COMBINED ACOUSTIC ECHO CONTROL AND NOISE REDUCTION FOR MOBILE COMMUNICATIONS

Stefan Gustafsson and *Rainer Martin*
Institute of Communication Systems and Data Processing
Aachen University of Technology
D-52056 Aachen, Germany
Tel: +49 241 806976; fax: +49 241 8888186
e-mail: gus@ind.rwth-aachen.de

ABSTRACT

In this paper an acoustic echo compensator with an additional frequency domain adaptive filter for combined residual echo and noise reduction is proposed. The algorithm delivers high echo attenuation as well as high near end speech quality over a wide range of signal-to-noise conditions. The system makes use of a standard time domain echo compensator of low order, after which the proposed adaptive filter, which is motivated by means of a minimum mean square error approach, is placed in the sending path. In contrast to other combined systems [1, 2, 3], our method uses an explicit estimate of the power spectral density of the residual echo after echo compensation. The separate estimations of the power spectral densities of the residual echo and the background noise, respectively, are then flexibly combined, such that in the processed signal a low level of intentionally left background noise will effectively mask the residual echo.

1 INTRODUCTION

A basic block diagram of our single-microphone system is shown in Fig. 1. It consists of the time domain echo compensator C and the additional adaptive filter, denoted by H, in the sending path. $x(k)$ denotes the signal from the far end speaker. The microphone signal $y(k)$ consists of the near end speech $s(k)$, the near end noise $n(k)$, and the echo $d(k)$. The estimated echo $\widehat{d}(k)$ is subtracted from $y(k)$ yielding the echo compensated signal $e(k)$. This can be written as $e(k) = s(k) + n(k) + b(k)$, where $b(k) = d(k) - \widehat{d}(k)$ is the residual echo.

In a car environment, we typically choose the order of the echo compensator to $N_C = 200$. The short compensator has, besides the lower implementation costs, some distinct advantages compared to a longer one: it will converge faster and is also more robust against noise. However, as the room impulse response in a medium size car usually has about 500 coefficients of substantial energy at a sampling frequency of 8000 Hz, it is obvious that the compensator will not be able to remove the echo $d(k)$ completely. One of the tasks of the filter H is to attenuate the residual echo $b(k)$. A time domain filter with this purpose has been examined in [4, 5, 6].

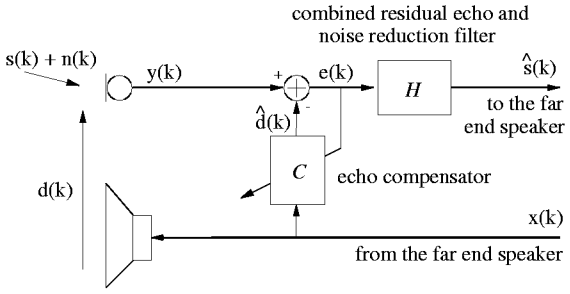

Figure 1: Block diagram of an echo compensator C with an additional adaptive filter H in the sending path.

The second task is to reduce the level of background noise $n(k)$, which is present in the microphone signal. This noise can be of very different characters. A large portion of the energy is, however, typically concentrated at the lower frequencies and the noise is fairly stationary compared to speech. The last property makes it possible to distinguish between speech and noise and is important to all single microphone speech enhancement algorithms.

It is seldom necessary, or even desirable, to completely remove the noise from the microphone signal. Most often, the atmosphere given by a natural sounding residual noise is prefered by the far end speaker. An even more important motive to preserve some level of background noise is that an attempt to a complete removal often leads to a very uncomfortable residual noise in form of "musical tones" and to severe distortions of the near end speech.

The residual noise as well as the near end speech will also, to some extent, mask the residual echo left by the echo compensator. To achieve the above goals the filter H therefore should balance the extent of noise reduction and residual echo suppression, such that a low amount of natural sounding background noise, but no residual echo, can be heard in the output signal $\widehat{s}(k)$. Any algorithm with this purpose needs some information about the noise and the residual echo. This will be discussed in the next section.

2 POWER SPECTRAL DENSITY ESTIMATION

For the combined reduction of residual echo and noise separate estimations of the power spectral densities (psd) of the background noise and the residual echo have to be performed. The noise psd – here denoted by $R_{nn}(\Omega_i)$, where $\Omega_i = \frac{i}{M}2\pi$, $i \in \{0, 1, 2, \ldots, M-1\}$ are the discrete frequencies – can be estimated by the "Minimum Statistics" and "Spectral Minima Tracking" methods outlined in [7, 8]. These methods have the advantage that the noise psd is estimated continuously, eliminating the need of a voice activity detector. They also allow some instationarity in the noise to be detected, which is vital for the noise reduction algorithm to perform well if $R_{nn}(\Omega_i)$ is changing slowly.

For the estimation of the residual echo psd $R_{bb}(\Omega_i)$ a method where the echo compensation is described by a transfer function $F(\Omega_i)$ of a possibly noncausal system is useful [9]. This is illustrated in Figure 2, and leads in the frequency domain to the identities

$$B(\Omega_i) = D(\Omega_i) - \widehat{D}(\Omega_i) \quad (1)$$

$$B(\Omega_i) = F(\Omega_i)D(\Omega_i). \quad (2)$$

The Eqs. (1) and (2) can be combined to yield

$$F(\Omega_i) = 1 - \frac{\widehat{D}(\Omega_i)}{D(\Omega_i)}. \quad (3)$$

The time domain echo compensation is performed by amplitude and phase. In general, the phase of $\widehat{D}(\Omega_i)$ is a good estimate of the phase of $D(\Omega_i)$. It has been verified by simulations that this statement holds whenever the magnitude of $\widehat{D}(\Omega_i)$ is a good estimate of the magnitude of $D(\Omega_i)$. With this knowledge we can make the assumption $\arg\{F(\Omega_i)\} = \arg\{\widehat{D}(\Omega_i)\} - \arg\{D(\Omega_i)\} \approx 0$, i.e. $F(\Omega_i)$ is a real valued function.

By combining the Eqs. (1) and (2), the psds of the echo and the residual echo, respectively, can be written as a function of the transfer function $F(\Omega_i)$ and the psd of the estimated echo, $R_{\widehat{dd}}(\Omega_i)$,

$$R_{dd}(\Omega_i) = \frac{1}{(1 - F(\Omega_i))^2} R_{\widehat{dd}}(\Omega_i) \quad (4)$$

$$R_{bb}(\Omega_i) = \left(\frac{F(\Omega_i)}{1 - F(\Omega_i)}\right)^2 R_{\widehat{dd}}(\Omega_i)\,. \quad (5)$$

Thereby, the problem of estimating $R_{bb}(\Omega_i)$ has changed into the estimation of the transfer function $F(\Omega_i)$.

Assuming statistical independence between the near end speech $s(k)$, the noise $n(k)$, and the echo $d(k)$ respectively the residual echo $b(k)$, we can write the power spectral densities of the microphone signal $y(k)$ and the compensated signal $e(k)$ as

$$\begin{aligned} R_{yy}(\Omega_i) &= R_{ss}(\Omega_i) + R_{nn}(\Omega_i) + R_{dd}(\Omega_i) \\ R_{ee}(\Omega_i) &= R_{ss}(\Omega_i) + R_{nn}(\Omega_i) + R_{bb}(\Omega_i)\,. \end{aligned} \quad (6)$$

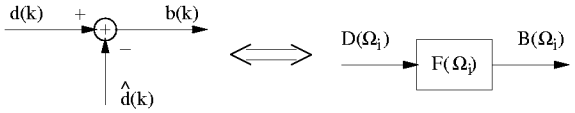

Figure 2: Interpretation of the echo compensation as a transfer function $F(\Omega_i)$

Combining the above equations with Eqs. (4) and (5) we arrive at an expression for $F(\Omega_i)$, which can now be calculated from known signals,

$$F(\Omega_i) = \frac{R_{yy}(\Omega_i) - R_{ee}(\Omega_i) - R_{\widehat{dd}}(\Omega_i)}{R_{yy}(\Omega_i) - R_{ee}(\Omega_i) + R_{\widehat{dd}}(\Omega_i)}\,. \quad (7)$$

Having an estimation of $F(\Omega_i)$, the power spectral density $R_{bb}(\Omega_i)$ can be estimated using Eq. (5).

3 SPECTRAL WEIGHTING RULES

For the purpose of noise reduction several weighting rules $H_n(\Omega_i)$, which modify only the spectral amplitudes of the input signal, leaving the phase unchanged, have been developed. Among them are the familiar Minimum Mean Square Error (MMSE) Wiener filter, newer methods such as the Minimum Mean Square Error Short-Time Spectral Amplitude estimator (MMSE-STSA) [10] and its derivative, the Logarithmic Spectral Amplitude estimator (MMSE-LSA) [11]. The choice of a weighting rule depends strongly on the objectives of the system. With the Wiener filter and modifications thereof [12] a high noise reduction and a relatively low distortion of the near end speech can be achieved. This, however, requires that the noise psd can be estimated very well, as otherwise the processing will lead to disturbing musical tones in the output signal. The MMSE-STSA and -LSA weighting rules are more robust to estimation errors, but will on the other hand in general not give that high a noise reduction.

3.1 Weighting Rules as a Function of SNR

A practical way of describing some common weighting rules is as functions of the *a priori* and *a posteriori* signal-to-noise ratios [13]. In our context the frequency dependent a priori SNR is defined as

$$SNR_n^s(\Omega_i) = \frac{\mathrm{E}\{|S(\Omega_i)|^2\}}{\mathrm{E}\{|N(\Omega_i)|^2\}} \quad (8)$$

and the a posteriori SNR as

$$SNR_n^e(\Omega_i) = \frac{|E(\Omega_i)|^2}{\mathrm{E}\{|N(\Omega_i)|^2\}}, \quad (9)$$

where $\mathrm{E}\{\cdot\}$ denotes the expectation operator. The Wiener weighting rule for noise supression can then be written as

$$H_n(\Omega_i) = \frac{SNR_n^s(\Omega_i)}{SNR_n^s(\Omega_i) + 1}. \quad (10)$$

To attenuate the residual echo, $N(\Omega_i)$ in Eqs. (8) and (9) can be substituted by the Fourier transform of the residual echo, $B(\Omega_i)$. We get the two corresponding a priori and a posteriori SNR expressions refering to the residual echo,

$$SNR_b^s(\Omega_i) = \frac{\mathrm{E}\{|S(\Omega_i)|^2\}}{\mathrm{E}\{|B(\Omega_i)|^2\}} \tag{11}$$

$$SNR_b^e(\Omega_i) = \frac{|E(\Omega_i)|^2}{\mathrm{E}\{|B(\Omega_i)|^2\}}. \tag{12}$$

The a posteriori SNRs refering to the noise or the residual echo are calculated using instantaneous spectral components of $E(\Omega_i)$ and estimates of the psds $R_{nn}(\Omega_i)$ and $R_{bb}(\Omega_i)$, respectively. The a priori SNRs are commonly estimated by a "decision directed" approach [10]. With m as the frame index, $SNR_n^s(\Omega_i)$ is recursively estimated by

$$\begin{aligned} &SNR_n^{s\,(m)}(\Omega_i) = \\ &= (1-\alpha_n)P(SNR_n^{e\,(m)}(\Omega_i) - 1) + \\ &\quad + \alpha_n \frac{|H^{(m-1)}(\Omega_i)E^{(m-1)}(\Omega_i)|^2}{R_{nn}^{(m)}(\Omega_i)}, \end{aligned} \tag{13}$$

where $P(x) = \frac{1}{2}(|x| + x)$. In this estimation the smoothing constant α_n is a decisive factor. The choice of this parameter depends strongly on the characteristics of the signal component to be removed. For the purpose of noise reduction, experiences have shown that α_n should be chosen to $\alpha_n = 0.97 \ldots 0.99$, depending on such factors as sampling frequency, FFT-length, overlap-length etc. This will lead to a satisfying level of noise reduction without attenuating near end speech transients too much.

As the residual echo is a speech-like signal with characteristics different from those of ambient noise, the parameter α_b for estimating $SNR_b^s(\Omega_i)$ must be optimized anew. Here $\alpha_b \approx 0.90$ has been found to lead to a good compromise between near end speech quality and residual noise attenuation [9].

3.2 Combined Reduction of Residual Echo and Noise

For the combined reduction of residual echo and noise we notice that $b(k)$ and $n(k)$ are statistically independent and define the a priori SNR and the a posteriori SNR with respect to both components,

$$SNR_{b+n}^s(\Omega_i) = \frac{\mathrm{E}\{|S(\Omega_i)|^2\}}{\mathrm{E}\{|B(\Omega_i)|^2\} + \mathrm{E}\{|N(\Omega_i)|^2\}} \tag{14}$$

$$SNR_{b+n}^e(\Omega_i) = \frac{|E(\Omega_i)|^2}{\mathrm{E}\{|B(\Omega_i)|^2\} + \mathrm{E}\{|N(\Omega_i)|^2\}}. \tag{15}$$

These equations can be rewritten as functions of the previously defined SNRs,

$$SNR_{b+n}^s(\Omega_i) = \frac{1}{(SNR_b^s(\Omega_i))^{-1} + (SNR_n^s(\Omega_i))^{-1}} \tag{16}$$

$$SNR_{b+n}^e(\Omega_i) = \frac{1}{(SNR_b^e(\Omega_i))^{-1} + (SNR_n^e(\Omega_i))^{-1}}. \tag{17}$$

Eqs. (16) and (17) are used as parameters for the chosen weighting rule. They give us a powerful and flexible way of treating the residual echo and noise reduction. This will now be discussed in detail.

3.3 Limiting of Estimated SNR

We will first look at the noise reduction case. In the first section of this paper, we mentioned that it is often desirable to leave a low level of natural sounding residual noise in the processed signal. This can be achieved by limiting the estimated SNR to a minimum threshold T_n,

$$\begin{aligned} SNR_n^s(\Omega_i) &:= \max(SNR_n^s(\Omega_i), T_n) \\ SNR_n^e(\Omega_i) &:= \max(SNR_n^e(\Omega_i), T_n), \end{aligned} \tag{18}$$

Especially at low SNR the limiting will have a significant effect – thus the stronger the background noise level is, the higher the noise level in the processed signal will be. This is in fact an advantage, as for speech enhancement in general performs less well in a low SNR-environment. The limiting prevents too high an attenuation and therefore it also reduces the distortions of the near end speech, which otherwise might lose intelligibility. A proper range is $T_n = 0.01 \ldots 0.1$, where the chosen value eventually depends on criterias such as desired noise reduction and admissible speech distortion. If the threshold T_n is chosen too high, the amount of noise reduction will be very low; if it is chosen too low, the limiting will have almost no influence.

Now consider an equivalent limiting of the SNR refering to the residual echo,

$$\begin{aligned} SNR_b^s(\Omega_i) &:= \max(SNR_b^s(\Omega_i), T_b) \\ SNR_b^e(\Omega_i) &:= \max(SNR_b^e(\Omega_i), T_b). \end{aligned} \tag{19}$$

This will have the effect that some residual echo always will be left in the signal $\widehat{s}(k)$. Of course, this is not desirable in a noise-free situation, as the echo then might be audible. However, when noise is present, some limiting is necessary as otherwise the attenuation by the filter H might be too high. This would lead to disturbing modulations of the residual noise.

The idea at this point is to attenuate the residual echo $b(k)$ where it is not masked by the residual noise. In the processed signal $\widehat{s}(k)$ only the near end speech and an attenuated, natural sounding background noise should be audible, but no echo. This can be achieved by a frequency dependent threshold T_b, which is a

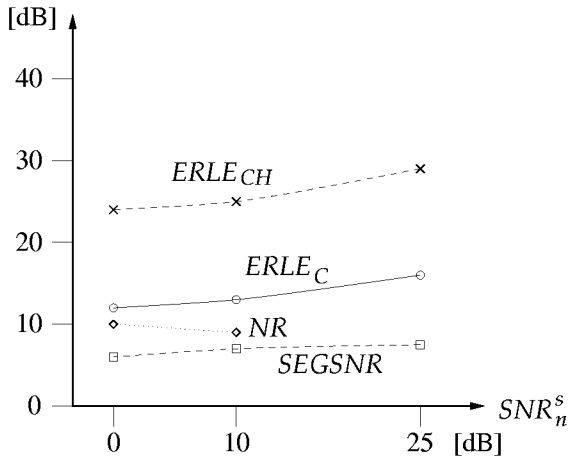

Figure 3: Simulation results for the double talk situation.

function of the chosen threshold T_n, and of the psds of the residual echo and the noise, for example

$$T_b(\Omega_i) = \frac{2T_n}{1 + \frac{R_{bb}(\Omega_i)}{R_{nn}(\Omega_i)}}. \qquad (20)$$

With the above adaptive limiting and a subsequent combination of the different a priori and a posteriori SNRs, the speech enhancement algorithm will work well over a wide range of input signal-to-noise conditions, effectively reducing the background noise and the residual echo with only minor impacts on the near end speech quality.

4 RESULTS AND CONCLUSIONS

Results from simulations show a very significant reduction of noise and residual echo for a wide range of signal-to-noise conditions. In Figure 3 the Echo Return Loss Enhancement for the echo compensator C ($ERLE_C$), for the combined system $C+H$ ($ERLE_{CH}$), the noise reduction (NR) and the *segmental SNR* as a measure of the near end speech distortion ($SEGSNR$) [14] are plotted as a function of the signal to noise ratio at the microphone. It can be seen that the higher the SNR is, the more the echo will be attenuated, whereas the noise reduction decreases. In the almost noise free case ($SNR \approx 25$ dB), where the residual echo may be masked by the near end speech only, 30 dB echo attenuation is achieved. In the single talk situation this figure will rise to 50 dB. Thus, the proposed system presents a good tradeoff between echo and noise reduction, near end speech distorsion, and computational complexity.

5 REFERENCES

[1] R. Martin and P. Vary, "Combined Acoustic Echo Cancellation, Dereverberation, and Noise Reduction: A Two Microphone Approach", Annales des Télécommunications, Vol. 49, No. 7-8, pp. 429-438, 1994.

[2] G. Faucon and R. Le Bouquin Jeannes, "Joint System for Acoustic Echo Cancellation and Noise Reduction", Proc. EUROSPEECH '95, Madrid, pp. 1525-1528, 18-21 September, 1995.

[3] J. Boudy, F. Capman, and P. Lockwood, "A Globally Optimised Frequency Domain Acoustic Echo Canceller for Adverse Environment Applications", Proc. Fourth Int. Workshop on Acoustic Echo and Noise Control, pp. 95-98, Røros, Norway, June 1995.

[4] R. Martin, "Combined Acoustic Echo Cancellation, Spectral Echo Shaping, and Noise Reduction", Proc. Fourth Int. Workshop on Acoustic Echo and Noise Control, pp. 48-51, Røros, Norway, June 1995.

[5] R. Martin and S. Gustafsson, "An Improved Echo Shaping Algorithm for Acoustic Echo Control", Proc. EUSIPCO-96, Trieste, September 1996.

[6] R. Martin and S. Gustafsson, "The Echo Shaping Approach to Acoustic Echo Control", Speech Communication, Vol. 20, No. 3-4, December 1996.

[7] R. Martin, "Spectral Subtraction Based on Minimum Statistics", Proc. EUSIPCO-94, Edinburgh, pp. 1182-1185, September 12-16, 1994.

[8] G. Doblinger, "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands", Proc. EUROSPEECH'95, pp. 1513-1516, Madrid, September 1995.

[9] S. Gustafsson, "Combined Frequency Domain Acoustic Echo Attenuation and Noise Reduction", Proc. 9th Aachener Kolloquium, Aachen, Germany, March 1997.

[10] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. Acoustics, Speech, Signal Processing, Vol. 32, No. 6, pp. 1109-1121, December 1984.

[11] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator", IEEE Trans. Acoustics, Speech, Signal Processing, Vol. 33, No. 2, pp. 443-445, April 1985.

[12] V. Turbin, A. Gilloire, and P.Scalart, "Comparision of Three Post-Filtering Algorithms For Residual Acoustic Echo Reduction", Proc. Int. Conf. Acoustics, Speech, Signal Processing '97, pp. 307-310, München, Germany, April 1997.

[13] P. Scalart and J. Vieira Filho, "Speech Enhancement Based on a Priori Signal-to-Noise Estimation", Proc. Int. Conf. Acoustics, Speech, Signal Processing '96, pp. 629-632, May 7-10, Atlanta, 1996.

[14] S. Gustafsson, R. Martin, and P. Vary, "On the Optimization of Speech Enhancement Systems Using Instrumental Measures", Proc. Workshop on Quality Assessment in Speech, Audio and Image Communication, Darmstadt, Germany, March 11-13, 1996.