AN MMSE SOFT-DECISION ESTIMATOR FOR COMBINED NOISE AND RESIDUAL ECHO REDUCTION

Rainer Martin

Institute of Communication Systems and Data Processing
Aachen University of Technology, D-52056 Aachen, Germany
phone: +49 241 806984, fax: +49 241 8888 186, e-mail: martin@ind.rwth-aachen.de

ABSTRACT

This contribution presents a framework for combined noise and acoustic echo reduction which is based on minimum mean square error (MMSE) short time spectral amplitude estimation and soft-decision weighting. This framework is derived by means of quadratic cost functions and allows to use two separate signal estimators, one for near end single talk and one for double talk. The soft-decision weighting of both estimators is based on the activity of the near end and the far end speakers. We derive the estimation algorithm and compare this system to a more conventional approach which uses only one estimator.

1. INTRODUCTION

With the wide dissemination of mobile communications and the increased use of natural man-machine interfaces, acoustic echo cancellation and noise reduction algorithms are in high demand. In harsh acoustic environments, however, acoustic echo cancellation and noise reduction algorithms by themselves do not always perform satisfactorily. Thus, additional measures must be taken. It has been shown that there are advantages when the echo and noise reduction tasks are jointly treated by means of an echo canceller of reduced order and a postfilter which reduces residual echoes as well as noise. Previously, this postfilter had been optimized according to MMSE and psychoacoustic criteria [1, 2, 3, 4, 5].

In this contribution we derive a new solution for the joint reduction of noise and residual echoes. This solution takes advantage of the fact that the near end and the far end speakers are not always active at all frequencies and that the estimation of the enhanced output signal can be improved by soft-decision switching between different estimators. The new algorithm is based on the MMSE (log) spectral amplitude estimator [6, 7] and speech presence uncertainty tracking as proposed in [8]. This new estimator will be compared to a soft-decision estimator which does not take the activity of the far end speaker explicitly into account.

2. DEFINITION OF SIGNALS

Fig. 1 depicts the combination of an echo canceller C and an adaptive postfilter H. We will not treat the problem of designing a robust echo canceller C itself and assume that a canceller (e.g. [9]) is given which achieves sufficient echo reduction. The postfilter is implemented in the frequency domain by means of a Discrete Fourier Transform (DFT) analysis, a spectral modification algorithm, and an overlap/add synthesis system [4].

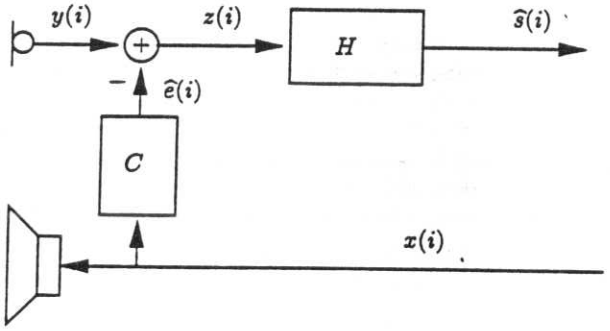

Figure 1: Combined acoustic echo and noise reduction.

In what follows, we consider bandlimited, sampled signals $x(i)$, $y(i)$, $z(i)$, and $\hat{s}(i)$ where i denotes the sampling time index. $x(i)$ is the far end speech signal and $y(i)$ is the microphone signal which is the sum of a clean near end speech signal $s(i)$, a disturbing noise $n(i)$, and an echo signal $e(i)$, $y(i) = s(i) + n(i) + e(i)$. The compensated signal $z(i)$ is equal to the microphone signal minus the estimated echo $\hat{e}(i)$, $z(i) = y(i) - \hat{e}(i) = s(i) + n(i) + \tilde{e}(i)$. $\tilde{e}(i)$ denotes the residual echo after echo compensation. We further assume that $s(i)$, $x(i)$, and $n(i)$ are statistically independent and zero mean. The noisy compensated signal $z(i)$ is transformed into the frequency domain by applying a window $h(i)$ to a frame of L consecutive samples of $z(i)$ and by computing the DFT of size L on the windowed data. Before the next DFT computation the window is shifted by R samples. This sliding window DFT analysis results in a set

of frequency domain signals which can be written as

$$Z_k(\lambda) = \sum_{\mu=0}^{L-1} z(\lambda R - \mu) h(\mu) e^{-j2\pi k\mu/L}, \qquad (1)$$

where λ is the subsampled time (frame) index, $\lambda \in \mathbb{Z}$, and k is the frequency bin index, $k \in \{0, 1..., L-1\}$, which is related to the normalized center frequency ω_k by $\omega_k = k2\pi/L$. Typically, we use a sampling rate of $f_A = 8000$ Hz and $L = 2R = 256$.

Likewise, we obtain the complex Fourier coefficients of the kth bin of the other signals in Fig. 1 where we have also dropped the frame index λ for clarity. We denote the coefficients of

- the clean near end speech by $S_k = |S_k| \exp(j\alpha_k)$,
- the noise signal by $N_k = |N_k| \exp(j\phi_k)$,
- the noisy near end speech by $Y_k = |Y_k| \exp(j\vartheta_k)$,
- the residual echo signal by $\widetilde{E}_k = |\widetilde{E}_k| \exp(j\beta_k)$,
- the compensated signal by $Z_k = |Z_k| \exp(j\zeta_k)$,
- the far end signal by $X_k = |X_k| \exp(j\theta_k)$,
- and the estimated near end speech by $\widehat{S}_k = |\widehat{S}|_k \exp(j\widehat{\alpha}_k)$.

We note that for all practical purposes and for $k \notin \{0, L/2\}$ the real and imaginary part of the Fourier transform coefficients of $Z_k(\lambda)$ (and all other signals) can be modelled as independent, zero mean Gaussian random variables [10]. Under this assumption and when the far end and the near end speakers are active each periodogram bin $|Z_k(\lambda)|^2$ is an exponentially distributed random variable with probability density function (pdf)

$$f_{|Z_k|^2}(x) = \frac{U(x) \exp(-x/(\Phi_{nn}(k) + \Phi_{ss}(k) + \Phi_{\widetilde{e}\widetilde{e}}(k)))}{\Phi_{nn}(k) + \Phi_{ss}(k) + \Phi_{\widetilde{e}\widetilde{e}}(k)}, \qquad (2)$$

where $\Phi_{ss}(k) = E\{|S_k|^2\}$, $\Phi_{nn}(k) = E\{|N_k|^2\}$, and $\Phi_{\widetilde{e}\widetilde{e}}(k) = E\{|\widetilde{E}_k|^2\}$, are the power spectral densities of the speech, the noise, and the residual echo, respectively. $U(x)$ denotes the unit step function.

3. MMSE SOFT-DECISION ESTIMATION

To derive the new estimator we introduce the hypotheses H^1_{sk} and H^0_{sk} for the presence and the absence of the near speech signal S_k in the kth frequency bin as well as the corresponding hypotheses for the residual echo signal $\widetilde{E}_k$, $H^1_{\widetilde{e}k}$ and $H^0_{\widetilde{e}k}$, respectively. Similar to the approach of [11] we model the probability density functions of the magnitude and the phase of the clean near end speech, $p_s(|S_k|, \alpha_k)$, and the magnitude and the phase of the residual echo, $p_{\widetilde{e}}(|\widetilde{E}_k|, \beta_k)$, as

$$p_s(|S_k|, \alpha_k) = P(H^1_{sk}) p(|S_k|, \alpha_k \mid H^1_{sk}) + P(H^0_{sk}) \delta(|S_k|, \alpha_k) \qquad (3)$$

$$p_{\widetilde{e}}(|\widetilde{E}_k|, \beta_k) = P(H^1_{\widetilde{e}k}) p(|\widetilde{E}_k|, \beta_k \mid H^1_{\widetilde{e}k}) + P(H^0_{\widetilde{e}k}) \delta(|\widetilde{E}_k|, \beta_k), \qquad (4)$$

where $P(H^1_{sk})$ and $P(H^1_{\widetilde{e}k})$ are the probabilities that near end speech and residual echo are present in the kth DFT bin, respectively, and $P(H^0_{sk}) = 1 - P(H^1_{sk})$ and $P(H^0_{\widetilde{e}k}) = 1 - P(H^1_{\widetilde{e}k})$. $p(|S_k|, \alpha_k \mid H^1_{sk})$ and $p(|\widetilde{E}_k|, \beta_k \mid H^1_{\widetilde{e}k})$ are the probability density functions of the near end speech and the residual echo, conditioned on the signal presence. $\delta(\cdot)$ denotes the Dirac impulse. For the four cases of near end speech presence or absence and residual echo presence or absence we define quadratic cost functions in Table 1:

case	cost function
H^0_{sk}, $H^0_{\widetilde{e}k}$	$\widehat{\lvert S\rvert}_k^2$
H^0_{sk}, $H^1_{\widetilde{e}k}$	$\widehat{\lvert S\rvert}_k^2$
H^1_{sk}, $H^0_{\widetilde{e}k}$	$(\lvert S_k\rvert - \widehat{\lvert S\rvert}_k)^2$
H^1_{sk}, $H^1_{\widetilde{e}k}$	$(\lvert S_k\rvert - \widehat{\lvert S\rvert}_k)^2$

Table 1: Cost functions for the different cases of speech activity.

In what follows we will drop the DFT bin index k. For each frequency bin k we minimize the average cost function

$$\begin{aligned}\mathfrak{C} = \iint_{\Omega_Z} \Big\{ & P(H^0_s) P(H^0_{\widetilde{e}}) \widehat{|S|}^2 p(Z \mid H^0_s, H^0_{\widetilde{e}}) \\ & + P(H^0_s) P(H^1_{\widetilde{e}}) \widehat{|S|}^2 p(Z \mid H^0_s, H^1_{\widetilde{e}}) \\ & + P(H^1_s) P(H^0_{\widetilde{e}}) \int_{\Omega_S} (\widehat{|S|} - |S|)^2 \\ & \quad \cdot p(Z \mid |S|, \alpha, H^0_{\widetilde{e}}) p(|S|, \alpha \mid H^1_s) dS \\ & + P(H^1_s) P(H^1_{\widetilde{e}}) \int_{\Omega_S} \int_{\Omega_{\widetilde{E}}} (\widehat{|S|} - |S|)^2 p(Z \mid |S|, \alpha, |\widetilde{E}|, \beta) \\ & \quad \cdot p(|S|, \alpha \mid H^1_s) p(|\widetilde{E}|, \beta \mid H^1_{\widetilde{e}}) d\widetilde{E} dS \Big\} dZ . \end{aligned} \qquad (5)$$

This integral can be minimized by minimizing the integrand of the integral over the signal space Ω_Z [11]. After differentiating the integrand with respect to $\widehat{|S|}$ and some fairly straightforward calculations we obtain

$$\begin{aligned}\widehat{|S|} = & \frac{P(H^1_s) P(H^1_{\widetilde{e}}) p(Z \mid H^1_s, H^1_{\widetilde{e}})}{P_\Sigma} E\{|S| \mid Z, H^1_s, H^1_{\widetilde{e}}\} \\ & + \frac{P(H^1_s) P(H^0_{\widetilde{e}}) p(Z \mid H^1_s, H^0_{\widetilde{e}})}{P_\Sigma} E\{|S| \mid Z, H^1_s, H^0_{\widetilde{e}}\}, \end{aligned} \qquad (6)$$

where

$$\begin{aligned} P_\Sigma = & P(H^0_s) P(H^0_{\widetilde{e}}) p(Z \mid H^0_s, H^0_{\widetilde{e}}) \\ & + P(H^1_s) P(H^1_{\widetilde{e}}) p(Z \mid H^1_s, H^1_{\widetilde{e}}) \\ & + P(H^0_s) P(H^1_{\widetilde{e}}) p(Z \mid H^0_s, H^1_{\widetilde{e}}) \\ & + P(H^1_s) P(H^0_{\widetilde{e}}) p(Z \mid H^1_s, H^0_{\widetilde{e}}) \end{aligned} \qquad (7)$$

and

$$p(Z \mid H_s^0, H_\varepsilon^0) = \frac{1}{\pi\Phi_{nn}} \exp(-\gamma_n) \tag{8}$$

$$p(Z \mid H_s^1, H_\varepsilon^0) = \frac{1}{\pi\Phi_{nn}(1+\xi_n)} \exp(-\gamma_n \frac{1}{1+\xi_n}) \tag{9}$$

$$p(Z \mid H_s^0, H_\varepsilon^1) = \frac{1}{\pi\widetilde{\Phi}_{nn}} \exp(-\gamma_d) \tag{10}$$

$$p(Z \mid H_s^1, H_\varepsilon^1) = \frac{1}{\pi\widetilde{\Phi}_{nn}(1+\xi_d)} \exp(-\gamma_d \frac{1}{1+\xi_d}) . \tag{11}$$

γ_n and γ_d are the *a posteriori* SNR values for near end single talk and for double talk, respectively. They are defined as

$$\gamma_n = \frac{|Z|^2}{\Phi_{nn}} \tag{12}$$

$$\gamma_d = \frac{|Z|^2}{\widetilde{\Phi}_{nn}} . \tag{13}$$

$\xi_n = E\{\gamma_n - 1\}/P(H_s^1)$ and $\xi_d = E\{\gamma_d - 1\}/P(H_s^1)$ are the corresponding *a priori* SNR values. The noise power spectral density for the double talk case is defined as the sum of the noise power spectral density Φ_{nn} and the power spectral density of the residual echo, conditioned on the presence of the residual echo, i.e.

$$\widetilde{\Phi}_{nn} = \Phi_{nn} + E\{|\widetilde{E}|^2 \mid H_\varepsilon^1\} . \tag{14}$$

$E\{|S| \mid Z, H_s^1, H_\varepsilon^1\}$ and $E\{|S| \mid Z, H_s^1, H_\varepsilon^0\}$ are the optimal MMSE estimators for double talk and near end single talk, respectively, and are given by

$$E\{|S| \mid Z, H_s^1, H_\varepsilon^1\} = \frac{1}{p(Z \mid H_s^1, H_\varepsilon^1)} \cdot \iint_{\Omega_S} |S|\, p(Z \mid |S|, \alpha, H_\varepsilon^1) p(|S|, \alpha \mid H_s^1) dS \tag{15}$$

and

$$E\{|S| \mid Z, H_s^1, H_\varepsilon^0\} = \frac{1}{p(Z \mid H_s^1, H_\varepsilon^0)} \cdot \iint_{\Omega_S} |S|\, p(Z \mid |S|, \alpha, H_\varepsilon^0) p(|S|, \alpha \mid H_s^1) dS . \tag{16}$$

The conditional densities $p(Z \mid |S|, \alpha, H_\varepsilon^0)$ and $p(Z \mid |S|, \alpha, |\widetilde{E}|, \beta)$ can be derived from the exponential distribution model for the DFT coefficients, i.e.

$$p(Z \mid |S|, \alpha, H_\varepsilon^0) = \frac{1}{\pi\Phi_{nn}} \exp(-\frac{|Z - |S| \exp(j\alpha)|^2}{\Phi_{nn}}) \tag{17}$$

$$p(Z \mid |S|, \alpha, H_\varepsilon^1) = \frac{1}{\pi\widetilde{\Phi}_{nn}} \exp(-\frac{|Z - |S| \exp(j\alpha)|^2}{\widetilde{\Phi}_{nn}}) \tag{18}$$

Hence, the optimal estimators $E\{|S| \mid Z, H_s^1, H_\varepsilon^1\}$ and $E\{|S| \mid Z, H_s^1, H_\varepsilon^0\}$ are given by a variation of the well known MMSE spectral amplitude estimator [6]. However, similiar to [8], improved performance is obtained when the estimators $E\{|S| \mid Z, H_s^1, H_\varepsilon^1\}$ and $E\{|S| \mid Z, H_s^1, H_\varepsilon^0\}$ are replaced by the log spectral amlitude estimators $\exp(E\{\log(|S|) \mid Z, H_s^1, H_\varepsilon^1\})$ and $\exp(E\{\log(|S|) \mid Z, H_s^1, H_\varepsilon^0\})$, respectively, as derived in [7]. In fact, all results reported below where achieved using the log spectral amplitude estimators, but other estimators could be used within this framework as well.

The unconditional *a priori SNR* values $\eta_n = E\{\gamma_n - 1\}$ and $\eta_d = E\{\gamma_d - 1\}$ are easier to estimate than ξ_n and ξ_d. η_n and η_d are estimated using the 'decision directed' approach [6]:

$$\begin{aligned} \widehat{\eta}_k(\lambda) &= a_n \frac{|S_k(\lambda-1)|^2}{\Phi_{nn}} + (1-a_n)\Xi(\gamma_n(\lambda) - 1) \\ \widehat{\eta}_d(\lambda) &= a_d \frac{|S_k(\lambda-1)|^2}{\widetilde{\Phi}_{nn}} + (1-a_d)\Xi(\gamma_d(\lambda) - 1) \end{aligned} \tag{19}$$

where $\Xi(\cdot)$ is defined as

$$\Xi(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{20}$$

These SNR values are limited to values above 0.08 (η_n) and above 0.01 (η_d).

To compute the probabilities of near end speech and residual echo presence we use two VAD's and the probability tracking approach of [8]. The VAD's are based on the mean *a posteriori* SNR values of the compensated and the far end signals [12, 13]. Near end speaker activity is detected if

$$\overline{\gamma}_d = \frac{1}{L} \sum_{k=0}^{L-1} \gamma_{dk} \tag{21}$$

is larger than a preselected threshold Γ_{dav}, e.g. $\Gamma_{dav} = 1.4$. The far end speech VAD is based on the same principle.

The probability of near end speech is tracked for all frequency bins individually. As in [8], we compare the *a posteriori* SNR γ_{nk} to a preselected threshold Γ_d (e.g. $\Gamma_d = 0.8$) and store the result of the test in an index function I_k, i.e. $I_k = 1$ if $\gamma_{nk} < \Gamma_d$ and $I_k = 0$ otherwise. Whenever speech is present in a frame this index function is smoothed to yield an estimate of the probabilities for speech absence

$$\widehat{P(H_{sk}^0)}(\lambda) = a_q \widehat{P(H_{sk}^0)}(\lambda - 1) + (1 - a_q) I_k(\lambda) , \tag{22}$$

where the smoothing parameter is set to $a_q = 0.95$. The probabilities for the residual echo are derived from the far end speech signal in the same fashion.

The power spectral density of the near end noise was estimated using the Minimum Statistics approach [14] and the power spectral density of the residual echo using the estimation method proposed in [4].

We also investigated an alternative solution which uses only one estimator and does not take the presence or absence of the far end speaker explicitly into account. This algorithm uses the same estimator for the near end single talk and the double talk case. The estimator can be derived from eq. 6 by assuming $P(H_\varepsilon^1) = 1$ at all times and for all frequencies and can be therefore written as

$$\widehat{|S|} = \frac{P(H_s^1) p(Z \mid H_s^1, H_\varepsilon^1) E\{|S| \mid Z, H_s^1, H_\varepsilon^1\}}{P(H_s^1) p(Z \mid H_s^1, H_\varepsilon^1) + P(H_s^0) p(Z \mid H_s^0, H_\varepsilon^1)} \tag{23}$$

where in this case a modified noise power spectral density is estimated as

$$\widetilde{\Phi}_{nn} = \Phi_{nn} + E\{|\widetilde{E}|^2\}. \qquad (24)$$

4. RESULTS AND CONCLUSIONS

To evaluate the performance of the proposed MMSE framework the above soft-decision methods were implemented and tested at various SNR levels. The sampling frequency of the speech material was 8 kHz. The echo signal was generated by convolving speech signals with a measured room impulse response of 512 taps and the echo canceller was simulated by truncating this impulse response to 300 taps and by adding a small perturbation to all remaining impulse response taps. The simulated echo compensated signal $z(i)$ contained an audible residual echo at a level of approximately 10 dB below the level of the near end speech signal. To avoid audible artifacts at low SNR conditions the soft-decision multipliers in eqs. 6 and 23 were limited to values above 0.1. This also limits the maximum echo and noise reduction. For noise free conditions more echo reduction is desirable. Ideally, these limiting factors should be adjusted according to the SNR of the speech sample.

The smoothing parameter of the decision directed *a priori* SNR estimators were chosen to $a_n = 0.94$ for the estimation of η_n and $a_d = 0.85$ for the estimation of η_d. The single estimator algorithm was operated with $a_d = 0.94$ at all times.

We found that both algorithms gave very little distortion of the near end speech signals and a very natural sounding residual noise. 'Musical noise' occured only sometimes at low SNR conditions (below 6 dB), mostly due to non-stationary noise. During speech pause the noise and the residual echo were attenuated by about 15 dB.

All other conditions being equal, there were only small audible differences between the algorithms of eqs. 6 and 23. Overall, the algorithm of eq. 6 achieved better echo reduction during double talk, less near end speech distortions, and less artifacts. This can be attributed to the additional flexibility of the two estimator solution. However, as outlined above, for lower SNR conditions this flexibility is somewhat reduced since the maximum attenuation must be limited in order to avoid annoying artifacts. Nevertheless, different estimators can be used for near end single talk and for double talk and parameters can be tuned in different ways for these two cases. E.g. the near end speech distortion and echo reduction during double talk can be adjusted independently from the performance during near end single talk. Hence, the proposed framework helps to improve the overall performance of the combined noise and echo reduction system.

5. REFERENCES

[1] R. Martin and S. Gustafsson, "The Echo Shaping Approach to Acoustic Echo Control." Speech Communication, Vol. 20, pp. 181-190, 1996.

[2] R. Martin and P. Vary, "Combined Acoustic Echo Control and Noise Reduction for Hands-Free Telephony - State of the Art and Perspectives." Proc. *EUSIPCO, 1996*, pp. 1107-1110, Trieste, 1996.

[3] B. Ayad and R. Le Bouquin-Jeannes, "Acoustic echo and noise reduction: A novel approach," in *Proc. International Workshop on Acoustic Echo and Noise Control*, pp. 168–171, 1997.

[4] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, pp. 21–32, 1998.

[5] S. Gustafsson, P. Jax, A. Kamphausen, and P. Vary, "A postfilter for echo and noise reduction avoiding the problem of musical tones," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1999.

[6] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, December 1984.

[7] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, pp. 443–445, April 1985.

[8] D. Malah, R. Cox, and A. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1999.

[9] C. Antweiler, "Orthogonalizing Algorithms for Digital Compensation of Acoustic Echoes." Dissertation (in German), Aachener Beiträge zu Digitalen Nachrichtensystemen, Ed. P. Vary, Vol. 1 (ISBN 3-86073-430-X), April 1995.

[10] D. Brillinger, *Time Series: Data Analysis and Theory.* Holden-Day, 1981.

[11] D. Middleton and R. Esposito, "Simultaneous Optimum Detection and Estimation of Signals in Noise," *IEEE Trans. Information Theory*, vol. 14, no. 3, pp. 434–444, 1968.

[12] A. Vähätalo and I. Johansson, "Voice Activity Detection for GSM Adaptive Multi-Rate Codec," in *Proc. IEEE Workshop on Speech Coding*, pp. 55–57, 1999.

[13] J. Yang, "Frequency Domain Noise Suppression Approaches in Mobile Telephone Systems," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 363–366, 1993.

[14] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. Euro. Signal Processing Conf. (EUSIPCO)*, pp. 1182–1185, 1994.