AEC capability verification of varying loudspeaker-microphone configurations

Simon Grimm[1], Markus Kasemann[1], Jürgen Freudenberger[2]

[1] *beyerdynamic GmbH & Co. KG, 74072 Heilbronn, Deutschland, Email: {simon.grimm, markus.kasemann}@beyerdynamic.de*

[2] *Institute for System Dynamics, 78462 Konstanz, Deutschland, Email: jfreuden@htwg-konstanz.de*

Abstract

Acoustic Echo Cancellation (AEC) plays a crucial role in speech communication devices to enable full-duplex communication. AEC algorithms have been studied extensively in the literature. However, device specific details like microphone or loudspeaker configurations are often neglected, despite their impact on the echo attenuation or near-end speech quality. In this work, we propose a method to investigate different loudspeaker-microphone configurations with respect to their contribution to the overall AEC performance. A generic AEC system consisting of an adaptive filter and a Wiener post filter is used for a fair comparison between different setups. We propose the near-end-to-residual-echo ratio (NRER) and the attenuation-of-near-end (AON) as quality measures for the full-duplex AEC performance.

Introduction

Acoustic echo cancellation is an important part of the signal enhancement in speech communication devices and conferencing systems. The ICASSP 2021 Acoustic Echo Cancellation Challenge demonstrates that AEC is still an active research area [1] with new approaches, e.g., based on deep learning techniques and supervised learning algorithms [2].

The purpose of speech communication devices is to pick up near-end speech signals and reproduce far-end signals without disturbance. AEC is employed to remove echos of the far-end signal in the near-end signal path [3, 4, 5]. The typical building blocks of AEC systems are adaptive filters for adaptive identification of the acoustic echo path, control mechanisms for the step-size control, and post filters for the residual echo suppression. The popular gradient-based normalized least-mean-square (NLMS) algorithm, the affine projection algorithm (APA), and the recursive least-squares (RLS) algorithm are thoroughly described in [6]. The residual echo suppression is typically based on frequency-selective post-processing. In principle, the operation of a residual echo suppression post filter is very similar to that of a noise suppression filter [5], where the post filter requires an estimate of the power spectral density (PSD) of the residual echo signal. AEC has been studied extensively in the literature, where the research typically considers the mentioned system components.

However, the objective performance evaluation of AEC systems is still an open issue, because conventional objective metrics such as echo return loss (ERL), echo return loss enhancement (ERLE), and objective speech quality evaluations like PESQ do not correlate well with subjective speech quality tests [1]. Moreover, device specific details like microphone or loudspeaker placement are often neglected in the system analysis, despite their contribution to the near-end speech quality and the AEC performance. For instance, the microphone characteristic impacts the near-end speech quality and the echo return loss. Common measures like ERL and ERLE can be used to judge the AEC contribution of different setups, but these measures do not reflect the influence of the near-end signal.

In this work, we address the influence of the microphone and loudspeaker setup on the AEC performance. We propose a method to compare different device configurations with respect to their overall double talk performance. Using recordings of echo and near-end signals, we employ a generic AEC algorithm consisting of an adaptive filter and an omniscient Wiener post filter [3, 4, 5]. The near-end signal quality depends on the pickup quality of the microphone and the degradation by the post filter. Since the post filter directly depends on power spectral densities of the near-end and the residual echo, we propose the near-end-to-residual-echo ratio (NRER) and the related attenuation-of-near-end (AON) as measures for the setup evaluation. Experimental results for different loudspeaker-microphone setups and varying microphone polar patterns are shown.

In the following sections, the signal model and the notation are described. Furthermore, the reference AEC system model and the performance measures are introduced. This is followed by a discussion of the experimental results.

Signal model and notation

For the acoustic echo cancellation (see Figure 1), we consider a linear and time-invariant acoustic system consisting of a loudspeaker, the room and a microphone, which is modeled by the acoustic transfer function H. The echo signal D can be written in the frequency domain as

$$D(\kappa, \nu) = X(\kappa, \nu) \cdot H(\nu) \tag{1}$$

where κ and ν correspond to the block and frequency index and $X(\kappa, \nu)$ is the AEC reference signal.

The microphone signal $Y(\kappa, \nu)$ can be expressed as the superposition of the echo $D(\kappa, \nu)$ and the near-end signal $S(\kappa, \nu)$ as

$$Y(\kappa, \nu) = D(\kappa, \nu) + S(\kappa, \nu). \tag{2}$$

The estimated echo signal $\hat{D}(\kappa, \nu)$ is related to the estimate of the acoustic transfer function $\hat{H}(\kappa, \nu)$ by

$$\hat{D}(\kappa, \nu) = X(\kappa, \nu) \cdot \hat{H}(\kappa, \nu) \tag{3}$$

and the residual echo $R(\kappa, \nu)$, which remains after the echo cancellation, depends on the mismatch between $H(\nu)$ and $\hat{H}(\kappa, \nu)$

$$R(\kappa, \nu) = X(\kappa, \nu) \cdot \left\{ H(\nu) - \hat{H}(\kappa, \nu) \right\}. \quad (4)$$

The output $E(\kappa, \nu)$ of the acoustic echo canceler is described as

$$E(\kappa, \nu) = Y(\kappa, \nu) - \hat{D}(\kappa, \nu) \quad (5)$$
$$= R(\kappa, \nu) + S(\kappa, \nu). \quad (6)$$

The generic AEC algorithm

In order to evaluate different hardware configurations regarding their AEC capability, we propose an AEC algorithm as a tool to fulfill this task. The algorithmic structure is therefore designed as generically as possible to allow a fair comparison between different setups. The generic AEC algorithm is depicted in Figure 1. Since most commonly, an AEC algorithm consists of an adaptive filter for echo cancellation in combination with a post filter for residual echo suppression, we decided for this combination. For the adaptive filter, we use a normalized-least-mean-square (NLMS) algorithm in combination with an omniscient Wiener post filter for residual echo suppression.

In a typical AEC system, only the signals $Y(\kappa, \nu)$, $X(\kappa, \nu)$ and $E(\kappa, \nu)$ are available, while all other signals have to be estimated. We overcome this by recording the echo and near-end signals separately. This allows us to have all signals in the AEC structure available for performance evaluation. The algorithm is applied in parallel, which means:

- The NLMS filter is adapted only on the echo signal using an echo voice activity detection
- The omniscient post filter is calculated based on the near-end and residual echo PSDs
- The post filter is only applied to the near-end signal to verify the full duplex capability of the AEC system

Omniscient Wiener post filter

For the omniscient Wiener post filter, the required near-end and residual echo PSDs are calculated blockwise as

$$\phi_S(\kappa, \nu) = \gamma \cdot \phi_S(\kappa - 1, \nu) + (1 - \gamma) \cdot |S(\kappa, \nu)|^2 \quad (7)$$
$$\phi_R(\kappa, \nu) = \gamma \cdot \phi_R(\kappa - 1, \nu) + (1 - \gamma) \cdot |R(\kappa, \nu)|^2 \quad (8)$$

where γ is a recursive smoothing factor. For the post filter law, the required short time near-end-to-residual-echo ratio $\zeta(\kappa, \nu)$ is calculated as

$$\zeta(\kappa, \nu) = \frac{\phi_S(\kappa, \nu)}{\phi_R(\kappa, \nu)} \quad (9)$$

and the resulting Wiener filter function is obtained by

$$G_{WF}(\kappa, \nu) = \frac{\zeta(\kappa, \nu)}{\zeta(\kappa, \nu) + \mu} = \frac{\phi_S(\kappa, \nu)}{\phi_S(\kappa, \nu) + \mu \cdot \phi_R(\kappa, \nu)} \quad (10)$$

where μ is an overestimation factor used to emulate the behavior of real-world residual echo PSD estimates. The post filter affected near-end signal $\tilde{S}(\kappa, \nu)$ is obtained by

$$\tilde{S}(\kappa, \nu) = G_{WF}(\kappa, \nu) \cdot S(\kappa, \nu). \quad (11)$$

Note that the Wiener filter law in Eq.(10), and therefore the spectral degradation of $S(\kappa, \nu)$ in Eq.(11), is closely related to the ratio of the near-end and the residual echo PSDs.

AEC performance measures

In the following section, measures to evaluate the AEC performance are introduced. Since all measures are calculated in the time domain, we use lower case letters for the time domain representation of the signals instead of the corresponding upper case letters in the frequency domain. We assume that all signals have a length of L samples and k is used as the discrete time index. At first, the echo related measures are presented

Echo return loss (ERL) [7]:

$$ERL = 10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{L-1} x(k)^2}{\sum_{k=0}^{L-1} d(k)^2} \right) \text{dB} \quad (12)$$

The echo return loss can be interpreted as an echo coupling factor from the reference to the microphone input.

Echo return loss enhancement (ERLE) [4]:

$$ERLE = 10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{L-1} d(k)^2}{\sum_{k=0}^{L-1} r(k)^2} \right) \text{dB} \quad (13)$$

The echo return loss enhancement is the contribution of the acoustic echo canceler to the overall echo reduction.

ACOM:

$$ACOM = ERL + ERLE \quad (14)$$

The ACOM is the overall echo reduction at the output of the acoustic echo canceler and is calculated as the sum of ERL and ERLE.

The presented measures only consider the echo signal and its acoustic properties, but neglect the near-end signal despite its contribution to the full-duplex performance. Therefore some new measures are introduced in the following, which explicitly take the near-end signal into account:

Near-end-to-echo ratio (NER):

$$NER = 10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{L-1} s(k)^2}{\sum_{k=0}^{L-1} d(k)^2} \right) \text{dB} \quad (15)$$

The NER can be interpreted as the "input SNR" at the microphone regarding the echo disturbance.

Near-end-to-residual-echo ratio (NRER):

$$NRER = 10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{L-1} s(k)^2}{\sum_{k=0}^{L-1} r(k)^2} \right) \text{dB} \quad (16)$$

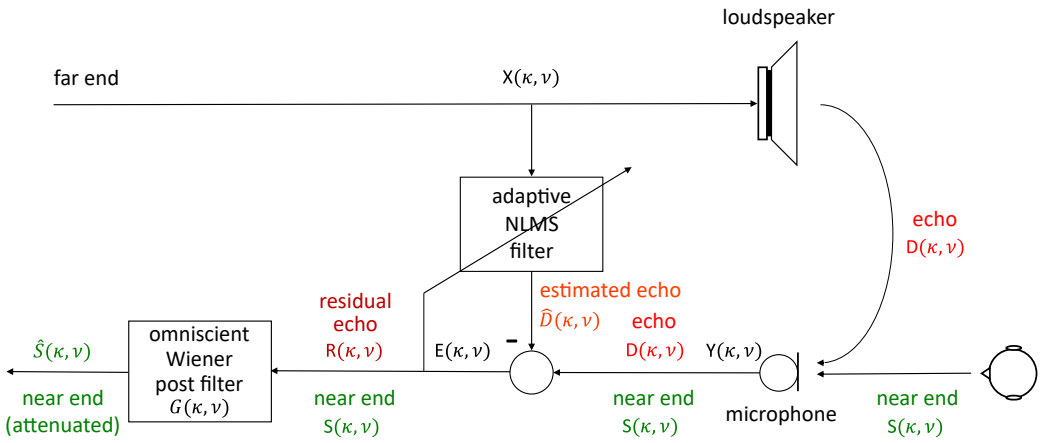

Figure 1: The proposed generic algorithm block structure for AEC capability verification of different acoustic hardware setups

The NRER can be interpreted as the "output SNR" of the echo canceler regarding the echo disturbance. Note that this measure is crucial, since it directly relates to the full-duplex capability, which depends on the near-end and the residual echo signals as has already been shown in Eq.(10) and (11).

Attenuation-of-near-end (AON):
The AON reflects the influence of the adaptive Wiener post filter on the near-end signal. It is calculated in mel-bands to acquire a measure for the spectral attenuation which is related to the human perception

$$AON = 10 \cdot \log_{10} \left(\sum_{\lambda=0}^{M-1} \frac{\phi_S(\lambda)}{\phi_{\tilde{S}}(\lambda)} \right) \text{dB}. \qquad (17)$$

λ denotes the mel-band index and M corresponds to the highest mel-band. ϕ_S and $\phi_{\tilde{S}}$ are the mel-band PSDs of the near-end and the post-filter affected near-end signals. The mel-band PSDs are calculated over all signal frames that contain double-talk (near-end and echo signals are present at the same time).

Experimental Results

In the following, the proposed generic AEC structure and the performance measures are used to examine the AEC capability of two different experimental setups named as Scenario 1 and 2. Scenario 1 investigates different distances between a loudspeaker and a microphone, while in Scenario 2 the influence of different polar patterns for the microphone of a speakerphone are investigated. All measures were performed in a typical office room with a reverberation time $RT_{60} \approx 0.5s$ and an approximately diffuse sound field. The sampling rate was 16kHz and the adaptive filter length was set to 512 samples with the NLMS calculated in the frequency domain. The PSD smoothing factor was set to $\gamma = 0.95$ and the overestimation factor of the post filter was $\mu = 24$. The ITU-T P.501 signals for AEC performance (single and double talk files of clause 7.3) are used for near-end and echo speech. The near-end signal was played back by a Genelec 8040 BMP loudspeaker.

Scenario 1

Scenario 1 is depicted in Figure 2. It is examined how different distances d between the loudspeaker and a microphone affect the AEC capability of the setup. Even if this setup comparison is quite trivial, it can be helpful to examine the best microphone position, e.g. in a car environment or a conferencing room, where the location of the near-end speaker and the loudspeaker are assumed to be known.

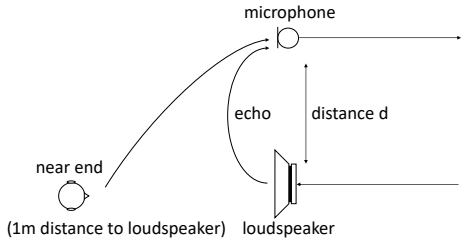

Figure 2: Scenario1 - varying distance between loudspeaker and microphone

By looking at the results in Table 1, it can be observed that the ERL increases with increased distance d. Simultaneously the ERLE, and therefore the performance of the acoustic echo canceler, decreases. By looking at the ACOM, it can be observed that the overall echo reduction is improved the further the distance d is increased, which suggests, that for $d = 80$cm the best AEC performance is achieved. However, by looking at the measures in Table 2, which take the near-end signal into account, it can be seen that the best NRER is achieved for a distance of $d = 40$cm and the performance gets worse for $d = 80$cm. This is also shown by the AON, which highlights that the near-end signal is least affected by the post filter for a distance of $d = 40$cm.

Scenario 2

For Scenario 2, as depicted in Figure 3, it is investigated which microphone polar pattern is best suited regarding

Table 1: Scenario1 - AEC performance measures 1

distance d	ERL	ERLE	ACOM
5cm	1.09 dB	19.09 dB	20.18 dB
7.5cm	5.98 dB	18.32 dB	24.30 dB
20cm	15.61 dB	14.36 dB	29.97 dB
40cm	18.92 dB	13.45 dB	32.37 dB
80cm	22.84 dB	10.19 dB	33.03 dB

Table 2: Scenario1 - AEC performance measures 2

distance d	NER	NRER	AON
5cm	-30.86 dB	-11.77 dB	13.27 dB
7.5cm	-22.24 dB	-7.73 dB	9.70 dB
20cm	-16.06 dB	-1.71 dB	5.76 dB
40cm	-13.12 dB	0.33 dB	5.07 dB
80cm	-11.42 dB	-1.23 dB	5.45 dB

the AEC performance for a speakerphone with a down-firing loudspeaker in a typical office situation.

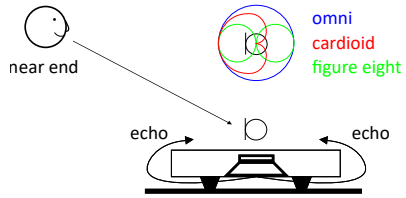

Figure 3: Scenario2 - varying microphone polar pattern

As can be observed by the results in Table 3, the ERL is increased by nearly 6 dB for the cardioid pattern compared with the omnidirectional pattern and by almost 15 dB for the figure eight pattern. However, the ERLE is the worst for the figure eight pattern. By looking at the ACOM, it can be observed that the cardioid and figure eight patterns achieve a similar echo reduction performance. However, by taking the near-end signal into account, it can be observed from Table 4, that the NRER, and therefore the full-duplex capability, is the highest for the figure eight polar pattern. This can also be observed from the AON, which shows that the figure eight pattern achieves the least spectral degradation of the near-end signal in the generic AEC algorithm.

Conclusions

In this work, a method that allows a fair comparison regarding the AEC capability of different hardware configurations was presented. By using a generic AEC algorithm structure that allows access to all necessary signals, the required AEC performance measures can be calculated. Besides the well-known values ERL, ERLE, and

Table 3: Scenario2 - AEC performance measures 1

polar pattern	ERL	ERLE	ACOM
omni directional	10.35 dB	21.17 dB	31.52 dB
cardioid	16.07 dB	20.47 dB	36.54 dB
figure eight	25.02 dB	11.48 dB	36.50 dB

Table 4: Scenario2 - AEC performance measures 2

polar pattern	NER	NRER	AON
omni directional	-31.10 dB	-9.94 dB	9.92 dB
cardioid	-27.36 dB	-6.89 dB	8.02 dB
figure eight	-17.26 dB	-5.78 dB	6.96 dB

ACOM, the new performance measures NER, NRER, and AON are introduced. These values take the near-end energy into account. Experimental results for specific hardware configurations and acoustic scenarios show that the echo-dependent performance measures ERL, ERLE, and ACOM are not sufficient for full-duplex AEC performance evaluation. It is demonstrated that the NRER and AON are more suited since they are directly related to the speech degradation of the near end by the post-filtering during double talk periods. Therefore, the new measures offer a more accurate perspective on the full-duplex AEC performance.

References

[1] Kusha Sridhar, Ross Cutler, Ando Saabas, Tanel Parnamaa, Markus Loide, Hannes Gamper, Sebastian Braun, Robert Aichner, and Sriram Srinivasan, "ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 151–155.

[2] Hao Zhang, Ke Tan, and DeLiang Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions," in *Proc. Interspeech 2019*, 2019, pp. 4255–4259.

[3] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Wiley & Sons, New Jersey, 2004.

[4] E. Hänsler and G. Schmidt, Eds., *Topics in Acoustic Echo and Noise Control: Selected Methods for the Cancellation of Acoustical Echoes, the Reduction of Background Noise, and Speech Processing*, Springer, Berlin, Heidelberg, 2006.

[5] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, *Acoustic echo control*, Academic press library in signal processing, 2014.

[6] S. Haykin, *Adaptive filter theory*, Prentice Hall, 2001.

[7] *IEEE Standard Method for Measuring Transmission Performance of Handsfree Telephone Sets*, IEEE Std 1329-1999 , pp.1-96, 2 May 2000