An Objective Technique for Evaluating Doubletalk Detectors in Acoustic Echo Cancelers

Jun H. Cho, Member, IEEE, Dennis R. Morgan, Senior Member, IEEE, and Jacob Benesty, Member, IEEE

Abstract—Echo cancelers commonly employ a doubletalk detector (DTD), which is essential to keep the adaptive filter from diverging in the presence of near-end speech and other disruptive noise. There have been numerous algorithms to detect doubletalk in an acoustic echo canceler (AEC). In those applications, typically, the threshold is chosen only by some heuristic method and the performance evaluation is very subjective. In this study, we develop a way to objectively evaluate DTD algorithms based on the standard statistical methods of detection theory. A receiver operating characteristic (ROC) is derived to characterize DTD performance. Several DTD algorithms are examined and simulated under typical real-world operating conditions using measured room responses and signals taken from a digital speech database. The DTD methods are then evaluated and compared using the ROC metric.

Index Terms—Acoustic echo cancellation, adaptive filter, cross-correlation, doubletalk detection.

I. INTRODUCTION

A n acoustic echo canceler (AEC) [1] removes the undesired echo signal that feeds back from the loudspeaker to the microphone in full-duplex hands-free telecommunication. The cancellation is done by modeling the echo path impulse response with an adaptive finite impulse response (FIR) filter and subtracting the echo from the microphone output signal. A basic AEC diagram is shown in Fig. 1. The far-end speech signal $x$ goes through the echo path represented by a filter $\hat{h}$ and adds to the microphone signal $y$ together with the near-end talker signal $v$ and noise $n$:

$$y(k) = \hat{h}^T x(k) + v(k) + n(k)$$  \hspace{1cm} (1)

where

$$\hat{h} = [\hat{h}_0 \hat{h}_1 \cdots \hat{h}_{M-1}]^T,$$

$$x(k) = [x(k) x(k-1) \cdots x(k-M+1)]^T$$

and $M$ is the length of the echo path response. The error signal is defined as

$$e(k) = y(k) - [\hat{h}^T 0] x(k) = \Delta \hat{h}^T x(k) + v(k) + n(k)$$  \hspace{1cm} (2)

where

$$\Delta \hat{h} = [\hat{h}_0 \hat{h}_1 \cdots \hat{h}_{I-1}]^T$$  \hspace{1cm} (3)

is the adaptive filter coefficient vector of length $L$ (generally less than $M$), and

$$\Delta \hat{h} = h - [\hat{h} 0].$$  \hspace{1cm} (4)

When the near-end talker is silent, i.e., $v = 0$, and there is little noise, the adaptive filter $\hat{h}$ can converge to a good estimate of the echo path response $h$ and successfully cancel the echo. However, when the two talkers on both sides speak at the same time—the doubletalk situation—the near-end speech acts as uncorrelated noise to the adaptive algorithm and the filter may diverge, causing annoying audible echo to pass through to the far-end. The best way known so far to remedy this problem is to detect the presence of near-end speech and use that to halt filter adaptation. This is the important role of the doubletalk detector (DTD).

The basic doubletalk detection scheme starts with computing a detection statistic and comparing it with a preset threshold. Different methods have been proposed to form the detection statistic. The Geigel algorithm [2] has proven successful in line echo cancelers, however, it does not always provide reliable performance when used in AEC’s. Recently, cross-correlation based methods [3]–[5] have been studied which appear to be more appropriate for AEC applications. For these DTD algorithms, however, only heuristic methods have been used to select the threshold $T$ with little justification for the choice. Also, there has not been an objective way to evaluate and compare these methods. In this work, we develop a quantitative method of setting the threshold and evaluating the performance of different methods based on detection theory concepts. We simulate the above mentioned DTD methods and study their behavior using this procedure. This article is organized as follows. Section II describes the DTD methods simulated in this work. Section III explains...
the proposed evaluation method and simulation procedure. Section IV presents the simulation results and compares these DTD methods. Section V draws conclusions and suggests possible further research.

II. DOUBLETALK DETECTION METHODS

In general, doubletalk is handled in the following way.

1) A detection statistic $\xi$ is formed using available signals, e.g., $x, y, e$, etc., and the estimated filter coefficients $\hat{h}$.
2) The detection statistic $\xi$ is compared to a preset threshold $T$; and doubletalk is declared if $\xi > T$.
3) Once doubletalk is declared, the detection is held for a minimum period of time $T_{\text{h持}}$. While the detection is held, the filter adaptation is disabled.
4) If $\xi \leq T$ consecutively over a time $T_{\text{h持}}$, the filter resumes adaptation, while the comparison of $\xi$ to $T$ continues until $\xi > T$ again.

The hold time $T_{\text{h持}}$ in Steps 3 and 4 is necessary to suppress detection dropouts due to the noisy behavior of the detection statistic. Although there are some possible variations, most of the DTD algorithms keep this basic form and only differ in how to form the detection statistic. In this paper, we consider three different methods to compute the detection statistic.

A. Geigel Algorithm

One simple algorithm due to A. A. Geigel is to declare the presence of near-end speech whenever

$$\xi(g) = \frac{|y(k)|}{\max\{|x(k-1)|, \ldots, |x(k-N)|\}} > T$$

where $N$ and $T$ are suitably chosen constants [2]. This detection scheme is based on a waveform level comparison between the microphone signal $y$ and the far-end speech $x$ assuming the near-end speech $v$ in the microphone signal will be typically stronger than the echo $h$. The maximum or $l_\infty$ norm of the $N$ most recent samples of $x$ is taken for the comparison because of the undetermined delay in the echo path. The threshold $T$ is to compensate for the energy level of the echo path response $h$, and is often set to 1/2 for line echo cancelers because the hybrid loss is typically about 6 dB. For an AEC, however, it is not easy to set a universal threshold to work reliably in all the various situations because the loss through the acoustic echo path can vary greatly depending on many factors. For $N$, one easy choice is to set it the same as the adaptive filter length $L$.

B. Cross-Correlation Method

Ye and Wu [3] proposed a double-talk detection algorithm based on the cross-correlation between $x$ and $e$. A similar idea using the cross-correlation between $x$ and $y$ is also suggested by Wesel [4].

The cross-correlation vector between $x$ and $y$ and between $x$ and $e$ are defined as

$$c_{xy}^{(1)} = [c_{x,y,0}^{(1)} \ c_{x,y,1}^{(1)} \ \cdots \ c_{x,y,L-1}^{(1)}]^T$$

where

$$c_{x,y}^{(1)} = \frac{E\{x(k-i)y(k)\}}{\sqrt{E\{x^2(k-i)\}E\{y^2(k)\}}}$$

and

$$c_{xe}^{(1)} = [c_{x,e,0}^{(1)} \ c_{x,e,1}^{(1)} \ \cdots \ c_{x,e,L-1}^{(1)}]^T$$

where

$$c_{x,e}^{(1)} = \frac{E\{x(k-i)e(k)\}}{\sqrt{E\{x^2(k-i)\}E\{e^2(k)\}}}$$

The operator $E\{\cdot\}$ denotes statistical expectation.

The detection statistic $\xi$ is formed by taking the inverse norm of the cross-correlation vector. Any scalar metric is possible in taking the norm such as the $l_1$ metric used in [3], $l_2$, or $l_\infty$. In this simulation, the $l_\infty$ norm is used to form the detection statistics

$$\xi_{xy}^{(1)} = \left[\max_i |c_{x,y,i}^{(1)}| \right]^{-1}$$

and

$$\xi_{xe}^{(1)} = \left[\max_i |c_{x,e,i}^{(1)}| \right]^{-1}$$

where $c_{x,y,i}^{(1)}$ and $c_{x,e,i}^{(1)}$ are estimates of $\hat{c}_{x,y,i}^{(1)}$ and $\hat{c}_{x,e,i}^{(1)}$, respectively. A time average or exponentially windowed sum is used for the estimation of these statistical quantities, e.g.,

$$E\{x(k-i)y(k)\} \approx (1 - e^{-1/W}) \sum_{j=0}^{\infty} x(k-j) \cdot y(k-j)e^{-j/W}$$

and other statistical expectations are estimated analogously. The effective window length $W$ needs be long enough for smooth estimation, but should not be too long because of the nonstationary nature of the speech signal and the desirability of rapid response.

C. Normalized Cross-Correlation Method

We also consider a new method based on the cross-correlation vector proposed in [6]. That method achieves proper normalization in the sense that the detection statistic is equal to one when the near end signal is zero. The normalized cross-correlation vector is defined as

$$c_{xy}^{(2)} = (\sigma_y^2 \sigma_x^{-1/2})^{-1/2} c_{xy}^{(1)}$$

where $\sigma_y^2 = E\{y^2\}$ is the variance of $y$, $\sigma_x = E\{xX^T\}$ is the autocorrelation matrix of $x$, and $\Sigma_{xy} = E\{xy^T\}$ is the cross-correlation vector between $x$ and $y$. The corresponding detection statistic is obtained by taking the inverse $l_2$ norm of the estimated correlation vector

$$\xi_{xy}^{(2)} = ||\hat{c}_{xy}^{(2)}||^{-1}.$$  (14)

Now if $\xi_{xy}^{(2)} \approx \xi_{xy}^{(2)}$, we have

$$\xi_{xy}^{(2)} \approx \left[f_{xy}^{(1)} (\sigma_y \sigma_x)^{-1} \Sigma_{xy} \right]^{-1/2}.$$  (15)

For computational simplicity, it can be modified as follows. Assuming that the length of the adaptive filter is long enough
to accurately model the room response, i.e., \( L = M \), we have 
\[
\mathbf{R}^{-1} \mathbf{r}_{xy} = \mathbf{h} \approx \hat{\mathbf{h}}
\]
when the filter is converged. In this case, (15) can be rewritten as
\[
\epsilon_{xy}^{(2)} = \frac{\sigma_y}{\sqrt{\mathbf{R}^{-1}_{xy} \mathbf{h}}} \approx \frac{\sigma_y}{\sqrt{\mathbf{R}_{xy}^{-1} \hat{\mathbf{h}}}}.
\]
(16)
This form is significantly easier to compute than (15) but may incur some loss in accuracy substituting \( \hat{\mathbf{h}} \) for \( \mathbf{h} \) when the filter is not yet converged. Again, we use exponentially windowed time averages to estimate \( \mathbf{r}_{xy} \) and \( \sigma_y \) as in (12).

### III. EVALUATION METHOD AND SIMULATION PROCEDURE

In the DTD methods discussed in the previous section, the role of threshold \( T \) is essential to the performance. However, there hasn’t been a systematic approach to select the value of \( T \). Also, it has been very hard to compare different methods objectively. To solve these difficulties, we view DTD as a binary detection problem and apply the detection theory concepts that have already been developed for radar and communication applications. We note that a similar detection theory approach was adopted by Gänssler [5], however it was only applied to theoretical models; here we extend these techniques to evaluating DTD performance in actual operating environments with real speech signals and room acoustics.

The general characteristics of a binary detection scheme are as follows.

- **Probability of False Alarm (\( P_f \))**: Probability of declaring detection when a target is not present.
- **Probability of Detection (\( P_d \))**: Probability of successful detection when a target is present.
- **Probability of Miss (\( P_m = 1 - P_d \))**: Probability of detection failure when a target is present.

A “good” detection method should maximize \( P_d \) while minimizing \( P_f \) even in a low signal-to-noise ratio (SNR) situation. In general, higher \( P_d \) is achieved at the cost of higher \( P_f \). There should be a tradeoff in performance depending on the penalty or cost function of a false alarm and of a miss. One common approach to characterize different detection methods is to represent the detection characteristic \( P_d \) as a function of SNR under a given constraint on the false alarm probability \( P_f \). This is known as a receiver operating characteristic (ROC) and is widely used to characterize detection schemes in radar and communication applications. The \( P_f \) constraint can be interpreted as the maximum tolerable false alarm rate.

We use a similar approach to evaluate a DTD, with special considerations specific to the AEC application. The evaluation is based on the probability of miss \( P_m \) performance for a given false alarm probability \( P_f \). The \( P_f \) is measured as the proportion of the far-end speech in which doubletalk remains declared when there is no near-end speech. For the AEC application, the penalty of false alarm is small because it simply halts the filter adaptation for a period of time \( T_{halo} \). When the adaptive filter has converged, halting filter adaptation does not incur a penalty. Even while converging, there is no penalty other than a \( T_{halo} \) percentage delay of the convergence time. For example, if \( P_f = 0.1 \), then the algorithm still converges 90% as fast as it would without false alarms. Therefore, a moderately high false alarm probability \( P_f \) is tolerable. In this work, \( P_f \) in the range 0.1–0.3 is chosen as a constraint. The miss probability \( P_m \) is measured as the proportion of near-end speech duration that remains undetected at different levels of near-end to far-end speech ratio (NFR, \( \sigma_x/\sigma_y \)) in the range from -20 dB to 10 dB. The \( P_m \) characteristic is a meaningful criterion to fairly compare different DTD methods of this kind, because the disruptive effect of undetected doubletalk on an adaptive filter depends on the time-averaged near-end speech that goes undetected. Speed of initial detection is also important; however, it is not necessarily a good indicator of performance because of possible subsequent detection dropouts. That is the reason we prefer \( P_m \) as the performance criterion. The magnitude of the disruption of course depends on the near-end signal level and that is why we calculate and plot as a function of NFR, so that the ultimate performance can be determined from the NFR distribution for any particular application by visual interpretation of these plots.

In order to measure these characteristics in an absolute sense, exhaustive simulations of the DTD methods with a real AEC are required. However, it is not necessary to simulate all possible diverse situations for **comparative** evaluation of DTD algorithms. Making reasonable assumptions and confining the situation to representative cases is essential to moderate the complexity of the simulation and to better analyze the operation of a DTD.

The far-end and near-end speech signals were taken from a database sampled at 8 kHz. The far-end speech (male talker) is 4.9 s long or 39200 samples at 8 kHz sampling rate. For the near-end, four different sentences (two male, two female) were chosen, each about 2 s long. Different levels of NFR are generated by attenuating the near-end speech \( \psi \) correspondingly. For the room response \( h_r \), we use a measured response shown in Fig. 2. The performance of DTD’s will be dependent on the choice of the room echo path response. However, since the main objective of this work is to compare different detection methods and analyze their behavior under...
For all the simulations in this study, we have included a Gaussian white noise at the near-end side, as shown in Fig. 1. The noise power was set at 30 dB relative to the power of the far-end signal (which is also the echo power in this case). The adaptive filter used for the AEC here is assumed to be of length corresponding to 64 ms at the 8 kHz sampling rate. We do not simulate the adaptive algorithm here, but rather assume for simplicity that it is converged throughout the simulation, since the filter remains converged most of the time in an actual situation. The misalignment of the filter is assumed to be 30 dB, which is a typical value. If necessary, 20 dB, 10 dB or even higher values can be chosen to simulate a still-adapting filter. The converged adaptive filter coefficient is generated by perturbing the actual room response samples

where is an uncorrelated Gaussian noise sequence with zero mean and 30 dB variance.

A miss or false alarm is counted only during the active portion of the far-end speech because the effect of a pause or inactive portion of the speech on the filter update is minimal. An activity detector (Fig. 3) is used to sense this condition. An identical activity detector is also applied to the near-end speech because silence in also does not usually cause the filter to diverge. In both cases, the constant is chosen to be to ensure inactivity over the full filter length interval . We chose a lagging window here to match the presumed lagging response of the AEC, in order to properly assess activity effects. The activity threshold was selected as 30 dB.

The overall simulation procedure is illustrated in Fig. 4, which is basically explained by the four steps in Section II. The hold time is a parameter which normally is set to 20–30 ms. The detection statistic is calculated using the detection methods described in Section II. For the Geigel algorithm, (5) is used. For the correlation based methods explained in Sections II-B and II-C, (10), (11), and (16) are used. Before measuring , the threshold is predetermined to meet the given constraint as follows. First, the detection statistic is calculated with as a function of the threshold. The probability of false alarm at each threshold point is calculated as

where is the DTD output, is the activity detector output (Fig. 3), and is the length of the entire far-end speech signal . The logical AND with the activity of is necessary to disregard false alarms during innocuous periods of inactivity. Then, the threshold is determined to achieve the given . Once the threshold is determined, the near-end speech is applied at different attenuation levels, and the detection procedure runs again. The miss probability is calculated as

\[
P_m = 1 - \frac{\sum \phi \cdot \bar{\tau} \cdot \bar{v}}{\sum \bar{\tau} \cdot \bar{v}}
\]
where $\phi$ is the DTD output and $x$ and $\bar{v}$ are the activity detector outputs for $x$ and $\bar{v}$ respectively (Fig. 4). The logical AND's in the numerator ensures that the miss probability is counted only when both $x$ and $\bar{v}$ are active, as previously discussed.

In order to gain better statistical significance, the $P_1$ and $P_{1m}$ calculations are averaged over 16 different conditions: four different 2-s near-end speech samples located at different positions within the 4.9 s far-end speech. The complete DTD evaluation technique is summarized as follows.

1) Set $v = 0$.
   a) Select threshold $T_1$.
   b) Compute $P_1$ using (18).
   c) Repeat steps a, b over a range of threshold values.
   d) Select threshold value that corresponds to $P_1 = 0.1$ or 0.3.

2) Select NFR value.
   a) Select one of four 2-s near-end speech samples.
   b) Select one of four positions within 4.9-s far-end speech.
   c) Compute $P_{1m}$ using (19).
   d) Repeat steps a, b, c over all sixteen conditions.
   e) Average $P_{1m}$ over all sixteen conditions.

3) Plot average $P_{1m}$ as a function of NFR.

IV. SIMULATION RESULTS AND COMPARISON

Simulations were carried out for the three different DTD methods discussed in Section II following the procedure explained in Section III. The $P_{1m}$ characteristics were measured under the constraints $P_1 = 0.1$ and $P_1 = 0.3$. Different settings of parameters such as the vector length $N$ for Geigel algorithm, the estimation window length $W$ for the cross-correlation methods, and the hold time $T_{\text{holding}}$ for all were observed to observe their effect on the performance of the DTD’s. The $P_{1m}$ characteristics simulated with $T_{\text{holding}} = 30$ ms are plotted with respect to NFR in Figs. 5–9.

The $P_{1m}$ characteristics of the Geigel algorithm are shown in Fig. 5. The vector length was set to $N = 512$ for the simulation. It can be easily observed that the $P_{1m}$ naturally converges to $1 - P_1$ as NFR gets smaller, i.e., $v \to 0$. The $P_{1m}$ remains quite high and only reached 0.1–0.2 at NFR as high as 10 dB. Simulation results with different parameter settings showed that detection improves ($P_{1m}$ reduces) with increasing $T_{\text{holding}}$ and degrades ($P_{1m}$ increases) with decreasing $T_{\text{holding}}$ in the high NFR region. However, too long a value of $T_{\text{holding}}$ is not desirable in practice because it would delay the filter update when room response changes coincide with either false alarms or detected doubletalk. Decreasing the vector length $N$ worsened the performance, because the smoothness of the detection statistic deteriorates, resulting in a higher threshold $T$ to meet the same $P_1$.

Figs. 6 and 7 show the characteristics of the cross-correlation method using (10) and (11) respectively, with window length $W = 500$. Both showed high probability of miss over the entire range of NFR. The poor performance is due to a lack of proper normalization of $\xi_{\phi x}^{(1)}$ and $\xi_{\bar{v} x}^{(1)}$ with and without the near-end speech signal. There was no improvement for other choices of window length $W$. 

Fig. 5. Performance of Geigel algorithm, $N = 512$.

Fig. 6. Performance of cross-correlation method, $\xi_{\phi x}^{(1)}$, $W = 500$.

Fig. 7. Performance of cross-correlation method, $\xi_{\bar{v} x}^{(1)}$, $W = 500$. 

Fig. 8. Performance of normalized cross-correlation method, $\xi_{2y}$. $W = 500$.

Fig. 9. Performance of normalized cross-correlation method, $\xi_{2y}$. $W = 60$.

For the normalized cross-correlation method using (16), ideally, we expect $P_m = 0$ and $P_f = 0$ all the time with the threshold $T = 1$, if the theoretical normalization $\xi_{2y}^{(2)} = 1$ when $v = 0$ can be achieved. In practice, however, the ideal behavior of $\xi_{2y}^{(2)}$ is perturbed by using $\hat{h}$ in place of $h$ in (16). The length of $h$, $M$, is generally larger than that of the modeling filter $\hat{h}$, $L$, thus the tail part of $h$ is not modeled by the adaptive filter. This unmodeled tail portion generates extra terms and corrupts the behavior of $\xi_{2y}^{(2)}$. See Appendix A for a detailed discussion of this “tail effect.” Fig. 8 is the result with estimation window length $W = 500$. The performance is greatly improved over all other methods. Changing the hold time affected the performance in the same general manner as with Geigel algorithm: $P_m$ improved with longer $T_{\text{hold}}$ and degraded with shorter $T_{\text{hold}}$, though the changes were minimal. More drastic change is observed when the estimation window length $W$ is changed. Fig. 9 shows the results when $W$ is decreased to 60 while $T_{\text{hold}} = 30$ ms is kept the same. This is because $\xi_{2y}^{(2)}$ becomes less smooth as $W$ is reduced. Larger $W$ enables a more accurate estimate of the statistical expectation values in (16), thus reducing the tail effect. However, making $W$ too large is not desirable because speech signals are nonstationary over a long period of time, thereby exacerbating the tail effect.

Among the DTD methods simulated in this work, the normalized cross-correlation method using $\xi_{2y}^{(2)}$ showed the best performance with appropriate estimation window length $W$. It achieved the smallest miss probability over a wide range of NFR without making the hold time too long.

V. CONCLUSION

We have proposed a method to evaluate DTD algorithms using the standard setup of detection theory to derive an ROC. The proposed method is effective in characterizing DTD performance and enables objective comparison of different algorithms. The Geigel algorithm and two cross-correlation based methods were simulated and evaluated using the proposed method. Among these DTD’s, the normalized cross-correlation method proposed in [6] exhibited the best performance.

This study was primarily concerned with the steady-state performance of the DTD after the AEC has converged. For the simulations, it was assumed that the adaptive algorithm had converged to $-30$ dB misalignment. However, more work is necessary to fully evaluate DTD’s in a dynamic situation where doubletalk occurs while the AEC is adapting. As mentioned in the text, the proposed evaluation method could be applied to this situation by selecting other misalignment levels to simulate a still-adapting filter. This would be particularly important for the simplified normalized cross-correlation DTD algorithm, where the room response is approximated by substituting the adaptive filter coefficients. Some degradation would be expected in that case and remedies are needed.

APPENDIX A

TAIL EFFECT

The ideal behavior of the normalized cross-correlation method discussed in Section II-C is valid only under the assumption that the length $L$ of the adaptive filter is equal to the length of the echo path response $M$. However, since $L$ is generally shorter than $M$, some degradation is inevitable. Here, we quantify this degradation in terms of the unmodeled “tail” of the room response.

Let the actual room response $h$ be decomposed as

\[ h = [\hat{h} \; \bar{h}] \tag{A1} \]

where $\hat{h}$ is the vector comprising the first $L$ samples of the room response and $\bar{h}$ is the $(M - L) \times 1$ tail vector. The far-end speech signal vector $x$ and its autocorrelation matrix are decomposed likewise:

\[ x = [\hat{x} \; \bar{x}] \tag{A2} \]

\[ R_x = \begin{bmatrix} R_{\hat{x}\hat{x}} & R_{\hat{x}\bar{x}} \\ R_{\bar{x}\hat{x}} & R_{\bar{x}\bar{x}} \end{bmatrix} \tag{A3} \]
The ideal detection statistic (15) in squared form is

\[ (\xi_{xy}^{(2)})^2 = \frac{\sigma_y^2}{t_x^T P_x r_{xy} t_y}. \]  

(A4)

In our implementation, we replace

\[ R_x^{-1} r_{xy} \approx \hat{h} \]

and use the fact that \( r_{xy} = R_x \hat{h} \) to get the denominator of (A4)

\[ t_x^T R_x^{-1} r_{xy} \approx t_x^T \hat{h} \]

\[ = [\hat{h}^T \hat{h}^T] [R_x R_x R_x] \hat{h} \]

\[ = \hat{h}^T R_x \hat{h} + \hat{h}^T R_x R_x \hat{h}. \]  

(A5)

The numerator of (A4) is

\[ \sigma_y^2 = \hat{h}^T R_x \hat{h} \]

\[ = [\hat{h}^T \hat{h}^T] [R_x R_x R_x] \hat{h} \]

\[ = \hat{h}^T R_x \hat{h} + \hat{h}^T R_x R_x \hat{h} + \hat{h}^T R_x \hat{h}. \]  

(A6)

Combining (A4), (A5), and (A6), we get

\[ (\xi_{xy}^{(2)})^2 = \frac{\hat{h}^T R_x \hat{h} + \hat{h}^T R_x R_x \hat{h} + \hat{h}^T R_x \hat{h}}{\hat{h}^T R_x \hat{h} + \hat{h}^T R_x R_x \hat{h}}. \]  

(A7)

If the estimation of the autocorrelation matrix \( R_x \) is accurate, \( R_x R_x \approx 0 \) and \( R_x R_x \approx 0 \) due to the limited correlation time of speech. Then assuming that the filter is converged, i.e., \( \hat{h} \approx \hat{h} \) (true to -30 dB for the simulations in this paper), (A7) can be simplified to

\[ (\xi_{xy}^{(2)})^2 \approx 1 + \frac{\hat{h}^T R_x \hat{h}}{\hat{h}^T R_x \hat{h}}. \]  

(A8)

The extra term together with the inaccuracy of the correlation matrix estimate perturb the ideal behavior of the detection statistic described in [6]. The magnitude of this effect depends primarily on the normalized tail energy \( ||\hat{h}||/||\hat{h}|| \) and to a lesser extent on the statistics of \( x \).

REFERENCES


Jun H. Cho (M’99) was born in Seoul, Korea, on September 1, 1970. He received the B.S. degree in control and instrumentation engineering from Seoul National University in 1993, the M.E.S. degree from the University of New South Wales, Sydney, Australia, in 1995, and the Ph.D. degree from the University of Pennsylvania, Philadelphia, in 1998, both in electrical engineering.

During his Ph.D. program (from June 1997 to August 1997), he worked on acoustic echo cancellation at Bell Laboratories, Lucent Technologies, Murray Hill, NJ. He is now with Aware, Inc., Bedford, MA, where he is engaged in research and development of digital subscriber line technologies. His research interests include speech and audio processing, wavelet analysis, radar target identification, and broadband communication systems.

Dennis R. Morgan (S’63–S’68–M’69–SM’92) was born in Cincinnati, OH, on February 19, 1942. He received the B.S. degree in 1965 from the University of Cincinnati and the M.S. and Ph.D. degrees from Syracuse University, Syracuse, NY, in 1968 and 1970, respectively, all in electrical engineering.

From 1965 to 1984, he was with the General Electric Company, Electronics Laboratory, Syracuse, NY, specializing in the analysis and design of signal processing systems used in radar, sonar, and communications. He is now a Distinguished Member of Technical Staff at Bell Laboratories, Lucent Technologies (formerly AT&T), Murray Hill, NJ, where he has been employed since 1984. From 1984 to 1990, he was with the Special Systems Analysis Department, Whippany, NJ, where he was involved in the analysis and development of advanced signal processing techniques associated with communications, array processing, detection and estimation, and adaptive systems; since 1990, he has been with the Acoustics Research Department, Murray Hill, where he is engaged in research on adaptive signal processing techniques applied to electroacoustic systems. He has authored numerous journal publications and is co-author of Active Noise Control Systems: Algorithms and DSP Implementations (New York: Wiley, 1996).

Dr. Morgan has served as Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING since 1995.

Jacob Benesty (M’98) was born in Marrakesh, Morocco, on April 8, 1963. He received the M.S. degree in microwaves from Pierre and Marie Curie University, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, France, in April 1991.

During his Ph.D. program (from November 1989 to April 1991), he worked on adaptive filters and fast algorithms at the Centre National d’Etudes des Telecommunications (CNET), Paris, France. From January 1994 to July 1995, he worked at Telecom Paris on multichannel adaptive filters and acoustic echo cancellation. He then joined Bell Labs, Lucent Technologies (formerly AT&T), Murray Hill, NJ, in October 1995, first as a Consultant and then as Member of Technical Staff. Since then, he has been working on stereophonic acoustic echo cancellation, adaptive filters, source localization, robust network echo cancellation, and blind deconvolution.