COHERENCE BASED DOUBLE TALK DETECTOR WITH ADAPTIVE THRESHOLD

Ivan Tashev

Microsoft Research, One Microsoft Way, Redmond, WA 98034, USA
ivantash@microsoft.com

ABSTRACT

Acoustic echo cancellation is one of the oldest applications of the adaptive filters and today part of each speakerphone. An important block of each acoustic echo canceller is the double talk detector. It blocks the adaptation of the filter when near end voice is present and thus preventing the adaptive filter from diverging from the optimal position. In this paper we present an improved version of coherence based double talk detector with improved precision compared to the base algorithm.

***Index Terms* —** Acoustic echo cancellation, double talk detector, coherence

1. INTRODUCTION

Acoustic echo cancellers (AEC) [1] are designed to remove the captured loudspeaker signal from the microphone channel of a speakerphone or another telecommunication device. The AEC consists of an adaptive filter, which estimates the transfer path between the loudspeaker channel and the microphone channel, convolves the loudspeaker signal with this transfer path and subtracts it form the microphone channel. Under absence of near end speech the adaptive filter converges to the closest estimation of the transfer path. The precision of this convergence depends on the noise in the microphone channel. When we have a local speech the adaptive filter diverges from this optimal position. The purpose of the double talk detector (DTD) is to detect the segments with the local speech and block the adaptation of the acoustic echo canceller.

The generic DTD computes a certain statistical parameter ξ, preferably data independent, which is compared with a threshold η. If the value is higher than the threshold, double talk is detected, if it is below – there is no double talk. The threshold value can be adjusted using the ROC (receiver operating characteristics) curves to provide maximum performance. Good overview for DTD evaluation criteria is given in [2]. One of the first DTD algorithms is the Geigel algorithm, which evaluates the proportion of the largest magnitude of the microphone signal for a given time interval and the magnitude of the loudspeaker signal. The optimal threshold is highly variable and the reliability of the DTD is low. Cross-correlation based algorithms are considered more robust and reliable. The problem with this class of algorithms is that the cross-correlation function is not very well normalized and it is not quite robust when noise is present. A DTD algorithm using the normalized cross-correlation function is derived in [3]. While more precise it is computationally expensive, which led to publishing a faster version of it [4] based on tracking with a Kalman filter. While substantially faster it is still computationally expensive. Instead of using the cross-correlation function as a statistical variable, the coherence function can be used [5]. The coherence function between the loudspeaker and microphone channels is easy to compute and is well normalized. Values close to one mean that microphone and loudspeaker signals are coherent and there is no local speech. Under presence of local speech the values of the coherence function decrease and approach zero, which makes it a good statistical parameter for DTD. Unfortunately the coherence function value decreases under the presence of noise or strong reverberation, which makes this method less suitable for cases when the microphone is away from the loudspeaker and/or high levels of noise are presented.

In this paper we present a modified version of the coherence based DTD. When the loudspeaker signal is presented adaptively we track the maximal values of the coherence function and scale the decision threshold between zero and the maximal value. The new algorithm is more robust to noise and reverberation. It was evaluated against a data corpus with wide range of noise levels and compared with the original version of the algorithm. The proposed approach improves the precision of the DTD by 5.5% relatively compared to the original version of the algorithm.

2. MODELING

A schematic diagram of an AEC is shown in Figure 1. The far end signal $z(t)$ is sent to the loudspeaker. The microphone captures this signal convolved with the impulse response of the transfer path speaker-microphone $h(t)$. It captures the local voice $s(t)$ and noise $n(t)$. The transfer path local speaker-microphone is omitted for simplicity. The microphone signal is:

$$x(t)=z(t)*h(t)+s(t)+n(t). \tag{1}$$

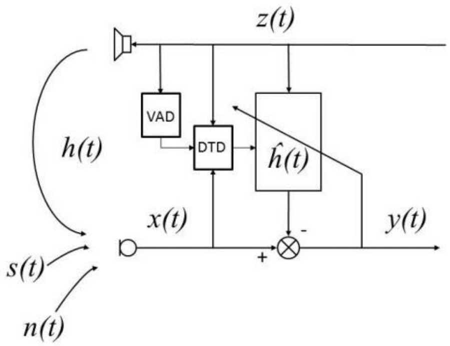

Figure 1. Schematic diagram of acoustic echo canceller.

The acoustic echo canceller estimates the transfer path loudspeaker-microphone $\hat{h}(t)$ and subtracts the estimated portion of the loudspeaker signal from the microphone signal. At the acoustic echo canceller output we have:

$$\begin{aligned} y(t) &= x(t) - z(t) * \hat{h}(t) = \\ &= z(t) * h(t) - z(t) * \hat{h}(t) + s(t) + n(t). \end{aligned} \tag{2}$$

In this paper we consider processing in frequency domain and then the convolution converts to multiplication and we have:

$$\begin{aligned} Y_k^{(n)} &= Z_k^{(n)} H_k^{(n)} - Z_k^{(n)} \hat{H}_k^{(n)} + S_k^{(n)} + N_k^{(n)} = \\ &= Z_k^{(n)} \left(H_k^{(n)} - \hat{H}_k^{(n)} \right) + S_k^{(n)} + N_k^{(n)}. \end{aligned} \tag{3}$$

Here k is the frequency bin and n is the frame number. The modeling described so far assumes that the audio frame is longer than the reverberation process, which is incorporated in $h(t)$, and we model it with one tap filter for each frequency bin. This is not the case with real systems with typical frame duration of 10-30 ms and reverberation times of 200-400 ms. To accommodate the longer impulse response the acoustic echo canceller uses an FIR filter with multiple taps for each frequency bin. This converts equation (3) to:

$$\begin{aligned} Y_k^{(n)} &= \sum_{i=0}^{L-1} Z_k^{(n-i)} H_k^{(n-i)} - \sum_{i=0}^{L-1} Z_k^{(n-i)} \hat{H}_k^{(n-i)} Z_k^{(n)} + S_k^{(n)} + N_k^{(n)} \\ &= \sum_{i=0}^{L-1} Z_k^{(n-i)} \left(H_k^{(n-1)} - \hat{H}_k^{(n-1)} \right) + S_k^{(n)} + N_k^{(n)}, \end{aligned} \tag{4}$$

where L is the number of taps in the FIR filter. Denoting:

$$\begin{aligned} \mathbf{Z}_k^{(n)} &= \left[Z_k^{(n)}, Z_k^{(n-1)}, \ldots, Z_k^{(n-L+1)} \right]^T \\ \mathbf{H}_k^{(n)} &= \left[H_k^{(n)}, H_k^{(n-1)}, \ldots, H_k^{(n-L+1)} \right]^T \\ \mathbf{X}_k^{(n)} &= \left[X_k^{(n)}, X_k^{(n-1)}, \ldots, X_k^{(n-L+1)} \right]^T \end{aligned} \tag{5}$$

the equation (4) can be rewritten in vector form:

$$Y_k^{(n)} = \left(\mathbf{H}_k^{(n)} \right)^T \mathbf{Z}_k^{(n)} - \left(\hat{\mathbf{H}}_k^{(n)} \right)^T \mathbf{Z}_k^{(n)} + S_k^{(n)} + N_k^{(n)}. \tag{6}$$

The squared magnitude of the coherence function between $\mathbf{Z}^{(n)}$ and $\mathbf{X}^{(n)}$ for the frequency bin k is:

$$\gamma_{ZX}^2(k) \triangleq \frac{\left| S_{ZX}(k) \right|^2}{S_{ZZ}(k) S_{XX}(k)}, \tag{7}$$

where $S_{AB} \triangleq \mathbf{A}\mathbf{A}^H$ are the spectral densities. Then the statistical parameter $\xi^{(n)}$ for the entire frame can be computed as a weighted sum $\xi^{(n)} = \sqrt{\mathbf{W} \left(\boldsymbol{\gamma}_{ZX}^{2\ (n)} \right)^T}$. Typically the weighting vector $\mathbf{W}$ is a band-pass filter and the statistical parameter is computed as a partial sum:

$$\xi^{(n)} = \sqrt{\frac{1}{K_{end} - K_{beg}} \sum_{k=K_{beg}}^{K_{end}-1} \gamma_{ZX}^{2\ (n)}(k)}. \tag{8}$$

Then the statistical parameter is compared to a threshold η to make the final decision:

$$D^{(n)} = \begin{vmatrix} 1 & \text{when} & \xi^{(n)} < \eta - \Delta\eta \\ 0 & \text{when} & \xi^{(n)} > \eta + \Delta\eta \\ D^{(n-1)} & \text{otherwise} & \end{vmatrix} \tag{9}$$

Here $\Delta\eta$ introduces a small hysteresis to prevent "ringing" in the slopes. If $D^{(n)}$ is 1 we have double talk detected in this frame, if zero – no double talk was detected.

3. PROPOSED ALGORITHM

The main problem with the algorithm above is that the statistical parameter $\xi^{(n)} \in [0,1]$ goes to one only in close to perfect conditions: no noise and reverberation. When noise is added to the microphone signal the value of $\xi^{(n)}$ is higher than when a double talk is present, but doesn't go to one and varies based on the noise and reverberation levels. This makes the optimal threshold η for one input SNR suboptimal for another. In low SNRs the DTD stops to work at all. This is why we propose to adaptively track the maximal value of the statistical parameter:

$$\xi_{MAX}^{(n)} = \begin{vmatrix} \left(1 - \dfrac{T}{\tau_{up}} \right) \xi_{MAX}^{(n-1)} + \dfrac{T}{\tau_{up}} \xi^{(n)} & \xi^{(n)} > \xi_{MAX}^{(n-1)} \\ \left(1 - \dfrac{T}{\tau_{down}} \right) \xi_{MAX}^{(n-1)} + \dfrac{T}{\tau_{down}} \xi^{(n)} & \xi^{(n)} \le \xi_{MAX}^{(n-1)} \end{vmatrix}. \tag{10}$$

Here T is the audio frame duration, τ_{up} and τ_{down} are the two different time constants. If $\tau_{up} > \tau_{down}$ this double time constant integrator will track the higher values of $\xi^{(n)}$. Then the threshold can be estimated as:

$$\eta = \max\left(\eta_{\min}, \xi_{MAX}^{(n)} \eta_{ABS} \right). \tag{11}$$

Here $\eta_{ABS} \in [0,1]$ is the absolute threshold, η_{MIN} is forcing the values of the threshold to stay above a certain minimum which can happen in very low SNRs. Under these conditions

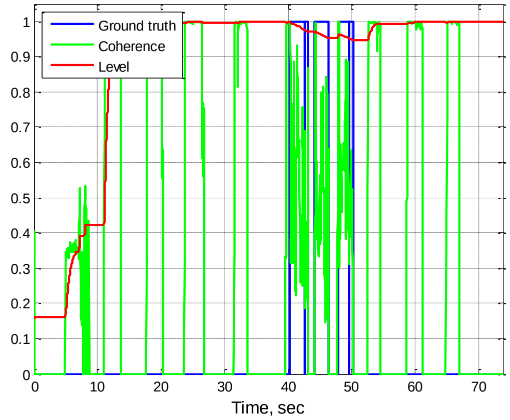

Figure 2. Coherence function and its tracking.

the threshold η_{ABS} is close to optimal in wider range of SNRs. Illustration how the proposed algorithm works is shown in Figure 2.

4. EXPERIMENTAL RESULTS

The proposed algorithm was evaluated and compared with a the baseline algorithm using a data corpus containing two noise levels (40 and 50 dBC SPL, automotive noise), two levels of the near end and far end signals (60 and 54 dBC SPL at 1 meter), played by two high quality loudspeakers in normal office reverberation conditions $\left(RT_{60} = 230 \text{ ms}\right)$. The loudspeakers and the microphones formed a triangle with sides of one meter each. All combinations of the noise, near, and far end signal levels produced eight combinations. Training and testing sets with all of the combinations were recorded. The near and far end signals were human speech, ten sentences each, equally mixed male and female voices, with pauses between them shifted in a way to produce partial and full overlap. The ground truth was established by running the clean near and far end speech signals trough a precise voice activity detector [6]. The binary decision "speech/no speech" for the two signals was compared and double talk marked for the frames where both VAD indicated speech activity.

The classification error was selected as evaluation parameter, defined as:

$$\varepsilon = \frac{N_{FP} + N_{FN}}{N_{Tot}}.100\%. \qquad (12)$$

Here N_{FP} is the number of false positives, N_{FN} is the number of false negatives, and N_{Tot} is the total number of audio frames.

The sampling rate was 16 kHz, 512 samples per audio frame, and we used overlap and add process as described in [7]. With the training set was conducted optimization to minimize the error rate by varying the values of the DTD parameters. After the optimization the values were $\tau_{up} = 0.55$ s, $\tau_{down} = 45$ s, $F_{beg} = 730$ Hz, $F_{end} = 7200$ Hz, $\eta = 0.95$, and $\eta_{MIN} = 0.15$. For the baseline algorithm the optimal threshold value was $\eta = 0.89$.

Table 1. Results, baseline and proposed algorithms

Algorithm	Equations	Error	N_{FP}	N_{FN}	T_{Tot}
Baseline	(8),(9)	2.94%	575	511	36920
Proposed	(8),(10),(11),(9)	2.45%	590	313	36920

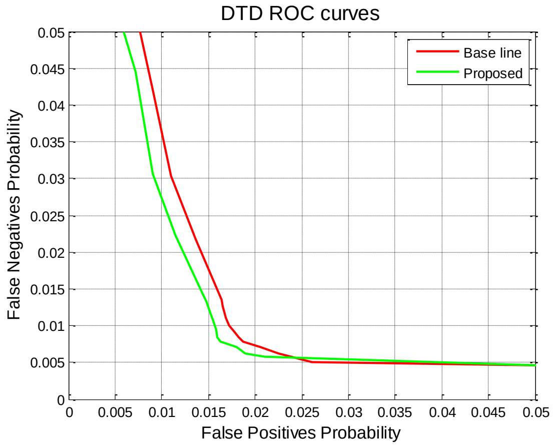

Figure 3. ROC curves

All further results were obtained against the testing set of recording. The results for the baseline and proposed algorithms are shown in Table 1, the ROC curves – in Figure 3. The proposed algorithm has lower error rate and is better in reducing the false negatives, preventing better AEC to diverge during the double talk situations.

5. REFERENCES

[1] M. Sondhi, "An adaptive echo canceller", Bell Syst. Tech. Journal, Vol. 46, pp. 497-511, March 1967.

[2] J. Cho, D. Morgan, J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancellers", IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 6, Nov. 1999.

[3] J. Benesty, D. Morgan, J. Cho, "A New Class of Doubletalk Detectors Based on Cross-Correlation", IEEE Transactions on Speech and Audio Processing, vol. 8, No. 2, pp. 168-172, March 2000.

[4] J. Benesty, T. Gänsler, "The fast cross-correlation double-talk detector". Signal Processing, vol. 86, No. 6, pp. 1124-1139, Elsevier 2006.

[5] T. Gänsler, M. Hansson, C.-J. Invarsson, G. Salomonsson, "A double-talk detector based on coherence", IEEE Transactions on Communications, Vol. 44, No. 11, pp. 1241-1247, Dec. 1996.

[6] I. Tashev, A. Lovitt, A. Acero, "Dual stage probabilistic voice activity detector", in proceedings of NOISE-CON 2010 and 159th Meeting of the Acoustical Society of America, Acoustical Society of America, 20 April 2010.

[7] I. Tashev, *Sound Capture and Processing: Practical Approaches*, pp. 388, Wiley, July 2009.