A PERCEPTUAL PERFORMANCE MEASURE FOR ADAPTIVE ECHO CANCELLERS IN PACKET-BASED TELEPHONY

J. D. Gordy and R. A. Goubran

Department of Systems and Computer Engineering, Carleton University
1125 Colonel By Drive, Ottawa, Canada K1S 5B6

ABSTRACT

This paper investigates performance measures of adaptive echo cancellers for packet-based telephony. It is shown that steady-state echo return loss enhancement (ERLE) does not accurately reflect perceived echo canceller convergence when background noise is present. An upper bound is derived for the maximum perceivable ERLE achievable in practice, and an algorithm is introduced for calculating ERLE that incorporates these masking effects based on a perceptual hearing model. Simulation and informal listening test results show a clear correspondence between the new performance measure and the perceptual upper bound induced by background noise.

1. INTRODUCTION

Adaptive echo cancellers are critical for removing annoying network and acoustic echoes in packet-based telephony and videoconferencing [1]. In these applications, a commonly used performance metric is echo return loss enhancement (ERLE), a simple measure of echo signal attenuation. An earlier study showed that the maximum ERLE achievable by an acoustic echo canceller is limited by under-modeling of the linear system and by loudspeaker non-linearity [2]. However, the study did not consider the effects of background noise on ERLE, nor did it investigate any relationship between ERLE and the *perceived* performance of an echo canceller. Recently psychoacoustics has seen increasing use in signal processing applications [3]. In the context of echo cancellation, postfiltering algorithms have been introduced that filter the error signal based on perceptual properties of near-end speech [4]. The idea is that near-end speech will tend to mask residual echo and background noise. However, in noisy environments without near-end speech, the presence of background noise will also mask residual echo.

This paper investigates the perceptual effects of background noise on ERLE when long round-trip delays are present in the network. In Section 2 an upper bound is shown for the maximum "perceptual" ERLE achievable by an echo canceller. A new algorithm for calculating ERLE that incorporates the perceptual effects of background noise is described in Section 3. Simulation and listening test results are shown in Section 4.

2. PERCEPTUAL ERLE

2.1. Echo canceller structure and conventions

A block diagram of a typical acoustic echo canceller is shown in Figure 1. Assume that the room impulse response can be perfectly modeled as a linear system with a finite impulse response (FIR) of length *N* with no nonlinearities introduced by the loudspeaker or microphone. In this case the echo signal *y*(*n*) is formed from the convolution of the far-end input signal *x*(*n*) with the room impulse response *h*(*n*). The reference signal *d*(*n*) recorded at the microphone includes the echo signal and near-end background noise *v*(*n*). In particular:

$$y(n) = x(n) \otimes h(n) \quad (1)$$

$$d(n) = y(n) + v(n) \quad (2)$$

The error signal *e*(*n*) at the output consists of the residual echo signal $\delta(n)$ and the near-end background noise:

$$\delta(n) = y(n) - y'(n) = x(n) \otimes [h(n) - h'(n)] \quad (3)$$

$$e(n) = \delta(n) + v(n) \quad (4)$$

where $h'(n)$ is an adaptive FIR filter of length $M \leq N$ samples. The residual echo signal is a result of the misadjustment between the room impulse response and the adaptive filter. In addition, the residual echo signal reflects the unmodeled "tail" if $M < N$. If the background noise is uncorrelated with the input signal, then the power spectrum of the reference signal is given by the sum of the echo signal and background noise power spectra. The error signal power spectrum is defined similarly:

$$S_{dd}(\omega) = S_{yy}(\omega) + S_{vv}(\omega) \quad (5)$$

$$S_{ee}(\omega) = S_{\delta\delta}(\omega) + S_{vv}(\omega) \quad (6)$$

0-7803-9332-5/05/$20.00 ©2005 IEEE

2.2. Reformulation of ERLE

A commonly used measure of echo canceller performance is echo return loss enhancement (ERLE), a broad measure of how much echo is reduced by the echo canceller. It is obtained from the ratio of the reference signal power to the error signal power, expressed in decibels:

$$ERLE = 10\log_{10}[\sigma_d^2 / \sigma_e^2] \qquad (7)$$

This equation assumes that the background noise $v(n)$ is sufficiently low that it can be ignored. However, in noisy environments it will skew the steady-state result. A more accurate measure is obtained by rewriting (7) in terms of the echo and residual echo signals (assumed available):

$$ERLE = 10\log_{10}[\sigma_y^2 / \sigma_\delta^2] \qquad (8)$$

The average signal powers in (8) can be expanded in terms of their power spectrum functions to obtain an alternative representation of ERLE. First define the difference $D(\omega)$ between the power spectrum functions of the echo and residual echo signals, then calculate ERLE as the difference function averaged over all frequencies:

$$D(\omega) = 10\log_{10}[S_{yy}(\omega) / S_{\delta\delta}(\omega)] \qquad (9)$$

$$ERLE = \frac{1}{2\pi}\int_{\omega=0}^{2\pi} D(\omega)d\omega \qquad (10)$$

It is important to note that (8) and (10) are not equivalent in general. However, (10) considers the contribution of the power spectrum functions at individual frequencies, which will be useful in the next section. In addition, (8) and (10) assume that the echo and residual echo signals can be measured, which is not the case in practice.

2.3. Perceptual limitations of ERLE

The annoyance of residual echo at the far end has traditionally been represented as a function of the round-trip delay [5]. However, during pauses in the far-end signal, and after long delays typical of mobile and packet-based telephony (≥ 200 ms), temporal masking effects have died out and the perceivability of the residual echo signal is more dependent on frequency masking effects of background noise. In particular, tonal and noise-like components in the background noise power spectrum tend to limit the audibility of residual echo around the same frequencies. In other words, the background noise induces a masking threshold $T_M(\omega)$ below which the residual echo signal will not be audible to the far-end listener [3]. A plot of the masking threshold in terms of sound pressure level (SPL) is shown in Figure 3 for background noise recorded in a small conference room. SPL is dependent upon the listening conditions at the far end. If $T_M(\omega)$ is properly calibrated, components of the residual echo below this threshold are not audible to the far-end listener. Therefore, by incorporating the masking threshold of the background noise signal, the maximum *perceptual* ERLE at each frequency ω is obtained by substituting $T_M(\omega)$ into (9):

$$D_{MAX}(\omega) = 10\log_{10}[S_{yy}(\omega) / T_M(\omega)] \qquad (11)$$

For a non-zero residual echo signal, the perceptual ERLE contribution at each frequency ω is obtained from the *maximum* of the residual echo power and the background noise masking threshold in (9):

$$D_P(\omega) = 10\log_{10}[S_{yy}(\omega) / \max\{S_{\delta\delta}(\omega), T_M(\omega)\}] \,(12)$$

Several important points can be observed from (9) – (12). Once the residual echo has been driven below the masking threshold of the background noise, no further *perceivable* improvement in echo cancellation can be achieved. It is also possible to have two echo cancellers producing the same ERLE when calculated using (7) or (8), but different perceptual ERLE using (12). As a result, the residual echo left by one echo canceller may be more *perceivable* at the far-end than the other.

3. CALCULATING THE PERCEPTUAL ERLE

3.1. Overview

A block diagram of the perceptual ERLE calculation is shown in Figure 2. At each time n, power spectrum estimates of the reference and error signals $d(n)$ and $e(n)$ are obtained from a windowed block of samples. Spectral subtraction is used to estimate the power spectrum of the echo and residual echo signals $y(n)$ and $\delta(n)$. The masking threshold is calculated from the background noise using a psychoacoustic model [5]. Finally, the estimates and masking threshold are used to calculate the perceptual ERLE from (10) and (12) for each block.

3.2. Implementation details

The background noise $v(n)$ is assumed to be stationary or slowly time-varying, and its power spectrum is estimated from the reference signal $d(n)$ during periods of quiet (no far-end speech). To that end, Welch's modified periodogram method is employed with $2K$-sample analysis blocks and a Hamming window applied [7]. Individual periodogram estimates are obtained from the discrete Fourier transform (DFT) of each block, and averaged over the set of L most recent blocks. Let $S_{vv}(k)$ represent the background noise power spectrum estimate, for $0 \le k \le K - 1$.

It is possible to estimate the power spectrum of $y(n)$ and $\delta(n)$ using (5) – (6) and spectral subtraction methods. First the power spectrum functions of the reference signal $d(n)$ and error signal $e(n)$ are estimated using the current windowed input block. They are represented by $S_{dd}(k)$ and $S_{ee}(k)$, respectively. One cannot employ averaging methods for these signals because real speech inputs can only be assumed to be stationary within periods of 20 – 30 ms. Let $S_{yy}(k)$ and $S_{\delta\delta}(k)$ represent the power spectrum functions of the echo and residual echo signals, respectively, estimated in accordance with (5) and (6):

$$S_{yy}(k) = \max\{S_{dd}(k) - S_{vv}(k), 0\} \qquad (13)$$

$$S_{\delta\delta}(k) = \max\{S_{ee}(k) - S_{vv}(k), 0\} \qquad (14)$$

The masking threshold $T_M(k)$, $0 \le k \le K - 1$, is calculated from $S_{vv}(k)$ using MPEG-1 Psychoacoustic Model 1 [3]. The model has been modified to accommodate the lower sampling rates more commonly associated with narrowband and wideband telephony (8 – 16 kHz). Finally, the perceptual ERLE for each block is calculated using discrete-time versions of (10) and (12):

$$D_P(k) = 10\log_{10}[S_{yy}(k) / \max\{S_{\delta\delta}(k), T_M(k)\}] \quad (15)$$

$$ERLE_P = \tfrac{1}{K}\sum_{k=0}^{K-1} D_P(k) \qquad (16)$$

4. SIMULATION RESULTS

4.1. Methodology

A room impulse response for a small conference room was measured and truncated to N = 2500 samples. Background noise from an overhead air conditioning fan in the room was recorded separately, and in all cases a sampling rate of f_s = 16 kHz was used. The reference signal $y(n)$ was formed by convolving a white Gaussian noise input signal $x(n)$ with the room impulse response, to which was added the recorded background noise $v(n)$. The reference signal was calibrated to have a power of 60 dB SPL. An adaptive echo canceller (M = 2500 taps) with the normalized least-mean-square (NLMS) algorithm was used to cancel the echo [6]. A step size of μ = 0.1 was employed, and data was collected during the initial convergence period of the adaptive filter. In particular, the reference and error signals $d(n)$ and $e(n)$ were measured directly, and the echo and residual echo signals $y(n)$ and $\delta(n)$ were obtained by subtracting the known background noise signal $v(n)$ from the former. In addition, informal listening tests were conducted with a panel of ten subjects to confirm the algorithm's validity.

4.2. Results and discussion

Figure 4(a) shows a plot of ERLE as a function of time, calculated using (7) and (8). It is clear that the presence of background noise limits the ERLE calculated using (7), and it reaches a steady state after approximately $n_1 \ge$ 30000 samples. However, when ERLE is calculated without the background noise using (8), the adaptive filter continues to adapt and finally reaches steady state after approximately $n_2 \ge$ 50000 samples. Figure 4(b) shows the residual echo signal power spectrum functions at times n_1 and n_2 along with the masking threshold of the background noise. From this plot it is clear that at time n_1 the residual echo is not below the masking threshold. As a result, some frequency components will be audible at the far end. At time n_2 the residual echo is far below the masking threshold, implying that it is inaudible at the far end. Therefore, neither (7) nor (8) determine the point at which the residual echo signal becomes inaudible.

Figure 5(a) shows a plot of ERLE as a function of time calculated using (9) and (12). Again it is clear that there is a difference between the steady state times revealed by these calculations. In particular, ERLE calculated using (12) reaches a steady state after approximately $n_1 \ge$ 35000 samples, whereas ERLE calculated using (9) reaches a steady state after approximately $n_2 \ge$ 45000 samples. Figure 5(b) again shows the residual echo signal power spectrum functions at times n_1 and n_2 along with the masking threshold of the background noise. From this plot it is clear that at time n_1 the residual echo is just at the masking threshold, and at time n_2 the residual echo is again far below the masking threshold. As a result, both residual echo signals are inaudible at the far end. Therefore, ERLE calculated using (12) can be used to determine the point at which the residual echo signal becomes inaudible.

Listening test results are shown in Table I, with subjects asked to detect residual echo perceivability over time using the corresponding adaptive filter coefficients and recorded background noise. It is important to note that at the perceptual ERLE steady state time (n = 35000), no subject reported a perceivable residual echo.

5. CONCLUSIONS

This paper introduced an algorithm for calculating ERLE incorporating the perceptual masking of background noise. It was shown that the proposed algorithm produces a more accurate measure of residual echo signal perceptibility than the traditional definition of ERLE.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from NSERC and CITO.

REFERENCES

[1] Y. Huang and R. A. Goubran, “Effects of vocoder distortion on network echo cancellation,” in *Proc. IEEE ICME*, July 2000, vol. 1, pp. 437 – 439.

[2] M. E. Knappe and R. A. Goubran, “Steady-state performance limitations of full-band acoustic echo cancellers,” in *Proc. IEEE ICASSP*, April 1994, vol. 2, pp. 73 – 76.

[3] ISO / IEC, JTC1/SC29/WG11 MPEG, “Information technology – coding of moving pictures and associated audio for digital storage media at up to 1.5 Mbit/s – Part 3: Audio,” IS11172-3, 1992.

[4] S. Gustafsson et al., “A psychoacoustic approach to combined acoustic echo cancellation and noise reduction,” *IEEE Trans. Speech Audio Processing*, vol. 10, no. 5, pp. 245 – 256, July 2002.

[5] International Telecommunications Union, *ITU-T G.131: Talker echo and its control*, ITU 2003.

[6] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 1996.

[7] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice Hall, 1989.

Table I – Proportion of subjects rating residual echo signal as perceivable against the background noise, versus time.

n	20000	25000	30000	35000
Ratio	9 / 10	5 / 10	2 / 10	0 / 10

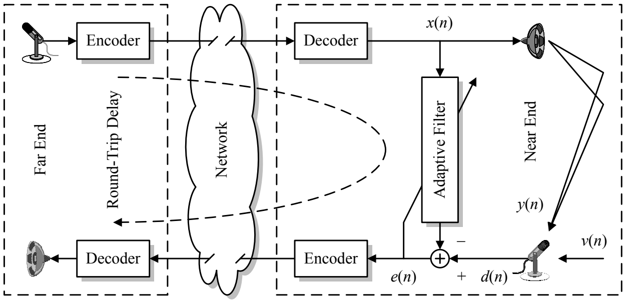

Figure 1 – An echo canceller in packet-based telephony.

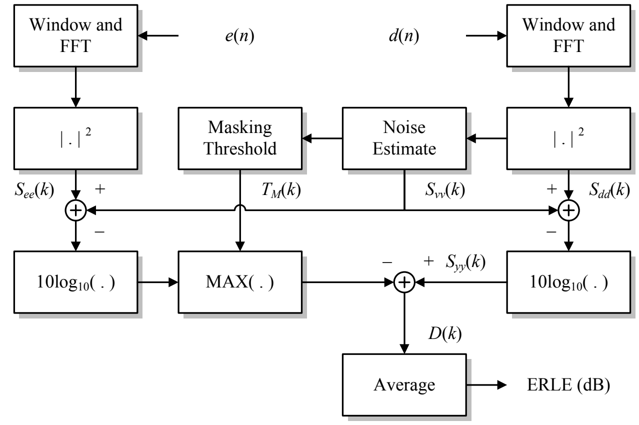

Figure 2 – Block diagram of perceptual ERLE calculation.

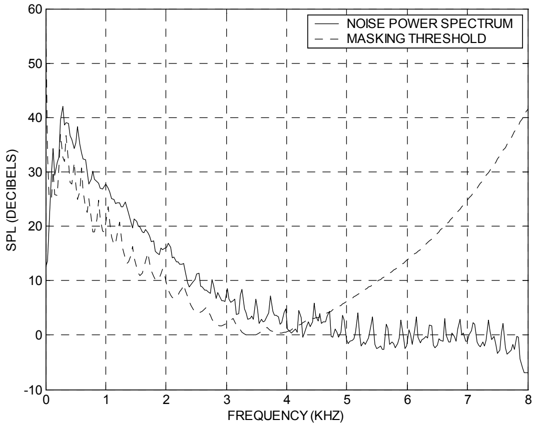

Figure 3 – Masking threshold $T_M(\omega)$ induced by conference room background noise.

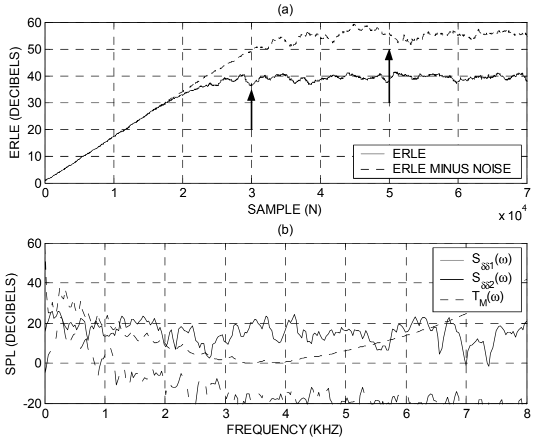

Figure 4 – (a) ERLE calculated using (7) and (8);
(b) Residual echo power spectrum at steady-state times.

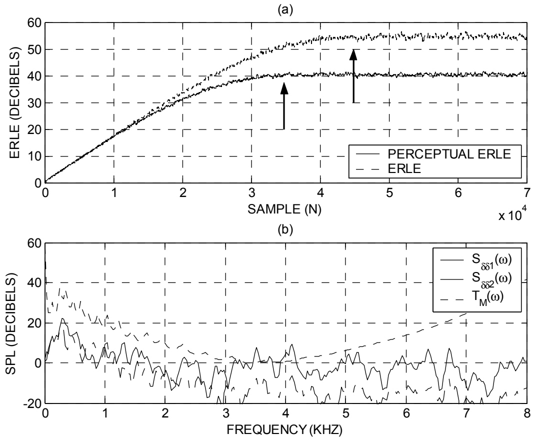

Figure 5 – (a) ERLE calculated using (9) and (12);
(b) Residual echo power spectrum at steady-state times.