Multi-Microphone acoustic echo cancellation using relative echo transfer functions

Conference Paper · October 2017

DOI: 10.1109/WASPAA.2017.8170029

2 authors:

María Luis Valero
Friedrich-Alexander-University of Erlangen-Nürnberg
12 PUBLICATIONS **66** CITATIONS
SEE PROFILE

Emanuël A. P. Habets
Friedrich-Alexander-University of Erlangen-Nürnberg
219 PUBLICATIONS **2,710** CITATIONS
SEE PROFILE

MULTI-MICROPHONE ACOUSTIC ECHO CANCELLATION USING RELATIVE ECHO TRANSFER FUNCTIONS

María Luis Valero and Emanuël A. P. Habets

International Audio Laboratories Erlangen*, Am Wolfsmantel 33, 91058 Erlangen, Germany
{maria.luis.valero,emanuel.habets}@audiolabs-erlangen.de

ABSTRACT

Modern hands-free communication devices, such as smart speakers, are equipped with several microphones, and one or more loudspeakers. The most straightforward solution to reduce acoustic echoes is to apply acoustic echo cancellation (AEC) to each microphone. Due to limited computational resources, the implementation of such a solution may not be feasible. To overcome this problem, a method is proposed that uses a primary estimated echo signal, obtained using state-of-the-art AEC, to compute the remaining, or secondary, acoustic echoes. To do this, relative transfer functions between primary and secondary acoustic echo signals, referred to as *relative echo transfer functions* (RETFs), are estimated and employed. In this work, the acoustic echo transfer functions (AETFs) and RETFs are modeled using convolutive transfer functions. Provided that the distance between microphones is small, the RETFs can be modeled using fewer partitions than the AETFs, which reduces the overall computational complexity.

Index Terms— Acoustic echo cancellation, relative transfer function estimation, adaptive filtering techniques

1. INTRODUCTION

Modern hands-free communication devices employ multiple microphones for, e.g., speech enhancement, room geometry inference or automatic speech recognition. Moreover, many smart devices, such as smart speakers or smart televisions, are equipped with one or more loudspeakers. Thus, microphones acquire, in addition to the desired near-end speech and background noise, the sound that is reproduced by the loudspeaker(s). The most commonly-used technique to reduce this particular electro-acoustic coupling is acoustic echo cancellation (AEC) [1]. AEC uses adaptive filtering techniques [2] to estimate the acoustic impulse responses (AIRs) between loudspeaker(s) and microphone(s). Subsequently, acoustic echo signals are computed, by filtering the loudspeaker signal with the estimated AIRs, and subtracted from the microphone signals. Given more than one microphone, the most straightforward solution to reduce the acoustic echoes is to place an acoustic echo canceler at the output of each microphone [3] as depicted in Fig. 1. Yet, the algorithmic complexity of such a solution is proportional to the number of microphones, and quickly exceeds the computational recourses that are available in current devices.

Provided that the microphones are relatively closely spaced, which is the case in many devices, the signals received by the microphones are very similar due to the relation between the AIRs [4]. In this work we exploit these similarities to reduce the complexity of multi-microphone AEC. More specifically, a primary echo signal is estimated using an AEC technique, and is then used to compute the secondary acoustic echoes. To do so, the relation in the frequency domain between the primary and secondary echo signals, in the following referred to as *relative echo transfer functions* (RETFs), are estimated and employed. It should be noted that RETFs model the relation between the frequency-domain representation of the acoustic echo paths, or acoustic echo transfer functions (AETFs). In this work, AETFs and RETFs are formulated in the short-time Fourier transform (STFT) domain using convolutive transfer functions (CTFs) [5]. Given closely spaced microphones, the advantage offered by the proposed approach resides in the fact that RETFs can be modeled using fewer partitions than AETFs [6]. Consequently, the computational complexity can be reduced at the cost of a moderate loss in performance.

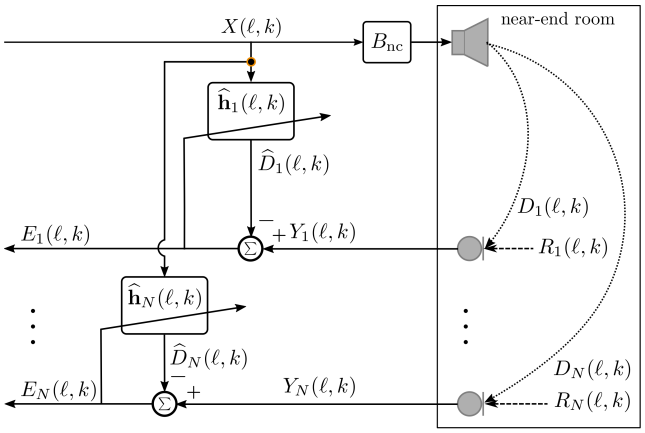

Figure 1: Multi-microphone AEC

Relative transfer functions (RTFs) [4, 5, 7] have been widely used in the context of microphone array processing [6,8–12]. Looking into more related applications, the RTFs between the primary error signal after cancellation and a secondary microphone signal were employed in [13] to estimate the residual echo power spectral density (PSD) for single-channel echo suppression. It is important to note that in contrast to the method presented in [13], the method proposed in the following uses RETFs to compute the secondary echo signals for multi-microphone AEC.

2. MULTI-MICROPHONE AEC

Given a hands-free communication set-up with one loudspeaker and N microphones, the n-th microphone signal can be expressed in the STFT domain as

$$Y_n(\ell,k) = D_n(\ell,k) + R_n(\ell,k),\ n \in \{1,\dots,N\}, \qquad (1)$$

*A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

978-1-5386-1631-4/17/$31.00 ©2017 IEEE

where ℓ and k are, respectively, the frame and frequency indexes. Further, $R_n(\ell,k)$ is the near-end signal, which comprises near-end speech and background noise, and $D_n(\ell,k)$ is the n-th acoustic echo. The latter is the result of the loudspeaker signal $X(\ell,k)$ being propagated through the room, and acquired by the n-th microphone. Its exact formulation in the STFT domain [14] is

$$D_n(\ell,k) = \sum_{b=-\infty}^{\infty} \underline{\boldsymbol{h}}_n^{\mathrm{H}}(b,k)\underline{\boldsymbol{x}}(\ell-b), \tag{2}$$

where $\underline{\boldsymbol{x}}(\ell) = [X(\ell,0),\ldots,X(\ell,K-1)]^{\mathrm{T}}$, superscripts $\cdot^{\mathrm{T}}$ and $\cdot^{\mathrm{H}}$ denote transpose and conjugate transpose, respectively, and K is the transform length. Further, the b-th partition of the n-th AETF is $\underline{\boldsymbol{h}}_n(b,k) = [H_n(b,k,0),\ldots,H_n(b,k,K-1)]^{\mathrm{T}}$, which is a vector containing all frequency dependencies $H_n(b,k,k')$, with $k' \in \{0,\ldots,K-1\}$.

It should be noted that AETFs in the STFT domain, which are extensively analyzed in [14], are non-causal. Moreover, the number of partitions, or input frames, that are necessary to estimate L AIR coefficients is $B = \lceil (L+K-1)/R \rceil + \lceil K/R \rceil - 1$, where R denotes the frame-shift between subsequent input frames. Due to the non-causality of the AETFs, $B_{\mathrm{nc}} = \lceil K/R \rceil - 1$ look-ahead frames of $X(\ell,k)$ are needed to compute the echo signals.

Let us assume that the frequency selectivity of the STFT analysis and synthesis windows is sufficient such that the frequency dependencies can be neglected. In addition, for notational brevity, we assume that a delay of B_{nc} frames is introduced to the reproduction path as depicted in Fig. 1. In practice, the capturing path is commonly delayed instead [5, 14]. Now, by using the convolutive transfer function (CTF) approximation [5], it is possible to write

$$D_n(\ell,k) \approx \sum_{b=-B_{\mathrm{nc}}}^{B-B_{\mathrm{nc}}-1} H_n^*(b,k)X(\ell-B_{\mathrm{nc}}-b,k), \tag{3}$$

where $\cdot^*$ denotes complex conjugation, and, for brevity, $H_n(b,k) \equiv H_n(b,k,k)$. Adaptive algorithms in AEC are driven by the error signal after cancellation, i.e.,

$$E_n(\ell,k) = Y_n(\ell,k) - \widehat{D}_n(\ell,k) = Y_n(\ell,k) - \widehat{\boldsymbol{h}}_n^{\mathrm{H}}(\ell,k)\boldsymbol{x}(\ell,k), \tag{4}$$

where $\widehat{\cdot}$ is used to denote estimates, $\widehat{\boldsymbol{h}}_n(\ell,k) = [\widehat{H}_n(\ell,-B_{\mathrm{nc}},k), \ldots, \widehat{H}_n(\ell,B-B_{\mathrm{nc}}-1,k)]^{\mathrm{T}}$ and $\boldsymbol{x}(\ell,k) = [X(\ell,k),\ldots,X(\ell-B+1,k)]^{\mathrm{T}}$. Most adaptive filters used in AEC are of gradient-descent type [2], thus a generic update equation is given by

$$\widehat{\boldsymbol{h}}_n(\ell+1,k) = \widehat{\boldsymbol{h}}_n(\ell,k) + \boldsymbol{M}_n(\ell,k)\boldsymbol{x}(\ell,k)E_n^*(\ell,k), \tag{5}$$

where $\boldsymbol{M}_n(\ell,k)$ is the step-size matrix of the adaptive filter, whose formulation depends on the specific adaptive algorithm used.

3. RETF-BASED MULTI-MICROPHONE AEC

Due to computational complexity restrictions, the implementation of multiple-microphone AEC as depicted in Fig. 1 is not always feasible. In this work, we propose to reduce the complexity by using a RETF-based approach, as depicted in Fig. 2.

3.1. Relative echo transfer functions

Let us denote, without loss of generality, the primary echo signal as $D_1(\ell,k)$ - defined as in (3). Under the previously made assumptions on the frequency dependencies, it is possible to write,

$$D_n(\ell,k) = \sum_{p=-\infty}^{\infty} A_n^*(p,k)D_1(\ell-p,k), \; n \in \{2,\ldots,N\} \tag{6}$$

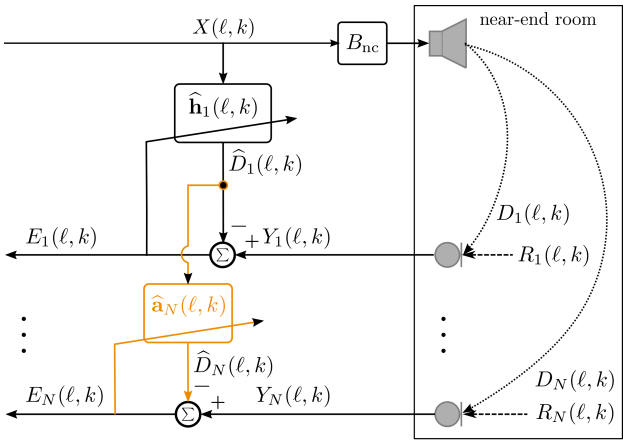

Figure 2: RETF-based multi-microphone AEC

where $A_n(p,k)$ is the p-th partition of the n-th *relative echo transfer function* (RETF). Provided that the distance between primary and secondary microphones is relatively small, it is possible to assume that the non-causal partitions of $A_n(p,k)\ \forall\ n$ are negligible. It is worth mentioning that a few non-causal time-domain coefficients are nevertheless modeled by $A_n(0,k)$. Under this assumption, no look-ahead is needed, and, consequently, no additional delay is introduced. Finally, using the CTF approximation leads to

$$D_n(\ell,k) \approx \sum_{p=0}^{P-1} A_n^*(p,k)D_1(\ell-p,k), \; n \in \{2,\ldots,N\} \tag{7}$$

where P is the number of RETF partitions.

As $D_1(\ell,k)$ is not observable, we propose to replace $D_1(\ell,k)$ by $\widehat{D}_1(\ell,k)$ in (7) that can be obtained using a state-of-the-art AEC. To estimate $A_n(p,k)\ \forall\ p$, we now minimize the error signal

$$E_n(\ell,k) = Y_n(\ell,k) - \widehat{\boldsymbol{a}}_{\boldsymbol{n}}^{\mathbf{H}}(\ell,k)\widehat{\boldsymbol{d}}_1(\ell,k), \tag{8}$$

where $\widehat{\boldsymbol{a}}_n(\ell,k) = [\widehat{A}_n(\ell,0,k),\ldots,\widehat{A}_n(\ell,P-1,k)]^{\mathrm{T}}$ is the n-th stacked vector of RETF partitions, and $\widehat{\boldsymbol{d}}_1(\ell,k) = [\widehat{D}_1(\ell,k),\ldots,\widehat{D}_1(\ell-P+1,k)]^{\mathrm{T}}$. The optimum filter in the mean-square-error sense, which is obtained by minimizing the quadratic cost-function $\mathcal{J}_n(\ell,k) = \mathrm{E}\{|E_n(\ell,k)|^2\}$, is equal to

$$\widehat{\boldsymbol{a}}_n^{\mathrm{opt}}(\ell,k) = \boldsymbol{\Psi}_1(\ell,k)^{-1}\boldsymbol{\psi}_{1n}(\ell,k), \tag{9}$$

where $\boldsymbol{\Psi}_1(\ell,k)$ is the covariance matrix of $\widehat{\boldsymbol{d}}_1(\ell,k)$, and $\boldsymbol{\psi}_{1n}(\ell,k)$ is the cross-correlation vector between $\widehat{\boldsymbol{d}}_1(\ell,k)$ and $Y_n(\ell,k)$, i.e.,

$$\boldsymbol{\Psi}_1(\ell,k) = \mathrm{E}\left\{\widehat{\boldsymbol{d}}_1(\ell,k)\widehat{\boldsymbol{d}}_1^{\mathrm{H}}(\ell,k)\right\},$$

$$\boldsymbol{\psi}_{1n}(\ell,k) = \mathrm{E}\left\{\widehat{\boldsymbol{d}}_1(\ell,k)Y_n^*(\ell,k)\right\},$$

where $\mathrm{E}\{\cdot\}$ denotes mathematical expectation. It should be noted that $\boldsymbol{\psi}_{1n}(\ell,k) \equiv \mathrm{E}\{\widehat{\boldsymbol{d}}_1(\ell,k)D_n^*(\ell,k)\}$, under the assumption that $\mathrm{E}\{\widehat{\boldsymbol{d}}_1(\ell,k)R_n^*(\ell,k)\} = \boldsymbol{0}$. Meaning that $\widehat{\boldsymbol{a}}_n^{\mathrm{opt}}(\ell,k)$ models the relation between the estimated primary AETF and the n-th secondary AETF. For instance, let us consider the trivial case $B = P = 1$, with $B_{\mathrm{nc}} = 0$, i.e., the multiplicative transfer function approximation [15], given which the n-th estimated RETF is equal to

$$\widehat{A}_n^{\mathrm{opt}}(\ell,0,k) = \widehat{H}_1(\ell,0,k)^{-1}H_n(\ell,0,k) \tag{10}$$

that, once the primary acoustic echo canceler has converged, is equal to $A_n(\ell,0,k)$ as defined in (7).

Compared to the problem of estimating RTFs from noisy observations [4, 5, 7], in our formulation there is no additional bias due to correlated noise components. Moreover, as the loudspeaker signal is known, the implementation of voice activity detectors (VADs) to control the estimation process is greatly simplified. However, as the previously made assumption on the statistical relationship between $\widehat{\boldsymbol{d}}_1(\ell,k)$ and $R_n(\ell,k)$ might be violated during double-talk situations, it is necessary to use a double-talk detector.

3.2. Adaptive RETF estimation

Adaptive filters can be used to track slowly time-varying RETFs. Due to the fact that $\widehat{\boldsymbol{d}}_1(\ell,k)$ is an estimate of the echo signal acquired by the primary microphone, it cannot be assumed to be uncorrelated across time. More precisely, the off-diagonals of $\boldsymbol{\Psi}_1(\ell,k)$ are not negligible if the STFT windows are short, or if the overlap between them is large. Taking this into consideration, Newton's method [2],

$$\widehat{\boldsymbol{a}}_n(\ell+1,k) = \widehat{\boldsymbol{a}}_n(\ell,k) + \eta\boldsymbol{\Psi}_1(\ell,k)^{-1}\widehat{\boldsymbol{d}}_1(\ell,k)E_n^*(\ell,k), \quad (11)$$

ensures a fast and stable convergence towards the optimum filter. In (11), η is a fixed step-size that is used to control the adaptation process. In practice, the covariance matrix $\boldsymbol{\Psi}_1(\ell,k)$ is approximated by averaging over time, e.g., by using a first-order recursive filter:

$$\widetilde{\boldsymbol{\Psi}}_1(\ell,k) = \beta\widetilde{\boldsymbol{\Psi}}_1(\ell-1,k) + (1-\beta)\widehat{\boldsymbol{d}}_1(\ell,k)\widehat{\boldsymbol{d}}_1^{\mathrm{H}}(\ell,k),$$

where time averages are denoted by $\widetilde{\cdot}$, and β is the forgetting factor.

4. COMPLEXITY ANALYSIS

In the following, the complexity in terms of additions and multiplications is analyzed. To this end, let us first look into the complexity per partition of an adaptive filter in the STFT domain, which is

$$\mathcal{O}(\mathrm{AF}) = \frac{3}{Q}\mathcal{O}(\mathrm{FFT}) + \underbrace{\mathcal{O}(\mathrm{CplxMult}) + 2K}_{\text{filtering \& cancellation}} + \mathcal{O}(\mathrm{Update}),$$

where $Q \in \{P, B\}$, $\mathcal{O}(\mathrm{FFT}) \approx 2K\log_2(K) - 4K$ is the complexity of a fast Fourier transform (FFT), $\mathcal{O}(\mathrm{CplxMult}) = 6K$ is the complexity of a complex multiplication of length K [16], and the complexity of the update equation $\mathcal{O}(\mathrm{Update})$ depends on the adaptive algorithm used. Hence, if N adaptive filters are used in parallel (one per microphone), the algorithmic complexity of multi-microphone AEC per partition is $N\mathcal{O}(\mathrm{AF})$.

The proposed method is able to reduce the algorithmic complexity if $P < B$. The reduction in algorithmic complexity is then given by the ratio

$$\frac{\mathcal{O}(\mathrm{Proposed})}{NB\mathcal{O}(\mathrm{AF})} = \frac{B\mathcal{O}(\mathrm{AF}) + (N-1)P\mathcal{O}(\mathrm{AF})}{NB\mathcal{O}(\mathrm{AF})}.$$

Consequently, if the same adaptive filter is used for the primary and secondary echo cancelers, the ratio is given by

$$\frac{\mathcal{O}(\mathrm{Proposed})}{NB\mathcal{O}(\mathrm{AF})} \approx \frac{1}{N} + \frac{N-1}{N}\frac{P}{B}.$$

If different adaptive filters are used for the AETF and RETF estimation, the computational complexity of the individual algorithms has to be carefully considered.

5. PERFORMANCE EVALUATION

To evaluate the proposed approach, three sets of experiments were conducted, for which the simulation set-up was designed as follows. Echo signals were generated by convolving a clean speech signal with simulated AIRs. The latter were generated using the image method [17] for a room of dimensions $3 \times 4 \times 2.5$ m^3, and reverberation time $T_{60} = 0.15$ and 0.35 s. The length of the simulated AIRs was $\mathcal{L} = 4096$ taps, at a sampling frequency of $F_\mathrm{s} = 16$ kHz. The AIRs were generated for a set-up with two microphones and one loudspeaker. The baseline set-up used a distance between loudspeaker and primary microphone of $\mathrm{l}_1 = 10$ cm, and between microphones of $\Delta = 1.5$ cm. The distance between the loudspeaker and the secondary microphone was $\mathrm{l}_2 = \mathrm{l}_1 + \Delta$ cm. The impact of these parameters on the performance was also analyzed. To this end, $\Delta = 3$ cm and $\mathrm{l}_1 = 20$ cm were also evaluated.

The signals were transformed to the STFT domain using Hamming analysis and synthesis windows of length $K = 512$ with 75% overlap, thus $R = 128$ samples. The adaptive algorithm used to estimate both the AETFs (5) and RETFs (11) was Newton's method. Thus, the step-size matrix in (5) was $\boldsymbol{M}_n(\ell,k) = \mu\widetilde{\boldsymbol{\Psi}}_\mathrm{x}(\ell,k)^{-1}$. As it is realistic to assume that the loudspeaker signal is uncorrelated across time, its covariance matrix was simplified by:

$$\widetilde{\boldsymbol{\Psi}}_\mathrm{x}(\ell,k) = \beta\widetilde{\boldsymbol{\Psi}}_\mathrm{x}(\ell-1,k) + (1-\beta)\boldsymbol{I} \odot \boldsymbol{x}(\ell,k)\boldsymbol{x}^{\mathrm{H}}(\ell,k),$$

where $\odot$ denotes element-wise multiplication, and I is the $B \times B$ identity matrix. Please note that in spite of this simplification, the normalization factors are still partition-dependent. The step-size factors were $\mu = 0.5/B$ and $\eta = 0.225/P$, and the forgetting factor was $\beta = 0.9$. Further, the adaptive filters and covariance matrices were not updated during speech pauses, and regularization was used to ensure the non-singularity of the covariance matrices. Finally, white Gaussian noise was added to the microphone signals to simulate a fixed segmental echo-to-noise ratio (SegENR). To make the differences in performance noticeable, a SegENR of 60 dB was used. Three sets of experiments were conducted:

1. The AIRs generated to simulate $T_{60} = 0.15$ s were truncated to length 256 taps, and used to generate the echo signals. The length of the estimated primary AIR was $L = 256$.
2. Simulated environment with $T_{60} = 0.15$ s, being the length of the estimated primary AIR $L = 256$ taps.
3. Simulated environment with $T_{60} = 0.35$ s, being the length of the estimated primary AIR $L = 1024$ taps.

Please recall that the number of AETF partitions that are necessary to completely estimate L AIR coefficients is $B = B_\mathrm{nc} + \lceil \frac{L+K-1}{R} \rceil$, thus at least K subsequent filter coefficients are partially estimated as well.

In all simulations, B partitions of the primary AETF were estimated, while the secondary AETFs and RETFs were estimated using different number of partitions $B_\mathrm{nc} < B' \leq B$ and P, respectively. The secondary echo signals were then obtained by convolving in the STFT domain the secondary AETFs with the loudspeaker signal, and the RETFs with the estimated primary echo signal. The echo return loss enhancement (ERLE) was used to measure the echo reduction in the secondary channel, with

$$\mathrm{ERLE}(\ell) = 10\log_{10}\frac{||\mathtt{d}_2(\ell)||_2^2}{||\mathtt{d}_2(\ell) - \widehat{\mathtt{d}}_2(\ell)||_2^2} \quad (12)$$

where $||\cdot||_2$ is the l_2-norm, and $\mathtt{d}_2(\ell) = [d_2(\ell R+1), \ldots, d_2(\ell R+K)]$ is the ℓ-th frame of the secondary acoustic echo in the time domain. The outcome of these simulations is depicted in Figs. 3 to 5,

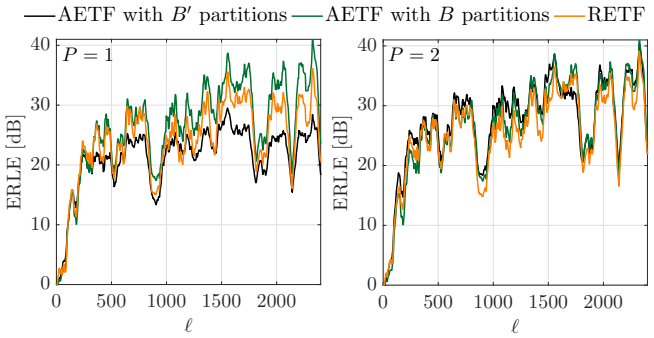

Figure 3: Comparison between AETF and RETF-based AEC with truncated AIRs and $L = 256$ taps

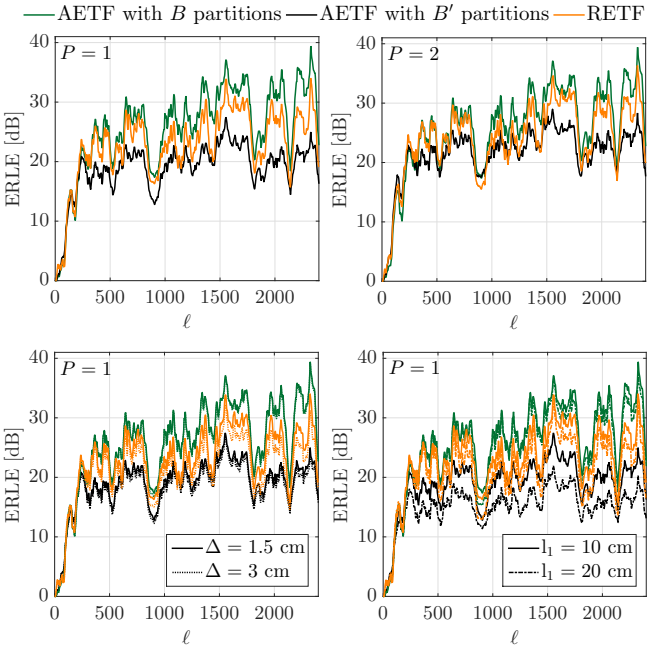

Figure 4: Comparison between AETF and RETF-based AEC with $T_{60} = 0.15$ s and $L = 256$ taps

where the ERLE measures were averaged over 60 frames for clarity. In these, the proposed RETF-based AEC is compared to state-of-the-art AEC using B and $B' = B_{\text{nc}} + P$ partitions for the AETF estimation. The latter condition is included to show a comparison with AETF-based AEC using fewer causal CTF partitions, which would also reduce the overall computational complexity.

Fig. 3 depicts the results corresponding to the simulations with truncated AIRs. The echo reduction obtained with $P = 1$ and 2, left and right sub-figures, are shown for all conditions under test. It can be observed that for $P = 1$, the RETF-based approach converges to a higher ERLE value than the AETF-based one with B' partitions, i.e., with only P causal partitions. Further, the performance is only moderately worse than that of the AETF-based approach with B partitions. For $P = 2$, all conditions under test perform similarly.

A performance comparison for $T_{60} = 0.15$ s, is shown in Fig. 4. The results depicted in the top-left and top-right sub-figures correspond to $P = 1$ and 2 for the baseline set-up. It can be observed that for $P = 1$, the RETF-based approach outperforms the AETF-based one with the same number of causal partitions. For $P = 2$,

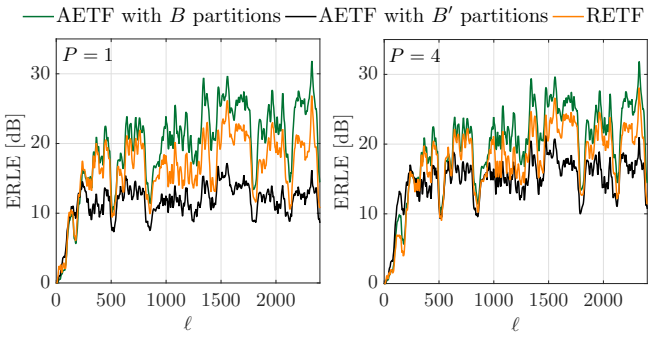

Figure 5: Comparison between AETF and RETF-based AEC with $T_{60} = 0.35$ s and $L = 1024$ taps

the performance of the AETF-based approach is visibly enhanced, and the advantage obtained by using the RETF-based approach is diminished. Nevertheless, the RETF-based approach still performs better, and nearly as well as the AETF-based one with $B = 9$ partitions. In the bottom, a comparison for different simulation set-ups is provided for $P = 1$. On the left, the results with different inter-microphone distances are shown. While, on the right, different distances between loudspeaker and primary microphone are evaluated. It can be observed that, for all conditions under test, enlarging any of these parameters impacts negatively on the canceler's performance. It should be noted that increasing the inter-microphone distance has a higher impact on the proposed approach, and that, in general, l_1 has a higher impact on the canceler's performance. Still, for the parameters used in these simulations, the proposed approach is able to outperform AETF-based AEC with equal number of causal partitions.

Finally, the results shown in Fig. 5 correspond to the simulated set-up with $T_{60} = 0.35$ s. The results obtained with $P = 1$ and 4 partitions are depicted in the left and right sub-figures. It can be observed that the proposed method outperforms, in both test-cases, the AETF-based approach with the same number of causal partitions. Further, for $P = 4$ it performs only moderately worse than the AETF-based AEC with $B = 15$.

6. CONCLUSIONS

A method to reduce the complexity of multi-microphone acoustic echo cancellation was proposed. The novel method uses state-of-the-art AEC techniques to estimate a primary acoustic echo signal, which is then used to compute all secondary acoustic echoes. To achieve this, *relative echo transfer functions* (RETFs) are estimated and employed. Provided that the distance between microphones is small, RETFs can be modeled using a lesser number of causal CTF partitions than AETFs, which leads to a reduction in computational complexity. It was shown that by using the proposed RETF-based AEC, the number of estimated partitions can be reduced, at the cost of a moderate loss in performance. In addition, it was shown that the RETF-based AEC outperforms the AETF-based AEC with equal number of causal CTF partitions. Finally, it was shown that by using a limited number of RETF partitions, the performance of the proposed RETF-based AEC is comparable to that of more computationally expensive state-of-the-art AEC implementations.

7. REFERENCES

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A practical Approach*. New Jersey, USA: Wiley, 2004.

[2] S. Haykin, *Adaptive Filter Theory*, 4th ed. New Jersey, USA: Prentice-Hall, 2001.

[3] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proc. IEEE ICASSP*, Munich, Germany, Apr. 1997, pp. 219–222.

[4] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sept. 2004.

[5] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[6] ——, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, Sept. 2009.

[7] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Process.*, vol. 44, no. 8, pp. 2055–2063, 1996.

[8] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[9] T. Dvorkind and S. Gannot, "Speaker localization in a reverberant environment," in *Proc. the 22nd convention of Electrical and Electronics Engineers in Israel (IEEEI)*, Tel-Aviv, Israel, Dec. 2002, pp. 7–7.

[10] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, Jan. 2005.

[11] 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. 7–8, pp. 623–635, Aug. 2007.

[12] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 4, no. 11, pp. 2171 – 2186, Nov. 2016.

[13] C. Yemdji, M. Mossi Idrissa, N. Evans, C. Beaugeant, and P. Vary, "Dual channel echo postfiltering for hands-free mobile terminals," in *Proc. IWAENC*, Aachen, Germany, Sept. 2012, pp. 1–4.

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

[15] ——, "On multiplicative transfer function approximation in the short-time fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337 – 340, May 2007.

[16] R. M. M. Derkx, G. P. M. Engelmeers, and P. C. W. Sommen, "New constraining method for partitioned block frequency-domain adaptive filters," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 2177–2186, 2002.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.