

NOISE-ROBUST DOUBLE-TALK DETECTION BASED ON NORMALIZED CROSS CORRELATION AND A NOISE OFFSET

Akihiko Sugiyama, Jérôme Berclaz[†], Miki Sato

Media and Information Research Labs., NEC Corporation
Kawasaki 211-8666, JAPAN

[†]School of Comp. and Commun. Sciences, Ecole Polytechnique Fédérale de Lausanne
CH-1015 Lausanne, Switzerland

ABSTRACT

This paper presents a noise-robust double-talk detection algorithm based on normalized cross-correlation and a noise offset. The noise offset alleviates undesirable influence by the background noise existing in the microphone signal. It is estimated from the echo-cancelled signal when its autocorrelation is low and its power is smaller than the echo-replica power. A detection threshold of the new normalized cross-correlation is adaptively controlled based on the echo-to-NES (near-end speech) ratio. Simulation results demonstrate superior performance of the new algorithm.

1. INTRODUCTION

An acoustic echo canceler [1, 2] electrically models the path from the loudspeaker to the microphone by an adaptive filter. By exciting the adaptive filter with the signal from the remote side, or far-end, an echo replica is generated. The echo replica is subtracted from the microphone signal, resulting in an echo-cancelled speech for transmission.

Double-talk, when the far-end and the near-end speech (NES) simultaneously exist, sometimes causes fatal degradation in echo cancellation because of the interference by the near-end speech [2]. Therefore, coefficient adaptation should be disabled during double-talk periods. Double-talk detection plays a key role in the overall performance and is carried out based on various measures [3]-[6].

The most basic algorithm for double-talk detection is the one originally developed by Geigel [3]. When the maximum of past samples multiplied by 0.5 is greater than the current microphone signal, double-talk is declared. The factor of 0.5 is based on the fact that a loss of 6 dB is typical in the case of network echo cancellation. Although it is simple with reasonable performance, it cannot be directly applied to acoustic echo cancellation where characteristics of the echo path is unknown in advance.

A more sophisticated algorithm by Ye et al. [4] utilizes cross-correlation of the reference signal and the error. Only when the correlation is non-zero, coefficient adaptation is carried out based on the fact that such a cross-correlation becomes zero after convergence. However, because the acoustic echo path keeps more or less changing, cross-correlation is likely to take a non-zero value and adaptation is performed even when the NES is present. Therefore, it is not suitable for acoustic echo cancellation.

Gänsler et al. proposed an algorithm based on coherence between the reference and the microphone signal [5]. However, its drawback is insufficient normalization of the coherence. As a result, a threshold heavily depend on the signal statistics and the echo-path characteristics [6].

To perform double-talk detection with a fixed threshold, Benesty et al. proposed a technique based on normalized cross-correlation of the echo and the microphone signal [6]. Although it exhibits good performance in most cases, it still has insufficient detection capability in the presence of noise, as the already mentioned algorithms. Acoustic echo cancelers are more exposed to open environment than network echo cancelers [7]. Therefore, it is important to pay attention to noise contaminating the echo and the NES.

This paper presents a noise-robust double-talk detection algorithm based on normalized cross-correlation and a noise offset. The noise offset, that is estimated from the echo-cancelled signal, alleviates undesirable influence by the background noise existing in the microphone signal. A detection threshold of the new normalized cross-correlation is adaptively controlled based on the echo-to-NES ratio (ENR). The next section reviews double-talk detection based on normalized cross-correlation. Section 3 develops a new noise-robust double-talk detection. Finally, in Section 4, simulation results are demonstrated to support improved performance.

2. DOUBLE-TALK DETECTION BASED ON NORMALIZED CROSS-CORRELATION [6]

This section reviews the normalized cross-correlation algorithm as the basis for the new algorithm. Figure 1 depicts a blockdiagram for an acoustic echo canceler with a double-talk detector.

The input signal vector, $\mathbf{x}(k)$, the $N \times 1$ echo-path impulse response vector, \mathbf{h} , and a time invariant $N \times 1$ coefficient vector, $\mathbf{w}(k)$, of the adaptive filter, are defined by the following equations. k is the time index for discrete signals.

$$\mathbf{x}(k) = [x(k) \ x(k-1) \ \cdots \ x(k-N+1)]^T \quad (1)$$

$$\mathbf{h} = [h_0 \ h_1 \ \cdots \ h_{N-1}]^T \quad (2)$$

$$\mathbf{w}(k) = [w_0(k) \ w_1(k) \ \cdots \ w_{N-1}(k)]^T \quad (3)$$

The echo, $y(k)$, and an echo replica, $\hat{y}(k)$, are given by

$$y(k) = \mathbf{h}^T \mathbf{x}(k), \quad (4)$$

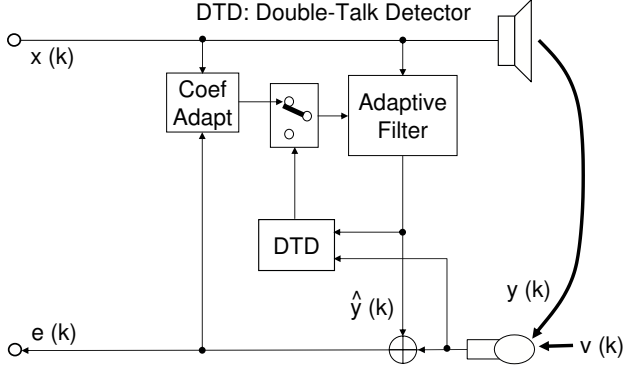


Fig. 1. Echo Canceled with Double-Talk Detection.

$$\hat{y}(k) = \mathbf{w}(k)^T \mathbf{x}(k). \quad (5)$$

The microphone signal, $m(k)$, is expressed as a sum of the echo $y(k)$ and the NES $v(k)$ as

$$m(k) = y(k) + v(k). \quad (6)$$

The output, $e(k)$, is obtained by subtracting the echo replica from the microphone signal as

$$e(k) = [\mathbf{h} - \mathbf{w}(k)]^T \mathbf{x}(k) + v(k). \quad (7)$$

When there is no NES ($v(k) = 0$), a mathematical expectation of the microphone signal power $m(k)^2$ becomes

$$E[m^2(k)] = \mathbf{h}^T \mathbf{R}_{xx} \mathbf{h} = \sigma_m^2, \quad (8)$$

where $\mathbf{R}_{xx} = E[\mathbf{x}(k)\mathbf{x}^T(k)]$. With a definition of

$$\mathbf{r}_{xm} = \mathbf{R}_{xx} \mathbf{h}, \quad (9)$$

(8) can be rewritten as

$$E[m^2(k)] = \mathbf{r}_{xm}^T \mathbf{R}_{xx}^{-1} \mathbf{r}_{xm}. \quad (10)$$

In reality, the microphone signal includes some NES ($v(k) \neq 0$) and (10) should be modified as follows.

$$E[m^2(k)] = \mathbf{r}_{xm}^T \mathbf{R}_{xx}^{-1} \mathbf{r}_{xm} + \sigma_v^2, \quad (11)$$

where σ_v^2 represents the NES power, *i.e.* $E[v(k)^2]$. The decision variable, $\xi_1(k)$, for the normalized cross-correlation is obtained by normalizing (10) by (11) and taking its square root.

$$\xi_1(k) = \sqrt{\mathbf{r}_{xm}^T (\sigma_m^2 \mathbf{R}_{xx})^{-1} \mathbf{r}_{xm}}. \quad (12)$$

For smaller computational load, it is often useful to approximate \mathbf{h} by $\mathbf{w}(k)$ to obtain

$$\mathbf{R}_{xx}^{-1} \mathbf{r}_{xm} = \mathbf{h} \approx \mathbf{w}(k). \quad (13)$$

This relationship holds when the adaptive filter has converged. Substituting (13) in (12) results in

$$\xi_1(k) = \sqrt{\mathbf{r}_{xm}^T \sigma_m^{-2} \mathbf{w}(k)}. \quad (14)$$

With (9) and $\mathbf{h} \approx \mathbf{w}(k)$, (14) can be further simplified as

$$\begin{aligned} \xi_1(k) &= \sqrt{\frac{\mathbf{w}^T \mathbf{R}_{xx}^T \mathbf{w}(k)}{\sigma_m^2}} \\ &= \sqrt{\frac{E[\hat{y}^2(k)]}{\sigma_m^2}} \\ &= \sqrt{\frac{\sigma_y^2}{\sigma_m^2}}. \end{aligned} \quad (15)$$

Assuming that the echo and the NES are not correlated, from (6) and (8), the following expression is obtained,

$$\begin{aligned} \xi_1(k) &= \sqrt{\frac{\sigma_y^2}{\sigma_v^2 + \sigma_y^2}} \\ &= \sqrt{\frac{1}{\sigma_v^2/\sigma_y^2 + 1}}. \end{aligned} \quad (16)$$

where it was naturally assumed that $\hat{y}(k) = y(k)$ after convergence of the adaptive filter.

When there is no NES, *i.e.* single talk, $\sigma_v^2 = 0$. Then, $\xi(k) = 1$ is obtained. On the other hand, if $v(k) \neq 0$, $\xi(k) = \alpha < 1$. Therefore, $\xi(k) = 1$ means single talk and otherwise, represents double-talk.

In the discussions so far, no attention has been paid to noise in the microphone signal. Defining the power of noise $n(k)$ as σ_n^2 , $\xi_1(k)$ in (16) is modified as

$$\begin{aligned} \xi_1(k) &= \sqrt{\frac{\sigma_y^2}{\sigma_v^2 + \sigma_y^2 + \sigma_n^2}} \\ &= \sqrt{\frac{1}{\sigma_v^2/\sigma_y^2 + 1 + \sigma_n^2/\sigma_y^2}}. \end{aligned} \quad (17)$$

With the noise term, it is clear that $\xi_1(k)$ is no longer equal to 1 even for a single-talk period ($v(k) = 0$). $\xi(k) = \beta < 1$. Therefore, coefficient adaptation may not be carried out when it is needed.

3. NEW DOUBLE-TALK DETECTION WITH A NOISE OFFSET AND AN ADAPTIVE THRESHOLD

Figure 1 depicts a blockdiagram of an acoustic echo canceler equipped with the noise-robust double-talk detection. Compared to Fig. 2, it newly contains noise estimation and a double talk detector with a noise offset. The new parts are highlighted by bold lines.

3.1. Measure for Double-Talk Detection with a Noise Offset

The new double-talk detection incorporates a noise offset in the value of $\xi_1(k)$ so that undesirable influence in its original form is alleviated. Let us assume that a good estimate of the noise power, σ_n^2 is available. Then, it is possible to use this estimate to offset the influence of the noise in the microphone signal in calculation of $\xi_1(k)$. For this purpose,

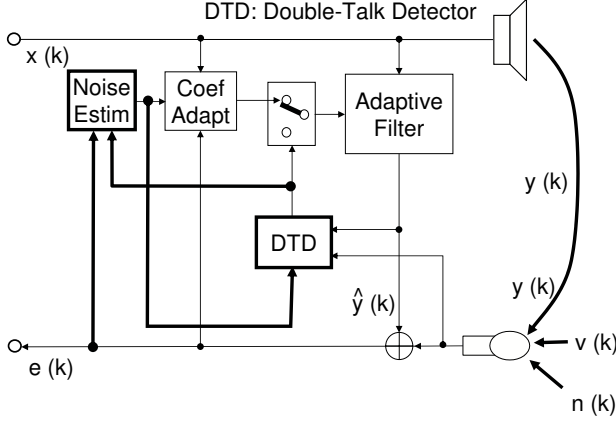


Fig. 2. Acoustic Echo Canceled with Proposed Double-Talk Detection.

(15) is modified to accommodate a noise offset in the numerator as

$$\begin{aligned} \xi_2(k) &= \sqrt{\frac{\sigma_y^2 + \sigma_n^2}{\sigma_m^2}} \\ &= \sqrt{\frac{\sigma_y^2 + \sigma_n^2}{\sigma_v^2 + \sigma_y^2 + \sigma_n^2}} \\ &= \sqrt{\frac{1 + \sigma_n^2/\sigma_y^2}{\sigma_v^2/\sigma_y^2 + 1 + \sigma_n^2/\sigma_y^2}}. \end{aligned} \quad (18)$$

It is clear that (18) takes a value exactly equal to 1 when there is no NES ($v(k) = 0$).

3.2. Noise Estimation with an Adaptive Averaging Constant

A good noise estimate controls the overall performance of the new double-talk detection. It is based on the output power, echo-replica power, and auto-correlation of the output. The noise is estimated when (19) and (20) are simultaneously satisfied.

$$\sigma_y^2 > \sigma_e^2 \quad (19)$$

$$\rho(k)/\rho_0(k) < \gamma \quad (20)$$

$$\rho(k+1) = \sum_{j=0}^J e(k-j)e(k-j-1) \quad (21)$$

$$\rho_0(k+1) = \delta_a \rho_0(k) + (1 - \delta_a) \cdot e(k)e(k-1), \quad (22)$$

where δ_a is a constant for the averaging operation in (22). (19) means that the output power is lower than the echo-replica power [8]. In such a case, it is likely that there is not significant residual echo nor NES. $\rho_0(k)$ is used to normalize $\rho(k)$ so that its value becomes independent of the correlation of the output and close to unity. Then, a threshold γ is set close to 1. $\rho_0(k)$ represents a global level of $\rho(k)$ due to the difference between an integer J

and a constant δ_n in (21) and (22). When there is NES, $\rho(k)/\rho_0(k)$ is likely to take a large value because NES has high autocorrelation. If only noise exists, its autocorrelation should be small.

An estimate of the noise power, $\sigma_n^2(k+1)$, at time $k+1$ is obtained by

$$\sigma_n^2(k+1) = (1 - \delta_n)\sigma_n^2(k) + \delta_n e^2(k). \quad (23)$$

For quick convergence of the noise estimate, a constant δ_n for the averaging operation in (23) is adaptively controlled based on variations of the noise-estimate gradient. The gradient of the noise estimate is evaluated to see if it changes frequently. If the change is frequent, the noise estimate is almost converged. In this case, a smaller value can be used for δ_n for more precise estimation. Otherwise, it is still in the convergence process. The value of δ_n is kept unchanged so that fast convergence is obtained. Such a control of δ_n can solve the trade-off between fast convergence and accuracy in noise estimation. δ_n is controlled by

$$\delta_n(k+1) = \begin{cases} \delta_n(k)/\tau & \text{frequent change} \\ \delta_n(k) & \text{otherwise} \end{cases}, \quad (24)$$

where $\tau > 1$ is a constant. Adaptation of $\delta_n(k)$ is continued until $\delta_n(k)$ reaches the minimum, δ_{min} .

3.3. Adaptive Threshold for $\xi_2(k)$

Due to the noise offset, the value of $\xi_2(k)$ is made equal to unity. However, some imperfections of elements and devices may cause some error in $\xi_2(k)$. It is likely that double-talk detection with a binary decision of 1 or 0 is not always correct. Therefore, it is natural that a continuous or soft decision is preferable with the value of $\xi_2(k)$ as a confidence measure. In this case, it is necessary to set up an appropriate threshold $T(k)$ somewhere between 1 and 0. When $\xi_2(k)$ takes a value below this threshold, a double-talk index, $\theta(k)$, is set to 0 so that coefficient adaptation is completely suppressed.

Considering that the NES and the echo are sufficiently larger in power than noise, (18) reduces to (16). For double-talk,

$$\xi_2(k) = \sqrt{\frac{1}{\sigma_v^2/\sigma_y^2 + 1}} = T(k). \quad (25)$$

$T(k)$ is clearly a function of a ratio of the NES power and the echo-replica power. Therefore, once these signals are known, it is possible to calculate $T(k)$.

The NES and echo replica powers can be estimated by an averaging operation.

$$\hat{\sigma}_v^2(k+1) = (1 - \delta_v)\hat{\sigma}_v^2(k) + \delta_v \cdot e^2(k) \quad (26)$$

$$\hat{\sigma}_y^2(k+1) = (1 - \delta_y)\hat{\sigma}_y^2(k) + \delta_y \cdot \hat{y}^2(k), \quad (27)$$

where $\hat{\sigma}_v^2(k)$ and $\hat{\sigma}_y^2(k)$ are estimated NES power and echo-replica power, respectively. It should be noted that (26) is performed only when confidence of double-talk is high. Values of $\xi_2(k)$ between 1 and $T(k)$ is mapped to a range of 1 and 0 for the double-talk index, $\theta(k)$, so that soft decision is implemented.

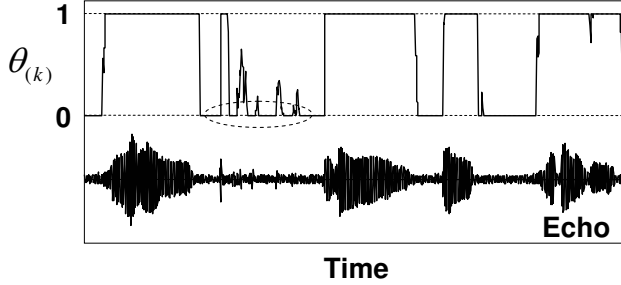


Fig. 3. Incorrect Double-Talk Detection by Noise.

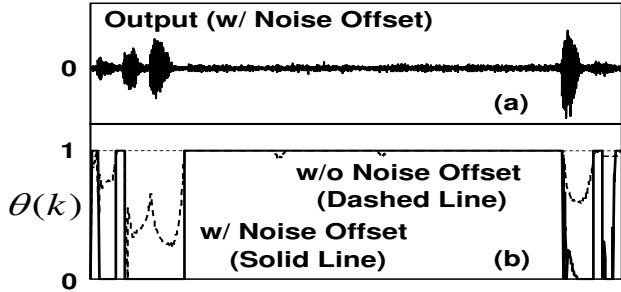


Fig. 4. $\theta(k)$ with and without Noise Offset.

4. EVALUATIONS

Evaluations were carried out using recorded speech sampled at 11.025 kHz. Coefficient adaptation was carried out by an adaptive step-size NLMS algorithm [8]. The number of adaptive filter taps was set to 1754. Parameters were set to $J = 128$, $\gamma = 1$, $\delta_a = 0.99995$, $\delta_n(0) = 0.5$, $\delta_v = \delta_y = 0.001$, and $\tau = 1.01$. Powers expressed by σ are averaged over 512 samples unless otherwise stated.

Figure 3 exhibits incorrect double-talk detection when there is no noise offset. The upper curve is $\theta(k)$ and the lower curve, the echo plus noise. The incorrect detection of double talk is highlighted by a dashed-line oval. Although the noise level is not significantly large in that section, incorrect double-talk detection does occur.

Shown in Fig. 4 (b) are curves of $\theta(k)$ with the noise offset (solid line) and that without the noise offset (dashed line). The corresponding output, $e(k)$, to the former is depicted in (a). Comparing the $\theta(k)$ with and without noise offset, its effect is obvious.

Figures 5 and 6 are devoted to comparison of fixed thresholds and an adaptive threshold. In Figure 5, there is no significant difference among fixed thresholds of $T = 0.4$, 0.7, 0.9, and 0.95 for $\theta(k)$ and an adaptive threshold $T(k)$. However, these specific values of the threshold do not work for the case of Fig. 6. An adaptive threshold achieves good double-talk detection in both cases of Figs. 5 and 6 without any adjustment.

5. CONCLUSION

A noise-robust double-talk detection algorithm based on normalized cross-correlation and a noise offset has been pro-

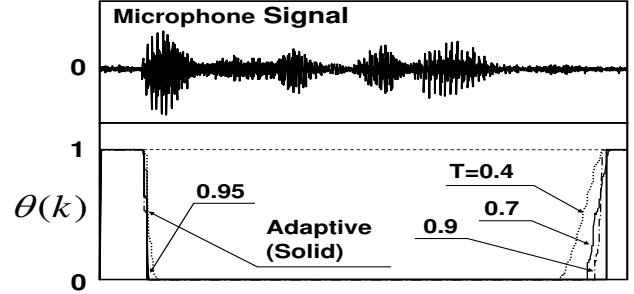


Fig. 5. $\theta(k)$ with Fixed and Adaptive Threshold (Ex. 1).

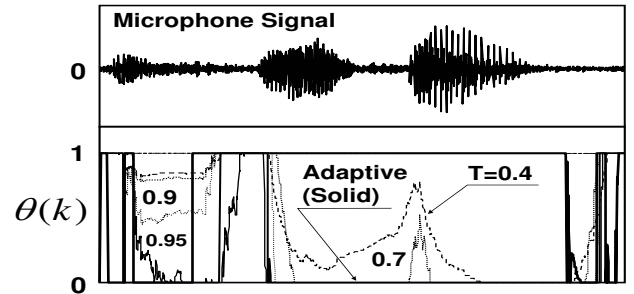


Fig. 6. $\theta(k)$ with Fixed and Adaptive Threshold (Ex. 2).

posed. The noise offset, estimated from the echo-cancelled signal, alleviates undesirable influence by the background noise existing in the microphone signal. A threshold for double-talk detection is adaptively controlled based on an estimate of the echo-to-NES ratio (ENR). Superior detection performance of the new algorithm has been demonstrated in comparison with the conventional algorithm.

6. REFERENCES

- [1] Adaptive Echo cancellation for Speech Signals," Chap. 11, Advances in Speech Signal Processing, Ed. S. Furui and M. M. Sondhi, Marcel-Dekker, 1991.
- [2] C. Breining, P. Dreiseitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Telp, "Acoustic Echo Control," IEEE Signal Processing Mag., pp. 42-69, Jul. 1999.
- [3] D. L. Duttweiler, "A Twelve-channel Digital Echo Canceler," IEEE Trans. Commun. Vol.26, No.5, pp.647-653, May 1978.
- [4] H. Ye and B.-X. Wu, "A New Double-Talk Detection Algorithm Based on the Orthogonality Theorem," IEEE Trans. Commun. Vol.39, No.11, pp.1542-1545, Nov. 1991.
- [5] T. Gansler, M. Hansson, C. J. Ivarsson, and G. Salomonsson, "A Double-Talk Detector Based on Coherence," IEEE Trans. Commun. Vol.44, No.11, pp.1421-1427, Nov. 1996.
- [6] J. Benesty, D. R. Morgan, and J. H. Cho, "A New Class of Doubletalk Detectors Based on Cross-Correlation," IEEE Trans. SAP., Vol.8, No.2, pp.168-172, Mar. 2000.
- [7] A. Sugiyama, A. Hirano, and K. Nakayama, "Acoustic Echo Cancellation for Conference Systems," Proc. EUSIPCO2004, pp.17-20, Sep. 2004.
- [8] A. Hirano and A. Sugiyama, "A Noise-Robust Stochastic Gradient Algorithm with an Adaptive Step-Size for Mobile Hands-Free Telephones," Proc. ICASSP'95, pp.1392-395, May 1995.