ONLINE DOUBLETALK DETECTOR CALIBRATION FOR ACOUSTIC ECHO CANCELLATION IN VIDEOCONFERENCING SYSTEMS

James D. Gordy and Rafik A. Goubran

Department of Systems and Computer Engineering
Carleton University, Ottawa, Canada
{jdgordy, goubran}@sce.carleton.ca

ABSTRACT

This paper addresses the problem of doubletalk detector calibration for acoustic echo cancellers in hands-free environments such as videoconferencing. A statistical model of a recently proposed doubletalk detector is used to show that optimal detection thresholds are dependent on the input signal and the adaptive filter error. A signal-adaptive algorithm is proposed for calculating an optimal threshold for arbitrary input signals and echo path environments. Simulation results verify the improvement in detection probability offered by the proposed algorithm compared to simple empirical calibration methods.

1. INTRODUCTION

Acoustic echo cancellation is a vital component of full-duplex videoconferencing systems. In such environments, undesirable acoustic echoes arise from direct-path coupling between the loudspeaker and microphone, and from reverberation within the room itself [1]. A block diagram of an acoustic echo canceller in a typical videoconferencing system is shown in Figure 1. In this configuration the adaptive echo canceller tracks the echo path system by modeling it as a linear system, and uses this model to cancel echo from the microphone signal [2]. In practice, a doubletalk detector is required to sense the presence of near-end speech in the microphone signal. The resulting doubletalk decision is used as a control mechanism to halt adaptation for the duration of the disturbance. If adaptation is not halted during doubletalk periods, most adaptation algorithms will diverge after only a few samples.

A number of doubletalk detection algorithms exist in the literature, several of which are reviewed and compared in [3]. Recently a doubletalk detector was proposed based on the normalized cross-correlation between the input and reference signals [4]. Simulations of the algorithm showed a higher detection probability than previous algorithms while maintaining a low complexity. However, a common problem among such algorithms is that of choosing optimal calibration parameters suitable for arbitrary acoustic environments. In this paper the problem of doubletalk detector calibration for the algorithm in [4] is addressed. A review of the doubletalk detector is presented in Section 2, and in Section 3 an online signal-adaptive calibration algorithm is derived. Simulation results with speech input and doubletalk signals are presented in Section 4.

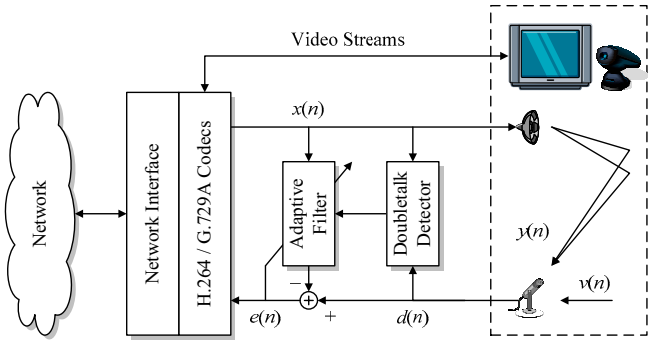

Figure 1 – Block diagram of a typical acoustic echo canceller in the context of a videoconferencing system.

2. DOUBLETALK DETECTION

2.1. Echo Canceller Structure and Conventions

A block diagram of a typical acoustic echo canceller and doubletalk detector is shown in Figure 1. The far-end input signal $x(n)$ is played over the loudspeaker at the near end, and an undesirable echo of the input signal, $y(n)$, is picked up by the microphone. The echo path consisting of the loudspeaker, room, and microphone is modeled as a linear system with an impulse response of length N samples. The microphone signal $d(n)$ consists of the echo signal, near-end speech $v(n)$, and background noise $\eta(n)$ as follows:

$$y(n) = \underline{x}^T(n)\underline{h}(n) \quad (1)$$

$$d(n) = y(n) + v(n) + \eta(n) \quad (2)$$

where $\underline{x}(n) = [x(n)\ x(n-1)\ \cdots\ x(n-N+1)]^T$ and $\underline{h}(n) = [h_0(n)\ h_1(n)\ \cdots\ h_{N-1}(n)]^T$ is the time-varying impulse response

vector at time n. Assuming a linear echo path impulse response, the adaptive echo canceller models and tracks the echo path as a finite impulse response (FIR) of length $M \leq N$ samples. The echo canceller output $e(n)$ is obtained from:

$$e(n) = d(n) - \underline{x}^T(n)\underline{\hat{h}}(n) = \underline{x}^T(n)\Delta\underline{h}(n) + v(n) + \eta(n) \quad (3)$$

where $\underline{\hat{h}}(n) = [\hat{h}_0(n) \quad \hat{h}_1(n) \quad \cdots \quad \hat{h}_{N-1}(n)]^T$ is the adaptive filter coefficient vector at time n and $\Delta\underline{h}(n)$ is the corresponding error vector. For simplicity it is assumed that $M = N$ and the adaptive filter coefficients are updated using an algorithm such as normalized LMS [5].

2.2. Cross-Correlation-Based Doubletalk Detection

In this section the normalized cross-correlation-based doubletalk detector of [4] is briefly reviewed. Assume that the echo path is stationary and there is no doubletalk. For a stationary input signal and room impulse response, the expected variance of (1) can be written in terms of the $N \times 1$ impulse response vector $\underline{h}$ and $\underline{R}_{xx}$, the $N \times N$ autocorrelation matrix of the input signal:

$$\sigma_d^2 = E\{d^2(n)\} = \underline{h}^T \underline{R}_{xx} \underline{h} \quad (4)$$

Assuming that the echo signal $y(n)$, near-end speech $v(n)$, and background noise $\eta(n)$ are uncorrelated, the $N \times 1$ cross-correlation vector $\underline{r}_{xd}$ between the input and microphone signals can be written in a similar manner:

$$\underline{r}_{xd} = E\{\underline{x}(n)d(n)\} = \underline{R}_{xx}\underline{h} \quad (5)$$

Solving for $\underline{h}$ in (5) and substituting the result in (4) yields the expected microphone signal variance in terms of the autocorrelation matrix and cross-correlation vector:

$$\sigma_d^2 = \underline{r}_{xd}^T \underline{R}_{xx}^{-1} \underline{r}_{xd} \quad (6)$$

A normalized doubletalk detection variable ξ is obtained by dividing the actual microphone signal variance into (6) and taking the square root of the result:

$$\xi = \sqrt{\underline{r}_{xd}^T \underline{R}_{xx}^{-1} \underline{r}_{xd} \Big/ \sigma_d^2} \quad (7)$$

In the absence of doubletalk the numerator and denominator terms are equal and $\xi = 1$. When doubletalk is present, the actual microphone signal variance will be larger and $\xi < 1$.

In practice the input and near-end speech signals are time-varying, so the parameters used to calculate (7) must be estimated and tracked. Note that a direct implementation requires constructing and updating an estimate of the autocorrelation matrix and its inverse. This is infeasible for long impulse responses typical of acoustic environments (upwards of 250 ms). One solution is to use the adaptive filter coefficients, assuming they have converged to a certain degree. In addition, the cross-correlation vector of (5) and the microphone signal variance may be estimated by averaging over a window of K samples. In particular:

$$\underline{R}_{xx}^{-1}\underline{r}_{xd} = \underline{h} \approx \underline{\hat{h}}(n) \quad (8)$$

$$\underline{\hat{r}}_{xd}(n) = \tfrac{1}{K}\textstyle\sum_{k=0}^{K-1} \underline{x}(n-k)d(n-k) \quad (9)$$

$$\hat{\sigma}_d^2(n) = \tfrac{1}{K-1}\textstyle\sum_{k=0}^{K-1}[d(n-k) - \tfrac{1}{K}\sum_{j=0}^{K-1} d(n-j)]^2 \quad (10)$$

Substituting (8) – (10) into (7) results in the estimated doubletalk detection variable at time n:

$$\xi(n) = \sqrt{\underline{\hat{r}}_{xd}^T(n)\underline{\hat{h}}(n) \Big/ \hat{\sigma}_d^2(n)} \quad (11)$$

3. DOUBLETALK DETECTOR CALIBRATION

3.1. Optimal Doubletalk Detection Threshold

Doubletalk detection is typically viewed as a statistical test with the hypotheses H_0 and H_1 that doubletalk is and is not present, respectively. The estimated parameters of (11) are noisy, so $\xi(n)$ is a random variable with a probability distribution function (PDF) ideally centered at unity in the absence of doubletalk. Therefore, a doubletalk decision can be made by comparing $\xi(n)$ to a threshold T:

$$\xi(n) < T \Rightarrow H_0 \quad \xi(n) > T \Rightarrow H_1 \quad (12)$$

An important question is how to select an appropriate detection threshold in practice. Too low or too high a threshold will increase either the probability of miss (Type I error) or the probability of false alarm (Type II error), respectively. If the near-end speech characteristics are not known, one approach is to choose T based on a given probability of false alarm P_F in the absence of doubletalk:

$$P(\xi(n) < T \mid H_1) = P_F \quad (13)$$

3.2. Noise and Bias Compensation

In practice, the echo path impulse response $\underline{h}(n)$ is at least slowly time-varying, introducing error into the adaptive filter coefficient vector of (8). In addition, the presence of continuous background noise in the environment will increase the actual microphone signal variance of (10) [6]. The resulting bias in the numerator of (11) due to the adaptive filter error can be written in terms of the cross-correlation vector between the input and error signals $x(n)$ and $e(n)$, and the residual echo signal variance:

$$\underline{r}_{xd}^{T}(n)\Delta\underline{h}(n) = \underline{r}_{xe}^{T}(n)\hat{\underline{h}}(n) + \sigma_{\delta}^{2}(n) \quad (14)$$

In the absence of doubletalk, the terms of (14) can be estimated using the error signal $e(n)$ over a window of samples in a manner similar to (9) and (10). The terms can be used to compensate for the bias in (11):

$$\hat{\underline{r}}_{xe}(n) = \tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\underline{x}(n-k)e(n-k) \quad (15)$$

$$\hat{\sigma}_{e}^{2}(n) = \tfrac{1}{K-1}\textstyle\sum_{k=0}^{K-1}[e(n-k) - \tfrac{1}{K}\sum_{j=0}^{K-1}e(n-j)]^{2} \quad (16)$$

$$\xi(n) = \sqrt{[\hat{\underline{r}}_{xd}^{T}(n)\hat{\underline{h}}(n) + \hat{\underline{r}}_{xe}^{T}(n)\hat{\underline{h}}(n) + \hat{\sigma}_{e}^{2}(n)]\big/\hat{\sigma}_{d}^{2}(n)} \quad (17)$$

Note that estimating σ_{δ}^{2} using the error signal also compensates for the presence of background noise in the microphone signal variance.

3.2. Online Detection Threshold Calculation

Equation (13) assumes knowledge of the doubletalk detection variable's conditional PDF in the absence of doubletalk. However, (17) shows that the PDF is typically time-varying and dependent on the input signal statistics, and as a result an optimal detection threshold constructed using (13) would be adaptive as well. In [3] an empirical calibration method is described for selecting a general detection threshold by averaging over one or more sets of training input signals. However, in [7] the authors derived a statistical model of the cross-correlation-based doubletalk detector of [4] by analyzing the parameter estimators of (8) – (10) and (15) – (17), and it is employed here.

Let $\Delta(n)$ be the error between the true and estimated bias introduced by the adaptive filter error and noise at time n:

$$\Delta(n) = [\underline{r}_{xe}(n) - \hat{\underline{r}}_{xe}(n)]^{T}\hat{\underline{h}}(n) + \sigma_{e}^{2}(n) - \hat{\sigma}_{e}^{2}(n) \quad (18)$$

The doubletalk detection variable of (13) can be represented as a function of the ratio of (18) to the actual microphone signal variance of (10):

$$\xi(n) \approx \sqrt{1 - \Delta(n)\big/\hat{\sigma}_{d}^{2}(n)} \quad (19)$$

In the absence of doubletalk, the two terms of (19) can be modeled as independent Gaussian random processes. This results in a simple expression for the corresponding conditional PDF of (13):

$$f_{\Xi}(\xi \mid H_{1}) = |2\xi| f_{Z}(1-\xi^{2}) \quad (20)$$

for $\xi > 0$, where $f_Z(z)$ is the PDF formed by A / B, the ratio of two independent Gaussian random variables with means μ_A and μ_B and variances σ_A^2 and σ_B^2, respectively [7]:

$$f_{Z}(z) = \frac{1}{\sqrt{2\pi}\sigma_{A}\sigma_{B}}\left\langle \frac{e^{-C/2}}{\sqrt{2\pi}A(z)} + \frac{B(z)}{A(z)^{3/2}}\exp\left\{-\frac{1}{2}\left[C - \frac{B^{2}(z)}{A(z)}\right]\right\} Erf\left[\frac{B(z)}{\sqrt{2A(z)}}\right]\right\rangle \quad (21)$$

$$A(z) = \sigma_{A}^{-2}z^{2} + \sigma_{B}^{-2} \quad (22)$$

$$B(z) = \mu_{A}\sigma_{A}^{-2}z + \mu_{B}\sigma_{B}^{-2} \quad (23)$$

$$C = \mu_{A}^{2}\sigma_{A}^{-2} + \mu_{B}^{2}\sigma_{B}^{-2} \quad (24)$$

To characterize the PDF of (20) it is necessary to know the first- and second-order statistics of the two terms in (19). If it is assumed that the cross-correlation vector terms and the residual echo signal are independent, then in general $\Delta(n)$ is approximately Gaussian with zero mean and dominated by the error signal variance. The denominator of (19) can be represented as Gaussian with mean and variance obtained from the microphone signal variance [7]:

$$\mu_{A} = E[\Delta(n)] = 0 \quad (25)$$

$$\sigma_{A}^{2} = VAR[\Delta(n)] \approx 2\sigma_{e}^{4}(n)\big/(K-1) \quad (26)$$

$$\mu_{B} = E[\hat{\sigma}_{d}^{2}(n)] = \sigma_{d}^{2}(n) \quad (27)$$

$$\sigma_{B}^{2} = VAR[\hat{\sigma}_{d}^{2}(n)] = 2\sigma_{d}^{4}(n)\big/(K-1) \quad (28)$$

Since the conditional PDF of the doubletalk detection variable is specified in terms of measured parameters, an adaptive threshold can be constructed using (13) and the cumulative density function (CDF) of (20).

Finally, it should be noted that the onset of doubletalk conditions will cause an abrupt increase in the error signal variance. To guard against this, the short-term correlation of $e(n)$ is estimated using the technique in [6] and used to freeze estimation during potential doubletalk conditions.

4. SIMULATION RESULTS

The goal of these experiments was to evaluate the signal-adaptive doubletalk detector calibration algorithm against simpler empirical calibration techniques. An impulse response was obtained from a conference room measuring approximately 4×6 meters in size, and truncated to a length of N = 500 samples. Five pairs of input and near-end speech samples were obtained from the TIMIT continuous speech database and downsampled to f_s = 8 kHz [8]. Variability in the echo path impulse response $\underline{h}(n)$ was simulated by modulating each coefficient by a zero-mean Gaussian noise process with variance 0.01. White background noise was added to produce a microphone SNR of -30 dB, and an estimation window of K = 200 samples was used to calculate the doubletalk detector parameters of (9), (10), (15), and (16). At each sample interval n the signal-adaptive technique of Section 3 was used to

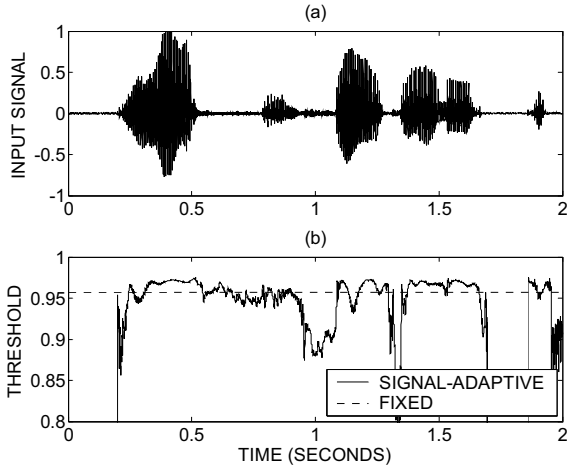

Figure 2 – (a) Speech input signal; (b) Fixed and signal-adaptive doubletalk detection thresholds ($P_F = 0.1$).

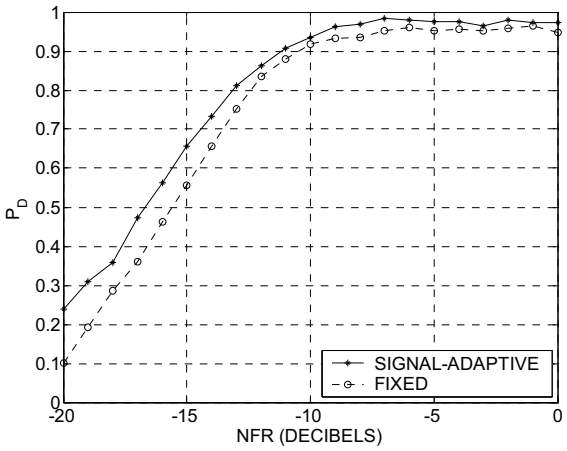

Figure 3 – Probability of detection (P_D) as a function of NFR for fixed and signal-adaptive detection thresholds ($P_F = 0.1$).

determine an optimal doubletalk detection threshold $T(n)$ for $P_F = 0.1$. For comparison, a fixed detection threshold T_{FIXED} was calculated using the approach in [3] by averaging over the set of five speech input signals. Near-end speech signals were added to the microphone signal to simulate doubletalk conditions. The near-end speech power was adjusted relative to that of the input signal to attain a desired segmental near- to far-end signal power ratio (NFR).

Figure 2(a) shows a plot of one of the speech input signals, and Figure 2(b) shows the corresponding signal-adaptive doubletalk detection threshold $T(n)$ compared to the fixed detection threshold T_{FIXED}. It is clear from this figure that the two thresholds differ considerably, and note that in several time intervals the fixed threshold is lower than the adaptive threshold. As a result, one would expect the actual false alarm probability to be lower using T_{FIXED}, which is confirmed by the results in Table I. One would also expect the resulting probability of detection P_D to be degraded in the presence of low-level doubletalk. Figure 3 shows a plot of the probability of detection P_D as a function of NFR ranging from -20 dB to 0 dB for $P_F = 0.1$, and averaged over the five pairs of input and near-end speech signals. From this figure it is clear that for low-to-moderate levels of near-end speech, the signal-adaptive detection threshold calibration algorithm provides a higher probability of detection than the fixed detection threshold.

5. CONCLUSIONS

This paper investigated the practical issue of doubletalk detector calibration in hands-free environments, and a signal-adaptive detection threshold algorithm was devised based on statistical analysis. Simulation results comparing the adaptive threshold with a fixed threshold revealed an improvement in detection probability.

Table I – Average probability of false alarm for five speech input signals using fixed detection threshold obtained for $P_F = 0.1$.

Input	1	2	3	4	5
P_F	0.116	0.084	0.093	0.089	0.087

ACKNOWLEDGEMENT

This work was supported by the Ontario Centres of Excellence and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] J. D. Gordy and R. A. Goubran, "A perceptual performance measure for adaptive echo cancellers in packet-based telephony," in *Proc. IEEE ICME*, July 2005, pp. 157 – 160.

[2] C. 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, July 1999.

[3] J. H. Cho, D. R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancellers," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 6, pp. 718 – 724, Nov. 1999.

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

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

[6] A. Sugiyama, J. Berclaz, and M. Sato, "Noise-robust double-talk detection based on normalized cross correlation and a noise offset," in *Proc. IEEE ICASSP*, Mar. 2005, vol. 3, pp. 153 – 156.

[7] J. D. Gordy and R. A. Goubran, "Statistical analysis of doubletalk detection for calibration," submitted to *IEEE Trans. Audio, Speech, Lang. Processing.*

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