Infinite Impulse Response Echo Canceller in STFT Domain for Reverberant Environments

Amos Schreibman and Shmulik Markovich-Golan

Communication and Devices Group, Intel corporation
Email: {amos.schreibman, shmulik.markovich-golan}@intel.com

***Abstract*—Acoustic echo cancellation and system identification in reverberant environments have been thoroughly studied in the literature. Theoretically, in a reverberant environment the Acoustic Impulse Response (AIR) relating the loudspeaker signal, denoted *reference*, with the corresponding signal component at the microphone, denoted *echo*, is of an infinite length and can be modeled as an Infinite Impulse Response (IIR) filter. Correspondingly, the echo signal can be modeled as an Auto Regressive Moving Average (ARMA) process. Yet, most methods for this problem adopt a Finite Impulse Response (FIR) system model or equivalently a Moving Average (MA) echo signal model due to their favorable simplicity and stability. Latter methods, denoted FIR-Acoustic Echo Canceller (AEC), employ an Adaptive Filter (AF) for tracking a possibly time-varying system and cancelling echo. Some contributions adopt an IIR system model and utilize it to derive a time-domain AEC and accurately analyze the room behaviour. An IIR system model has also been successfully applied in the Short Time Fourier Transform (STFT) domain for the dereverberation problem.**

In this contribution we consider an IIR model in the STFT domain and propose a novel online AEC algorithm, denoted IIR-AEC, which tracks the model parameters and cancels echo. The order of the feed-back filter, equivalent to the order of the Auto Regressive (AR) part of the echo signal model, can be designed to fit the acoustic model and the order of the feed-forward filter, equivalent to the order of the MA part of the echo signal model, is limited to a single tap, thereby requiring that the STFT window is longer than the *early* part of the AIR. The computational complexity of proposed IIR-AEC is comparable to a Recursive Least Squares (RLS) implementation of FIR-AEC. These methods are evaluated using real measured AIRs drawn from a recording campaign and the IIR-AEC is shown to outperform the FIR-AEC.

I. Introduction

Acoustic echo cancellation is a fundamental and imperative function in speech processing applications, such as hands-free voice-communication and speech recognition. Typically in these applications the system is comprised of one or more microphones and loudspeakers. Microphone signals pick-up the desired speech signal of a human user as well as acoustic echo, denoting a signal component which correspond to the signal emitted through the loudspeakers, denoted as the reference signal, and propagating in the environment [1]. Propagation of acoustic echo, affected by the frequency response of the loudspeakers, the direct-path between loudspeakers and microphones, reflections bouncing of objects and boundaries of the environment and the frequency response of the microphones, is intricate and time-varying. Sufficient cancellation of the echo signal is crucial for enhancing the quality, intelligibility and identifiability of desired speech.

In the current contribution reverberant environments are considered, in which the Acoustic Impulse Response (AIR) can be split into two parts. The first part, denoted *early*, contains the direct path and early reflections which are not *very dense* in time. The length of this part is denoted *mixing time* and an approximation of it is the square-root of the room volume [2]. The second part, denoted *late*, contains high-order reflections, is infinitely long and can be modeled as white-Gaussian process with exponentially decaying amplitude [2].

Adaptive filtering methods are widely used for cancelling echo [3], [4], [5], [6]. These methods model the AIR as an all-zero system, i.e., Finite Impulse Response (FIR) filter, thereby neglecting the contribution of the infinite AIR tail. Throughout this paper, these methods are referred to as FIR-Acoustic Echo Canceller (AEC). The performance of the AEC, measured by the echo cancellation level using the Echo Return Loss Enhancement (ERLE) metric, is a concave function of the FIR-AEC model order. Performance is impaired for too small model orders, since significant parts of the AIR are neglected, and for too large model orders, since the estimation error increases with the model order, assuming a finite and fixed estimation period. The latter behaviour is explained in [4].

Another modeling approach for the AIR is using a combined poles and zeros system model, i.e., Infinite Impulse Response (IIR) filter, see [7], [8], [9], [10], [11]. This model is capable of capturing the infinite AIR with a finite model order. The finite number of poles defines the order of the feed-back filter, which captures the infinite tail of the AIR and the finite number of zeros defines the order the feed-forward filter. Haneda et. al. [12] adopts an IIR model and propose a hybrid time-domain AEC which estimates the echo component by combing filtered versions of the reference and microphone signals. The authors claim that the filter that is applied to the microphone signal and models the tail of the infinite response of the IIR can be computed once per environment and is time- and space-invariant. Mohammed et. al. [13] propose a time-domain IIR structure and develop an efficient filter weights adaptation algorithm.

Time-domain IIR implementation does not outperform FIR counterpart [14]. Moreover, time-domain implementation of an AEC is computationally intensive due to the large model order. Alternative Short Time Fourier Transform (STFT) domain implementations benefit from the diagonalization property of Fourier transform, allowing them to treat frequencies independently [4]. Nakatani et al. [15] address the problem of dereverberating the desired speech signal, and propose applying an adaptive IIR filter in the STFT domain to the microphone signal, also denoted as the Weighted Prediction Error (WPE) dereverberation method. The dereverberation problem is more complicated than the echo cancellation problem, since it aims to equalize the AIR blindly, without knowing the reference signal. The WPE method obtains state-of-the-art performance for the dereverberation task.

In the current contribution we adopt an IIR model for the AIR in the STFT domain and suggest an online AEC which efficiently estimates the model parameters. The paper is organized as follows. The echo-cancellation problem is formulated in Section II, and common FIR-AEC solutions are presented in Section III. The proposed IIR-AEC method is described in Section IV and is evaluated and

978-9-0827-9703-9/19/$31.00

compared to an FIR-AEC in Section V. The paper is concluded in Section VI.

II. Problem Formulation

Consider the echo-cancellation problem of a single channel system, containing a single microphone and a single loudspeaker, situated in a reverberant environment with Reverberation Time (RT) RT_{60}. The problem is first formulated in the time domain, in which signals and systems, denoted by $\underline{\bullet}$, are sampled at a rate of f_s. Let

$$\underline{d}(t) \triangleq \underline{z}(t) + \underline{v}(t), \tag{1}$$

denote the microphone signal at discrete time-index t, which is comprised of a desired signal, denoted as $\underline{v}(t)$, and of an echo component, modeled as

$$\underline{z}(t) \triangleq \underline{h}(t) * \underline{x}(t) \tag{2}$$

where $\underline{x}(t)$ denotes the loudspeaker signal, also denoted as reference signal, $\underline{h}(t)$ denotes the AIR and $*$ denotes the convolution operator.

Polack [2] models the AIR as a white-noise process multiplied by an exponentially decaying envelope. This common model suggests that the duration of the AIR is infinite. For practical considerations, the AIR duration is typically assumed to be of the order of the RT, i.e., $\text{RT}_{60} f_s$ samples, and the signal is processed in the STFT domain. Let K denote the length of analysis and synthesis windows and let D denote the frame-shift. Following [16], for $K < f_s\text{RT}_{60}$, and neglecting cross-band terms, the time-domain formulation in (1) can be formulated in the STFT domain as:

$$d(n,k) = z(n,k) + v(n,k), \tag{3}$$

where n and k denote the time-frame and frequency-bin indices and

$$z(n,k) = \sum_{i=0}^{N_h-1} h(i,k)x(n-i,k), \tag{4}$$

$x(n,k)$, $v(n,k)$ and $h(n,k)$ respectively denote the STFT of $\underline{z}(t)$, $\underline{x}(t)$, $\underline{v}(t)$ and $\underline{h}(t)$. The length in frames of the echo transfer function (ETF) in the STFT domain, denoted by N_h, can be approximated as:

$$N_h \approx 1 + \frac{f_s\text{RT}_{60} - K}{D}. \tag{5}$$

The goal of echo-cancellation is to extract the desired signal $v(n,k)$ from the microphone signal $d(n,k)$, given the reference signal $x(n,k)$. Hereafter, the frequency-bin index k is omitted for brevity.

III. Background on FIR-AEC

The echo-cancellation problem is comprehensively surveyed in [3], [4]. Least-squares-based system identification methods (such as Least Mean Squares (LMS) and Recursive Least Squares (RLS)) are typically incorporated to estimate an FIR model of the ETF, defined as $\hat{h}(i)$ for $i = 0, \ldots, N_{\hat{h}}$. The reference signal is then filtered with the estimated system to yield the estimated echo component at the microphone, i.e.,

$$\hat{z}_{\text{MA}}(n) \triangleq \sum_{i=0}^{N_{\hat{h}}-1} \hat{h}(i)x(n-i). \tag{6}$$

An estimate of the desired signal is finally obtained by subtracting the estimated echo from the microphone signal

$$\hat{v}_{\text{MA}}(n) = d(n) - \hat{z}_{\text{MA}}(n). \tag{7}$$

Arrange the $N_{\hat{h}}$ coefficients of $\hat{h}(i)$ in a vector and define

$$\hat{\mathbf{h}} \triangleq \begin{bmatrix} \hat{h}(0) & \cdots & \hat{h}(N_{\hat{h}}-1) \end{bmatrix}^T, \tag{8}$$

where $\bullet^T$ denotes the transpose operator.

The optimal FIR-AEC, also denoted as the Minimum Mean Squared Error (MMSE) estimator, is designed to minimize the variance of the estimated desired signal, i.e.,

$$\hat{\mathbf{h}} \triangleq \underset{\hat{\mathbf{h}}}{\text{argmin}}\{\text{E}\left[|\hat{v}_{\text{MA}}(n)|^2\right]\}. \tag{9}$$

The solution to the echo-cancellation problem in (9) is the Wiener Filter (WF) (see [3], [4])

$$\hat{\mathbf{h}} = \left(\mathbf{R}_{xx}^{-1}\mathbf{r}_{xd}\right)^*, \tag{10}$$

where the cross-correlation vector is defined as

$$\mathbf{r}_{xd} \triangleq \begin{bmatrix} r_{xd}(0) & \cdots & r_{xd}(N_{\hat{h}}-1) \end{bmatrix}^T, \tag{11}$$

with

$$r_{xd}(l) \triangleq \text{E}\left[x(n-l)d^*(n)\right] \tag{12}$$

and the (i,j)-th element of the auto-correlation matrix $\mathbf{R}_{xx}$ is defined as

$$R_{xx}(i,j) \triangleq \text{E}\left[x(n-i)x^*(n-j)\right]. \tag{13}$$

For optimal performance, the length of the estimated ETF should be larger than its practical length, i.e., $N_{\hat{h}} \geq N_h$. The selection of the AEC parameters, namely K, D and $N_{\hat{h}}$, affect the complexity, delay and accuracy of its operation. Thanks to the diagonalization property of the Fourier transform, estimating the ETF can be done at each frequency independently. Optimal ETF identification involves solving an $N_{\hat{h}}$ linear equations system per frequency [4] and implies a computational complexity of $\mathcal{O}\left(N_{\hat{h}}\right)$ for a gradient-descent based LMS implementation or $\mathcal{O}\left(N_{\hat{h}}^2\right)$ for an RLS implementation. Minimizing the computational complexity is attained by increasing K, such that ETF length is minimized and reaches its asymptotic value of $N_{\hat{h}} = 1$. The latter complexity minimization is obtained at the cost of increasing the algorithmic delay, which is K. Moreover, as the residual echo due to system mismatch increases linearly with the model order $N_{\hat{h}}$, see [4], it is desirable to select $N_{\hat{h}} = N_h$, however, this requires obtaining an estimate of the RT.

In various applications, such as communication or hearing aid systems, the maximal algorithmic delay is constrained. In these cases, one can employ a Multi-Delay Filtering (MDF) technique [17]. The latter method sets K such that its corresponding delay is acceptable, and applies reduced computational complexity gradient descent methods, thereby however, sacrificing convergence time of the estimate. In other systems with fewer computational resources, under-modeling the length of the ETF, such that $N_{\hat{h}} < N_h$, is inevitable, which results in leakage of late echo components into the estimated desired signal. A block-diagram of the FIR-AEC is depicted in Figure 1.

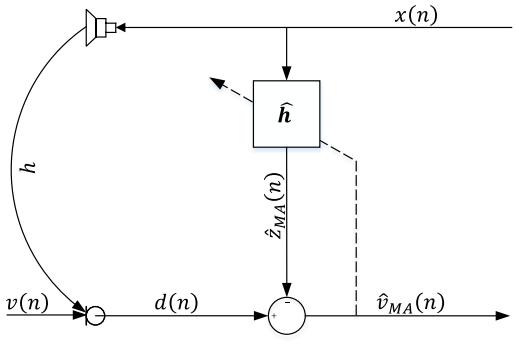

Figure 1: Block-diagram of FIR-AEC.

IV. Proposed IIR-AEC algorithm

As previously presented in Section I, the echo signal in a reverberant environment can be modeled as an Auto Regressive (AR) process in the time domain [2]. A similar model can be formulated in the STFT domain, e.g., in the state-of-the-art dereverberation method that was suggested in [15]. Here, we model the ETF as an IIR in the STFT domain and re-formulate the echo component $z(n)$ in (4) as an AR process, i.e.,

$$z_{\text{AR}}(n) = y(n) + q(n), \tag{14}$$

where

$$y(n) \triangleq bx(n), \tag{15}$$

is denoted the innovation process or the early component, assumed to be a white-noise process, i.e., $\mathrm{E}\left[y(n)y^*(n-i)\right] = 0$ for $i \neq 0$, and the component

$$q(n) \triangleq \sum_{i=1}^{N_a} a(i) z_{\text{AR}}(n-i) \tag{16}$$

is denoted the feed-back component and the AR model parameters, or equivalently the IIR parameters, are denoted $a(i)$, for $i = 1, \ldots, N_a$. Similarly to (3), the microphone signal is re-modeled as

$$d(n) \triangleq z_{\text{AR}}(n) + v(n). \tag{17}$$

Correspondingly, the desired signal is estimated by

$$\hat{v}(n) \triangleq d(n) - \hat{z}_{\text{AR}}(n). \tag{18}$$

We turn now to the computation of the IIR parameters $a(i)$ for $i = 1, \ldots, N_a$. Consider the cross-correlation between the early component $y(n-l)$ and the feed-back component $q(n)$ for $l > 0$, i.e., $r_{yq}(l) \triangleq \mathrm{E}\left[y(n-l)q^*(n)\right]$. Substituting (16), the cross-correlation can be formulated as

$$r_{yq}(l) = \sum_{i=1}^{N_a} a^*(i) r_{yz}(l-i). \tag{19}$$

where

$$r_{yz}(l) \triangleq \mathrm{E}\left[y(n-l)z^*_{\text{AR}}(n)\right]. \tag{20}$$

Considering (17) and since the reference signal and the desired signal are statistically independent, note that $r_{yd}(l) \triangleq \mathrm{E}\left[y(n-l)d^*(n)\right] = r_{yz}(l)$. Substituting the latter in (19) yields

$$r_{yq}(l) = \sum_{i=1}^{N_a} a^*(i) r_{yd}(l-i). \tag{21}$$

A linear set of N_a equations, derived from (21) for $l = 1, \ldots, N_a$, can be expressed in matrix notation as

$$\mathbf{r}_{yq} = \mathbf{R}_{yd}\mathbf{a}^*, \tag{22}$$

where

$$\mathbf{r}_{yq} \triangleq \begin{bmatrix} r_{yq}(1) & r_{yq}(2) & \cdots & r_{yq}(N_a) \end{bmatrix}^T \tag{23a}$$

$$\mathbf{a} \triangleq \begin{bmatrix} a(1) & a(2) & \cdots & a(N_a) \end{bmatrix}^T \tag{23b}$$

and

$$\mathbf{R}_{yd} \triangleq \begin{bmatrix} r_{yd}(0) & 0 & \cdots & 0 \\ r_{yd}(1) & r_{yd}(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r_{yd}(N_a-1) & r_{yd}(N_a-2) & \cdots & r_{yd}(0) \end{bmatrix}. \tag{24}$$

The lower triangular matrix form of $\mathbf{R}_{yd}$ results from the *causal* relation between $z_{\text{AR}}(n)$ and $y(n)$ in (14), (15), (16), i.e., $r_{yd}(l) = r_{yz}(l) = 0$, for $l < 0$.

The solution to (22) is given by

$$\hat{\mathbf{a}} = \left(\mathbf{R}_{yd}^{-1}\mathbf{r}_{yq}\right)^*, \tag{25}$$

Note that since $\mathbf{R}_{yd}$ is a lower triangular matrix, the matrix inversion in (25) can be avoided and (22) can be solved by back substitution [18] with a computational complexity of $\mathcal{O}(N_a^2)$.

In practice, the signals $y(n)$ and $q(n)$ are not observable, and the second order moments in (25) should be estimated. As presented in Section III, b and $y(n)$ are estimated by employing any standard adaptive filtering technique for $N_{\hat{h}} = 1$ (see [3], [4]) and substituting:

$$\hat{b} \triangleq \hat{h}(0) \tag{26a}$$

$$\hat{y}(n) \triangleq \hat{z}_{\text{MA}}(n). \tag{26b}$$

Extending the method to the generic ARMA case, i.e., $N_{\hat{h}} > 1$ and $N_a > 1$, is out of the current scope and will be treated in future work.

Considering (17) and (14), the feed-back component is estimated by

$$\hat{q}(n) = d(n) - \hat{y}(n). \tag{27}$$

Finally, the required second-order-statistics are recursively estimated by

$$\hat{r}_{yq}(l,n) = \alpha\hat{r}_{yq}(l,n-1) + (1-\alpha)\,\hat{y}(n-l)\,\hat{q}^*(n) \tag{28a}$$

$$\hat{r}_{yd}(l,n) = \alpha\hat{r}_{yd}(l,n-1) + (1-\alpha)\,\hat{y}(n-l)\,d^*(n) \tag{28b}$$

with $0 < \alpha < 1$ being the recursive-averaging factor.

The estimated IIR filter in (25) might become unstable if any of its poles reside outside of the unit circle. Instability in a certain frequency will result in exponentially increasing energy which will eventually corrupt the estimated desired signal. Instability can be avoided by checking the positions of the IIR filter poles, however this operation is computationally intensive. Here, we propose a simpler method for detecting instability. Considering (18) and (14), we expect that $\mathrm{Var}\{\hat{v}(n)\} \leq \mathrm{Var}\{d(n) - \hat{y}(n)\}$, where $\mathrm{Var}\{\bullet\}$ stands for short-term variance estimate. If the latter condition fails we reset the estimated IIR filter coefficients. Consequently, the echo is temporarily estimated by $\hat{z}_{\text{AR}}(n) \approx \hat{y}(n)$, until stability is reached.

A block-diagram of the proposed method is depicted in Figure 2.

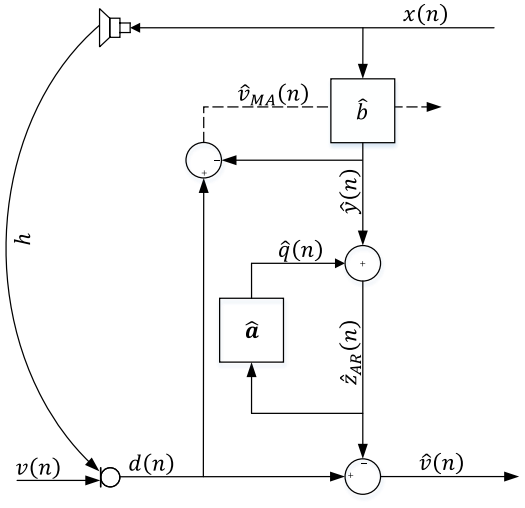

Figure 2: Proposed IIR-AEC algorithm.

V. Experimental Results

We evaluate the proposed IIR-AEC algorithm, where the Moving Average (MA) component $\hat{b}$ is estimated using an RLS-based AEC, and compare it to the an FIR-AEC, based on a multiple delay RLS [4], using a database of room impulse responses [19], obtained from the acoustic lab at Bar-Ilan University, Israel. The room dimensions are 6m $\times$ 6m $\times$ 2.4m, and its acoustic properties can be controlled by opening and closing various panels that are mounted on the walls, ceiling and floor, thereby changing their reflectivity. Two room configurations with $\text{RT}_{60} = 360$ms and $\text{RT}_{60} = 610$ms, and where the distance between the loudspeaker and microphone is 0.72m are evaluated.

The measured Energy Decay Curves (EDCs) [20] of the AIRs of $\text{RT}_{60} = 360$ms and of $\text{RT}_{60} = 610$ms as well as linear fitted curves which are theoretically expected from the exponentially decaying amplitude model [2] are respectively depicted in Fig. 3. Also depicted in Fig. 3 are the differences between the empirical EDCs and the theoretically expected ones. The transition time between the early and late reverberant parts of the AIR is approximated as the time in which the difference becomes low and the theoretical exponentially decaying EDC model matches the empirical one. Examining Figs. 3a, 3b, the early part of the $\text{RT}_{60} = 360$ms and of the $\text{RT}_{60} = 610$ms room configurations is approximately 10ms and 100ms, respectively.

As a reference signal we use either a synthetic white Gaussian process or a real speech signal, both sampled at 16kHz. The microphone signal is constructed by filtering the reference signal by the AIR and adding spatially-white noise, modeling the microphone noise, at an Signal-to-Noise Ratio (SNR) of 80dB. The performance of the AEC is measured by the ERLE metric, defined as

$$\text{ERLE}(n) \triangleq \frac{\sum_{k=0}^{K-1} \text{Var}\left\{z(n,k)\right\}}{\sum_{k=0}^{K-1} \text{Var}\left\{z(n,k) - \hat{z}(n,k)\right\}} \tag{29}$$

where the variance is estimated by recursive averaging over a window of 60 samples, which is equivalent to ~ 4ms.

The proposed IIR-AEC and the FIR-AEC are evaluated for orders of $N_a = 1,..,8$ and $N_{\hat{h}} = 1,..8$ and STFT window lengths of $K = 512, 1024, 2048, 4096$ for the $\text{RT}_{60} = 360$ms scenario, and of $K = 1024, 2048, 4096, 8192$ for the $\text{RT}_{60} = 610$ms scenario. The recursive averaging factor that is used by both algorithms is set to $\alpha = 0.99$.

The mean ERLE, averaged over time, for various STFT window lengths and model orders for the case of $\text{RT}_{60} = 360$ms is depicted in Fig. 4. As deduced from Fig. 3a, the late reverberation part of the AIR begins after 10ms, therefore the IIR-AEC is expected to model the AIR correctly for an STFT window that is longer than 160 samples. Clearly, the proposed IIR-AEC outperforms the FIR-AEC for all model orders and STFT window lengths. For model orders larger than 1 the improvement in ERLE is approximately 1.0dB, 1.5dB, 2.5dB and 4.5dB for STFT window lengths of 512, 1024, 2048 and 4096, respectively. Note that the ERLE of the FIR-AEC begins to drop for higher model orders. This performance drop can be related to the misadjustment noise which linearly increases with the over-estimated model order [4]. Interestingly, the proposed IIR-AEC does not seem to suffer from this performance drop when the model order is over-estimated. A theoretical analysis of the performance of the IIR-AEC is beyond the scope of the current contribution.

The mean ERLE, averaged over time, for various STFT window lengths and model orders for the case of $\text{RT}_{60} = 610$ms is depicted in Fig. 5. As deduced from Fig. 3b, the late reverberation part of the AIR begins after 100ms, therefore the IIR-AEC is expected to model the AIR correctly for an STFT window that is longer 1600 samples. The proposed IIR-AEC outperforms the FIR-AEC for STFT windows of $K = 4096, 8192$ and for all model orders. For model orders larger than 1 the improvement in ERLE is approximately 3.5dB and 5.5dB for STFT window lengths of 4096 and 8192, respectively. For STFT window lengths of $K = 1024, 2048$ there is no clear advantage to neither IIR-AEC or FIR-AEC.

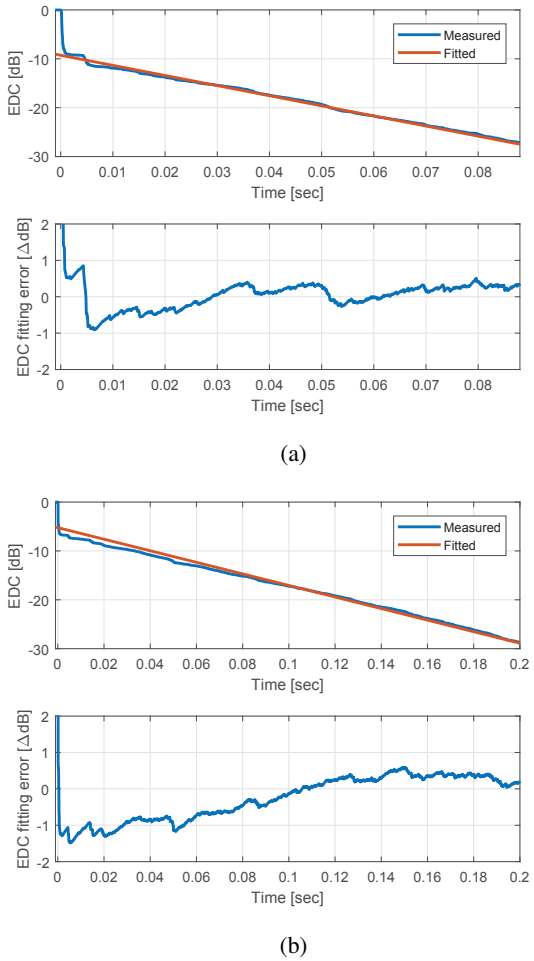

Figure 3: Measured Energy Decay Curves (EDCs), linear fits and their differences for $\text{RT}_{60} = 360$ms (a) and for $\text{RT}_{60} = 610$ms (b).

The ERLE versus time of the proposed IIR-AEC and of the FIR-AEC for a speech signal emitted in a room with $\text{RT}_{60} = 360$ms is depicted in Fig. 6 for model order of 4 and STFT window length of $K = 2048$. Evidently, the proposed IIR-AEC consistently outperforms the FIR-AEC, and the average improvement is 10dB. In our experiments, approximately 0.5% of the frequency-bins where detected as instable and reset for preventing divergence.

VI. Conclusions

Adopting an IIR model for the AIR in the STFT domain, we developed a novel online AEC, denoted IIR-AEC. The order of the feed-back filter, denoted N_a, can be designed to fit the acoustic environment, whereas the order of the feed-forward filter is restricted to a single tap. Consequently, the STFT window length should be larger than the early part of the AIR. The model parameters are recursively tracked with a computational complexity of $\mathcal{O}(N_a^2)$, which is comparable to an RLS implementation of an FIR-AEC. The proposed IIR-AEC and an FIR-AEC are evaluated using real AIRs drawn from a recording campaign [19], and the proposed method is shown to outperform the FIR-AEC. A theoretical analysis of the

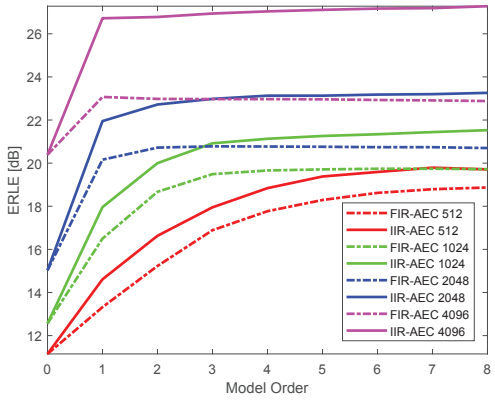

Figure 4: ERLE for proposed IIR-AEC and FIR-AEC with $\mathrm{RT}_{60} = 360$ms for model orders of $1, \ldots, 8$ and STFT window lengths of $K = 512, \ldots, 4096$.

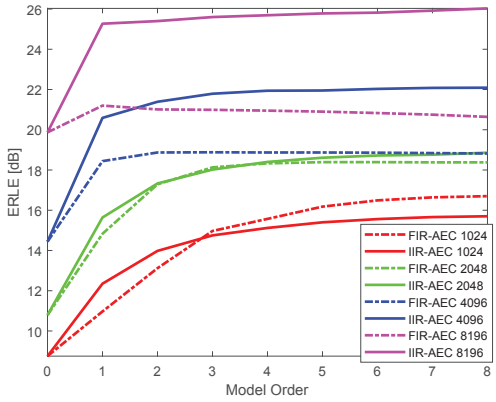

Figure 5: ERLE for proposed IIR-AEC and FIR-AEC with $\mathrm{RT}_{60} = 610$ms for model orders of $1, \ldots, 8$ and STFT window lengths of $K = 1024, \ldots, 8192$.

proposed method is beyond the scope of the current contribution and will be presented in future work.

References

[1] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[2] J.-D. Polack, "La transmission de l'energie sonore dans les salles," Ph.D. dissertation, 1988, thèse de doctorat dirigée par Bruneau, Michel Physique Le Mans 1988. [Online]. Available: http://www.theses.fr/1988LEMA1011

[3] B. Widrow and S. Stearns, *Adaptive Signal Processing*, Jan. 1985.

[4] S. Haykin, *Adaptive Filter Theory*, 4th ed. New Jersey: Prentice-Hall, 2002.

[5] M. Dentino, J. McCool, and B. Widrow, "Adaptive filtering in the frequency domain," *Proceedings of the IEEE*, vol. 66, no. 12, pp. 1658–1659, Dec 1978.

[6] E. Ferrara, "Fast implementations of LMS adaptive filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 474–475, August 1980.

[7] J. J. Shynk, "Adaptive IIR filtering," *IEEE Assp Magazine*, vol. 6, no. 2, pp. 4–21, 1989.

[8] M. Karjalainen, P. A. A. Esquef, P. Antsalo, A. Makivirta, and V. Valimaki, "AR/ARMA analysis and modeling of modes in resonant and reverberant systems," in *Audio Engineering Society Convention 112*, Apr 2002. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=11329

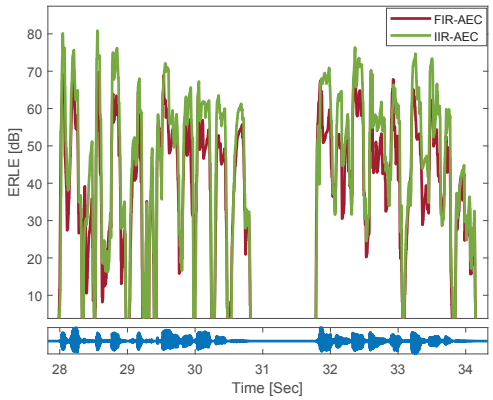

Figure 6: ERLE vs. time of proposed IIR-AEC (in green) and of FIR-AEC (in red) for a speech signal (in blue) in a room with $\mathrm{RT}_{60} = 610$ms for model order of 4 and STFT window length of $K = 2048$.

[9] S. J. Schlecht and E. A. Habets, "Modal decomposition of feedback delay networks," *arXiv preprint arXiv:1901.08865*, 2019.

[10] D. Beaton and N. Xiang, "Room acoustic modal analysis using Bayesian inference," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, p. 4480–4493, Jun 2017. [Online]. Available: http://dx.doi.org/10.1121/1.4983301

[11] S. J. Schlecht and E. A. Habets, "Time-varying feedback matrices in feedback delay networks and their application in artificial reverberation," *The Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1389–1398, 2015.

[12] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, p. 320–328, Apr 1994. [Online]. Available: http://dx.doi.org/10.1109/89.279281

[13] J. R. Mohammed and G. Singh, "An efficient RLS algorithm for output-error adaptive IIR filtering and its application to acoustic echo cancellation," in *2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing*. IEEE, 2007, pp. 139–145.

[14] A. P. Liavas and P. A. Regalia, "Acoustic echo cancellation: Do IIR models offer better modeling capabilities than their FIR counterparts?" *IEEE Transactions on signal processing*, vol. 46, no. 9, pp. 2499–2504, 1998.

[15] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 85–88.

[16] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[17] J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[18] G. H. Golub and G. Meurant, *Matrices, moments and quadrature with applications*. Princeton University Press, 2009, vol. 30.

[19] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*. IEEE, 2014, pp. 313–317.

[20] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 497–500, 1979.