COMBINED RESIDUAL ECHO AND NOISE REDUCTION: A NOVEL PSYCHOACOUSTICALLY MOTIVATED ALGORITHM

Stefan Gustafsson and Peter Jax

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 focus on the problem of acoustic echo cancellation and noise reduction for hands-free telephony devices. A standard echo canceller is combined with a frequency domain post-filter, which applies a novel psychoacoustically motivated weighting rule.
The algorithm makes use of the masking threshold of the human auditory system to achieve a perceived reduction of noise and residual echo equal to some pre-defined levels.
In contrast to conventional methods, the proposed one preserves the nature of the original background noise and doesn't introduce any audible artifacts. At the same time it can attain a very high reduction of the residual echo.

1 INTRODUCTION

Until some years ago, noise reduction algorithms for speech signals were in general based upon some form of spectral subtraction [1, 2]. The drawback of these methods is that a very unpleasant residual noise in form of musical tones remains in the processed signal, and that the speech is distorted. Later proposals [3, 4, 5], differing primarily in the way the signal-to-noise ratio (SNR) is estimated, reduce the amount of musical tones, but the residual noise still sounds unnatural.

Still in an early stage are speech enhancement methods relying on psychoacoustical considerations. Most contributions in this area exploit the masking properties of the auditory system. In principle, they make use of various linear or nonlinear weighting rules, which are adjusted according to the noise masking threshold [6, 7].

For the problem of acoustic echo cancellation, as necessary in hands-free telephony devices, several solutions are proposed. Common to these is that they in general never succeed in a complete acoustic echo cancellation – there will always be some residual echo left audible in the output signal.

Studies have been made to combine noise reduction with a *residual echo attenuation* algorithm. These algorithms combine a conventional echo compensator with an adaptive filter in the sending path, see for example [8, 9, 10]. Alternatively, the combined task is accomplished by a single filter [11].

Common to both the pure noise reduction algorithms and those for the combined reduction of noise and residual echos is that their performance strongly depends on how well the power spectral densities (PSDs) of the noise and the residual echo can be estimated. The better the estimation is, the more natural the residual noise sounds (with fewer musical tones), the less residual echo is left, and the lower the distortion of the speech is.

In this paper we will discuss a new approach to speech enhancement based on psychoacoustics. In contrast to previous methods, the proposed one does not use the masking threshold to modify a standard spectral weighting rule, but uses it in a direct manner to calculate the weighting coefficients, such that the perceived noise suppression will always be equal to a predefined level. The residual echo is attenuated to be inaudible, i.e. masked by the near end speech and/or the remaining low-level background noise.

2 PSYCHOACOUSTICS

Some models to describe the perception of an audio signal have been developed in the past [12]. Especially, the known phenomenon of auditory masking has been exploited successfully in signal processing systems, e.g. in the field of wide-band audio coding [13, 14]. They build upon the fact that a human listener will not perceive any additive signal components as long as their power spectral density lies completely below the so called *masking threshold*. It must, however, be emphasized that conclusions about the subjective perception of partially masked signals can not be easily drawn from the knowledge of the masking threshold alone.

In most situations a complete removal of the noise is neither necessary nor desirable. In a telephone application, for example, a retained low-level natural sounding background noise will give the far end user a feeling of the atmosphere at the near end, and also avoids the impression of an interrupted transmission. Consequently, it is only desired to reduce the noise level by some predefined amount. However, in this step the spectral char-

acteristics (i.e. the colour) of the noise shall be preserved.

In contrast, no acoustic echo should be left audible in the processed signal. This does not necessarily impose a complete cancellation of the echo, as there may also be speech and noise present at the near end which can mask some residual echo. The fact that the far end speaker himself might mask parts of the echo is not considered here.

3 ALGORITHM

A block diagram of the proposed system for combined residual echo and noise reduction is illustrated in Fig. 1. $x(k)$ denotes the signal from the far end speaker, $s(k)$ the near end speech, and $n(k)$ an additive near end noise. These are assumed to be mutually statistically independent. The echo compensator C estimates the echo $d(k)$ and subtracts it from the microphone signal $y(k)$, yielding the echo compensated signal $e(k) = s(k)+n(k)+b(k)$, where $b(k) = d(k)-\hat{d}(k)$ is the residual echo. The filter H, which is implemented in the frequency domain, performs the subsequent noise and residual echo attenuation.

In terms of short-time spectral analysis, let $S(\Omega_i)$, $N(\Omega_i)$, and $B(\Omega_i)$ denote the discrete and complex Fourier transformations of the speech $s(k)$, the noise $n(k)$, and the residual echo $b(k)$, respectively, with $\Omega_i = 2\pi\frac{i}{M}, i \in \{0, 1, \ldots, M-1\}$. The power spectral densities (PSD) are denoted with $R_s(\Omega_i)$, $R_n(\Omega_i)$, and so on.

The algorithm which we present in this paper, designed to reduce residual echo as well as noise, is an extension of our noise reduction algorithm described in [15, 16]. We first define a desired residual noise level ζ_n and a desired residual echo level ζ_b, e.g. -20 and -40 dB, respectively. If only noise $n(k)$ is present in the microphone signal $y(k)$, then ideally the output signal should be $\zeta_n n(k-T)$, where T is the algorithmic delay. Similarly, if only echo $d(k)$ is present, the output should be $\zeta_b b(k-T)$. Note that ζ_b only refers to the additional echo reduction performed by the filter H.

3.1 A Two-Step Approach: First Reducing Noise and then Echo

The above process can be performed in one or two steps. For the two-step alternative, the masking threshold of the near end speech is first estimated. This is done by performing a spectral weighting, reducing both noise and residual echo [10], and then calculating the masking threshold $R_T(\Omega_i)$ for the resulting speech signal estimation. In this first step the noise component is then reduced with the method described in [16]. In this way only the desired amount of noise is left audible. The rest is either reduced or masked by the near end speech.

In the second step, the residual echo is attenuated much in the same way as the noise is in the first step; the weighting rule is chosen to attenuate the residual

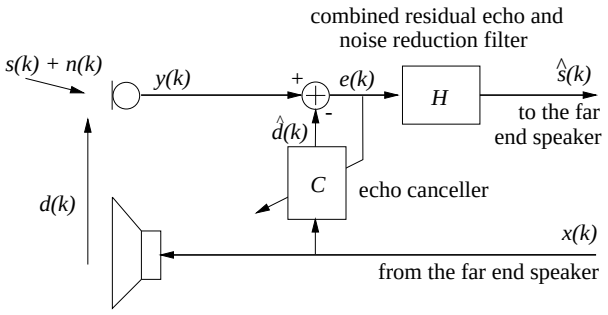

Figure 1: Block diagram of the combined echo cancelling and noise reduction system.

echo such that it is masked by the near end speech and the desired residual noise.

The disadvantage of this two-step processing is of course that two spectral weightings and two masking threshold calculations must be performed. Further, we found it difficult to correctly estimate the masking threshold of the residual noise after the first step. The result was a signal where either the residual noise level fluctuated or the residual echo could still be heard.

3.2 Reducing Noise and Residual Echo Simultaneously

The above task can also be performed with a single filter by precisely defining how the attenuation of the noise and the residual echo should be done. We will now describe this algorithm in detail.

With the attenuation factors ζ_n and ζ_b as described above, we can write the *desired* output signal of the system as

$$\tilde{S}(\Omega_i) = S(\Omega_i) + \zeta_n N(\Omega_i) + \zeta_b B(\Omega_i). \quad (1)$$

The actual output of the system with the real-valued weighting coefficients $H(\Omega_i)$ is

$$\hat{S}(\Omega_i) = H(\Omega_i)\left(S(\Omega_i) + N(\Omega_i) + B(\Omega_i)\right). \quad (2)$$

The error $E(\Omega_i) = \hat{S}(\Omega_i) - \tilde{S}(\Omega_i)$ can be expressed as

$$\begin{aligned} E(\Omega_i) = {} & S(\Omega_i)(H(\Omega_i)-1) + N(\Omega_i)(H(\Omega_i)-\zeta_n) \\ & + B(\Omega_i)(H(\Omega_i)-\zeta_b), \end{aligned} \quad (3)$$

and its PSD as

$$\begin{aligned} R_e(\Omega_i) = {} & R_s(\Omega_i)(H(\Omega_i)-1)^2 + R_n(\Omega_i)(H(\Omega_i)-\zeta_n)^2 \\ & + R_b(\Omega_i)(H(\Omega_i)-\zeta_b)^2. \end{aligned} \quad (4)$$

The PSD of the error contains three components, $R_{e_s}(\Omega_i) = R_s(\Omega_i)(H(\Omega_i)-1)^2$, $R_{e_n}(\Omega_i) = R_n(\Omega_i)(H(\Omega_i)-\zeta_n)^2$, and $R_{e_b}(\Omega_i) = R_b(\Omega_i)(H(\Omega_i)-\zeta_b)^2$, which in a non-trivial case cannot be equal to zero simultaneously. All error components are quadratic functions of $H(\Omega_i)$ and are minimized by choosing $H(\Omega_i)$ to 1, ζ_n or ζ_b, respectively. The total error will thus be minimized for some $H_{opt}(\Omega_i)$ in the range $\min(\zeta_n, \zeta_b) \le H_{opt}(\Omega_i) \le 1$. With $\zeta_n = \zeta_b = 0$ this $H_{opt}(\Omega_i)$ is equal to the Wiener solution.

With the same argumentation as in [16] our spectral weighting rule is defined by putting the sum of the error components $R_{e_n}(\Omega_i)$ and $R_{e_b}(\Omega_i)$ exactly on the masking threshold $R_T(\Omega_i)$ of the near end speech,

$$R_n(\Omega_i)(H(\Omega_i)-\zeta_n)^2 + R_b(\Omega_i)(H(\Omega_i)-\zeta_b)^2 \stackrel{!}{=} R_T(\Omega_i). \quad (5)$$

We call the solution of this second order equation $H^{JND}(\Omega_i)$, JND standing for Just Notable Distortion. Solving it with the constraint $H(\Omega_i) \leq 1$ we obtain Eq. (6), found at the bottom of this page. $H(\Omega_i)$ must not be negative if both ζ_n and ζ_b are chosen to zero, so we choose the "+"-solution.

Looking at the argument of the square root in Eq. (6) and assuming that $R_b(\Omega_i)$ and $R_n(\Omega_i)$ are not much larger than $R_T(\Omega_i)$, we can see that for $\zeta_n, \zeta_b \ll 1$ the first term is dominating. Eq. (6) then simplifies to Eq. (7).

Eq. (7) consists of three terms. The last term is only a function of the masking threshold of the near end speech and the power spectral densities of the noise and the residual echo. The sum of the first two terms is the minimum value of $H^{JND}(\Omega_i)$ for the given parameters.

The factors $R_n(\Omega_i)/(R_n(\Omega_i) + R_b(\Omega_i))$ and $R_b(\Omega_i)/(R_n(\Omega_i) + R_b(\Omega_i))$ indicate an *adaptive adjustment* of the desired reduction levels. For example, if the residual echo is much stronger than the noise, $R_b(\Omega_i) \gg R_n(\Omega_i)$, the first term will be very small, thus effectively reducing the minimum value of $H(\Omega_i)$. This will lead to a greater attenuation; the hearable effect is that the residual echo is hidden by the background noise.

If the noise is much stronger than the residual echo, i.e. when $R_n(\Omega_i) \gg R_b(\Omega_i)$ and no extra attenuation to reduce the residual echo is necessary, then the middle term will be comparatively small with ζ_n dominating, approaching the noise-only solution.

We can thus see that the factors before ζ_n and ζ_b are key elements to obtain a constant level residual noise. The speech distortions are not explicitly considered by the weighting rule, yet the solution reduces them to the smallest possible value for the specified noise and residual echo reduction. A greater $H^{JND}(\Omega_i)$ will doubtlessly reduce the distortions, but at the same time more noise and/or residual echo will be audible. A smaller weighting factor would lead to a larger distortion without any perceivable improvement of the noise and residual echo reduction.

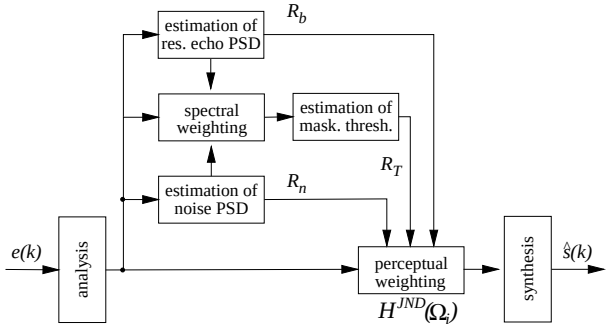

Figure 2: Block diagram of the arrangement of the noise and residual echo reduction filter.

4 SIMULATIONS

4.1 Implementation

Simulations were performed with the combined system in Fig. 1 for a car environment. We used an echo canceller C consisting of an FIR filter of order 200 adapted using the NLMS-algorithm. The finite loudspeaker-room-microphone impulse response had a length of about twice the one of C. The near end speech power was equal to the power of the echo. The sampling frequency was 8 kHz.

Fig. 2 shows an arrangement of the filter H. The spectral analysis/synthesis is based on the *overlap-and-add* method, using FFT/IFFT of length 512, data frame length 256, decimation ratio 128, and a Hamming window function for input signal weighting. An estimation of the near end speech is first performed with the spectral weighting method described in [10]. Then the masking threshold $R_T(\Omega_i)$ is calculated using a mixture of the Johnston and the ISO models [13, 14].

4.2 Results

We compare the new weighting rule Eq. (7), simply called H^{JND}, with the MMSE LSA [4], and thereby consider four instrumental measures:

- ERLE_C (the Echo Return Loss Enhancement achieved by the echo canceller alone),
- ERLE_CH (the total ERLE),
- NA (the attenuation of the noise $n(k)$),
- SA (the attenuation of the near end speech $s(k)$).

Mean values of the simulation results are plotted in Fig. 3 for different input SNR, defined as the segmental SNR between the near end speech and the noise. The ERLE_CH is a few decibel higher for H^{JND} than for MMSE LSA, whereas the noise attenuation is almost identical for both methods.

$$H^{JND}(\Omega_i) = \min\left(\frac{\zeta_n R_n(\Omega_i) + \zeta_b R_b(\Omega_i) \pm \sqrt{(R_n(\Omega_i)+R_b(\Omega_i))R_T(\Omega_i) - R_n(\Omega_i)R_b(\Omega_i)(\zeta_n-\zeta_b)^2}}{R_n(\Omega_i)+R_b(\Omega_i)},\ 1\right) \quad (6)$$

$$H^{JND}(\Omega_i) \approx \min\left(\frac{R_n(\Omega_i)}{R_n(\Omega_i)+R_b(\Omega_i)}\zeta_n + \frac{R_b(\Omega_i)}{R_n(\Omega_i)+R_b(\Omega_i)}\zeta_b + \sqrt{\frac{R_T(\Omega_i)}{R_n(\Omega_i)+R_b(\Omega_i)}},\ 1\right) \quad (7)$$

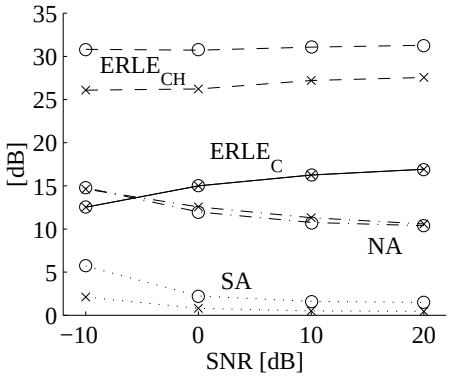

Figure 3: $\mathrm{ERLE_C}$, $\mathrm{ERLE_{CH}}$, SA, and NA (see text) as a function of input SNR for H^{JND} ($\circ$) and MMSE LSA ($\times$).

As far as the attenuation SA of the near end speech is concerned, especially for H^{JND} it gets larger as the noise level increases. However, looking at the difference NA - SA, the effective noise reduction for H^{JND} remains almost constant. Should only the near end speaker be active, the system acts as a dedicated noise reduction system, whose performance is described in [16].

When only the far end speaker is active, the $\mathrm{ERLE_{CH}}$ for $H^{JND}(\Omega_i)$ is about 40 to 55 dB, depending on background noise levels. This is 5 to 10 dB higher than for MMSE LSA. The noise attenuation is then between 15 and 20 dB for both algorithms.

It was found from informal listening tests that the H^{JND} method retains the natural character of the near end speech, whereas MMSE LSA makes the speech sound somewhat artificial at low SNR. On the other hand, we also studied several instrumental measures (segmental SNR, cepstral distance, and basilar distance) for determining the distortion of the near end speech caused by the residual echo and noise reduction. No significant difference was found between the H^{JND} and MMSE LSA methods, except at very low SNR, where the higher speech attenuation of H^{JND} was reflected in somewhat worse figures.

However, the audible impression of the low level background noise comprises the most important difference between the two methods. Although the MMSE LSA method produces few artifacts in form of musical tones, the residual noise definitely sounds different from the original one. The new H^{JND} weighting rule fully retains the spectral character of the noise – it sounds very much like the original noise but at a much lower level and no artifacts are audible.

5 CONCLUSIONS

The algorithm discussed in this paper efficiently reduces the background noise and residual echo by utilizing the psychoacoustic properties of the human ear. Thereby it preserves the characteristics of the noise and avoids annoying artifacts such as musical tones.

Although the standard echo canceller only attenuates the echo with some 10 to 20 dB, the overall echo attenuation will be 40 to 55 dB, depending on the background noise level. This is significantly higher than for previous, non-psychoacoustic methods, and is deemed sufficiently high.

References

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 27, no. 2, pp. 113 – 120, April 1979.

[2] R.J. McAulay and M.L. 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.

[3] 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.

[4] 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.

[5] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 345 – 349, 1994.

[6] D.E. Tsoukalas, J. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 6, pp. 497 – 514, November 1997.

[7] N. Virag, "Speech enhancement based on masking properties of the auditory system", in *Proceedings ICASSP*, May 1995, pp. 796–799, Detroit, USA.

[8] R. Martin and J. Altenhöner, "Coupled adaptive filters for acoustic echo control and noise reduction", in *Proceedings ICASSP*, May 1995, pp. 3043 – 3046, Detroit, USA.

[9] R. Martin and S. Gustafsson, "The echo shaping approach to acoustic echo control", *Speech Communication*, vol. 20, no. 3–4, December 1996.

[10] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony", *Signal Processing*, vol. 64, no. 1, January 1998.

[11] C. Beaugeant, V. Turbin, P. Scalart, and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals", *Signal Processing*, vol. 64, no. 1, January 1998.

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

[13] 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.

[14] J. Johnston, "Transform coding of audio signals using perceptual noise criteria", *IEEE Journal on Selected Areas of Communication*, vol. 6, pp. 314–323, February 1988.

[15] P. Jax, "Entwicklung und Untersuchung von Algorithmen für eine gemeinsame Echo- und Störgeräuschreduktion im Frequenzbereich", diploma thesis, IND, RWTH Aachen, May 1997.

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