Copyright © 2012 IEEE.
Citation for the published paper:

Title:
A Delay-Based Double-Talk Detector

Author:
Christian Schüldt, Fredric Lindström, Ingvar Claesson

Journal:
IEEE Transactions on Audio, Speech, and Language Processing

Year:
2012
Vol:
20
Issue:
6
Pagination:
1725-1733
URL/DOI to the paper:
10.1109/TASL.2012.2188512

This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of BTH's products or services Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by sending a blank email message to pubs-permissions@ieee.org.
By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

A Delay-based Double-talk Detector

Christian Schüldt, *Student Member, IEEE,* Fredric Lindstrom, and Ingvar Claesson, *Member, IEEE*

***Abstract*—When an adaptive filter is used for echo cancellation, it is essential to prevent the filter from diverging in situations when the echo signal is contaminated with near-end disturbance, i.e. during double-talk. This paper presents an extension of a previously proposed double-talk detector for improved performance. It is shown that the computational complexity of the proposed detector is lower than that of the well-used normalized cross correlation (NCC) double-talk detector, at the cost of performance. Further, it is shown that there can be a significant performance difference, in terms of detecting double-talk, between having a fixed echo cancellation filter, which is a common strategy in objective evaluation techniques, and an adaptive filter, which is more close to realistic conditions.**

***Index Terms*—Echo cancellation, adaptive filters, double-talk, double-talk detection.**

I. Introduction

The purpose of an echo canceller is to remove echo of a known output (far-end) signal from an input signal. This is in practice typically achieved with an adaptive finite impulse response (FIR) filter set to model the echo path, creating a replica of the echo which is then subtracted from the input signal. To prevent the adaptive filter from diverging when local (near-end) disturbance is present, a double-talk detector (DTD) can be used.

A scheme of an adaptive echo cancellation filter controlled by a double-talk detector is shown in figure 1. In this case, $\mathbf{h} = [h_0, h_1, \cdots, h_{N-1}]^T$ is the unknown echo path and $\hat{\mathbf{h}}(k) = [\hat{h}_0(k), \hat{h}_1(k), \cdots, \hat{h}_{N-1}(k)]^T$ is the adaptive filter, both assumed, for the sake of simplicity, to be of length N and k is the sample index. Also, $\mathbf{h}$ is considered time-invariant or slowly changing for the sake of simplicity. The driving far-end signal $x(k)$ is filtered with the echo path, forming an echo which in turn is summed with local near-end noise and/or speech $v(k)$, yielding the input signal $y(k) = \mathbf{h}^T\mathbf{x}(k) + v(k)$, where $\mathbf{x}(k) = [x(k), x(k-1), \cdots, x(k-N+1)]^T$ is the regressor vector. The echo is subtracted from the input signal $y(k)$ to obtain an echo cancelled signal

$$e(k) = y(k) - \hat{\mathbf{h}}^T(k)\mathbf{x}(k). \tag{1}$$

Updating of the adaptive filter $\hat{\mathbf{h}}(k)$ can be achieved in many ways [1]. In this paper, the normalized least mean square (NLMS) algorithm is used owing to its simplicity. An NLMS filter update is performed as

$$\hat{\mathbf{h}}(k+1) = \hat{\mathbf{h}}(k) + \mu\frac{e(k)\mathbf{x}(k)}{\mathbf{x}^T(k)\mathbf{x}(k) + \epsilon}, \tag{2}$$

C. Schüldt and I. Claesson are with Blekinge Institute of Technology, Department of Electrical Engineering, SE-37179, Karlskrona, Sweden. (e-mail: christian.schuldt@bth.se; ingvar.claesson@bth.se)

F. Lindstrom is with Limes Audio AB, Tvistevägen 47, SE-90729, Umeå, Sweden. (e-mail: fredric.lindstrom@limesaudio.com)

The funding from the Swedish Knowledge Foundation (KKS) is gratefully acknowledged.

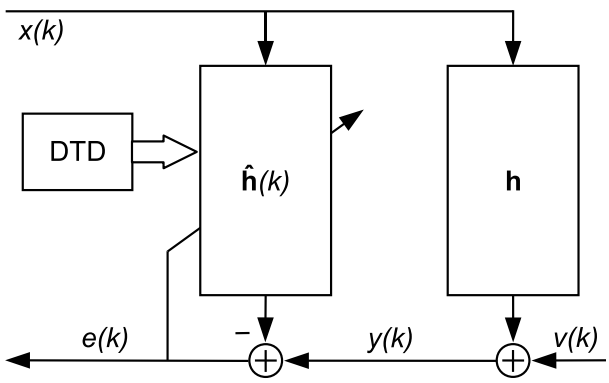


Fig. 1. A generic echo canceller controlled by a DTD.

where ϵ is a regularization term to avoid division by zero and μ is the step-size control parameter [1].

In the case of a significant near-end signal, $v(k)$, risking to interfere with the update of the adaptive filter, the updating of the adaptive filter should be halted to avoid filter divergence. Halting of the filter update in case of near-end disturbance is commonly handled by a DTD. Typically, the DTD calculates a detection statistic ξ, and double-talk is said to be active when ξ is lower than some threshold T. Thus, the filter is updated normally according to equation (2) when $\xi > T$ and when $\xi \leq T$ the update is halted.

Perhaps the most basic double-talk detector is the Geigel detector [2], which compares the far-end and the near-end signal and decides that double-talk is present when the near-end energy is larger than the far-end energy. Other, more recent approaches have been power comparison using cepstral analysis [3] as well as coherence and cross-correlation based techniques [4], [5].

It should be noted that in addition to the mentioned adaptive filter and double-talk detector, a complete echo cancellation solution typically also require components for residual echo removal, feedback estimation, estimation of filter misalignment and rescue detection to prevent the filter from longlasting misadjustment [3], [6]. The focus of this paper is however only the problem of correlation-based double-talk detection. It should also be noted that the proposed solution, and basically any type of DTD, could be used together with a parallel "two-path" adaptive filter structure [7], [3], [8] for preventing erroneous filter updating and for controlling the residual echo suppression.

The outline of the paper is as follows. In section II normalized cross-correlation based double-talk detection is briefly described and the fundamental problem with the so-called

MECC detector is shown analytically. Section III introduces the proposed double-talk detector denoted D-MECC and discusses practical implementation issues in section III-A and computational complexity in section III-B. Then, in section IV double-talk detector evaluation is discussed, and the problem of trying to separate the DTD from the adaptive filtering algorithm is shown. In this section it is also shown by an example that the claim that the performance of the MECC DTD and the NCC DTD are exactly similar [9] does not hold when the DTD operates together with an adaptive echo cancellation filter. Simulations to compare performance of the different DTDs are then described in sections V and VI, showing that the detection performance of the proposed D-MECC lies between that of MECC and the more computationally demanding NCC. Finally, conclusions are presented in section VII.

II. Normalized cross-correlation based double-talk detection

The normalized cross correlation (NCC) double-talk detector, presented in [5], uses the detection statistic

$$\xi_{\text{NCC}} = \sqrt{\mathbf{r}_{\mathbf{x}y}^T(\sigma_y^2\mathbf{R}_{\mathbf{xx}})^{-1}\mathbf{r}_{\mathbf{x}y}}, \tag{3}$$

where $\mathbf{r}_{\mathbf{x}y} = \mathbb{E}[\mathbf{x}(k)y(k)]$ is the cross correlation vector between $\mathbf{x}(k)$ and $y(k)$, $\mathbf{R}_{\mathbf{xx}} = \mathbb{E}\left[\mathbf{x}(k)\mathbf{x}^T(k)\right]$ is the auto-correlation matrix of $x(k)$, $\sigma_y^2 = \mathbb{E}\left[y^2(k)\right]$ is the variance of $y(k)$ (assuming zero mean) and $\mathbb{E}[\cdot]$ denotes expected value. One of the main advantages with this detection statistic is that it achieves normalization in the sense that ξ_{NCC} is 1 when no near-end disturbance is present and $0 < \xi_{\text{NCC}} < 1$ when near-end disturbance is present.

It can be seen that $\mathbf{R}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{r}_{\mathbf{x}y} = \mathbf{h}$ and when in a converged state it is clear that $\mathbf{h} \approx \hat{\mathbf{h}}(k)$. Thus, a common way to reduce the computational complexity is to substitute $\mathbf{R}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{r}_{\mathbf{x}y}$ for $\hat{\mathbf{h}}(k)$ in equation (3) [5], [10], [11], [9]. Equation (3) can then be rewritten as

$$\xi_{\text{NCC}} \approx \sqrt{\frac{\mathbf{r}_{\mathbf{x}y}^T\hat{\mathbf{h}}(k)}{\sigma_y^2}}. \tag{4}$$

To further reduce the computational complexity, one can use the approximation [11]

$$\mathbf{r}_{\mathbf{x}y}^T\hat{\mathbf{h}}(k) \approx r_{y\hat{y}}, \tag{5}$$

where $r_{y\hat{y}} = \mathbb{E}[y(k)\hat{y}(k)]$. Using this approximation and removing the square root, one obtains the double-talk detection statistic [11], [9]

$$\xi_{\text{MECC}} = \frac{r_{y\hat{y}}}{\sigma_y^2} = 1 - \frac{r_{ye}}{\sigma_y^2}, \tag{6}$$

where $r_{ye} = \mathbb{E}[y(k)e(k)]$. In [11], this double-talk detector is denoted Cheap-NCR variant 2 (and the double-talk detector corresponding to equation (3) in this paper is denoted Cheap-NCR variant 1). The approximation in equation (5) was also used in [9], and in that paper the corresponding detector was denoted MECC. Worth noting is that no clear distinction between NCC and MECC in terms of performance is made in neither [11] nor [9]. In fact [9] even states that the performance of NCC and MECC are exactly similar.

By combining equations (1) and (6), assuming that $x(k)$ and $v(k)$ are independent and zero mean and using that $y(k) = \mathbf{h}^T\mathbf{x}(k) + v(k)$ and $\mathbf{r}_{\mathbf{x}y} = \mathbf{h}^T\mathbf{R}_{\mathbf{xx}}$, it can be seen that

$$\xi_{\text{NCC}}^2 \approx \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{xx}}\hat{\mathbf{h}}(k)}{\mathbf{h}^T\mathbf{R}_{\mathbf{xx}}\mathbf{h} + \sigma_v^2}, \tag{7}$$

$$\xi_{\text{MECC}} = \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{xx}}\hat{\mathbf{h}}(k) + \rho(k)}{\mathbf{h}^T\mathbf{R}_{\mathbf{xx}}\mathbf{h} + \sigma_v^2}, \tag{8}$$

where $\rho(k) = \mathbb{E}\left[\hat{\mathbf{h}}^T(k)\mathbf{x}(k)v(k)\right]$. If the adaptive filter does not update during near-end disturbance $v(k)$ it is clear that the adaptive filter coefficients and $v(k)$ are independent, and thus $\rho(k) = 0$ and $\xi_{\text{NCC}}^2 \approx \xi_{\text{MECC}}$. Since $\mathbf{h} \approx \hat{\mathbf{h}}(k)$ when the filter is converged, it can be seen that $\xi_{\text{MECC}} \approx 1$ when no near-end disturbance is present and $\xi_{\text{MECC}} \ll 1$ during strong near-end noise. However, in a situation where the DTD misses the near-end disturbance and updates the filter, the performance of the MECC DTD will deteriorate due to the influence of $\rho(k)$ in equation (8).

The nature of $\rho(k)$ will entirely depend on the filter adaptation algorithm. In the case of NLMS, $\rho(k)$ can be evaluated as follows. By using the NLMS filter update equation (2) rewritten as

$$\begin{aligned}\hat{\mathbf{h}}^T(k) &= \hat{\mathbf{h}}^T(k-1) + \\ &+ \mu\left(\mathbf{h} - \hat{\mathbf{h}}(k-1)\right)^T\frac{\mathbf{x}(k-1)\mathbf{x}^T(k-1)}{\mathbf{x}^T(k-1)\mathbf{x}(k-1)} + \\ &+ \mu v(k-1)\frac{\mathbf{x}^T(k-1)}{\mathbf{x}^T(k-1)\mathbf{x}(k-1)}\end{aligned} \tag{9}$$

(assuming $\epsilon = 0$ for simplicity) to expand the expression for $\rho(k)$, one obtains

$$\begin{aligned}\rho(k) &= \mathbb{E}\left[\hat{\mathbf{h}}^T(k)\mathbf{x}(k)v(k)\right] = \\ &= \mathbb{E}\left[\hat{\mathbf{h}}^T(k-1)\mathbf{X}_1\mathbf{x}(k)v(k)\right] + \\ &+ \mu\mathbb{E}[v(k-1)v(k)]\,\mathbb{E}\left[\frac{\mathbf{x}^T(k-1)\mathbf{x}(k)}{\mathbf{x}^T(k-1)\mathbf{x}(k-1)}\right],\end{aligned} \tag{10}$$

where $\mathbf{X}_i = \left(\mathbf{I} - \mu\frac{\mathbf{x}(k-i)\mathbf{x}^T(k-i)}{\mathbf{x}^T(k-i)\mathbf{x}(k-i)}\right)$ and $\mathbf{I}$ is the $N\times N$ identity matrix. It should be noted that

$$\mu\mathbb{E}\left[\mathbf{h}^T\frac{\mathbf{x}(k-1)\mathbf{x}^T(k-1)}{\mathbf{x}^T(k-1)\mathbf{x}(k-1)}\mathbf{x}(k)v(k)\right] = 0, \tag{11}$$

owing to the independence and zero-mean of $x(k)$ and $v(k)$ and the fact that $\mathbf{h}$ is considered constant. Again, using the NLMS filter update equation (9) to expand equation (10) yields the expression

$$\begin{aligned}\rho(k) &= \mathbb{E}\left[\hat{\mathbf{h}}^T(k-2)\mathbf{X}_2\mathbf{X}_1\mathbf{x}(k)v(k)\right] + \\ &+ \mu\mathbb{E}[v(k-2)v(k)]\,\mathbb{E}\left[\frac{\mathbf{x}^T(k-2)\mathbf{X}_1\mathbf{x}(k)}{\mathbf{x}^T(k-2)\mathbf{x}(k-2)}\right] \\ &+ \mu\mathbb{E}[v(k-1)v(k)]\,\mathbb{E}\left[\frac{\mathbf{x}^T(k-1)\mathbf{x}(k)}{\mathbf{x}^T(k-1)\mathbf{x}(k-1)}\right],\end{aligned} \tag{12}$$

It can thus be seen that continuing to expand the expression

for $\rho(k)$ using the NLMS filter update equation (9) gives

$$\begin{aligned}\rho(k) &= \mathbb{E}\left[\hat{\mathbf{h}}^T(k-M)\left(\prod_{i=1}^{M}\mathbf{X}_i\right)\mathbf{x}(k)v(k)\right] + \\ &+ \mu\sum_{i=1}^{M}\mathbb{E}\left[v(k-i)v(k)\right]\times \\ &\times \mathbb{E}\left[\frac{\mathbf{x}^T(k-i)}{\mathbf{x}^T(k-i)\mathbf{x}(k-i)}\left(\prod_{j=0}^{i-1}\mathbf{X}_j\right)\mathbf{x}(k)\right], \quad (13)\end{aligned}$$

where $\mathbf{X}_0 = \mathbf{I}$ and M is the number of expansions. By considering $k = 0$ as the starting index and thus $\hat{\mathbf{h}}(0)$ as the initial adaptive filter vector, it is clear that the first term is 0 since $x(k)$ and $v(k)$ are assumed to be independent and zero-mean. (Also, the adaptive filter typically has all coefficients set to zero initially.) The relation $k = M$ will hold for all $k > 0$, since the reference is always the chosen starting point $\rho(0) = 0$. Hence, the resulting expression becomes

$$\begin{aligned}\rho(k) &= \mu\sum_{i=1}^{k}\mathbb{E}\left[v(k-i)v(k)\right]\times \\ &\times \mathbb{E}\left[\frac{\mathbf{x}^T(k-i)}{\mathbf{x}^T(k-i)\mathbf{x}(k-i)}\left(\prod_{j=0}^{i-1}\mathbf{X}_j\right)\mathbf{x}(k)\right], \\ &k \geq 1. \quad (14)\end{aligned}$$

Several conclusions can be drawn from equation (14). First of all, it is obvious that the disturbance $\rho(k)$ is directly proportional to the step-size parameter μ. Further, if any of the signals $x(k)$ and $v(k)$ are white, then $\rho(k) = 0$. In the case of speech, the magnitude of the first factor in equation (14) is likely to decrease as i increases, since the autocorrelation of a speech signal usually decrease rapidly as the lag increases [12], [13], [14]. Further, since all eigenvalues if $\mathbf{X}_i$ are non-negative and ≤ 1, the magnitude of the eigenvalues of the matrix resulting of the product $\prod_{j=0}^{i-1}\mathbf{X}_j$ are monotonically decreasing as i increases. This is intuitive since it can be argued that recent activity should have more influence on the disturbance than earlier activity.

III. Proposed double-talk detector

The detection statistic proposed in this paper is based on the same idea as in [12], where a delay is introduced to reduce the influence of near-end disturbance in a filter deviation measure, although in this paper the idea is used in a DTD context. The proposed detection statistic is

$$\xi_{\text{D-MECC}} = 1 - \frac{r_{y_D e_D}}{\sigma^2_{y_D}}, \quad (15)$$

where $r_{y_D e_D} = \mathbb{E}\left[y(k-D)e_D(k)\right]$, $\sigma^2_{y_D} = \mathbb{E}\left[y(k-D)y(k-D)\right]$ and $e_D(k) = y(k-D) - \hat{\mathbf{h}}^T(k)\mathbf{x}(k-D)$.

As in the previous section, using equations (1) and (6), equation (15) can be rewritten as

$$\xi_{\text{D-MECC}} = \frac{\mathbf{h}^T\mathbf{R}_{\mathbf{x_D x_D}}\hat{\mathbf{h}}(k) + \rho_D(k)}{\mathbf{h}^T\mathbf{R}_{\mathbf{x_D x_D}}\mathbf{h} + \sigma^2_{v_D}}, \quad (16)$$

where $\rho_D(k) = \mathbb{E}\left[\mathbf{x}^T(k-D)\hat{\mathbf{h}}(k)v(k-D)\right]$ and $\mathbf{R}_{\mathbf{x_D x_D}} = \mathbb{E}\left[\mathbf{x}(k-D)\mathbf{x}^T(k-D)\right]$. Using the same recursive approach, inserting the NLMS update equation equation (9), as previously for $\rho(k)$, an expression for $\rho_D(k)$ can be obtained as

$$\begin{aligned}\rho_D(k) &= \mu\sum_{i=1}^{k}\mathbb{E}\left[v(k-i)v(k-D)\right]\times \\ &\times \mathbb{E}\Big[\frac{\mathbf{x}^T(k-i)}{\mathbf{x}^T(k-i)\mathbf{x}(k-i)}\left(\prod_{j=0}^{i-1}\mathbf{X}_j\right)\times \\ &\times \mathbf{x}(k-D)\Big], \quad k \geq 1. \quad (17)\end{aligned}$$

It is obvious that $\rho(k) = \rho_D(k)$ for $D = 0$. As in [12], it can be argued that $|\rho_D(k)| < |\rho(k)|$ should hold in most cases since the auto-correlation for speech decreases as the lag increases. For $D < 0$ this is trivial to realize by comparing equations (14) and (17). For $D > 0$ however, as shown in [13], it does not always hold that $|\rho_D(k)| < |\rho(k)|$. For example, $D = 1$ and $\mu = 1$ will result in the first term of $\rho_D(k)$ being $\mathbb{E}\left[v^2(k-1)\right]$ which is likely to give even worse performance than $D = 0$. Nevertheless, as D increases, the disturbance term $|\rho_D(k)|$ is likely to decrease since then the largest first factor $\mathbb{E}\left[v^2(k-i)\right]$ will be multiplied with a second factor of smaller magnitude (since, as argued before, the magnitude of the eigenvalues of the matrix resulting of the product $\prod_{j=0}^{i-1}\mathbf{X}_j$ are monotonically decreasing as i increases). This agrees well with the simulated results in [13].

To illustrate the behavior of $\rho_D(k)$, simulations to estimate the components of equation (17) were conducted. The signal $v(k)$ was chosen as $v(k) = 0.9v(k-1) + w_1(k)$ where $w_1(k) \sim \mathcal{N}(0,1)$ and the signal $x(k)$ was chosen as $x(k) = 0.95x(k-1) + w_2(k)$ where $w_2(k) \sim \mathcal{N}(0,1)$. The parameters μ and N were set to 0.95 and 16, respectively. The reason for choosing a comparatively short filter length was to avoid excessive computations. Ensemble averages were taken over 10^5 runs and the results are shown in figures 2, 3 and 4. From the figures, it can clearly be seen that a low D reduces the magnitude of $\rho_D(k)$.

Thus, $\xi_{\text{D-MECC}}$ should then be a better choice for detection statistic than ξ_{MECC} for $D < 0$. A setting of $D = -32$ was chosen in this paper.

A. Practical considerations

In practice, $\mathbf{r}_{\mathbf{x}y}$, r_{ye} and σ^2_y can be estimated using a running average over a time-window [5], [11] or exponential recursive weighting [9], [12] as

$$\begin{aligned}\hat{\mathbf{r}}_{\mathbf{x}y}(k) &= \lambda\hat{\mathbf{r}}_{\mathbf{x}y}(k-1) + (1-\lambda)\mathbf{x}(k)y(k), \\ \hat{r}_{ye}(k) &= \lambda\hat{r}_{ye}(k-1) + (1-\lambda)y(k)e(k), \\ \hat{\sigma}^2_y(k) &= \lambda\hat{\sigma}^2_y(k-1) + (1-\lambda)y^2(k), \quad (18)\end{aligned}$$

where $\hat{\mathbf{r}}_{\mathbf{x}y}(k)$ is to approximate $\mathbf{r}_{\mathbf{x}y}$, $\hat{r}_{ye}(k)$ is to approximate r_{ye} and $\hat{\sigma}^2_y(k)$ is to approximate σ^2_y, respectively, and λ is a forgetting factor. This is the approach used hereinafter, with a forgetting factor $\lambda = 0.995$ used for all three DTD and for all simulations. Variables used in the proposed detection statistic

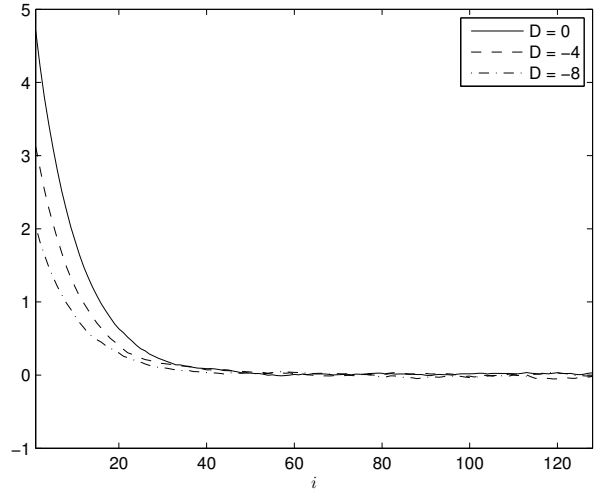


Fig. 2. Estimation of $\mathbb{E}\left[v(k-i)v(k-D)\right]$ for $v(k) = 0.9v(k-1)+w_1(k)$ where $w_1(k) \sim \mathcal{N}(0,1)$, $\mu = 0.95$ and $N = 16$. Ensemble average is taken over 10^5 runs.

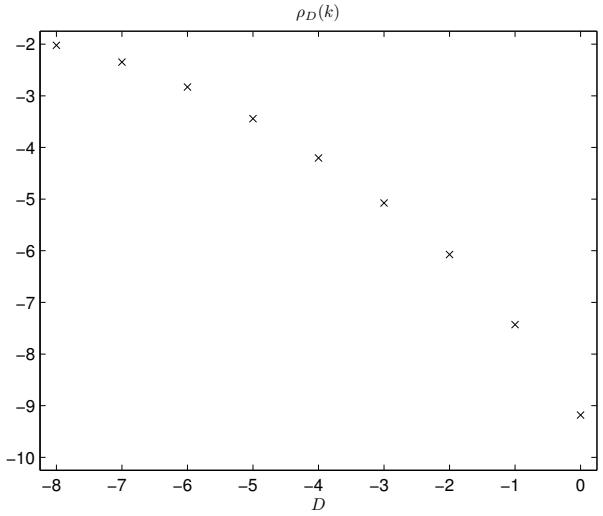


Fig. 4. Estimation of $\rho_D(k)$ for $v(k) = 0.9v(k-1) + w_1(k)$ where $w_1(k) \sim \mathcal{N}(0,1)$ and $x(k) = 0.95x(k-1) + w_2(k)$ where $w_2(k) \sim \mathcal{N}(0,1)$, $\mu = 0.95$ and $N = 16$. Ensemble average is taken over 10^5 runs.

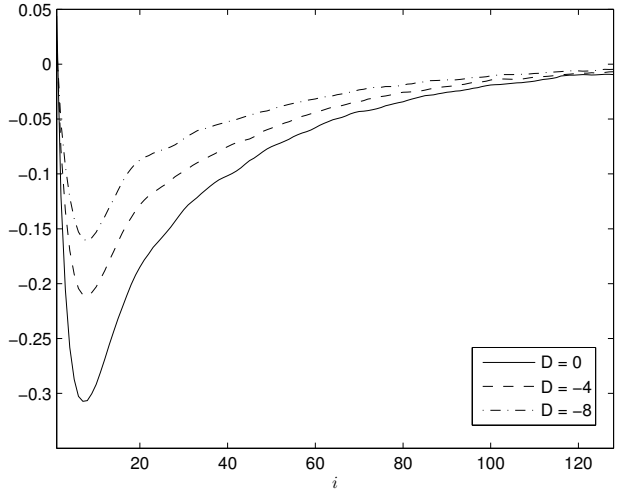


Fig. 3. Estimation of $\mathbb{E}[\frac{\mathbf{x}^T(k-i)}{\mathbf{x}^T(k-i)\mathbf{x}(k-i)}\left(\prod_{j=0}^{i-1}\mathbf{X}_j\right)\mathbf{x}(k-D)]$ for $x(k) = 0.95x(k-1) + w_2(k)$ where $w_2(k) \sim \mathcal{N}(0,1)$, $\mu = 0.95$ and $N = 16$. Ensemble average is taken over 10^5 runs.

are estimated analogously. The variable λ determines the trade-off between sensitivity and robustness, i.e. a small forgetting factor results in averages that change rapidly over time and quickly adapt to changes, while a large forgetting factor gives averages which are more consistent (robust). Hence, a small λ would imply rapid but not so accurate detection of double-talk, while a large λ would imply more accurate, but less rapid double-talk detection.

As discussed in the previous section, the proposed algorithm is based on a delay D. One approach for implementation with $D < 0$ is to introduce a delay $|D|$ in the signal path of $y(k)$, yielding a causal process. The proposed DTD then operates on the "early" signals $y(k-D)$ and $\mathbf{x}(k-D)$ preceding the delay, and the adaptive filter operates on the delayed signals $y(k)$ and $\mathbf{x}(k)$. The downside of this implementation is of course the delay introduced in the signal path, but since $|D|$ typically is comparatively small this delay is acceptable in some applications such as Voice-over-IP endpoints. If the delay is not acceptable (for instance, the ITU-T recommendation G-168 establishes that the delay in the "receive path" should not exceed 250 μs), an alternative approach would be to store D previous versions of the adaptive filter $\hat{\mathbf{h}}(k)$. On the other hand, this significantly increases the amount of required memory.

Nevertheless, a third, much more efficient, approach is to calculate $\hat{\mathbf{h}}^T(k)\mathbf{x}(k-D)$ directly using knowledge about previous filter updates. First of all the index is changed using $n = k - D$ for the sake of clarity, yielding $\hat{\mathbf{h}}^T(n+D)\mathbf{x}(n)$. It should be kept in mind that only the case of $D < 0$ is considered here. First of all, inserting the NLMS update equation (2) into the expression for $\hat{\mathbf{h}}^T(n)\mathbf{x}(n)$ gives

$$\begin{aligned}\hat{\mathbf{h}}^T(n)\mathbf{x}(n) &= \left(\hat{\mathbf{h}}(n-1) + \beta(n-1)\mathbf{x}(n-1)\right)^T\mathbf{x}(n) \\ &= \hat{\mathbf{h}}^T(n-1)\mathbf{x}(n) + \\ &\quad +\beta(n-1)\mathbf{x}^T(n-1)\mathbf{x}(n), \end{aligned} \tag{19}$$

where $\beta(n) = \mu\frac{e(n)}{\mathbf{x}^T(n)\mathbf{x}(n)+\epsilon}$. Continuing to recursively expand equation (19) using the NLMS update equation (2) yields

$$\hat{\mathbf{h}}^T(n)\mathbf{x}(n) = \hat{\mathbf{h}}^T(n+D)\mathbf{x}(n) + \sum_{i=1}^{|D|}\beta(n-i)\alpha_i(n), \tag{20}$$

where $\alpha_i(n) = \mathbf{x}^T(n-i)\mathbf{x}(n)$. It is thus obvious that

$$\hat{\mathbf{h}}^T(n+D)\mathbf{x}(n) = \hat{\mathbf{h}}^T(n)\mathbf{x}(n) - \boldsymbol{\beta}^T(n)\boldsymbol{\alpha}^T(n), \tag{21}$$

where the vectors $\boldsymbol{\beta}(n) = [\beta(n-1), \beta(n-2), \cdots, \beta(n-|D|)]^T$ and $\boldsymbol{\alpha}(k) = [\alpha_1(n), \alpha_2(n), \cdots, \alpha_{|D|}(n)]^T$ are both of length $|D|$. It should be noted that $\hat{\mathbf{h}}^T(n)\mathbf{x}(n)$ is calculated in the adaptive filtering update and is thus available without additional computational cost. It should also be noted that

this approach introduces no signal delay and avoids storing of previous filter coefficients.

Since the proposed DTD uses an "old" copy of the adaptive filter, $\hat{\mathbf{h}}(n+D)$, it is in a sense more sensitive to an echo path change than the related methods. This is likely to become apparent in situations with short filters, large $|D|$ and abrupt changes of the echo-path. A straight forward approach to avoid problems related to this, such as e.g. dead-lock, is to use the proposed DTD together with a parallel two-path adaptive filter structure [7], [3], [8].

B. Computational complexity

In terms of computational complexity, the NCC double-talk detector requires $2N+2$ multiplications and $N+1$ additions just for calculating $\hat{\mathbf{r}}_{\mathbf{x}y}(k)$ and $\hat{\sigma}_y^2(k)$ in equation (18). Further, an additional N multiplications and additions as well as one division are required for evaluating equation (6) (ignoring the square-root), resulting in a total of $3N+2$ multiplications, $2N+1$ additions and one division per evaluation, i.e. typically per input sample.

Using similar exponential recursive weighing as in equation (18), it is clear that ξ_{MECC} can be evaluated very efficiently, using a total of 4 multiplications, two additions, one subtraction and one division per evaluation.

As a comparison, D-MECC with the direct approach of storing previous filter coefficients or introducing a signal delay requires, in addition to the complexity of the MECC, N multiplications, N additions and one subtraction to calculate $e_D(k)$. On the other hand, D-MECC with the approach of calculating $\hat{\mathbf{h}}^T(k)\mathbf{x}(k-D)$ as explained in the previous section requires evaluation of equation (21) together with one subtraction (for obtaining $e_D(k)$) in addition to the complexity of the MECC.

The vector $\boldsymbol{\beta}(n)$, used in the scalar product in equation (21) can be obtained at practically no additional computational cost, since $\beta(n)$ which is calculated in the NLMS update equation (2) just has to be delayed/stored in the vector. The elements of the vector $\boldsymbol{\alpha}(k)$ can be obtained recursively at low computational cost as $\alpha_i(n) = \alpha_i(n-1) - x(n-N)x(n-N-i) + x(n)x(n-i)$, i.e. using just two multiplications, one addition and one subtraction per element. Thus, in addition to the complexity of the MECC, D-MECC with the approach of calculating $\hat{\mathbf{h}}^T(k)\mathbf{x}(k-D)$ requires $3|D|$ multiplications and $3|D|+1$ additions/subtractions per evaluation.

IV. Evaluation of double-talk detection performance

An objective technique for evaluating the performance of DTDs based on receiver operating characteristics (ROC) was presented in [10] and has since been used in numerous publications. The technique is carried out by first selecting a *probability of false alarm*, P_{f}, i.e. the probability of declaring detection when double-talk is not present, and finding appropriate detection thresholds $\{T_{\text{NCC}}, T_{\text{MECC}}, T_{\text{D-MECC}}\}$ that correspond to the selected P_{f} setting by using speech signals where double-talk is not present. Then, simulations using speech signals with double-talk are carried out, using the respective thresholds corresponding to the chosen P_{f} setting, and the *probability of miss*, P_{m}, i.e. the probability of failing to detect double-talk when double-talk is present, is calculated. The probability of miss is evaluated over a range of near-end to acoustic echo ratios (NER) to give an indication of the performance of the double-talk detection algorithms in different situations. For a more detailed description of the evaluation technique, the reader is referred to [10].

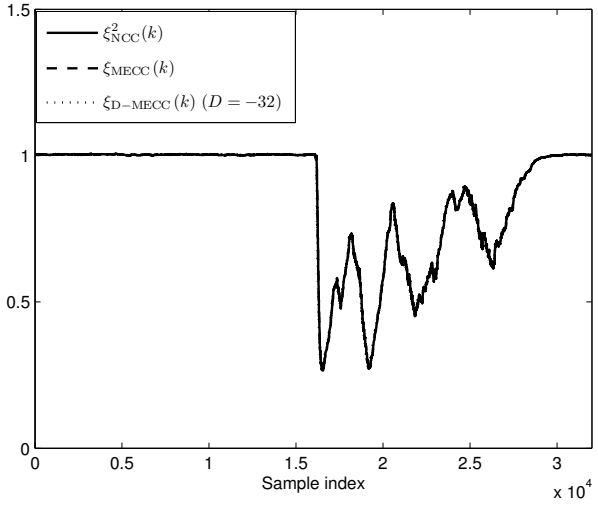


Fig. 5. Comparison of DTD statistics in a situation with a fixed echo cancellation filter. Double-talk occurs after sample index 16200.

One important aspect of the evaluation technique, as originally presented, is that the adaptive filter is assumed to be converged throughout the simulation. Thus, a fixed filter with a pre-determined misalignment at -30 dB, generated by perturbing the actual room response samples, is used [10]. Naturally, this is done to remove the dependence of the adaptive algorithm. However, as will later become apparent, this dependence can in some cases be crucial for the performance of the DTD. Thus, in those cases, using a fixed filter will not reflect the true performance of the double-talk detector in a real environment together with an actual adapting filter.

In figure 5, the detection statistics for the three considered DTDs; NCC, MECC and D-MECC in a simulation with a fixed filter, at a pre-determined misalignment at -30 dB, are shown. The far-end signal was a colored stationary signal generated as $x(n) = 0.9x(n-1) + w(n)$, where $w(n) \sim \mathcal{N}(0, 4 \times 10^{-4})$ and the near-end signal was a speech signal, becoming active after 16200 samples, corresponding to approximately 2 seconds with 8 kHz sample rate. As can be seen from figure 5, all three double-talk detection statistics perform identical. This is indeed in agreement with the results of [9].

Shown in figure 6 is the resulting detection statistics from the same simulation as described above, with the single difference that a constantly updating NLMS-based adaptive filter with step-size parameter $\mu = 1$ was used instead of a fixed filter. The filter was allowed to converge to a steady-state before activating the detection statistics, and is still updating as double-talk occurs, simulating a situation where the DTD fails to detect double-talk or a parallel filter ("two-path")

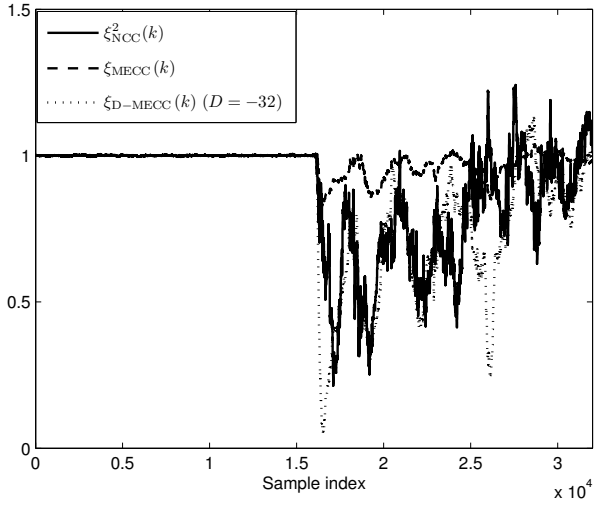


Fig. 6. Comparison of DTD statistics in a situation with a constantly updating adaptive echo cancellation filter. Double-talk occurs after sample index 16200.

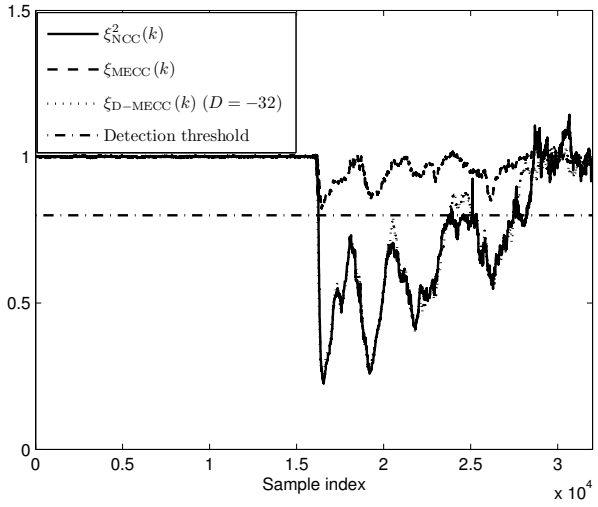


Fig. 7. Comparison of DTD statistics in a situation where the adaptive echo cancellation filter is halted when the DTD statistic is < 0.8. Double-talk occurs after sample index 16200.

implementation where the DTD is coupled to the constantly updating background filter [8]. From figure 6, it can be seen that all compared detection statistics behave very differently during double-talk, when the adaptive filter is updating. Worth noting is that the detection statistic for MECC in this case is significantly worse than the NCC and D-MECC detection statistics - a threshold setting of 0.8 would in this case mean that the MECC completely misses to detect the double-talk. This situation is shown in figure 7 where it clearly can be seen that MECC fails to detect doubletalk, while the detection statistics for NCC and D-MECC both behave similar to the situation with no filter updates in figure 5.

Further, a double-talk detection threshold of 0.8 would also mean that both NCC and D-MECC detect doubletalk approximately at the same time, after $25 - 75$ samples (depending on exactly at which sample near-end speech is considered active). However, using the implementation of D-MECC with the delay D in the signal path of $y(k)$, as described in section III-A, will result in a delayed update of the adaptive filter, i.e. the DTD will operate on $y(k - D)$ and the adaptive filter will operate on $y(k)$. Thus, the adaptive filter will in the D-MECC case update $|D|$ less iterations during actual double-talk than NCC (since the NCC DTD operates on $y(k)$), resulting in a reduced risk of a misadjusted filter.

In conclusion; NCC, MECC and D-MECC all show identical performance in the case of a fixed filter. In a more realistic scenario with an adaptive filter however, it is apparent that the performance of the three algorithms are very different. Therefore, further simulations to compare the performance of the algorithms are performed.

V. Simulations

To evaluate the performance of the double-talk detectors, the evaluation method in [10] was used, with the modification that an adaptive NLMS-based filter was used for echo cancellation instead of a fixed filter. The adaptive filter was constantly updating with step-size parameter $\mu = 0.95$ during far-end single-talk and was halted when the evaluated algorithm declared double-talk. The length of the adaptive filter was set to $N = 500$, which was the same length as the fixed echo-generating filter which was obtained by measurement in a standard office. The forgetting factor was set to $\lambda = 0.995$ for all averages, see section III-A.

Like in [10], the far-end signal with duration of 12.5 seconds was from a male talker, sampled at 8 kHz and four different speech signals (two male and two female) of approximately 2 s each and also sampled at 8 kHz, were used as near-end speech. The near-end speech was set to occur at four different positions in time (at 6.25 s, 7.5 s, 8.75 s or 10 s) within the 12.5 s far-end speech. Independent flat spectrum noise with different intensity was added to the near-end signal, resulting in three different cases with echo-to-noise ratio (ENR) of 10 dB, 20 dB and 30 dB. In all cases the adaptive filter did reach a steady-state in less than 5 seconds and after this the double-talk evaluation was initiated. Hence the adaptive filter misalignment depended only on the near-end noise intensity during the double-talk evaluation.

A. Results

The simulation results for $P_{\mathrm{f}} = 0.1$ over a range of NERs and ENRs are shown in figures 8, 9 and 10. It can be seen that the NCC detector performs significantly better than the MECC detector, while the performance of the proposed D-MECC detector lies close to that of NCC for low NERs and relatively high misalignment (NER 10 dB shown in figure 8) and goes down towards (and occasionally surpasses) NCC for high NERs and for less misalignment (NER 20 dB and 30 dB shown in figures 9 and 10).

Simulations were also performed for $P_{\mathrm{f}} = 0.3$, and the results are visible in figures 11, 12 and 13. In these cases NCC still shows the best performance, although the D-MECC performance is close. MECC shows the worst performance.

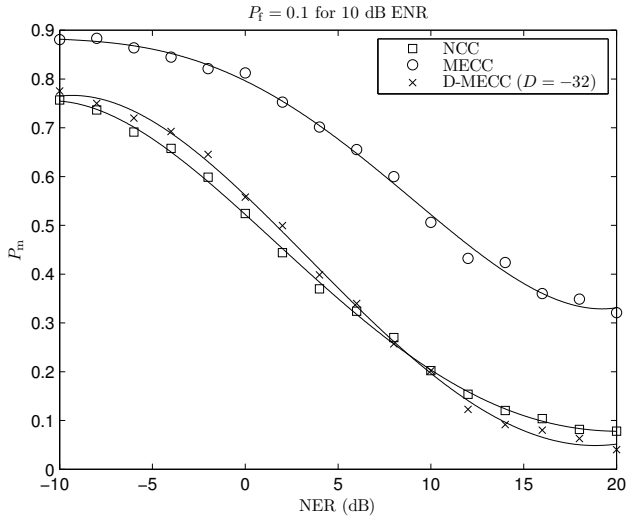


Fig. 8. Performance of the three DTDs for $P_\mathrm{f} = 0.1$ with 10 dB echo-to-noise ratio (ENR).

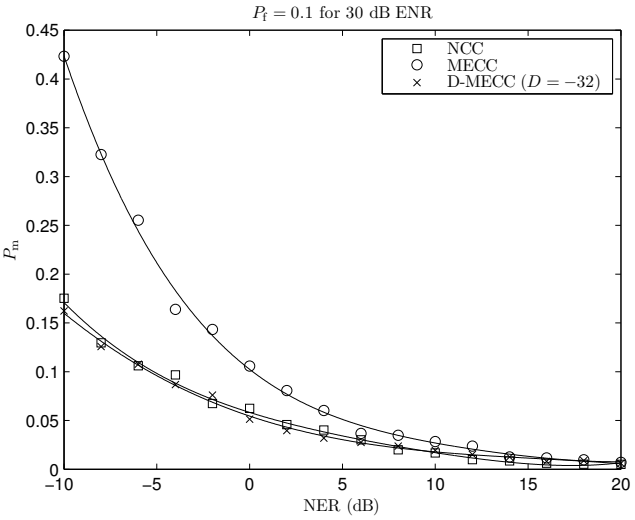


Fig. 10. Performance of the three DTDs for $P_\mathrm{f} = 0.1$ with 30 dB echo-to-noise ratio (ENR).

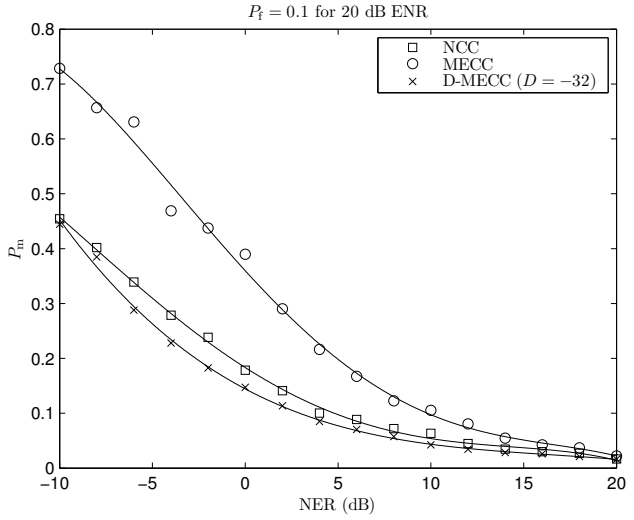


Fig. 9. Performance of the three DTDs for $P_\mathrm{f} = 0.1$ with 20 dB echo-to-noise ratio (ENR).

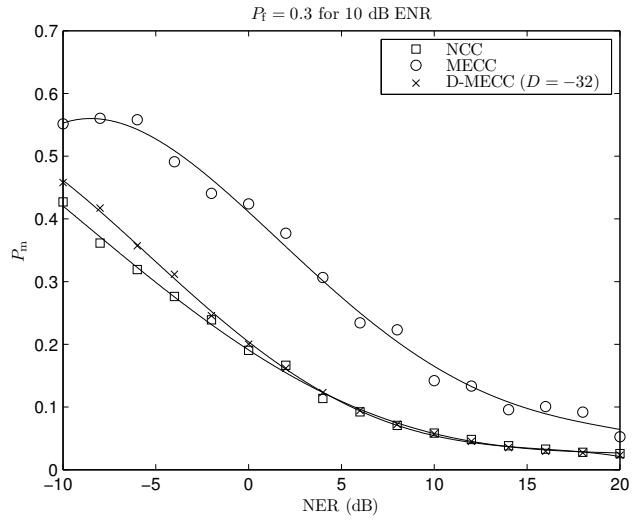


Fig. 11. Performance of the three DTDs for $P_\mathrm{f} = 0.3$ with 10 dB echo-to-noise ratio (ENR).

It should be noted that the difference in performance between MECC and NCC decrease for increased ENR and NER. The performance difference also seems to be smaller for $P_\mathrm{f} = 0.3$ than for $P_\mathrm{f} = 0.1$. This is probably owing to the fact that the adaptive filter is halted more often in these cases (as the DTD detects double-talk more frequently), and once the adaptive filter is halted, the performance of NECC and MECC are identical in theory given the same adaptive filter.

VI. Experiments with recorded signals

Experiments were also carried out with signals recorded in a small office. The loudspeaker and microphone were placed with approximately 50 cm distance from each other on a desk and the loudspeaker volume was set so that the ENR was approximately 24 dB. In this case it was also necessary to increase the adaptive filter length to $N = 1000$ in order to capture most of the echo tail (all other parameters were unchanged). As in the simulations in section V, the adaptive filter was allowed to converge for 5 seconds before double-talk was applied. Double-talk was applied in the same manner as previously.

The results of the experiments are shown in figures 14 and 15. Figure 14 shows the result for $P_\mathrm{f} = 0.1$ and figure 15 shows the result for $P_\mathrm{f} = 0.3$. It can be seen that all three DTDs in figure 14 show very similar results to the simulation with ENR set to 20 dB displayed in figure 9. The reason for the experiments not showing improved results, despite an ENR increase of 4 dB, is probably due to the colored background noise in the case of the recordings.

As in the simulations, NCC and D-MECC are fairly similar, while MECC exhibits worse performance.

It is thus clear that in realistic situations, with an adaptive

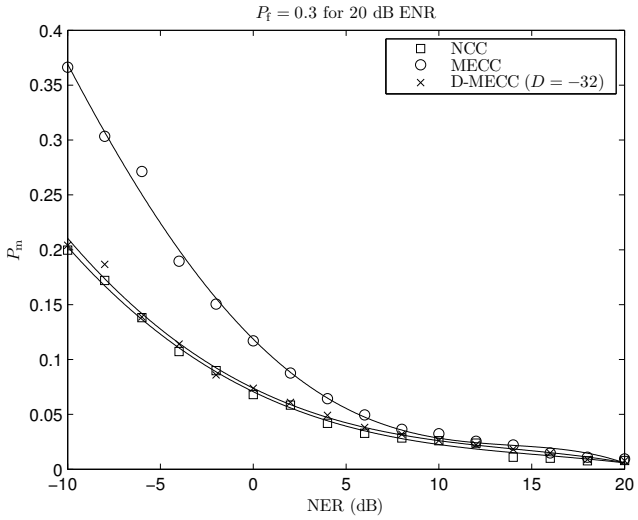


Fig. 12. Performance of the three DTDs for $P_f = 0.3$ with 20 dB echo-to-noise ratio (ENR).

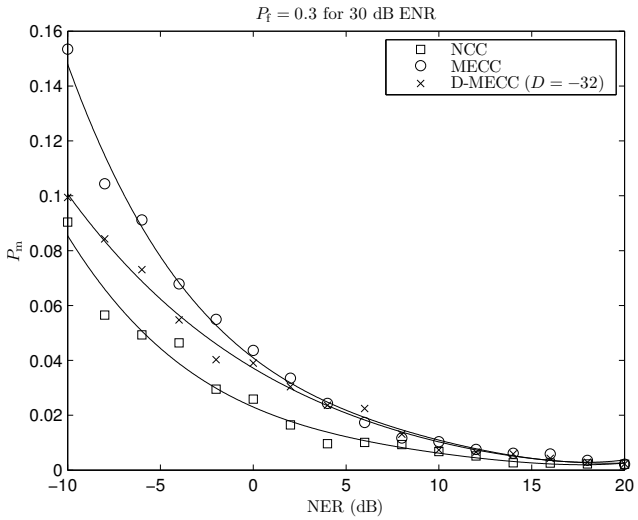


Fig. 13. Performance of the three DTDs for $P_f = 0.3$ with 30 dB echo-to-noise ratio (ENR).

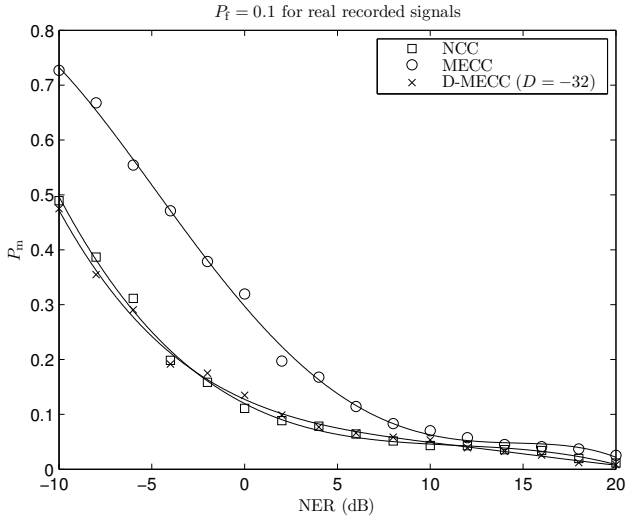


Fig. 14. Performance of the three DTDs for $P_f = 0.1$ with recorded signals.

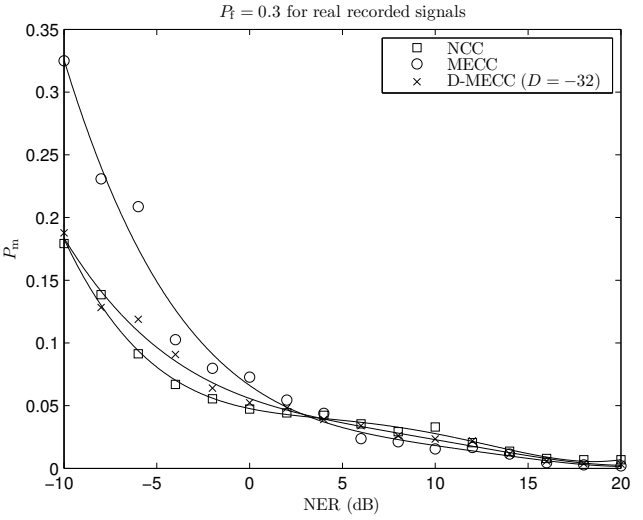


Fig. 15. Performance of the three DTDs for $P_f = 0.3$ with recorded signals.

NLMS filter which updates when the DTD does not indicate double-talk, the NCC detector is superior to the MECC detector. The proposed D-MECC detector performs overall significantly better than the MECC detector and slightly worse than the NCC detector. On the other hand, in terms of computational complexity, MECC has by far the lowest, followed by D-MECC, while the NCC requires the most, see section III-A.

VII. Conclusions

In [9] is it claimed that the performance of the MECC double-talk detector and the NCC double-talk detector are exactly similar. In this paper it was shown that this holds only under the assumption of a fixed echo cancellation filter. In a realistic situation with an adaptive NLMS filter updating when the DTD does not indicate double-talk, the MECC performs significantly worse than the NCC detector. This has been verified by simulations. Further, a novel DTD named D-MECC with computational complexity slightly higher than the MECC but much lower than NCC, has been proposed. It has been shown through simulations that the D-MECC performance is significantly better than MECC and appears to become comparable to that of NCC when the adaptive filter misalignment is low.

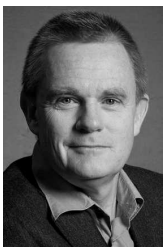
References

[1] S. Haykin, *Adaptive Filter Theory*, 4th ed. Prentice-Hall, 2002.
[2] D. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Transactions on Communications*, vol. COM-26, pp. 647–653, May 1978.
[3] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic cancellation filters - an overview," *Signal Processing*, vol. 80, pp. 1697–1719, 2000.
[4] T. Gansler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Transactions on Communication*, vol. 44, pp. 1421–1427, November 1996.
[5] J. Benesty, D. Morgan, and J. 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.

[6] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Wiley, 2004.
[7] K. Ochiai, T. Araseki, and T. Ogihara, "Echo canceler with two echo path models," *IEEE Transactions on Communications*, vol. COM-25, no. 6, pp. 8–11, June 1977.
[8] F. Lindstrom, C. Schüldt, and I. Claesson, "An improvement of the two-path algorithm transfer logic for acoustic echo cancellation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1320–1326, May 2007.
[9] M. Iqbal, J. Stokes, and S. Grant, "Normalized double-talk detection based on microphone and AEC error cross-correlation," in *Proceedings of IEEE International Conference on Multimedia and Expo*, July 2007, pp. 360–363.
[10] J. H. Cho, D. R. Morgan, and J. 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.
[11] P. Åhgren, "On system identification and acoustic echo cancellation," Ph.D. dissertation, Uppsala University, 2004.
[12] C. Schüldt, F. Lindstrom, and I. Claesson, "An improved deviation measure for two-path echo cancellation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 305–308.
[13] ——, "Evaluation of an improved deviation measure for two-path echo cancellation," in *Proceedings of IWAENC International Workshop on Acoustic Echo and Noise Control*, September 2010.
[14] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

Christian Schüldt was born in Stockholm, Sweden. He received his B.Sc. degree in electrical engineering with specialization in mobile communication from the Royal Institute of Technology, Stockholm, Sweden, in 2001 and his M.Sc. degree in electrical engineering with specialization in signal processing and computer vision from the Royal Institute of Technology, Stockholm, Sweden, in 2004. Currently he is pursuing his Ph.D. studies at the Department of Electrical Engineering, Blekinge Institute of Technology, Karlskrona, Sweden. His areas of professional interest include signal processing techniques and real-time programming.

Fredric Lindström was born in Skellefteå, Sweden. He received his M.Sc. degree in applied physics from Uppsala University, Uppsala, Sweden, in 2001 and the Ph.D. degree from Blekinge Institute of Technology, Ronneby, Sweden, in 2007. Since 2007, he is the CEO of Limes Audio AB, Umeå, Sweden. His current research interest is adaptive signal processing with applications in hands-free systems, e.g. acoustic echo canceling, acoustic echo suppression techniques, and algorithms for finite precision implementations.



Ingvar Claesson (M'91) received the M.Sc. degree in 1980 and the Ph.D. degree in 1986 in Electrical Engineering at University of Lund, Sweden. He was appointed Senior Lecturer in Telecommunication Theory at Lund University in 1986, and was appointed Associate Professor in 1992. In 1990, he was one of the founders of the Department of Signal Processing, Blekinge Institute of Technology, and is currently Head of Research and Principal Supervisor in Signal Processing. Since 1998, he holds the chair of Applied Signal Processing at Blekinge Institute of Technology, Karlskrona, Sweden, and also served as Research Dean 2005-2011. His current research interests are in adaptive signal processing, blind equalization, adaptive beamforming, speech enhancement, blind signal separation, active noise control, health applications, filter design and remote laboratories. A keen interest for applications has lead to more than 20 patents and he regularly serves as industry consultant.