A MULTI-CHANNEL ACOUSTIC ECHO CANCELER DOUBLE-TALK DETECTOR BASED ON A NORMALIZED CROSS-CORRELATION MATRIX

Jacob Benesty and Tomas Gänsler†

Bell Laboratories, Lucent Technologies
† Agere Systems
700 Mountain Avenue
Murray Hill, NJ 07974, USA
jbenesty@bell-labs.com, gaensler@agere.com

ABSTRACT

Multi-channel acoustic echo cancellation is basically composed of two parts. One part is a multi-channel system identification problem which is nontrivial to solve. The other part, which is also nontrivial, is the so-called *double-talk* detection problem. Near-end speech detection is based on a test statistic. Recently, a new double-talk test statistic based on the normalized cross-correlation vector was proposed for the single-channel case. Obviously, in the multi-channel case, there are several solutions, but what should be the optimal one is not yet known. Then, a fundamental question arises: how do we deal optimally with multiple channels? In this paper, we generalize the idea of normalized cross-correlation vector (single-channel) to the matrix case (multi-channel), derive a frequency-domain version, and show how to combine both the multi-channel frequency-domain adaptive filter and the multi-channel double-talk detector.

1. INTRODUCTION

Multi-channel acoustic echo cancellation is basically composed of two parts. One part is a multi-channel system identification problem which is nontrivial to solve; see [1], [2] for more details. The other part, which is also nontrivial, is the so-called *double-talk* detection problem [3]. When the multiple echo paths are identified by a multi-channel adaptive filter, a function should be included to freeze the adaptation whenever a near-end signal is detected, and thereby avoid the divergence of the adaptive algorithm. This is the role of a double-talk detector (DTD) [4]. Consequently, a suitable DTD decision variable has to be found.

An "optimum"[1] decision variable ξ for double-talk detection should behave as follows:

(i) if double-talk is not present, $\xi \geq T$;

(ii) if double-talk is present, $\xi < T$.

The threshold T must be a constant (in this application), independent of the data. Moreover, ξ must be insensitive to echo path variations when there is no double-talk. A decision variable that exhibits this behavior was proposed in [5] for the single-channel case, using a new normalized cross-correlation vector.

In many acoustic signal processing problems such as, to name a few, voice activity detection, time delay estimation, source localization, double-talk detection, etc, multiple microphone signals are available. However, it is not obvious as to how all of this information should be taken into account for best performance.

In the case of (multi-channel) double-talk detection, there are at least three options. The first one is to consider only a single microphone and build a test statistic based on the normalized cross-correlation vector between the loudspeaker signals and the chosen microphone signal. This is the approach that was taken in [6], [7]. However, this approach may increase the probability of miss since the microphones pick the near-end speech up with different levels and different SNRs, depending on the position of the talker and microphones. Furthermore, not all of this information is taken into account in the decision variable. The second option is to use a test statistic for each microphone, so that the decision made on one microphone is independent of the others. But a large amount of tests may not be easy to handle and we may have some contradictory results on the presence of a near-end talker. In this paper, a third option is proposed which consists of using one test statistic that combines all the (spatial sampling) information of the microphone signals. This approach leads to a nice generalization of the normalized cross-correlation vector to the normalized cross-correlation matrix. By its nature, this method considers all the microphone signals equally. Other possibilities based on this approach can be derived but what will be the optimal choice is not yet known.

2. MULTI-CHANNEL ACOUSTIC ECHO CANCELLATION

First, we assume that we have Q loudspeakers and P microphones. We also assume that the system (room) is linear and time-invariant. Acoustic echo cancellation consists of identifying Q echo paths at each microphone so that in total, PQ echo paths need to be estimated. We have P output (microphone) signals:

$$y_p(n) = \sum_{q=1}^{Q} \mathbf{h}_{qp}^T \mathbf{x}_q(n) + v_p(n), \;\; p = 1, 2, ..., P, \qquad (1)$$

where superscript ^{T} denotes transpose of a vector or a matrix,

$$\mathbf{h}_{qp} = \begin{bmatrix} h_{qp,0} & h_{qp,1} & \cdots & h_{qp,L-1} \end{bmatrix}^T$$

is the echo path – of length L – between loudspeaker q and microphone p,

$$\begin{aligned} \mathbf{x}_q(n) &= \begin{bmatrix} x_q(n) & x_q(n-1) & \cdots & x_q(n-L+1) \end{bmatrix}^T, \\ q &= 1, 2, ..., Q, \end{aligned}$$

is the qth reference (loudspeaker) signal (also called the far-end speech), and v_p is the near-end speech added at microphone p.

[1] Optimum in the sense that for a given probability of false alarm (of double-talk), the probability of miss is minimized.

We define the error signal at time n for microphone p as

$$\begin{aligned} e_p(n) &= y_p(n) - \hat{y}_p(n) \\ &= y_p(n) - \sum_{q=1}^{Q} \hat{\mathbf{h}}_{qp}^T \mathbf{x}_q(n), \qquad (2) \end{aligned}$$

where

$$\hat{\mathbf{h}}_{qp} = \begin{bmatrix} \hat{h}_{qp,0} & \hat{h}_{qp,1} & \cdots & \hat{h}_{qp,L-1} \end{bmatrix}^T$$

are the model filters. It is more convenient to define an error signal vector for all the microphones:

$$\begin{aligned} \mathbf{e}(n) &= \mathbf{y}(n) - \hat{\mathbf{y}}(n) \\ &= \mathbf{y}(n) - \hat{\mathbf{H}}^T \mathbf{x}(n), \qquad (3) \end{aligned}$$

where

$$\begin{aligned} \mathbf{y}(n) &= \mathbf{H}^T \mathbf{x}(n) + \mathbf{v}(n), \\ \mathbf{v}(n) &= \begin{bmatrix} v_1(n) & v_2(n) & \cdots & v_P(n) \end{bmatrix}^T, \\ \mathbf{e}(n) &= \begin{bmatrix} e_1(n) & e_2(n) & \cdots & e_P(n) \end{bmatrix}^T, \\ \mathbf{y}(n) &= \begin{bmatrix} y_1(n) & y_2(n) & \cdots & y_P(n) \end{bmatrix}^T, \\ \hat{\mathbf{y}}(n) &= \begin{bmatrix} \hat{y}_1(n) & \hat{y}_2(n) & \cdots & \hat{y}_P(n) \end{bmatrix}^T, \\ \hat{\mathbf{H}} &= \begin{bmatrix} \hat{\mathbf{h}}_{11} & \hat{\mathbf{h}}_{12} & \cdots & \hat{\mathbf{h}}_{1P} \\ \hat{\mathbf{h}}_{21} & \hat{\mathbf{h}}_{22} & \cdots & \hat{\mathbf{h}}_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{h}}_{Q1} & \hat{\mathbf{h}}_{Q2} & \cdots & \hat{\mathbf{h}}_{QP} \end{bmatrix}, \\ \mathbf{x}(n) &= \begin{bmatrix} \mathbf{x}_1^T(n) & \mathbf{x}_2^T(n) & \cdots & \mathbf{x}_Q^T(n) \end{bmatrix}^T. \end{aligned}$$

Having written the error signal, we now define the cost function

$$\begin{aligned} J &= E\{\mathbf{e}^T(n)\mathbf{e}(n)\} \qquad (4) \\ &= \sum_{p=1}^{P} E\{e_p^2(n)\} \\ &= \sum_{p=1}^{P} J_p. \end{aligned}$$

$E\{\cdot\}$ denotes the statistical expectation operator. The minimization of (4) leads to the multi-channel Wiener-Hopf equation:

$$\mathbf{R}_{xx}\hat{\mathbf{H}} = \mathbf{R}_{xy}, \qquad (5)$$

where

$$\mathbf{R}_{xx} = E\{\mathbf{x}(n)\mathbf{x}^T(n)\} \qquad (6)$$

is the covariance matrix – of size $(QL \times QL)$ – of the reference signals $\mathbf{x}$, and

$$\mathbf{R}_{xy} = E\{\mathbf{x}(n)\mathbf{y}^T(n)\} \qquad (7)$$

is the cross-correlation matrix – of size $(QL \times P)$ – between $\mathbf{x}$ and $\mathbf{y}$.

It can easily be seen that the multi-channel Wiener-Hopf equation (5) can be decomposed in P independent Wiener-Hopf equations, each one corresponding to a microphone signal:

$$\mathbf{R}_{xx}\hat{\mathbf{h}}_p = \mathbf{r}_{xy_p}, \; p = 1, 2, ..., P, \qquad (8)$$

where $\hat{\mathbf{h}}_p$ (resp. $\mathbf{r}_{xy_p}$) is the pth column of matrix $\hat{\mathbf{H}}$ (resp. $\mathbf{R}_{xy}$). This result implies that minimizing J or minimizing each J_p independently gives the same results. This makes sense from an identification point of view, since the identification of the impulse responses for one microphone is completely independent of the others. However, as far as double-talk is concerned, it is preferable to have a global and unique test statistic that takes into account the information of all the microphone signals, since the near-end speech is picked-up by all the microphones with different levels. Choosing one microphone signal and using a single test statistic based on this signal is not enough. On the other hand, using P independent decision variables will be much harder to handle (computational complexity, inconsistency among the different tests, ...). Thus, the approach taken here is to develop a unique test statistic by looking at the covariance matrix $\mathbf{R}_{yy}$ of the microphone signals $\mathbf{y}$.

In the following, we suppose that the covariance matrix $\mathbf{R}_{xx}$ is invertible. To almost guarantee that, we may add (for $Q > 1$) a non-linear (NL) transformation (or add perceptually acceptable uncorrelated noise) to each input signal x_q in order to reduce the coherence of the signals two-by-two [2].

3. A NORMALIZED CROSS-CORRELATION MATRIX FOR MULTI-CHANNEL DOUBLE-TALK DETECTION

For the single-channel case, it has been shown [5] that the so-called normalized cross-correlation vector is well suitable for DTD. In this section, we derive a decision variable based on a normalized cross-correlation matrix for the multi-channel situation.

Suppose that $\mathbf{v} = \mathbf{0}_{P\times 1}$ (no near-end speech). In this case:

$$\begin{aligned} \mathbf{R}_{yy} &= E\{\mathbf{y}(n)\mathbf{y}^T(n)\} \\ &= \mathbf{H}^T\mathbf{R}_{xx}\mathbf{H}. \qquad (9) \end{aligned}$$

Since $\mathbf{y}(n) = \mathbf{H}^T\mathbf{x}(n)$, we have:

$$\mathbf{R}_{xy} = \mathbf{R}_{xx}\mathbf{H} \qquad (10)$$

and (9) may be re-written as

$$\mathbf{R}_{yy} = \mathbf{R}_{xy}^T\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}. \qquad (11)$$

Now, in general for $\mathbf{v} \neq \mathbf{0}_{P\times 1}$,

$$\mathbf{R}_{yy} = \mathbf{R}_{xy}^T\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} + \mathbf{R}_{vv}, \qquad (12)$$

where

$$\mathbf{R}_{vv} = E\{\mathbf{v}(n)\mathbf{v}^T(n)\} \qquad (13)$$

is the covariance matrix of the near-end speech v. Then from (11) and (12), the following decision statistic is proposed,

$$\begin{aligned} \xi &= \frac{1}{\sqrt{P}}\|\mathbf{C}_{xy}\|_E \\ &= \frac{1}{\sqrt{P}}\sqrt{\mathrm{tr}(\mathbf{C}_{xy}^T\mathbf{C}_{xy})} \\ &= \frac{1}{\sqrt{P}}\sqrt{\mathrm{tr}(\mathbf{R}_{xy}^T\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1})}, \qquad (14) \end{aligned}$$

where

$$\mathbf{C}_{xy} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1/2} \qquad (15)$$

is the *normalized cross-correlation matrix* between the two vectors $\mathbf{x}$ and $\mathbf{y}$.

Substituting (10) and (12) into (14),

$$\xi \quad = \quad \frac{1}{\sqrt{P}}\sqrt{\mathrm{tr}[\mathbf{H}^T\mathbf{R}_{xx}\mathbf{H}(\mathbf{H}^T\mathbf{R}_{xx}\mathbf{H}+\mathbf{R}_{vv})^{-1}]}. \qquad (16)$$

It is easily deduced from (16) that for $\mathbf{v} = \mathbf{0}_{P\times 1}$, $\xi = 1$ and for $\mathbf{v} \neq \mathbf{0}_{P\times 1}$, $\xi < 1$. So that we can set the threshold $T = 1$.

For the single-channel case ($Q = P = 1$), (15) becomes the normalized cross-correlation vector between vector $\mathbf{x}_1$ and scalar y_1:

$$\mathbf{c}_{x_1y_1} \quad = \quad (\sigma_{y_1}^2\mathbf{R}_{x_1x_1})^{-1/2}\mathbf{r}_{x_1y_1}, \qquad (17)$$

with $\sigma_{y_1}^2 = E\{y_1^2(n)\}$. The test statistic is now:

$$\begin{aligned} \xi &= \frac{1}{\sqrt{1}}\|\mathbf{c}_{x_1y_1}\|_E \\ &= \sqrt{\mathrm{tr}(\mathbf{c}_{x_1y_1}^T\mathbf{c}_{x_1y_1})} \\ &= \|\mathbf{c}_{x_1y_1}\|_2 \\ &= \sqrt{\mathbf{r}_{x_1y_1}^T(\sigma_{y_1}^2\mathbf{R}_{x_1x_1})^{-1}\mathbf{r}_{x_1y_1}}. \qquad (18) \end{aligned}$$

Thus, the normalized cross-correlation matrix is a natural and elegant extension of the normalized cross-correlation vector [5] to the multi-channel case.

4. A FREQUENCY-DOMAIN APPROACH

In this section, a frequency-domain DTD is derived that will be more useful in practice than its time-domain counterpart, because it is much more efficient from a computational complexity point of view.

We define the block error signal (of length L) for microphone p as:

$$\begin{aligned} \mathbf{e}_p(m) &= \mathbf{y}_p(m) - \hat{\mathbf{y}}_p(m) \\ &= \mathbf{y}_p(m) - \sum_{q=1}^{Q}\mathbf{X}_q(m)\hat{\mathbf{h}}_{qp}, \;\; p = 1, 2, ..., P, \; (19) \end{aligned}$$

where m is the block time index, and

$$\begin{aligned} \mathbf{e}_p(m) &= \left[\begin{array}{ccc} e_p(mL) & \cdots & e_p(mL+L-1) \end{array}\right]^T, \\ \mathbf{y}_p(m) &= \left[\begin{array}{ccc} y_p(mL) & \cdots & y_p(mL+L-1) \end{array}\right]^T, \\ \mathbf{X}_q(m) &= \left[\begin{array}{ccc} \mathbf{x}_q(mL) & \cdots & \mathbf{x}_q(mL+L-1) \end{array}\right]^T. \end{aligned}$$

In the frequency domain, we have:

$$\underline{\mathbf{e}}_p(m) \quad = \quad \underline{\mathbf{y}}_p(m) - \mathbf{G}_1\sum_{q=1}^{Q}\mathbf{D}_q(m)\underline{\hat{\mathbf{h}}}_{qp}, \qquad (20)$$

where

$$\begin{aligned} \underline{\mathbf{e}}_p(m) &= \mathbf{F}\left[\begin{array}{c} \mathbf{0}_{L\times 1} \\ \mathbf{e}_p(m) \end{array}\right], \\ \underline{\mathbf{y}}_p(m) &= \mathbf{F}\left[\begin{array}{c} \mathbf{0}_{L\times 1} \\ \mathbf{y}_p(m) \end{array}\right], \\ \mathbf{G}_1 &= \mathbf{F}\mathbf{W}_1\mathbf{F}^{-1}, \\ \mathbf{W}_1 &= \left[\begin{array}{cc} \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L} \\ \mathbf{0}_{L\times L} & \mathbf{I}_{L\times L} \end{array}\right], \\ \mathbf{D}_q(m) &= \mathbf{F}\mathbf{C}_q(m)\mathbf{F}^{-1}, \\ \mathbf{C}_q(m) &= \left[\begin{array}{cc} \mathbf{X}'_q(m) & \mathbf{X}_q(m) \\ \mathbf{X}_q(m) & \mathbf{X}'_q(m) \end{array}\right], \\ \underline{\hat{\mathbf{h}}}_{qp} &= \mathbf{F}\left[\begin{array}{c} \hat{\mathbf{h}}_{qp} \\ \mathbf{0}_{L\times 1} \end{array}\right]. \end{aligned}$$

$\mathbf{F}$ is the Fourier matrix of size ($2L \times 2L$) and $\mathbf{D}_q$, $q = 1, 2, ..., Q$, are diagonal matrices.

Minimizing the frequency-domain criterion

$$J_\mathrm{f} \quad = \quad \sum_{p=1}^{P} E\{\underline{\mathbf{e}}_p^H(m)\underline{\mathbf{e}}_p(m)\} \qquad (21)$$

leads to the multi-channel Wiener-Hopf equation in the frequency domain:

$$\mathbf{S}_{xx}\underline{\hat{\mathbf{H}}} \quad = \quad \mathbf{S}_{xy}, \qquad (22)$$

where ^{H} denotes conjugate transpose and

$$\begin{aligned} \mathbf{S}_{xx} &= E\{\mathbf{D}^H(m)\mathbf{G}_1\mathbf{D}(m)\}, \\ \mathbf{D}(m) &= \left[\begin{array}{cccc} \mathbf{D}_1(m) & \mathbf{D}_2(m) & \cdots & \mathbf{D}_Q(m) \end{array}\right], \\ \underline{\hat{\mathbf{H}}} &= \left[\begin{array}{cccc} \underline{\hat{\mathbf{h}}}_{11} & \underline{\hat{\mathbf{h}}}_{12} & \cdots & \underline{\hat{\mathbf{h}}}_{1P} \\ \underline{\hat{\mathbf{h}}}_{21} & \underline{\hat{\mathbf{h}}}_{22} & \cdots & \underline{\hat{\mathbf{h}}}_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\hat{\mathbf{h}}}_{Q1} & \underline{\hat{\mathbf{h}}}_{Q2} & \cdots & \underline{\hat{\mathbf{h}}}_{QP} \end{array}\right], \\ \mathbf{S}_{xy} &= E\{\mathbf{D}^H(m)\underline{\mathbf{Y}}(m)\}, \\ \underline{\mathbf{Y}}(m) &= \left[\begin{array}{cccc} \underline{\mathbf{y}}_1(m) & \underline{\mathbf{y}}_2(m) & \cdots & \underline{\mathbf{y}}_P(m) \end{array}\right]. \end{aligned}$$

Following the same philosophy as in Section 3, we define the *pseudo-coherence* matrix:

$$\mathbf{\Gamma}_{xy} = \mathbf{S}_{xx}^{-1/2}\mathbf{S}_{xy}\mathbf{R}'^{\,-1/2}_{yy} \qquad (23)$$

with $\mathbf{R}'_{yy} = E\{\underline{\mathbf{Y}}^H(m)\underline{\mathbf{Y}}(m)\}$. For $Q = P = 1$, $\mathbf{\Gamma}_{xy}$ becomes:

$$\boldsymbol{\gamma}_{x_1y_1} = \mathbf{S}_{x_1x_1}^{-1/2}\mathbf{s}_{x_1y_1}E^{-1/2}\{\underline{\mathbf{y}}_1^H(m)\underline{\mathbf{y}}_1(m)\}, \qquad (24)$$

which is also called the pseudo-coherence vector [6], [7]. The difference between the true coherence and the pseudo-coherence is a different normalization with respect to the signal y_1.

We define the multi-channel frequency-domain decision variable as:

$$\begin{aligned} \xi_\mathrm{f} &= \frac{1}{\sqrt{P}}\|\mathbf{\Gamma}_{xy}\|_E \\ &= \frac{1}{\sqrt{P}}\sqrt{\mathrm{tr}(\mathbf{\Gamma}_{xy}^H\mathbf{\Gamma}_{xy})} \\ &= \frac{1}{\sqrt{P}}\sqrt{\mathrm{tr}(\mathbf{S}_{xy}^H\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{R}'^{\,-1}_{yy})}, \qquad (25) \end{aligned}$$

and it can be checked that for $\mathbf{v} = \mathbf{0}_{P\times 1}$, $\xi_\mathrm{f} = 1$ and for $\mathbf{v} \neq \mathbf{0}_{P\times 1}$, $0 \leq \xi_\mathrm{f} < 1$.

5. COMBINATION OF MULTI-CHANNEL IDENTIFICATION AND DOUBLE-TALK DETECTION IN THE FREQUENCY DOMAIN

In this part, it is shown how to combine both the adaptive identification of the echo paths and the near-end speech detection. Since the DTD will also be adaptive, two different multi-channel model filters (one foreground and one background) will be used, like the two-path model [8]. The background filter will be updated permanently to estimate the test statistic. Each time double-talk is detected, the adaptation of the foreground filter (for identification) will be halted.

A multi-channel frequency-domain adaptive algorithm can be easily derived from the Wiener-Hopf equation. In this section, only the algorithm is given [9], which is:

$$
\begin{aligned}
\hat{\mathbf{S}}_{xx}(m) &= \lambda_{\mathrm{f}}\hat{\mathbf{S}}_{xx}(m-1) + (1-\lambda_{\mathrm{f}})\mathbf{D}^{H}(m)\mathbf{D}(m), && (26)\\
\underline{\mathbf{e}}_{\mathrm{f},p}(m) &= \underline{\mathbf{y}}_{p}(m) - \mathbf{G}_1 \sum_{q=1}^{Q} \mathbf{D}_q(m)\underline{\hat{\mathbf{h}}}_{\mathrm{f},qp}(m-1) && \\
&= \underline{\mathbf{y}}_{p}(m) - \mathbf{G}_1\mathbf{D}(m)\underline{\hat{\mathbf{h}}}_{\mathrm{f},p}(m-1), && (27)\\
\underline{\hat{\mathbf{h}}}_{\mathrm{f},p}(m) &= \underline{\hat{\mathbf{h}}}_{\mathrm{f},p}(m-1) + \mu\mathbf{G}_2\hat{\mathbf{S}}_{xx}^{-1}(m)\mathbf{D}^{H}(m)\underline{\mathbf{e}}_{\mathrm{f},p}(m), && \\
& \quad p = 1, 2, ..., P, && (28)
\end{aligned}
$$

where the subscript f stands for “foreground,” λ_{f}, $0 < \lambda_{\mathrm{f}} < 1$, is an exponential forgetting factor, $\mu = \mu'(1-\lambda_{\mathrm{f}})$, $0 < \mu' \leq 2$, is the adaptation step size, and

$$
\begin{aligned}
\mathbf{G}_2 &= \mathbf{F}\mathbf{W}_2\mathbf{F}^{-1},\\
\mathbf{W}_2 &= \begin{bmatrix} \mathbf{I}_{L\times L} & \mathbf{0}_{L\times L} \\ \mathbf{0}_{L\times L} & \mathbf{0}_{L\times L} \end{bmatrix}.
\end{aligned}
$$

The decision variable should be estimated as follows:

$$
\begin{aligned}
\hat{\xi}_{\mathrm{f}}^{2}(m) &= \frac{1}{P}\mathrm{tr}[\hat{\mathbf{S}}_{xy}^{H}(m)\hat{\mathbf{S}}_{xx}^{-1}(m)\hat{\mathbf{S}}_{xy}(m)\hat{\mathbf{R}}_{yy}^{\prime -1}(m)] && \\
&= \frac{1}{P}\mathrm{tr}[\hat{\mathbf{S}}_{xy}^{H}(m)\underline{\hat{\mathbf{H}}}_{\mathrm{b}}(m)\hat{\mathbf{R}}_{yy}^{\prime -1}(m)], && (29)
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{\mathbf{S}}_{xy}(m) &= \lambda_{\mathrm{b}}\hat{\mathbf{S}}_{xy}(m-1) + (1-\lambda_{\mathrm{b}})\mathbf{D}^{H}(m)\underline{\mathbf{Y}}(m) && (30)\\
\underline{\mathbf{e}}_{\mathrm{b},p}(m) &= \underline{\mathbf{y}}_{p}(m) - \mathbf{G}_1\mathbf{D}(m)\underline{\hat{\mathbf{h}}}_{\mathrm{b},p}(m-1) && (31)\\
\underline{\hat{\mathbf{h}}}_{\mathrm{b},p}(m) &= \underline{\hat{\mathbf{h}}}_{\mathrm{b},p}(m-1) + (1-\lambda_{\mathrm{b}})\mathbf{G}_2\hat{\mathbf{S}}_{xx}^{-1}(m)\mathbf{D}^{H}(m)\underline{\mathbf{e}}_{\mathrm{b},p}(m) && \\
& \quad p = 1, 2, ..., P, && (32)\\
\hat{\mathbf{R}}_{yy}^{\prime}(m) &= \lambda_{\mathrm{b}}\hat{\mathbf{R}}_{yy}^{\prime}(m-1) + (1-\lambda_{\mathrm{b}})\underline{\mathbf{Y}}^{H}(m)\underline{\mathbf{Y}}(m), && (33)
\end{aligned}
$$

subscript b stands for “background,” and λ_{b}, $0 < \lambda_{\mathrm{b}} < 1$, is an exponential forgetting factor. We must choose $\lambda_{\mathrm{b}} < \lambda_{\mathrm{f}}$ (for a faster tracking of the background filter used in the DTD) in order that the DTD alerts the foreground filter before it diverges.

6. CONCLUSIONS

In the literature, a lot of attention has been given to the identification part of multi-channel acoustic echo cancellation but little has been done for near-end speech detection. Multi-channel double-talk detection is not trivial and has to be investigated more deeply. In this paper, some possibilities to handle this problem have been proposed, centered on the use of the normalized cross-correlation matrix as the test statistic. This is a natural extension of the normalized cross-correlation vector used in the single-channel case which has been previously discussed [5]. The advantage of the proposed approach is that all the P microphone signals are taken into account in one decision variable. We also applied this method in the frequency domain and obtained a pseudo-coherence matrix. Finally, an efficient way to combine acoustic echo cancellation and double-talk detection in the frequency domain was proposed.

Acknowledgment

The authors would like to thank Dennis R. Morgan for his helpful suggestions.

7. REFERENCES

[1] M. M. Sondhi, D. R. Morgan, and J. L. Hall, “Stereophonic acoustic echo cancellation—An overview of the fundamental problem,” *IEEE Signal Processing Lett.*, vol. 2, pp. 148–151, Aug. 1995.

[2] J. Benesty, D. R. Morgan, and M. M. Sondhi, “A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation,” *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 156–165, Mar. 1998.

[3] J. H. Cho, D. R. Morgan, and J. Benesty, “An objective technique for evaluating doubletalk detectors in acoustic echo cancelers,” *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 718–724, Nov. 1999.

[4] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, “Acoustic echo control, an application of very-high-order adaptive filters,” *IEEE Signal Processing Mag.*, vol. 16, pp. 42–69, July 1999.

[5] 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, pp. 168–172, Mar. 2000.

[6] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. Springer-Verlag, April 2001.

[7] T. Gänsler and J. Benesty, “A frequency-domain doubletalk detector based on a normalized cross-correlation vector,” *Signal Processing*, 2001, to Appear.

[8] K. Ochiai, T. Araseki, and T. Ogihara, “Echo canceler with two echo path models,” *IEEE Trans. Commun.*, vol. 25, pp. 589–595, June 1977.

[9] J. Benesty and D. R. Morgan, “Multi-channel frequency-domain adaptive filtering,” in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds., Kluwer Academic Publishers, 2000, ch. 7.