COHERENCE BASED DOUBLE TALK DETECTOR WITH SOFT DECISION

Ivan J. Tashev

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

ABSTRACT

Acoustic echo cancellation is one of the oldest applications of adaptive filters and today is a 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. It provides estimation of the double talk presence probability per bin and per frame and has better precision compared to the baseline 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 function between the loudspeaker channel and the microphone channel, convolves the loudspeaker signal with this transfer function, and subtracts it from the microphone channel. Under absence of near end speech the adaptive filter converges to the closest estimation of the transfer function. The precision of this convergence depends on the noise in the microphone channel. When we have local speech the adaptive filter diverges from this optimal position. The purpose of the double talk detector (DTD) is to detect the segments of 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 receiver operating characteristics (ROC) curves to provide maximum performance. A 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 this algorithm 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 value of the coherence function decreases and approaches 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 present.

In this paper we present a modified version of the coherence based DTD. We build statistical models of the coherence function distribution, a classifier, and use first order HMM filter for smoothing. 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 more than two times 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 function speaker-microphone $h(t)$. It captures the near end voice $s(t)$ and noise $n(t)$. The transfer function near end speaker-microphone is omitted for simplicity. The microphone signal is:

$$x(t) = z(t) * h(t) + s(t) + n(t). \quad (1)$$

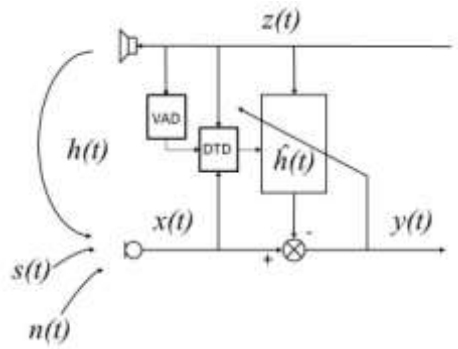

Figure 1. Schematic diagram of acoustic echo canceller.

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 where 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 a one tap filter. 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:

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

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:

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

where $S_{AB} \triangleq \mathbf{A}\mathbf{B}^H$ 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:

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

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. When a loudspeaker signal is present, two hypotheses can be considered for the current frame and bin:

$$\begin{aligned} H_0: &\quad \text{no double talk:} \quad X_k^{(n)} = \mathbf{H}_k^{(n)T} \mathbf{Z}_k^{(n)} + N_k^{(n)} \\ H_1: &\quad \text{double talk:} \quad X_k^{(n)} = \mathbf{H}_k^{(n)T} \mathbf{Z}_k^{(n)} + S_k^{(n)} + N_k^{(n)} \end{aligned}. \tag{10}$$

We model the distribution of the statistical parameter $\xi_k^{(n)} = \gamma_{ZX}^2(k)$ as Gaussian in both cases:

$$p\left(\xi_k^{(n)} \middle| H_0\right) = \frac{1}{\sqrt{2\pi\lambda_N}} \exp\left(-\frac{\left(\xi_k^{(n)} - \bar{\xi}_N\right)^2}{2\lambda_N} \right), \tag{11}$$

$$p\left(\xi_k^{(n)} \middle| H_1\right) = \frac{1}{\sqrt{2\pi\lambda_D}} \exp\left(-\frac{\left(\xi_k^{(n)} - \bar{\xi}_D\right)^2}{2\lambda_D} \right). \tag{12}$$

Here $\bar{\xi}_N, \lambda_N, \bar{\xi}_D$, and λ_D are the means and variances of the statistical parameter without and with double talk. Then given value of the statistical parameter ξ_k, after applying the Bayesian rule, the probability to have double talk is:

$$P_k\left(H_1 \mid \xi_k\right) = \frac{p\left(\xi_k \mid H_1\right)P_k\left(H_1\right)}{p\left(\xi_k \mid H_1\right)P_k\left(H_1\right) + p\left(\xi_k \mid H_0\right)P_k\left(H_0\right)} = \frac{\varepsilon_k \Lambda_k}{1+\varepsilon_k \Lambda_k}. \quad (13)$$

Here $P_k\left(H_1\right)$ and $P_k\left(H_0\right) = 1 - P_k\left(H_1\right)$ are the prior probabilities, $\varepsilon_k = \frac{P_k\left(H_1\right)}{P_k\left(H_0\right)}$, and $\Lambda_k = \frac{p\left(\xi_k \mid H_1\right)}{p\left(\xi_k \mid H_0\right)}$ is the likelihood ratio. The frame indexes are omitted for simplicity.

The estimation so far was based on the assumption of statistically independent consecutive audio frames, which in the case of speech and music is not quite correct. To express this property explicitly, we model the sequence of frame states as a first-order Markov process. The full derivation is presented in [6] and the smoothed likelihood $\hat{\Lambda}_k^{(n)}$ is:

$$\tilde{\Lambda}_k^{(n)} = \frac{a_{01} + a_{11}\tilde{\Lambda}_k^{(n-1)}}{a_{00} + a_{10}\tilde{\Lambda}_k^{(n-1)}}\Lambda_k^{(n)} \qquad \hat{\Lambda}_k^{(n)} = \frac{P_k^{(n)}\left(H_0\right)}{P_k^{(n)}\left(H_1\right)}.\tilde{\Lambda}_k^{(n)} \quad (14)$$

Here a_{01} and a_{10} are the prior probabilities for changing the state, $a_{00} = 1 - a_{01}$ and $a_{11} = 1 - a_{10}$ are the prior probabilities to stay in the same state. In general this is a smoothing filter, lower are the priors for change – higher is the smoothing. After substituting $\hat{\Lambda}_k^{(n)}$ in (13) the prior probabilities cancel nicely and for the probability for double talk in given bin and frame we have:

$$P_k^{(n)}\left(H_1 \mid \xi_k^{(n)}\right) = \frac{\tilde{\Lambda}_k^{(n)}}{1+\tilde{\Lambda}_k^{(n)}}. \quad (15)$$

There are two main ways to combine the likelihoods from all frequency bins to estimate the likelihood for the entire frame. The first is the geometric mean, also known as the log-likelihood ratio test:

$$\Lambda_{GM}^{(n)} = \exp\left(\frac{1}{K}\sum_k \log \tilde{\Lambda}_k^{(n)}\right). \quad (16)$$

The second is the arithmetic mean:

$$\Lambda_{AM}^{(n)} = \frac{1}{K}\sum_k \tilde{\Lambda}_k^{(n)}. \quad (17)$$

The first expects that all frequency bins should have double talk to trigger double talk for the entire frame; the second can have high likelihood even if just a few frequency bins have double talk. The reality is somewhere in between: the speech signal is quite sparse, so (16) will not work well; on the other hand (17) is less robust to noise. We compute the likelihood ratio for the entire frame as a weighted sum of the geometric and arithmetic means:

$$\Lambda^{(n)} = \beta\Lambda_{GM}^{(n)} + \left(1-\beta\right)\Lambda_{AM}^{(n)}, \quad (18)$$

hoping that with a properly selected value of the coefficient β we can combine the advantages of both approaches.

Using the same methodology as above we derive the smoothing filter and the probability for double talk for the entire frame:

$$\tilde{\Lambda}^{(n)} = \frac{b_{01} + b_{11}\tilde{\Lambda}^{(n-1)}}{b_{00} + b_{10}\tilde{\Lambda}^{(n-1)}}\Lambda^{(n)} \qquad P^{(n)}\left(H_1 \middle| \boldsymbol{\xi}^{(n)}\right) = \frac{\tilde{\Lambda}^{(n)}}{1+\tilde{\Lambda}^{(n)}} \quad (19)$$

Here b_{01} and b_{10} are the prior probabilities for changing the frame state, $b_{00} = 1 - b_{01}$ and $b_{11} = 1 - b_{10}$ are the prior probabilities to stay in the same frame state. The soft decision in (19) can be converted to a binary decision by comparing with a threshold according to equation (9).

Once we have estimated the probabilities for double talk for each frequency bin and for the entire frame we can update the estimates for the means and variances:

$$\begin{aligned}
\nu_k^{(n)}\left(k\right) &= \frac{T}{\tau_N}\left(1 - P^{(n)}\left(H_1 \middle| \boldsymbol{\xi}^{(n)}\right)P_k^{(n)}\left(H_1 \mid \xi_k^{(n)}\right)\right) \\
\bar{\xi}_N^{(n)}\left(k\right) &= \left(1-\nu_k^{(n)}\right)\bar{\xi}_N^{(n-1)}\left(k\right) + \nu_k^{(n)}\xi_k^{(n)} \qquad (20) \\
\lambda_N^{(n)}\left(k\right) &= \left(1-\nu_k^{(n)}\right)\lambda_N^{(n-1)}\left(k\right) + \nu_k^{(n)}\left(\xi_k^{(n)} - \bar{\xi}_N^{(n)}\left(k\right)\right)^2
\end{aligned}$$

$$\begin{aligned}
\upsilon_k^{(n)} &= \frac{T}{\tau_D}P^{(n)}\left(H_1 \middle| \boldsymbol{\xi}^{(n)}\right)P_k^{(n)}\left(H_1 \mid \xi_k^{(n)}\right) \\
\bar{\xi}_D^{(n)}\left(k\right) &= \left(1-\upsilon_k^{(n)}\right)\bar{\xi}_D^{(n-1)}\left(k\right) + \upsilon_k^{(n)}\xi_k^{(n)} \qquad (21) \\
\lambda_D^{(n)}\left(k\right) &= \left(1-\upsilon_k^{(n)}\right)\lambda_D^{(n-1)}\left(k\right) + \upsilon_k^{(n)}\left(\xi_k^{(n)} - \bar{\xi}_D^{(n)}\left(k\right)\right)^2
\end{aligned}$$

Here T is the audio frame duration, τ_N and τ_D are the adaptation time constants. The adaptation speed also depends on the double talk probabilities.

4. EXPERIMENTAL RESULTS

The proposed algorithm was evaluated and compared with 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 (54 and 60 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 microphone formed a triangle with sides of one meter each. All combinations of the noise, near, and far end signal levels produced eight recordings. Training and testing sets with all of the combinations were recorded separately. 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, i.e. double talk. The ground truth was established by running the clean near and far end speech signals trough a precise voice activity detector [7]. 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 the evaluation parameter, defined as:

$$\varepsilon = \frac{N_{FP} + N_{FN}}{N_{Tot}}.100\%. \quad (22)$$

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 frames.

Table 1. Optimal values of the DTD parameters

Parameter	Value	Unit
f_{beg}	853.33	Hz
f_{end}	6090.00	Hz
a_{01}	0.0000123	
a_{10}	0.0000433	
β	0.285	
b_{01}	0.0000010	
b_{10}	0.0000035	
η	0.95	
τ_N	4.33	sec
τ_D	10.00	sec

Table 2. Results, baseline and proposed algorithms

Algorithm	Error	N_{FP}	N_{FN}	T_{tot}
Baseline	2.94%	575	511	36920
Proposed	1.26%	251	214	36920

The sampling rate was 16 kHz, we used 512 samples per audio frame, and the overlap and add process was as described in [8] with 50% overlapping and Hann weight window. As the duration of the audio frame was 16 ms we used a ten taps filters for each frequency bin, i.e. L=10.

With the training set an optimization was conducted to minimize the error rate by varying the values of the DTD parameters, using the same approach as described in [7]. The vector of the optimization parameters is:

$$\mathbf{V} = \left[f_{beg}, f_{end}, a_{01}, a_{10}, \beta, b_{01}, b_{10}, \eta, \tau_N, \tau_D \right]. \quad (23)$$

Here f_{beg} and f_{end} are the beginning and ending frequencies to process in equations (8), (16), and (17) rounded to the closest frequency bins K_{beg} and K_{end}. As optimization criterion we selected to minimize the classification error after the binary decision, which is a function of the optimization parameters. Then:

$$\mathbf{V}_{OPT} = \arg\min_{\varepsilon} \left(\varepsilon(\mathbf{V}) \right). \quad (24)$$

The optimal values of these parameters are shown in Table 1. Note the relatively high value of f_{beg} – the optimization program lifted it because of the high energy of the automotive noise in the lower part of the frequency band. For the baseline algorithm the optimal threshold value was determined to be $\eta = 0.960$ after a similar optimization.

All further results were obtained against the testing set of recordings, which the optimization procedure hasn't used. The results for the baseline and proposed algorithms are shown in Table 2, the ROC curves – in Figure 2.

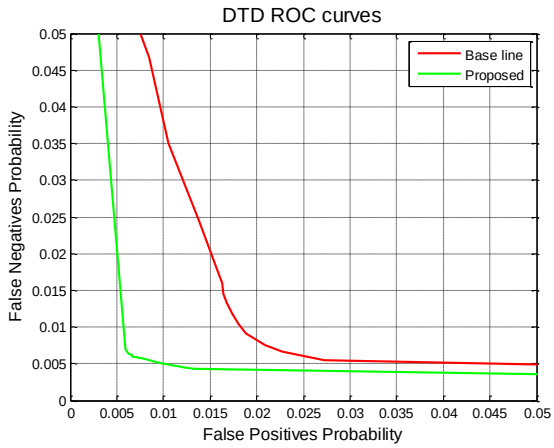

Figure 2. ROC curves

5. CONCLUSIONS AND FUTURE WORK

The proposed algorithm has a lower error rate and is substantially better in reducing the false negatives, preventing AEC from diverging during the double talk situations.

The proposed DTD algorithm provides estimation of DTD probability for each frequency bin separately. This makes possible controlling the AEC adaptation speed for each frequency bin separately based on DTD probability. This will keep the adaptation on for the frequency bins without double talk and improve the AEC parameters considering the sparse nature of the speech signal.

6. 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] J. Sohn, N. Kim, W. Sung, "A statistical model based voice activity detector". IEEE Signal Processing Letters, vol. 6, No. 1, pp. 1-3, January 1999.

[7] 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, 20 April 2010.

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