QRD based minimum variance distortionless response (MVDR) beamformer [1], briefly presented in Sec. III. We assume that the first-stage echo-canceller is able to track the multichannel echo paths and cancel the *linear echo component*. We further assume that the residual echo at the output of this stage is dominated by the *non-linear echo components* which are modeled to undergo the same acoustic echo paths as the linear echo component. This assumption allows us to derive a simpler solution compared to the previously mentioned methods which jointly address echo cancellation and beamforming and aim to reduce the echo also as a spatial interference. The MVDR at the second stage is designed to minimize a weighted sum of the powers of the noise and of the non-linear echo while maintaining the desired talker undistorted. This is accomplished by splitting the beamformer into a *whitening* step, which spatially whitens the noise, followed by a multichannel filter which passes the desired talker undistorted and reduces the residual echo. Note that utilizing spatial degrees of freedom for reducing the residual echo comes at the expense of noise reduction. We introduce a design parameter to help control the latter trade-off by applying a weight to the power of the non-linear echo component. Unlike in [1], the relative transfer function (RTF) of the desired talker is estimated in the whitened domain, and it does not require transforming it back to the *microphone signals domain*. This saves computations compared to [1]. Specifically, the proposed method requires the application of the inverse QRD (IQRD) method, instead of applying both IQRD and QRD methods. Thus, the computational and memory complexities of the proposed method are sufficiently low, making it practical for implementation in mobile devices.

The performance of the proposed method is tested in Sec. V by using real recordings from two devices, a mobile-phone and a smart-speaker. The loudspeaker of the mobile-phone generates high non-linear components whereas the smart-speaker loudspeaker is more linear. By analyzing the signals at the output of these two different

ISBN 978-90-827970-1-5 © EURASIP 2018

cases, we show the robustness of the proposed algorithm and the contribution of its two stages to the overall performance depending on the level of the non-linear echo. The proposed algorithm is able to reduce noise and echo while maintaining the quality of the desired talker in both cases.

II. Problem Formulation

Let us consider the problem of enhancing the speech signal of a desired talker, denoted $\underline{s}(t)$, that is picked up by a system comprising of M microphones and a loudspeaker. The received microphone signals are contaminated by echo signals, originating from the reference signal, denoted $\underline{r}(t)$, which drives the system loudspeaker, and noise components, denoted $\underline{v}_m(t)$, for $m = 1, \dots, M$. In practice, due to non-ideal amplifier and loudspeaker, the signal emitted from the loudspeaker contains non-linear distortions of $\underline{r}(t)$. Without loss of generality, we assume that the non-linear distorted component, denoted as $\underline{\tilde{r}}(t)$, is statistically independent of $\underline{r}(t)$. Denote the summation of the reference signal and its non-linear distortion, i.e. $\underline{r}(t) + \underline{\tilde{r}}(t)$, as the *emitted reference*. The signals are given in the discrete-time domain with t denoting the time-index and the sample-rate given by f_s. Following these definitions, the received signal at the m-th microphone, denoted $\underline{x}_m(t)$, is

$$\underline{x}_m(t) = \underline{c}_m(t) + \underline{e}_m(t) + \underline{v}_m(t) \tag{1}$$

where

$$\begin{aligned}\underline{c}_m(t) &\triangleq \underline{h}_{s,m}(t) * \underline{s}(t)\\ \underline{e}_m(t) &\triangleq \underline{h}_{e,m}(t) * (\underline{r}(t) + \underline{\tilde{r}}(t))\end{aligned}$$

are the received desired talker and echo components, respectively, with $\underline{h}_{s,m}(t)$ and $\underline{h}_{e,m}(t)$ being the acoustic impulse responses (AIRs) relating the desired talker source and the emitted reference, respectively, with the m-th microphone, where $*$ denotes the convolution operator. Note that we model that the non-linear distortion components of the reference signal undergo the same AIRs as the reference signal before reaching the microphones.

Transforming the microphone signals to the short time Fourier transform (STFT) domain, using an equal length of F for analysis and synthesis windows with an overlap of D between frames, and arranging them in vectors yields:

$$\mathbf{x}(n,f) \triangleq \mathbf{c}(n,f) + \mathbf{e}(n,f) + \mathbf{v}(n,f) \tag{2}$$

where

$$\begin{aligned}\mathbf{c}(n,f) &\triangleq [c_1(n,f), \dots, c_M(n,f)]^T = \mathbf{h}_s(n,f)s(n,f)\\ \mathbf{e}(n,f) &\triangleq [e_1(n,f), \dots, e_M(n,f)]^T = \mathbf{h}_e(n,f)(r(n,f) + \tilde{r}(n,f)),\end{aligned}$$

are the speech and the echo components vectors, respectively, with

$$\begin{aligned}\mathbf{h}_s(n,f) &\triangleq [h_{s,1}(n,f), \dots, h_{s,M}(n,f)]^T\\ \mathbf{h}_e(n,f) &\triangleq [h_{e,1}(n,f), \dots, h_{e,M}(n,f)]^T\end{aligned}$$

defined to be the desired talker and echo acoustic transfer functions (ATFs) vectors, respectively, and n and f denote the time-frame and frequency-bin indices. Note that we assume that the AIRs are significantly shorter than the window length F, such that a linear convolution in the time-domain can be approximated as a multiplication in the STFT domain. Throughout the paper we distinguish between terms in the STFT domain and in the discrete-time domain by underlining the latter. Hereon, we omit the frequency-bin index f for simplicity and all derivations should be interpreted as for a particular frequency.

Let us analyze the second-order statistics of the received signals. Considering (2), the spatial covariance matrix of $\mathbf{x}(n)$ is:

$$\mathbf{\Phi}_x(n) = \mathbf{\Phi}_c(n) + \mathbf{\Phi}_e(n) + \mathbf{\Phi}_v(n) \tag{3}$$

where

$$\begin{aligned}\mathbf{\Phi}_c(n) &= \phi_s(n)\mathbf{h}_s(n)\mathbf{h}_s^H(n)\\ \mathbf{\Phi}_e(n) &= (\phi_r(n) + \phi_{\tilde{r}}(n))\mathbf{h}_e(n)\mathbf{h}_e^H(n)\end{aligned} \tag{4}$$

are the spatial covariance matrices of the received desired talker and echo components, respectively, $\mathbf{\Phi}_v(n)$ is the spatial covariance matrix of the noise components and we assume that the desired talker source, the emitted reference and the noise components are statistically independent. The time-varying spectra of the desired talker source, the reference signal and its non-linear artifact are denoted as $\phi_s(n)$, $\phi_r(n)$ and $\phi_{\tilde{r}}(n)$, respectively.

Herein, we aim to enhance the desired talker component from the observed signals $\mathbf{x}(n)$ which are contaminated by echo and noise.

III. QRD-based MVDR beamformer [1]

The MVDR [11], [12] is a common criterion for beamformers which is designed to minimize the noise variance at the output while maintaining the desired source undistorted. Its closed-form solution is given by

$$\mathbf{w}_v(n) = \frac{\mathbf{\Phi}_v^{-1}(n)\tilde{\mathbf{h}}_s(n)}{\tilde{\mathbf{h}}_s^H(n)\mathbf{\Phi}_v^{-1}(n)\tilde{\mathbf{h}}_s(n)} \tag{5}$$

where

$$\tilde{\mathbf{h}}_s(n) \triangleq \mathbf{h}_s(n)/h_{s,1}(n) \tag{6}$$

is the RTFs vector, comprising of the desired source ATFs to all microphones normalized by the desired source ATF to the reference microphone, selected here as the first microphone.

In a recent contribution in [1], an efficient QRD based implementation of the MVDR beamformer, capable of fast-tracking of speech and noise sound fields in dynamic scenarios, has been proposed. Define the decomposition of the inverted noise covariance matrix $\mathbf{\Phi}_v^{-1}(n)$ to a product of its square-root matrices as

$$\mathbf{\Phi}_v^{-1}(n) \triangleq \mathbf{S}^{-1}(n)\mathbf{S}^{-H}(n) \tag{7}$$

where $\mathbf{S}^{-1}(n)$ is the *square-root matrix* of $\mathbf{\Phi}_v^{-1}(n)$. Note that the latter decomposition is not unique. Next, by defining

$$\mathbf{b}_s(n) \triangleq \mathbf{S}^{-H}(n)\tilde{\mathbf{h}}_s(n) \tag{8}$$

the MVDR in (5) can be reformulated as:

$$\mathbf{w}_v(n) = \frac{\mathbf{S}^{-1}(n)\mathbf{b}_s(n)}{||\mathbf{b}_s(n)||^2}.$$

The efficiency of the algorithm results from reformulating the compute-intensive operations to reuse the matrix $\hat{\mathbf{S}}^{-H}(n)$, which is a recursive estimate of $\mathbf{S}^{-H}(n)$, computed from the microphone signals using the IQRD method. Explicitly these operations are: 1) tracking the inverse square-root of the noise covariance matrix; 2) estimating the desired source RTFs vector, $\hat{\mathbf{b}}_s(n)$; and 3) estimating the voice activity detection (VAD) for controlling the adaption. Note that another QRD procedure which tracks $\hat{\mathbf{S}}^H(n)$, an estimate of $\mathbf{S}^H(n)$, is required for tracking the desired source RTFs vector.

IV. Proposed Method

We propose the following two-stage method for addressing the echo cancellation and noise reduction problems in a joint approach. In the first stage, see Sec. IV-A, a linear echo-canceller based on an efficient implementation of the multichannel IQRD-RLS is applied. The multichannel output of the first stage as well as the estimated echo paths are fed-forward to the second stage, which comprises a weighted MVDR beamformer that is described in Sec. IV-B. The latter is based on the efficient QRD-based MVDR beamformer [1] and its design is modified to minimize a weighted sum of the powers of the noise and residual-echo components at the output while maintaining the desired source undistorted. We assume that the residual echo component is mainly dominated by the non-linearly distorted reference component $\tilde{r}(n)$, which undergoes the same multichannel ATF as the reference signal $r(n)$ before reaching the microphones. Therefore, we conjecture that by adding a penalty term with the spatial properties of the echo signal $\mathbf{e}(n)$ to the noise covariance matrix we will be able to utilize spatial degrees of freedom to further suppress the residual echo. Note that other echo cancellation methods can be used with this principle. Also note that the power of the residual echo at the input of the second stage can be reduced by incorporating non-linear echo cancellation techniques [2]–[4], however we do not consider them in the current contribution and leave that for future research. We replace the speech tracking method in [1] which requires both the inverse square-root and square-root of the matrix $\mathbf{\Phi}_v$ by a more efficient method, described in Sec. IV-C, that requires only its inverse square-root. A high-level block-diagram of the proposed method is depicted in Fig. 1.

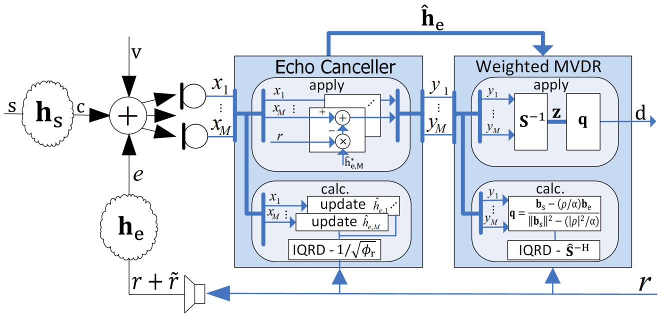

Figure 1: High-level block diagram of the proposed method.

A. IQRD-RLS based multichannel echo-canceller

The optimal estimate for the echo path to the m-th microphone, i.e. $h_{e,m}(n)$, is the Wiener filter [13]:

$$\hat{h}_{e,m}(n) = \phi^*_{rx_m}(n)/\phi_r(n)$$

where $\phi_{rx_m}(n) \triangleq \mathrm{E}\left[r(n)x^*_m(n)\right]$ is the cross-correlation of the reference signal and the m-th microphone signal, respectively, and $\cdot^*$ denotes the conjugate operation. We adopt the IQRD-RLS method [14] for recursively estimating $\hat{h}_{e,m}(n)$. This flavor of the RLS method is more stable numerically and requires a reduced word length for representing numbers since it tracks $1/\sqrt{\phi_r(n)}$ unlike the standard RLS which tracks $1/\phi_r(n)$. Furthermore, as the standard RLS, it converges much faster than the least mean squares (LMS) method [15]. Even though the complexity of the LMS is lower than that of the RLS, note that the main additional computation is of $1/\sqrt{\phi_r(n)}$, which is independent on the number of microphones.

The output of the multichannel echo-canceller signal is denoted as:

$$\mathbf{y}(n) \triangleq \mathbf{x}(n) - \hat{\mathbf{h}}_e(n)r(n) = \mathbf{c}(n) + \dot{\mathbf{e}}(n) + \mathbf{v}(n)$$

where

$$\dot{\mathbf{e}}(n) \triangleq (\mathbf{h}_e(n) - \hat{\mathbf{h}}_e(n))r(n) + \mathbf{h}_e(n)\tilde{r}(n). \tag{9}$$

denotes the residual echo component and $\hat{\mathbf{h}}_e(n) \triangleq \left[\hat{h}_{e,1}(n), \ldots, \hat{h}_{e,M}(n)\right]^T$ is the estimated multichannel echo path.

B. A QRD-based Weighted MVDR beamformer

The weighted MVDR beamformer criterion is defined as:

$$\mathbf{w}_{v\dot{e}}(n) \triangleq \mathrm{argmin}_{\mathbf{w}} \mathbf{w}^H \mathbf{\Phi}(\mathbf{n})\mathbf{w} \text{ s.t. } \mathbf{w}^H \tilde{\mathbf{h}}_s(n) = 1 \tag{10}$$

where

$$\mathbf{\Phi}(n) \triangleq \mathbf{\Phi}_v(n) + \mu \mathbf{\Phi}_{\dot{e}}(n) \tag{11}$$

is the modified noise covariance matrix defined as a weighted sum of the noise covariance matrix and the residual echo covariance matrix with the weight parameter μ. Examining the residual echo component in (9) and assuming that the linear echo-canceller of the first stage has converged, we approximate that the residual echo is dominated by the non-linearly distorted reference component, i.e.

$$\dot{\mathbf{e}}(n) \approx \mathbf{h}_e(n)\tilde{r}(n).$$

In this case the residual echo covariance matrix is approximated as

$$\mathbf{\Phi}_{\dot{e}}(n) \approx \phi_{\tilde{r}}(n)\mathbf{h}_e(n)\mathbf{h}^H_e(n). \tag{12}$$

Recalling the closed-form formula of the MVDR beamformer in (5) and substituting (11) and (12) in the similar closed-form solution of (10) yields:

$$\mathbf{w}_{v\dot{e}}(n) = \frac{\left(\mathbf{\Phi}_v(n) + \mu\phi_{\tilde{r}}(n)\mathbf{h}_e(n)\mathbf{h}^H_e(n)\right)^{-1}\tilde{\mathbf{h}}_s(n)}{\tilde{\mathbf{h}}^H_s(n)\left(\mathbf{\Phi}_v(n) + \mu\phi_{\tilde{r}}(n)\mathbf{h}_e(n)\mathbf{h}^H_e(n)\right)^{-1}\tilde{\mathbf{h}}_s(n)}. \tag{13}$$

Next, we derive a simplified expression for (13) based on the square-root matrix of $\mathbf{\Phi}^{-1}_v(n)$. Let us consider the expression $\mathbf{\Phi}^{-1}(n) = \left(\mathbf{\Phi}_v(n) + \phi_{\tilde{r}}(n)\mathbf{h}_e(n)\mathbf{h}^H_e(n)\right)^{-1}$ that appears in (13). By applying the Woodbury identity, the latter expression can be formulated as:

$$\mathbf{\Phi}^{-1}(n) = \mathbf{\Phi}^{-1}_v(n) - \frac{\mathbf{\Phi}^{-1}_v(n)\mathbf{h}_e(n)\mathbf{h}^H_e(n)\mathbf{\Phi}^{-1}_v(n)}{1/(\mu\phi_{\tilde{r}}(n)) + \mathbf{h}^H_e(n)\mathbf{\Phi}^{-1}_v(n)\mathbf{h}_e(n)}. \tag{14}$$

By substituting (7), (8) and defining the whitened echo transfer functions (TFs) vector

$$\mathbf{b}_e(n) \triangleq \mathbf{S}^{-H}(n)\mathbf{h}_e(n),$$

the expression in (14) can be further simplified as:

$$\mathbf{\Phi}^{-1}(n) = \mathbf{S}^{-1}(n)\mathbf{S}^{-H}(n) - \frac{\mathbf{S}^{-1}(n)\mathbf{b}_e(n)\mathbf{b}^H_e(n)\mathbf{S}^{-H}(n)}{1/(\mu\phi_{\tilde{r}}(n)) + \|\mathbf{b}_e(n)\|^2}. \tag{15}$$

Finally, by substituting (15), (8) and defining $\rho(n) \triangleq \mathbf{b}^H_e(n)\mathbf{b}_s(n)$ and $\alpha(n) \triangleq 1/\left(\mu\phi_{\tilde{r}}(n)\right) + \|\mathbf{b}_e(n)\|^2$, the weighted MVDR beamformer in (13) can be reformulated as:

$$\mathbf{w}_{v\dot{e}}(n) = \mathbf{S}^{-1}(n)\mathbf{q}(n)$$

where

$$\mathbf{q}(n) \triangleq \frac{\mathbf{b}_s(n) - \left(\rho(n)/\alpha(n)\right)\mathbf{b}_e(n)}{\|\mathbf{b}_s(n)\|^2 - \left(|\rho(n)|^2/\alpha(n)\right)}.$$

The output of the proposed method is given by

$$d(n) \triangleq \mathbf{q}^H(n)\mathbf{S}^{-H}(n)\mathbf{y}(n). \tag{16}$$

Defining the *whitened* multichannel output of the echo-canceller as

$$\mathbf{z}(n) = \mathbf{S}^{-H}(n)\mathbf{y}(n)$$

and substituting it into (16) yields:

$$d(n) = \mathbf{q}^H(n)\mathbf{z}(n).$$

The output signal is transformed back to the time domain and denoted $\underline{d}(t)$. Note that the weighted MVDR is split into two stages, the *whitening* stage which yields the multichannel signal $\mathbf{z}(n)$ and the spatial filtering stage which yields the output signal $d(n)$.

In practice, constructing the weighted MVDR beamformer requires terms that are unknown and should be estimated. The square-root matrix $\mathbf{S}^{-1}(n)$ is estimated from the multichannel output of the echo-canceller stage, i.e. $\mathbf{y}(n)$, using the IQRD method as in [1]. The whitened desired source RTFs vector, i.e. $\mathbf{b}_s(n)$, is estimated by the procedure described in Sec. IV-C. Note that this procedure replaces the RTF estimation procedure in [1] and is less complicated since it does not require applying the QRD method. The whitened echo TFs vector is computed from the estimated echo TFs vector (see Sec. IV-A) as $\hat{\mathbf{b}}_e(n) \triangleq \hat{\mathbf{S}}^{-H}(n)\hat{\mathbf{h}}_e(n)$. The spectrum of the non-linearly distorted reference is modeled as a frequency dependent scaled version of the spectrum of the reference signal:

$$\hat{\phi}_{\tilde{r}}(n) = \hat{\phi}_r(n)\eta_r,$$

where $\hat{\phi}_r(n) = \lambda_r\hat{\phi}_r(n-1)+(1-\lambda_r)|r(n)|^2$ is a recursive average estimate of the spectrum of the reference signal $\phi_r(n)$, λ_r is a forgetting factor and η_r is a pre-calibrated time-invariant frequency scaling. Note that the estimation of $\phi_{\tilde{r}}(n)$ can be improved by modeling the non-linear distortion function, however this is beyond the scope of the current contribution. The estimation of the various terms is controlled by desired talker activity, estimated as in [1], and by echo signal activity.

C. Whitened desired source RTFs vector tracking

The whitened desired source RTFs vector is estimated in a similar way to the covariance whitening (CW) method [16]–[18].

Let us consider the whitened multichannel output of the echo-canceller $\mathbf{z}(n)$ in time periods where the desired talker is active and the reference signal is inactive. By substituting (3),(4),(6) and (8), its covariance matrix during these periods equals

$$\mathbf{\Phi}_z(n) = \mathbf{S}^{-H}(n)\mathbf{\Phi}_x(n)\mathbf{S}^{-1}(n) = |h_{s,1}(n)|^2\phi_s(n)\mathbf{b}_s(n)\mathbf{b}_s^H(n)+\mathbf{I} \tag{17}$$

where $\mathbf{I}$ denotes an identity matrix. Let $\mathbf{g}(n)$ denote the principal eigenvector of $\mathbf{\Phi}_z(n)$. From (17), it can be derived that the principal eigenvector is a normalized version of $\mathbf{b}_s(n)$, with a normalization factor of $\delta(n) = 1/||\mathbf{b}_s(n)||$ with an arbitrary phase, i.e.

$$\mathbf{g}(n) = \delta(n)\mathbf{b}_s(n). \tag{18}$$

The scalar $\delta(n)$ can be determined from (18), (8) and (6) as

$$\delta(n) = \mathbf{j}_1^T\mathbf{S}^H(n)\mathbf{g}(n) \tag{19}$$

where $\mathbf{j}_m \triangleq [\mathbf{0}_{1\times(m-1)}, 1, \mathbf{0}_{1\times(M-m)}]^T$ is a selection vector that is used for extracting the m-th column of an $M \times M$ matrix and we utilize that $\tilde{h}_{s,1} = 1$.

Next, assuming that the whitening matrix $\mathbf{S}^{-H}(n)$ is a lower triangular matrix, as in the case of the Cholesky based square-root decomposition, we note that its inverse matrix $\mathbf{S}^H(n)$ is also lower triangular and thus conclude that the first row of the matrix $\mathbf{S}^H(n)$ equals $\left[1/\left(\mathbf{S}^{-H}(n)\right)_{1,1}, 0, \ldots, 0\right]$, where $(\cdot)_{1,1}$ denotes the $(1,1)$ element of a matrix. Now, re-evaluating (19), $\delta(n)$ can be computed by

$$\delta(n) = g_1(n)/\left(\mathbf{S}^{-H}(n)\right)_{1,1}$$

and $\mathbf{b}_s(n)$ can be computed from $\mathbf{g}(n)$ and $\mathbf{S}^{-H}(n)$ by

$$\mathbf{b}_s(n) = \left(\mathbf{S}^{-H}(n)\right)_{1,1}\mathbf{g}(n)/g_1(n). \tag{20}$$

In practice, the whitened desired source RTFs vector in (20) is computed using estimated values of $\mathbf{g}(n)$ and $\mathbf{S}^{-H}(n)$. As mentioned earlier, the matrix $\mathbf{S}^{-H}(n)$ is recursively estimated using the IQRD method, which guarantees that the estimated square-root matrix is lower-triangular as required. For estimating the principal eigenvector of $\mathbf{\Phi}_z(n)$ we follow the high SNR based approximation that was developed in [1]. Explicitly, the covariance matrix in (17) is estimated by recursive averaging as

$$\hat{\mathbf{\Phi}}_z(n) = \lambda_z\hat{\mathbf{\Phi}}_z(n-1) + (1-\lambda_z)\mathbf{z}(n)\mathbf{z}^H(n)$$

with a forgetting factor of λ_z. Finally, the principal eigenvector is approximated as

$$\hat{\mathbf{g}}(n) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{\theta_m^*(n)}\left(\hat{\mathbf{\Phi}}_z(n) - \mathbf{I}\right)\mathbf{j}_m \tag{21}$$

with $\boldsymbol{\theta}(n) \triangleq \left(\hat{\mathbf{\Phi}}_z(n) - \mathbf{I}\right)\frac{\left(\hat{\mathbf{\Phi}}_z(n)-\mathbf{I}\right)\mathbf{e}_1}{\|\left(\hat{\mathbf{\Phi}}_z(n)-\mathbf{I}\right)\mathbf{e}_1\|}$. Note that the complexity of approximating the principal eigenvector using (21) is $O(M^2)$, which is significantly lower than $O(M^3)$, the complexity of performing an eigenvalue decomposition (EVD).

V. Experimental Results

We evaluate the performance of the proposed algorithm when applied in two scenarios, in a commercial mobile-phone and in a commercial smart-speaker. The performance, in terms of SNR and SER, is evaluated at the reference microphone, $\underline{x}_1(t)$, after the first stage echo-canceller, transformed for the sake of evaluation to the time-domain and denoted $\underline{y}_1(t)$, and at the output of the proposed method, $\underline{d}(t)$. Note that the performance is evaluated with real recordings comprising a mixture of desired talker, echo and noise. Therefore, the SNR and SER measured are approximated as the ratios of the speech power, estimated in an echo-free time-segment, and the noise power and the echo power, respectively estimated during a noise only time segment and an echo and noise time-segment. The signals are given at a sampling-rate of 16 kHz, the STFT window size and the overlap between frames are respectively set to $F = 256$ samples and $D = 96$ samples. The parameters of the method are set to $\mu = 1$, $\lambda_z = 0.99$ and $\eta_r = 0.0631$.

In the first scenario, a mobile-phone is used in a speaker-phone mode during a voice-call and is positioned at the center of the room at a distance of 30 cm from a desired talker. The mobile-phone comprises two microphones, positioned at the top and at the bottom edges, with a spacing of 7 cm, and a loudspeaker, positioned at the center of the back of the device, at a distance of 3.5 cm from the microphones. The recorded microphone signals comprise a mixture of a desired talker, an echo signal and a low ambient noise. The measured SNR and SER at the reference microphone are 32.0 dB and -8.9 dB, respectively. The signals of a partial time-segment at the reference microphone, at the corresponding output of the first-stage echo-canceller and at the output of the proposed method are depicted in Fig. 2. This time-segment includes: 1) a speech and noise segment; 2) an echo and noise segment; and 3) a speech, echo and noise segment. The spectra during an echo and noise time-segment and during a desired talker and noise time segment, computed at the reference microphone, at the corresponding output of the first-stage echo-canceller and at the output of the proposed method, are depicted in Fig. 3.

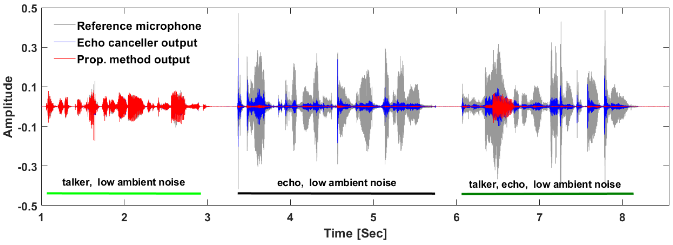

Figure 2: Signals from a real recording of a mobile-phone in low ambient noise: reference microphone signal in grey; echo-canceller output in blue; and proposed method output in red.

In the second scenario, a smart-speaker playing back an audio book is placed at the center of a room and located at a distance of 2.5 m from a desired talker. The cylinder shaped smart-speaker comprises 8 microphones, placed on its 8 cm diameter top face, and 4 loudspeakers, placed 18 cm below and at equal angles on its face. The recorded microphone signals comprise a mixture of a desired talker, an echo signal and diffuse living room noise. The measured SNR and SER at the reference microphone are 15.9 dB and -6 dB, respectively. The signals of a partial time-segment, containing desired talker, echo and noise, at the reference microphone, at the

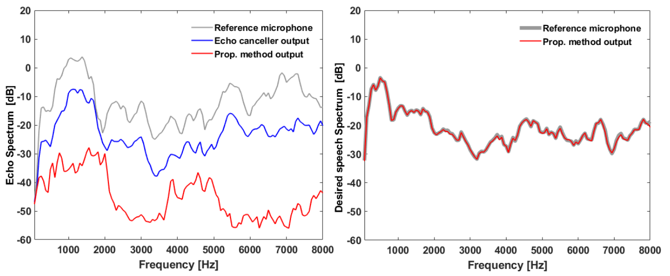

Figure 3: Spectra of the desired talker signal, on the right, and of the echo and noise signals, on the left, at different stages of the proposed method applied to a recording from a mobile-phone.

corresponding output of the first-stage echo-canceller and at the output of the proposed method, are depicted in Fig. 4.

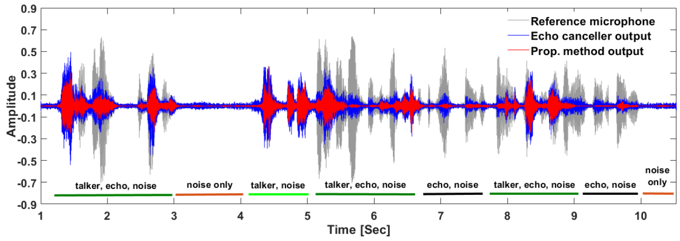

Figure 4: Signals from a real recording of a smart-speaker in noisy living room environment: reference microphone signal in grey; echo-canceller output in blue; and proposed method output in red.

In Table I we summarize the SNR and the SER, as measured at the reference microphone, at the output of the echo-canceller and at the output of the proposed method, for the two devices that were tested. The SER improvement is 13 dB – 13.7 dB at the output of the echo-canceller stage for both scenarios, and is 26.1 dB – 27.2 dB at the output of the proposed method for both scenarios. The SNR improvement at the output of the proposed method for the mobile-phone and smart-speaker is 4.1 dB and 11.3 dB, respectively. Evidently, from the depicted signals, their spectra and the average SNR and SER improvements, the proposed method is capable of successfully coping with the problems of echo and noise in realistic conditions and commercial devices, such as a mobile-phone and a smart-speaker.

Table I: Summary of the SNR and the SER as measured at the reference microphone, at the output of the echo-canceller and at the output of the proposed method for the mobile-phone and smart-speaker devices.

Stage/Criterion	SNR [dB]		SER [dB]	
	Meas.	Imp.	Meas.	Imp.
Mobile-phone				
Reference mic.	32	-	-8.9	-
Echo-canceller out.	32	0	4.1	13
Prop. method out.	36.7	4.7	17.2	26.1
Smart-speaker				
Reference mic.	15.9	-	-6	-
Echo-canceller out.	15.9	0	7.7	13.7
Prop. method out.	27.2	11.3	21.2	27.2

VI. Conclusions

The problems of echo and noise contaminating a desired talker signal in a communication or an entertainment device were considered, and a combined method comprising a linear echo-canceller followed by a weighted MVDR beamformer, designed to reduce noise and echo residues, was presented. The computational and memory complexities of the proposed algorithm are sufficiently low, making it appropriate for implementation in mobile devices. The RTF of the desired talker is estimated in the whitened domain, and does not require transforming it back to the *microphones signals* domain. This saves computations compared to [1]. Specifically, the proposed method requires the application of the IQRD method, instead of applying both IQRD and QRD methods. The proposed method was successfully applied to real recordings from two commercial devices, a mobile-phone and a smart-speaker.

References

[1] A. Barnov, V. Bar Bracha, and S. Markovich-Golan, “QRD based MVDR beamforming for fast tracking of speech and noise dynamics,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017.

[2] M. Zeller and W. Kellermann, “Fast and robust adaptation of DFT-domain Volterra filters in diagonal coordinates using iterated coefficient updates,” *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1589–1604, 2010.

[3] S. Malik and G. Enzner, “State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation,” *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 7, pp. 2065–2079, 2012.

[4] C. Hofmann, C. Huemmer, M. Guenther, and W. Kellermann, “Significance-aware filtering for nonlinear acoustic echo cancellation,” *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 113, 2016.

[5] W. Kellermann, “Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays,” in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 1. IEEE, 1997, pp. 219–222.

[6] K.-D. Kammeyer, M. Kallinger, and A. Mertins, “New aspects of combining echo cancellers with beamformers,” in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP’05). IEEE International Conference on*, vol. 3. IEEE, 2005, pp. iii–137.

[7] G. Reuven, S. Gannot, and I. Cohen, “Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller,” *Speech communication*, vol. 49, no. 7, pp. 623–635, 2007.

[8] S. Doclo, M. Moonen, and E. De Clippel, “Combined acoustic echo and noise reduction using GSVD-based optimal filtering,” in *Acoustics, Speech, and Signal Processing, 2000. ICASSP’00. Proceedings. 2000 IEEE International Conference on*, vol. 2. IEEE, 2000, pp. II1061–II1064.

[9] W. Herbordtt, S. Nakamura, and W. Kellermann, “Joint optimization of LCMV beamforming and acoustic echo cancellation for automatic speech recognition,” in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP’05). IEEE International Conference on*, vol. 3. IEEE, 2005, pp. iii–77.

[10] M. Kallinger, J. Bitzer, and K.-D. Kammeyer, “Interpolation of MVDR beamformer coefficients for joint echo cancellation and noise reduction,” 2001.

[11] O. L. Frost, “An algorithm for linearly constrained adaptive array processing,” vol. 60, no. 8, pp. 926–935, Aug. 1972.

[12] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[13] B. Widrow and S. D. Stearns, “Adaptive signal processing,” *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1985, 491 p.*, vol. 1, 1985.

[14] J. A. Apolinário, *QRD-RLS adaptive filtering*. Springer, 2009.

[15] S. S. Haykin, *Adaptive filter theory*. Pearson Education India, 2008.

[16] S. Markovich, S. Gannot, and I. Cohen, “Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[17] A. Bertrand and M. Moonen, “Distributed node-specific LCMV beamforming in wireless sensor networks,” *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 233–246, 2012.

[18] S. Markovich-Golan and S. Gannot, “Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 544–548.