A COMBINED LPC-BASED SPEECH CODER AND FILTERED-X LMS ALGORITHM FOR ACOUSTIC ECHO CANCELLATION

J. D. Gordy and R. A. Goubran

Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada

ABSTRACT

This paper presents a novel acoustic echo canceller structure based on combining the Filtered-X LMS algorithm with an LPC-based speech coder for use in videoconferencing and VoIP. The algorithm updates coefficients using filtered versions of the input and error signals obtained by directly tapping the short-term excitation signal from the speech decoder, and by filtering the error signal with a bank of FIR decorrelation filters constructed from the LPC synthesis filter coefficients. The proposed algorithm was implemented using ITU G.729, and simulation results with 2000-tap room impulse responses show a faster and more constant rate of convergence than NLMS using speech input signals and an average 10 dB greater ERLE observed during convergence.

1. INTRODUCTION

A key requirement in Voice-over-Internet-Protocol (VoIP) and videoconferencing applications is the ability to transmit toll-quality speech at very low bit rates. This requirement is satisfied by the use of code excited linear prediction (CELP) based speech compression standards such as ITU G.729 and the more recent ITU G.722.2 [1], [2]. In such applications, speech signals are typically compressed and decompressed at the interface between a communications network and IP-enabled handsets, desktop speakers and microphones, and videoconferencing room equipment. However, such systems are still subject to the problem of acoustic echoes caused by mechanical coupling between a speaker and microphone or, more commonly, by acoustic coupling and room reflections. To compensate for these effects in the near-end signal, we know that the optimal location for an adaptive echo canceller is at the network interface after the far-end speech signal has been decompressed and before the near-end signal is compressed [3].

The stochastic gradient algorithms and variants such as the normalized least-mean-square (NLMS) remain the most commonly used techniques for acoustic echo cancellation because of their simplicity and low computational complexity [4]. However, the convergence rate of these algorithms is dependent upon the input signal's autocorrelation function, and generally degrades in the presence of colored input signals such as speech. In addition, for long impulse responses encountered in acoustic echo cancellation the convergence time of these algorithms is unacceptably low [5]. To improve the rate of convergence, there exist many recently proposed techniques for improving the convergence rate such as the computationally expensive frequency domain adaptive filtering (FDAF) [6], employing lower complexity affine projection algorithms (APA) [7], and adaptively varying the NLMS step size parameter [8]. In [9] a modified version of the Filtered-X LMS algorithm is introduced whereby the input signal is whitened using linear prediction, but that technique uses a single decorrelation filter the same length as the target impulse response that has to be updated at every sample period. In this paper we build upon the algorithm in [9] by employing a bank of short decorrelation filters whose coefficients are obtained from an LPC-based compressed representation of the input signal, which is typically available at the network interface mentioned above in a VoIP or videoconferencing system.

The paper is organized as follows. In Section 2 we describe our combined LPC-based speech coder and echo canceller structure. This is followed in Section 3 with simulation results of the proposed algorithm applied to acoustic echo cancellation both in a stationary environment and in the presence of a changing room impulse response.

2. COMBINED ECHO CANCELLER STRUCTURE

2.1. Overview

A block diagram of the echo canceller structure is shown in Figure 1. A stream of compressed speech frames is received at a network interface and the input signal $x(n)$ is reconstructed using an LPC-based speech decoder. An adaptive filter $h'_n(j)$ is used to compensate for the acoustic channel and produces an error signal $e(n)$. However, the adaptation algorithm employs the short-term excitation signal from the speech decoder as an input signal, and preprocesses the error signal with a bank of FIR filters constructed from the current and previous sets of LPC synthesis filter coefficients from the speech decoder. We describe the system in more detail in the following subsections.

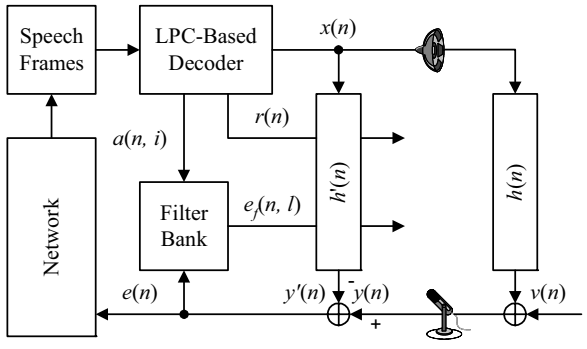

Figure 1 – Block diagram of echo canceller structure

2.2. LPC-based speech synthesis model

Most low-bit-rate LPC-based speech coders segment signals into frames of 10 – 30 ms in duration, and for each frame determine a representative set of parameters which are quantized and sent to the destination decoder. To reduce encoder complexity the parameters are often updated for smaller subframes of 5 – 7.5 ms in duration.

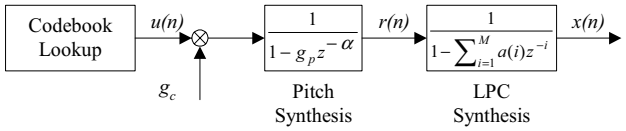

Figure 2 – LPC-based speech synthesis model

Figure 2 shows a typical speech synthesis model used by an LPC-based decoder. An excitation signal $u(n)$ is obtained from a structured or stochastic codebook and scaled by a gain parameter $g_c(n)$. This scaled excitation signal is spectrally shaped by an all-pole pitch synthesis filter represented by a single pitch lag $\alpha(n)$ and gain $g_p(n)$ to form the short-term excitation signal $r(n)$. Finally, an M^{th}-order all-pole LPC synthesis filter is applied, represented by the set of coefficients $a(n, i)$ for $1 \le i \le M$. Using this model the reconstructed speech signal $x(n)$ is given by:

$$x(n) = \sum_{i=1}^{M} a(n,i)x(n-i) + r(n) \quad (1)$$

$$r(n) = g_p(n)r(n-\alpha(n)) + g_c(n)u(n) \quad (2)$$

Note that the parameters are depicted here as time-varying in n, but in practice they are constant for the duration of a subframe.

2.3. Filtered-X LMS algorithm

We base our filter adaptation on a reformulated version of the Filtered-X LMS algorithm introduced in [9], as shown in Figure 1 and reviewed as follows. Let $\Delta h_n(j) = h(j) - h'_n(j)$ be the difference between the target and estimated impulse responses at time n. Combining this with (1) gives the error signal as a function of the current and previous sets of LPC synthesis filter coefficients over the duration of the target impulse response:

$$\begin{aligned} e(n) &= \sum_{j=0}^{N-1} \Delta h_n(j)x(n-j) + v(n) \\ &= \sum_{j=0}^{N-1} \Delta h_n(j) \sum_{i=1}^{M} a(n-j,i)x(n-i-j) + r(n-j) \\ &\quad + v(n) \end{aligned} \quad (3)$$

where N is the length of the impulse response and $v(n)$ is uncorrelated additive noise. In [9] an adaptive FIR decorrelation filter $f(n)$ is applied to the input $x(n)$ and the error $e(n)$ to form the filtered input and filtered error, respectively:

$$x_f(n) = \sum_{k=0}^{L-1} f(k)x(n-k) \quad (4)$$

$$e_f(n) = \sum_{k=0}^{L-1} f(k)e(n-k) \quad (5)$$

where L is the length of $f(n)$. These filtered signals are employed in the normalized LMS filter coefficient update instead of the original input and error signals. Ideally $x_f(n)$ is completely decorrelated and $e_f(n)$ is proportional to the convolution of the filtered input and the desired impulse response:

$$e_f(n) \approx \sum_{j=0}^{N-1} \Delta h_n(j)x_f(n-j) + \sum_{k=0}^{L-1} f(k)v(n-k) \quad (6)$$

In this case the expected value of the stochastic gradient estimate in the LMS filter coefficient update is directly proportional to the current filter coefficient difference:

$$\begin{aligned} E\left[e_f(n)x_f(n-l)\right] &= \sum_{j=0}^{N-1} \Delta h_n(j) E\left[x_f(n-j)x_f(n-l)\right] \\ &= \sum_{j=0}^{N-1} \Delta h_n(j)\sigma_x^2 \delta(l) \\ &= \Delta h_n(l)\sigma_x^2 \end{aligned} \quad (7)$$

where σ_x^2 is the variance of $x_f(n)$, and $0 \le l \le N - 1$. This is equivalent to filtering uncorrelated noise, and the convergence rate is no longer dependent upon the eigenvalue spread of the input signal's correlation matrix [4].

From (1) it is clear that a good choice for the decorrelation filter is the inverse of the all-pole LPC synthesis filter at time n:

$$f(0) = 1; \quad f(k) = -a(n,k) \quad 1 \le k \le M \quad (8)$$

Substituting (1) and (8) into (4) shows that a good candidate for the filtered input signal $x_f(n)$ is simply the short-term excitation signal tapped from the speech decoder:

$$\begin{aligned} x_f(n) &= \sum_{k=0}^{L-1} f(k)\left[\sum_{i=1}^{M} a(n-k,i)x(n-i-k) + r(n-k)\right] \\ &= \sum_{k=1}^{M} \left[a(n,k) - a(n,k)\right]x(n-k) + r(n) \\ &= r(n) \end{aligned} \quad (9)$$

Recall that the short-term excitation signal is formed from a "random" codebook signal and a single-tap pitch synthesis filter. Therefore we predict that $r(n)$ will be more decorrelated than $x(n)$, an assumption we investigate in Section 3.

From (3) we see that $e(n)$ is a function of the current and previous LPC synthesis filter coefficients which vary considerably in time, particularly over the duration of impulse responses common in acoustic echo cancellation ($N \ge 2000$). Therefore, we define a new filtered error signal $e_f(n, l)$ that is a function of both the current time n and the desired filter tap l to be updated in the LMS coefficient update equation. In this case, we apply a bank of decorrelation filters constructed by applying the inverse of the all-pole LPC synthesis filter in effect at the decoder at time $n – l$:

$$f(0) = 1; \quad f(k) = -a(n-l,k) \quad 1 \le k \le M \quad (10)$$

Substituting (10) into (3) gives us the new filtered error signal $e_f(n,l)$ for use in the LMS filter coefficient update equation:

$$
\begin{aligned}
e_f(n,l) \;=\; & \sum_{j=0}^{N-1} \Delta h_n(j) \cdot \\
& \left[\sum_{i=1}^{M} a(n-j,i)x(n-i-j) + r(n-j) \right] - \\
& \sum_{k=1}^{M} a(n-l,k) \left\{ \sum_{j=0}^{N-1} \Delta h_{n-k}(j) \cdot \right. \\
& \left. \left[\sum_{i=1}^{M} a(n-j-k,i)x(n-i-j-k) + r(n-j-k) \right] \right\} + \\
& v(n) - \sum_{k=1}^{M} a(n-l,k)v(n-k)
\end{aligned}
\tag{11}
$$

$$
h'_{n+1}(l) = h'_n(l) + \frac{\mu e_f(n,l) r(n-l)}{\sum_{k=0}^{N-1} r^2(n-k)} \tag{12}
$$

where μ is the LMS step size parameter and $0 \le l \le N-1$. In order to simplify (11) we make the following assumptions. A practical vocoder fixes the LPC synthesis filter coefficients for the duration of a subframe, and typically $M = 10$. Since $M << N$ for long impulse responses, then the LPC synthesis filter is approximately constant in the short term: $a(n-j-k, i) \approx a(n-j, i)$ for $1 \le k \le M$ and $0 \le j \le N-1$. Furthermore, if $\mu << 1$ then $\Delta h_{n-k}(j) \approx \Delta h_n(j)$ for $1 \le k \le M$, and it can be shown that $e_f(n, l)$ assumes the simplified form in accordance with (6) and (7):

$$
e_f(n,l) \approx \sum_{j=0}^{N-1} \Delta h_n(j) r(n-j) + \sum_{k=0}^{L-1} f(k) v(n-k) \tag{13}
$$

Note that the proposed algorithm differs from a conventional pre-whitened NLMS ([7]) in that the presence of LPC synthesis filter coefficients at the speech decoder avoids the need to calculate decorrelation filter coefficients. In addition, a bank of such filters is employed to produce a coefficient-varying filtered error signal for use in the LMS coefficient update equation.

2.4. Computational complexity

If the short term excitation signal can be extracted directly from the decoder, then no computation is required to compute $x_f(n)$ from the input signal. Calculating $e_f(n,l)$ would require NM multiplications per sample if the LPC synthesis filter coefficients were time varying in n. In practice the filter is constant for the duration of a subframe and interpolated within frames of F samples, so an approximation is to use the per-frame LPC coefficients to calculate $e_f(n, l)$ in F-sample blocks. In this case only $ceil(N / F)$ filters are necessary to calculate $e_f(n, l)$ per sample period, each with a cost of M multiplications. For example, ITU G.729 uses $F = 80$ and $M = 10$, so for $N = 2000$ the algorithm requires $(2000 / 80) \times 10 = 250$ multiplications per sample to calculate $e_f(n, l)$.

3. SIMULATION RESULTS

3.1. Methodology

The proposed algorithm was implemented using the LPC-based speech coder defined in ITU G.729 [1]. The reference code was modified to extract the short-term excitation signal $r(n)$ and LPC synthesis filter coefficients $a(n, i)$ from the decoder, and the speech signal $x(n)$ was calculated with postfiltering disabled. Tests were conducted using the room impulse responses shown in Figure 3 consisting of $N = 2000$ samples (250 ms). The performance was compared to NLMS, and to pre-whitened NLMS employing a first-order decorrelation filter [7]. A step size of $\mu = 0.1$ was employed for all algorithms, and performance measured using the system distance and echo return loss enhancement (ERLE), defined respectively by [5]:

$$
DIST(n) = 10 \log_{10} \left[\sum_{j=0}^{N-1} \Delta h_n(j) \Big/ \sum_{j=0}^{N-1} h(j) \right] \tag{14}
$$

$$
ERLE(n) = 10 \log_{10} \left\{ E[y^2(n)] \Big/ E[e^2(n)] \right\} \tag{15}
$$

Test input consisted of continuous speech from the DARPA TIMIT database [10] formed by concatenating training sequences from a male speaker downsampled to 8 kHz and compressed using the ITU G.729 encoder. White noise was added to the near-end signals to produce a segmental signal-to-noise ratio (SEGSNR) of 45 dB.

3.2. Statistics of short-term excitation signal

In order to verify that the short-term excitation signal is more decorrelated than the reconstructed speech signal as assumed earlier, we estimated the autocorrelation matrices of $r(n)$ and $x(n)$ for segments of 250 samples (~30 ms) with 50-sample overlap, and for each matrix the condition number was determined. Note that for a completely uncorrelated signal the condition number is one. Figure 4 shows a plot of these results, revealing that $r(n)$ has a very low average condition number, and as such is a good candidate for the filtered input signal $x_f(n)$.

3.3. Convergence and tracking performance

Figure 5 shows a plot of the system distance and ERLE as a function of time for the first room impulse response. It is clear that the proposed algorithm produces a more constant rate of convergence with respect to system distance, which is in agreement with the theoretical performance shown in (7) and (13). Note also that after ten seconds of adaptation the system distance of the proposed algorithm is 10 dB lower than NLMS and pre-whitened NLMS. With respect to ERLE, the proposed algorithm achieves an average 10 dB improvement in ERLE during the convergence phase, approaching the steady-state bound of 45 dB faster than NLMS and pre-whitened NLMS.

Figure 6 shows the system distance and ERLE as a function of time after switching from the first to second room impulse responses after ten seconds of adaptation. It is clear from this plot that the depth and rate of convergence of the proposed algorithm is similar during both initial convergence and after the change in impulse response. In addition, the improvement in ERLE remains 5 dB on average during re-convergence.

4. CONCLUSIONS

A novel echo canceller was described combining a Filtered-X LMS algorithm with an LPC-based speech coder. Simulation using the ITU G.729 standard shows a more constant and faster

rate of convergence than NLMS and pre-whitened NLMS when applied to room impulse responses, with a greater average ERLE observed during both initial convergence and while tracking. In closing, we note that the algorithm is applicable to other LPC-based vocoders including the recent ITU G.722.2 standard.

5. ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada.

6. REFERENCES

[1] International Telecommunications Union, *ITU-T G.729: Coding of speech at 8 kbit/s using CS-ACELP*, ITU 1996.

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

[3] V. V. Krishna, J. Rayala, and B. Slade, "Algorithmic and implementation aspects of echo cancellation in packet voice networks," *Proc. 36th Asilomar Conf.*, vol. 2, pp. 1252 – 1257, Nov. 2002.

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

[5] J. Homer, R. R. Bitmead and I. Mareels, "Quantifying the effects of dimension on the convergence rate of the LMS adaptive FIR estimator," *IEEE Trans. Signal Processing*, vol. 46, no. 10, pp. 2611 – 2615, Oct. 1998.

[6] K. Eneman and M. Moonen, "Iterated partitioned block frequency-domain adaptive filtering for acoustic echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 2, Mar. 2003.

[7] Breining et al, "Acoustic echo control: An application of very-high-order adaptive filters," *IEEE Signal Processing Mag.*, vol. 16, no 4, pp. 42 – 69, Jul. 1999.

[8] S. Emura and Y. Haneda, "A method of coherence-based step-size control for robust stereo echo cancellation," *Proc. IEEE ICASSP'03*, vol. 5, pp. 592 – 595, Apr. 2003.

[9] M. Mboup, M. Bonnet and N. Bershad, "LMS coupled adaptive prediction and system identification: a statistical model and transient mean analysis," *IEEE Trans. Signal Processing*, vol. 42, no. 10, pp. 2607 – 2614, Oct. 1994.

[10] J. Garofolo et al, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM.* NIST, 1990.

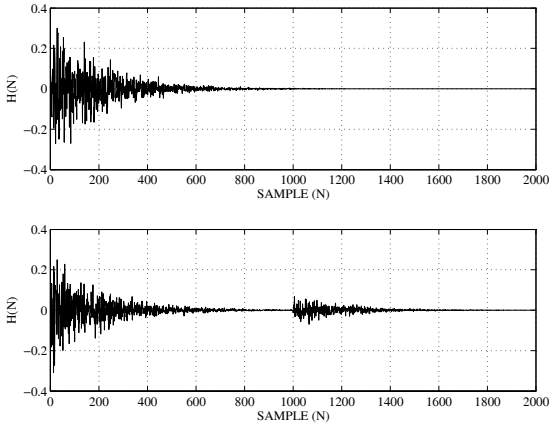

Figure 3 – Plot of two test room impulse responses

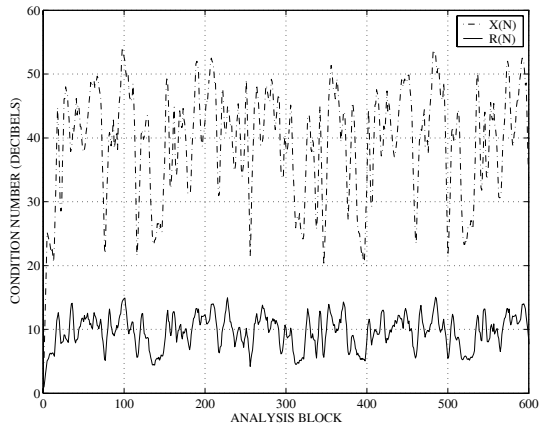

Figure 4 – Autocorrelation matrix condition number as a function of time for $x(n)$ and $r(n)$

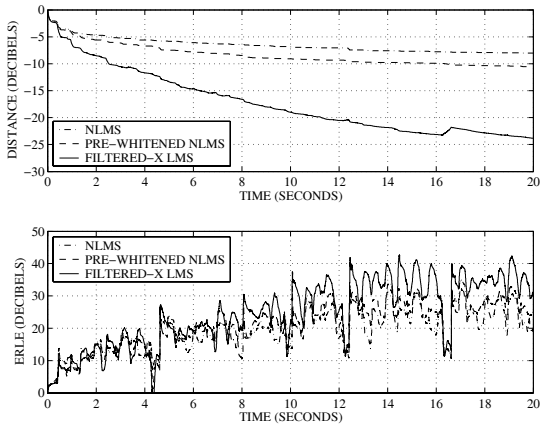

Figure 5 – System distance and ERLE as a function of time for the test signal applied to the first room impulse response

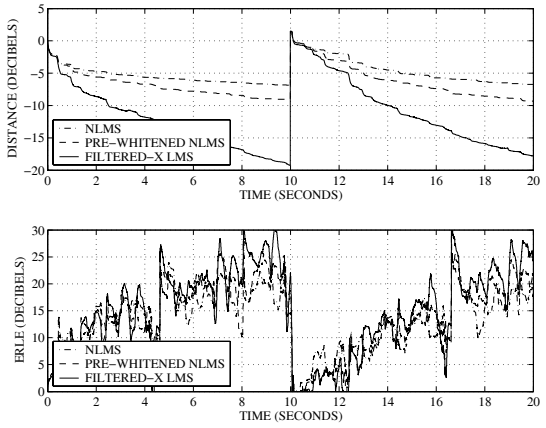

Figure 6 – System distance and ERLE as a function of time for tracking between the first and second impulse responses