ACOUSTIC ECHO CANCELLATION OVER A NON-LINEAR CHANNEL

Xiaojian Lu and Benoît Champagne

Department of Electrical & Computer Engineering, McGill University
3480 University Street, Montreal, Quebec, H3A 2A7, Canada
{xlu,champagne}@tsp.ece.mcgill.ca

ABSTRACT

A new acoustic echo canceller (AEC) for use over a non-linear channel is proposed in this paper. This AEC shows improved performance when low bit rate codecs are present along the echo path which is often the case in digital network applications. In our experiments, the new AEC not only suppresses the acoustic echo significantly, but also works well in double talk situation, where no double talk detection is required.

1. INTRODUCTION

Speech codecs are widely employed in digital channels today in order to compress the transmitted data. Although codecs such as G.729 [1] or GSM can significantly reduce the transmission rate of the speech signal, they introduce severe waveform distortion to the output signal. For example, it can be verified that the Signal-to-Noise Ratio (SNR) of the reconstructed speech signal from G.729 is only about a few decibels.

A typical setup of AEC in the context of wireless telecommunications is shown in Figure 1. The AEC devices are placed in the central stations or the base stations instead of the mobile user terminals in order to minimize the system costs and to simplify the implementation of the mobile terminals. Since the low bit rate codecs are cascaded with the acoustic echo path, the non-linear properties of these codecs cause the whole echo path to present non-linearity. The non-linear effect of the codecs on the AEC is still on the way of study [2].

The conventional AEC that usually employs an adaptive filter has been studied extensively [3]. However, its performance is severely degraded when the codecs are present along the echo path. This is because the linear structure of the adaptive filter cannot handle the non-linear characteristics of the codecs very well. Although the acoustic echo can be suppressed to some degree if a conventional AEC is used in the configuration illustrated in Figure 1, the attenuation of the echo is not large enough to make the echo residual unperceivable.

In a conventional AEC setting, where no codecs are involved, the coefficients of the adaptive filter need to be frozen when double talk occurs. In this case, there is still some attenuation of the echo since in practice the acoustic echo path does not change too fast. However, if codecs are present along the echo path, the conventional AEC must be deactivated to avoid the possibility of the echo residual becoming larger than the echo, because the whole echo path containing codecs changes rapidly. The required detection of double talk is by itself a very difficult issue of AEC.

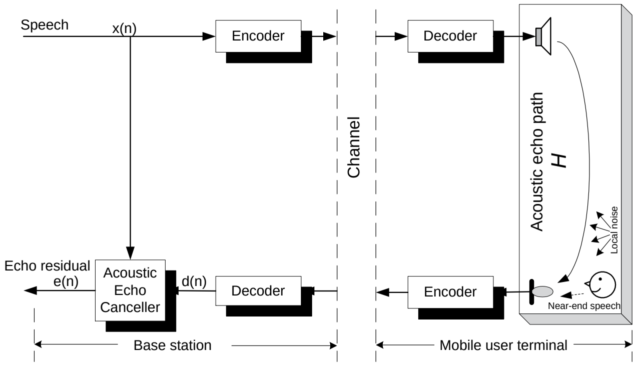

Figure 1: A digital communication network with AEC located at base station.

In this paper, a new structure of AEC is proposed. This approach not only significantly reduces the echo residual in the nonlinear channels, but also eliminates the need of the double talk detector.

2. THE NON-LINEAR CHANNEL AEC

The structure of the non-linear channel acoustic echo canceller (NLCC) that we propose is illustrated in Figure 2, where $x(n)$ is the input signal of the first codec (upper branch in Figure 1), $d(n)$ is the output of the second codec (lower branch in Figure 1), $y(n)$ is the estimated echo replica, and $e(n)$ is the output of the AEC. In effect, this structure is made up of a linear part and a non-linear part.

The linear part consists of an FIR filter combined with an adaptive cross-spectral algorithm, which is a modified version of the adaptive cross-spectral technique in [4]. The latter is used to adjust the tap weights

This work was supported by Nortel Networks.

of the FIR filter whose output provides the echo replica $y(n)$. The non-linear part uses the spectral subtraction technique [5, 6] to suppress the nonlinear echo components introduced by the codecs from the microphone signal.

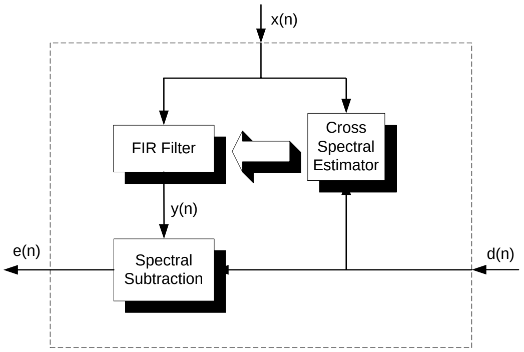

Figure 2: A new acoustic echo canceller.

2.1. Interpretation of the Echo Path

The characteristics of the codec present strong nonlinearities. However, based on observations of the codec output, the net effect of the codec on its input may be approximated as the superposition of a non-linear component on a slowly-varying linear component. Mathematically, this may be expressed as follows. Let $h_l(n)$ denote the impulse response of the linear portion over some limited time interval, and let $F_{nl}(\cdot)$ denote the non-linear function of the codec. Then, the codec output signal, $s_{out}(n)$, can be expressed in terms of its input, $s_{in}(n)$, as

$$s_{out}(n) = s_{in}(n) * h_l(n) + F_{nl}(s_{in}(\cdot)), \qquad (1)$$

where $*$ denotes the convolution.

We define the effective frequency response of the codec as

$$H_{eff}(k;m) = \frac{S_{out}(k;m)}{S_{in}(k;m)}, \qquad (2)$$

where, $S_{in}(k;m)$ and $S_{out}(k;m)$ are the short-term Discrete Fourier Transform (stDFT) [6] of the input and output signals of the codec, respectively. The effective short-term impulse response, $h_{eff}(n;m)$, is obtained by taking the inverse stDFT of $H_{eff}(k;m)$. Note that the parameter k is the index of the frequency bins, m is the index of the blocks in time domain, and n is the time index.

From a practical viewpoint, the coefficients of the effective short-term impulse response of the codec can be regarded as the superposition of linear and non-linear components. The former may be regarded as a slowly varying mean value that corresponds to $h_l(n)$ in (1); the latter as a faster fluctuation that corresponds to $F_{nl}(s_{in}(\cdot))$ in (1). We have been able to confirm this behavior experimentally.

Because the acoustic echo path which is supposed to be a linear system cascades with the codecs, the impulse response of the entire echo path is obtained by the convolution of the effective impulse response of the codecs and the system impulse response of the acoustic echo path. Consequently, the short-term impulse response of the entire echo path consists of linear and nonlinear parts, where the former can be estimated by an adaptive filtering algorithm, and the latter can be handled by a non-linear procedure, as explained in the following sections.

2.2. Estimation of the Echo Path

Many algorithms can be employed to estimate an echo path in AEC [3], but most of them just work well in the single talk situation. When the near-end speech is present, i.e. in the double talk case, the adaptation of AEC must be frozen. Since the double talk detection still remains a very difficult issue for AEC, it is better for us to find an algorithm that can avoid this problem. The adaptive cross-spectral technique [4] has the advantage of working well without the double talk detection even in the high disturbance case. Furthermore, this kind of block processing technique is suitable for estimating the mean coefficient values of the echo path in the setup shown in Figure 1.

Referring to Figure 2, let $\mathbf{x}(n)$ denote the N-dimensional input signal vector, i.e.

$$\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-N+1)]^T, \qquad (3)$$

and let $\hat{\mathbf{h}}(n)$ denote the estimated channel coefficient vector in a conventional N-tap adaptive filter:

$$\hat{\mathbf{h}}(n) = [h_0(n), h_1(n), \ldots, h_{N-1}(n)]^T. \qquad (4)$$

Then, the error signal $e_p(n)$ can be written as

$$e_p(n) = d(n) - \hat{\mathbf{h}}(n)^T \mathbf{x}(n), \qquad (5)$$

where the microphone signal $d(n) = \mathbf{h}(n)^T \mathbf{x}(n) + v(n)$, $v(n)$ is the near-end signal. Introducing $\Delta\mathbf{h}(n) = \mathbf{h}(n) - \hat{\mathbf{h}}(n)$, the error signal can be rewritten as

$$e_p(n) = \Delta\mathbf{h}(n)^T \mathbf{x}(n) + v(n), \qquad (6)$$

Taking the stDFT of (6), then

$$E_p(k;m) = \Delta H(k;m) X(k;m) + V(k;m), \qquad (7)$$

where, $E_p(k;m)$, $X(k;m)$, and $V(k;m)$ are the stDFT of $e_p(n)$, $x(n)$, and $v(n)$ at block m, respectively, while $\Delta H(k;m)$ is the error of the estimated frequency response of the echo path.

Multiplying both sides of (7) by the complex conjugate of the input signal spectrum $X^*(k;m)$, and taking the expectation, one can easily find the recursion of the adaptive filter coefficients in the general form:

$$\Delta H(k;m) = \frac{\mathrm{E}[X^*(k;m)E_p(k;m)]}{\mathrm{E}[|X(k;m)|^2]}, \tag{8}$$

$$H(k;m+1) = H(k;m) + \Delta H(k;m), \tag{9}$$

where the assumption of no correlation between the far-end speech $X(k;m)$ and the near-end signal $V(k;m)$ has been made. Actually, correlation does exist between these signals in some cases. However, since the correlation is weak in most situations, this assumption is reasonable.

In practice, the expectation $\mathrm{E}[\cdot]$ in (8) is estimated by taking average over the time. Here, M non-overlapping blocks are used to compute the average value each time, and the block length is N. Empirical results show that block number M should not be too small in order to ensure the convergence of the algorithm [4]. Thus, each update of the echo path needs to gather MN samples. Since the acoustic echo path is relatively long, the slow initial convergence rate of the cross-spectral algorithm becomes significantly deficient for its use in the application of AEC.

To overcome this drawback, we propose a modified version of the original adaptive cross-spectral algorithm. Our approach focuses on the initial transient period where both expectation and weight update are computed differently, so that the adaptation starts as early as possible. The modified adaptive cross-spectral algorithm can be stated as follows.

During the initial period, i.e. for $m = 1, 2, \ldots, M$, $H(k;m)$ is computed every block as is

$$H(k;m+1) = \Delta H(k;m). \tag{10}$$

Following this period, $H(k;m)$ is updated every M blocks as is the original approach [4]. Thus, for $m = lM + p$, where $l = 1, 2, \ldots$; $p = 1, 2, \ldots, M$:

$$H(k;lM+p) = H(k;lM) + \Delta H(k;lM). \tag{11}$$

The formula for computing $\Delta H(k;m)$, for both periods, is given by

$$Y(k;m) = X(k;m)H(k;m), \tag{12}$$

$$E_p(k;m) = D(k;m) - Y(k;m), \tag{13}$$

$$\Delta H(k;m) = \frac{\sum_{i=p_1}^{p_2} [X^*(k;i)E_p(k;i)]}{\sum_{i=p_1}^{p_2} [|X(k;i)|^2]}, \tag{14}$$

$$p_1 = \left\lfloor \frac{m-1}{M} \right\rfloor M + 1;$$

$$p_2 = \left\lfloor \frac{m-1}{M} \right\rfloor M + min\{m, M\}.$$

2.3. Spectral Subtraction Method

The Spectral Subtraction Method was developed for the suppression of acoustic noise in speech [5]. This approach and its variants have been widely used in the speech enhancement. The core of the spectral subtraction method is to find a noise-suppressed spectral estimator. This estimator is obtained by subtracting an estimate of the noise spectrum from the spectrum of the noisy speech. The noise spectrum is estimated during the silence period of the speech since there is only one signal input in most cases of speech enhancement.

In the application of AEC, unlike the speech enhancement, there are two signals provided simultaneously, namely, microphone signal and far-end signal. The estimated echo signal is obtained by convolving the far-end speech and the estimated linear portion of the echo path; the microphone signal includes acoustic echo and near-end speech. Recall from the previous section that the spectra of those two signals, which are computed through the stDFT, are denoted as $Y(k;m)$ and $D(k;m)$, respectively. Hence, the spectrum amplitude of the echo-suppressed estimator is obtained:

$$|E(k;m)| = |D(k;m)| - |Y(k;m)|. \tag{15}$$

For the simplicity of the algorithm, the phase of the microphone signal is employed as the phase of the echo-suppressed estimator. This is the usual procedure in the spectral subtraction approach, since the replacement of the phase spectrum is sufficient for all practical purposes [6].

Because the signal obtained from (15) can only remove the echo caused by the linear portion of the echo path, the echo residual is still too high to achieve the goal of echo cancellation. Referring to the weighted spectral subtraction method which was originally developed for the speech enhancement [6], we transplant this approach to AEC to suppress the echo introduced by the non-linearity of the echo path:

$$E(k;m) = [|D(k;m)|^\alpha - k|Y(k;m)|^\alpha]^{1/\alpha} e^{j\varphi_D(k;m)}, \tag{16}$$

where the parameters α and k control the echo suppression and the signal distortion, and $\varphi_D(k;m)$ is the phase of the microphone signal $d(n)$.

Most conventional AECs do not use the spectral subtraction approach because this process could cause some distortion of the near-end speech. However, the distortion is unavoidable when the codecs are present along the transmission channel shown in Figure 1, and to the best of our knowledge, the echo caused by the channel non-linearity is very difficult to remove if only by conventional means. Our experiments suggest it is preferable to achieve a large attenuation of echo at the price of a little distortion. To take advantage of this approach, the parameters α and k should be carefully chosen so as to optimize the trade-off between the echo suppression and the distortion of the near-end speech.

3. SIMULATION RESULTS

In our simulation platform, real speech was used as the test signal, and the codec was G.729. White noise was added as the local noise signal, so that the Echo-to-Noise Ratio (ENR) was 30dB at the near-end. Another speech signal was also added to the near-end signal to simulate the double talk situation. The Loudspeaker-Enclosure-Microphone (LEM) system was mimic to the cab of a vehicle, whose impulse response was about 40ms, corresponding to 300 taps at 8kHz sampling rate.

For the first part of our simulations, the modified adaptive cross-spectral algorithm was compared with the original one in the above mentioned platform. As we expected, the former shows 5 $\sim$ 10dB of MSE lower than the latter for the initial convergence. After initialization, both algorithms are identical, with staircase-like convergence curves.

In the second part of the simulations, the NLCC was implemented and tested in the platform. The analysis and synthesis windows used in the algorithm were the Hanning window and the rectangular window, respectively. Half-overlap data segments were necessary in this case, where the window length was 300 in the simulation. Empirical results indicate that a high attenuation is achieved while the distortion of the near-end speech is acceptable, when the parameters are chosen as $k \approx 1.0$, and $\alpha = 0.6 \sim 1.0$.

The Affine Projection Algorithm (APA) [7] was also implemented for comparison with NLCC, since the former has very attractive properties for the conventional AEC application (e.g., fast convergence rate for colored signals such as speech). The coefficients of APA were frozen to avoid the divergence of the algorithm during the double talk period. In the simulation of APA, the projection order was set to 3; no obvious improvement was observed with higher projection orders when the codecs are present along the echo path.

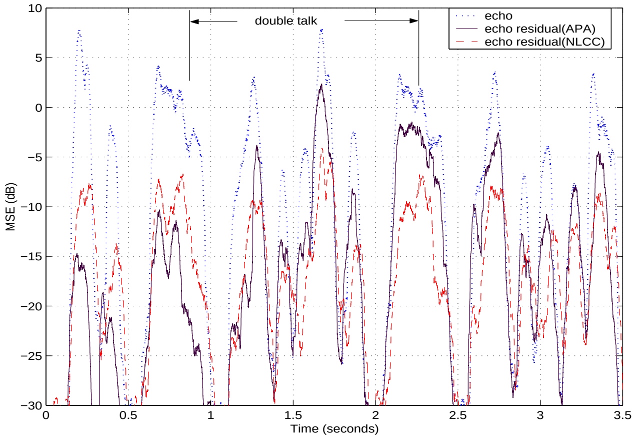

Figure 3: MSE versus time in the steady-state for NLCC ($k = 1.0, \alpha = 0.8$) and APA (step-size $\mu = 0.9$).

Comparing the MSE curves plotted in Figure 3 in three periods: initial period, double talk period, and after double talk period, one can find some interesting results. With a linear filter structure, APA regards the non-linear echo path as a linear time-varying system. Therefore, it desperately tracks the change of the echo path. Because of its merit of fast tracking, APA has achieved a lower MSE in the first period, although it does not estimate the average coefficients of the system response. In the double talk period, NLCC shows very desirable properties: MSE achieved 5 $\sim$ 10dB lower than APA; no double talk detection needed. Note that APA will diverge without the double talk detector while such detector is still a challenge of the AEC application. In the third period, APA has a high MSE at first because the entire echo path has changed significantly during the double talk period due to the codecs.

4. CONCLUSIONS

We have been able to verify that when codecs are present along the echo path, the NLCC approach proposed here has better performance than the conventional AEC based purely on the linear adaptive filtering. Indeed, the NLCC significantly reduces the echo residual in the nonlinear channel during both single talk *and* double talk periods, without the need of a double talk detector. As a result, the NLCC ensures smooth transition between double talk and single talk so as to make users more comfortable.

5. REFERENCES

[1] ITU-T Recommendation G.729, Mar. 1996.

[2] ITU-T Recommendation G.168, Apr. 2000.

[3] C. Breining, *et al.*, “Acoustic Echo Control: An Application of Very-High-Order Adaptive Filters,” *IEEE SP Magazine*, pp. 42-69, Jul. 1999.

[4] T. Okuno, *et al.*, “Adaptive Cross-Spectral Technique for Acoustic Echo Cancellation,” *IEICE Trans. Fundamentals*, vol. E82-A, no. 4, pp. 634-639, Apr. 1999.

[5] S. F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” *IEEE Trans. ASSP*, vol. 27, no. 2, pp. 113-120, Apr. 1979.

[6] J. R. Deller, *et al.*, *Discrete-time processing of speech signals*, IEEE Press, New York, 2000.

[7] K. Ozeki, *et al.*, “An Adaptive Filtering Algorithm Using an Orthogonal Projection to an Affine Subspace and its Properties,” *Electronics and Communications in Japan*, vol. 67-A, no. 5, pp. 19-27, 1984.