PITCH ANALYSIS-BASED ACOUSTIC ECHO CANCELLATION OVER A NONLINEAR CHANNEL

Xiaojian Lu and Benoît Champagne
Department of Electrical & Computer Engineering, McGill University
3480 University Street, Montreal, Quebec, H3A 2A7, Canada
e-mail: {xlu,champagne}@tsp.ece.mcgill.ca

ABSTRACT

This paper proposes a new acoustic echo cancellation (AEC) system which combines a conventional AEC algorithm with the pitch analysis technique. Departing from the use of pitch prediction in speech coding, we use two signals, i.e. the estimated echo and the residual echo, to fulfill the pitch extraction in the AEC system. This system is suitable for the applications of digital networks where the nonlinearities of the echo path significantly degrade the performance of conventional AEC algorithms. Simulation results using a nonlinear channel show that, compared to conventional AEC algorithms, the proposed system remarkably suppresses the acoustic echo during double-talk.

1 INTRODUCTION

Acoustic echo cancellation (AEC) has been extensively studied in the past decades [1]. Most AEC systems, based on the assumption that the acoustic echo path can be modelled as a linear system, employ an adaptive filter to estimate the acoustic echo from the loudspeaker signal and the microphone signal, then subtract the estimated echo from the microphone signal. In practice, the centralized AEC systems, i.e. located in central stations or base stations instead of user terminals, are more attractive. This is because such configuration can minimize the whole system costs and simplify the implementation of the user terminals.

As a result of advances made in telecommunications, digital communication networks prevail nowadays. Speech codecs are increasingly used to reduce the speech transmission rate in these modern networks, especially in the applications of cellular telephone and mobile radio. These low-bit-rate codecs introduce severe distortions of the speech signal in terms of waveforms. Since the codecs are cascaded along the echo path, the entire echo path presents strong non-linearities which significantly degrade the performance of the conventional AEC systems [2].

One of the serious problems of the conventional AEC system caused by the nonlinearities of the echo path is the echo suppression during the double-talk period. In a conventional AEC system, the coefficients of the adaptive filter are frozen to avoid the divergence of the adaptive filtering algorithm when both near-end and far-end speech are active [3]. However, in the nonlinear channel, the acoustic echo is difficult to suppress and the residual echo may be larger than the echo if the adaptation of the conventional AEC system is stopped, due to the complex nonlinear characteristics of the low-bit-rate codecs [4].

In this paper, we propose a new AEC system for the use over the nonlinear channel. Combined with a linear adaptive filter, the new AEC system exploits the speech analysis technique, namely pitch extraction from the residual echo, to further suppress the residual echo produced by the linear adaptive filter.

2 PITCH PREDICTION FILTER

2.1 Structure of the pitch filter

Speech signal is highly correlated, and thus has redundancies in either near-sample or distant-sample [5]. The near-sample redundancies can be removed by the formant filter, while the distant-sample waveform similarities can be extracted by the pitch filter.

In the application of AEC, the acoustic echo path is modelled as a linear system. That is, the echo is the output of a linear filter representing the loudspeaker-enclosure-microphone (LEM) system, with the loudspeaker signal as input. Compared to the loudspeaker signal, the formant of the acoustic echo usually changes notably because the formant is easily affected by the spectrum of the LEM system. However, the pitch is represented as periodic impulses in the frequency domain, and hence the characters of the pitch can be preserved for the output signal, i.e. echo, from the linear filter, i.e. LEM system. Consequently, the pitch information of the loudspeaker signal is similar to that of the echo signal, as we have verified experimentally.

There are different pitch prediction filters such as multi-lag pitch filters and fractional delay pitch filters [5]. We are only interested in the one-lag pitch filter which is shown in Figure 1, due to its simplicity and robustness. This pitch filter has only one coefficient

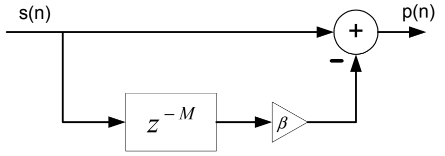

Figure 1: The pitch prediction for speech signal.

and is expressed

$$P(z) = \beta z^{-M}, \tag{1}$$

where β is a scaling factor related to the degree of waveform similarity and the integer M is the estimated period.

2.2 Estimation of the pitch parameters

The basic method of finding the pitch parameters, i.e. the lag M and the correlation coefficient β, is the open-loop analysis [5]. Using this approach, the pitch parameters M and β are chosen to minimize the mean-squared residual $p(n)$ in each N-sample frame:

$$\arg\min_{M,\beta} \sum_{n=0}^{N-1} p^2(n), \tag{2}$$

where, from Figure 1,

$$p(n) = s(n) - \beta s(n-M). \tag{3}$$

In the case of narrow-band speech where the sampling rate is 8kHz, the lag M ranges between 20 to 147 samples [6]. The pitch coefficient β varies from 0 to 1. For a signal with no detectable periodic structure such as unvoiced speech, β is 0 and M is irrelevant; for a well-structured periodic signal such as steady-state voiced speech, β is close to 1; the value of β lies between 0 and 1 for other cases.

For a given lag M, (2) leads to the optimal value of pitch coefficient in terms of M

$$\beta_{opt}(M) = \frac{\sum_{n=0}^{N-1} s(n)s(n-M)}{\sum_{n=0}^{N-1} s^2(n-M)}. \tag{4}$$

Clearly, in order to find the optimal gain β_{opt} among the values of $\beta_{opt}(M)$ and the corresponding M, an exhaustive search is necessary in the range of the pitch lag. Considering of the non-stationary property of the speech signal, the frame size N should not be too large in case of the reduction of prediction gain [7] and large delay. But too small frame size may cause inaccurate estimation of the pitch lag. In our research, we take the advantage of the use of different frame size N to estimate the pitch lag and the coefficient respectively. A larger frame size $N = 80$ is used for the lag M estimation, while a shorter frame size $N = 40$ is used to find the gain β and the updating of the output residual.

In order to avoid the pitch multiples issue of the pitch filter, the search range for the parameters is divided into three regions [8]. In practice, the computational complexity of a pitch predictor is remarkably reduced by searching the pitch parameters in the following steps.

Loop: for each frame ($N_1 = 40$)

- Step 1: Find three maxima $r(m_i),\ i = 1, 2, 3$, from the correlations

$$r(m) = \sum_{n=0}^{N_2-1} s(n)s(n-m), \tag{5}$$

where the larger frame size $N_2 = 80$ is used to obtain a better estimation of the correlation and the three ranges are

$$\begin{array}{ll} i = 1, & 80 \le m \le 147 \\ i = 2, & 40 \le m \le 79 \\ i = 3, & 20 \le m \le 39 \end{array} \tag{6}$$

- Step 2: Normalize the three candidates

$$\tilde{r}(m_i) = \frac{r(m_i)}{\sqrt{\sum_{n=0}^{N_2-1} s^2(n-m_i)}},\ i = 1, 2, 3. \tag{7}$$

- Step 3: Search for the proper pitch lag M among above candidates, where the smaller one is preferable to avoid the pitch multiples:

```
Initialization: $M = m_1$;
Loop: for $i = 2, 3$
        if $\tilde{r}(m_i) \ge \rho \tilde{r}(M)$
            $M = m_i$
        end
    end
```

where the weighting parameter ρ is set to 0.85 experimentally.

- Step 4: Compute the pitch gain β by using a shorter frame whose size is N_1 samples

$$\beta = \frac{\sum_{n=0}^{N_1-1} s(n)s(n-M)}{\sum_{n=0}^{N_1-1} s^2(n-M)} \tag{8}$$

End loop.

3 PROPOSED AEC SYSTEM

The AEC system that we propose for the nonlinear channel, where low-bit-rate codecs are present along the echo path, is illustrated in Figure 2. It consists of two components: an echo estimator and a pitch extractor. The estimated echo $\hat{y}(n)$ produced by the echo estimator is subtracted from the microphone signal $d(n)$, resulting in the residual echo $e(n)$. The processed residual echo $e_p(n)$ is obtained by attenuating the residual echo $e(n)$ through pitch extraction where the pitch parameters are computed from the estimated echo $\hat{y}(n)$.

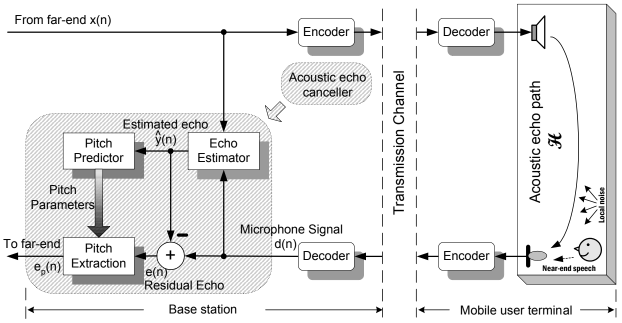

Figure 2: Pitch analysis-based acoustic echo cancellation over a nonlinear channel.

3.1 Echo Estimator

An adaptive filter is used to estimate the echo in most AEC applications, but its performance is significantly degraded by the nonlinearities of the codecs when they are present along the echo path. Since the affine projection (AP) algorithm [9] shows the best performance among some popular adaptive filtering algorithms in the nonlinear channel [2], it is thus used to estimate the echo in the new AEC system.

The AP algorithm, in a relaxed and regularized form, can be written as follows:

$$\hat{\mathbf{y}}_n = \mathbf{X}_n \mathbf{w}_n \tag{9}$$

$$\mathbf{e}_n = \mathbf{d}_n - \hat{\mathbf{y}}_n \tag{10}$$

$$\varepsilon_n = [\mathbf{X}\mathbf{X}^T + \delta \mathbf{I}]^{-1} \mathbf{e}_n \tag{11}$$

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu \mathbf{X}_n^T \varepsilon_n \tag{12}$$

In these equations, $\mathbf{w}_n$ represents the length-N coefficient vector of the adaptive filter at discrete time n, denoted by

$$\mathbf{w}_n = [w_0(n), w_1(n), \cdots, w_{N-1}(n)]^T, \tag{13}$$

where the superscript T denotes the transpose of the vector. The microphone signal vector $\mathbf{d}_n$, the estimated echo vector $\hat{\mathbf{y}}_n$, and the residual echo signal vector $\mathbf{e}_n$ are respectively defined as

$$\mathbf{d}_n = [d(n), d(n-1), \cdots, d(n-p+1)]^T, \tag{14}$$

$$\hat{\mathbf{y}}_n = [\hat{y}(n), \hat{y}(n-1), \cdots, \hat{y}(n-p+1)]^T, \tag{15}$$

$$\mathbf{e}_n = [e(n), e(n-1), \cdots, e(n-p+1)]^T, \tag{16}$$

where p denotes the projection order. The excitation signal matrix $\mathbf{X}_n$ is defined as

$$\mathbf{X}_n = [\mathbf{x}_n, \mathbf{x}_{n-1}, \cdots, \mathbf{x}_{n-p+1}]^T, \tag{17}$$

with the far-end signal vector $\mathbf{x}_n$

$$\mathbf{x}_n = [x(n), x(n-1), \cdots, x(n-N+1)]^T. \tag{18}$$

A small diagonal matrix $\delta\mathbf{I}$ is added to $\mathbf{XX^T}$ in case of ill-conditioned problem. The relaxation factor must be chosen for $0 < \mu < 2$ to keep the algorithm stable.

The echo estimator provides two signals for the next processing, namely the estimated echo $\hat{y}(n)$ and the residual echo signal $e(n)$.

3.2 Pitch Extraction in AEC

Speech analysis indicates that most of the speech energy is concentrated on the voiced sounds whose power is often about 20dB larger than that of the unvoiced sounds [6]. Furthermore, the voiced sounds have a relative periodicity which is represented by the pitch. Based on these considerations, the power of the residual echo from a conventional acoustic echo canceller will be further reduced if the pitch of the residual echo is extracted.

In the new AEC system shown in Figure 2, the acoustic echo is attenuated by subtracting the estimated echo $\hat{y}(n)$ from the microphone signal $d(n)$ before it is further suppressed by the pitch filter. The residual echo signal $e(n)$ may contain the near-end speech and the remaining echo. When the near-end speech is active, it is almost impossible to obtain the correct pitch parameters of the echo component from the residual echo signal.

However, as discussed before, the pitch information of the far-end speech $x(n)$ is similar to that of the echo, but that information can not be directly applied to the residual echo due to the synchronization problem, i.e. the delay introduced by the codecs and the acoustic echo path. This problem can be solved if the estimated echo $\hat{y}(n)$ is used to obtain the pitch parameters. Then these parameters are applied to the pitch filter that attenuates the residual echo. This is because, the pitch information of the estimated echo and that of the echo component contained in the residual echo signal are very alike when the echo estimator is active, and secondly, based on the assumption that the delay of the entire echo path does not change significantly during double-talk period when the coefficients of the echo estimator are frozen, those two signals are still well synchronized.

The pitch parameters, i.e. the pitch lag M and the pitch gain β, of the estimated echo $\hat{y}(n)$, are obtained by using the algorithm in section 2.2 from step 1 to 4, where the $s(n)$ should be replaced by $e(n)$. The pitch of residual echo $e(n)$ is then extracted using (3). Similarly, the signals $p(n)$ and $s(n)$ in (3) are replaced by $e_p(n)$ and $e(n)$, respectively.

4 SIMULATION RESULTS

In order to test the proposed AEC system, we conducted a simulation based on the platform shown in Figure 2, where the codecs are G.729 [8]. A coloured noise, produced by passing a white noise through an IIR filter with the system function $H(z) = \frac{0.1}{1-0.9z^{-1}}$, was added as the background noise, so that the echo-to-noise ratio is 30dB. The LEM system of our test platform was simulated to represent the cab of a vehicle. The impulse response was about 40ms long, corresponding to 300 taps at the sampling rate of 8kHz. The relaxation factor μ for the AP echo estimator was set to 0.9; while the projection order p was 3, since a higher order can not lead to obvious improvement to the AP algorithm in this nonlinear channel [4].

The double-talk occurred between time $0.6s$ and $2.7s$. During this period, the coefficients of the echo estimator, i.e. the AP algorithm, were frozen to avoid the divergence. The simulation results are shown in Figure 3 and Figure 4 in terms of signal waveforms and signal powers, respectively. The residual echo of the conventional AEC system which only employs AP to suppress the echo was also plotted in these figures for comparison.

From the simulation results, we can find that, in the nonlinear channel, the pitch analysis-based AEC system yields 5 to 10dB of additional echo attenuation during double-talk period, as compared to the conventional AEC system which usually only use an adaptive filter for the echo suppression. In the case of single-talk, the new AEC system obtained the same results as the conventional AEC system, since the periodic similarities in the residual echo have been removed by the echo estimator when the AP algorithm is active due to its strong tracking capability. Minor distortions to the near-end speech, which may add a little extra noise to the near-end speech, was introduced by the new AEC system. However, since the local background noise exists in the most cases of hands-free applications, this distortion is acceptable.

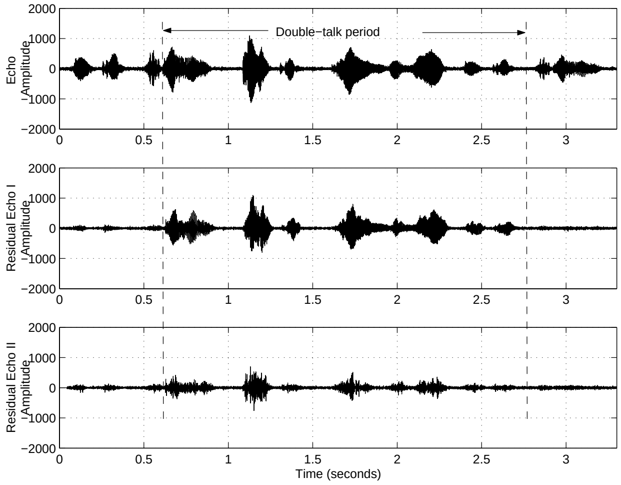

Figure 3: Waveforms of the echo, residual echo I (only AP is employed), and residual echo II (the proposed AEC system).

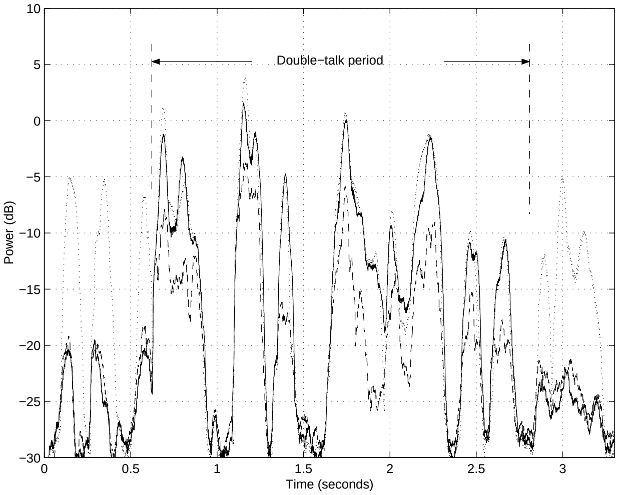

Figure 4: Power versus time for echo (dot), residual echo of AP (solid), and residual of the proposed AEC system (dash).

5 CONCLUSION

We have presented a new AEC system which combines a pitch extractor with a conventional echo estimator for the use of echo suppression over a nonlinear channel where low-bit-rate codecs are cascaded along the echo path. The simulation results shows that the proposed AEC system significantly outperforms the conventional AEC system during double-talk with a little acceptable distortions.

References

[1] C. Breining *et al.*, "Acoustic echo control," *IEEE Signal Processing Magazine*, pp. 42–69, Jul. 1999.

[2] Y. Huang and R.A. Goubran, "Effects of vocoder distortion on network echo cancellation," in *Proc. ICME'00*, 2000, vol. 1, pp. 437–439.

[3] P. Heitkämper, "An adaptation control for acoustic echo cancellers," *IEEE Signal Processing Letters*, vol. 4, no. 6, pp. 170–172, Jun. 1997.

[4] X. Lu and B. Champagne, "Acoustic echo cancellation over a non-linear channel," in *Proc. IWAENC'01*, Sep. 2001, pp. 139–142.

[5] R. P. Ramachandran and R. J. Mammone, Eds., *Modern Methods of Speech Processing*, Kluwer Academic Publishers, Boston, 1995.

[6] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, New York, 2nd edition, 2000.

[7] R. P. Ramachandran and P. Kabal, "Pitch prediction filter in speech coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 4, pp. 467–478, Apr. 1989.

[8] ITU-T Recommendation G.729, *Coding of Speech at 8 kbits/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, International Telecommunication Union, Mar. 1996.

[9] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Elec. and Comm. in Japan*, vol. 67-A, no. 5, pp. 19–27, 1984.