NORMALIZED DOUBLE-TALK DETECTION BASED ON MICROPHONE AND AEC ERROR CROSS-CORRELATION

Mohammad Asif Iqbal[1] *Jack W. Stokes*[2] *Steven L. Grant*[1]

[1]University of Missouri-Rolla, Rolla, MO 65409, {ammq2,sgrant}@umr.edu
[2]Microsoft Research, Redmond, WA 98052, jstokes@microsoft.com

ABSTRACT

In this paper, we present two different double-talk detection schemes for Acoustic Echo Cancellation (AEC). First, we present a novel normalized detection statistic based on the cross-correlation coefficient between the microphone signal and the cancellation error. The decision statistic is designed in such a way that it meets the needs of an optimal double-talk detector. We also show that the proposed detection statistic converges to the recently proposed normalized cross-correlation based double-talk detector [1], the best known cross-correlation based detector. Next, we present a new hybrid double-talk detection scheme based on a cross-correlation coefficient and two signal detectors. The hybrid algorithm not only detects double-talk but also detects and tracks any echo-path variations efficiently. We compare our results with other cross-correlation based double-talk detectors to show their effectiveness.

1. INTRODUCTION

Most teleconferencing conversations are conducted in the presence of acoustic echoes [2]; if the delay between the speech and its echo is more than a few tens of milliseconds, the echo is distinctly noticeable. An acoustic echo canceller (AEC) is used to remove the echo created due to the loudspeaker-microphone environment [3]. Echo cancellation is achieved by adaptively synthesizing a replica of the echo and subtracting the result from the echo-corrupted signal [2]. When the near-end talker is active or when the speech comes from both the far-end and near-end, the filter coefficients will diverge from the true echo path impulse response if adaptation is enabled. A double-talk detector is used to stop the AEC's filter adaptation during periods of near-end speech [3].

Double-talk detection plays a very important part in acoustic echo cancellation. A double-talk detection algorithm should be able to detect a double-talk condition quickly and accurately so as to freeze adaptation as soon as possible; at the same time it should be able to track any echo-path changes and should be able to distinguish double-talk from the echo-path variations [4]. To solve this problem, this paper presents two different techniques for double-talk detection. An optimum decision variable ξ for double-talk detection should behave as follows [3]:

1. If double-talk is not present i.e. $v = 0$, then $\xi \geq T$.
2. If double-talk is present i.e. $v \neq 0$, then $\xi < T$. The threshold T must be a constant independent of the data and the decision statistic ξ must be insensitive to echo-path variations when $v = 0$.

Figure 1 shows the basic structure of the adaptive acoustic echo canceller. The far-end signal x is filtered through the room impulse response $\mathbf{h}$ to get the echo signal

$$y(n) = \mathbf{h^T x} \tag{1}$$

where

$$\begin{aligned} \mathbf{h} &= [h_0 \;\; h_1 \;\; .... \;\; , \;\; h_{L-1}]^T, \\ \mathbf{x} &= [x(n) \;\; x(n-1) \;\; .... \;\; , \;\; x(n-L+1)]^T, \end{aligned}$$

and L is the length of the echo-path. This echo signal is added to the near-end speech signal v to get the microphone signal

$$m(n) = y(n) + v(n). \tag{2}$$

The error signal at time n is defined as

$$e(n) = m(n) - \hat{\mathbf{h}}^\mathbf{T}\mathbf{x} \tag{3}$$

and is used to adapt the L taps of the AEC's adaptive filter $\hat{\mathbf{h}}$.

This paper is structured as follows. In section 2, we review previous double-talk detection algorithms. In section 3, the novel normalized double-talk detection algorithm is formulated and we also show a link between the proposed algorithm and the one proposed in [1]. We propose the new hybrid double-talk detection scheme in section 4. Next, we do a comprehensive study on the proposed algorithms in section 5 which is followed by a summary and conclusions in section 6.

2. PREVIOUS WORK

Referring to Figure 1, Ye and Wu [4] first proposed using the cross-correlation vector between the far-end signal vector $\mathbf{x}$, which is played out of the speakers, and the AEC's cancellation error e, $r_{e\mathbf{x}} = E[e\mathbf{x}^T]$, as the basis for double-talk detection. In this paper, we will refer to this algorithm as XECC. Simulation results by Benesty [1] have shown that this approach does not work well for detecting double-talk, and a theoretical derivation provides further insight. Noting that the near-end speech v is independent of the far-end signal $\mathbf{x}$ and assuming all of the signals are zero mean, the cross-correlation between the AEC's error signal and the speaker signal is

$$\begin{aligned} r_{e\mathbf{x}} &= E[(y + v - \hat{\mathbf{h}}^\mathbf{T}\mathbf{x})\mathbf{x}^\mathbf{T}] \\ &= E[(\mathbf{h}^\mathbf{T}\mathbf{x} - \hat{\mathbf{h}}^\mathbf{T}\mathbf{x})\mathbf{x}^\mathbf{T}] \\ &= (\mathbf{h}^\mathbf{T} - \hat{\mathbf{h}}^\mathbf{T})R_{\mathbf{xx}} \end{aligned} \tag{4}$$

where $E[\bullet]$ denotes the mathematical expectation and $R_{\mathbf{xx}} = E[\mathbf{x}\mathbf{x}^\mathbf{T}]$. Clearly from equation 4 we observe $r_{e\mathbf{x}}$ is high only when there is a change in the echo-path; hence this approach is

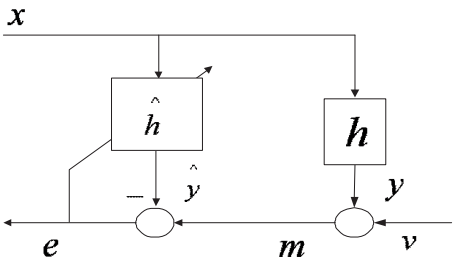

Fig. 1. Basic AEC Model

more suitable for tracking echo-path variations rather than detecting double-talk.

More recently, Benesty, et. al. [1] [5] proposed a double-talk detection algorithm based on the cross-correlation between the far-end signal vector $\mathbf{x}$ and the microphone signal scalar m, $r_{\mathbf{x}m} = E[\mathbf{x}m]$, which we refer to as XMCC in this paper. Benesty's decision statistic used to detect double-talk in [1] is given by

$$\xi_{XMCC} = \sqrt{r_{\mathbf{x}m}^T(\sigma_m^2 R_{\mathbf{xx}})^{-1} r_{\mathbf{x}m}} \tag{5}$$

where $R_{\mathbf{xx}}$ is defined earlier and the variance of the microphone signal (σ_m^2) is

$$\begin{aligned} \sigma_m^2 &= E[mm^T] = E[(y+v)(y+v)^T] \\ &= E[yy^T] + E[vv^T] = E[\mathbf{h^T x}(\mathbf{h^T x})^T] + \sigma_v^2 \\ &= \mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2 \end{aligned} \tag{6}$$

and σ_v^2 is the near-end speech power.

3. NORMALIZED DOUBLE-TALK DETECTION BASED ON THE MICROPHONE SIGNAL AND THE AEC ERROR CROSS-CORRELATION

Instead of using $r_{e\mathbf{x}}$ or $r_{\mathbf{x}m}$ as discussed in section 2, we propose using the cross-correlation between the cancellation error e and the microphone signal m, $r_{em} = E[em]$, as the basis for double-talk detection. This algorithm will be called MECC in this paper.

$$\begin{aligned} r_{em} &= E[(y + v - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x})(y+v)^T] \\ &= E[(\mathbf{h^T x} - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x} + v)(\mathbf{h^T x} + v)^T] \\ &= E[(\mathbf{h^T x} - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x})\mathbf{x^T h} + vv^T] \\ &= (\mathbf{h^T} - \hat{\mathbf{h}}^{\mathbf{T}}) R_{\mathbf{x},\mathbf{x}} \mathbf{h} + \sigma_v^2. \end{aligned} \tag{7}$$

We define our new normalized decision statistic to be:

$$\xi_{MECC} = 1 - \frac{r_{em}}{\sigma_m^2}. \tag{8}$$

Substituting equations 6 and 7 in 8 we get:

$$\begin{aligned} \xi_{MECC} &= 1 - \frac{(\mathbf{h^T} - \hat{\mathbf{h}}^{\mathbf{T}}) R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2} \\ &= \frac{\hat{\mathbf{h}}^{\mathbf{T}} R_{\mathbf{xx}} \mathbf{h}}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2}. \end{aligned} \tag{9}$$

We observe from equation 9, that for $v = 0$, $\xi_{MECC} \approx 1$ and for $v \neq 0$, $\xi_{MECC} < 1$. Thus, the proposed detection statistic meets the needs of an optimal double-talk detector.

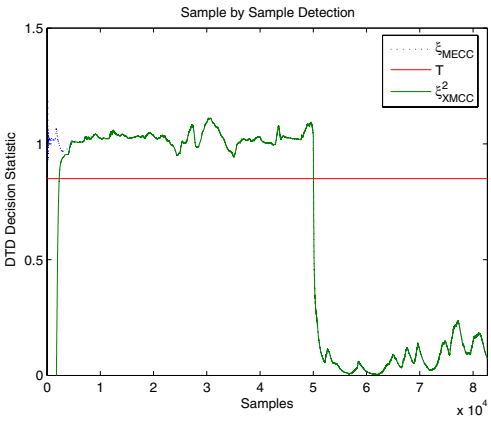

Fig. 2. Illustrating the convergence of the proposed MECC and the XMCC double-talk detectors.

The values for r_{em} and σ_m^2 in (8) are exact and not available in practice. As a result, the final decision statistic is given by:

$$\xi_{MECC} = 1 - \frac{\hat{r}_{em}}{\hat{\sigma}_m^2} \tag{10}$$

which is based on the estimates $\hat{r}_{em}[n]$ and $\hat{\sigma}_m^2[n]$. The estimates are found using the exponential recursive weighting algorithm, [6] [7]:

$$\begin{aligned} \hat{r}_{em}[n] &= \lambda \hat{r}_{em}[n-1] + (1-\lambda) e[n] m^T[n] \\ \hat{\sigma}_m^2[n] &= \lambda \hat{\sigma}_m^2[n-1] + (1-\lambda) m[n] m^T[n] \end{aligned}$$

where $e[n]$ is the captured cancellation error sample at time n, $m[n]$ is the captured microphone signal sample at time n, and λ is the exponential weighting factor. If

$$\xi_{MECC} < T \tag{11}$$

we conclude that the captured sample of the microphone signal is corrupted by the near-end speech and halt adaptation of the AEC's adaptive filter(s). Otherwise, we continue adapting.

In addition to its simplicity, the main advantage of the proposed detection statistic is that only the maximum cross-correlation needs to be computed instead of computing the entire cross-correlation vector required by the other algorithms. This results in significant computational savings as compared to the other algorithms; we only require 2 multiplications, 2 additions, 1 subtraction and a division to compute the decision statistic at each sample (i.e. 6 operations per sample), whereas for the Benesty's test statistic $3L + 3$ operations are required to compute the detection statistic at each sample where L is the frame size (typically $L \geq 512$).

3.1. Relationship Between New and Benesty's Test Statistic

The proposed decision statistic is given by (10), which theoretically can be rewritten as in (9), and Benesty's double-talk decision statistic is given in (5). The decision statistics are different as the former is based on r_{em}, and the latter is based on $r_{\mathbf{x}m}$. Although the decision statistics are different, they can be shown to result in a similar expression. Substituting $r_{\mathbf{x}m} = R_{\mathbf{xx}}\mathbf{h}$ and $\sigma_m^2 = \mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2$ in (5), we get

$$\begin{aligned} \xi_{XMCC}^2 &= \mathbf{h^T} R_{\mathbf{xx}} (\sigma_m^2 R_{\mathbf{xx}})^{-1} R_{\mathbf{xx}} h \\ &= \frac{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h}}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2} \end{aligned} \tag{12}$$

and from (9) we have

$$\xi_{MECC} = \frac{\hat{\mathbf{h}}^{\mathbf{T}} R_{\mathbf{xx}} \mathbf{h}}{\mathbf{h}^{\mathbf{T}} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2}. \quad (13)$$

In addition to the square root, the other difference between the decision statistics is in the numerator; we have the taps of the AEC filter $\hat{\mathbf{h}}^{\mathbf{T}}$ in ξ_{MECC} and the true echo-path impulse response $\mathbf{h}^{\mathbf{T}}$ in ξ_{XMCC}. However for practical implementation and computational simplicity, the authors in [1] substitute $\hat{\mathbf{h}}^{\mathbf{T}}$ for $\mathbf{h}^{\mathbf{T}}$ resulting in similar decision statistics. Simulations in Figure 2 shows that the proposed decision statistic has similar performance compared to Benesty's test statistic. However, our algorithm is significantly simpler and computationally efficient.

4. HYBRID DOUBLE-TALK DETECTION

In this section, we introduce a hybrid double-talk detector based on a cross-correlation measure between the microphone and AEC cancellation error similar to the idea presented in section III and the double-talk detection algorithm based on speech detection and discriminator based on real-time recurrent learning (RTRL) presented in [8]. The architecture for the hybrid double-talk detection algorithm is shown in Figure 3

In this algorithm, we use a different cross-correlation measure between the cancellation error e and the microphone signal m given by the estimated cross-correlation function,(ECC):

$$ECC[t] = \frac{P_{m,e}[t]}{P_e[t] P_m[t]}. \quad (14)$$

The ECC is the maxima of the correlation in a frame and is updated using the exponential recursive weighting algorithm [6] [7]

$$P_e^2[t] = \lambda P_e^2[t-1] + (1-\lambda)\mathbf{e}[t]\mathbf{e}^{\mathbf{T}}[t] \quad (15)$$

$$P_m^2[t] = \lambda P_m^2[t-1] + (1-\lambda)\mathbf{m}[t]\mathbf{m}^{T}[t] \quad (16)$$

$$P_{m,e}[t] = P_{m,e}[t-1] + (1-\lambda)\mathbf{e}[t]\mathbf{m}^{T}[t] \quad (17)$$

where $\mathbf{e}[t]$ is the captured cancellation error vector in the time frame t, $\mathbf{m}[t]$ is the captured microphone signal vector at the time frame t and λ is the exponential weighting factor. Alternatively, we could also use the MECC test statistic given in (10). Smaller values of λ provide better tracking capability but worse estimation accuracy. In practice for slowly time varying signals, $0.9 \leq \lambda \leq 1$ is usually chosen [4]. We observe from (14), the cross-correlation is high whenever there is a change in the echo-path and/or when the near-end speech is present. To differentiate the near-end speech from the echo-path variations we use a speech detector and signal discrimator based on real time recurrent learned (RTRL) [8] which is described next.

Frequency domain logistic discriminative speech detectors are used to detect the presence of speech [9]. The class probability is estimated as

$$P_t = \frac{1}{1 + \exp(-\mathbf{W}^{\mathbf{T}} \chi_{\mathbf{t}})} \quad (18)$$

where P_t is the probability of speech at time frame t, $\mathbf{W}^{\mathbf{T}}$ are the trained weights $(1 \times frequencybins)$ and χ_t is a vector of extracted features in each frequency bin at the time frame t. The trained weights $\mathbf{W}^T$ are obtained using Real Time Recurrent Learning [10] and are obtained by off-line training. For a detailed discussion on speech detectors and their training process, see [8].

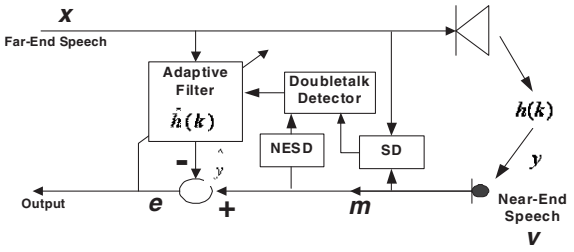

Fig. 3. Hybrid Double-talk Detection Model

We use two detectors at the microphone to detect the presence of the near-end speech as shown in Figure 3. For the microphone signal detector (NESD) we use the logarithm of the estimated posterior SNR as the feature [9]:

$$\chi_{NESD}(k,t) = 10\{\log|M(k,t)|^2 - \log N_{NE}(k,t)\} \quad (19)$$

where N_{NE} is the noise energy in frequency bin k and time-frame t at the near-end. The noise power N can be tracked using [11]. In this paper we use a minima tracker (for each frequency bin we look back a few frames (e.g. 25) and choose the lowest value of the signal) followed by smoothing, to track the noise floor [11]. This NESD detector gives the presence of speech at the microphone, which can be due to near-end speech or the far-end echo.

To differentiate the near-end speech from the far-end echo we use a special detector/discriminator SD which requires features that differentiate the near-end speech from the far-end echo. Thus we use the logarithm of the ratio of the instantaneous power of the microphone signal M to the instantaneous power of the far-end signal X as the feature, i.e.

$$\chi_{SD}(k,t) = 10\{\log|M(k,t)|^2 - \log|X(k,t)|^2\}. \quad (20)$$

It was observed in [8] that the extracted features are distinct for different scenarios. The extracted features are typically largest for only the near-end speech, smallest for the echo-only case, and in between for the case of double-talk. Different feature levels correspond to different probability levels; larger features correspond to higher probabilities. For the echo-only case, the extracted features are always low independent of the echo-path; hence the special detector/disriminator is independent of the echo-path in the absence of near-end speech.

We confirm the presence of the near-end speech when both the detectors indicate the presence of speech. A speech detection based double-talk detector [8] when used alone for double-talk detection does not give superior performance. However, the performance can be improved by combining it with the proposed cross-correlation measure. The hybrid double-talk detector works as follows:

1. When both the detectors indicate a high probability of the presence of speech (i.e. $P_{NESD}(t) \geq P_{Threshold_1}$ and $P_{SD}(t) \geq P_{Threshold_2}$) and the estimated cross-correlation $ECC(t) \geq R_{th}$ then the captured frame of the microphone signal is corrupted by the near-end speech.
2. When $P_{NESD}(t) \geq P_{Threshold_1}$, $P_{SD}(t) < P_{Threshold_2}$ and $ECC(t) \geq R_{th}$ then the signal at the microphone is echo only due to echo-path change.
3. When $P_{NESD}(t) \geq P_{Threshold_1}$, $P_{SD}(t) < P_{Threshold_2}$ and $ECC(t) < R_{th}$ then the signal at the microphone is echo only without echo-path change.

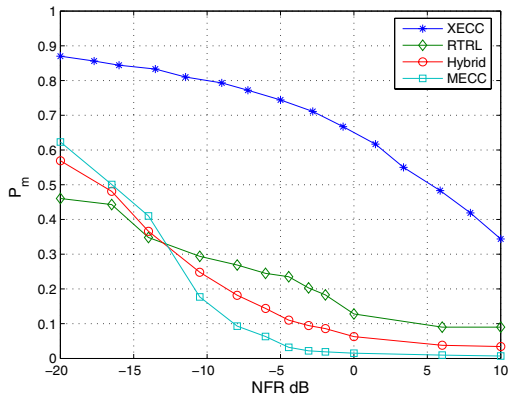

Fig. 4. P_m as function of NFR for the proposed MECC and the CC-SD double-talk detectors and XECC double-talk detectors at $P_f = 0.1$.

In the first condition, we halt adaptation of the adaptive filter coefficients, but continue adapting in the last two conditions. The results in Figure 4 use the ECC, but using the MECC test statistic (8) in the hybrid double-talk detector may perform equal to or slightly better than using the ECC test statistic.

5. EXPERIMENTS AND RESULTS

We now introduce simulation results for the proposed double-talk detectors. The performance is characterized in terms of the probability of miss (P_m) as a function of near-end to far-end speech ratio (NFR) under a probability of false alarm (P_f) constraint [5]. The probability of miss (P_m) is the probability of not detecting (miss) double-talk when it is present; therefore a smaller value of P_m indicates better performance. To evaluate the proposed double-talk detectors we follow [5].

The recorded digital speech sampled at 16 KHz is used as far-end speech $\mathbf{x}$ and near-end speech v and a measured $L = 8000$ sample (500 ms) room impulse response of a $10' \times 10' \times 8'$ room is used as the loudspeaker-microphone environment $\mathbf{h}$. We compare our results with the conventional cross-correlation (XECC) based double-talk detector proposed in [4] and the RTRL based double-talk detector proposed in [8]. The P_m characteristics of all the four methods under the constraint of $P_f = 0.1$ are shown in Figure 4. It is clear that the hybrid and the proposed normalized detection statistic (MECC) significantly outperform the conventional (XECC) double-talk detector over a full-range of NFR values. Also it can be observed that the hybrid double-talk detection scheme outperforms the RTRL based double-talk detector for most of the NFR values. Thus, we conclude that the performance of the RTRL based double-talk detector [8] is improved by combining it with the proposed cross-correlation measure. At low values of NFR, the RTRL and hybrid double-talk detector perform better than the MECC algorithm based on the optimal test statistic. Most likely, the increased performance is due to improved speech detection capabilities of RTRL in the presence of noise.

It should be noted that the performance of the proposed normalized decision statistic (MECC) is exactly similar to the Benesty's test statistic (XMCC), the best known cross-correlation based double-talk detector. However, our detection statistic is computationally very efficient, the detection threshold T is independent of the data and is insensitive to echo-path variations.

6. CONCLUSIONS

We have proposed two different techniques for double-talk detection. First, we introduced the novel normalized decision statistic; the proposed detection statistics meets the needs of an optimal double-talk detector, is computationally very efficient and converges to the best known cross-correlation based double-talk detector. Next, we formulated the hybrid double-talk detection scheme. The hybrid double-talk detector works on a frame by frame basis; the algorithm not only detects double-talk but also detects and tracks any echo-path variations. This is achieved at the cost of increased computational complexity.

7. REFERENCES

[1] Jacob Benesty, Dennis R. Morgan, and Juan H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 168–172, March 2000.

[2] Simon Haykin, *Adaptive Filter Theory*, Prentice Hall, Englewood Cliffs, NJ, 1991.

[3] J.Benesty, T. Gansler, D.R. Morgan, M.M. Sondhi, and S.L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer, Inc., New York, 2001.

[4] Hua Ye and Bo-Xiu Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Transactions on Communications*, vol. 39, pp. 1542–1545, November 1991.

[5] Juan H. Cho, Dennis R. Morgan, and Jacob Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 718–724, November 1999.

[6] B. Porat, "Second-order equvalence of rectangular and exponential windows in least-squares estimation of autoregressive processes," *IEEE Transactions on Acoustics,Speech and Signal Processing*, 1985.

[7] Y. Hua, *Adaptive Filter Theory and applications*, Ph.D. thesis, South-East University, Tiangsu, China, March 1989.

[8] Mohammad Asif, Jack W. Stokes, John. C. Platt, Arun Surendran, and Steven L. Grant, "Doubletalk detection using real time recurrent learning," in *International Workshop on Acoustic Echoes and Noise Control*, Paris, France, September 2006.

[9] Arun C. Surendran, Somsak Sukittanon, and John Platt, "Logistic discriminative speech detectors using posterior snr," in *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, pp. 625–628.

[10] Ronald J. Williams and David Zipser, "Experimental analysis of real-time recurrent learning algorithm," in *Connection Science,Vol 1, No 1*, 1989, pp. 87–111.

[11] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of the 7th European Signal Processing Conference*, Edinburgh, Scotland, September 1994, pp. 1182–1185.