Realtime Improvement of Audio Signals in Underwater Communication

Michael Hochmuth [1], Tim Owe Wisch[2], Peter Eisert[3], Gerhard Schmidt[2]

[1] *Christian-Albrechts-Universität, 24118 Kiel, E-Mail: mail@michaelhochmuths.de*
[2] *Christian-Albrechts-Universität, 24118 Kiel, E-Mail: {gus,timw}@tf.uni-kiel.de*
[3] *Humboldt Universität zu Berlin, 10099 Berlin, E-Mail: eisert@informatik.hu-berlin.de*

Abstract

In this work, we implement speech signal enhancements in the context of underwater telephony. The system mostly addresses issues of the “dry end”, covering acoustic echoes, local stationary noise, and gain fluctuations. It consists of a linear acoustic echo canceller with a sophisticated timevariant and nonlinear control strategy, a postfilter for suppressing noise and residual echoes, and both automatic and noise-dependent gain control algorithms. The system relies on estimations of noise spectra, coupling factors, and delays as well as voice activity detections for different signals. The acoustic echo canceller includes a detector for room changes for quick adaptation and recovery from errors. It was tested successfully with both regular and underwater transmission channels, in conversational and simulated setups, and shown to significantly improve the speech quality in conversations.

Introduction

Underwater telephony is a key element for operations involving submarines, divers and other underwater entities. While the technology is constantly being evolved to increase the quality of this very limited transmission channel, the situation at each end is often not as advanced as for regular telephony setups. In this work, the operation of a hands-free terminal on both sides of such a underwater telephony system, as illustrated in Figure 1, is achieved by implementing key signal enhancements to mostly eliminate acoustic echoes and other distractions from the speech signals.

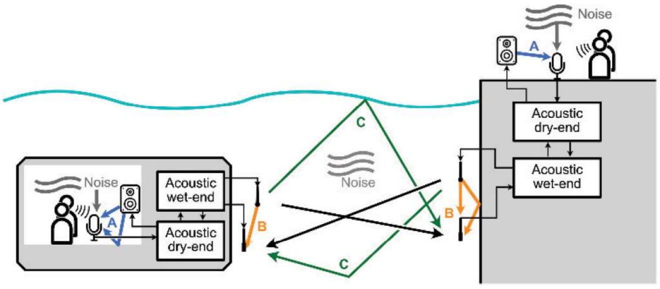

Figure 1: Illustration of the context of this work, illustrating the three types of acoustic echoes present. Here we deal with acoustic dry-end echoes (A).

Notation

We denote time-domain signals with lowercase and frequency-domain quantities with uppercase letters. The variable n indexes samples, k processing frames, μ frequency subbands and m partitions of a frequency response which spans multiple frames.

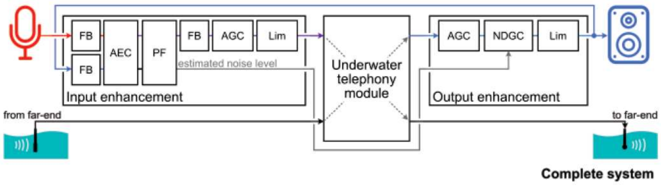

Figure 2: Algorithmic structure overview for the whole system with Filterbanks, Acoustic Echo Cancellation, Postfilter, Automatic Gain Control, Limiter and Noise Dependent Gain Control.

Algorithmic structure

The purpose of the developed sytem is to enhance on one hand the local microphone signal $y(n)$ before transmitting it underwater, and on the other hand the received far-end signal $x(n)$ before playing back locally. As depicted in Figure 2, the system is thus split into *Input Enhancement* and *Output Enhancement* modules. The first module performs Acoustic Echo Cancellation and Postfiltering in the frequency domain and additionally Automatic Gain Control and limiting. The second block also features the last two steps, with an additional Noise Dependent Gain Control unit.

For the echo cancellation, the signal $x(n)$ *exciting* the echo needs to be known as well, and both are transformed into their frequency-domain representations $X(\mu,k), Y(\mu,k)$ via overlap-add *Filterbanks*. A synthesis Filterbank is then used on the enhanced signal before feeding it into the AGC, and an estimate of the local noise level is in turn forwarded to output enhancement.

Acoustic Echo Cancellation (AEC)

Acoustic echoes are by far the most irritating disturbance in hands-free terminals. A well established way of reducing them is also chosen here and relies on modeling the disturbed signal Y as result of an Loudspeaker-Enclosure-Microphone system (see Figure 3) as

$$Y = X * H + N,$$

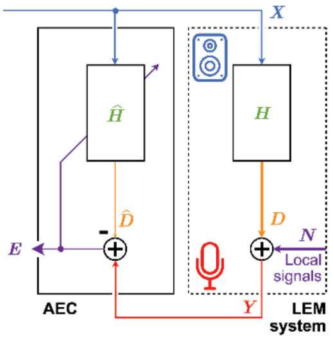

Figure 3: Loudspeaker-Enclosure-Microphone system.

where H denotes the *frequency response* between speaker and microphone, N the *local signals* and $*$ point-wise multiplication. To obtain an estimate of the undisturbed local signals, the frequency response is estimated as $\hat{H}_m(\mu, k)$ via an *adaptive filter* and the estimated echo

$$\hat{D}(\mu, k) = \sum_m X(\mu, k - m) * \hat{H}_m(\mu, k)$$

is subtracted from the microphone signal to yield our first enhancement, also called the *error signal*

$$E(\mu, k) = Y(\mu, k) - X(\mu, k) * \hat{H}(\mu, k).$$

Echo estimation

For the adaptive filter, the *Normalized Least Mean Square (NLMS)* algorithm is chosen due to its robustness and modest complexity. In each processing frame k, it produces an update ΔH to be added to $\hat{H}$, which can be given as

$$\Delta H_m(\mu, k) = \mu_F(\mu, k) \cdot \frac{E^*(\mu, k) X(\mu, k - m)}{\sum_{l=0}^{K-1} |X(\mu, k - l)|^2}$$

with the complex conjugated error signal E^* [1]. The factor $\mu_F(\mu, k)$ is called the *stepsize* parameter and is determined by control mechanisms which are crucial for using the NLMS for Acoustic Echo Cancellation.

Stepsize calculation

The quickest adaptation of the NLMS filter is achieved by a stepsize of $\mu_F = 1$, in which case the estimated echo $\hat{D}$ would always try to converge to the full microphone signal Y. This is however not always the best choice for our purposes, as it would try to adapt to local speech signals as well. Thus the stepsize needs also to be adapted, to which end we follow a formulation of the optimal stepsize for LEM systems with local signals [2, p 80] as

$$\mu_{\mathrm{opt}}(\mu, k) \approx \frac{\mathbb{E}[|E_u(\mu, k)|^2]}{\mathbb{E}[|E(\mu, k)|^2]},$$

where E_u stands for the *undisturbed error*, defined as the distance between estimated and true echo

$$E_u(\mu, k) = D(\mu, k) - \hat{D}(\mu, k).$$

As the optimal stepsize term only requires the power spectral density (PSD) of E_u, we can approximate it as $\hat{E}_u$ based on coupling factors with known quantities and an additional voice activity detection term V_x, leading to the following stepsize term:

$$\mu_F(\mu, k) = V_x(\mu, k) \cdot \frac{\hat{E}_u^2(\mu, k)}{\overline{E}^2(\mu, k)} \text{ limited to } [0,1]$$

Undisturbed error

In this work we estimate the undisturbed error PSD through its estimated coupling to both the smoothed far-end signal $\overline{X}^2$ and the smoothed estimated echo $\overline{\hat{D}}^2$, expressed as time and frequency selective quantities $\beta_{\mathrm{xe}}(\mu, k)$ and $\beta_{\mathrm{de}}(\mu, k)$ respectively (see also figure 4). This leads to the following approximation:

$$\hat{E}_u^2(\mu, k) = \max\left[\overline{X}^2(\mu, k - \hat{k}_{\mathrm{dly}})\beta_{\mathrm{xe}}(\mu, k) \, , \overline{\hat{D}}^2(\mu, k)\beta_{\mathrm{de}}\right]$$

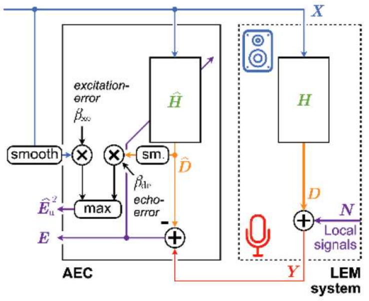

Figure 4: Estimation of the undisturbed error with coupling factors.

Here, the far-end signal is also delayed by $\hat{k}_{\mathrm{dly}}$ frames to account for the initial delay between excitement and microphone signals, stemming from technical latencies and the shortest acoustic path in the LEM system. This delay can be estimated by a fairly simple procedure from the estimated frequency response and avoids certain unstable behaviour especially during transitions of silence and speech.

Voice activity detection

Both coupling factors are updated only during remote single-talk, when the undisturbed error can be assumed to be equal to the error signal E. This can be performed per-subband, allowing the algorithm to adapt to frequency ranges containing only echoes at the same time as local speech is present in others.

Remote voice activity – or more specifically *any* far-end activity beyond background noise – can be detected as

$$\overline{X}(\mu, k) > \hat{B}_x(\mu, k) \cdot \mathrm{SNR}_{V_x}$$

with a threshold SNR_{V_e} and a remote background noise estimate $\hat{B}_x(\mu. k)$, which in turn is obtained for this and other purposes by multiplicative constants based methods as described in [3].

Local voice activity is not as clearly detectable, as it is only known as a mix with acoustic echoes and noise. The estimation used here uses the error signal E and a local noise estimate $\hat{B}_{e|y}(\mu, k)$ to obtain

$$\overline{E}(\mu, k) > \hat{B}_{e|y}(\mu, k) \cdot \mathrm{SNR}_{V_e}$$

with a higher signal to background noise ratio, to account for residual echoes in E.

Coupling factors

$\beta_{\mathrm{xe}}(\mu, k)$ and $\beta_{\mathrm{de}}(\mu, k)$ are updated by fixed increase or decrease factors in each frame, depending on whether the corresponding estimate is larger or smaller than E which we assume to momentarily resemble the true undisturbed error. A much smaller increase factor can in this case be beneficial to underestimate the coupling factors and thus keep the stepsize more stable during residual echoes.

Rescue mechanisms

The presented coupling factor estimation method relies on system distance estimations, which are at risk of getting stuck in long-lasting misadjustments [2, p 333]. For the mitigation of this problem, an *Enclosure Dislocation Detection* method is added, which can enforce a strong re-adaptation of the coupling factors and help the echo estimation in keeping up with sudden changes of the LEM system.

As illustrated in figure 5, this is realized by a second instance of the NLMS filter, which is run in parallel but with a very simple and much more aggressive control strategy, which sets the stepsize to $\mu_F = 1$ whenever remote voice activity is detected in a given subband. The filter operates only on a small subset of frequency bands to reduce the complexity and would be too divergent to be used for echo estimation. This divergence however allows it to recover from room changes quickly, and when the shadow filter error E_{sh} is consistently smaller than the regular error E for some amount of time, the coupling factors can be reset, leading to a strong readaptation of the main filter. The opposite case of $E(\mu, k) < E_{\mathrm{sh}}(\mu, k)$ happens very frequently as the main filter is designed to adapt more robustly, and leads to a reset of the shadow filter coefficients to the correponding values of the main filter.

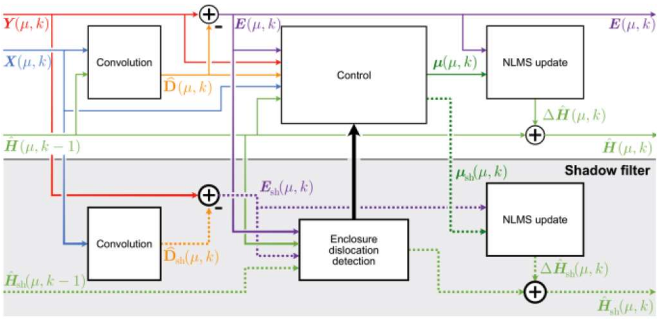
$Y(\mu, k)$
$X(\mu, k)$
Convolution
$\hat{D}(\mu, k)$
$E(\mu, k)$
Control
$\mu(\mu, k)$
NLMS update
$E(\mu, k)$
$\Delta\hat{H}(\mu, k)$
$\hat{H}(\mu, k-1)$
$\hat{H}(\mu, k)$
Shadow filter
$\mu_{\mathrm{sh}}(\mu, k)$
$E_{\mathrm{sh}}(\mu, k)$
Convolution
$\hat{D}_{\mathrm{sh}}(\mu, k)$
Enclosure dislocation detection
NLMS update
$\Delta\hat{H}_{\mathrm{sh}}(\mu, k)$
$\hat{H}_{\mathrm{sh}}(\mu, k-1)$
$\hat{H}_{\mathrm{sh}}(\mu, k)$

Figure 5: Schematic depiction of the *Enclosure Dislocation Detection* module. The coupling factor reset is illustrated by the bold arrow.

Apart from Enclosure Dislocation Detection, more direct constraints are imposed by strongly attenuating individual bands of the frequency response $\hat{H}_m(\mu, k)$ if either the error signal surpasses the microphone signal in subband μ, which should theoretically never be the case as echo subtraction is not meant to add something to the signal, of if the energy of a subband μ throughout all partitions m reaches a threshold itself.

Postfilter

As stated in [2, p 246], an AEC can only subtract the part of the echoes modeled by the frequency response $\hat{H}(\mu, k)$, whose length is limited as it impacts the convergence speed of the NLMS. This leaves *residual echo* components in place, which still can be fairly disturbing. Using an estimated PSD of those components, we can suppress them further, now introducing slight degradations of the speech quality, with a recursive Wiener Filter. This filter can also additionally deal with stationary background noise, whose PSD can also be estimated and is already used in the previous processing steps.

Residual echo estimation

The true residual echo D_{res} is per definition equal to the undisturbed error E_{u}, thus we can use the coupling factor based estimation as starting point. Compared to $\hat{E}_{\mathrm{u}}$, which is rather *under*estimated to avoid unstable behaviour of the filter, we add more signal components to $\hat{D}_{\mathrm{res}}$ to reduce the chance of the Wiener Filter "missing" disturbances in the signal:

The estimated echo is also used without smoothing as

$$\hat{D}_{\mathrm{res}}(\mu, k) \geq |\hat{D}(\mu, k)| \cdot \beta_{\mathrm{de}}(\mu. k)$$

to catch onsets of residual echoes faster. A exponential falloff of $\hat{D}_{\mathrm{res}}$ is used as lower bound to capture echo tails through

$$\hat{D}_{\mathrm{res}}(\mu, k) \geq \hat{D}_{\mathrm{res}}(\mu, k-1) \cdot \gamma_{\mathrm{res}}.$$

A smoothing along the frequency axis is added to avoid close-to-zero outliers in $\hat{D}_{\mathrm{res}}$, which are problematic as they introduce a large bias [2, p 360]. And finally, the resulting residual echo estimate is overestimated during remote single-talk or speech pauses to strike a balance between reducing disturbances and impairing the resulting signal's quality.

Wiener Filter

The Wiener Filter acts as an attenuation on the signal, giving

$$\tilde{E}(\mu, k) = G_{\mathrm{w}}(\mu, k) \cdot E(\mu, k)$$

where G_{w} incorporates noise $\hat{B}$ and echo $\hat{D}_{\mathrm{res}}$ estimations as

$$G_{\mathrm{w}}(\mu, k) = \max[G'_{\mathrm{w}}(\mu, k)\,, G_{\min}(\mu, k)],$$

$$G'_{\mathrm{w}}(\mu, k) = 1 - \frac{\beta_{\mathrm{noise}} \hat{B}_{\mathrm{e}}^2(\mu, k) + \hat{D}_{\mathrm{res}}^2(\mu, k)}{\max\left[G(\mu, k-1), G_{\min,\mathrm{r}}\right] |E(\mu, k)|^2}.$$

This term also includes the *spectral floor* $G_{\min}(\mu, k)$ which can be adapted based on voice activity, the fixed overestimation of the noise estimate β_{noise} and the recursive term.

Gain control

Apart from added disturbances, the volume of the signal can also be outside the optimal range, leading to either distortions or a low signal to noise ratio.

To compensate for this, first an Automatic Gain Control (AGC) unit is employed in both input and output enhancements, which operates in the time-domain, estimates the signal peak and boosts or attenuates the signal to reach a desired peak. A gain change is only made during voice activity, which in this case is detected whenever a fast envelope tracker surpasses a slow one by a threshold. By allowing for a faster decrease of the gain we allow the algorithm to quickly react to volume increases while skipping over speech pauses instead of falsely interpreting the noise as speech.

While the AGC reacts very quickly, a Limiter is added to catch any remaining peaks and serve as a definitive constraint on the signal amplitude.

During output enhancement, a Noise Dependent Gain Control (NDGC) unit is also used, which receives the estimated local background noise and maps its power to a certain amount of additional gain, to allow the listening level to exceed the noise enough to be well audible while not being unreasonably loud during its absence.

Evaluation

The presented algorithms were developed in C/C++ within the KiRaT framework developed at the chair of Digital Signal Processing and System Theory at CAU Kiel. They were tested subjectively in an office environment and outdoors with a short underwater path, with both conversations and recorded signals.

Instead of a study with human listeners, we used a neural network model developed by Purin et al. [4] to predict mean opinion score (MOS) values for echo and other impairments as defined by ITU-T P.831 and P.832 recommendations. Their model has a pearson correlation coefficient to true human ratings of up to 0.847, far outperforming conventional metrics like ERLE (echo return loss enhancement).

A similar approach was used to also judge the performance of the noise suppression in specific, with a neural network model called DNSMOS by Reddy et al. [5], which computes scores for the signal before and after enhancement, and allows their difference to be used as a metric with a PCC of up to 0.98 with human ratings.

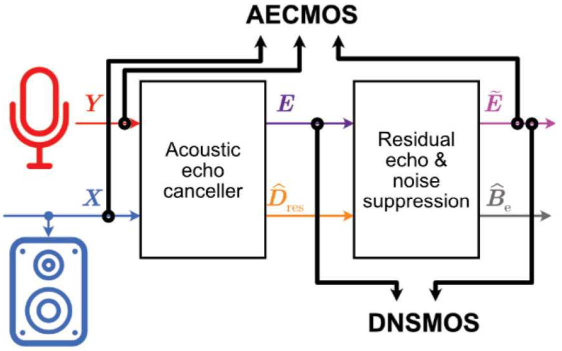

Figure 6: Both neural-network metrics illustrated by the signals they use.

Results

Subjectively the system performed very well, allowing unimpaired conversations in various environments and noise levels. The results from AECMOS were obtained on a set of self-recorded clips and compared with the distribution of results from the participants of the 2022 ICASSP Acoustic Echo Cancellation Challenge [6] and are shown in figure 7. Even though our approach is rather conventional, its results lie well within the competitors, except for near-end single talk, in which case the lack of more sophisticated noise suppression comes to play.

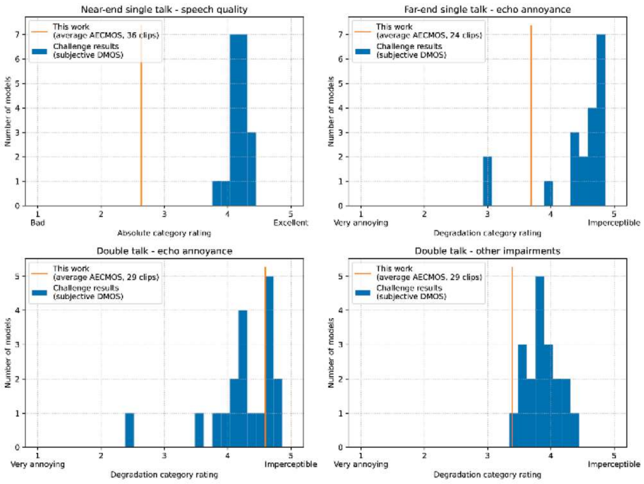

Figure 7: AECMOS results for this work, compared to distribution of results from [6].

The DNSMOS results are shown in figure 8 and also show a clear improvement in all categories, while not reaching the best scores due to, again, a lack of more advanced nosie suppression.

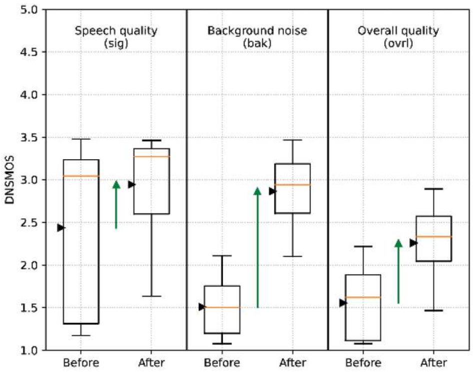

Figure 8: DNSMOS results for this work, the improvement of average values shown with the arrows.

Conclusion

A system for enhancing the dry-end situation of an underwater telephony setup with hands-free terminals was developed and evaluated. It runs in real-time and was shown to perform well in various scenarios, allowing distraction-less conversations with consistent listening levels and without disturbing acoustic echoes or stationary noise components. The good results were also confirmed with more objective metrics.

Literature

[1] Haykin, S.: Adaptive Filter Theory. Pearson Education India (2008), 334

[2] Hänsler, E., Schmidt, G.: Acoustic Echo and Noise Control: A Practical Approach. John Wiley & Sons (2005), 80-86

[3] Maschmann, T., Gimm, M., Kandade Rajan, V., Schmidt, G.: Implementaion of a New Method for Noise Suppression in Automotive Environments. Proc. DAGA, Kiel, 2017

[4] Purin, M., Sootla, S., Sponza, M., Saabas, A., Cutler, R.: A Speech Quality Assessment Metric for Echo Impairment. ICASSP 2022, IEEE International Conference on Acoustics, Speech and Signal Processing. 901-905

[5] Reddy, C., Gopal, V., Cutler, R.: DNSMOS: A Non-intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. ICASSP 2021, IEEE International Conference on Acoustics, Speech and Signal Processing. 6493-6497

[6] Cutler, R, Saabas, A., Parnamaa, T., Purin, M., Gamper, H., Braun, S., Sørensen, K., Aichner, R.: ICASSP 2022 acoustic echo cancellation challenge. ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. 9107-9111