A POSTFILTER FOR ECHO AND NOISE REDUCTION AVOIDING THE PROBLEM OF MUSICAL TONES

Stefan Gustafsson, Peter Jax, Axel Kamphausen, and Peter Vary

Institute of Communication Systems and Data Processing,
RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany
E-mail: {gus, jax}@ind.rwth-aachen.de

ABSTRACT

In this paper we address the problem of acoustic echo cancellation and noise reduction for narrow and wide band telephone applications. We combine a conventional echo canceller with a postfilter implemented in the frequency domain and derive an algorithm for the simultaneous attenuation of residual echo and noise.

The main goals are a low level natural sounding background noise without artifacts such as musical tones, and an inaudible residual echo. This is achieved by considering the masking properties of the human auditory system.

Simulation results verify that these goals are reached, while the distortion of the near end speech is comparable to conventional algorithms. Audio demonstrations are available via Internet from http://www.ind.rwth-aachen.de.

1. INTRODUCTION

The possibility of hands-free operation of a telephone set is appreciated in a growing field of applications. An important area for narrow band systems ($\approx 0 - 3.4$ kHz) is the one of mobile telephony, where hands-free capability is demanded when using a mobile phone in a car. In such an environment the echo cancellation problem is the smaller issue, because the echo is relatively short, but noise reduction may be necessary due to high ambient noise levels. A typical wide band ($\approx 0 - 7$ kHz) application is teleconferencing, where several participants in one room use a single hands-free telephone set. In such an environment the noise levels can be expected to be lower than for the car environment, but even higher demands must be put on the speech quality and the echo cancellation.

Recently, several proposals have been made for algorithms addressing both the acoustic echo cancellation and the noise reduction problem, for example [2, 3, 5, 9, 7, 11].

Common to most speech enhancement methods is that there remain annoying artifacts (e.g. musical tones) in the processed signal. These are in general the effects of the statistical properties of the noise and arise when an estimate of the noise power spectral density (PSD) is subtracted from the input power spectrum. One way to reduce the artifacts is to limit the noise reduction so that a "noise floor" is retained, or to use other techniques to estimate the instantaneous signal-to-noise ratio (SNR), see for example [6, 4]. Nevertheless, such methods still lead to a loss of naturalness in the remaining residual noise.

In this paper a psychoacoustically motivated method for the combined attenuation of noise and residual echo left after a conventional echo canceller is presented. The main difference to previous algorithms is that the primary goal is to retain a constant level of natural sounding background noise in the output signal and an inaudible residual echo. In this process the speech distortion is only implicitly considered.

An overview of the system is shown in Fig. 1. We use a conventional echo canceller C, consisting of a time variant FIR-filter adapted by the NLMS algorithm with a variable step-size, and of a combined residual echo and noise reduction filter H implemented in the frequency domain. $x(k)$ denotes the far end speech, $s(k)$ the near end speech and $n(k)$ the noise. The microphone signal $y(k)$ is made up of the echo $d(k)$ as well as of the near end speech and noise,

$$y(k) = s(k) + n(k) + d(k). \tag{1}$$

The estimated echo $\hat{d}(k)$ is subtracted from $y(k)$ forming the echo compensated signal $e(k)$,

$$\begin{aligned} e(k) &= s(k) + n(k) + d(k) - \hat{d}(k) = \\ &= s(k) + n(k) + b(k). \end{aligned} \tag{2}$$

Depending on the effectiveness of the echo canceller, the residual echo $b(k) = d(k) - \hat{d}(k)$ must be more or less attenuated by the filter H. The output signal of the system is denoted by $\hat{s}(k)$.

In the following we denote the discrete time Fourier transform of a signal with its capital representative, e.g. $X(\Omega)$ denotes the discrete time Fourier transform of the far end speech $x(k)$.

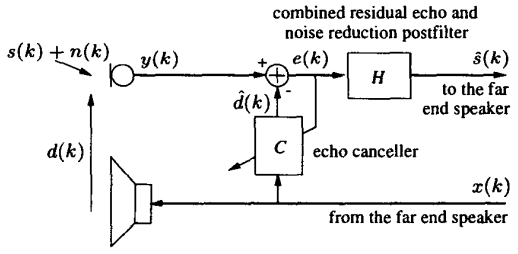

Figure 1: Block diagram of a combined echo cancelling and noise reduction system.

2. COMBINED RESIDUAL ECHO AND NOISE REDUCTION

In [8] our first approach to a psychoacoustically motivated *noise reduction* technique which succeeds in preserving the background noise characteristics was presented. Similar to other algorithms which make use of an estimate of the masking threshold [12], the near end speech was taken as the masker, but instead of using the masking threshold to modify an existing, conventional weighting rule, a completely new weighting rule was derived.

In this paper the algorithm proposed in [8] is extended to cope with the *reduction of a residual echo as well as noise* (compare [7]). The general arrangement of the filter H is shown in Fig. 2. A frame of the input signal is transformed into the frequency domain by means of the discrete Fourier transform. A conventional spectral weighting [9] is done to obtain the preliminary near end speech estimate $\check{S}(\Omega)$. From this the masking threshold $\hat{R}_T(\Omega)$

is estimated using an algorithm described in [1]. Estimations of the residual echo PSD $\hat{R}_b(\Omega)$ and the noise PSD $\hat{R}_n(\Omega)$ are performed with the methods outlined in [9] and [10], respectively. The actual reduction of the residual echo and the noise is performed in the block called "perceptual weighting", and consists of multiplying the spectral components $E(\Omega)$ with real-valued weighting coefficients $0 \leq H(\Omega) \leq 1$. Finally, the result $\hat{S}(\Omega)$ is transformed back into the time domain. Necessary functions such as windowing and overlap-and-add are not included in the figure for reasons of simplicity.

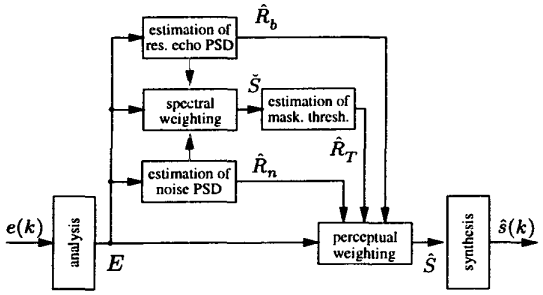

Figure 2: Block diagram of the postfilter.

2.1. Definition of Speech and Noise Distortions

We begin with defining the desired noise attenuation factor ζ_n and the desired *residual* echo attenuation factor ζ_b. Note that the attenuation ζ_b is in addition to the one already achieved by the echo canceller C. Ideally, when disregarding any signal delay caused by the system, the noise in the output signal should then be identical to $\zeta_n n(k)$, and the remaining residual echo should be $\zeta_b b(k)$. Now, in terms of short-time spectral analysis, the desired output signal of the system is

$$\tilde{S}(\Omega) = S(\Omega) + \zeta_n N(\Omega) + \zeta_b B(\Omega). \quad (3)$$

The output of the system is the result of multiplying the input spectrum $E(\Omega) = S(\Omega) + N(\Omega) + B(\Omega)$ with the real-valued weighting coefficients $H(\Omega)$,

$$\hat{S}(\Omega) = H(\Omega)[S(\Omega) + N(\Omega) + B(\Omega)]. \quad (4)$$

We now define the error $Q(\Omega)$ as the difference between the desired and the actual output spectra,

$$Q(\Omega) = \tilde{S}(\Omega) - \hat{S}(\Omega). \quad (5)$$

Assuming that the speech, noise and residual echo can be described by mutually independent and stationary statistical processes, the power spectral density $R_q(\Omega) = \mathcal{E}\{|Q(\Omega)|^2\}$, where $\mathcal{E}\{\cdot\}$ denotes the expectation operator, can be split into three components,

$$R_q(\Omega) = R_{q_s}(\Omega) + R_{q_n}(\Omega) + R_{q_b}(\Omega), \quad (6)$$

where

$$R_{q_s}(\Omega) = [1 - H(\Omega)]^2 R_s(\Omega) \quad (7)$$

$$R_{q_n}(\Omega) = [\zeta_n - H(\Omega)]^2 R_n(\Omega) \quad (8)$$

$$R_{q_b}(\Omega) = [\zeta_b - H(\Omega)]^2 R_b(\Omega). \quad (9)$$

$R_s(\Omega)$, $R_n(\Omega)$, $R_b(\Omega)$ denote the PSDs of the speech, the noise and the residual echo, respectively. All error components are quadratic functions of $H(\Omega)$. The first component $R_{q_s}(\Omega)$ is the distortion of the speech and is minimized by $H(\Omega) = 1$. The second component $R_{q_n}(\Omega)$ is the "noise distortion", i.e. the power of the difference between the desired and the actual noise, and is minimized by $H(\Omega) = \zeta_n$. Similarly, $R_{q_b}(\Omega)$ is the power of the difference between the desired and the actual residual echo. It is minimized by choosing $H(\Omega) = \zeta_b$.

The total error $R_q(\Omega)$ will be minimized for some $H^{opt}(\Omega)$ in $\min\{\zeta_n, \zeta_b\} \leq H^{opt}(\Omega) \leq 1$. The solution can be written as the ratio between the cross power spectral density $\mathcal{E}\{E(\Omega)\tilde{S}^*(\Omega)\}$ and the auto power spectral density $\mathcal{E}\{|E(\Omega)|^2\}$, and can be regarded as a modified Wiener rule,

$$H^{opt}(\Omega) = \frac{R_s(\Omega) + \zeta_n R_n(\Omega) + \zeta_b R_b(\Omega)}{R_s(\Omega) + R_n(\Omega) + R_b(\Omega)}. \quad (10)$$

We will return to this interpretation later on.

2.2. Design Object: Mask the Noise and Residual Echo Distortions

As mentioned in section 1, a problem with many noise reduction algorithms is that there remains an annoying, unnatural sounding noise in the processed signal. This problem also exists for algorithms combining the reduction of residual echo and noise.

A common way to use the masking threshold $R_T(\Omega)$ in the context of speech enhancement is to first calculate the weighting coefficients according to some conventional rule (e.g. H^{opt} in Eq. (10)) and then in some way modify the weighting coefficients, for example to only attenuate frequency components where the noise or the residual echo are not already masked by the near end speech. This approach may lead to a reduction of the artifacts, but will not remove them completely.

In [8] it was argued that to achieve a perceived noise attenuation in accordance with ζ_n, the noise distortion power $R_{q_n}(\Omega)$ should be placed at the masking threshold $R_T(\Omega)$ of the speech. The same argumentation can be used for the case of combined residual echo and noise reduction: if $H(\Omega)$ is chosen such that the sum $R_{q_n}(\Omega) + R_{q_b}(\Omega)$ is equal to $R_T(\Omega)$, then the "distortion" of the noise and the residual echo will be masked by the speech. To calculate $H(\Omega)$ we then have to solve the second order equation

$$[\zeta_n - H(\Omega)]^2 R_n(\Omega) + [\zeta_b - H(\Omega)]^2 R_b(\Omega) = R_T(\Omega). \quad (11)$$

The solution, which we call $H^{JND}(\Omega)$, JND standing for "Just Noticeable Distortion", can be found in Eq. (12) at the bottom of the next page. For $\zeta_n = \zeta_b = 0$ the weighting rule must not be negative, so we choose the "+"-solution.

It is not guaranteed that the argument of the square root is positive. However, assuming that near end speech is present and that $R_T(\Omega)$ is not too small compared to $R_n(\Omega)$ and $R_b(\Omega)$, and because ζ_n and ζ_b in general are much smaller than 1, the negative term can be neglected in favour of the dominating, positive one. With this simplification Eq. (12) reduces to the approximation $\tilde{H}^{JND}$ in Eq. (13).

2.3. Discussion of the New Weighting Rule

The weighting rule $\tilde{H}^{JND}(\Omega)$ in Eq. (13) resembles the one derived in [8]. The first term is a function of the masking threshold of the near end speech and the PSDs of the noise and the residual echo. The second term defines the minimum value of $\tilde{H}^{JND}(\Omega)$ for a given set of parameters. To understand the effect of the second term it is helpful to consider the situation when the near end

speaker is inactive. Then $s(k)$, $R_s(\Omega)$, and $R_T(\Omega)$ are all zero. The consequence for $\tilde{H}^{JND}(\Omega)$ is that the first term vanishes and the solution is equal to the modified Wiener rule in Eq. (10) for $R_s(\Omega) = 0$.

The second term balances the amount of residual echo and noise in the output signal. In general, the residual echo attenuation factor ζ_b is chosen to be much smaller than the noise attenuation factor ζ_n. Then, if on the one hand the noise is much stronger than the residual echo, $\tilde{H}^{JND}(\Omega)$ will approximately attain the value of ζ_n. Because the residual noise presumably already masks the residual echo, no extra attenuation is necessary. On the other hand, if $R_n(\Omega) \ll R_b(\Omega)$, then $\tilde{H}^{JND}(\Omega)$ will decrease towards the minimum level ζ_b. The audible effect is that when residual echo as well as noise are present, both components are mixed to achieve a nearly constant level of background noise, which to a high degree retains the characteristics of the original noise $n(k)$.

If neither near end speech nor residual echo are present, $\tilde{H}^{JND}(\Omega)$ simplifies to ζ_n. The output signal is identical to the input noise up to a scalar factor and as a consequence no artifacts of any kind are introduced. Actually, artifacts (such as musical tones) due to a subtraction of the noise power spectral density estimate $R_n(\Omega)$ from the input power spectrum $|E(\Omega)|^2$ are avoided when using $\tilde{H}^{JND}$.

In the derivation of H^{JND} the speech distortion is not considered explicitly, yet it is minimized for the given set of parameters; if a larger weighting factor is chosen, the speech distortion will certainly be reduced, but more residual echo and noise reduction will be audible; if the weighting factor is smaller, then the speech distortion will increase without any perceivable improvement of the residual echo and noise reduction.

3. APPLICATIONS AND RESULTS

3.1. Narrow Band Application – Car Environment

To evaluate the proposed algorithm it was implemented in a simulation system. For a narrow band application with sampling frequency 8 kHz we used a decimation rate of 128, a frame length of 256 samples, and a 512-point FFT, with Eq. (13) considered for $\Omega = \Omega_i = 2\pi\frac{i}{512}, i \in \{0, 1, \ldots, 511\}$. A Hamming window of length 256 was used for input data weighting. The loudspeaker-room-microphone system (LRM-system) of a medium sized car was measured and modelled with an FIR-filter of order 400. The echo canceller order was only 200 and the attenuation factors were chosen to $20\lg\zeta_n = -15$ and $20\lg\zeta_b = -35$. Owing to its shortness, the echo canceller converges fast and is very robust even when the disturbances (near end speech and noise) are strong. However, there always remains an audible residual echo in the echo compensated signal $e(k)$. The near end input signals were made up of a set of phonetically balanced sentences mixed with car noise to achieve signal-to-noise ratios from -10 dB to 25 dB. The far end signals consisted of another set of speech sentences.

The Echo Return Loss Enhancement was measured for the echo canceller ($ERLE_C$) and for the complete system ($ERLE_{CH}$). The noise attenuation NA is defined as the mean noise power in the input signal $y(k)$ divided by the mean noise power in the output signal $\hat{s}(k)$. Since the case of an inactive far end speaker is identical to the noise reduction situation described in [8], we only present the single talk (of the far end speaker) and double talk results. The $\tilde{H}^{JND}$ weighting rule is compared with the MMSE-LSA (Minimum Mean Square Error - Log-Spectral Amplitude) rule [6] which is denoted here by H^{LSA}. It is applied for a combined residual echo and noise reduction as described in [9] and adjusted to achieve the same noise and residual echo attenuation as $\tilde{H}^{JND}$ during double talk.

The instrumental measures of the single talk situation are plotted in the upper diagram of Fig. 3. We see that the low order echo canceller is relatively insensitive to the background noise. However, an echo attenuation of 15 to 20 dB is not sufficient, especially when there is no near end signal which can mask the residual echo. For both methods, the total echo attenuation depends on the noise level. At high SNR a strong attenuation (more than 40 dB) is necessary, whereas at low SNR the residual noise will already mask parts of the residual echo.

The main difference between $\tilde{H}^{JND}$ and the reference H^{LSA} is the audible impression of the residual noise. This is not accounted for at all in the instrumental measures. As with most conventional noise reduction methods, the residual noise resulting from H^{LSA} suffers from artifacts and loss of naturalness (although the result is much better than when using a standard spectral subtraction). In contrast, the $\tilde{H}^{JND}$ weighting rule achieves a residual noise which preserves the characteristics of the original noise. In the output signal the residual echo and noise are well balanced, so that only slight fluctuations of the residual noise hint the presence of a residual echo.

In a double talk situation the near end speech will already mask a great deal of the residual echo, so the total echo attenuation is allowed to be much lower. Because of the stronger disturbance, the echo canceller doesn't work as well as for single talk. The results are plotted in the lower diagram of Fig. 3.

Several speech distortion measures such as the segmental SNR, the cepstral distance and the basilar distance were used to evaluate the distortion of the near end speech. Although the two last-mentioned ones take the properties of the auditory system into account, the segmental SNR had often a higher correlation with the perceived results obtained from informal listening tests. These listening tests revealed that the speech distortions from applying either of the two considered weighting rules are similar, although the measured segmental SNR was somewhat higher for MMSE-LSA.

Summed up, the instrumental results and the audible speech distortions are equivalent for both methods. Once again the main difference lies in the characteristics of the residual noise: just as in the single talk and noise reduction situations, the result from using $\tilde{H}^{JND}$ is considerably more pleasant.

3.2. Wide Band Application – Tele-Conference

A typical wide band application for hands-free telephony is a conference telephone used in a relatively large office room. Not only

$$H^{JND}(\Omega) = \min\left(\frac{\zeta_n R_n(\Omega) + \zeta_b R_b(\Omega) \pm \sqrt{(R_n(\Omega)+R_b(\Omega))R_T(\Omega) - R_n(\Omega)R_b(\Omega)(\zeta_n-\zeta_b)^2}}{R_n(\Omega)+R_b(\Omega)}, 1\right) \tag{12}$$

$$\tilde{H}^{JND}(\Omega) = \min\left(\sqrt{\frac{R_T(\Omega)}{R_n(\Omega)+R_b(\Omega)}} + \frac{\zeta_n R_n(\Omega) + \zeta_b R_b(\Omega)}{R_n(\Omega)+R_b(\Omega)}, 1\right) \tag{13}$$

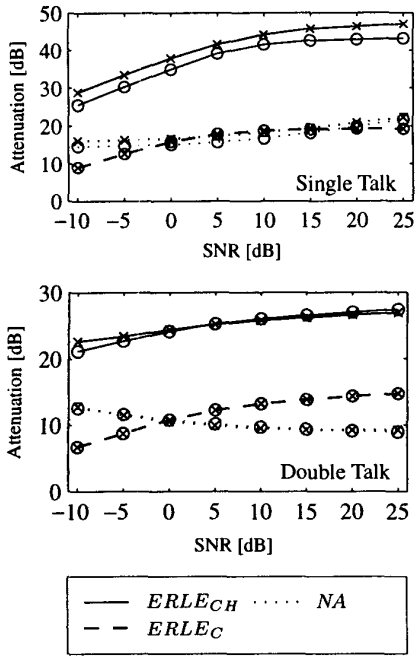

Figure 3: Single Talk (upper diagram) and double talk (lower diagram) results for $\tilde{H}^{JND}$ (×) and for H^{LSA} (○), as a function of input SNR.

because the impulse response of the LRM-system is longer, but also because the sampling frequency is higher, we have to deal with an LRM-system model which is of a much higher order than in the typical narrow band case. Consequently, the adaptive filter of the echo canceller must also be of an higher order. This leads to slower convergence and higher sensitivity to disturbances.

For the simulations, the LRM-system was modelled with an FIR-filter of order 2000, which corresponds to a duration of only 125 ms. The values for the decimation rate, frame length, and FFT size were all doubled compared to the narrow band system. Wide band speech samples and computer fan noise were used. The signal-to-noise ratio can be expected to be much higher in an office environment than in a car, so we limited our study to SNRs between 0 and 25 dB.

The instrumental results (which are not presented here) indicate that for noise reduction, the wide band system works as satisfactorily as the narrow band system, but that when the far end speaker is active, it performs considerably less well. This was also verified by the listening tests.

In most situations an echo canceller of much lower order than 2000 had significant convergence problems resulting in strongly fluctuating residual echo and background noise.

With an echo canceller of full length (i.e. 2000 coefficients), the system worked relatively well during single talk and during double talk if there was only a weak noise present at the near end. However, the remaining residual echo and background noise wasn't that well balanced as in the narrow band case. Although the residual echo couldn't be clearly heard, the presence of the echo could be perceived from the fluctuations of the background noise and/or higher distortions of the near end speech.

4. CONCLUSIONS

The combined residual echo and noise attenuation is in many aspects an extension of the noise reduction algorithm proposed in [8]. For narrow band signals, the combined algorithm can be used in connection with a standard echo canceller of relatively short order to achieve a very high overall echo attenuation. In many situations even a reduction of the total system complexity is possible because an echo canceller of lower order can be used. Compared to conventional weighting rules, the strength of $\tilde{H}^{JND}$ is that the original noise characteristics are preserved and no musical tones are introduced.

The conclusions which can be drawn from the wide band experiments are that an echo canceller producing a reliable estimate of the echo is absolutely necessary – an echo canceller of significantly reduced order might only be applicable in some special situations – and that the procedure for estimating the residual echo PSD must be modified to better estimate the PSD of the late residual echo components. This is indeed a formidable task, because the later part of the echo often has statistical properties different to the early part.

5. REFERENCES

[1] ISO/IEC 11172-3:1993. Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3, Audio, 1993.

[2] B. Ayad and R. Le Bouquin-Jeannès. Acoustic echo and noise reduction: A novel approach. In *Proc. Int. Workshop on Acoustic Echo and Noise Control*, 1997. London.

[3] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire. New optimal filtering approaches for hands-free telecommunication terminals. *Signal Processing*, 64(1), 1998.

[4] O. Cappé. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech and Audio Processing*, 2(2), 1994.

[5] P. Dreiseitel and H. Puder. A combination of noise reduction and improved echo cancellation. In *Proc. Int. Workshop on Acoustic Echo and Noise Control*, 1997. London.

[6] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. ASSP*, 33(2), 1985.

[7] S. Gustafsson and P. Jax. Combined residual echo and noise reduction: A novel psychoacoustically motivated algorithm. In *Proc. EUSPICO*, 1998. Rhodos, Greece.

[8] S. Gustafsson, P. Jax, and P. Vary. A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics. In *Proc. ICASSP*, 1998. Seattle, USA.

[9] S. Gustafsson, R. Martin, and P. Vary. Combined acoustic echo control and noise reduction for hands-free telephony. *Signal Processing*, 64(1), 1998.

[10] R. Martin. Spectral subtraction based on minimum statistics. In *Proc. EUSIPCO*, 1994. Edinburgh, UK.

[11] V. Turbin, A. Gilloire, P. Scalart, and C. Beaugeant. Using psychoacoustic criteria in acoustic echo cancellation algorithms. In *Proc. Int. Workshop on Acoustic Echo and Noise Control*, 1997. London.

[12] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. Springer-Verlag, New York, 1990.