Combined Frequency Domain Acoustic Echo Attenuation and Noise Reduction

Stefan Gustafsson
Institut für Nachrichtengeräte und Datenverarbeitung
Aachen University of Technology
52056 Aachen, Germany
Tel: +49 241 806976; fax: +49 241 8888186
e-mail: gus@ind.rwth-aachen.de

ABSTRACT

In the hands free telephone environment the reduction of the acoustic echo of the far speaker is a major problem. Recent developments of algorithms use a conventional echo compensator and in addition a time domain filter in the sending path to attenuate the residual echo [1, 2, 3].

A further problem is the reduction of the background noise in the microphone signal. As the signal-to-noise ratio (SNR) in the hands free telephone environment can be very low, the reduction of noise is hard to perform.

Most noise reduction methods of today work in the frequency domain. An interesting approach is the combination of the echo attenuation and the noise reduction in a single frequency domain filter. In this paper a structure consisting of a conventional echo canceller and a combined echo attenuation and noise reduction frequency domain filter is proposed.

1 Introduction

We assume that an echo canceller working either in the time domain or in the frequency domain is available. In most practical applications the echo canceller alone will not deliver sufficient echo attenuation. We therefore use an additional echo attenuation and noise reduction filter in the sending path. In Figure 1 the combination of this filter with a time domain echo compensator $c(k)$ is illustrated. $x(k)$ is the signal from the far 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 compensated signal $e(k)$. This can be written as $e(k) = s(k) + n(k) + b(k)$, where $b(k)$ is the residual echo, $b(k) = d(k) - \widehat{d}(k)$. Our aim is to make an estimation $\widehat{s}(k)$ of the near end speech by using a frequency domain filter $H(\Omega_i)$, which is calculated for each time frame of N samples (e.g. $N = 256$). The purpose of this paper is to discuss estimation procedures for the noise and residual echo power spectral densities, and methods of a combined treatment of residual echo and noise.

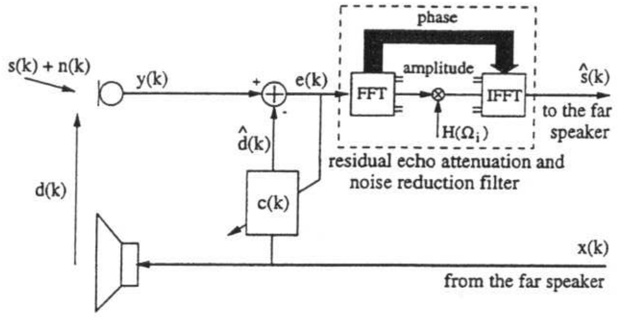

Figure 1: Block diagram of the proposed structure.

2 Algorithms

Applying the filter $H(\Omega_i)$ to the compensated signal, the spectrum $\widehat{S}(\Omega_i)$ of the estimated speech signal becomes

$$\widehat{S}(\Omega_i) = H(\Omega_i)E(\Omega_i) , \qquad (1)$$

where $E(\Omega_i)$ is the discrete Fourier transform of a frame of N samples of the signal $e(k)$ and $i = 0, 1, \dots, M-1$ denotes the frequency index, $\Omega_i = \frac{i}{M}2\pi$.

Several weighting rules $H(\Omega_i)$ which modify only the amplitude of the input signal, leaving the phase unchanged, have been developed for noise reduction. Among them the familiar Wiener filter [4] and newer methods such as the Minimum Mean Square Error Short Time Estimator (MMSE) [5] are subject to investigations for the proposed combined echo attenuation and noise reduction filter.

In case of the Wiener filter, the weighting rule for a noise reduction filter $H(\Omega_i)$ can be written as

$$H(\Omega_i) = \frac{R_{ss}(\Omega_i)}{R_{ss}(\Omega_i) + R_{nn}(\Omega_i)}, \qquad (2)$$

where $R_{ss}(\Omega_i)$ is the power spectral density of the near speech $s(k)$ and $R_{nn}(\Omega_i)$ is the power spectral density of the noise $n(k)$.

Because of the statistical independence of the signals $s(k)$, $n(k)$ and $b(k)$, the weighting rule (2) can be modified to perform as a combined echo attenuation and noise reduction filter by replacing the power spectral density $R_{nn}(\Omega_i)$ with the sum of $R_{nn}(\Omega_i)$ and the power spectral density of the residual echo, $R_{bb}(\Omega_i)$, [6]

$$H(\Omega_i) = \frac{R_{ss}(\Omega_i)}{R_{ss}(\Omega_i) + R_{nn}(\Omega_i) + R_{bb}(\Omega_i)}. \qquad (3)$$

The weighting rule (3) is the optimal echo attenuation and noise reduction filter in the sense of minimizing the error $\mathcal{E}\{(s(k) - \widehat{s}(k))^2\}$. $\mathcal{E}\{\cdot\}$ denotes expectation.

The modifications of the MMSE method are analogous, as the residual echo can be incorporated in the statistical model as follows. Based on the assumption that the real and imaginary parts of each Fourier coefficient of the near speech, the near noise as well as the residual echo are statistically independent, zero mean, and normally distributed with the variances $\frac{\sigma_s^2}{2}$, $\frac{\sigma_n^2}{2}$ and $\frac{\sigma_b^2}{2}$, respectively, the sum of the noise and the residual echo has the same properties, but now with the variance $\frac{\sigma_n^2+\sigma_b^2}{2}$. Therefore, the noise component in the derivation of the weighting rule can be interpreted as the sum of the noise *and* residual echo components.

Having a theoretical support for the combination of residual echo attenuation and noise reduction, we can now go a step further and look at implementation aspects.

2.1 Combination Using SNR Estimations

We consider again a noise reduction filter $H(\Omega_i)$. The MMSE weighting rule can be written as functions of *a priori* and *a posteriori* signal-to-noise ratios [7, 8]. The *a priori SNR* is defined as

$$SNR_n^s(\Omega_i) = \frac{\mathcal{E}\{|S(\Omega_i)|^2\}}{\mathcal{E}\{|N(\Omega_i)|^2\}}, \qquad (4)$$

and the *a posteriori SNR* as

$$SNR_n^e(\Omega_i) = \frac{|E(\Omega_i)|^2}{\mathcal{E}\{|N(\Omega_i)|^2\}}. \qquad (5)$$

The Wiener weighting rule (2) can be written as

$$H(\Omega_i) = \frac{SNR_n^s(\Omega_i)}{SNR_n^s(\Omega_i) + 1}. \qquad (6)$$

To attenuate residual echo, $N(\Omega_i)$ in Eqs. (4) and (5) can be substituted by $B(\Omega_i)$,

$$SNR_b^s(\Omega_i) = \frac{\mathcal{E}\{|S(\Omega_i)|^2\}}{\mathcal{E}\{|B(\Omega_i)|^2\}} \qquad (7)$$

$$SNR_b^e(\Omega_i) = \frac{|E(\Omega_i)|^2}{\mathcal{E}\{|B(\Omega_i)|^2\}}. \qquad (8)$$

The *a posteriori SNR* is calculated by using the instantaneous spectral components of $E(\Omega_i)$ and estimations of the power spectral densities $R_{nn}(\Omega_i)$ and $R_{bb}(\Omega_i)$, respectively. The *a priori SNR* is commonly estimated by a "decision directed" approach [5], effectively by a recursive smoothing of the *a posteriori SNR* which also takes into account the estimated filter of the previous frame. With m as the frame index, $SNR_n^{s\,(m)}(\Omega_i)$ is estimated by

$$\begin{aligned} SNR_n^{s\,(m)}(\Omega_i) &= \\ &= (1-\alpha_n)P(SNR_n^{e\,(m)}(\Omega_i) - 1) + \\ &\quad + \alpha_n \frac{|H^{(m-1)}(\Omega_i)E^{(m-1)}(\Omega_i)|^2}{R_{nn}^{(m)}(\Omega_i)}, \end{aligned} \qquad (9)$$

with $P(x) = \frac{1}{2}(|x| + x)$. The calculation of $SNR_b^s(\Omega_i)$ is analogous to (9) with the smoothing parameter α_b. The choice of the parameters α_n and α_b depends strongly on the characteristics of the signal component to be removed. For the noise reduction case, the variations of $SNR_n^e(\Omega_i)$ depends mainly on the variations of the power spectral density of the near speech, whereas $SNR_b^e(\Omega_i)$ will fluctuate more rapidly because of the instationarity of the residual echo. This leads to different choices of the constants α_n and α_b. $\alpha_n = 0.97$ and $\alpha_b = 0.90$ have been found by experiment to give high speech quality as well as good reduction of noise or residual echo, respectively.

For the combination of echo attenuation and noise reduction the different *SNR* expressions have to be combined in $SNR_{b+n}^s(\Omega_i)$ and $SNR_{b+n}^e(\Omega_i)$. As the *a posteriori SNR* is an instantaneous value and $b(k)$ and $n(k)$ can be considered as independent, the calculation of $SNR_{b+n}^e(\Omega_i)$ can be made straightforward by adding $R_{bb}(\Omega_i)$ and $R_{nn}(\Omega_i)$,

$$SNR_{b+n}^e(\Omega_i) = \frac{|E(\Omega_i)|^2}{R_{bb}(\Omega_i) + R_{nn}(\Omega_i)} \,. \tag{10}$$

Using $SNR_{b+n}^e(\Omega_i)$ for calculating $SNR_{b+n}^s(\Omega_i)$ in the same way as in Eq. (9) forces us to a compromise regarding the parameter α. It can then be choosen either for good noise reduction *or* residual echo attenuation performance. A solution can be found by computing $SNR_b^s(\Omega_i)$ and $SNR_n^s(\Omega_i)$ separately and combine them using the relation

$$\begin{aligned} SNR_{b+n}^s(\Omega_i) &= \\ &= \frac{\mathcal{E}\{|S(\Omega_i)|^2\}}{\mathcal{E}\{|B(\Omega_i)|^2\} + \mathcal{E}\{N(\Omega_i)|^2\}} = \\ &= \frac{1}{(SNR_b^s(\Omega_i))^{-1} + (SNR_n^s(\Omega_i))^{-1}} \,. \end{aligned} \tag{11}$$

Eq. (11) gives us a powerful and flexible way of treating the combination of residual echo attenuation and noise reduction. As $SNR_b^s(\Omega_i)$ and $SNR_n^s(\Omega_i)$ are calculated independently, optimal parameters α_b and α_n can be used. It is easy to see, that if for example no residual echo is present, i.e. $SNR_b^s(\Omega_i) \gg 1$, the system will act as a dedicated noise reduction system as $SNR_{b+n}^s(\Omega_i) \approx SNR_n^s(\Omega_i)$.

A common feature of all noise reduction methods and of the calculation of the different *SNR* expressions is that they require estimations of the noise power spectral density. For the combined filter, estimations of both the echo and the noise power spectral densities are necessary.

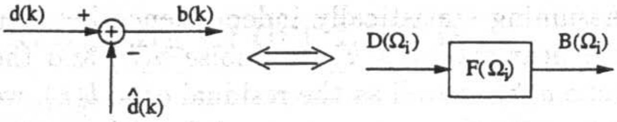

Figure 2: Interpretation of the echo compensation in terms of a transfer function $F(\Omega_i)$.

2.2 Estimation of the Residual Echo

For the calculation of $SNR_b^e(\Omega_i)$ using Eq. (8) an estimate of the power spectral density of the residual echo is needed. As the residual echo is only a function of the echo itself and the estimated echo from the compensator, a model where the compensation is considered as a transfer function is useful. This is illustrated in Figure 2. It leads to the identities

$$b(k) = d(k) - \widehat{d}(k) \tag{12}$$

$$b(k) = f * d(k) \tag{13}$$

and in the frequency domain

$$B(\Omega_i) = D(\Omega_i) - \widehat{D}(\Omega_i) \tag{14}$$

$$B(\Omega_i) = F(\Omega_i) D(\Omega_i) \,. \tag{15}$$

As verified by simulations the time domain compensator estimates the echo with just small phase errors. Consequently, we can assume that $\arg\{\widehat{D}(\Omega_i)\} = \arg\{D(\Omega_i)\}$, from which follows that $F(\Omega_i)$ is a real function. Using Eqs. (14) and (15) we can write the echo $D(\Omega_i)$ and the residual echo $B(\Omega_i)$ as a possibly noncausal function of $F(\Omega_i)$ and the estimated echo $\widehat{D}(\Omega_i)$,

$$D(\Omega_i) = \frac{1}{1 - F(\Omega_i)} \widehat{D}(\Omega_i) \tag{16}$$

$$B(\Omega_i) = \frac{F(\Omega_i)}{1 - F(\Omega_i)} \widehat{D}(\Omega_i) \,, \tag{17}$$

and with $F(\Omega_i) \in \mathbb{R}$ the power spectral densities can be calculated as

$$R_{dd}(\Omega_i) = \frac{1}{(1 - F(\Omega_i))^2} R_{\widehat{d}\widehat{d}}(\Omega_i) \tag{18}$$

$$R_{bb}(\Omega_i) = \left(\frac{F(\Omega_i)}{1 - F(\Omega_i)} \right)^2 R_{\widehat{d}\widehat{d}}(\Omega_i) \,. \tag{19}$$

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

Assuming statistically independence between the near speech $s(k)$, the noise $n(k)$ and the echo $d(k)$ as well as 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

$$R_{yy}(\Omega_i) = R_{ss}(\Omega_i) + R_{nn}(\Omega_i) + R_{dd}(\Omega_i) \tag{20}$$

$$R_{ee}(\Omega_i) = R_{ss}(\Omega_i) + R_{nn}(\Omega_i) + R_{bb}(\Omega_i) \, . \tag{21}$$

Combining the above equations with Eqs. (18) and (19) we arrive at an expression for estimating $F(\Omega_i)$, which can be calculated from known signals,

$$F(\Omega_i) = \frac{R_{yy}(\Omega_i) - R_{ee}(\Omega_i) - R_{\widetilde{dd}}(\Omega_i)}{R_{yy}(\Omega_i) - R_{ee}(\Omega_i) + R_{\widetilde{dd}}(\Omega_i)} \, . \tag{22}$$

Having an estimation of $F(\Omega_i)$ the power spectral density $R_{bb}(\Omega_i)$ can be estimated using Eq. (19).

2.3 Estimation of Noise Power

To combine the echo attenuation with noise reduction we will also need an estimation of the noise power spectral density. Most noise reduction methods make use of a voice activity detector to obtain an estimation during speech pauses. The disadvantage of this method is obvious: it requires that the noise is stationary and that there will be enough speech pauses to perform an estimation. As both these conditions are seldom met, especially not in a mobile telephone environment, where the background noise may change rapidly, we use the Minimum Statistics estimate algorithm [9]. It constantly gives an estimation $R_{nn}(\Omega_i)$ of the noise power at each discrete frequency by searching power minima in a time frame. Alternatively, the Minima Tracking method [10] can be used.

3 Results

Preliminary results obtained from MATLAB simulations show a very significant echo reduction and noise attenuation for a wide range of signal-to-noise ratios. When the filter is used to attenuate only the residual echo, the Wiener and MMSE rules perform equally well, whereas the latter is superior in the noise reduction role, as it introduces less "musical noise".

Further tests are needed for optimizing the estimation methods and the weighting rules, and for performing instrumental evaluations.

Acknowledgement

This research is supported by the German Telekom AG, Berlin. I would also like to thank Dr. Rainer Martin and Prof. Peter Vary, IND, Aachen University of Technology, for support and valuable discussions.

References

[1] 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.

[2] R. Martin und S. Gustafsson, "The Echo Shaping Approach to Acoustic Echo Control." To be published in Speech Communication.

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

[4] S. Haykin, "Adaptive Filter Theory." Prentice Hall, 1986.

[5] Y. Ephraim und 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.

[6] R. Martin, "Algorithms for Hands-Free Voice Communication in Noisy Environments." Proc. 9th Aachener Kolloquium Signaltheorie, Aachen, 1997.

[7] R. McAulay und M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter." IEEE Trans. Acoustics, Speech, Signal Processing, Vol. 28, No. 2, pp. 137-145, April 1980.

[8] P. Scalart und J. V. 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.

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

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