COMBINED ACOUSTIC ECHO CONTROL AND NOISE REDUCTION BASED ON RESIDUAL ECHO ESTIMATION

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 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 of 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 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 the adaption 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 (sampling frequency 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].

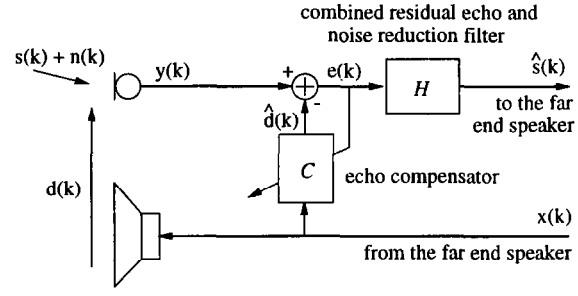

Figure 1: Block diagram of an echo compensator C with an additional adaptive filter H in the sending path.

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 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 fundamental 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.

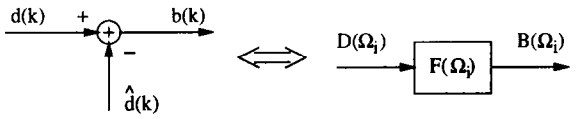

Figure 2: Interpretation of the echo compensation as a transfer function $F(\Omega_i)$.

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 discrete frequencies – can be estimated by the "Minimum Statistics" and "Spectral Minima Tracking" methods outlined in [6, 7]. 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.

As the residual echo is only a function of the echo itself and the estimated echo, a model where the echo compensation is described by a transfer function $F(\Omega_i)$ of a possibly noncausal system is useful [8]. This is illustrated in Figure 2. It leads to the identities

$$b(k) = d(k) - \widehat{d}(k) \tag{1}$$

$$b(k) = f * d(k), \tag{2}$$

and in the frequency domain

$$B(\Omega_i) = D(\Omega_i) - \widehat{D}(\Omega_i) \tag{3}$$

$$B(\Omega_i) = F(\Omega_i)D(\Omega_i). \tag{4}$$

The time domain echo compensation is perfomed by amplitude and phase. That is, the phase of $\widehat{D}(\Omega_i)$ is a good estimation 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 $D(\Omega_i)$. A sample frame of $|D(\Omega_i)|$, of the magnitude error $|D(\Omega_i)| - |\widehat{D}(\Omega_i)|$, and of the phase error $\arg\{D(\Omega_i)\} - \arg\{\widehat{D}(\Omega_i)\}$ are shown in Figure 3. It can clearly be seen that at frequencies where the echo is strong, the phase error is very close to zero. A large phase error can only be found at frequencies where $|D(\Omega_i)|$ is very small or zero.

With this knowledge we can make the assumption

$$\arg\{\widehat{D}(\Omega_i)\} \approx \arg\{D(\Omega_i)\}, \tag{5}$$

from which follows $\arg\{F(\Omega_i)\} \approx 0$, i.e. $F(\Omega_i)$ can be approximated by a real valued function.

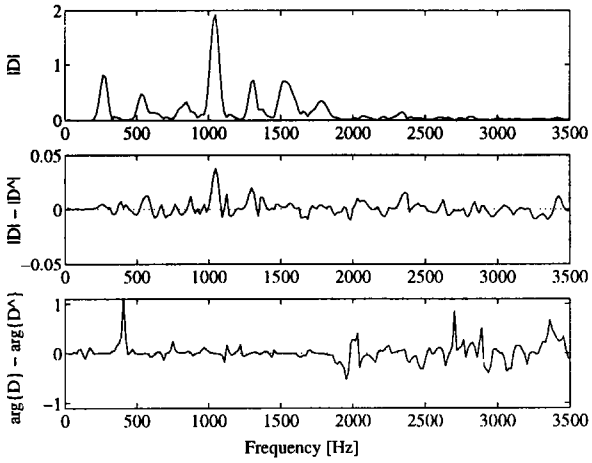

Figure 3: Time domain echo compensation: the magnitude of the echo, the magnitude error, and the phase error (in radians), for a sample speech frame.

By combining the Eqs. (3) and (4), the psd of the echo, $R_{dd}(\Omega_i)$, and psd of the the residual echo, $R_{bb}(\Omega_i)$, can be written as functions of the transfer function $F(\Omega_i)$ and the psd of the estimated echo, $R_{\widehat{d}\widehat{d}}(\Omega_i)$,

$$R_{dd}(\Omega_i) = \frac{1}{(1 - F(\Omega_i))^2} R_{\widehat{d}\widehat{d}}(\Omega_i) \tag{6}$$

$$R_{bb}(\Omega_i) = \left(\frac{F(\Omega_i)}{1 - F(\Omega_i)}\right)^2 R_{\widehat{d}\widehat{d}}(\Omega_i)\,. \tag{7}$$

The problem of estimating $R_{bb}(\Omega_i)$ then changes into the estimation of the transfer function $F(\Omega_i)$.

If no near end speech and no near end noise is present, i.e. a noise free single talk situation where $y(k) = d(k)$ and $e(k) = b(k)$, $F(\Omega_i)$ can be calculated from Eq. (3),

$$F(\Omega_i) = \frac{B(\Omega_i)}{D(\Omega_i)}. \tag{8}$$

However, as this situation seldom prevails, another solution must be found.

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

$$\begin{aligned} R_{yy}(\Omega_i) &= R_{ss}(\Omega_i) + R_{nn}(\Omega_i) + R_{dd}(\Omega_i) \\ R_{ee}(\Omega_i) &= R_{ss}(\Omega_i) + R_{nn}(\Omega_i) + R_{bb}(\Omega_i)\,. \end{aligned} \tag{9}$$

Combining the above equations with Eqs. (6) and (7) we arrive at an expression for estimating $F(\Omega_i)$, which can now be calculated from known signals,

$$F(\Omega_i) = \frac{R_{yy}(\Omega_i) - R_{ee}(\Omega_i) - R_{\widehat{d}\widehat{d}}(\Omega_i)}{R_{yy}(\Omega_i) - R_{ee}(\Omega_i) + R_{\widehat{d}\widehat{d}}(\Omega_i)}\,. \tag{10}$$

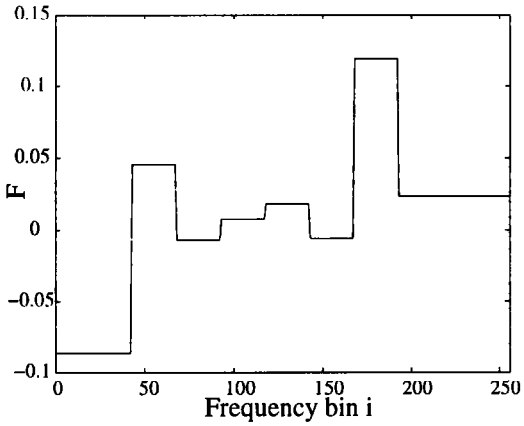

Figure 4: A sample of the estimated transfer function $\tilde{F}(\Omega_i)$.

Eq. (10) is only valid when $R_{\widehat{dd}}(\Omega_i) \neq 0$. Under some circumstances, for example when the estimated echo psd is very weak compared to the microphone signal psd, Eq. (10) can, owing to psd estimation errors and finite numerical accuracy, lead to wrong results. Therefore, potential errors must be excluded from the calculation. In our algorithm this is done in five steps:

1. Limit $F(\Omega_i)$ to some reasonable range $[F_{min}, F_{max}]$.
2. Consider only $F(\Omega_i)$ at frequencies where $R_{\widehat{dd}}(\Omega_i)$ is not too small.
3. Split the frequency range in N subbands.
4. In each subband calculate the mean value $\bar{F}_m$ of those $F(\Omega_i)$ which satisfies the condition in step 2.
5. At each frequency Ω_i, set $\tilde{F}(\Omega_i)$ to the corresponding mean value $\bar{F}_m$.

The transfer function $\tilde{F}(\Omega_i)$ estimated this way will then be used for the estimation of $R_{bb}(\Omega_i)$ using Eq. (7). It will possess a step-shape as illustrated in Figure 4.

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) [9] and its derivative, the Logarithmic Spectral Amplitude estimator (MMSE-LSA) [10].

3.1 Weighting Rules as 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 [11]. In our context the frequency dependent a priori SNR is defined as

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

and the a posteriori SNR as

$$SNR_n^e(\Omega_i) = \frac{|E(\Omega_i)|^2}{\mathrm{E}\{|N(\Omega_i)|^2\}}, \tag{12}$$

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

$$H_n(\Omega_i) = \frac{SNR_n^s(\Omega_i)}{SNR_n^s(\Omega_i) + 1}. \tag{13}$$

To attenuate the residual echo, $N(\Omega_i)$ in Eqs. (11) and (12) can be substituted by the Fourier transform of the residual echo, $B(\Omega_i)$. We then get the two corresponding a priori and a posteriori SNR expressions refering to the residual echo.

The a posteriori SNRs refering to the noise or the residual echo are calculated using instantaneous spectral components of $E(\Omega_i)$ and estimations 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 [9]. In this estimation the smoothing constant α 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 (the index n denotes that this constant belongs to the estimation of $SNR_n^s(\Omega_i)$) 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 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 echo attenuation [8].

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. For the a priori SNR this is

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

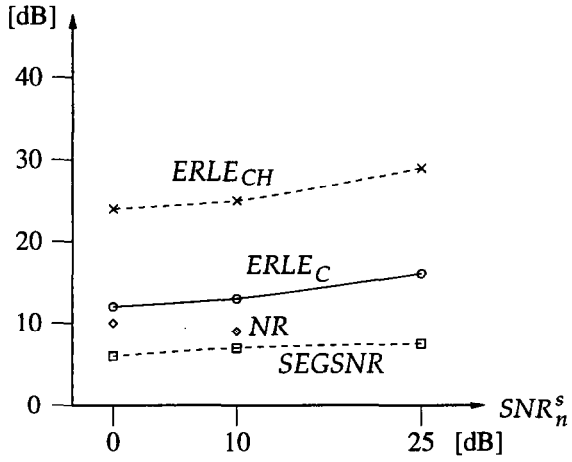

Figure 5: Simulation results for the double talk situation.

This equation can be rewritten as functions of the previously defined SNRs refering to either $b(k)$ or $n(k)$,

$$SNR^s_{b+n}(\Omega_i) = \frac{1}{(SNR^s_b(\Omega_i))^{-1} + (SNR^s_n(\Omega_i))^{-1}}. \quad (15)$$

The a posteriori SNR is calculated analogous. $SNR^s_{b+n}(\Omega_i)$ and $SNR^e_{b+n}(\Omega_i)$ are now used as parameters for the chosen weighting rule. This kind of combination gives us a powerful and flexible way of treating the residual echo and noise reduction [12]. For example, the SNRs refering to the noise can be downward limited to some value to retain a low level of natural sounding background noise in the processed signal.

4 RESULTS

Results from simulations show a very significant reduction of noise and residual echo for a wide range of signal-to-noise conditions. In Figure 5 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 ($SEGSNR$) 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. The robust estimation of the residual echo psd presented in Section 2 and the flexible combination of noise and residual echo reduction as outlined in Section 3 are both vital components of the algorithm.

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, "The Echo Shaping Approach to Acoustic Echo Control", Speech Communication, Vol. 20, No. 3-4, December 1996.

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

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

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

[9] 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.

[10] 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.

[11] 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.

[12] S. Gustafsson and R. Martin, "Combined Acoustic Echo Control and Noise Reduction for Mobile Communications", To be presented at EUROSPEECH '97, Rhodes, Greece, September 1997.