A New Class of Doubletalk Detectors Based on Cross-Correlation

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

***Abstract*—A doubletalk detector (DTD) is used with an echo canceler to sense when far-end speech is corrupted by near-end speech. Its role is to freeze the adaptation of the model filter when near-end speech is present in order to avoid divergence of the adaptive algorithm. Several authors have proposed to use the cross-correlation coefficient vector between the input signal vector x and the scalar output y for a DTD. We show in this paper that this measure is not appropriate and propose a modified form that meets, in an optimal way, the needs for an efficient DTD. By extension, we also propose a definition of the normalized cross-correlation matrix between two vectors and show a link with the coherence function.**

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

I. Introduction

AN ECHO canceler [1] removes echo due to echo path h (see Fig. 1), where h represents coupling between a loudspeaker and microphone for the case of an acoustic echo canceler or the hybrid mismatch in the case of a network echo canceler. A doubletalk detector (DTD) [1] is used with an echo canceler to sense when far-end speech is corrupted by near-end speech. The role of this important function is to freeze adaptation of the model filter $\hat{h}$ when near-end speech v is present, in order to avoid divergence of the adaptive algorithm. The far-end talker signal x is filtered with the impulse response h and the resulting signal (the echo) is added to the near-end speech signal v to give the corrupted signal

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

where

$$\mathbf{h} = [h_0 \quad h_1 \quad \cdots \quad h_{L-1}]^T,$$

$$\mathbf{x}(n) = [x(n) \quad x(n-1) \quad \cdots \quad x(n-L+1)]^T,$$

and L is the length of the echo path. We define the error signal at time n as

$$e(n) = y(n) - \hat{\mathbf{h}}^T\mathbf{x}(n). \tag{2}$$

Manuscript received March 30, 1998; revised March 12, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Kahrs.

J. Benesty and D. R. Morgan are with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974-0636 USA (e-mail: jbenesty@bell-labs.com; drrm@bell-labs.com).

J. H. Cho was with the Electrical Engineering Department, University of Pennsylvania, Philadelphia, PA 19131 USA. He is now with Aware, Inc., Bedford, MA 01730 USA (e-mail: jhcho@aware.com).

Publisher Item Identifier S 1063-6676(00)01723-5.

This error signal is used in the adaptive algorithm to adapt the L taps of the filter $\hat{h}$.

For simplicity, we have assumed here that the length of the signal vector $\mathbf{x}$ is the same as the effective length of the echo path h. In reality, the length of h is infinite, thereby resulting in an unmodeled "tail" for any finite value of L. This effect will be discussed later in Section IV.

When v is not present, with any adaptive algorithm, $\hat{h}$ will quickly converge to an estimate of h and this is the best way to cancel the echo. When x is not present, or very small, adaptation is halted by the nature of the adaptive algorithm. When both x and v are present, the near-end talker signal could disrupt the adaptation of $\hat{h}$ and cause divergence. So, the goal of an effective doubletalk detection algorithm is to stop the adaptation of $\hat{h}$ as fast as possible when the level of v becomes appreciable in relation to the level of x, and to keep the adaptation going when the level of v is negligible.

Ye and Wu [2] proposed to use the cross-correlation coefficient vector between $\mathbf{x}$ and e as a means for doubletalk detection. A similar idea using the cross-correlation coefficient vector between $\mathbf{x}$ and y has proven more robust and reliable [3], [4]. Accordingly, we will limit this study to the cross-correlation coefficient vector between $\mathbf{x}$ and y which is defined as

$$\begin{aligned}\mathbf{c}_{xy}^{(1)} &= \frac{E\{\mathbf{x}(n)y(n)\}}{\sqrt{E\{x^2(n)\}E\{y^2(n)\}}} \\ &= \frac{\mathbf{r}_{xy}}{\sigma_x\sigma_y} \\ &= [c_{xy,0}^{(1)} \quad c_{xy,1}^{(1)} \quad \cdots \quad c_{xy,L-1}^{(1)}]^T\end{aligned} \tag{3}$$

where $E\{\cdot\}$ denotes mathematical expectation and $c_{xy,i}^{(1)}$ is the cross-correlation coefficient between $x(n-i)$ and $y(n)$. (We discuss estimation of these quantities for a practical detector in Section IV.)

The idea here is to compare

$$\begin{aligned}\xi^{(1)} &= ||\mathbf{c}_{xy}^{(1)}||_\infty \\ &= \max_i |c_{xy,i}^{(1)}|, \qquad i = 0, 1, \cdots, L-1\end{aligned} \tag{4}$$

to a threshold level T. The decision rule will be very simple: if $\xi^{(1)} \geq T$, then doubletalk is not present; if $\xi^{(1)} < T$, then doubletalk is present.

Although the l_∞ norm used in (4) is perhaps the most natural, other scalar metrics, e.g., l_1, l_2, could alternatively be used to assess the cross-correlation coefficient vectors. However, there is a fundamental problem here which is not linked to the type of metric used. The problem is that these cross-correlation coefficient vectors are not well normalized. Indeed, we can only say

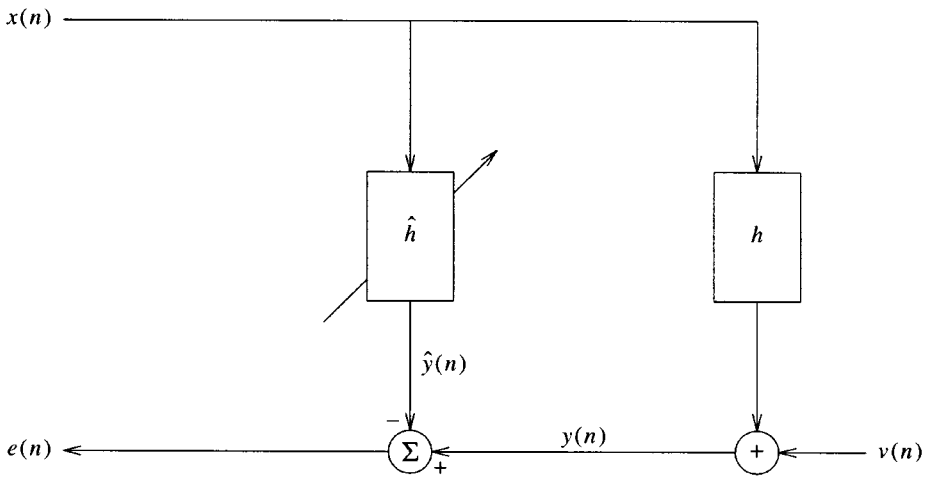


Fig. 1. Block diagram of generic echo canceler.

in general that $\xi^{(1)} \leq 1$. Thus if $v = 0$, that does not imply that $\xi^{(1)} = 1$ or any other known value. We do not know the value of $\xi^{(1)}$ in general. The amount of correlation will depend a great deal on the statistics of the signals and of the echo path. As a result, the best value of T will vary a lot from one experiment to another. So there is no "natural" threshold level associated with the variable $\xi^{(1)}$ when $v = 0$.

An "optimum" decision variable ξ for doubletalk detection will behave as follows:

1) if $v = 0$ (doubletalk is not present), $\xi \geq T$;
2) if $v \neq 0$ (doubletalk is present), $\xi < T$.

The threshold T must be a constant, independent of the data. Moreover, ξ must be insensitive to echo path variations when $v = 0$. In the following we derive a new decision variable that exhibits this behavior. To do this, we present a new way of normalizing the cross-correlation vector between $\mathbf{x}$ and y.

II. A Normalized Cross-Correlation Vector

We now derive in a simple way a new normalized cross-correlation vector between a vector $\mathbf{x}$ and a scalar y. Suppose that $v = 0$. In this case

$$\sigma_y^2 = \mathbf{h}^T \mathbf{R}_{xx} \mathbf{h} \tag{5}$$

where $\mathbf{R}_{xx} = E\{\mathbf{x}(n)\mathbf{x}^T(n)\}$. Since $y(n) = \mathbf{h}^T\mathbf{x}(n)$, we have

$$\mathbf{r}_{xy} = \mathbf{R}_{xx}\mathbf{h} \tag{6}$$

and (5) can be re-written as

$$\sigma_y^2 = \mathbf{r}_{xy}^T \mathbf{R}_{xx}^{-1} \mathbf{r}_{xy}. \tag{7}$$

Now, in general for $v \neq 0$,

$$\sigma_y^2 = \mathbf{r}_{xy}^T \mathbf{R}_{xx}^{-1} \mathbf{r}_{xy} + \sigma_v^2. \tag{8}$$

If we divide (7) by σ_y^2 and take the square root, we obtain the new decision variable

$$\begin{aligned} \xi^{(2)} &= \sqrt{\mathbf{r}_{xy}^T (\sigma_y^2 \mathbf{R}_{xx})^{-1} \mathbf{r}_{xy}} \\ &= \|\mathbf{c}_{xy}^{(2)}\|_2 \end{aligned} \tag{9}$$

where

$$\mathbf{c}_{xy}^{(2)} = (\sigma_y^2 \mathbf{R}_{xx})^{-1/2} \mathbf{r}_{xy} \tag{10}$$

is what we will call the normalized cross-correlation vector between $\mathbf{x}$ and y.

Substituting (6) and (8) into (9), we show that the decision variable is

$$\xi^{(2)} = \frac{\sqrt{\mathbf{h}^T \mathbf{R}_{xx} \mathbf{h}}}{\sqrt{\mathbf{h}^T \mathbf{R}_{xx} \mathbf{h} + \sigma_v^2}}. \tag{11}$$

We easily deduce from (11) that for $v = 0$, $\xi^{(2)} = 1$ and for $v \neq 0$, $\xi^{(2)} < 1$. Note also that $\xi^{(2)}$ is not sensitive to changes of the echo path when $v = 0$. Moreover, a fast version of this algorithm can be derived by recursively updating $\mathbf{R}_{xx}^{-1}\mathbf{r}_{xy}$ using the Kalman gain $\mathbf{R}_{xx}^{-1}\mathbf{x}$ [5].

1) Particular Case: x is white Gaussian noise. For this kind of signal, the autocorrelation matrix is diagonal: $\mathbf{R}_{xx} = \sigma_x^2 \mathbf{I}$. Then (10) becomes

$$\begin{aligned} \mathbf{c}_{xy}^{(2)} &= \frac{\mathbf{r}_{xy}}{\sigma_x \sigma_y} \\ &= \mathbf{c}_{xy}^{(1)}. \end{aligned} \tag{12}$$

Note that, in general, what we are doing in (9) is equivalent to prewhitening the signal $\mathbf{x}$, which is one of many known *generalized cross-correlation* techniques [6]. Thus, when $\mathbf{x}$ is white, no prewhitening is necessary and $\mathbf{c}_{xy}^{(2)} = \mathbf{c}_{xy}^{(1)}$. This suggests a more practical implementation, whereby matrix operations are replaced by an adaptive prewhitening filter [7].

III. A Normalized Cross-Correlation Matrix

Everyone is familiar with the cross-correlation coefficient between two scalars x and y; we have given a new normalized cross-correlation vector between a vector $\mathbf{x}$ and a scalar y, and we now propose to extend this definition to the cross-correlation between two vectors $\mathbf{x}$ and $\mathbf{y}$. We define the normalized cross-correlation matrix $\mathbf{C}_{xy}$ between two vectors $\mathbf{x}$ and $\mathbf{y}$ as follows:

$$\mathbf{C}_{xy} = \mathbf{R}_{xx}^{-1/2} \mathbf{R}_{xy} \mathbf{R}_{yy}^{-1/2} \tag{13}$$

where

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

is a vector of size N. There are two interesting cases:

1) $N = 1$, $\mathbf{C}_{xy} = \mathbf{c}_{xy}^{(2)}$ (normalized cross-correlation vector between $\mathbf{x}$ and y).
2) $N = L = 1$, $\mathbf{C}_{xy} = c_{xy,0}^{(1)}$ (cross-correlation coefficient between x and y).

By extension to (9), we then form the detection statistic

$$\xi^{(3)} = \frac{1}{\sqrt{N}}\|\mathbf{C}_{xy}\|_E = \frac{1}{\sqrt{N}}\sqrt{\text{tr}(\mathbf{C}_{xy}^T\mathbf{C}_{xy})}. \tag{14}$$

We note that for case 1), $\xi^{(3)} = \xi^{(2)}$ as before. Again, we can interpret this formulation as a generalized cross-correlation, where now both $\mathbf{x}$ and $\mathbf{y}$ are prewhitened, which is also known as the *smoothed coherence transform* (SCOT) [6].

We now show that there is a link between the normalized cross-correlation matrix and the coherence. Suppose that $N = L \to \infty$. In this case, a Toeplitz matrix is asymptotically equivalent to a circulant matrix if its elements are absolutely summable [8], which is the case for the intended application. Hence we can decompose $\mathbf{R}_{ab}$ as

$$\mathbf{R}_{ab} = \mathbf{F}^{-1}\mathbf{S}_{ab}\mathbf{F} \tag{15}$$

where $\mathbf{F}$ is the discrete Fourier transform (DFT) matrix and

$$\mathbf{S}_{ab} = \text{diag}\{S_{ab}(0), S_{ab}(1), \cdots, S_{ab}(L-1)\} \tag{16}$$

is a diagonal matrix formed by the first column of $\mathbf{F}\mathbf{R}_{ab}$, and

$$\begin{aligned} S_{ab}(k) &= \sum_{m=-\infty}^{+\infty} E\{a(n)b(n-m)\}e^{-i2\pi km/L} \\ &= \sum_{m=-\infty}^{+\infty} R_{ab}(m)e^{-i2\pi km/L} \end{aligned} \tag{17}$$

is the DFT cross-power spectrum. Now

$$\begin{aligned} \text{tr}(\mathbf{C}_{xy}^T\mathbf{C}_{xy}) &= \text{tr}(\mathbf{R}_{yy}^{-1/2}\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1/2}) \\ &= \text{tr}(\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1}) \end{aligned} \tag{18}$$

since $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$. Using (15), we easily find that

$$\begin{aligned} \text{tr}(\mathbf{C}_{xy}^T\mathbf{C}_{xy}) &= \text{tr}(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1}) \\ &= \sum_{k=0}^{L-1} |\gamma_{xy}(k)|^2 \end{aligned} \tag{19}$$

where

$$\gamma_{xy}(k) = \frac{S_{xy}(k)}{\sqrt{S_{xx}(k)S_{yy}(k)}} \tag{20}$$

is the discrete coherence function. Thus, asymptotically we have

$$\begin{aligned} \xi^{(3)} &\approx \sqrt{\frac{1}{L}\sum_{k=0}^{L-1} |\gamma_{xy}(k)|^2} \\ &= \sqrt{\frac{1}{L}\sum_{k=0}^{L-1} \frac{|H(k)|^2}{|H(k)|^2 + \kappa(k)}} \end{aligned} \tag{21}$$

where $H(k)$ is the transfer function of h and

$$\kappa(k) = \frac{S_{vv}(k)}{S_{xx}(k)} \geq 0 \tag{22}$$

is the near-end talker to far-end talker spectral ratio at frequency k. Except for an unrestricted frequency range, this form is identical to the coherence-based doubletalk detector proposed by Gänsler [9]. (Because all frequencies are not equally important, it is generally advantageous to limit the frequency range in (21) or, more generally, apply weighting over frequency.) This idea seems to be very appropriate since when $v = 0$, the two signals x and y are completely coherent and then $|\gamma_{xy}(k)| = 1, \forall k$, and $\xi^{(3)} \approx 1$; when $v \neq 0$, $|\gamma_{xy}(k)| < 1, \forall k$, and $\xi^{(3)} < 1$.

IV. Practical Considerations and Simulations

Up until now, we have formulated the double-talk decision variables in terms of the various auto-correlation and cross-correlation signal statistics, taking those as a given. However, in practice, we have to estimate these quantities in real time from the only available signals that we have, namely $x(n)$ and $y(n)$.

Estimation of auto-correlation and cross-correlation signal statistics necessarily involves averaging over a suitable time interval, and that then becomes a key problem because of the inevitable tradeoff between response time and accuracy. Response time is crucial for double-talk detection, so we would like to minimize it. On the other hand, if we try to make the response time too fast, insufficient smoothing of the statistical estimates will lead to unreliable performance.

The usual procedure to derive estimates of statistical quantities like $\mathbf{r}_{xy}$ and $\mathbf{R}_{xx}$ is to form a running average of the signal products over a window that moves with time. The length of the window, i.e., the number of samples that form the running average, then determines the response time of the estimate, which is intended to be not too long. Thus, for example, we have

$$\hat{\mathbf{r}}_{xy}(n) = \sum_{m=0}^{M-1} \mathbf{x}(n-m)y(n-m) \tag{23}$$

which averages over M samples.

It is possible that one could sidestep the estimation of certain quantities involved in the decision variables by substituting estimates that have been derived for other purposes. For example, from (6) we know that $\mathbf{R}_{xx}^{-1}\mathbf{r}_{xy} = \mathbf{h}$. Therefore, in (9), we could substitute $\hat{\mathbf{h}}$ for $\mathbf{R}_{xx}^{-1}\mathbf{r}_{xy}$, where $\hat{\mathbf{h}}$ is copied from the echo canceler adaptive filter. This will perturb the ideal performance of the normalized cross-correlation DTD even when the filter is converged, due to the unmodeled "tail" of h [4]. However, the computational advantage of avoiding matrix inversion (or the calculation of the Kalman gain for the fast version) makes the substitution attractive for a practical implementation.

We now briefly introduce some simulation results for our proposed method using detection statistic $\xi^{(2)}$ defined by (9), and compare these with the conventional cross-correlation method using $\xi^{(1)}$ defined by (4). The doubletalk detector performance is characterized in terms of the probability of miss (P_{m}) as a function of near-end to far-end speech ratio (NFR, σ_v/σ_x) under a probability of false alarm (P_{f}) constraint [4]. The miss probability is the portion of the doubletalk interval during which the

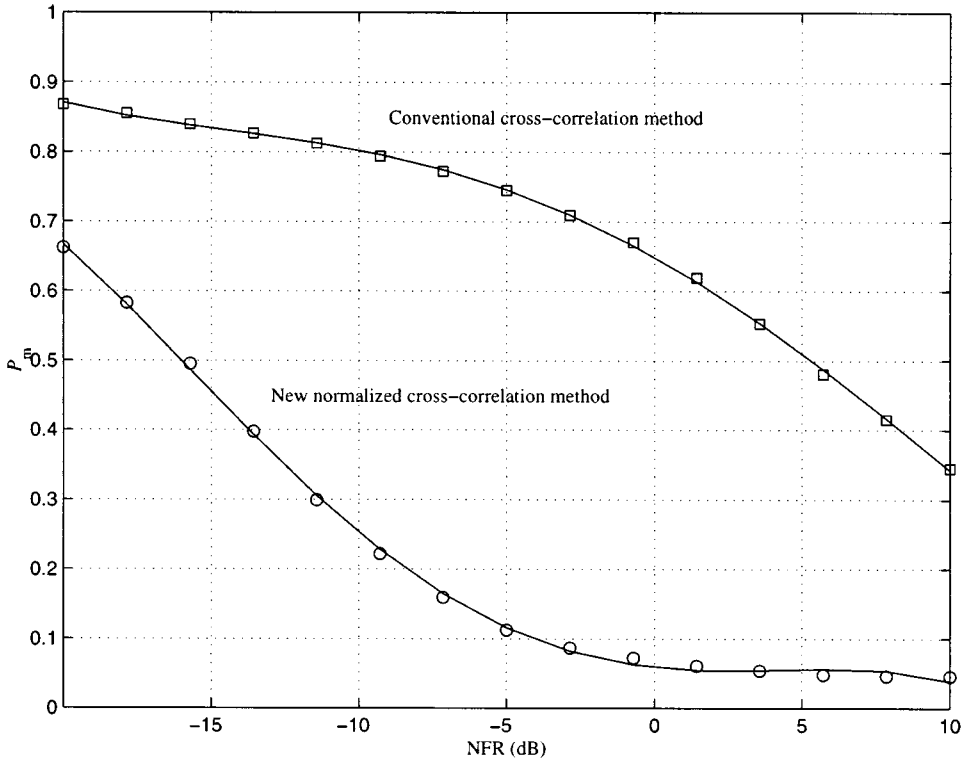


Fig. 2. Probability of miss ($P_{\rm m}$) as a function of the near-end speech to far-end speech ratio (NFR) for doubletalk detectors using new normalized cross-correlation ($\xi^{(2)}$) and conventional cross-correlation ($\xi^{(1)}$); probability of false alarm, $P_{\rm f} = 0.1$.

double talk detector fails to detect the presence of the near-end speech. Therefore, a smaller value of $P_{\rm m}$ indicates better performance of the doubletalk detector. The complete DTD evaluation technique is summarized as follows (see [4] for further details):

1) Set $v = 0$
    a) Select threshold T.
    b) Compute $P_{\rm f}$.
    c) Repeat steps a,b over a range of threshold values.
    d) Select threshold value that corresponds to $P_{\rm f} = 0.1$.
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 $P_{\rm m}$.
    d) Repeat steps a,b,c over all sixteen conditions.
    e) Average $P_{\rm m}$ over all sixteen conditions.
3) Repeat step 2 over a range of NFR values.
4) Plot average $P_{\rm m}$ as a function of NFR.

We used recorded digital speech sampled at 8 kHz for x and v and a measured $L = 2048$-sample (256 ms) room impulse response for h. For $\xi^{(2)}$, we have substituted $\hat{\mathbf{h}} \approx \mathbf{R}_{xx}^{-1}\mathbf{r}_{xy}$ in (9) and have used (23) to estimate $\mathbf{r}_{xy}$ over a window of $M = 500$ samples. For $\xi^{(1)}$, we also estimated $\mathbf{r}_{xy}$ over a window of $M = 500$ samples. The $P_{\rm m}$ characteristics of these two methods under the constraint $P_{\rm f} = 0.1$ are shown in Fig. 2. It is clear that the new normalized cross-correlation method proposed here shows significantly better performance over the full range of NFR.

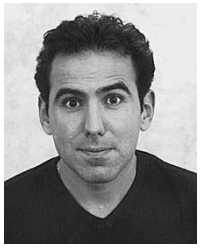
V. Conclusion

We have proposed a new normalized cross-correlation vector for doubletalk detection and have shown that the conventional cross-correlation coefficient vector is just an approximation of this. Simulations demonstrate the superiority of this new technique. We have also generalized this concept to a normalized cross-correlation matrix and have shown a relationship to the coherence technique proposed by Gänsler. While Gänsler's method may result in further improvement, it comes at a much higher price of computational complexity. We have instead concentrated on computationally-simpler methods. For the simplified form of the normalized cross-correlation DTD, it is assumed that the AEC has already converged so that $\hat{\mathbf{h}}$ well approximates $\mathbf{R}_{xx}^{-1}\mathbf{r}_{xy}$. However, some degradation would be expected in a dynamic situation where doubletalk occurs while the AEC is adapting. Further work is necessary to assess this problem and propose remedies.

References

[1] M. M. Sondhi, "An adaptive echo canceler," *Bell Syst. Tech. J.*, vol. 46, pp. 497–510, Mar. 1967.
[2] H. Ye and B.-X. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Trans. Commun.*, vol. 39, pp. 1542–1545, Nov. 1991.
[3] R. D. Wesel, "Cross-correlation vectors and doubletalk control for echo cancellation," unpublished.
[4] 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.
[5] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1991, ch. 16.
[6] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 320–327, Aug. 1976.

[7] J. R. Zeidler, "Performance analysis of LMS adaptive prediction filters," *Proc. IEEE*, vol. 78, pp. 1781–1806, Dec. 1990.
[8] R. M. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 725–730, Nov. 1972.
[9] T. Gänsler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Trans. Commun.*, vol. 44, pp. 1421–1427, Nov. 1996.



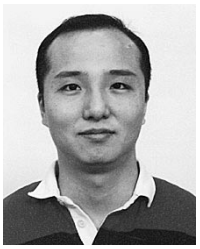
Jacob Benesty (M'98) was born in Marrakesh, Morocco, on April 8, 1963. He received the M.S. degree in microwaves from Pierre & Marie Curie University, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, France, in April 1991.

While pursuing the Ph.D. degree, 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 was with Telecom Paris, working on multichannel adaptive filters and acoustic echo cancellation. He joined Bell Labs, Lucent Technologies (formerly AT&T) in October 1995, first as a Consultant and then as Member of Technical Staff. He has been working on stereophonic acoustic echo cancellation, adaptive filters, source localization, robust network echo cancellation, and blind deconvolution.

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, Cincinnati, OH, 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 Electronics Laboratory, General Electric Company, Syracuse, NY, specializing in the analysis and design of signal processing systems used in radar, sonar, and communications. He is now Distinguished Member of Technical Staff at Bell Laboratories, Lucent Technologies (formerly AT&T), Murray Hill, NJ, where he has been 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, where he is engaged in research on adaptive signal processing techniques applied to electroacoustic systems. He has authored numerous journal publications and is coauthor 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.



Jun H. Cho 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.

While pursuing the Ph.D. degree, 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 (DSL) technologies. His research interests include speech and audio processing, wavelet analysis, radar target identification, and broadband communication systems.