DOUBLE-TALK DETECTION IN ACOUSTIC ECHO CANCELLERS USING ZERO-CROSSINGS RATE

Muhammad Z. Ikram

Embedded Processing Systems Lab
Texas Instruments Incorporated
12500 TI Boulevard, Dallas, TX 75243
mzi@ti.com

ABSTRACT

We propose a new method to detect double-talk and control filter adaptation in an acoustic echo canceller (AEC). The method is based on computing the zero-crossings rate (ZCR) of the AEC output and comparing it against a suitably-chosen threshold. As the ZCR values falls below the threshold, double talk is declared and the AEC filter adaptation is either slowed down or halted. The zero crossings are very easy to compute by observing the sign changes of two consecutive samples from the output of the AEC. In contrast to most existing methods, the computational burden of the proposed method is minimal and it can, therefore, be conveniently implemented on a low-power, low-resource processor. This computational simplicity is enjoyed without sacrificing for any AEC performance. We will illustrate effectiveness of the proposed method by comparing against the existing state of the art and present guidelines on choosing parameters for computing the sample-by-sample ZCR.

Index Terms— Speech Enhancement, Zero Crossing Rate, Low-Power Processing.

1. INTRODUCTION

The history of acoustic echo cancellation (AEC) dates back to 1960s, and since then a number of attractive methods have been proposed to address various aspects of this problem [1]. One of the primary issues that has remained at the forefront of AEC development is its handling of double talk (DT) [2]. DT happens when both near-end and far-end speakers talk at the same time. The adaptive filter in an AEC is designed to cancel only the far-end echo, and any presence of near-end signal strongly influences its convergence. The DT results in divergence of the adaptive filter and causes the far-end listener to hear its own echo, which is annoying and undesirable. Ever since the development of the first adaptive-filter based echo canceller several methods have been proposed to detect DT and thus avoid filter divergence. DT detectors have now become an integral part of almost all commercially available echo cancellers. A review of classical DT detection methods can be found in [3]. Certain methods handle the DT by restricting the communication to only one way; i.e., half-duplex. In other cases, a DT detector is used that freezes filter adaptation in the presence of DT [1]. While half-duplex communication is not desirable in many situations the DT detector based AEC bring along its own issues. First, in order to benefit from frozen adaptation, the DT detector should correctly estimate the start and end of the near-end speech. Any misdetection may lead to echo leakage to the far end. Secondly, the echo cancellation may suffer if the echo path changes during the time of frozen adaptation.

In this paper, we present an effective and a cost-efficient solution to the DT problem. The simplicity of the proposed method allows it to be implemented on a low-power processor without any loss in performance. This solution is based on measuring the zero-crossing rates (ZCR) of the AEC output and comparing it against a threshold. We show that the ZCR serves as a classifier in discriminating between the presence and absence of DT. We employ this detector in the normalized least-mean squared (NLMS) based AEC and control the filter adaptation based on its decisions. The ZCR is measured over a window of samples and its estimate can be updated with each incoming sample or after every block of samples. As a result, no additional delay is introduced, and the ZCR is updated by comparing sign of the incoming sample against that of the previous sample. The main contribution of the paper is low complexity of the proposed method, which shines in contrast to the existing methods, where in some cases expensive correlation is computed over a window of samples. Furthermore, the memory requirement is minimal since only a bit is needed to store a zero-crossing decision. We compare performance of the proposed method against the well-known normalized cross-correlation method of [4] as the speech mode changes from far-end speaker (single talker) only to DT and then back to the single talker case. We also provide guidelines to choose the window size over which ZCR is measured and to select the threshold. We would like to emphasize that although the results that we present in the paper pertain to

NLMS-based implementation, the theory and implementation is applicable to any AEC configuration where DT control is desired.

2. AEC AND THE DT PROBLEM

Let us consider the AEC setup shown in Fig. 1. The far-end and near-end speakers are denoted by $s_1(n)$ and $s_2(n)$, respectively. The echo path from the loudspeaker to the microphone is modeled by a length-L FIR filter $\mathbf{h}(n) = \left[h^0(n), h^1(n), \ldots, h^{L-1}(n)\right]^T$, where the superscript on the filter coefficient denotes the tap index and $[\cdot]^T$ denotes transposition. Likewise, the adaptive filter of length P is denoted by $\mathbf{w}(n) = \left[w^0(n), w^1(n), \ \ldots, w^{P-1}(n)\right]^T$.

There are several method available to update the filter coefficients $\mathbf{w}(n)$. In this paper, we will use the well-known normalized least-mean squares (NLMS) algorithm [1], because of its simplicity.

$$w^k(n) = w^k(n-1) + \alpha \frac{e(n)s_1^*(n-k)}{\sum_{i=0}^{P-1} |s_1(n-i)|^2}, \qquad (1)$$

for $k = 0, \ldots, P-1$, where α is the adaptation constant and $(\cdot)^*$ denotes complex conjugation. The echo is cancelled when $\mathbf{w}(n) = \mathbf{h}(n)$ at convergence. As noted earlier, DT occurs when the far-end and near-end speakers talk simultaneously; i.e., $s_1(n) \neq 0$ and $s_2(n) \neq 0$. In this event there is a tendency for the AEC to diverge resulting in annoying experience for the far-end listener. Handling of DT in AEC has always remained an area of interest within the speech research community. Several notable methods have been proposed for this purpose. In [3], [5], [6], and [7] an extra component is employed in the AEC to help avoid filter divergence during DT. This component could be in the form of a DT detector or a step-size controller. The DT detector detects the presence of DT by comparing some signal statistics against a pre-set threshold. Most common forms of DT detectors are Geigel detector and detectors based on coherence or cross-correlation, and their variants. Both these detectors are sensitive to variations in echo path; moreover the cross-correlation based detectors are also computationally expensive and an attempt to reduce the complexity comes at an expense of loss in performance [7]. In [8] the authors presented a variable step-size NLMS (VSS-NLMS) algorithm that is robust against DT. The automatic step-size control mechanism of this method halts the adaptation during instances of DT. The algorithm did not require explicit DT detection; however its convergence rate was slow. In [9], the authors proposed to wipe off the frequency contents in a spectral slit of the downlink signal, and they detected DT if the same spectral slit of the microphone input had any frequency content. This method required frequency-domain transformation and was also sensitive to near-end noise.

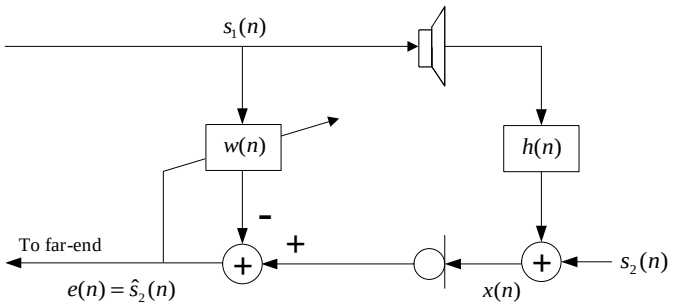

Fig. 1. AEC setup.

3. ZCR FOR DT DETECTION

For digital signals, zero crossing occurs if two consecutive samples have opposite signs. The ZCR is defined as the number of zero crossings per sample, and, for an analysis window of M samples, it is computed by dividing the number of zero crossings by M. At time n, the ZCR for digital signal $y(n)$ is given by

$$\mathrm{ZCR}(n) = \frac{1}{2M} \sum_{m=n-M+1}^{n} \left|\mathrm{sgn}\left(y(m)\right) - \mathrm{sgn}\left(y(m-1)\right)\right| \omega(n-m), \qquad (2)$$

where the sign operator is defined as

$$\mathrm{sgn}\left(y(n)\right) = \begin{cases} 1 & y(n) \geq 0 \\ -1 & y(n) < 0, \end{cases}$$

and $\omega(n)$ is the window operator. The term within the modulus operator is equal to 2 when a zero crossing occurs and the two signal samples have opposite sign. The summation on the right side of (2) is, therefore, divided by 2 to obtain the total number of zero crossings. After measuring the ZCR over a window of M samples the short-time window is advanced by K samples to obtain the next estimate. As we will see later, the parameters M and K play a key role in the operation of the DT detector.

Along with other well-known measures, such as short-time energy and short-time auto-correlation, the ZCR is also used to characterize time-domain signals. In the realm of speech processing, zero crossings have been used in the past for speech recognition and speech-music discrimination [10]. One of the classical uses of ZCR is the frequency estimation of sinusoidal signals. Signals with higher frequency contents result in higher short-time ZCR, whereas low-frequency signals have low short-time ZCR. It is also known that high-energy signals have lower ZCR and low-energy signals have higher ZCR [10]. It is this property of ZCR that we will exploit to detect the occurrence of near-end speech.

4. DT DETECTION IN AEC USING ZCR

In the event of DT, the near-end signal $s_2(n)$ acts as disturbing noise to the adaptive filter that cancels the acoustic echo. As a result, the filter diverges and unwanted echo escapes to the far-end. Mathematically, the AEC output in the presence of DT is given by

$$e(n) = [\mathbf{h}(n) - \mathbf{w}(n)] \star s_1(n) + s_2(n), \tag{3}$$

where $\star$ is convolution operator. In the absence of any corrective measure taken to counter against the impact of DT, $\mathbf{w}(n) \neq \mathbf{h}(n)$, and the energy of the AEC output remains high during this time. Consequently, the ZCR stays low during the time that the near-end speaker is active. We monitor the ZCR to detect DT and propose appropriate measures to avoid filter divergence.

Let us look at an example to analyze the use of ZCR for detecting DT. We consider a conversation between a male speaker at the far-end and a female speaker at the near-end. The two speech signals are shown in Fig. 2. It is seen from the figure that the far-end speaker talks for the first 10 sec, followed by about 11 sec of DT, with both $s_1(n) \neq 0$ and $s_2(n) \neq 0$. Finally, the last 4 sec of conversation again constitutes single talk. Such an example is suitable to evaluate a DT detection algorithm as the conversation mode changes from single talk to DT and then back to single talk. In our example, the echo length is about 30 sec and the average far-end to DT ratio is 5dB. To evaluate the AEC performance we compute adaptive-filter misalignment given by

$$20 \log 10 \frac{\|\mathbf{h} - \mathbf{w}\|}{\|\mathbf{h}\|}. \tag{4}$$

With reference to (1), we used a step size of $\alpha = 0.5$.

Fig. 3 shows the misalignment of the AEC adaptive filter. The filter diverged as soon as the near-end speaker started speaking. At the end of the near-end speech, the filter is again seen to converge. It should be noted that the filter misalignment requires a knowledge of the echo impulse response $\mathbf{h}(n)$, which is not available in practice; it is shown here to illustrate the impact of DT on the filter AEC performance. Also shown in Fig. 3 is the ZCR at the output of the AEC. It is clearly seen that the ZCR falls with the introduction of DT and stays low as long as the two speakers continue talking simultaneously. The ZCR was updated with each incoming sample ($K = 1$) and computed using a rectangular window of $M = 1000$ samples, which corresponds to 125 msec at 8kHz. ZCR has been used in the past for DTD. For example, the inventors in [11] compute two ZCRs, one for the near-end signal and the other for the far-end signal, which are then compared for the occurrence of double talk. This method is based on spectral differences between the two signals and, thus, will only work for voiced speech. On the other hand, our method is based on signal energy and works for both voiced and unvoiced parts of speech.

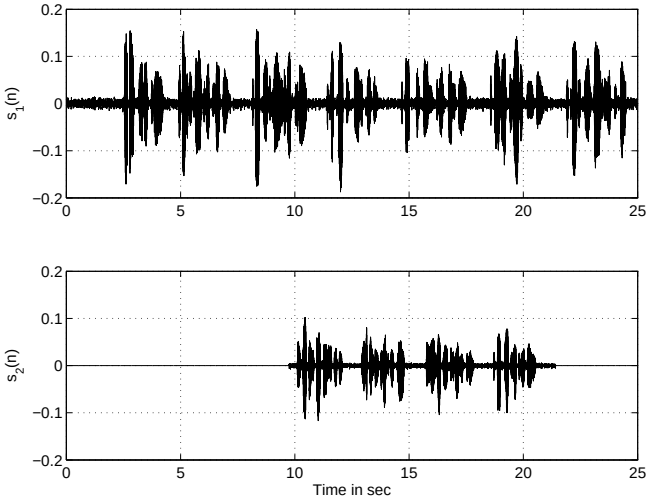

Fig. 2. The two speech signals used in the experiment. $s_1(n)$ is the far-end speech, whereas $s_2(n)$ is the near-end speech.

5. NEW AEC USING ZCR-BASED DT DETECTOR

In order to avoid filter divergence during DT we place the ZCR-based DT in AEC as shown in Fig. 4. The logic within the detector consists of measuring the short-time ZCR and comparing it against a suitably-chosen threshold γ ($0 \leq \gamma < 1$). As the ZCR falls below γ, DT is declared and filter adaptation is halted. Otherwise, normal processing takes place. This can be summarized as follows:

```
if ZCR(n) > γ
    w̃ = w(n)
else if ZCR(n) ≤ γ
    e(n) = x(n) − w̃ ⋆ s1(n)
end
```

Note that $\tilde{\mathbf{w}}$ is used to store the last "good" set of filter coefficients that do not let the filter diverge. The coefficients in $\tilde{\mathbf{w}}$ are updated only in the absence of DT. During DT, $\tilde{\mathbf{w}}$ is used to compute the AEC output. Let us use this detector to control DT in the NLMS-based AEC. Using the values of M and K that we used to plot ZCR in Fig. 3, we ran the AEC with the ZCR-based DT detector of Fig. 4 using $\gamma = 0.45$. The resulting AEC filter misalignment is shown in Fig. 5. Also shown in the same plot is the filter misalignment from Fig. 3 with no control for double talk. It is seen that the DT is correctly detected using ZCR at 10 sec and the filter adaptation is halted during the time that the detector indicates the presence of DT. Normal processing resumes at around 22 sec when the near-end speech ends.

Let us now use the same test signals $s_1(n)$ and $s_2(n)$ and run them through the normalized cross-correlation (NCC) based DT detector [4]. We used a window size of 550 in the NCC based detector, which resulted in best performance during the occurrence of DT. The results are presented in Fig. 5.

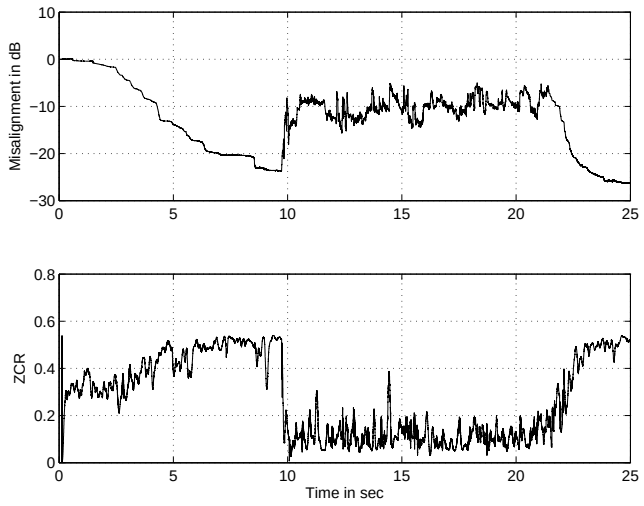

Fig. 3. Behavior of the NLMS-based AEC in the presence of DT. The adaptive-filter misalignment and the ZCR of the AEC output are shown.

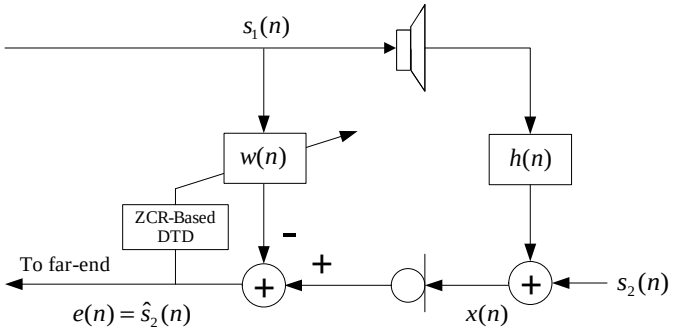

Fig. 4. AEC with ZCR-Based DT Detector.

We first note that the NCC detector miscalculates the far-end only speech as DT during the time interval from 1 msec to about 8 msec. This halts the learning of the adaptive filter. The performance of NCC-based detector is very close to that of ZCR-based detector during the double talk interval. However, as the ZCR-based detector continues to further converge after about 22 sec, the NCC-based detector does not. As far as complexity and memory requirements are concerned, the cross-correlation based methods, require at least $O(L^2)$ multiplications at each iterations. The ZCR-based method, on the other hand, requires only sign comparison at each iteration.

6. PRACTICAL CONSIDERATIONS

The performance of ZCR detector is dependent on the parameters M, K, and γ. A shorter window length M provides less smoothing and is thus prone to making an incorrect decision when the ZCR happens to be close to the threshold. On the other hand, longer window will result in smoothed ZCR estimate and may help in suppressing any error due to ZCR being close to the threshold. However, a longer window may result in false decisions, especially in locating the true start and end of DT regions. Our experiments showed that a window of size 1000 samples (125 msec at 8kHz) is suitable to ease the trade-off. The parameter K determines the number of samples after which the ZCR is updated. Ideally, $K = 1$ is desired when the ZCR is updated with each incoming sample. For $K = 1$, M bits are needed to store M zero crossing decisions. On the other hand, memory and computational savings can be made if $K > 1$ is chosen. Finally, the threshold γ is carefully chosen to avoid any mis-detection or false alarm. Our experiments with various test signals showed that $\gamma = 0.45$ to 0.5 is suitable for a window size of $M = 1000$.

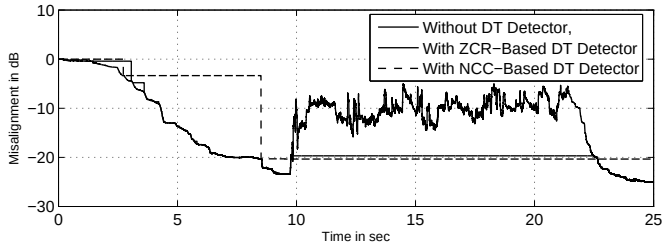

Fig. 5. Misalignment of the NLMS-based AEC (a) without a DT detector, (2) with a ZCR-based DTD, and (3) with a NCC-based DT detector. Note that the result without the DT detector are the same as in Fig. 3.

7. CONCLUSION

We employed short-time ZCR to detect and control DT in time-domain AEC. The zero crossings are extremely simple to compute and proves to be an effective discriminant of DT. The DT detector based on the ZCR is used to drive the AEC filter, halting its adaptation when DT is detected. The detector operation is governed by three parameters that are easy to adjust and requires minimal tuning. Because of its simplicity in computation and ease of operation, the ZCR-based AEC is an excellent choice for implementation on a low-power and low-resource DSP. It has also shown improved performance over the well-known NCC-based detector.

8. REFERENCES

[1] S. Gay and J. Benesty, Eds., *Acoustic Signal Processing for Telecommunication,* Norwell, MA: Kluwer Academic, 2000.

[2] M. Z. Ikram, "Blind source separation and acoustic echo cancellation: A unified framework," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2012, pp. 1701–1704.

[3] T. Gänsler, S. L. Gay, M. M. Sondhi, and J. Benesty, "Double-talk robust fast converging algorithms for network echo cancellation," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 656–663, Nov. 2000.

[4] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-Correlation", *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 2, pp. 168–172, Mar. 2000.

[5] K.-H. Lee, J.-H. Chang, N. S. Kim, S. Kang, and Y. Kim, "Frequency-domain double-talk detection based on the Gaussian mixture model," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 453–456, May 2010.

[6] H. Buchner, J. Benesty, T. Gänsler, and W. Kellermann, "Robust extended multidelay filter and double-talk detector for acoustic echo cancellation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1633–1643, Sept. 2006.

[7] C. Schüldt, F. Lindstrom, and I. Claesson, "A delay-based double talk detector," *IEEE Trans. Speech and Audio Processing*, vol. 20, no. 6, pp. 1725–1733, Aug. 2012.

[8] C. Paleologu, S. Ciochină, and J. Benesty, "Double-talk robust VSS-NLMS algorithm for under-modeling acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2008, pp. 245–248.

[9] S. Y. Low, S. Vekatesh, and S. Nordholm, "A spectral slit approach to doubletalk detection," *IEEE Trans. Speech and Audio Processing*, vol. 20, no. 3, pp. 1074–1080, March 2012.

[10] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, Upper Saddle River, NJ: Prentice-Hall, 2011.

[11] R. Cheng, C. Zhang, and C. Wei, "Method and apparatus for double-talk detection," U.S. Patent 8 160 238 B2, Apr. 17, 2012.