POSTFILTERING FOR SUPPRESSION OF RESIDUAL ECHO FROM VOCODER DISTORTION IN PACKET-BASED TELEPHONY

James D. Gordy and Rafik A. Goubran

Department of Systems and Computer Engineering, Carleton University
1125 Colonel By Drive, Ottawa, Ontario, Canada K1S 5B6
{jdgordy, goubran}@sce.carleton.ca

ABSTRACT

This paper investigates postfiltering for residual echo suppression in networks employing low-bit-rate speech compression in the echo path. Simulations show that the residual echo from nonlinear vocoder distortion with ITU G.729 is proportional to the input signal LPC spectrum. An algorithm is proposed to estimate the residual echo power spectrum using a frequency-dependent scaling factor. The algorithm is incorporated into a psychoacoustic postfilter for residual echo suppression and compared to an existing estimator with a fixed scaling factor. Experiments with speech input and near-end signals show an average 0.85 dB lower spectral distortion and 0.4 higher estimated mean opinion score.

1. INTRODUCTION

A current trend in telecommunications is the migration of voice services from dedicated networks to integrated voice-and-data packet-switched networks. This has led to various interconnection topologies between next-generation networks (VoIP) and the legacy public switched telephone network (PSTN). An additional trend is the adoption of videoconferencing and mobile telephony with their inherent problem of acoustic echoes. Digital echo cancellers are typically deployed as close as possible to echo sources to mitigate their effects on speech quality. Unfortunately existing echo cancellers may not provide a sufficient level of cancellation given the increased round-trip delays introduced by VoIP and mobile networks [1]. One solution is to employ a *centralized* echo canceller at IP gateways providing additional echo cancellation or suppression. However, to reduce bandwidth in packet-switched and mobile networks, speech signals are compressed using low-bit-rate speech compression algorithms (vocoders) such as the ITU G.729 or G.722.2 standards [2] – [3]. As shown in Figure 1, vocoders introduce distortions into the network limiting the echo cancellation achievable [4]. Vocoder distortion appears at the far end as residual echo.

One approach to handling residual echo is to employ a frequency-domain postfilter to suppress residual echo while enhancing near-end speech. A psychoacoustic postfilter was proposed in [5] for suppressing residual echo in undermodeled acoustic echo cancellers. However, their approach employs linear models for estimating the residual echo power spectrum which are ineffective for the time-varying, nonlinear distortion introduced by vocoders. Recently a psychoacoustic postfilter was proposed to reduce the residual echo from vocoder distortion [6]. However, their structure employs a very simple model of the residual echo power spectrum which we improve.

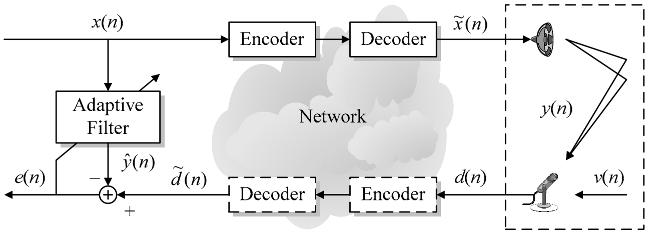

Figure 1 – Echo cancellation in a network with vocoder distortion introduced into the input and reference signals.

It is shown in this paper that psychoacoustic postfiltering depends on the accuracy of the near-end speech power spectrum estimate, which in turn depends on the accuracy of the estimated residual echo power spectrum. We introduce an improved method for estimating the residual echo power spectrum arising from nonlinear vocoder distortion. Psychoacoustic postfiltering for residual echo suppression is briefly reviewed in Section 2. In Section 3 we analyze the spectral characteristics of residual echo from nonlinear vocoder distortion, and describe the proposed technique for estimating the residual echo power spectrum. Experimental results are provided in Section 4.

2. RESIDUAL ECHO POSTFILTERING

2.1. Echo Canceller Structure and Conventions

A block diagram of the echo canceller is shown in Figure 1. The input signal $x(n)$ is transmitted through the network to the near end, where an echo path impulse response $h(n)$ of length N samples is applied. Distortions are introduced by the presence of vocoders along the send and, optionally, receive paths. Neglecting vocoder distortion for now, the reference signal $d(n)$ consists of the near-end speech signal $v(n)$ and the echo signal $y(n)$ as follows:

$$d(n) = v(n) + y(n) \tag{1}$$

$$y(n) = x(n) \otimes h(n) \tag{2}$$

where ⊗ denotes convolution. The echo path is modeled by the echo canceller as a finite impulse response $g(n)$ of length M samples updated using the normalized least-mean-square (NLMS) algorithm. The echo canceller error signal $e(n)$ consists of the near-end speech signal $v(n)$ and the residual echo signal $\delta(n)$:

1-4244-0367-7/06/$20.00 ©2006 IEEE

$$e(n) = d(n) - \hat{y}(n) = v(n) + \delta(n) \tag{3}$$

$$\delta(n) = x(n) \otimes [h(n) - g(n)] \tag{4}$$

In a practical echo canceller implementation, the structure of Figure 1 requires additional components which are out of the scope of this paper but still assumed to be present. For example, a doubletalk detector is required to sense the presence of near-end speech and halt adaptation of $g(n)$ for the duration of the disturbance [7]. If adaptation is not halted during such conditions, most adaptation algorithms tend to diverge after a few samples.

2.2. Postfiltering for Residual Echo Suppression

A block diagram of the postfilter is shown in Figure 2. The error signal $e(n)$ is transformed by an analysis stage into the frequency spectrum $E(\omega)$ consisting of the near-end speech $V(\omega)$ and residual echo $\Delta(\omega)$. A weighting $H(\omega)$ is applied to estimate the near-end speech spectrum before reconstruction by a synthesis stage:

$$E(\omega) = V(\omega) + \Delta(\omega) \tag{5}$$

$$\hat{V}(\omega) = H(\omega)E(\omega) = H(\omega)[V(\omega) + \Delta(\omega)] \tag{6}$$

A key question is how to construct $H(\omega)$. As shown in Figure 2 we follow [5] which employs a *psychoacoustic* model to minimize audible near-end speech distortion. In this approach a preliminary estimate of the near-end speech power spectrum is constructed from the error signal using an estimate of the residual echo power spectrum. The masking threshold $T_M(\omega)$ of the near-end speech is then obtained using a psychoacoustic model such as [8]. Finally, $H(\omega)$ is constructed under the assumption that residual echo below the masking threshold will be inaudible to the listener [9].

Define a cost function $J(\omega)$ as the squared error between the true and estimated near-end speech spectra. Assuming statistical independence between the near-end speech and residual echo, $J(\omega)$ can be written as the sum of two cost functions $J_V(\omega)$ and $J_\Delta(\omega)$:

$$J(\omega) = [V(\omega) - \hat{V}(\omega)]^2 = J_V(\omega) + J_\Delta(\omega) \tag{7}$$

$$J_V(\omega) = [1 - H(\omega)]^2 S_{VV}(\omega) \tag{8}$$

$$J_\Delta(\omega) = H^2(\omega) S_{\Delta\Delta}(\omega) \tag{9}$$

where $S_{VV}(\omega)$ and $S_{\Delta\Delta}(\omega)$ are the power spectra of the near-end speech and residual echo, respectively. $J_V(\omega)$ and $J_\Delta(\omega)$ represent the distortion of the near-end speech and residual echo, respectively. Minimizing $J_V(\omega)$ such that that $J_\Delta(\omega)$ is at the masking threshold $T_M(\omega)$ of the near-end speech results in a real-valued transfer function given by [5]:

$$H(\omega) = \max\left\{\sqrt{T_M(\omega)/S_{\Delta\Delta}(\omega)}, 1\right\} \tag{10}$$

In order to accurately construct the masking threshold of the near-end speech, first it is important to have a "good" estimate of the residual echo power spectrum. From (10) it is clear that inaccuracies in $T_M(\omega)$ or $S_{\Delta\Delta}(\omega)$ may lead to audible distortion of the near-end speech and / or insufficient residual echo suppression. In the sequel we discuss the problem of estimating the residual echo power spectrum resulting from nonlinear vocoder distortion.

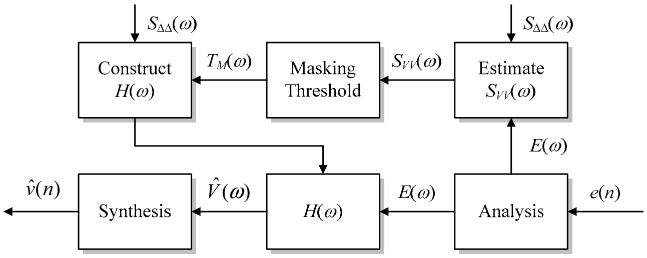

Figure 2 – Block diagram of a psychoacoustic postfilter for suppressing residual echo from near-end speech.

3. RESIDUAL ECHO POWER SPECTRUM ESTIMATION

3.1. Effect of Vocoder Distortion

Low-bit-rate speech encoders introduce nonlinear distortion into speech signals through coding of parameters such as LPC coefficients, pitch period and gains, and excitation signal modeling [2] – [3]. In addition, decoders often employ a harmonic postfilter stage which introduces further nonlinearity. In this section we consider the configuration of Figure 1 with a cascade of vocoders present in the network *only* along the send path. In this case vocoders introduce distortion into the input signal $x(n)$, and from (1) and (2) the result can be written as the sum of contributions from clean and distortion signals:

$$\widetilde{x}(n) = x(n) + x_{NL}(n) \tag{11}$$

$$\widetilde{d}(n) = d(n) = [x(n) + x_{NL}(n)] \otimes h(n) + v(n) \tag{12}$$

where $x_{NL}(n)$ represents the nonlinear distortion in the input signal. From (3) and (4), the resulting error signal and residual echo signals can be represented as follows:

$$e(n) = v(n) + \delta(n) \tag{13}$$

$$\delta(n) = x(n) \otimes [h(n) - g(n)] + x_{NL}(n) \otimes h(n) \tag{14}$$

To effectively employ the postfilter described in Section 2.2, it is necessary to have a power spectrum estimate for the residual echo signal of (14). This issue is addressed in the following section.

3.2. Residual Echo Power Spectrum Estimation

Methods exist for estimating the power spectrum of residual echo caused by adaptive filter misadjustment and undermodeling [10]. These are applicable for the first term of (14). Unfortunately, the distortion signal $x_{NL}(n)$ in the second term resulting from vocoders is difficult to model due to the complexity of low-bit-rate speech compression algorithms. However, encoders typically minimize a perceptually weighted squared error, allowing greater error in high-energy (formant) regions corresponding to peaks in the spectrum, and lower error in non-formant regions. The weighting filter $W(z)$ is constructed from the input signal LPC spectrum with:

$$W(z) = A(z/\gamma_1)/A(z/\gamma_2) \tag{15}$$

where $A(z)$ is the input signal LPC spectrum and γ_1 and γ_2 are parameters controlling the weighting, with typical values of $\gamma_1 = 1$ and $\gamma_2 = 0.9$. To investigate the spectral characteristics of (14), we examined the steady-state conditions of the echo canceller from Figure 1 with the ITU G.729 encoder and decoder cascaded along the send path [2]. An echo path N = 250 samples in length was modeled with an appropriate filter (N = M plus codec delay) adapted using NLMS. A stationary 10th-order autoregressive process was used as a speech-like input signal. Figure 3(a) shows a plot of the corresponding echo return loss enhancement (ERLE) as a function of time. In this case the vocoder distortion limits the achievable ERLE to approximately 9 dB. At this point we froze the echo canceller adaptation and measured the power spectrum functions for the two components of (14) for typical speech signals. Figure 3(b) shows power spectrum of the residual echo signal for a frame of speech, along with spectra of its constituent adaptive filter misadjustment and nonlinear distortion components. Although the contribution from misadjustment is still significant, it is clear that the residual echo is dominated by the contribution of the nonlinear distortion. For this paper we make the assumption that once the adaptive filter has converged as much as possible, the residual echo signal is dominated by $x_{NL}(n)$ and approximated as:

$$\delta(n) \approx x_{NL}(n) \otimes h(n) \qquad (16)$$

Figure 4(a) shows the power spectra for the echo signal $y(n)$ and residual echo signal $\delta(n)$ corresponding to a frame of voiced speech, while Figure 4(b) shows the power spectra for an unvoiced speech frame. Superimposed are the corresponding 10th-order LPC spectra of the input signal $x(n)$. For voiced speech the residual echo power spectrum is generally at the *same* level as that of the echo at high-energy (formant) frequencies. At non-formant frequencies and for unvoiced speech frames, the residual echo power spectrum is generally uniformly below the echo power spectrum. Figure 4(c) shows the ratio of residual echo to echo power spectra as a function of the normalized LPC spectrum averaged over a large number of inputs. A distinct linear trend is visible over most of the spectrum.

Given these observations it is reasonable to approximate the residual echo power spectrum as a scaled version of the *estimated* echo power spectrum once the adaptive filter has converged sufficiently. To accommodate the observed relationship with the LPC spectrum, we estimate the power spectrum of (16) using the estimated echo power spectrum and a frequency scaling factor:

$$S_{\Delta\Delta}(\omega) \approx \mu(\omega) S_{XX}(\omega) \,|\, \hat{H}(\omega) \,|^2 = \mu(\omega) S_{\hat{Y}\hat{Y}}(\omega) \qquad (17)$$

Let $A(\omega)$ represent the LPC spectrum of the input signal and let A_{MAX} and A_{MIN} be the corresponding maximum and minimum values. The scaling factor $\mu(\omega)$ is modeled as a linear function dependent on the normalized LPC spectrum:

$$\mu(\omega) = \alpha \frac{A(\omega) - A_{MIN}}{A_{MAX} - A_{MIN}} [\mu_{MAX} - \mu_{MIN}] + \mu_{MIN} \qquad (18)$$

where μ_{MAX} and μ_{MIN} are maximum and minimum scaling factors and $\alpha = \{1, 0\}$ indicates voiced and unvoiced states, respectively, for the current block. It was found that attenuation could be estimated adaptively during periods of no near-end speech by calculating the maximum and minimum attenuation with a smoothing factor $0 < \lambda < 1$ and the power spectra of the error and estimated echo signals:

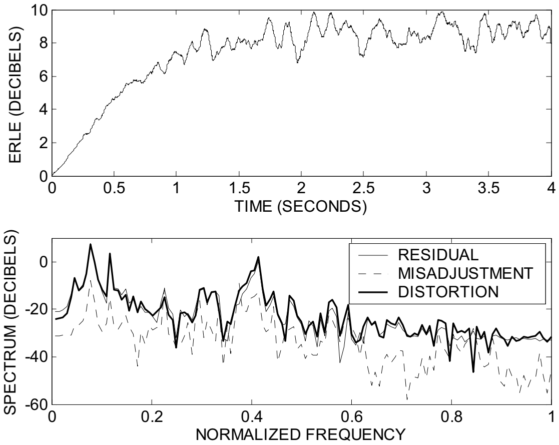

Figure 3 – (a) ERLE with vocoder distortion in the input signal; (b) Residual echo power spectrum compared to contributions from misadjustment and nonlinear distortion.

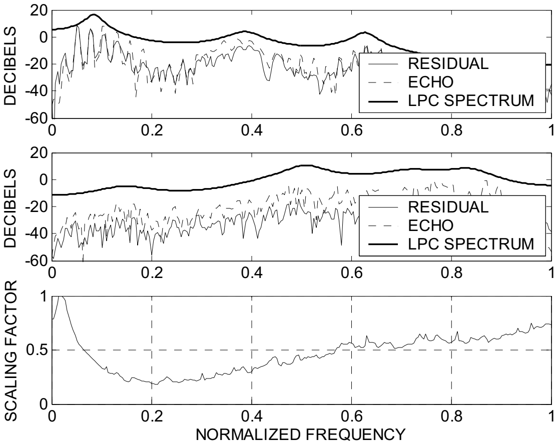

Figure 4 – Echo and residual echo power spectra for (a) voiced and (b) unvoiced speech frames; (c) Average residual echo scaling factor as a function of LPC spectrum magnitude.

$$\mu_{MAX} = \lambda \mu_{MAX} + (1-\lambda) \max_{\omega} [S_{\hat{Y}\hat{Y}}(\omega) / S_{EE}(\omega)] \qquad (19)$$

$$\mu_{MIN} = \lambda \mu_{MIN} + (1-\lambda) \min_{\omega} [S_{\hat{Y}\hat{Y}}(\omega) / S_{EE}(\omega)] \qquad (20)$$

4. SIMULATION RESULTS

Experimental results of the proposed residual echo power spectrum estimation method were obtained by incorporating it into the postfilter of Section 2. The analysis and synthesis stages were implemented using the Fast Fourier Transform (FFT) on 256-sample blocks with 50 percent overlap and zero-padded to 512 samples. The preliminary near-end speech power spectrum estimate was obtained with the MMSE-LSA estimator using the estimated residual echo power spectrum [5]. The masking threshold was calculated using the MPEG-1 Psychoacoustic Model 1 modified for a sampling rate of 8 kHz [8]. Input and near-end speech signals were obtained from the TIMIT database [11]. For these experiments the environment from Section 3.2 was employed again. The power spectrum estimation algorithm was evaluated by measuring the spectral distortion between the original and estimated near-end speech signals:

$$SD^2 = \frac{1}{2\pi}\int_{\omega=0}^{2\pi}[10\log_{10}P(\omega)-10\log_{10}P'(\omega)]^2\,d\omega \qquad (21)$$

where $P(\omega)$ and $P'(\omega)$ are the LPC spectra of the original and estimated near-end speech signals for the current block, respectively. We also evaluated the quality of the estimated near-end speech using the mean opinion score estimate provided by ITU-T P.862 [12]. For comparison we implemented the algorithm in [6] which employs a *fixed* scaling factor to approximate the residual echo power spectrum. To find the maximum improvement possible with postfiltering, we also considered the (idealized) case where the power spectrum of the residual echo signal, including the adaptive filter misadjustment, is known.

Table I shows the average spectral distortion and estimated mean opinion score (MOS) for near-end speech after postfiltering with the three test configurations. The proposed method results in an average 0.85 dB lower spectral distortion and an average 0.4 higher estimated MOS, which is clearly an improvement over the fixed scaling factor. Informal listening tests confirmed that there was less residual echo and less distortion of the postfiltered near-end speech using the proposed algorithm. To illustrate the improvement afforded by postfiltering, Figures 5(a) and 5(b) show spectrograms of the original and estimated near-end speech signals produced using the proposed residual echo power spectrum estimator. Figure 5(c) shows a plot of ERLE during singletalk conditions with postfiltering using the proposed and fixed scaling factor algorithms and compared to no postfiltering. The plot also shows an improvement of 5 – 7 dB for the proposed algorithm over the estimator employing a fixed scaling factor.

5. CONCLUSIONS

An investigation of residual echo due to vocoder distortion with ITU G.729 was performed. An algorithm was described for estimating the residual echo power spectrum as part of a psychoacoustic postfilter, and shown to produce higher echo suppression and less near-end speech distortion. Further work is required to model the more complicated case of vocoder distortion in the receive path.

REFERENCES

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

[2] ——, *ITU-T G.729: Coding of speech at 8 kbit/s using CS-ACELP*, ITU 1996.

[3] ——, *ITU-T G.722.2: Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*, ITU 2001.

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

[5] 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, Jul. 2002.

[6] X. Lu and B. Champagne, "A centralized acoustic echo canceller exploiting masking properties of the human ear," in *Proc. IEEE ICASSP*, Apr. 2003, vol. 5, pp. 377 – 380.

[7] J. D. Gordy and R. A. Goubran, "A low-complexity doubletalk detector for echo cancellers in packet-based telephony," in *Proc. IEEE WASPAA*, Oct. 2005, pp. 74 – 77.

[8] 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.

[9] J. D. Gordy and R. A. Goubran, "A perceptual performance measure for adaptive echo cancellers in packet-based telephony," in *Proc. IEEE ICME*, Jul. 2005, vol. 1, pp. 431 – 434.

[10] G. Enzner, R. Martin and P. Vary, "Unbiased residual echo power estimation for hands-free telephony," in *Proc. IEEE ICASSP*, May 2002, vol. 2, pp. 1893 – 1896.

[11] J. Garofolo et al., *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM.* National Institute of Science and Technology, 1990.

[12] International Telecommunication Union, *ITU-T P.862: Perceptual evaluation of speech quality (PESQ)*, ITU 2001.

Table I – Average spectral distortion and estimated MOS of postfiltered near-end speech compared to a fixed scaling factor ([6]) and compared to the ideal case with known residual echo.

Test	Proposed		Fixed Scaling		Ideal	
Pair	SD	MOS	SD	MOS	SD	MOS
1	2.23	3.38	2.90	3.08	1.22	3.96
2	3.37	2.87	4.69	2.26	2.11	3.17
3	3.55	2.78	4.73	2.33	2.24	3.23
4	2.66	3.26	3.30	2.92	1.70	3.59
5	2.90	2.94	3.86	2.47	1.79	3.27
6	3.15	2.65	4.12	2.25	1.97	3.15
7	2.29	3.19	3.04	3.00	1.38	3.47
8	3.76	3.12	4.00	2.71	2.29	3.38
9	2.84	2.79	3.90	2.43	1.83	3.29
10	2.50	3.15	3.51	2.87	1.74	3.45

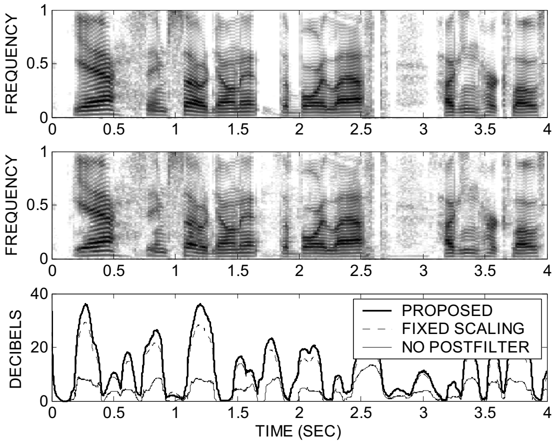

Figure 5 – (a) Original and (b) estimated near-end speech spectrograms using the proposed algorithm; (c) ERLE using proposed and fixed scaling factors compared to no postfiltering.