Summer 2007

Simple and efficient solutions to the problems associated with acoustic echo cancellation

Asif Iqbal Mohammad

Department: Electrical and Computer Engineering

SIMPLE AND EFFICIENT SOLUTIONS TO THE PROBLEMS ASSOCIATED WITH ACOUSTIC ECHO CANCELLATION

by

ASIF IQBAL MOHAMMAD

A DISSERTATION

Presented to the Faculty of the Graduate School of the

UNIVERSITY OF MISSOURI–ROLLA

in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

ELECTRICAL ENGINEERING

2007

Dr. Steven L. Grant, Advisor	Dr. Kurt Kosbar
Dr. Randy H. Moss	Dr. Levent Acar

Dr. Ilene H. Morgan

ABSTRACT

This dissertation is a collection of papers that addresses several important problems associated with acoustic/line echo cancellation (AEC/LEC), specifically double-talk and echo-path change detection. A double-talk detector is used to freeze AEC filter's adaptation during periods of near-end speech. This dissertation presents three different novel double-talk detection schemes. Simulations demonstrate the efficiency of the proposed algorithms. The novel normalized cross-correlation based double-talk detector proposed in Section 3, outperforms the best existing algorithms and is computationally of the order of magnitude simpler. Next, this novel double-talk detector is extended to the frequency domain adaptive algorithms and the proposed technique is also generalized for the multi-channel case.

Echo-path variations in acoustic case are common. In general, a detector's increased sensitivity towards double-talk also increases its probability of falsely declaring echo-path changes as double-talk. This adversely affects the performance of the acoustic echo canceller (AEC) as the filter coefficients are frozen precisely when they should be adapting. To remedy this, an efficient explicit echo-path change (EPC) statistic is derived to help differentiate between echo-path variations and double-talk. The combination of the new double-talk and echo-path change statistics yield an effective low-complexity solution to the AEC adaptation control problem.

The key to a good two-path method performance lies in the definitions of the download tests. A novel download test is also proposed that improves the overall performance of the system.

ACKNOWLEDGMENTS

I would like the thank my committee for their help and instructions over the last several years. The classes I have taken with you provided me with the necessary technical background to pursue my work and provided motivation for many of the problems I have considered.

I would especially like to thank my advisor, Dr. Steven L. Grant, for his help and accepting me as his Ph.D. student in midway of my Ph.D. programme. Our research meetings were always enjoyable and great learning experiences. I must admit, I learned a lot from him both technically and morally. I hope we can continue our research together and develop breakthrough technology.

A special thanks to my first advisor Dr. William Weeks. Courses with you laid the foundation of my knowledge, you have been an excellent mentor. I am really thankful to God for providing me with so many excellent mentors in my career. I would also like to thank Dr. Kurt Kosbar, for being a motivating factor and some excellent advice.

I would like to give very special thanks to my parents for their consistent and unconditional support throughout my life. Their encouragement and dedication to me and their prayers have eased all the burdens of my life. I would like to dedicate this work to my father Mr. Mohammad Iqbal and my mother Ms. Ameerunisa.

Finally, I would like to thank my brothers and sister for all the encouraging discussions. I am really fortunate to have their unconditional and consistent support. I would like to thank God for so many wonderful blessings.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. DISSERTATION OVERVIEW

This dissertation examines the problems associated with acoustic/line echo cancellation. A brief summary of each section is presented here and a list of main contributions can be found in Section 8. Main contribution of this work is towards the double-talk detection and in detecting variations in the echo-path(s).

1.1. DOUBLE-TALK DETECTION USING SPEECH DETECTORS

A new system for double-talk detection that uses multiple signal detectors / discriminators based on recurrent networks is presented in this section. The goal was to build a simple system that learns to combine information from different signal sources to make robust decisions even under changing noise conditions. Three detectors are used - two of these are frequency domain signal detectors, one at the far-end and one at the microphone channel. The third detector determines the relative level of near-end speech vs. far-end echo in the microphone signal. The new double-talk detector combines information from all these detectors to make its decision. An important part of this proposed design is that the features used by these detectors can be easily tracked online in the presence of noise. Results were compared with other cross-correlation based double-talk detectors to show its effectiveness. Portions of this work were published in the Proceedings of the 2006 International Workshop on Acoustics, Echo and Noise Control, Paris, France [2].

1.2. A NEW CLASS OF DOUBLE-TALK DETECTORS

Two different double-talk detection schemes for acoustic echo cancellation (AEC) are presented here. First, a novel normalized detection statistic based on the cross-correlation coefficient between the microphone signal and the cancellation error is introduced. The decision statistic is designed in such a way that it meets the needs of an optimal double-talk detector. It is also shown that the proposed detection statistic converges to the recently proposed normalized cross-correlation based double-talk detector [3], the best known cross-correlation based detector. Next, a new hybrid double-talk detection scheme based on a cross-correlation coefficient and a near-end signal detector is formulated. The proposed algorithm not only detects double-talk but also tracks any echo-path variations efficiently. Results were compared with other cross-correlation based double-talk detectors to show their effectiveness. Portions of this work were published in the Proceedings of the 2007 IEEE International Conference on Multi-media and Expo, Beijing, China [4].

1.3. A FREQUENCY DOMAIN DOUBLE-TALK DETECTOR

Most teleconferencing conversations are conducted in the presence of acoustic echoes. Typically an adaptive filter is used to cancel the echo, with a control device called the double-talk detector which controls the adaptation. A novel test statistic for the double-talk detection based on the cross-correlation between the microphone signal and the cancellation error for the frequency domain adaptive algorithm is derived. The main advantage of the proposed algorithm is its simplicity and computational efficiency. Results are compared with the normalized cross-correlation based double-talk detector proposed in [3]. The idea of the proposed double-talk detector (single-channel) is also generalized to the matrix case (multi-channel). Portions of this work were submitted to the European Signal Processing Journal, Elsevier [5].

1.4. A NOVEL NORMALIZED ECHO-PATH CHANGE DETECTOR

A double-talk detector is used to freeze acoustic echo canceller's (AEC) filter adaptation during periods of near-end speech. Increased sensitivity towards double-talk results in declaring echo-path changes as double-talk which adversely effects the performance of an AEC as adaptation is frozen when it really needs to be on. Thus, one needs an efficient and simple echo-path change detector so as to differentiate any echo-path variations from double-talk condition. A novel test statistic for echo-path change detection is derived. The proposed decision statistic detects any echo-path variations, is normalized properly and is computationally very efficient as compared to existing techniques. Simulations demonstrate the efficiency of the proposed algorithm. Portions of this work were published in the Proceedings of the 2007 IEEE Region-5 Conference, Fayettville, Arkansas [6].

1.5. SIMPLE AND EFFICIENT SOLUTIONS TO AEC PROBLEMS

With rare exceptions, teleconferencing conversations are conducted in the presence of acoustic echoes. Typically, an adaptive filter is used to remove the echo created by the loudspeaker-microphone environment. When the near-end talker is active or when speech comes from both the far-end and near-end simultaneously, identification of the echo-path becomes problematic because the adaptive filter coefficients diverge from the true echo-path if adaptation is continued. To avoid this problem, a double-talk detector is used to inhibit the filter's adaptation during periods of near-end speech. Some of the most successful detectors use the cross-correlation between the far-end signal ($\mathbf{x}(n)$) and the microphone output ($m(n)$) as the basis for a decision statistic; others have used the cross-correlation between ($\mathbf{x}(n)$) and the cancellation error ($e(n)$), where n is the time index. In this section, a novel double-talk detection algorithm based on the cross-correlation between ($m(n)$) and ($e(n)$) is proposed.

The resulting algorithm has the same performance as the most effective known techniques but with an order of magnitude decrease in computational complexity.

In general, a detector's increased sensitivity towards double-talk also increases it's probability of falsely declaring echo-path changes as double-talk. This adversely effects the performance of the acoustic echo canceller (AEC) as the filter coefficients are frozen precisely when they should be adapting. To remedy this, the addition of an efficient explicit echo-path change (EPC) statistic is proposed to help differentiate between echo-path variations and double-talk. The combination of the new double-talk and echo-path change statistics yield an effective low-complexity solution to the AEC adaptation control problem. Portions of this work were submitted to the journal of IEEE transactions on Speech and Audio Processing [7].

1.6. A NOVEL DOWNLOAD TEST FOR TWO PATH ECHO CANCELLER

The two-path technique is an algorithm for acoustic / line echo cancellation (AEC/LEC) based on two sets of parallel filters, background and foreground, that predict the echo. The key to a good two-path method performance lies in the definitions of the download tests. In this section, a novel download test for a two-path approach is proposed. The new download test is a good measure of the adaptive filter's convergence, and is computationally efficient. With the aid of the proposed download test, significant improvement in the overall performance of the system was observed. Portions of this work were submitted to the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics in Mohonk, New York [8].

2. DOUBLE-TALK DETECTION USING SPEECH DETECTORS

2.1. INTRODUCTION

Acoustic echo cancellers (AEC) are an important part of teleconferencing systems, they are necessary to mitigate the deleterious effect of acoustic feedback from the speaker signal to the microphone input [9]. In an AEC, the echo path is adaptively modelled using a filter, which is then used to synthesize a replica of the echo and subtract it from the echo-corrupted microphone signal [10]. When the near-end talker is active, or when there is no far-end signal, the filter coefficients will diverge from the true echo path impulse response; hence, it is crucial to have a good *double-talk detector* which indicates periods of simultaneous far-end and near-end speech. During these periods, the adaptation of the filter coefficients is stopped [9].

Double-talk detection can use statistics computed from both the microphone and the far-end signal. Typically, a cross-correlation based statistic is used in these scenarios [3]. In addition, some statistics based on each individual signal may also be computed which can assist in the detection. In this section a machine learning based approach is proposed.

In this new approach, multiple speech detectors / discriminators (D/D) at various points are used, and then combined for effective double-talk detection. The system is modular in nature, so it is extendable to multi-channel scenarios. But in this section the idea on a system with a single microphone channel is demonstrated. In this system, three different D/D units are used. Two of them are signal detectors and are used to detect the presence of a signal at the far-end (FESD) and at the near-end (NESD) as shown in Figure 2.1. At the near-end, the signal can be due to near-end speech or due to echo from the far-end talker. Thus, a third unit is needed, which is a discriminator, it estimates the relative influence of far-end echo vs. the near-end speech in the microphone signal. For lack of a better term, this third unit is labelled as a "signal discriminator" (SD). The final part of the double-talk detector combines the output of all these units to make robust decision regarding double-talk. Since the detectors have to be robust to changing noise conditions, SNR dependent features which have been shown to be effective for speech detection [1] are used, and can be easily tracked online in the presence of noise.

This section is structured as follows: In Section 2.2, the proposed method for signal detectors/discriminators and for double-talk detection is presented. In Section 2.3, the experiments and results are discussed which is followed by a summary and conclusion in Section 2.4.

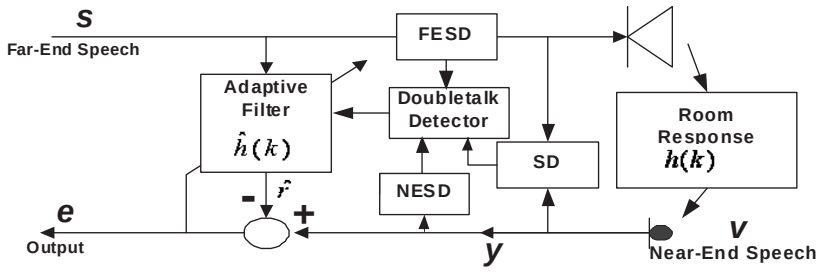

Figure 2.1: An AEC system showing various modules of proposed double-talk detector.

2.2. SIGNAL DETECTORS/DISCRIMINATORS

One of the primary goals is to make the overall system have low complexity, this requires that the D/D units themselves to be very simple. Recently logistic [1] networks were shown to be very simple and effective for speech detection even in changing noise conditions. This idea can be easily carried over to detecting other types of signals in noise.

In acoustic application, all the signals are influenced by reverberation, whose effect typically lasts for hundreds of milliseconds; further speech itself is a highly correlated signal. Hence, it is important that the detectors incorporate this long-term effect in them automatically. One way to achieve this is to take multiple frames of data (spanning the desired time-length of interest) and use them as inputs to the network. One problem with this approach is that the correct number to include will depend upon the situation, and will have to be determined by trial and error. This also makes the network more complex. Another option is to use past decisions rather than features. *Recurrent networks* [10] are excellent examples of systems that achieve this - they dynamically re-use information about the state of the network from the past (these typically constitute the previous outputs of the network) as inputs to the current decision.

Combining the above two ideas, a single layer network with recurrent feedback (shown in Figure 2.2) is used. The state space model of the proposed system can be written as:

$$x(n) = (1-\alpha)(\sum_{i=1}^{N} w_i u_i) + \alpha x(n-1) \tag{2.1}$$

$$y(n) = \frac{1}{1+\exp(-x(n))} \tag{2.2}$$

where $[u_1(n)u_2(n)\ldots u_{N-1}(n)1]$ is the current input data and w_is and α are the parameters of the system. $y(n)$ is a value between 0 and 1, and, hence, can be interpreted as a probability.

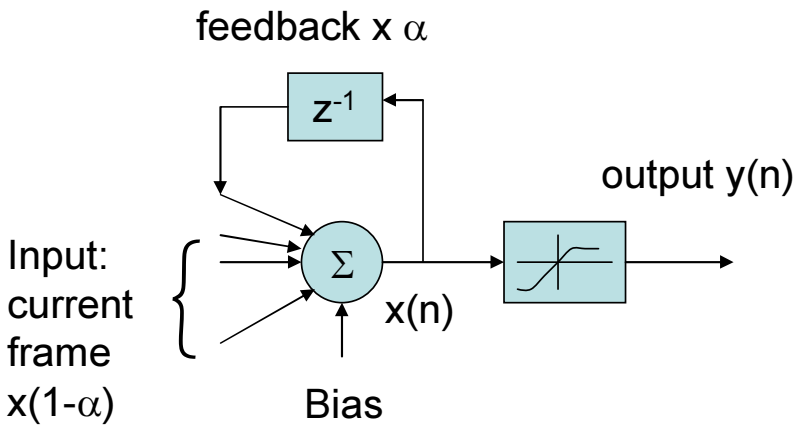

Figure 2.2: Recurrent network architecture.

Since the input features are time-dependent and arrive one per time-segment, it is appropriate to train this network continuously in on-line fashion after every frame of data arrives. This type of learning is appropriate for a non-stationary signal like speech, and is called *real-time recurrent learning* (RTRL) [11]. RTRL uses stochastic gradient descent to train this network to minimize the cross-entropy error [12]. This error metric makes the network discriminative, and provides the maximum likelihood estimate of the class probability for a wide variety of class conditional densities of the data [12]. The reason this is useful for us is that, since the outputs represent probabilities, it is easy for us to make decisions based on them, or combine their decisions with others.

2.3. FEATURE DESIGN

One of the desired characteristics of any detector is that its features are sufficiently simple, easy to calculate, have discriminatory power and work well under changing noise conditions. Estimated posterior SNR $\chi(k,t)$ are used as the feature set for the NESD and FESD (these have been shown to have all the above desirable properties [1]). $\chi(k,t)$ is the ratio of the energy in a given time-frequency atom S to the noise energy N $\chi(k,t) = \frac{|S(k,t)|^2}{N(k,t)}$ where k,t are the frequency bin and time indices respectively. The FESD uses the speaker signal S as the target signal, and the NESD uses the microphone signal Y. The short term spectra of speech are well modelled by log-normal distributions; hence the logarithm of the SNR estimate is used rather than the SNR estimate itself. Thus, the inputs used are:

$$\chi_{FESD}(k,t) = \{\log|S(k,t)|^2 - \log N_{FE}(k,t)\} \tag{2.3}$$

and

$$\chi_{NESD}(k,t) = \{\log|Y(k,t)|^2 - \log N_{NE}(k,t)\} \tag{2.4}$$

where N_{FE} and N_{NE} are the noise energies in frequency bin k and time-frame t at the far-end and near-end respectively. The noise power N can be tracked using various algorithms such as [13],[14]. Here a minima tracker is used (for each frequency bin look back a few frames e.g. 25, and choose the lowest value of the signal) followed by smoothing, to track the noise floor [14].

The features for the speech discriminator (SD) are described next. SD is trying to look at the microphone signal, and it is trying to figure out how much of it is dominated by the near-end speech (as opposed to the far-end echo). Thus, it is trying to discriminate the level of near-end speech. Thus, for this system, the logarithm of the ratio of the microphone instantaneous power Y to the far-end instantaneous power S for each frequency bin per frame is used as the feature i.e.

$$\chi_{SD}(k,t) = \log|Y(k,t)|^2 - \log|S(k,t)|^2. \tag{2.5}$$

As can be seen in Figure 2.3, the extracted features are clearly distinct for different scenarios. As expected, the extracted features are typically largest for only the near-end speech, smallest for the echo-only case, and in between for the case of double-talk. Different feature levels

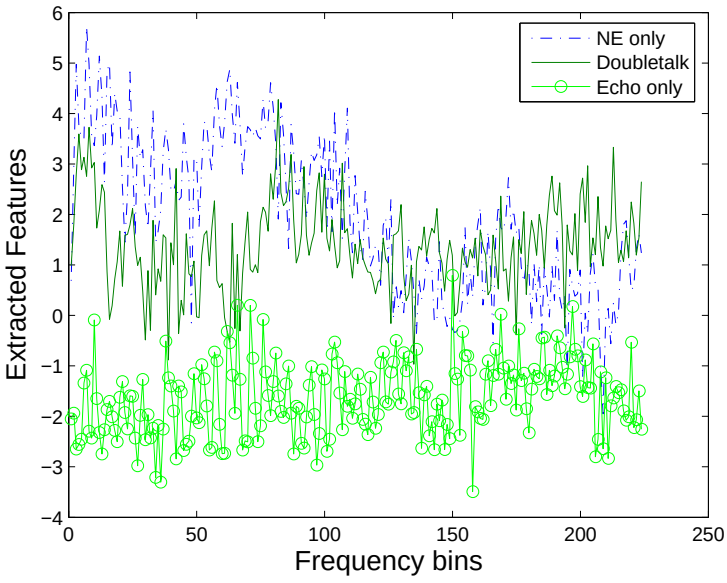

Figure 2.3: Extracted features for the SD.

correspond to different probability levels; larger features correspond to higher probabilities. For the echo-only case, the extracted features are always low and independent of the echo-path; hence the discriminator performance is relatively independent of the echo-path. This has been empirically verified under a wide variety of situations. The decision from this discriminator is combined with decisions from NESD and FESD for double-talk detection. It is probably best to build another learner which combines all these three decisions into one. In this work, a simple approach is used (as outlined below).

When the NESD and the SD of Figure 2.1 both indicate a high probability of the presence of speech, above the selected threshold, the presence of near-end speech is confirmed. If the FESD of Figure 2.1 indicates the presence of speech and a confirmed near-end talker, then the current-frame of the captured signal is declared to be a double-talk. In short, double-talk is declared when all the three detectors indicate the presence of speech.

2.4. EXPERIMENTS AND RESULTS

The well known AURORA database [15] is used for experiments. The recorded digital speech is sampled at 16 KHz and is used for the far-end speech s and the near-end speech v of Figure 2.1. Room impulse response of a $10' \times 10' \times 8'$ room is measured using a stereo system; the truncated 8000 sample (500 ms) room response is used as the loudspeaker-microphone environment h in Figure 2.1. A subset of the Aurora data base was used for training the FESD of Figure 2.1, precisely 75 signals (50000 frames) consisting of a mixture of male and female speakers. These signals were filtered through the left channel of the measured room impulse response to create the echo part of the microphone signals; near-end speech signals (different signals taken from the Aurora database) were added to simulate the microphone signals for training the NESD and the SD of Figure 2.1. Near-end speech was added at different near-end to far-end ratios to improve training.

For testing, a completely different set of 120 signals taken from the Aurora data-base [15] were used to simulate the far-end speech. These signals were filtered using the right channel of the measured room impulse response to simulate a different channel for testing. To these artificially created echo signals, near-end speech is added from a second different set of 120 signals taken from Aurora data-base at 12 different near-end to far-end ratios (NFR).

The true labels on the speech signals were generated by thresholding the energy in each time frame of the clean data; the threshold was selected so that all the speech events were retained, which was verified by listening to a small fraction of the training data. To study the performance of the speech detectors, the ROC curve is plotted (correct detection of speech versus false alarm). As can be observed from Figure 2.4, results are compatible with the

speech detector of [1], which was trained with a 8 KHz sampled speech. This confirms that the training is done appropriately for the FESD.

The presence of near-end speech is confirmed when both the NESD and the SD indicate presence of speech. The combined ROC curve for the NESD and SD is shown in Figure 2.5 at different values of NFR. At a false alarm rate of 0.1, the near-end speech is detected with a detection probability of 0.89 at 0 dB NFR; as expected the near-end speech is detected with a lower detection rate of 0.7 at -10.5 dB NFR. The axes are truncated to highlight the upper left quadrant of the plot. Thresholds corresponding to $P_f = 0.1$ (probability of false alarm = 0.1) were obtained by following [3]:

1. Set $v = 0$ (No near-end speech).
2. Select thresholds for all the speech detectors.
3. Compute P_f.
4. Repeat steps 2, 3 over a range of threshold values.
5. Select the thresholds that correspond to $P_f = 0.1$.

These thresholds were used to compute the probability of miss, P_m , for the test signals. For the ten signals at each NFR, the average of the P_m over the respective signals is used to calculate the average probability of miss P_m. The new RTRL double-talk detector

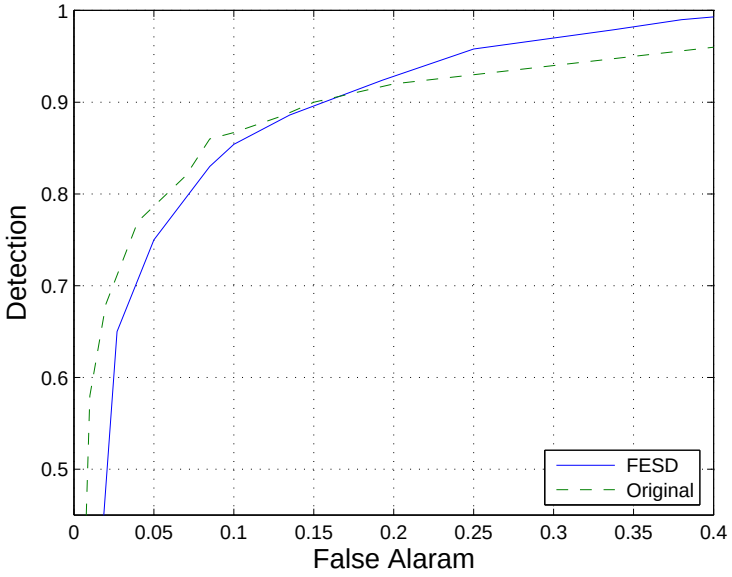

Figure 2.4: ROC curve for the FESD, original curve taken directly from [1].

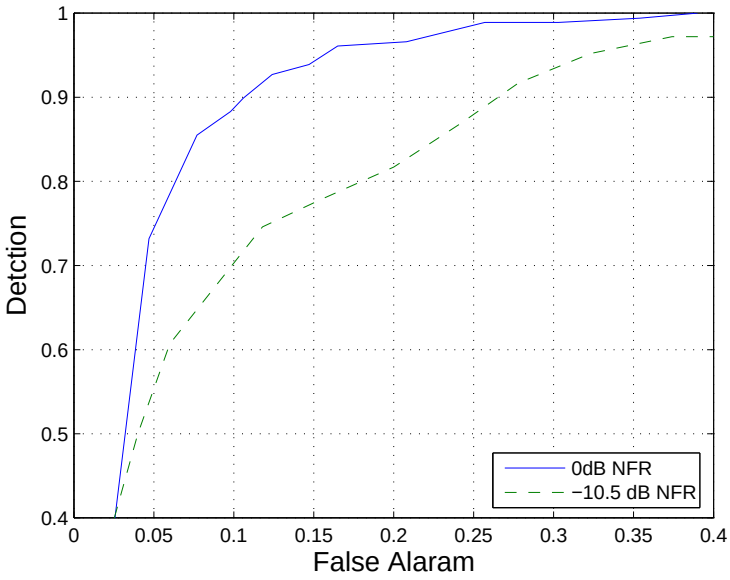

Figure 2.5: ROC curve for detecting NE speech at different NFR.

was evaluated by closely following [16]. Results are compared with the new normalized cross-correlation based detector [3] and the conventional cross-correlation based detector [17]. The P_m characteristics of all three methods under the constraint of $P_f = 0.1$ are shown in Figure 2.6. The RTRL double-talk detector proposed here clearly outperforms the conventional cross-correlation based detector over a full range of NFR. The new algorithm outperforms the normalized cross-correlation based detector for lower values of NFR and is comparable over the remaining region. It must be noted that the RTRL based double-talk detector works with a frame size of 16 ms (256 samples at 16 KHz) whereas the other methods use a frame of size 62.5 ms (500 samples at 8 KHz).

Next a bi-level architecture is implemented by aggregating 4 frames into a single frame so as to have a frame of duration 64 ms comparable to that of the normalized cross-correlation based detector's 62.5 ms. It is observed in Figure 2.6, that the RTRL double-talk detector outperforms the normalized cross-correlation based detector in almost half of the range of NFR values and is very close in the remaining region.

The FESD has a detection rate of 0.88 at 15 dB SNR (Figure 2.4); thus the RTRL based double-talk detector is bounded by a miss probability of 0.1 even at higher NFR values (Figure 2.6). Typically in a teleconferencing device such as the Microsoft RingCam [18], the loudspeaker is located very close to the microphone, and the near-end talkers are relatively further away from the microphone. Thus, low NFR values are prominent in such devices. As

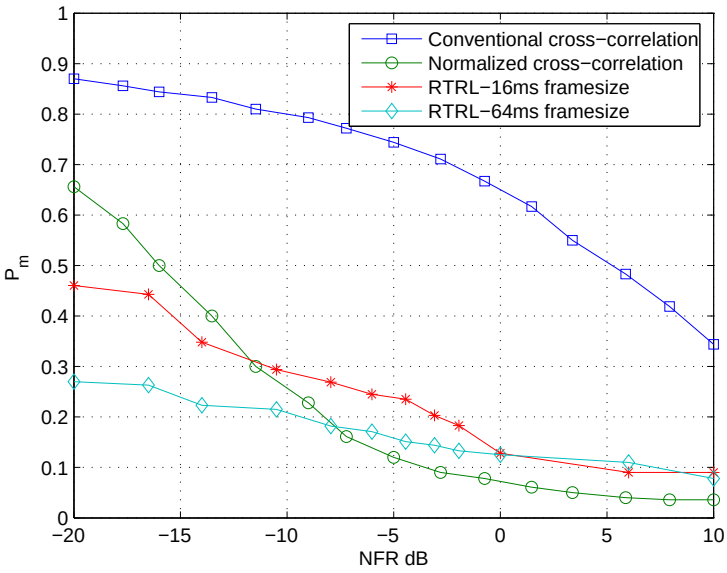

Figure 2.6: P_m as function of NFR for double-talk detectors using RTRL method, normalized cross-correlation based detector and the conventional cross-correlation based detector at $P_f = 0.1$.

can be observed from Figure 2.6, the RTRL based double-talk detector significantly outperforms the normalized cross-correlation based detector over such lower NFR values making it suitable for such applications.

2.5. CONCLUSION

A new double-talk detector based on a novel speech discriminator is proposed; which significantly outperforms the conventional cross-correlation based detector and is comparable to the normalized cross-correlation based detector.

Echo is a delayed speech signal; typically the spectrum of the echo is very similar to the spectrum of a speech signal with a quicker falloff from the maxima. Thus, in the frequency domain, the trained coefficients are equally applicable to any room responses. Similar results were observed for different room responses and even better results were observed with real data collected using the RingCam project at Microsoft Research [18]. Based on these observations it can be concluded that the trained weights are equally applicable to any room responses if not independent of room responses.

3. A NEW CLASS OF DOUBLE-TALK DETECTORS

3.1. INTRODUCTION

Most teleconferencing conversations are conducted in the presence of acoustic echoes [10]; if the delay between the speech and its echo is more than a few tens of milliseconds, the echo is distinctly noticeable. An acoustic echo canceller (AEC) is used to remove the echo created due to the loudspeaker-microphone environment [9]. Echo cancellation is achieved by adaptively synthesizing a replica of the echo and subtracting the result from the echo-corrupted signal [10]. When the near-end talker is active or when the speech comes from both the far-end and near-end, the filter coefficients will diverge from the true echo path impulse response if adaptation is enabled. A double-talk detector is used to stop the AEC's filter adaptation during periods of near-end speech [9].

Double-talk detection plays a very important part in acoustic echo cancellation. A double-talk detection algorithm should be able to detect a double-talk condition quickly and accurately so as to freeze adaptation as soon as possible; at the same time it should be able to track any echo-path changes and should be able to distinguish double-talk from the echo-path variations [17]. To solve this problem, this section presents two different techniques for double-talk detection. An optimum decision variable ξ for double-talk detection should behave as follows [9]:

1. If double-talk is not present i.e. $v = 0$, then $\xi \geq T$.

2. If double-talk is present i.e. $v \neq 0$, then $\xi < T$. The threshold T must be a constant independent of the data and the decision statistic ξ must be insensitive to echo-path variations when $v = 0$.

Figure 3.1 shows the basic structure of the adaptive acoustic echo canceller. The far-end signal $\mathbf{x}$ is filtered through the room impulse response $\mathbf{h}$ to get the echo signal

$$y(n) = \mathbf{h^T x} \tag{3.1}$$

where

$$\begin{aligned} \mathbf{h} &= [h_0 \;\; h_1 \;\; .... \;\; , \;\; h_{L-1}]^T, \\ \mathbf{x} &= [x(n) \;\; x(n-1) \;\; .... \;\; , \;\; x(n-L+1)]^T, \end{aligned}$$

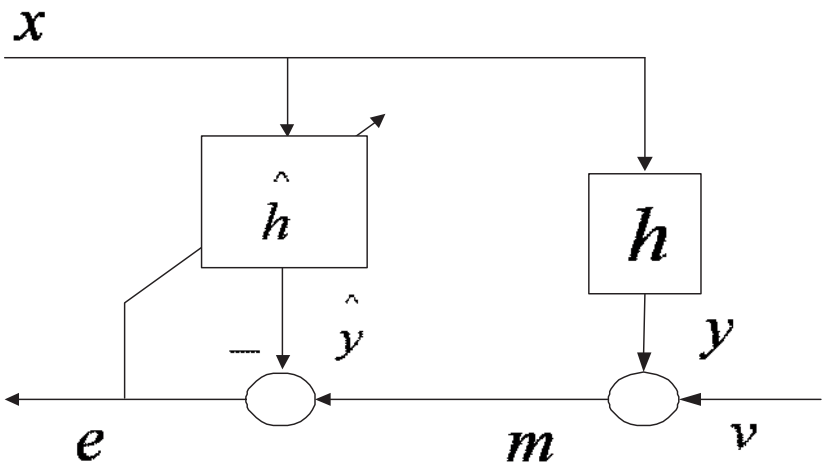

Figure 3.1: Basic AEC model.

and L is the length of the echo-path. This echo signal is added to the near-end speech signal v to get the microphone signal

$$m(n) = y(n) + v(n). \tag{3.2}$$

The error signal at time n is defined as

$$e(n) = m(n) - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x} \tag{3.3}$$

and is used to adapt the L taps of the AEC's adaptive filter $\hat{\mathbf{h}}$.

This section is structured as follows. In Section 3.2, previous double-talk detection algorithms are revisited. In Section 3.3, the novel normalized double-talk detection algorithm is formulated and a link between the proposed algorithm and the one proposed in [3] is derived. The new hybrid double-talk detection scheme is proposed in Section 3.5. Next, a comprehensive study on the proposed algorithms is done in Section 3.6 which is followed by a summary and conclusions in Section 3.7.

3.2. PREVIOUS WORK

Referring to Figure 3.1, Ye and Wu [17] first proposed using the cross-correlation vector between the far-end signal vector $\mathbf{x}$, which is played out of the speakers, and the AEC's cancellation error e, $\mathbf{r}_{e\mathbf{x}} = E[e\mathbf{x}^T]$, as the basis for double-talk detection. In this section, this algorithm is referred as XECC. Simulation results by Benesty [3] have shown that this approach does not work well for detecting double-talk, and a theoretical derivation provides further insight. Noting that the near-end speech v is independent of the far-end signal $\mathbf{x}$ and assuming all of the signals are zero mean, the cross-correlation between the AEC's error signal

and the speaker signal is given by:

$$\begin{aligned} \mathbf{r}_{e\mathbf{x}} &= E[(y + v - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x})\mathbf{x}^{\mathbf{T}}] \\ &= E[(\mathbf{h}^{\mathbf{T}}\mathbf{x} - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x})\mathbf{x}^{\mathbf{T}}] \\ &= (\mathbf{h}^{\mathbf{T}} - \hat{\mathbf{h}}^{\mathbf{T}})R_{\mathbf{xx}} \end{aligned} \tag{3.4}$$

where $E[\bullet]$ denotes the mathematical expectation and $R_{\mathbf{xx}} = E[\mathbf{x}\mathbf{x}^{\mathbf{T}}]$. Clearly from equation 3.4 it is observed that $\mathbf{r}_{e\mathbf{x}}$ is high only when there is a change in the echo-path; hence this approach is more suitable for tracking echo-path variations rather than detecting double-talk.

More recently, Benesty, et al. [3] [16] proposed a double-talk detection algorithm based on the cross-correlation between the far-end signal vector $\mathbf{x}$ and the microphone signal scalar m, $\mathbf{r}_{\mathbf{x}m} = E[\mathbf{x}m]$, which is referred as XMCC in this section. Benesty's decision statistic used to detect double-talk in [3] is given by

$$\xi_{XMCC} = \sqrt{\mathbf{r}_{\mathbf{x}m}^T(\sigma_m^2 R_{\mathbf{xx}})^{-1}\mathbf{r}_{\mathbf{x}m}} \tag{3.5}$$

where $R_{\mathbf{xx}}$ is as defined earlier and the variance of the microphone signal (σ_m^2) is

$$\begin{aligned} \sigma_m^2 &= E[m^2] \\ &= E[(y + v)^2] \\ &= E[y^2] + E[v^2] \\ &= E[\mathbf{h}^{\mathbf{T}}\mathbf{x}(\mathbf{h}^{\mathbf{T}}\mathbf{x})^T] + \sigma_v^2 \\ &= \mathbf{h}^{\mathbf{T}}R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2. \end{aligned} \tag{3.6}$$

where σ_v^2 is the near-end speech power.

3.3. NORMALIZED DOUBLE-TALK DETECTION

Instead of using $\mathbf{r}_{e\mathbf{x}}$ or $\mathbf{r}_{\mathbf{x}m}$ as discussed in Section 3.2, the new statistic uses the cross-correlation between the cancellation error e and the microphone signal m, $r_{em} = E[em]$, as the basis for double-talk detection. This algorithm will be called MECC in this section.

$$\begin{aligned} r_{em} &= E[(y + v - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x})(y + v)] \\ &= E[(\mathbf{h}^{\mathbf{T}}\mathbf{x} - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x} + v)(\mathbf{h}^{\mathbf{T}}\mathbf{x} + v)^T] \\ &= E[(\mathbf{h}^{\mathbf{T}}\mathbf{x} - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x})\mathbf{x}^{\mathbf{T}}\mathbf{h} + v^2] \\ &= (\mathbf{h}^{\mathbf{T}} - \hat{\mathbf{h}}^{\mathbf{T}})R_{\mathbf{x},\mathbf{x}}\mathbf{h} + \sigma_v^2 \end{aligned} \tag{3.7}$$

The new normalized decision statistic is defined as:

$$\xi_{MECC} = 1 - \frac{r_{em}}{\sigma_m^2}. \tag{3.8}$$

Substituting equations 3.6 and 3.7 in 3.8 yields:

$$\begin{aligned} \xi_{MECC} &= 1 - \frac{(\mathbf{h^T} - \hat{\mathbf{h}}^\mathbf{T})R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2}{\mathbf{h^T}R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2} \\ &= \frac{\hat{\mathbf{h}}^\mathbf{T}R_{\mathbf{xx}}\mathbf{h}}{\mathbf{h^T}R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2}. \end{aligned} \tag{3.9}$$

It can be observed from equation 3.9, that for $v = 0$, $\xi_{MECC} \approx 1$ and for $v \neq 0$, $\xi_{MECC} < 1$. Thus, the proposed detection statistic meets the needs of an optimal double-talk detector.

The values for r_{em} and σ_m^2 in 3.8 are exact and not available in practice. As a result, the final decision statistic is given by:

$$\xi_{MECC} = 1 - \frac{\hat{r}_{em}}{\hat{\sigma}_m^2} \tag{3.10}$$

which is based on the estimates $\hat{r}_{em}[n]$ and $\hat{\sigma}_m^2[n]$. The estimates are found using the exponential recursive weighting algorithm, [19] [20]:

$$\hat{r}_{em}[n] = \lambda\hat{r}_{em}[n-1] + (1-\lambda)e[n]m[n] \tag{3.11}$$

$$\hat{\sigma}_m^2[n] = \lambda\hat{\sigma}_m^2[n-1] + (1-\lambda)m^2[n] \tag{3.12}$$

where $e[n]$ is the captured cancellation error sample at time n, $m[n]$ is the captured microphone signal sample at time n, and λ is the exponential weighting factor. If

$$\xi_{MECC} < T \tag{3.13}$$

it is concluded that the captured sample of the microphone signal is corrupted by the near-end speech and the adaptation of the AEC's adaptive filter(s) is frozen. Otherwise, adaptation continues.

In addition to its simplicity, the main advantage of the proposed detection statistic is that only the maximum cross-correlation needs to be computed instead of computing the entire cross-correlation vector required by the other algorithms. This results in significant computational savings as compared to the other algorithms; requiring 2 multiplications, 2 additions, 1 subtraction and a division to compute the decision statistic at each sample (i.e. 6

operations per sample), whereas for the Benesty's test statistic $3L+3$ operations are required to compute the detection statistic at each sample where L is the frame size (typically $L \geq 512$).

3.4. RELATIONSHIP BETWEEN NEW AND BENESTY'S TEST STATISTIC

The proposed decision statistic is given by equation 3.10, which theoretically can be rewritten as in equation 3.9, and Benesty's double-talk decision statistic is given in equation 3.5. The decision statistics are different as the former is based on r_{em}, and the latter is based on $\mathbf{r}_{\mathbf{x}m}$. Although the decision statistics are different, they can be shown to result in a similar expression. Substituting $\mathbf{r}_{\mathbf{x}m} = R_{\mathbf{xx}}\mathbf{h}$ and $\sigma_m^2 = \mathbf{h^T}R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2$ in equation 3.5, yields

$$\begin{aligned} \xi_{XMCC}^2 &= \mathbf{h^T}R_{\mathbf{xx}}(\sigma_m^2 R_{\mathbf{xx}})^{-1}R_{\mathbf{xx}}\mathbf{h} \\ &= \frac{\mathbf{h^T}R_{\mathbf{xx}}R_{\mathbf{xx}}^{-1}R_{\mathbf{xx}}\mathbf{h}}{\sigma_m^2} \\ &= \frac{\mathbf{h^T}R_{\mathbf{xx}}\mathbf{h}}{\mathbf{h^T}R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2} \end{aligned} \tag{3.14}$$

and from equation 3.9 they have

$$\xi_{MECC} = \frac{\hat{\mathbf{h}}^{\mathbf{T}}R_{\mathbf{xx}}\mathbf{h}}{\mathbf{h^T}R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2}. \tag{3.15}$$

In addition to the square root, the other difference between the decision statistics is in the numerator; the taps of the AEC filter $\hat{\mathbf{h}}^{\mathbf{T}}$ are used in ξ_{MECC} and the true echo-path impulse response $\mathbf{h^T}$ in ξ_{XMCC}. However, for practical implementation and computational simplicity, the authors in [3] substitute $\hat{\mathbf{h}}^{\mathbf{T}}$ for $\mathbf{h^T}$ resulting in similar decision statistics. Thus the proposed decision statistic is exactly analogous to the Benesty's test statistic, and simulations (Figure 3.2) further demonstrate the convergence. However, the proposed algorithm is significantly simpler and computationally efficient.

3.5. HYBRID DOUBLE-TALK DETECTION BASED ON MECC AND RTRL

In this section, a hybrid double-talk detector based on a cross-correlation measure between the microphone signal and the AEC cancellation error, and the double-talk detection algorithm based on speech detection and discriminator (based on real-time recurrent learning (RTRL) presented in [2]) is formulated.

3.5.1. Cross-correlation Measure. The cross-correlation measure between the cancellation error e and the microphone signal m is used. It can be observed from equation 3.7, cross-correlation is high whenever there is a change in the echo-path and/or when the

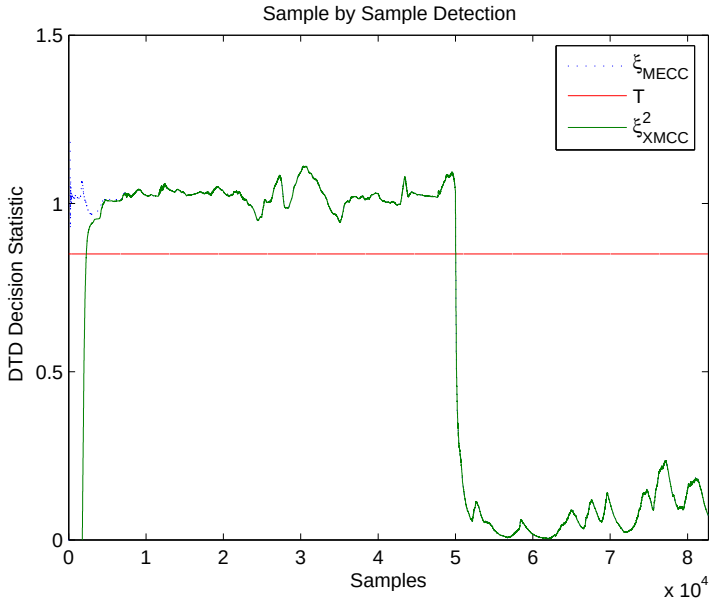

Figure 3.2: Illustrating the convergence of the proposed MECC and the XMCC double-talk detectors.

near-end speech is present. To differentiate the near-end speech from the echo-path variations real time recurrent learned (RTRL) speech detectors are used [2].

An estimated cross-correlation function is used; the estimated cross-correlation function (ECC) which is the maxima of the correlation in a frame, is updated using the exponential recursive weighting algorithm [19] [20]:

$$P_e^2[t] = \lambda P_e^2[t-1] + (1-\lambda)\mathbf{e}[t]\mathbf{e}^T[t] \tag{3.16}$$

$$P_m^2[t] = \lambda P_m^2[t-1] + (1-\lambda)\mathbf{m}[t]\mathbf{m}^T[t] \tag{3.17}$$

$$P_{m,e}[t] = \lambda P_{m,e}[t-1] + (1-\lambda)\mathbf{e}[t]\mathbf{m}^T[t] \tag{3.18}$$

where $\mathbf{e}[t]$ is the captured cancellation error vector in the time frame t and $\mathbf{m}[t]$ is the captured microphone signal vector at the time frame t and λ is the exponential weighting factor. Smaller values of λ provide better tracking capability but worse estimation accuracy. In practice, for slowly time varying signals; $0.9 \leq \lambda \leq 1$ is usually chosen [17]. The estimated cross-correlation function (ECC) is given by:

$$ecc[t] = \frac{P_{m,e}[t]}{P_e[t]P_m[t]}. \tag{3.19}$$

3.5.2. Near-end Speech Detector and Speech Discriminator. Frequency domain logistic discriminative speech detectors are used to detect the presence of speech [1].

The class probability is estimated as

$$P_t = \frac{1}{1 + \exp(-\mathbf{W^T}\chi_\mathbf{t})} \tag{3.20}$$

where P_t is the probability of speech at time frame t, $\mathbf{W^T}$ are the trained weights ($1 \times frequencybins$) and χ_t is a vector of extracted features in each frequency bin at the time frame t. The trained weights $\mathbf{W}^T$ are obtained using Real Time Recurrent Learning [11], these weights are obtained by training off-line. For a detailed discussion on speech detectors and their training process see [2].

Two detectors are used at the microphone to detect the presence of the near-end speech as shown in Figure 3.3. For the microphone signal detector (NESD), the logarithm of the estimated posterior SNR is used as the feature [1]

$$\chi_{NESD}(k,t) = 10\{\log|M(k,t)|^2 - \log N_{NE}(k,t)\} \tag{3.21}$$

where N_{NE} is the noise energy in frequency bin k and time-frame t at the near-end. The noise power N can be tracked using [14]. In this section, a minima tracker is used (for each frequency bin look back a few frames e.g. 25 and choose the lowest value of the signal) followed by smoothing, to track the noise floor [14]. This NESD detector gives the presence of speech at the near-end; which can be the near-end speech or the far-end echo. To differentiate the near-end speech from the far-end echo, a special detector/discriminator SD is used.

To distinguish the near-end speech from the far-end echo, features that differentiate the near-end speech from the far-end echo are required; thus,the logarithm of the ratio of the microphone instantaneous power M to the far-end instantaneous power X is employed as the

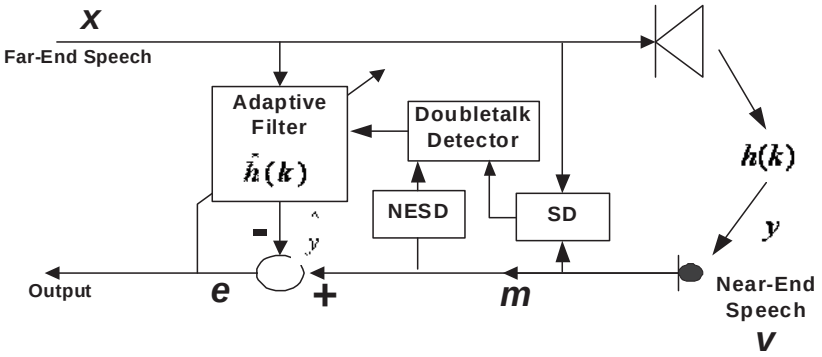

Figure 3.3: Hybrid double-talk detection model.

feature, i.e.

$$\chi_{SD}(k,t) = 10\{\log|M(k,t)|^2 - \log|X(k,t)|^2\}. \tag{3.22}$$

It was observed in [2] that the extracted features are distinct for different scenarios. The extracted features were typically largest for only the near-end speech, smallest for the echo-only case, and in between for the case of double-talk. Different feature levels correspond to different probability levels; larger features correspond to higher probabilities. For the echo-only case, the extracted features were always low independent of the echo-path; hence the special detector/disriminator is independent of the echo-path in the absence of near-end speech.

The presence of the near-end speech is confirmed when both the detectors indicate the presence of speech. Speech detection based double-talk detector [2], when used by itself for double-talk detection does not give superior performance. However, the performance can be improved by combining it with the proposed cross-correlation measure. The hybrid double-talk detector works as follows:

1. When both the detectors indicate a high probability of the presence of speech i.e. $P_{NESD}(t) \geq P_{Threshold_1}$ and $P_{SD}(t) \geq P_{Threshold_2}$ and the estimated cross-correlation $ECC(t) \geq R_{th}$ then it is concluded that the captured frame of the microphone signal is corrupted by the near-end speech.

2. Else, it is concluded that either there is a change in the echo-path or the echo signal is present and adapting the filter taps continues.

The results in Figure 3.4 use the ECC, but using the test statistic 3.8 may perform equal or slightly better than using the ECC.

3.6. EXPERIMENTS AND RESULTS

The performance is characterized in terms of the probability of miss (P_m) as a function of near-end to far-end speech ratio (NFR) under a probability of false alarm (P_f) constraint [16]. The probability of miss (P_m) is the probability of not detecting (miss) double-talk when it is present; therefore a smaller value of P_m indicates better performance. The proposed double-talk detectors are evaluated using [16].

The recorded digital speech sampled at 16 KHz is used as far-end speech $\mathbf{x}$ and near-end speech v and a measured $L = 8000$ sample (500 ms) room impulse response of a $10' \times 10' \times 8'$ room is used as the loudspeaker-microphone environment $\mathbf{h}$. Results are compared with the conventional cross-correlation (XECC) based double-talk detector proposed in [17] and the

RTRL based double-talk detector proposed in [2]. The P_m characteristics of all the four methods under the constraint of $P_f = 0.1$ are shown in Figure 3.4. It is clear that the hybrid and the proposed normalized detection statistic (MECC) significantly outperform the conventional (XECC) double-talk detector over a full-range of NFR values. Also it can be observed that the hybrid double-talk detection scheme outperforms the RTRL based double-talk detector for most of the NFR values. Thus, it can be concluded that the performance of the RTRL based double-talk detector [2] is improved by combining it with the proposed cross-correlation measure.

It should be noted that the performance of the proposed normalized decision statistic (MECC) is similar to the Benesty's test statistic (XMCC) the best known cross-correlation based double-talk detector. However, the detection statistic is computationally of the order of magnitude simpler, the detection threshold ($T \approx 1$) is independent of the data and is insensitive to echo-path variations.

3.7. CONCLUSION

Two different techniques for double-talk detection are proposed. First, the novel normalized decision statistic is introduced, the proposed detection statistic meets the needs of an optimal double-talk detector, is computationally very efficient and converges to the best known cross-correlation based double-talk detector. Next, the hybrid double-talk detection

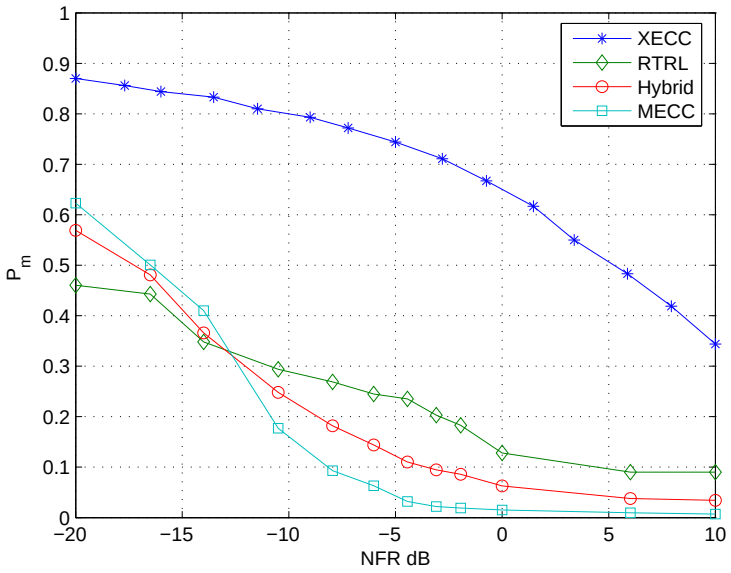

Figure 3.4: P_m as function of NFR for the proposed MECC and the CC-SD and XECC double-talk detectors at $P_f = 0.1$.

scheme is formulated. The hybrid double-talk detector works on a frame by frame basis; the algorithm not only detects double-talk but also tracks any echo-path variations. This is achieved at the cost of increased computational complexity.

4. A FREQUENCY DOMAIN DOUBLE-TALK DETECTOR

4.1. INTRODUCTION

Most teleconferencing conversations are conducted in the presence of acoustic echoes [10]. An acoustic echo canceller (AEC) is used to remove the echo created due to the loudspeaker-microphone environment [9]. Figure 4.1 shows the basic block diagram of an AEC. The far-end signal $\mathbf{x}$ is filtered through the room impulse response $\mathbf{h}$ to get the echo signal

$$y(n) = \mathbf{h^T x} \tag{4.1}$$

where

$$\mathbf{h} = [h_0 \;\; h_1 \;\; .... \;\; , \;\; h_{l-1}]^T,$$

$$\mathbf{x} = [x(n) \;\; x(n-1) \;\; .... \;\; , \;\; x(n-l+1)]^T,$$

and l is the length of the echo-path. The signal picked up by the microphone is denoted by m. Typically, the microphone signal is composed of an echo, the near-end speech v and the surrounding noise w. Hence,

$$m(n) = y(n) + v(n) + w(n) \tag{4.2}$$

The error signal at time n is defined as

$$e(n) = m(n) - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x} \tag{4.3}$$

where $\hat{\mathbf{h}}$ is the adaptive AEC filter tap vector. In an echo canceller, one adaptively synthesizes a replica of the echo and subtracts it from the echo-corrupted signal [10]. When the near-end talker is active or when the speech comes from both the far-end and near-end, identification of the echo-path becomes problematic and the adaptive filter coefficients diverge from the true echo-path. To avoid this problem, a double-talk detector is used to stop the AEC's adaptation during periods of near-end speech [9].

In a double-talk detector, a decision variable ξ is formed from the available signals x, m, and e. This variable is compared to a preset threshold T. An optimum decision variable for double-talk detection should behave as follows [9]:

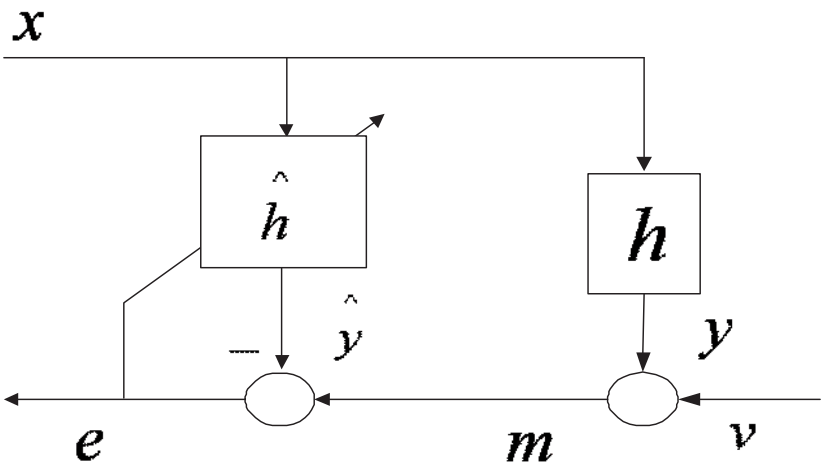

Figure 4.1: Basic AEC model.

1. If $v(n) = 0$ (double-talk is not present) then $\xi \geq T$.

2. If $v(n) \neq 0$ (double-talk is present) then $\xi < T$.

3. ξ is insensitive to echo-path variations.

A decision statistic that meets in an efficient way the needs of an optimal double-talk detector was proposed in [4]. The decision statistic is based on the cross-correlation between the microphone signal and the cancellation error. In this section, a frequency domain computation scheme for this statistic is presented. The frequency domain approach is chosen because of its desirable properties such as low computational complexity, inherent stability and proven performance for the echo cancellation problem [21]. Results are compared with the normalized cross-correlation based double-talk detector proposed in [22], which also meets the needs of an optimal double-talk detector and whose performance was shown to be superior compared to other double-talk detectors based on the cross-correlation coefficient. This algorithm will be referred as the Benesty's algorithm in this section. However, the proposed algorithm is very attractive because of its computational efficiency. The main advantage of the proposed technique is that only the maximum cross-correlation needs to be computed instead of computing the entire cross-correlation vector required by other algorithms. Computational complexity of the proposed double-talk detector is of the order of L ($O(L)$) whereas for the Benesty's double-talk detector proposed in [22] it is of the order of L^2 ($O(L^2)$), where L is the block size. Next, the idea of the proposed double-talk detector is extended to the multi-channel case by defining a global test statistic based on the cross-correlation matrix between the microphone

signals and the cancellation errors that takes into account all the microphone signals. Finally, it is shown that the proposed double-talk detector converges to the Benesty's detection statistic and simulation results verify this convergence.

This section is structured as follows: In Section 4.2, the frequency domain adaptive algorithm is given and the novel double-talk detection statistic is formulated. In Section 4.3, the idea of the proposed double-talk detector is extended to the multi-channel case. Simulation results are discussed in Section 4.4 which is followed by a summary in Section 4.5. For notational convenience through out this section, all the boldface lower case letters correspond to vectors, all the boldface uppercase letter correspond to matrices, under-bars denote frequency domain, and scalars are not boldfaced.

4.2. FREQUENCY DOMAIN ADAPTIVE ALGORITHM

In this section, the basic frequency domain adaptive algorithm is given [23] and the novel test statistic for double-talk detection is introduced. The frequency domain echo canceller model is shown in Figure 4.2.

4.2.1. Frequency Domain Adaptive Algorithm. Here, the frequency-domain adaptive algorithm is briefly described by minimizing an error signal in the frequency-domain [23]. First, the following block signals are defined:

$$\mathbf{e}_{L\times 1} = [e(nL)....,e(nL+L-1)]^T$$

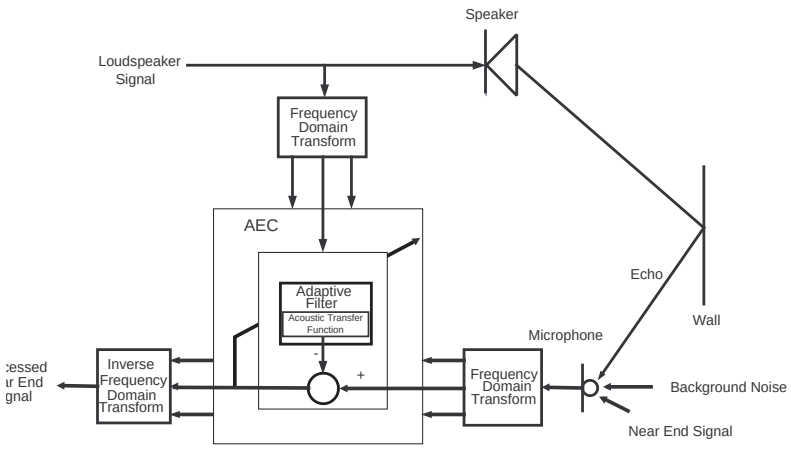

Figure 4.2: Frequency domain basic AEC model.

and

$$\mathbf{m}_{L\times 1} = [m(nL)....,m(nL+L-1)]^T \tag{4.4}$$

where $\mathbf{e}_{L\times 1}$ is the error signal block and $\mathbf{m}_{L\times 1}$ is the microphone signal block at time n. The error signal in the frequency domain is defined by:

$$\underline{\mathbf{e}}(n) = \underline{\mathbf{m}}(n) - \mathbf{G}\mathbf{D}_1(n)\hat{\underline{\mathbf{h}}}\hat{\underline{\mathbf{h}}} \tag{4.5}$$

where under bars denote the frequency domain,

$$\mathbf{D}_1(n) = diag\{\mathbf{F}[x[nL-L]....x[nL+L-1]]^T\} \tag{4.6}$$

and

$$\mathbf{G} = \mathbf{F}\mathbf{W}\mathbf{F}^{-1} \tag{4.7}$$

where

$$\mathbf{W} = \begin{bmatrix} 0_{L\times 1}0_{L\times 1} \\ 0_{L\times 1}I_{L\times 1} \end{bmatrix} \tag{4.8}$$

and $\mathbf{F}$ is the Fourier matrix. The error signal in the frequency domain is obtained by multiplying the corresponding time domain equation with the Fourier matrix $\mathbf{F}$ (of size $2L \times 2L$) whose entries are given by

$$F_{j,k} = \exp\{-2\pi ijk/2L\} \tag{4.9}$$

for $j,k = 0,1,2,...,2L-1$, $i = \sqrt{-1}$ and L is the block size i.e.

$$\underline{\mathbf{e}}(n) = \mathbf{F}[0_{L\times 1}\mathbf{e}_{L\times 1}]^T$$

$$\underline{\mathbf{m}}(n) = \mathbf{F}[0_{L\times 1}\mathbf{m}_{L\times 1}]^T \tag{4.10}$$

Minimizing the error signal in the frequency domain yields the following update equations for the frequency domain adaptive algorithm:

$$\mathbf{S}_1(n) = \lambda\mathbf{S}_1(n-1) + (1-\lambda)\mathbf{D}_1^H(n)\mathbf{D}_1(n) \tag{4.11}$$

$$\hat{\underline{\mathbf{h}}}(n) = \hat{\underline{\mathbf{h}}}(n-1) + 2(1-\lambda)\mathbf{S}_1^{-1}(n)\mathbf{D}_1^H(n)\underline{\mathbf{e}}(n) \tag{4.12}$$

where $\mathbf{S}_1$ is a diagonal matrix and $2(1-\lambda)$ is a positive number. This algorithm is similar to the unconstrained frequency-domain adaptive algorithm proposed by Mansour and Gray in [24]. This approach is computationally very attractive since $\mathbf{S}_1$ is diagonal.

4.2.2. Double-talk Detection in Frequency Domain. In this section, the novel test statistic for double-talk detection in the frequency domain analogous to the time-domain method proposed in [4] is introduced.

The variance of the microphone signal in the frequency domain is given by $\sigma_m^2 = E[\underline{\mathbf{m}}^{\mathbf{H}}\underline{\mathbf{m}}]$ where $\underline{\mathbf{m}} = \mathbf{G}\mathbf{D}_1(n)\underline{\mathbf{h}} + \underline{\mathbf{v}}$. Now

$$\begin{aligned} \underline{\mathbf{m}}^{\mathbf{H}}\underline{\mathbf{m}} &= (\mathbf{G}\mathbf{D}_1(n)\underline{\mathbf{h}} + \underline{\mathbf{v}})^H(\mathbf{G}\mathbf{D}_1(n)\underline{\mathbf{h}} + \underline{\mathbf{v}}) \\ &= (\underline{\mathbf{h}}^{\mathbf{H}}\mathbf{D}_1(n)\mathbf{G} + \underline{\mathbf{v}}^{\mathbf{H}})(\mathbf{G}\mathbf{D}_1(n)\underline{\mathbf{h}} + \underline{\mathbf{v}}) \end{aligned} \tag{4.13}$$

Now, since near-end speech is independent of the far-end signal:

$$\begin{aligned} \sigma_m^2 &= E[\underline{\mathbf{m}}^{\mathbf{H}}\underline{\mathbf{m}}] \\ &= \underline{\mathbf{h}}^{\mathbf{H}}E[\mathbf{D}_1(n)\mathbf{G}\mathbf{D}_1(n)]\underline{\mathbf{h}} + E[\underline{\mathbf{v}}^{\mathbf{H}}\underline{\mathbf{v}}] \\ &= \underline{\mathbf{h}}^{\mathbf{H}}\mathbf{S}\underline{\mathbf{h}} + \sigma_v^2 \end{aligned} \tag{4.14}$$

where $\mathbf{S} = E[\mathbf{D}_1(n)\mathbf{G}\mathbf{D}_1(n)]$ is the spectral matrix of the far-end signal and σ_v^2 is the variance of the near-end signal.

Next, the cross-spectral coefficient S_{em} (maximum cross-correlation) between the microphone signal and the cancellation error is computed.

$$\begin{aligned} S_{em} &= E[\underline{\mathbf{e}}^{\mathbf{H}}\underline{\mathbf{m}}] = E[(\underline{\mathbf{m}} - \mathbf{G}\mathbf{D}_1(n)\hat{\underline{\mathbf{h}}})^H\underline{\mathbf{m}}] \\ &= E[\underline{\mathbf{m}}^{\mathbf{H}}\underline{\mathbf{m}}] - E[\hat{\underline{\mathbf{h}}}^{\mathbf{H}}\mathbf{D}_1(n)\mathbf{G}\underline{\mathbf{m}}] \\ &= \sigma_m^2 - E[\hat{\underline{\mathbf{h}}}^{\mathbf{H}}\mathbf{D}_1(n)\mathbf{G}(\mathbf{G}\mathbf{D}_1(n)\underline{\mathbf{h}} + \underline{\mathbf{v}})] \\ &= \sigma_m^2 - \hat{\underline{\mathbf{h}}}^{\mathbf{H}}E[\mathbf{D}_1(n)\mathbf{G}\mathbf{D}_1(n)]\underline{\mathbf{h}} \\ &= \sigma_m^2 - \hat{\underline{\mathbf{h}}}^{\mathbf{H}}\mathbf{S}\underline{\mathbf{h}} \end{aligned} \tag{4.15}$$

Substituting equation 4.14 yields,

$$S_{em} = (\underline{\mathbf{h}}^{\mathbf{H}} - \hat{\underline{\mathbf{h}}}^{\mathbf{H}})\mathbf{S}\underline{\mathbf{h}} + \sigma_v^2 \tag{4.16}$$

The new decision variable is defined as:

$$\xi_{Asif} = 1 - \frac{S_{em}}{\sigma_m^2} \tag{4.17}$$

Substituting equations 4.14 and 4.16 yields:

$$\begin{aligned}\xi_{Asif} &= 1 - \frac{(\underline{\mathbf{h}}^{\mathbf{H}} - \underline{\hat{\mathbf{h}}}^{\mathbf{H}})\mathbf{S}\underline{\mathbf{h}} + \sigma_v^2}{\underline{\mathbf{h}}^{\mathbf{H}}\mathbf{S}\underline{\mathbf{h}} + \sigma_v^2} \\ &= \frac{\underline{\hat{\mathbf{h}}}^{\mathbf{H}}\mathbf{S}\underline{\mathbf{h}}}{\underline{\mathbf{h}}^{\mathbf{H}}\mathbf{S}\underline{\mathbf{h}} + \sigma_v^2} \end{aligned} \tag{4.18}$$

It can be observed from equation 4.18, that for $v = 0$, $\xi_{Asif} \approx 1$ and for $v \neq 0$, $\xi_{Asif} < 1$. The values of S_{em} and σ_m^2 in (4.17) are exact and are not available in practice. As a result, the final decision statistic is given by:

$$\xi_{Asif} = 1 - \frac{\hat{S}_{em}}{\hat{\sigma}_m^2} \tag{4.19}$$

which is based on the estimates $\hat{S}_{em}$ and $\hat{\sigma}_m^2$. The estimates are found using the estimated cross-correlation function, which is the maxima of the correlation in a frame and is updated using the exponential recursive weighting algorithm, [19] [20]:

$$\begin{aligned}\hat{\sigma}_m^2(n) &= \lambda_1\hat{\sigma}_m^2(n-1) + (1-\lambda_1)\underline{\mathbf{m}}^{\mathbf{H}}(n)\underline{\mathbf{m}}(n) \\ \hat{S}_{em}(n) &= \lambda_1\hat{S}_{em}(n-1) + (1-\lambda_1)\underline{\mathbf{e}}^{\mathbf{H}}(n)\underline{\mathbf{m}}(n) \end{aligned} \tag{4.20}$$

In practice, for slowly time varying signals, $0.9 \leq \lambda_1 \leq 1$ is usually chosen [17]. If

$$\xi_{Asif} < T \tag{4.21}$$

it can be concluded that the captured frame of the microphone signal is corrupted by the near-end speech and adaptation of the AEC's adaptive filter(s) is frozen. Otherwise, adaptation continues.

4.3. EXTENSION TO MULTI-CHANNEL CASE

In this section, the idea of the proposed double-talk detector is extended to the multi-channel case. Assuming that there are Q loudspeakers and P microphones, the acoustic echo cancellation problem now consists of identifying Q echo paths at each microphone i.e. in total PQ echo paths need to be estimated. However, as far as double-talk detection is concerned, it is better to have a global test statistic that takes into account the information of all the microphone signals [25]. Selecting a single microphone signal and using a single test statistic based on this signal is not enough since the near-end speech is picked up by various microphones at different amplitude levels. Also, using P independent decision variables will

be computationally expensive. Thus, a global test statistic is developed by looking at the cross-correlation matrix between the microphone signals and the cancellation errors.

An error signal vector for all the microphones is defined by:

$$\mathbf{e}_{P\times 1}(n) = \mathbf{m}_{P\times 1}(n) - \hat{\mathbf{H}}^T\mathbf{X}(n) \tag{4.22}$$

where

$$\mathbf{e}_{P\times 1}(n) = [e_1(n), ..., e_P(n)]^T$$

and

$$\mathbf{m}_{P\times 1}(n) = [m_1(n), ..., m_P(n)]^T \tag{4.23}$$

where $e_p(n)$ and $m_p(n)$ are the cancellation error and the signal collected at the p^{th} microphone respectively at time n. Furthermore,

$$\hat{\mathbf{H}} = \begin{bmatrix} \hat{\mathbf{h}}_{1,1} & \hat{\mathbf{h}}_{1,2} & ... & \hat{\mathbf{h}}_{1,P} \\ .. & .. & ... & .. \\ .. & .. & ... & .. \\ \hat{\mathbf{h}}_{Q,1} & \hat{\mathbf{h}}_{Q,2} & ... & \hat{\mathbf{h}}_{Q,P} \end{bmatrix}$$

is the estimated echo-path channel matrix of size $LQ \times P$, where $\hat{\mathbf{h}}_{q,p} = [\hat{h}_{qp,0},\ \hat{h}_{qp,1}\ ...,\ \hat{h}_{qp,L-1}]^T$ is an estimate of the echo-path from the q^{th} loudspeaker to the p^{th} microphone (L taps) and the far-end signal matrix $\mathbf{X}(n)$ is given by

$$\mathbf{X}(n) = [\mathbf{x}_1^T(n),\ \ ...,\ \ \mathbf{x}_Q^T(n)]^T_{LQ\times 1} \tag{4.24}$$

where $\mathbf{x}_q(n) = [x_q(n),\ \ ....,\ \ x_q(n-L+1)]^T$ is the q^{th} loudspeaker signal. In addition, the block error matrix is defined as:

$$\mathbf{E}_{L\times P} = \begin{bmatrix} e_1(nL) & e_2(nL) & .. & e_P(nL) \\ e_1(nL+1) & e_2(nL+1) & .. & e_P(nL+1) \\ .. & .. & .. & .. \\ e_1(nL+L-1) & e_2(nL+L-1) & .. & e_P(nL+L-1) \end{bmatrix}$$

Similarly, $\mathbf{M}_{L\times P}$ is defined as the block microphone signal matrix. The corresponding frequency domain matrices are obtained by multiplying with the Fourier matrix $\mathbf{F}$ i.e.

$$\underline{\mathbf{E}}(n) = \mathbf{F}\begin{bmatrix} \mathbf{0}_{L\times P} \\ \mathbf{E}_{L\times P} \end{bmatrix}$$

and

$$\underline{\mathbf{M}}(n) = \mathbf{F} \begin{bmatrix} \mathbf{0}_{L\times P} \\ \mathbf{M}_{L\times P} \end{bmatrix}$$

Next, the cross-correlation matrix of size $(P \times P)$ between the microphone signal and the cancellation error matrices is computed by $\mathbf{S}_{EM} = E[\underline{\mathbf{E}}^H \underline{\mathbf{M}}]$ and the covariance matrix of the microphone signals by $\mathbf{S}_{MM} = E[\underline{\mathbf{M}}^H \underline{\mathbf{M}}]$. The global decision statistic for double-talk detection is defined to be

$$\xi_{AsifMIMO} = \frac{1}{P}\ tr(\mathbf{I}_{P\times P}\ -\ \mathbf{S}_{EM}\ \ \mathbf{S}_{MM}^{-1}) \tag{4.25}$$

The scaling factor $\frac{1}{P}$ is used for normalizing the decision statistic and it can be verified that for $\mathbf{v} = \mathbf{0}_{P\times 1}$, $\xi_{AsifMIMO} \approx 1$ and for $\mathbf{v} \neq \mathbf{0}_{P\times 1}$, $\xi_{AsifMIMO} < 1$ as follows: Extending from the single channel case, it can be shown that

$$\mathbf{S}_{EM} = \left(\underline{\mathbf{H}}^H - \underline{\hat{\mathbf{H}}}^H\right) \underline{\mathbf{S}}\ \underline{\mathbf{H}} + \mathbf{R}_{\underline{\mathbf{vv}}}$$

$$\mathbf{S}_{MM} = \underline{\mathbf{H}}^H \underline{\mathbf{S}}\ \underline{\mathbf{H}} + \mathbf{R}_{\underline{\mathbf{vv}}} \tag{4.26}$$

where $\underline{\hat{\mathbf{H}}}$ is the estimated echo-path matrix, $\underline{\mathbf{H}}$ is the true channel matrix, $\underline{\mathbf{S}} = E[\mathbf{D}^H(n)\ \mathbf{G}\ \mathbf{D}(n)]$ where $\mathbf{D}(n) = [\mathbf{D}_1(n)\ ....,\ \mathbf{D}_Q(n)]$ where $\mathbf{D}_i(n)$ is the i^{th} loudspeaker signal as defined in equation 4.6 and $\mathbf{R}_{\underline{v}\ \underline{v}} = E[\underline{\mathbf{v}}\ \underline{\mathbf{v}}]$. Next, let $\mathbf{I}_{P\times P} = \mathbf{S}_{MM}\ \mathbf{S}_{MM}^{-1}$ substituting in (4.25) yields:

$$\begin{aligned} \xi_{AsifMIMO} &= \frac{1}{P}\ tr\ \left[\mathbf{S}_{MM}\ \mathbf{S}_{MM}^{-1} - \mathbf{S}_{EM}\ \mathbf{S}_{MM}^{-1}\right] \\ &= \frac{1}{P}\ tr\ \left[(\mathbf{S}_{MM} - \mathbf{S}_{EM})\ \mathbf{S}_{MM}^{-1}\right], \end{aligned} \tag{4.27}$$

which after substituting equation (4.26) yields,

$$\xi_{AsifMIMO} = \frac{1}{P}\ tr\left[\underline{\hat{\mathbf{H}}}^H \underline{\mathbf{S}}\ \underline{\mathbf{H}}\ \ \left(\underline{\mathbf{H}}^H \underline{\mathbf{S}}\ \underline{\mathbf{H}} + \mathbf{R}_{\underline{\mathbf{vv}}}\right)^{-1}\right] \tag{4.28}$$

Now it is clear that for $\mathbf{v} = \mathbf{0}_{P\times 1}$ i.e. $\mathbf{R}_{\underline{\mathbf{v}}\ \underline{\mathbf{v}}} = \mathbf{0}_{P\times P}$,

$$\begin{aligned} \xi_{AsifMIMO} &= \frac{1}{P}\ tr\left[\underline{\hat{\mathbf{H}}}^H \underline{\mathbf{S}}\ \underline{\mathbf{H}}\ \ \left(\underline{\mathbf{H}}^H \underline{\mathbf{S}}\ \underline{\mathbf{H}}\right)^{-1}\right] \\ &\approx 1 \end{aligned} \tag{4.29}$$

and similarly for $\mathbf{v} \neq \mathbf{0}_{P\times 1}$ i.e. $\mathbf{R}_{\underline{\mathbf{v}}\ \underline{\mathbf{v}}} \neq \mathbf{0}_{P\times P}$, $\xi_{AsifMIMO} < 1$. Hence, it can be concluded that the decision statistic and the threshold $T \approx 1$ are independent of the excitation signals and the echo-path variations.

4.4. SIMULATION RESULTS

To evaluate the proposed double-talk detector, the probability of miss P_m (not detecting double-talk when it is present) is plotted, versus the probability of false alarm P_f (declaring double-talk when it is not present). This is a standard technique for evaluating a double-talk detector [16]. The estimation of false-alarm P_f and the miss probability P_m was made according to [16].

If $\xi_{Asif} < T$, it is concluded that the captured frame of the microphone signal is corrupted by the near-end speech else the adaptation continues. The recorded digital speech sampled at 16 KHz is used as the far-end and near-end speech and a measured $l = 8000$ sample (500 ms) room impulse response of a $10' \times 10' \times 8'$ room is used as the loudspeaker-microphone environment $\mathbf{h}$. The room response is also normalized so that $\sigma_y^2 = \sigma_x^2$. The performance of the proposed double-talk detector is evaluated when the echo to background ratio ($EBR = \sigma_y^2/(\sigma_v^2 + \sigma_w^2)$) is set to 0 dB and the echo to ambient noise ratio ($ENR = \sigma_y^2/\sigma_w^2$) is set to 30 dB. The average probability of miss is estimated using a 5 second speech signal as the far-end speech and 12 sentences each about 2 seconds long are used as the near-end speech signals.

Results are compared with the Benesty's double-talk detector proposed in [22] for the single-channel case. The probability of miss (P_m) is the probability of not detecting (miss) double-talk when it is present, therefore a smaller value of P_m indicates better performance. From Figure 4.3, it can be observed that both the double-talk detectors have similar performance. For the Benesty's double-talk detector proposed in [22], the detection statistic is computed by:

$$\xi^2_{Benesty} \approx \frac{\mathbf{s}^H \hat{\underline{\mathbf{h}}}}{\sigma_m^2} \tag{4.30}$$

where $\mathbf{s}$ is the cross-correlation vector between the far-end and the microphone signal given by $\mathbf{s} = \mathbf{S}\underline{\mathbf{h}}$ substituting in equation 4.30 yields:

$$\xi^2_{Benesty} \approx \frac{\underline{\mathbf{h}}^{\mathbf{H}} \mathbf{S} \hat{\underline{\mathbf{h}}}}{\underline{\mathbf{h}}^{\mathbf{H}} \mathbf{S} \underline{\mathbf{h}} + \sigma_v^2} \tag{4.31}$$

since the detection statistic is a scalar:

$$\xi^2_{Benesty} \approx \frac{\hat{\underline{\mathbf{h}}}^{\mathbf{H}} \mathbf{S} \underline{\mathbf{h}}}{\underline{\mathbf{h}}^{\mathbf{H}} \mathbf{S} \underline{\mathbf{h}} + \sigma_v^2} \tag{4.32}$$

It can be observed from equations 4.18 and 4.32 that $\xi_{Asif} \approx \xi^2_{Benesty}$ and simulations (Figure 4.3) demonstrate this convergence as well. It should be noted that both the detection statistics are computed differently. The proposed algorithm is based on the cross-correlation

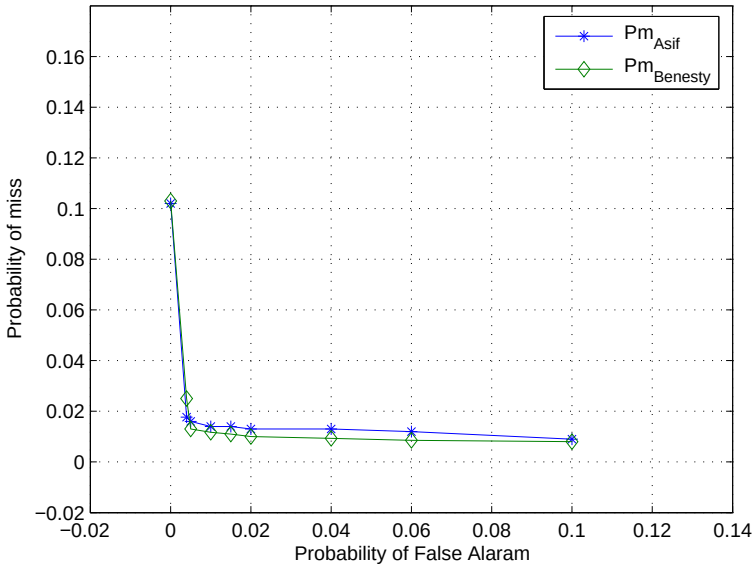

Figure 4.3: Receiver Operating Characteristics (ROC).

coefficient between the microphone signal and the cancellation error whereas in the later case it is based on the cross-correlation *vector* between the far-end and the microphone signal. However, the proposed algorithm is computationally very efficient as only the maximum cross correlation needs to computed whereas in the later case the entire cross correlation vector is required.

4.5. SUMMARY

In this section, a frequency domain calculation scheme for a novel cross-correlation based double-talk detector proposed in [4] was discussed. The proposed technique is computationally very simple, as only $2L$ multiplications, $2L+1$ additions and a division are required to compute the decision statistic at each frame i.e. the computational complexity is $O(L)$ where as for the Benesty's double-talk detector it is $O(L^2)$ and it is also shown that the proposed detection statistic converges to the Benesty's double-talk detector. Multi-channel double-talk detection is non trivial, yet the proposed technique for double-talk detection is successfully extended to multi-channel case by defining a global test statistic based on the cross-correlation matrix between the microphone signals and the cancellation errors that takes into account all P microphone signals.

5. A NOVEL NORMALIZED ECHO-PATH CHANGE DETECTOR

5.1. INTRODUCTION

Most teleconferencing conversations are conducted in the presence of acoustic echoes [10]; if the delay between the speech and its echo is more than a few tens of milliseconds, the echo is distinctly noticeable. An acoustic echo canceller (AEC) is used to remove the echo created due to the loudspeaker-microphone environment ($\mathbf{h}$) [9]. In an AEC the echo-path (loudspeaker microphone path $\mathbf{h}$) is adaptively modelled using a filter ($\hat{\mathbf{h}}$), which is then used to synthesize a replica of the echo ($\hat{y}$). This synthesized replica of the echo is subtracted from the echo-corrupted microphone signal (m) to get an echo-free signal (e). When the near-end talker (v) is active or when the speech comes from both the far-end ($\mathbf{x}$) and near-end (v), the adaptive filter coefficients diverge from the true echo path impulse response if the adaptation is not halted. A double-talk detector is used to stop the AEC's filter adaptation during periods of near-end speech [9]. A double-talk detector should be able to detect a double-talk condition quickly and accurately so as to freeze adaptation as soon as possible; at the same time it should be able to track any echo-path changes and should be able to distinguish the double-talk from the echo-path variations [17]. Typically, better immunity towards double-talk results in declaring echo-path changes as double-talk, which adversely affects the performance of an AEC as the adaptation is frozen when it really needs to be on. Thus, an efficient and simple echo-path change detector is required so as to differentiate any echo-path variations from double-talk.

An optimum decision variable for echo-path change detection should behave as follows:

1. If no echo-path variations i.e. when the adaptive filter is converged $\xi_{EP} < T_{EP}$.
2. During echo-path variations i.e. when the adaptive filter is not converged $\xi_{EP} \geq T_{EP}$ and
3. ξ_{EP} is insensitive to near-end speech v.

Figure 5.1 shows the basic structure of the adaptive acoustic echo canceller. The far-end signal $\mathbf{x}$ is filtered through the room impulse response $\mathbf{h}$ to get the echo signal

$$y(n) = \mathbf{h^T x} \tag{5.1}$$

where

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

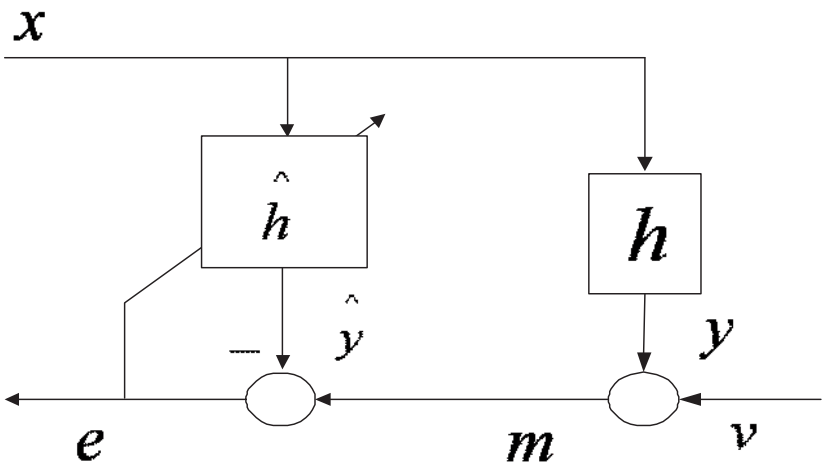

Figure 5.1: Basic AEC model.

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

and L is the length of the echo-path. This echo signal is added to the near-end speech signal v to get the microphone signal

$$m(n) = y(n) + v(n) \tag{5.2}$$

The error signal at time n is defined as

$$e(n) = m(n) - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x} \tag{5.3}$$

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

This section is structured as follows: In Section 5.2, the proposed echo-path change detection statistic is derived. A comprehensive study on the proposed algorithm for echo-path change detection is done in Section 5.3 which is followed by a summary and conclusion in Section 5.4.

5.2. ECHO-PATH CHANGE DETECTION ALGORITHM

In this section, a novel normalized cross-correlation based echo-path change detector is derived. Referring to Figure 5.1, the cross-correlation between the microphone signal m, and the cancellation error e is given by:

$$\begin{aligned} r_{em} &= E[em] \\ &= E[(y + v - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x})(y + v)] \end{aligned}$$

$$\begin{aligned} r_{em} &= E[(\mathbf{h^T x} - \hat{\mathbf{h}}^\mathbf{T}\mathbf{x} + v)(\mathbf{h^T x} + v)] \\ &= (\mathbf{h} - \hat{\mathbf{h}})^T E[\mathbf{x}\mathbf{x^T}]\mathbf{h} + E[v^2] \\ &= (\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2 \end{aligned} \tag{5.4}$$

where σ_v^2 is the variance of the near-end speech, the far-end speech vector $\mathbf{x}$ the near-end signal v are independent and are assumed to be of zero mean. Variance of the microphone signal is given by:

$$\begin{aligned} \sigma_m^2 &= E[m^2] = E[(y+v)^2] \\ &= E[y^2] + E[v^2] = E[\mathbf{h^T x}(\mathbf{h^T x})^T] + \sigma_v^2 \\ &= \mathbf{h^T} R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2. \end{aligned} \tag{5.5}$$

and, finally, the variance of the cancellation error e is given by:

$$\begin{aligned} \sigma_e^2 &= E[e^2] \\ &= E[((\mathbf{h} - \hat{\mathbf{h}})^T\mathbf{x} + v)((\mathbf{h} - \hat{\mathbf{h}})^T\mathbf{x} + v)^T] \\ &= (\mathbf{h} - \hat{\mathbf{h}})^T E[\mathbf{x}\mathbf{x}^T](\mathbf{h} - \hat{\mathbf{h}}) + E[v^2] \\ &= (\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}}(\mathbf{h} - \hat{\mathbf{h}}) + \sigma_v^2 \end{aligned} \tag{5.6}$$

The new normalized decision statistic is defined as

$$\xi_{AsifEPD} = \left| \frac{r_{em} - \sigma_e^2}{\sigma_m^2 - r_{em}} \right| \tag{5.7}$$

substituting equations 5.4 , 5.5 and 5.6 in 5.7 yields:

$$\xi_{AsifEPD} = \left| \frac{(\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}}\hat{\mathbf{h}}}{\mathbf{h^T} R_{\mathbf{xx}}\hat{\mathbf{h}}} \right| \tag{5.8}$$

It can be observed from equation 5.8, for $\mathbf{h} \approx \hat{\mathbf{h}}$, $\xi_{AsifEPD} \approx 0$ and for $\mathbf{h} \neq \hat{\mathbf{h}}$, $\xi_{AsifEPD} > 0$. Thus, the proposed echo-path change detector meets the needs of an optimal echo-path change detector.

The proposed algorithm is computationally very efficient, as only 9 operations per sample are required as compared to $6l$+4 operations for the decision statistic proposed in [17]. Further, the proposed decision statistic is normalized appropriately i.e. it is approximately zero in the absence of echo-path variations and is greater than zero during echo-path variations.

5.3. EXPERIMENTS AND RESULTS

The values of r_{em}, σ_m^2 and σ_e^2 in 5.7 are exact and not available in practice. As a result, the final decision statistic is given by:

$$\xi_{Asif} = \left| \frac{\hat{r}_{em} - \hat{\sigma}_e^2}{\hat{\sigma}_m^2 - \hat{r}_{em}} \right| \tag{5.9}$$

where the estimates denoted by a hat are obtained using the exponential recursive weighting algorithm, [19] [20]:

$$\begin{aligned} \hat{r}_{em}(t) &= \lambda \hat{r}_{em}(t-1) + (1-\lambda) e(t) m(t) \\ \hat{\sigma}_m^2(t) &= \lambda \hat{\sigma}_m^2(t-1) + (1-\lambda) m^2(t) \\ \hat{\sigma}_e^2(t) &= \lambda \hat{\sigma}_e^2(t-1) + (1-\lambda) e^2(t) \end{aligned}$$

The echo-path change detector works as follows:

1. When $\xi_{Asif} > T$ (T is a properly chosen detection threshold), it is declared that the echo-canceller has not converged i.e. the echo-path has changed, the adaptation is enabled even if the double-talk detector declares a double-talk.

2. Whenever $\xi_{Asif} < T$, the detector decides that the echo-canceller has converged i.e. there are no echo-path variations.

Detection threshold T is chosen to be slightly greater than the steady state value (the value of ξ_{Asif} in the absence of any echo-path variations) as shown in Figure 5.2. The recorded digital speech sampled at 16 KHz is used as far-end speech $\mathbf{x}$ and near-end speech v and a measured $L = 8000$ sample (500 ms) room impulse response of a $10' \times 10' \times 8'$ room is used as the loudspeaker-microphone environment $\mathbf{h}$. The room response was collected using a stereo system.

To create echo-path variations, the room response was changed from the collected left channel impulse response to the right channel response after 320 frames. As can be seen in Figure 5.3, these changes in echo-path were detected. The echo-path change statistic goes above the detection threshold T as observed, and hence the variations in echo-path are detected. Next, the echo-path gain was increased by 2 dB after 320 frames. Simulations demonstrate the efficiency of the proposed algorithm, even variations in filter coefficients by 2dB are detected as shown in Figure 5.4. Thus, it can be concluded that the proposed decision statistic detects any echo-path variations efficiently. It is computationally of the order of the magnitude simpler as compared to the conventional statistic and meets the needs of an optimal echo-path change detector.

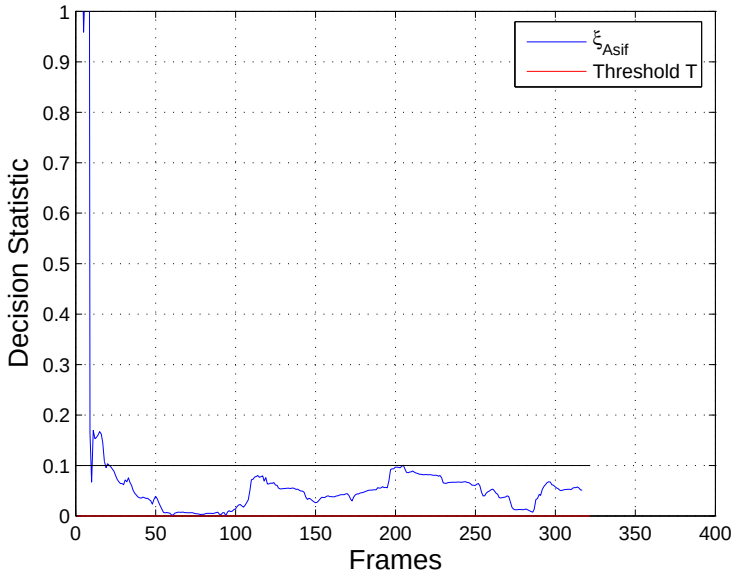

Figure 5.2: ξ_{Asif} as function of time frames, selecting detection threshold T.

5.4. CONCLUSION

A novel normalized sample by sample echo-path change detector is proposed. To summarize, the major advantages of the proposed echo-path change detector are listed:

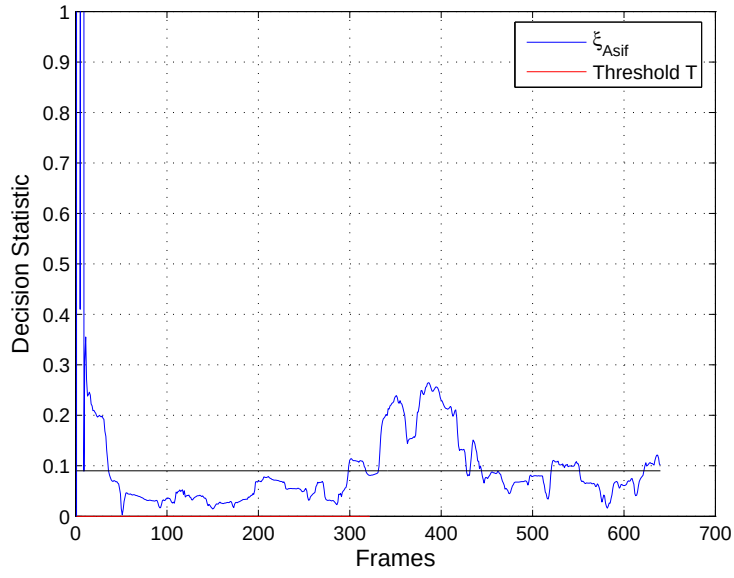

Figure 5.3: ξ_{Asif} as function of time frames.

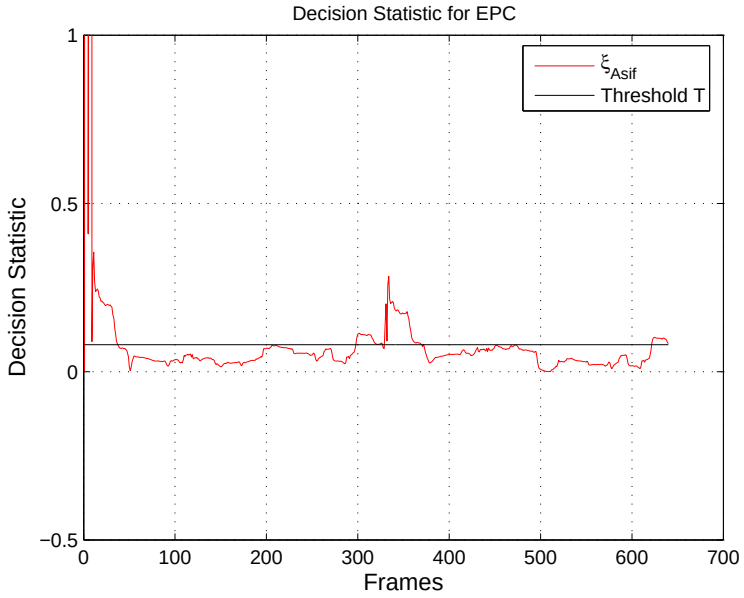

Figure 5.4: ξ_{Asif} as function of time frames.

- Detects any echo-path variations and is normalized appropriately i.e. the detection statistic is greater than zero only for echo-path variations.
- Low added complexity when implemented with any adaptive algorithm.
- Independent of the near-end speech/doubletalk.

The proposed echo-path change detector can also serve as a good download test for a two-path AEC, and, when used with a good double-talk detector makes, the complete system (AEC) very robust.

6. SIMPLE AND EFFICIENT SOLUTIONS TO AEC PROBLEMS

6.1. INTRODUCTION

An acoustic echo canceller (AEC) is used to remove the echo created due to the acoustic coupling ($\mathbf{h}$) between the loudspeaker-microphone environment [9]. Figure 6.1 shows the basic block diagram of an AEC. The far-end signal $\mathbf{x}$ is filtered through the echo-path $\mathbf{h}$ to get the echo signal

$$y(n) = \mathbf{h^T x} \tag{6.1}$$

where

$$\mathbf{h} = [h_0 \;\; h_1 \;\; .... \;\; , \;\; h_{l-1}]^T,$$

$$\mathbf{x}[n] = [x(n) \;\; x(n-1) \;\; .... \;\; , \;\; x(n-l+1)]^T,$$

and l is the length of the echo-path. This echo signal is added to the near-end speech signal v to get the microphone signal:

$$m(n) = y(n) + v(n) \tag{6.2}$$

The error signal at time n is defined as

$$e(n) = m(n) - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x} \tag{6.3}$$

and is used to adapt the l taps of the AEC's adaptive filter $\hat{\mathbf{h}}$ to generate an estimate of the echo $\hat{y}$.

In an AEC, echo cancellation is achieved by adaptively modelling the echo-path ($\mathbf{h}$) using an adaptive filter ($\hat{\mathbf{h}}$), which is then used to synthesize a replica of the echo ($\hat{y}$). This synthesized replica of the echo is subtracted from the echo-corrupted microphone signal to get an echo free signal (e). When the near-end talker (v) is active or when the speech comes from both the far-end and near-end identification of the echo-path becomes problematic and the adaptive filter coefficients diverge from the true echo-path if the adaptation is not halted. A double-talk detector is used to stop the AEC's filter adaptation during periods of near-end speech. A double-talk detector should be able to detect double-talk condition quickly

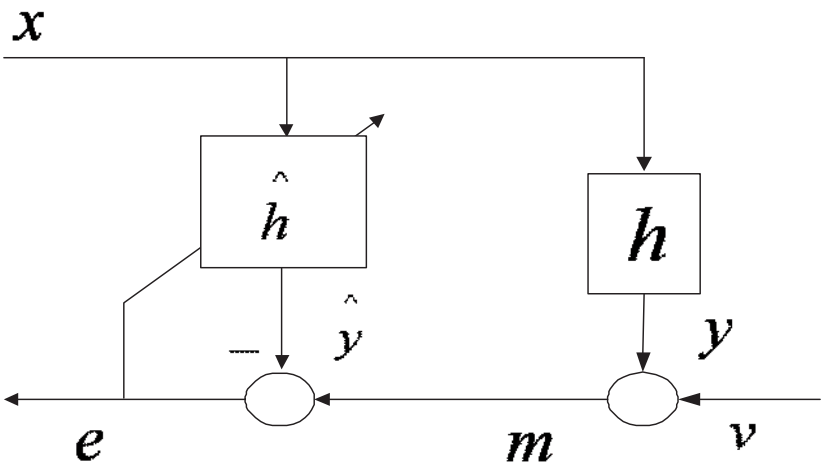

Figure 6.1: Basic AEC model.

and accurately so as to freeze AEC's filter adaptation as soon as possible. In a double-talk detector, a decision variable ξ_{DTD} is formed from the available signals $x(n), m(n)$, and $e(n)$. This decision variable is compared to a preset threshold T_{DTD}. An optimum decision variable for double-talk detection should behave as follows [9]:

1. If $v(n) = 0$ (no near-end speech) then $\xi_{DTD} \geq T_{DTD}$.

2. If $v(n) \neq 0$ (near-end speech is present) then $\xi_{DTD} < T_{DTD}$.

3. ξ_{DTD} is insensitive to echo-path variations when $v = 0$ i.e no near-end speech.

In this section, a novel sample by sample double-talk detection algorithm is presented based on cross-correlation between the microphone signal and the cancellation error. The double-talk detector is designed in such a way that it meets the needs of an optimal double-talk detector. It is shown both theoretically and by simulations that the proposed double-talk detector converges to the recently proposed double-talk detector based on a normalized cross-correlation vector between the far-end signal ($\mathbf{x}$) and the microphone scalar (m) in [3] whose performance was shown to be superior compared to other double-talk detectors based on the cross-correlation coefficient. However, the proposed algorithm is computationally very attractive as only 2 multiplications, 2 additions, 1 subtraction and a division are required to compute the decision statistic at each sample (i.e. 6 operations per sample) whereas in the later case $2l + 1$ multiplications, $l + 1$ additions and a division are required to compute the decision statistic at each sample i.e. $3l$+3 operations per sample are required, where l is the frame size (typically $l \geq 512$). A shorter version of the proposed double-talk detection algorithm was proposed in [4]. The proposed double-talk detector can be implemented in the frequency domain and can be extended to the multi-channel case as well [5].

False alarm rate of a double-talk detector increases when echo-path changes [9], this increases the convergence rate of an AEC since adaptation is frozen when it really needs to be on. Thus, an efficient and simple echo-path change detector is required so as to differentiate any echo-path variations from double-talk. In an echo-path change detector, a decision variable ξ_{EP} is formed from the available signals $x(n), m(n)$, and $e(n)$. This decision variable is compared to a preset threshold T_{EP}. An optimum decision variable for echo-path change detection should behave as follows:

1. If no echo-path variations i.e. when the adaptive filter is converged $\xi_{EP} < T_{EP}$.

2. During echo-path variations i.e. when the adaptive filter is not converged $\xi_{EP} \geq T_{EP}$ and

3. ξ_{EP} is insensitive to near-end speech v.

In this section, a novel test statistic for echo-path change detection based on the cross-correlation coefficient between the microphone signal and the cancellation error is also presented. The decision statistic is designed in such a way that it meets the needs of an optimal echo-path change detector efficiently. Results are compared with the echo-path change detector proposed in [17] which is based on the orthogonality theorem. The proposed algorithm is computationally very efficient, as only 3 multiplications, 3 additions, 2 subtractions and a division are required to compute the decision statistic at each sample (i.e. 9 operations per sample) as compared to $2l + 2$ multiplications, $l + 1$ divisions and $3l + 1$ additions (i.e. $6l+4$ operations) per sample are required for the decision statistic proposed in [17]. Further, the proposed decision statistic is normalized appropriately i.e. it is approximately zero in the absence of echo-path variations and is greater than zero during echo-path variations.

Finally, a robust fast recursive least squares algorithm [9] is successfully applied to the problem of acoustic echo cancellation by combining it with the proposed double-talk and echo-path change detection algorithms and the advantages of the proposed algorithms are also listed.

This section is structured as follows: In Section 6.2, the second order statistics for the available signals $x(n), m(n)$, and $e(n)$ are derived which subsequently are used for deriving the double-talk and echo-path change detection statistics. In Section 6.3, the previous algorithms for double-talk and echo-path change detection are reviewed. In Section 6.4, the proposed double-talk detector is introduced and formulated. In Section 6.6, the echo-path change detector is derived. A single channel AEC (using the fast recursive least squares for adaptive filtering) using the proposed algorithms for double-talk and echo-path change detection is

implemented in Section 6.7. A comprehensive study on the proposed algorithms for double-talk and echo-path change detection is done in Section 6.8 which is followed by a summary and conclusion in Section 6.9.

6.2. SECOND-ORDER STATISTICS

In this section, the second order statistics for the available signals $x(n), m(n)$, and $e(n)$ are derived. Referring to Figure 6.1, first the cross-correlation vector between the far-end signal vector $\mathbf{x}$ and the microphone scalar m is derived:

$$\begin{aligned} \mathbf{r}_{\mathbf{x}m} &= E[m\mathbf{x}^T] \\ &= E[(\mathbf{h^T x} + v)\mathbf{x}^T] \\ &= \mathbf{h^T}E[\mathbf{x}\mathbf{x^T}] + E[v\mathbf{x^T}] \end{aligned} \tag{6.4}$$

where $E[\bullet]$ denotes the mathematical expectation. Noting that the near-end speech v is independent of the far-end signal vector $\mathbf{x}$, the cross-correlation vector between the far-end signal vector $\mathbf{x}$ and the microphone scalar m is given by

$$\mathbf{r}_{\mathbf{x}m} = \mathbf{h^T} R_{\mathbf{xx}} \tag{6.5}$$

where $R_{\mathbf{xx}} = E[\mathbf{x}\mathbf{x^T}]$. Next, the cross-correlation vector between the far-end signal vector $\mathbf{x}$ and the cancellation error e is derived:

$$\begin{aligned} \mathbf{r}_{e\mathbf{x}} &= E[e\mathbf{x^T}] \\ &= E[(y + v - \hat{\mathbf{h}}^\mathbf{T}\mathbf{x})\mathbf{x^T}] \\ &= E[(\mathbf{h^T x} - \hat{\mathbf{h}}^\mathbf{T}\mathbf{x})\mathbf{x^T}] \\ &= (\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}} \end{aligned} \tag{6.6}$$

the cross-correlation coefficient between the microphone signal m and the cancellation error e is given by:

$$\begin{aligned} r_{em} &= E[em] \\ &= E[(y + v - \hat{\mathbf{h}}^\mathbf{T}\mathbf{x})(y + v)] \\ &= E[(\mathbf{h^T x} - \hat{\mathbf{h}}^\mathbf{T}\mathbf{x} + v)(\mathbf{h^T x} + v)] \\ &= (\mathbf{h} - \hat{\mathbf{h}})^T E[\mathbf{x}\mathbf{x^T}]\mathbf{h} + E[v^2] \\ &= (\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2 \end{aligned} \tag{6.7}$$

where σ_v^2 is the variance of the near-end speech. Variance of the microphone signal is given by:

$$\begin{aligned}
\sigma_m^2 &= E[m^2] \\
&= E[(y+v)^2] \\
&= E[y^2] + E[v^2] \\
&= E[\mathbf{h^T x}(\mathbf{h^T x})^T] + \sigma_v^2 \\
&= \mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2. \qquad (6.8)
\end{aligned}$$

and, finally, the variance of the cancellation error e is given by:

$$\begin{aligned}
\sigma_e^2 &= E[e^2] \\
&= E[((\mathbf{h}-\hat{\mathbf{h}})^T\mathbf{x}+v)((\mathbf{h}-\hat{\mathbf{h}})^T\mathbf{x}+v)] \\
&= (\mathbf{h}-\hat{\mathbf{h}})^T E[\mathbf{x}\mathbf{x}^T](\mathbf{h}-\hat{\mathbf{h}}) + E[v^2] \\
&= (\mathbf{h}-\hat{\mathbf{h}})^T R_{\mathbf{xx}}(\mathbf{h}-\hat{\mathbf{h}}) + \sigma_v^2 \qquad (6.9)
\end{aligned}$$

6.3. PREVIOUS WORK

Referring to Figure 6.1, Ye and Wu [17] first proposed using the cross-correlation vector between the far-end signal vector $\mathbf{x}$, which is played out of the speakers, and the AEC's cancellation error e, $\mathbf{r}_{e\mathbf{x}} = E[e\mathbf{x}^T]$, as the basis for double-talk detection. In this section, this algorithm is referred as the Conventional cross-correlation based detector. Simulation results by Benesty [3] have shown that this approach does not work well for detecting double-talk, and a theoretical derivation provides further insight. Clearly from equation 6.6 it can be observed that $\mathbf{r}_{e\mathbf{x}}$ is high only when there is a change in the echo-path; hence, this approach is more suitable for tracking echo-path variations rather than detecting double-talk. The decision statistic used to detect doubletalk/echo-path variations in [17] is given by

$$\xi_{Conventional} = \left| \frac{\mathbf{r}_{e\mathbf{x}}}{\sigma_x \ \sigma_e} \right| \qquad (6.10)$$

substituting equations 6.6 and 6.9 in 6.10 yields

$$\xi_{conventional} = \left| \frac{(\mathbf{h}-\hat{\mathbf{h}})^T R_{\mathbf{xx}}}{\sigma_x \ \sqrt{(\mathbf{h}-\hat{\mathbf{h}})^T R_{\mathbf{xx}}(\mathbf{h}-\hat{\mathbf{h}}) + \sigma_v^2}} \right| \qquad (6.11)$$

It can be observed from equation 6.11 that the decision statistic is not normalized properly i.e. it neither meets the needs of an optimal double-talk detector nor an optimal echo-path change detector defined in Section 6.1. More recently, Benesty, et al. [3] [16] proposed a double-talk detection algorithm based on the cross-correlation between the far-end signal vector $\mathbf{x}$ and the microphone signal scalar m, $\mathbf{r}_{\mathbf{x}m} = E[\mathbf{x}m]$, which is referred as Benesty in this section. Benesty's decision statistic used to detect double-talk in [3] is given by

$$\xi_{Benesty} = \sqrt{\mathbf{r}_{\mathbf{x}m}(\sigma_m^2 R_{\mathbf{xx}})^{-1}\mathbf{r}_{\mathbf{x}m}^T} \tag{6.12}$$

substituting equations 6.5 and 6.8 in 6.12 yields

$$\begin{aligned} \xi_{Benesty}^2 &= \frac{\mathbf{h^T} R_{\mathbf{xx}} R_{\mathbf{xx}}^{-1} R_{\mathbf{xx}} \mathbf{h}}{\sigma_m^2} \\ &= \frac{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h}}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2} \end{aligned} \tag{6.13}$$

It is clear from equation 6.13 that the decision statistic is normalized appropriately and meets the need of an optimal double-talk detector.

6.4. DOUBLE-TALK DETECTION

Instead of using $\mathbf{r}_{e\mathbf{x}}$ or $\mathbf{r}_{\mathbf{x}m}$ as discussed in section 6.3, using the cross-correlation between the microphone signal m and the cancellation error e, $r_{em} = E[em^T]$, as the basis for double-talk detection is proposed. The new decision statistic is defined to be

$$\xi_{AsifDTD} = 1 - \frac{r_{em}}{\sigma_m^2}. \tag{6.14}$$

Substituting equations 6.7 and 6.8 in 6.14 yields:

$$\begin{aligned} \xi_{AsifDTD} &= 1 - \frac{(\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2} \\ &= \frac{\hat{\mathbf{h}}^\mathbf{T} R_{\mathbf{xx}} \mathbf{h}}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2}. \end{aligned} \tag{6.15}$$

since the decision statistic is a scalar

$$\xi_{AsifDTD} = \frac{\mathbf{h^T} R_{\mathbf{xx}} \hat{\mathbf{h}}}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2}. \tag{6.16}$$

It can be observed from equation 6.16, that for $v = 0$ i.e. no near-end speech, $\xi_{AsifDTD} \approx 1$ and for $v \neq 0$ i.e. during near-end speech, $\xi_{AsifDTD} < 1$. Thus, the proposed double-talk detector meets the needs of an optimal double-talk detector.

The values for r_{em} and σ_m^2 in 6.14 are exact and not available in practice. As a result, the final decision statistic is given by:

$$\xi_{AsifDTD} = 1 - \frac{\hat{r}_{em}}{\hat{\sigma}_m^2} \tag{6.17}$$

which is based on the estimates $\hat{r}_{em}[n]$ and $\hat{\sigma}_m^2[n]$. The estimates are found using the exponential recursive weighting algorithm, [19] [20]:

$$\hat{r}_{em}[n] = \lambda \hat{r}_{em}[n-1] + (1-\lambda)e[n]m[n]$$

$$\hat{\sigma}_m^2[n] = \lambda \hat{\sigma}_m^2[n-1] + (1-\lambda)m^2[n] \tag{6.18}$$

where $e[n]$ is the captured sample of the cancellation error at time n, $m[n]$ is the captured microphone signal sample at time n, and λ is the exponential weighting factor. Smaller values of λ yield better time varying signal tracking capability at the expense of worse estimation accuracy. In practice for slowly time varying signals, $0.95 \leq \lambda \leq 1$ is usually chosen [17]. If

$$\xi_{AsifDTD} < T_{DTD} \tag{6.19}$$

it is concluded that the captured sample of the microphone signal is corrupted by the near-end speech and the AEC's adaptive filter adaptation is disabled. Otherwise, adaptation continues (where T_{DTD} is a preselected threshold (close to 1)).

6.5. RELATION BETWEEN PROPOSED AND BENESTY'S METHOD

The proposed decision statistic is given by equation 6.14, which can be rewritten as in equation 6.16, and Benesty's double-talk decision statistic is given in equation 6.12. The decision statistics are different as the former in based on r_{em}, and the latter is based on $\mathbf{r}_{\mathbf{x}m}$. Although the decision statistics are different, they can be shown to result in a similar expression. It can be observed from equation 6.13 :

$$\xi_{Benesty}^2 = \frac{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h}}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2} \tag{6.20}$$

and from equation 6.16

$$\xi_{AsifDTD} \;=\; \frac{\mathbf{h^T} R_{\mathbf{xx}} \hat{\mathbf{h}}}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2}. \tag{6.21}$$

In addition to the square root, the other difference between the decision statistics is in the numerator; the taps of the AEC filter $\hat{\mathbf{h}}^{\mathbf{T}}$ are used in $\xi_{AsifDTD}$ and the true echo-path impulse

response $\mathbf{h^T}$ in $\xi_{Benesty}$. However, for practical implementation and computational simplicity, the authors in [3] substitute $\hat{\mathbf{h}}^{\mathbf{T}}$ for $\mathbf{h^T}$ resulting in similar decision statistics i.e.

$$\xi^2_{Benesty} = \frac{\mathbf{h^T} R_{\mathbf{xx}} \hat{\mathbf{h}}}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2}. \tag{6.22}$$

and is computed by

$$\xi^2_{Benesty} = \frac{\hat{\mathbf{r}}_{\mathbf{x}m} \hat{\mathbf{h}}}{\hat{\sigma}^2_m} \tag{6.23}$$

where the estimates $\hat{\mathbf{r}}_{\mathbf{x}m}$ and $\hat{\sigma}^2_m$ are again found using the exponential recursive weighting algorithm [19] [20]:

$$\begin{aligned} \hat{\mathbf{r}}_{\mathbf{x}m}[n] &= \lambda \hat{\mathbf{r}}_{\mathbf{x}m}[n-1] + (1-\lambda)\mathbf{x}[n]m[n] \\ \hat{\sigma}^2_m[n] &= \lambda \hat{\sigma}^2_m[n-1] + (1-\lambda)m^2[n] \end{aligned} \tag{6.24}$$

simulations (Figure 6.2) further demonstrate the convergence of both the double-talk detectors.

In addition to its simplicity, another main advantage of the proposed algorithm for double-talk detection is that only the maximum cross-correlation needs to be computed instead of computing the entire cross-correlation vector required by the Benesty's test statistic. This results in significant computational savings; requiring only 6 operations to compute the decision statistic at each sample, where as in the later case $3l + 3$ operations are required at each sample where l is the frame size (typically $l \geq 512$).

6.6. ECHO-PATH CHANGE DETECTION ALGORITHM

Instead of using $\mathbf{r_{ex}}$ as the basis for echo-path change detection, the cross-correlation coefficient between the microphone signal (m) and the cancellation error (e) is used. The new normalized decision statistic is defined as:

$$\xi_{AsifEPD} = \left| \frac{r_{em} - \sigma_e^2}{\sigma_m^2 - r_{em}} \right| \tag{6.25}$$

substituting equations 6.7 , 6.8 and 6.9 in 6.25 yields:

$$\xi_{AsifEPD} = \left| \frac{(\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2 - (\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}} (\mathbf{h} - \hat{\mathbf{h}}) - \sigma_v^2}{\mathbf{h^T} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2 - (\mathbf{h^T} - \hat{\mathbf{h}}^{\mathbf{T}}) R_{\mathbf{xx}} \mathbf{h} - \sigma_v^2} \right|$$

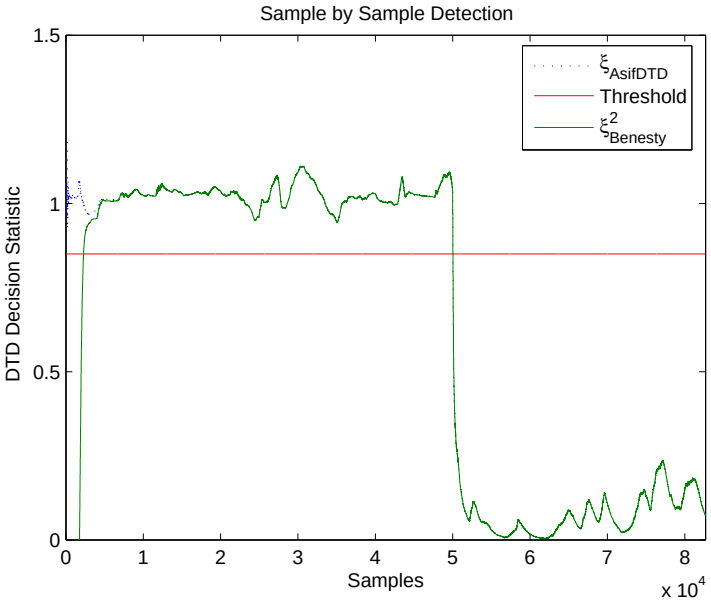

Figure 6.2: Double-talk detection statistics as a function of time (samples), showing the convergence of the proposed and Benesty's double-talk detection statistics.

$$\xi_{AsifEPD} = \left| \frac{(\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}} \hat{\mathbf{h}}}{\mathbf{h^T} R_{\mathbf{xx}} \hat{\mathbf{h}}} \right| \tag{6.26}$$

it can be observed from equation 6.26, for $\mathbf{h} \approx \hat{\mathbf{h}}$, $\xi_{AsifEPD} \approx 0$ and for $\mathbf{h} \neq \hat{\mathbf{h}}$, $\xi_{AsifEPD} > 0$. Thus, the proposed echo-path change detector meets the needs of an optimal echo-path change detector. Whereas, for the conventional echo-path change detector proposed in [17], in the absence of the near-end speech ($\sigma_v^2 = 0$), the decision statistic is not normalized properly i.e. $\xi_{Conventional}$ is not necessarily ≈ 0 for $\mathbf{h} \approx \hat{\mathbf{h}}$. Hence, it does not meets the needs of an optimal echo-path change detector.

The values of r_{em}, σ_m^2 and σ_e^2 in 6.25 are exact and not available in practice. As a result, the final decision statistic is given by:

$$\xi_{Asif} = \left| \frac{\hat{r}_{em} - \hat{\sigma}_e^2}{\hat{\sigma}_m^2 - \hat{r}_{em}} \right| \tag{6.27}$$

where the estimates denoted by a hat are again obtained using the exponential recursive weighting algorithm [19] [20]

$$\begin{aligned} \hat{r}_{em}[n] &= \lambda \hat{r}_{em}[n-1] + (1-\lambda) e[n] m[n] \\ \hat{\sigma}_m^2[n] &= \lambda \hat{\sigma}_m^2[n-1] + (1-\lambda) m^2[n] \\ \hat{\sigma}_e^2[n] &= \lambda \hat{\sigma}_e^2[n-1] + (1-\lambda) e^2[n]. \end{aligned} \tag{6.28}$$

The proposed echo-path change detector works as follows:

1. When $\xi_{AsifEPD} > T_{EPD}$ (T_{EPD} is a properly chosen threshold close to zero), it is declared that the echo-canceller has not converged i.e. echo-path has changed and the adaptation is enabled.

2. Whenever $\xi_{AsifEPD} < T_{EPD}$, the detector decides that the echo-canceller has converged i.e. there are no echo-path variations.

Simulations demonstrate the efficiency of the proposed algorithm for echo-path change detection, even variations in filter coefficients by 2dB are detected as shown in Figure 6.3.

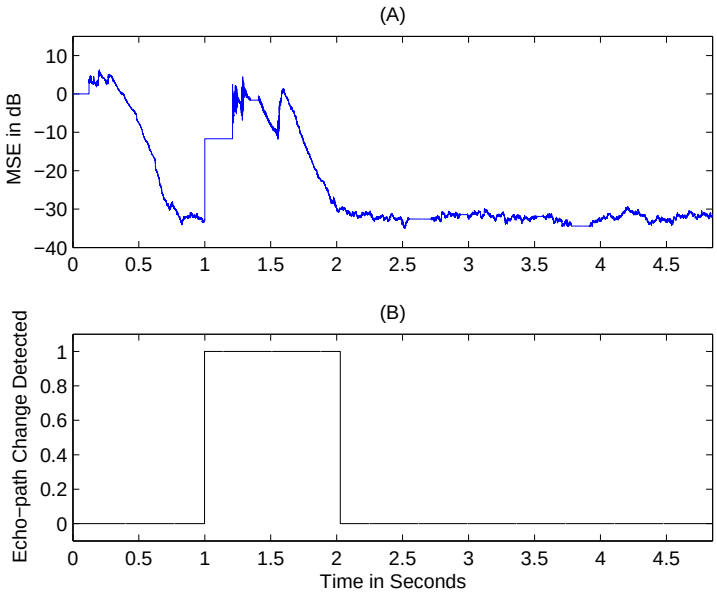

Figure 6.3: Normalized mean square error in the filter coefficients during echo-path change for the robust FRLS using the proposed double-talk detector.

The proposed detection statistic is given by equation 6.25 and is computed using equation 6.27, whereas the conventional echo-path change detection statistic is given by equation 6.10 and is computed by [17]:

$$\xi_{Conventional}[n] = \frac{\sum_{i=0}^{l-1} |C_i[n]|}{l} \tag{6.29}$$

where $C_i[n]$ is the cross-correlation coefficient between $x[n-i]$ and $e[n]$, l is the number of taps of the adaptive filter. The cross-correlation coefficients are updated using an exponential

recursive weighting algorithm [19] [20]:

$$\begin{aligned}
\hat{\sigma}_e^2[n] &= \lambda\hat{\sigma}_e^2[n-1] + (1-\lambda)e^2[n] \\
\hat{\sigma}_i^2[n] &= \lambda\hat{\sigma}_i^2[n-1] + (1-\lambda)x^2[n-i] \\
r_{e,i}[n] &= \lambda r_{e,i}[n-1] + (1-\lambda)e[n]x[n-i] \\
C_i[n] &= \frac{r_{e,i}[n]}{\sigma_e[n]\ \sigma_i[n]}, \quad i = 0, 1, ..., (l-1).
\end{aligned} \tag{6.30}$$

Clearly the proposed algorithm is very simple and computationally very efficient. Requiring only 9 operations/sample as compared to $(6l + 4)$ operations/sample for the conventional test statistic.

6.7. MONO-CHANNEL AEC IMPLEMENTATION

A mono-channel FRLS algorithm is implemented using Table 6.3 of reference [9] and successfully applied to the problem of AEC using the proposed algorithms for double-talk and echo-path change detection. The algorithm is altered using the proposed decision statistics in order to handle the problems of an AEC in a better way. For handling double-talk, the following steps are performed [16]:

1. Once double-talk is detected i.e. $\xi_{AsifDTD} < T_{DTD}$, it is declared for a minimum period of time t_{hold_1} and during this period the filter adaptation is disabled.

2. If the decision statistic $\xi_{AsifDTD} \geq T_{DTD}$ (i.e. no double-talk) continuously for an interval of t_{hold_1} seconds, the filter resumes adaptation. The comparison of $\xi_{AsifDTD}$ to T_{DTD} continues and double-talk is declared again when $\xi_{AsifDTD} < T_{DTD}$.

A major concern when an echo-path change occurs, is that the false alarm of a double-talk detector increases. This reduces the convergence rate of an AEC allowing annoying echo to persist. However, if an echo-path change detector is employed double-talk false alarms can be detected and the convergence rate of an AEC can be increased significantly. Thus, for handling echo-path changes, the following steps are incorporated:

3) If the echo-path change detection statistic $\xi_{AsifEPD} > T_{EPD}$ continuously for an interval of t_{hold_2} seconds ($t_{hold_2} > t_{hold_1}$), it is declared that the echo-path has changed and adaptation continues for a period of t_{hold_3} seconds ($t_{hold_3} > t_{hold_2}$) irrespective of the double-talk flag.

By following the above mentioned steps, the problems of double-talk and echo-path variations in an AEC are efficiently handled. For the purpose of showing how the system

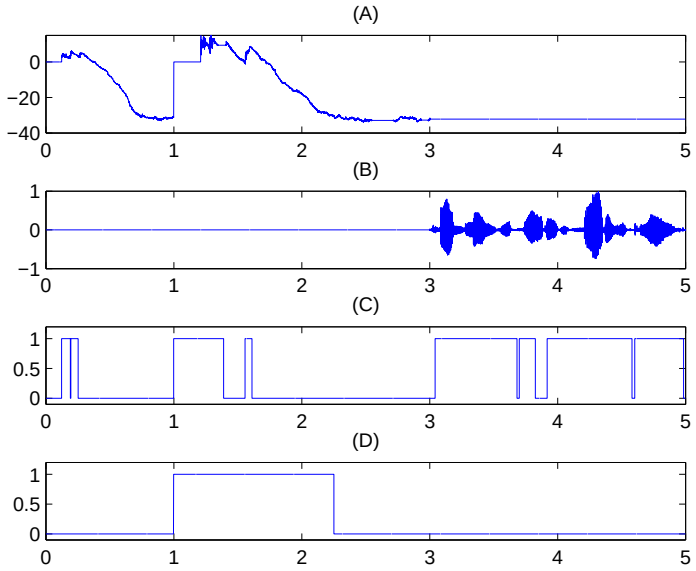

Figure 6.4: Normalized mean square error in the filter coefficients during echo-path change and double-talk situation for the robust FRLS using the proposed double-talk and echo-path change detectors.

handles the echo-path variations and double-talk situations, a special case is chosen. Echo-path variations were created after 1 second by increasing the echo-path gain by 6 dB, and the double-talk is introduced after 3 seconds by adding the near-end speech. The recorded digital speech sampled at 16 KhZ is used as the far-end speech $\mathbf{x}$ and a measured room impulse response of a $10' \times 10' \times 8'$ room is used as the loudspeaker microphone environment $\mathbf{h}$. The far-end speech signal is filtered through the measured echo-path $\mathbf{h}$ to create the echo signal y, near-end speech signal as shown in Figure 6.4 is added to the echo signal y to get the microphone signal m. Near-end speech is added such that the echo to background ratio is set to 0 dB ($EBR = \sigma_y^2/\sigma_v^2 = 1$ (0 dB)). It can be observed in Figure 6.4, that the echo-path changes are detected and the near-end speech is detected as well by the double-talk detector as shown in Figure 6.4. It can be further observed that, 6 dB increase in echo-path gain is declared as double-talk, but the echo-path change detector detects these false alarms and the adaptation is enabled after a period of t_{hold_2} seconds after the echo-path change occurs.

Robust FRLS parameters are set/initialized according to [9]. It can be observed from Figure 6.4 that the convergence rate of an AEC is increased as compared to the convergence rate of a Robust FRLS using Benesty's double-talk detector. The system converges in approximately 1 second whereas in the latter case it takes 2 seconds to converge. This increase in convergence rate is achieved, because the false alarms of the double-talk detector are detected by the proposed echo-path change detector.

6.8. SIMULATION RESULTS

In this section, the performance of the proposed algorithms for double-talk and echo-path change detection and the performance of AEC during double-talk and echo-path variations is evaluated.

6.8.1. Evaluation of the Proposed Double-talk Detector. Receiver operation characteristics (R.O.C) i.e. probability of detection (P_D) versus probability of false alarm (P_f) are typically used to evaluate any detectors. A similar approach is employed to evaluate the proposed double-talk detector, the performance is characterized in terms of probability of miss (P_m) as a function of near-end to far-end ratio (NFR) under a probability of false alarm constraint [16]. This approach is chosen because for the AEC application, the penalty of false alarm is small because it simply halts the adaptation for a duration of t_{hold_1} seconds. When the AEC has converged, freezing adaptation does not perturb the performance whereas while converging this increases the convergence time. Further, some of the false alarms are detected by the echo-path change detector particularly the important ones, when the filter is converging and thereby not deteriorating the convergence rate during echo-path variations. The probability of miss is the probability of not detecting near-end speech when it is present, therefore a smaller value of P_m indicates better performance. The probability of miss characteristics of the proposed and the conventional double-talk detector are shown in Figure 6.5. It is clear that the proposed double-talk detector significantly outperforms the conventional algorithm. It should be noted that the Benesty's performance is exactly similar to the proposed double-talk detector. It has been shown in Section 6.5, that the detection statistics converge theoretically and Figure 6.2 illustrates the convergence.

6.8.2. Evaluation of the Proposed Echo-path Change Detector and AEC Sensitivity to Echo-path Variations. The proposed echo-path change detector can be written as in equation 6.26 i.e.

$$\xi_{AsifEPD} = \left| \frac{(\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}} \hat{\mathbf{h}}}{\mathbf{h^T} R_{\mathbf{xx}} \hat{\mathbf{h}}} \right|. \tag{6.31}$$

It is clear that $\xi_{AsifEPD} > 0$ for $\mathbf{h} \neq \hat{\mathbf{h}}$. It has been observed in [6] that the proposed detection statistic detects any echo-path variations. To show how the AEC handles the echo-path variations, two basic changes in the echo-path that may occur are chosen:

1. Increase in echo-path gain (6 dB i.e. $\mathbf{h} \rightarrow 2\mathbf{h}$) and
2. Decrease in echo-path gain (-6 dB i.e. $\mathbf{h} \rightarrow 0.5\mathbf{h}$).

In each simulation, the echo-path changes after the AEC has converged. It can be observed from Figures 6.6 and 6.7 that these variations are detected by the proposed algorithm and

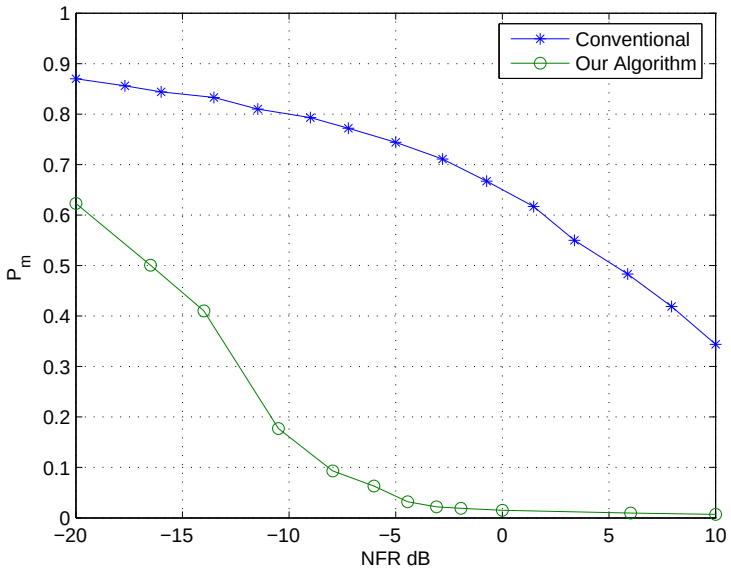

Figure 6.5: P_m as a function of NFR for the proposed and the conventional double-talk detector under a constraint of $P_f = 0.1$.

the filter adaptation is enabled. It was pointed out in [9] that, when an echo-path change occurs the false alarm rate of a double-talk detector increases this hurts the performance of an AEC by increasing the convergence time. It can be observed from Figures 6.6 and 6.7 that the proposed echo-path change detector detects these false alarms and hence the convergence rate is increased.

6.8.3. AEC Sensitivity to Double-talk Situations. Objective here is to show that the proposed double-talk detector is appropriate for the acoustic case. For these simulations, the following data is used [9]:

- *Far-end:* A 5 second speech from a female talker is used as the far-end speech. The standard deviation of the signal is set at 1900 ($\sigma_{\mathbf{x}} = 1900$).
- *Near-end:* Two cases are considered:
  1. A two second speech from a female talker beginning after 2 seconds is used as a near-end speech.
  2. A two second speech from a male talker beginning after 3 seconds is used as a near-end speech. Again the standard deviation of both the signals is set at 1900.
- *Levels:* Echo to Background Ratio (EBR) is set at 0 dB and Echo to Noise Ratio (ENR) is set at 30 dB.

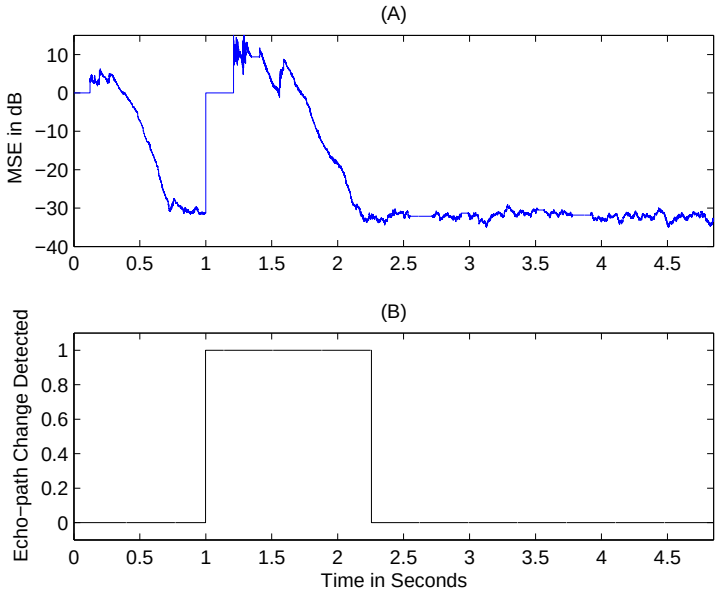

Figure 6.6: Mean square error during echo-path changes.

- *Echo-path:* A measured 8000 sample (500 msec) impulse response of a $10' \times 10' \times 8'$ room is used as the echo-path $\mathbf{h}$ and
- *FRLS parameters:* are set/initialized according to [9].

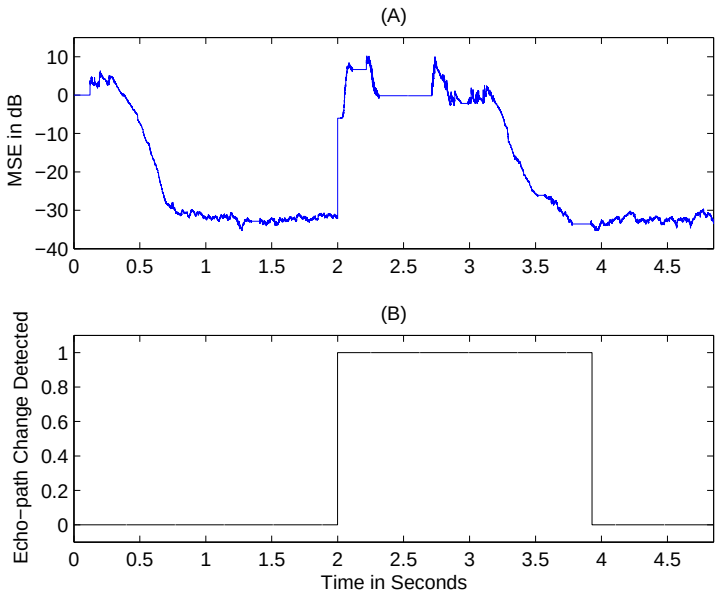

Figure 6.7: Mean square error during echo-path change.

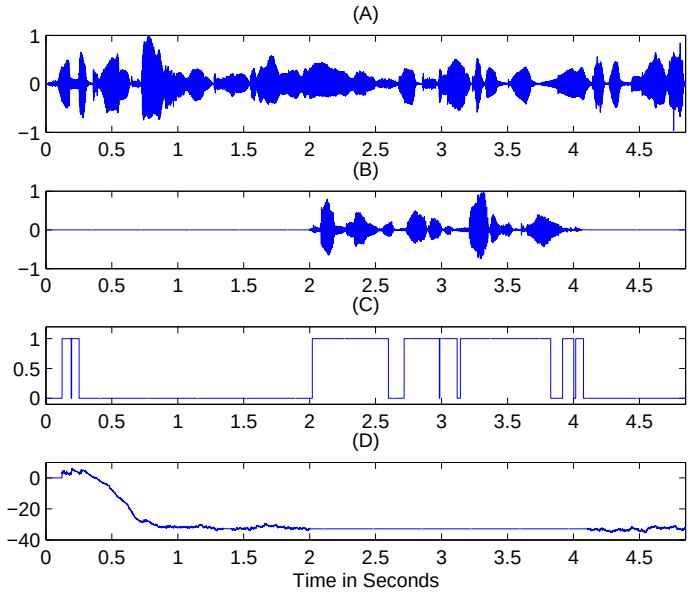

Figure 6.8: Double-talk situation of a robust FRLS using the proposed algorithms.

The threshold T_{DTD} of the proposed double-talk detector is chosen in such a way that the probability of miss $P_m \approx 0$ (too small to be reliably measured). In general, lower P_m is achieved at the cost of higher P_f. The penalty of false alarm is small as it simply halts the filter adaptation, where as low P_m is important to prevent divergence due to double-talk.

The performance is measured using the normalized mean square error in the filter coefficients, which is given by:

$$MSE[n] = 10\log_{10}\frac{|\mathbf{h} - \hat{\mathbf{h}}_n|^2}{|\mathbf{h}|^2}. \tag{6.32}$$

where $\hat{\mathbf{h}}_n$ are the adaptive filter coefficients at time n. Figures 6.8 and 6.9 show (A) far-end speech, (B) near-end speech, (C) Double-talk flag detecting near-end speech and (D) the mean square error performance of the robust FRLS using the proposed double-talk detector.

In both the cases, the near-end speech is detected and hence no divergence is observed in the mean square error. Hence, it is concluded that the proposed double-talk detector is absolutely suitable for acoustic case. With the aid of the proposed echo-path change detector, one can avoid the two-path method to handle the problems of an AEC efficiently and achieve similar and even better performance regardless of the environment (double-talk and/or echo-path variations) resulting in significant memory and computational savings.

6.9. CONCLUSION

A novel normalized cross-correlation based sample by sample double-talk detector is proposed. Next, a novel normalized sample by sample echo-path change detector is introduced.

Finally, the robust FRLS algorithm is combined with the proposed algorithms for double-talk and echo-path change detection to solve the problems that arise in an AEC in a better and efficient way. To summarize, the major advantages of the proposed double-talk and echo-path change detector are listed:

1. *Double-talk detector:*

   - Similar performance as compared to the best known existing technique, but with an order of magnitude improvement in computational complexity.
   - Independent of the echo-path variations, very desirable as the acoustic echo paths vary randomly.
   - Low added complexity when implemented with any adaptive algorithm such as FRLS, NLMS, etc.

2. *Echo-path detector:*

   - Detects any echo-path variations and is normalized appropriately i.e. the detection statistic is greater than zero only for echo-path variations.
   - Low added complexity when implemented with any adaptive algorithm.
   - Independent of the near-end speech/doubletalk.

The double-talk and echo-path change detector complement each other to handle the problems of an AEC in a better way. The threshold (T_{DTD}) of the proposed double-talk

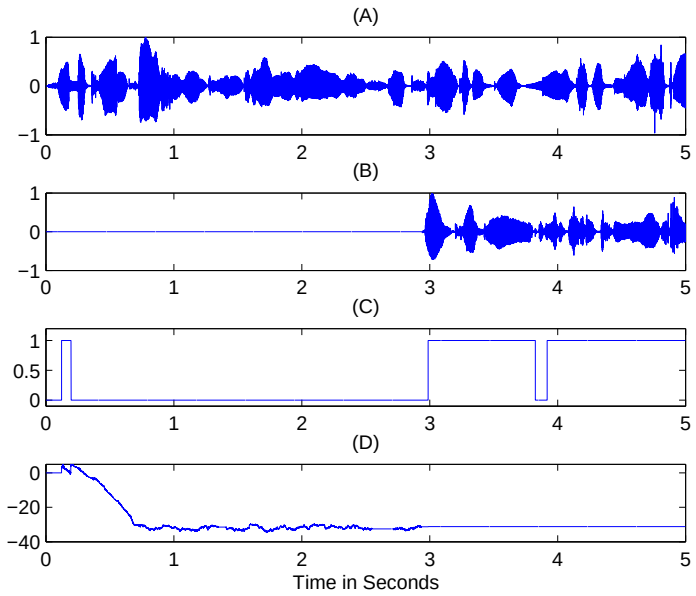

Figure 6.9: Double-talk situation of a robust FRLS using the proposed algorithms.

detector is chosen in such a way that the probability of miss is minimized ($P_m \approx 0.001$ too small to be reliably measured), this is achieved at an increased false alarm ($P_f \approx 0.1$) but most of these false alarms are detected by the proposed echo-path change detector. Hence, the AEC has an excellent double-talk interference protection as the probability of miss of the double-talk detector is close to zero $P_m \approx 0.001$ and the convergence rate is also increased as the false alarms of the double-talk detector during echo-path variations are detected by the proposed echo-path change detector.

7. A NOVEL DOWNLOAD TEST FOR TWO PATH ECHO CANCELLER

7.1. INTRODUCTION

In an echo-canceller, the adaptive filters are used to cancel the echo. The adaptive filters used are finite impulse response (FIR) filters that continuously adjust/adapt their coefficients in an attempt to predict the echo coming from the acoustic coupling/network based on the excitation signal. The acoustic case is considered here, but the same ideas can be easily extended to the LEC. The adaptive filter coefficients diverge from the true echo-path if the adaptation is continued during periods of near-end speech/signal. A double-talk detector is used to freeze the adaptation during periods of near-end speech/signal. Dynamic suppression and non linear processor (NLP) further attenuate the residual echo that leaves the adaptive filter. The suppressor and the NLP create an annoying modulation of the background noise. To abate this effect, comfort noise is added at the output [9].

In general, the effectiveness of these elements can be enhanced if they are implemented within a sub-band structure. The computational complexity of the adaptive filters decreases linearly with the number of sub-bands. Also, the adaptive filters tend to converge faster, the dynamic suppressor and NLP are less disturbing, and the comfort noise can easily be spectrally shaped to the ambient surroundings [9]. Figure 7.1 shows the arrangement of these elements.

Divergence due to double-talk can be alleviated by using the so-called two-path echo canceller where there are two sets of filters, background and foreground that predict the echo. The background filters almost always adapt their coefficients, regardless of double-talk. The foreground filters periodically receive their coefficients from the background filters when a series of tests indicate that it is favorable to do so. Only the error signals of the foreground filters are returned to the user. This allows the background filters to diverge during double-talk without affecting the observed performance of the system. Normalized least mean square (NLMS) based adaptive filters are used as the background adaptive filters. The two-path structure is shown in Figure 7.2.

This section is structured as follows. In Section 7.2, the conventional download tests for two-path echo cancellation are given. The novel optimal download test is introduced in Section 7.3. Next, a comprehensive study on the proposed download test is done in Section 7.4 which is followed by a summary and conclusions in Section 7.5.

7.2. TWO-PATH METHOD DOWNLOAD TESTS

The key to a good two-path AEC performance lies in the definitions of the download tests. Typical tests include but are not limited to a double-talk detector, echo return loss enhancement (ERLE) measure and more as described below. First, the following frame-based energy measures are defined:

- σ_x^2 is the far-end excitation signal energy.
- σ_m^2 is the microphone signal energy.
- σ_{ef}^2 is the fore-ground error energy and
- σ_{eb}^2 is the background error energy.

The download tests are defined as follows:

1. Is $\sigma_x^2 > T_1$? That is, is there sufficient excitation energy?

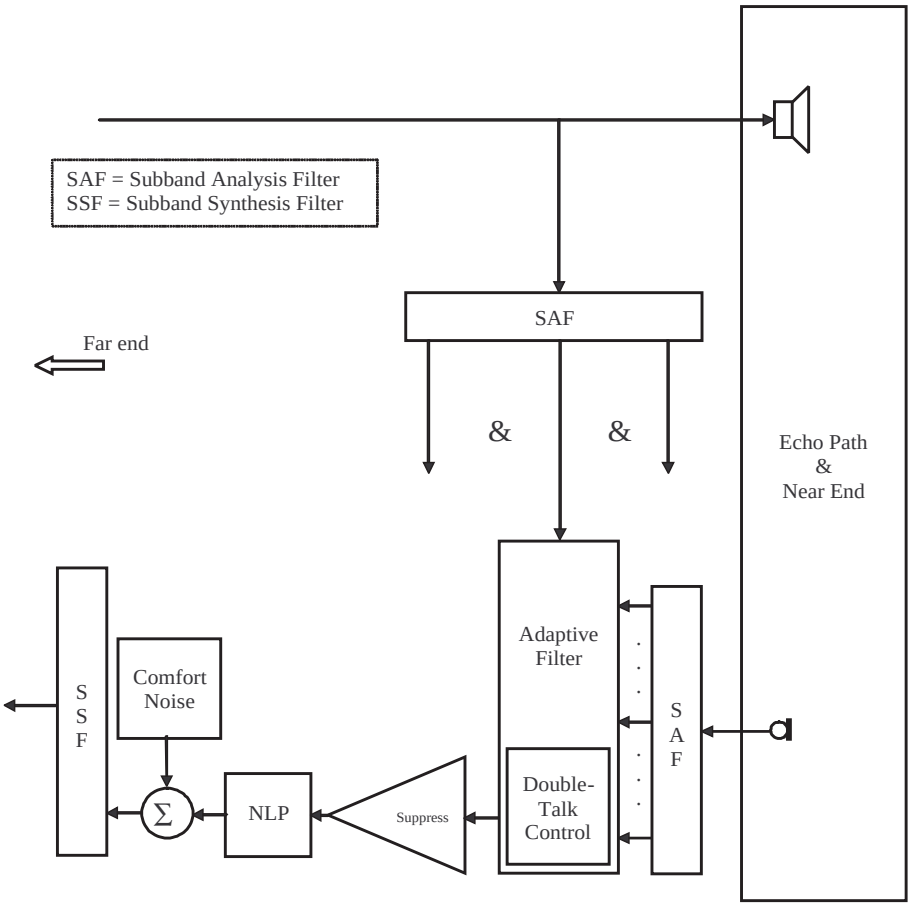

Figure 7.1: Complete AEC model.

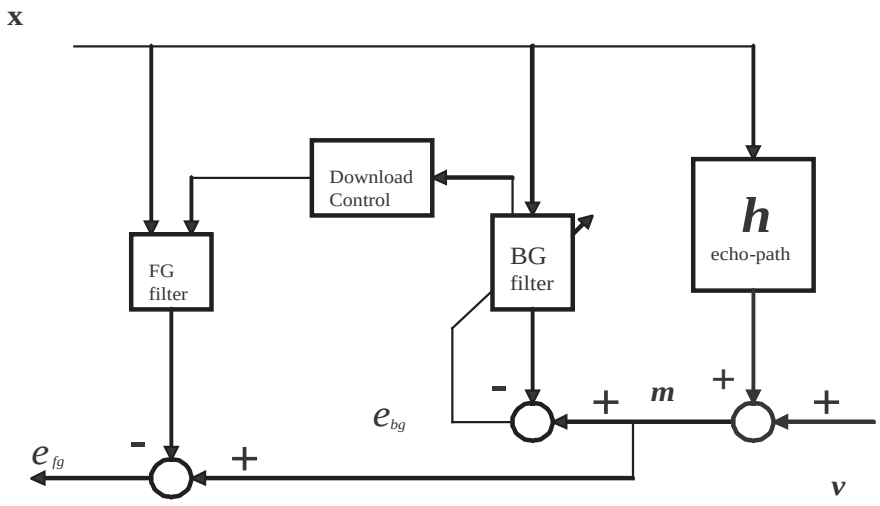

Figure 7.2: Two-path AEC model.

2. Is $\sigma_m^2 > T_2$? That is, is there sufficient signal at the microphone?

3. Is $\frac{\sigma_{ef}^2}{\sigma_{eb}^2} > T_3$? That is, is the background error energy smaller than the foreground error energy by a factor of T_3?

4. Is $\frac{\sigma_m^2}{\sigma_{eb}^2} > T_4$? That is, is the ERLE greater than the factor T_4? and

5. Is $\xi_{DTD} > T_5$? That is, is the microphone signal corrupted by near-end speech? Here, a novel normalized cross-correlation based double-talk detector proposed in [4] is used, the double-talk decision statistic is compared to a pre-selected threshold $T_5 \approx 1$.

When all these tests are passed say in three consecutive frames, then the background filter coefficients are downloaded. The idea behind these tests is as follows.

There is no reason to adapt when there is insufficient excitation/microphone signal energy. The first two tests address this problem. Third test guarantees that the background coefficients that give greater error energy than those in the foreground are not downloaded [9]. The idea behind the fourth test is that, if the background filter is giving minimum required ERLE then there is no reason to inhibit the downloading process. Further, during double-talk the background filter coefficients will diverge from the true echo-path impulse response by trying to drive its error signal to zero. However, since the far-end and near-end signals are uncorrelated the background filter will fail to make the error signal much smaller than the microphone signal. So the ratio of error to the microphone signal energy in this case will be near zero dB thereby inhibiting the download process [9]. Finally, in the last double-talk test

it is checked whether the microphone signal is corrupted by near-end speech or not? If so, it is concluded that the background coefficients are diverged due to double-talk and downloading in inhibited.

These are the typical download tests, that are used in a two-path echo canceller. The basic idea behind the download tests is to use the better converged filter among the background and the foreground filters. An optimal download test statistic should measure the convergence of both the filters, thereby recommending the better converged filter. Next, a novel download test that explicitly measures the convergence of the adaptive filters is proposed.

7.3. NOVEL DOWNLOAD TEST

In this section, a novel download test is derived, which is a direct measure of the adaptive filter's convergence. Referring to Figure 7.3, the cross-correlation between the microphone signal m, and the cancellation error e is given by:

$$\begin{aligned} r_{em} &= E[em] \\ &= E[(y + v - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x})(y + v)] \\ &= E[(\mathbf{h}^{\mathbf{T}}\mathbf{x} - \hat{\mathbf{h}}^{\mathbf{T}}\mathbf{x} + v)(\mathbf{h}^{\mathbf{T}}\mathbf{x} + v)] \\ &= (\mathbf{h} - \hat{\mathbf{h}})^T E[\mathbf{x}\mathbf{x}^{\mathbf{T}}]\mathbf{h} + E[v^2] \\ &= (\mathbf{h} - \hat{\mathbf{h}})^T R_{\mathbf{xx}}\mathbf{h} + \sigma_v^2 \end{aligned} \tag{7.1}$$

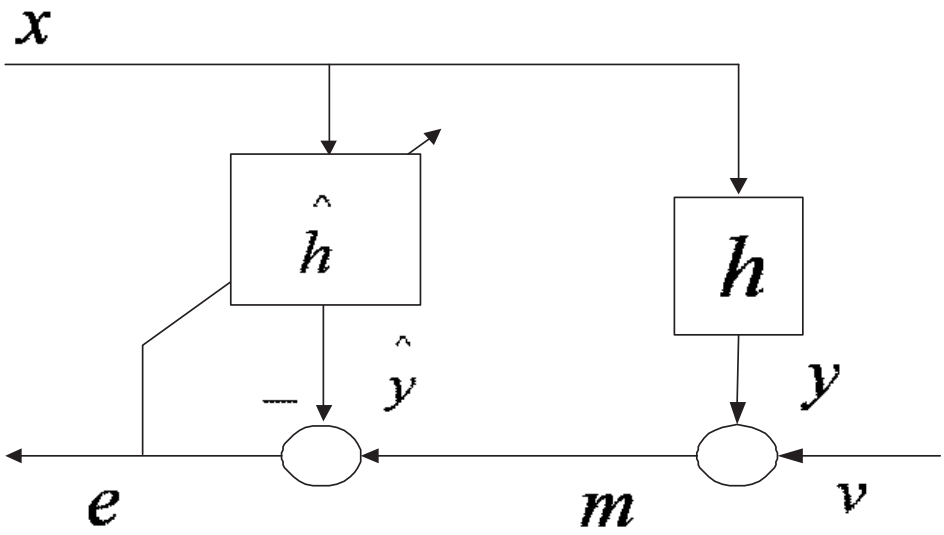

Figure 7.3: Basic AEC model.

where $E[\bullet]$ denotes the mathematical expectation and $R_{\mathbf{xx}} = E[\mathbf{x}\mathbf{x}^{\mathbf{T}}]$, $\mathbf{h}$ is the true echo-path, $\hat{\mathbf{h}}$ is an estimate of the echo-path and σ_v^2 is the variance of the near-end speech. The far-end speech vector $\mathbf{x}$ and the near-end signal v are independent and are assumed to be of zero mean. Variance of the microphone signal is given by:

$$\begin{aligned} \sigma_m^2 &= E[m^2] = E[(y+v)^2] \\ &= E[y^2] + E[v^2] = E[\mathbf{h}^{\mathbf{T}}\mathbf{x}(\mathbf{h}^{\mathbf{T}}\mathbf{x})^T] + \sigma_v^2 \\ &= \mathbf{h}^{\mathbf{T}} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2. \end{aligned} \tag{7.2}$$

and, finally, the variance of the cancellation error e is given by:

$$\begin{aligned} \sigma_e^2 &= E[e^2] \\ &= E[((\mathbf{h}-\hat{\mathbf{h}})^T\mathbf{x} + v)((\mathbf{h}-\hat{\mathbf{h}})^T\mathbf{x} + v)] \\ &= (\mathbf{h}-\hat{\mathbf{h}})^T E[\mathbf{x}\mathbf{x}^T](\mathbf{h}-\hat{\mathbf{h}}) + E[v^2] \\ &= (\mathbf{h}-\hat{\mathbf{h}})^T R_{\mathbf{xx}}(\mathbf{h}-\hat{\mathbf{h}}) + \sigma_v^2 \end{aligned} \tag{7.3}$$

The new test statistic is defined to be

$$\xi = \left| \frac{r_{em} - \sigma_e^2}{\sigma_m^2 - r_{em}} \right| \tag{7.4}$$

substituting equations 7.1 , 7.2 and 7.3 in 7.4 yields:

$$\begin{aligned} \xi &= \left| \frac{(\mathbf{h}-\hat{\mathbf{h}})^T R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2 - (\mathbf{h}-\hat{\mathbf{h}})^T R_{\mathbf{xx}}(\mathbf{h}-\hat{\mathbf{h}}) - \sigma_v^2}{\mathbf{h}^{\mathbf{T}} R_{\mathbf{xx}} \mathbf{h} + \sigma_v^2 - (\mathbf{h}^{\mathbf{T}} - \hat{\mathbf{h}}^{\mathbf{T}}) R_{\mathbf{xx}} \mathbf{h} - \sigma_v^2} \right| \\ &= \left| \frac{(\mathbf{h}-\hat{\mathbf{h}})^T R_{\mathbf{xx}} \hat{\mathbf{h}}}{\mathbf{h}^{\mathbf{T}} R_{\mathbf{xx}} \hat{\mathbf{h}}} \right| \end{aligned} \tag{7.5}$$

it can be observed from equation 7.5, for $\mathbf{h} \approx \hat{\mathbf{h}}$, $\xi \approx 0$ and for $\mathbf{h} \neq \hat{\mathbf{h}}$, $\xi > 0$. Thus the proposed statistic is a good measure of the adaptive filter's convergence.

The proposed algorithm is computationally very efficient, as only 3 multiplications, 3 additions, 2 subtractions and a division are required to compute the decision statistic at each sample (i.e. 9 operations per sample). First, the following measures are defined:

$$\xi_{BG} = \left| \frac{r_{e_{bg}m} - \sigma_{e_{bg}}^2}{\sigma_m^2 - r_{e_{bg}m}} \right| \tag{7.6}$$

$$\xi_{BG} = \left| \frac{(\mathbf{h} - \hat{\mathbf{h}}_{bg})^T R_{\mathbf{xx}} \hat{\mathbf{h}}_{bg}}{\mathbf{h^T} R_{\mathbf{xx}} \hat{\mathbf{h}}_{bg}} \right| \tag{7.7}$$

where $r_{e_{bg}m}$ is the maximum cross-correlation between the background error and the microphone signal, $\sigma^2_{e_{bg}}$ is the background error variance and $\hat{\mathbf{h}}_{bg}$ is the background coefficient vector and

$$\xi_{FG} = \left| \frac{r_{e_{fg}m} - \sigma^2_{e_{fg}}}{\sigma^2_m - r_{e_{fg}m}} \right| \tag{7.8}$$

$$= \left| \frac{(\mathbf{h} - \hat{\mathbf{h}}_{fg})^T R_{\mathbf{xx}} \hat{\mathbf{h}}_{fg}}{\mathbf{h^T} R_{\mathbf{xx}} \hat{\mathbf{h}}_{fg}} \right|. \tag{7.9}$$

where $r_{e_{fg}m}$ is the maximum cross-correlation between the foreground error and the microphone signal, $\sigma^2_{e_{fg}}$ is the foreground error variance and $\hat{\mathbf{h}}_{fg}$ is the foreground coefficient vector. The new download test is defined as: If

$$\xi_{BG} < \xi_{FG} \tag{7.10}$$

continuously say for five frames. Then, it is concluded that the background filter is better converged than the foreground and hence, the background coefficients are passed onto the foreground filter. Further, during double-talk background coefficients diverge from the true echo-path, making $\xi_{BG} > \xi_{FG}$. Hence, it can be concluded that the proposed test is immune to double-talk and simulations further demonstrate this behavior.

7.4. SIMULATION RESULTS

The values for $r_{e_{bg}m}$, $r_{e_{fg}m}$, $\sigma^2_{e_{bg}}$, $\sigma^2_{e_{fg}}$ and σ^2_m in equations 7.6 and 7.8 are exact and not available in practice. As a result, the final statistic is given by

$$\xi_{BG} = \left| \frac{\hat{r}_{e_{bg}m} - \hat{\sigma}^2_{e_{bg}}}{\hat{\sigma}^2_m - \hat{r}_{e_{bg}m}} \right| \tag{7.11}$$

and

$$\xi_{FG} = \left| \frac{\hat{r}_{e_{fg}m} - \hat{\sigma}^2_{e_{fg}}}{\hat{\sigma}^2_m - \hat{r}_{e_{fg}m}} \right| \tag{7.12}$$

where the estimates denoted by a hat are obtained using the exponential recursive weighting algorithm, [19] [20]:

$$\hat{r}_{e_{bg}m}[n] = \lambda \hat{r}_{e_{bg}m}[n-1] + (1-\lambda) e_{bg}[n] m[n]$$

$$
\begin{aligned}
\hat{r}_{e_{fg}m}[n] &= \lambda \hat{r}_{e_{fg}m}[n-1] + (1-\lambda) e_{fg}[n] m[n] \\
\hat{\sigma}^2_m[n] &= \lambda \hat{\sigma}^2_m[n-1] + (1-\lambda) m^2[n] \\
\hat{\sigma}^2_{e_{bg}}[n] &= \lambda \hat{\sigma}^2_{e_{bg}}[n-1] + (1-\lambda) e^2_{bg}[n] \\
\hat{\sigma}^2_{e_{fg}}[n] &= \lambda \hat{\sigma}^2_{e_{fg}}[n-1] + (1-\lambda) e^2_{fg}[n]
\end{aligned} \tag{7.13}
$$

where λ is the exponential weighting factor.

The complete AEC model shown in Figure 7.1 was simulated, including the NLP with the two-path method working independently in each sub-band. First, the immunity of the new download test towards double-talk is tested by creating different double-talk situations as shown in Figures 7.4 and 7.5. In Figure 7.4, double-talk situations are created at two

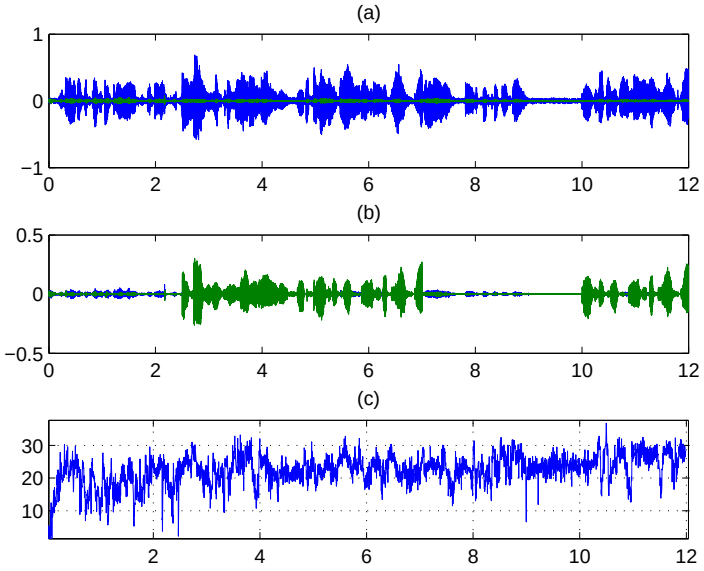

Figure 7.4: Near-end speech is introduced at the microphone from 2.5 to 7 seconds and for the last two seconds.

different instances. First, a double-talk situation is created after 2.5 seconds all the way till 7 seconds and then again the near-end speech is introduced for the last two seconds. In the uppermost plot, the microphone signal and the residual echo without the near-end speech are shown. This was done to observe divergence due to double-talk. None is observed. At the center, the microphone signal and the residual echo leaving the system with the near-end speech are shown, and finally at the bottom the ERLE without the NLP is plotted. It can be observed that there was absolutely no divergence due to double-talk. Also, no undesirable

artifacts were observed in listening during and after the periods of double-talk. The near-end speech was left unscathed.

Next, a near-end tone is introduced as shown in Figure 7.5. Here a composite source signal (CSS)is used as the excitation vector, CSS consists of different sequences including voiced and unvoiced sounds as well as pauses. The near-end tones were detected and adaptation was inhibited in the corresponding bands and no divergence was observed in Figure 7.5a and hence there is no degradation in the ERLE (Figure 7.5c). Based on these results, it can be concluded that the proposed download test is immune to near-end speech/signal.

The standard International Telecommunication Union (ITU) G.168 tests, with and without the proposed download test were performed. Significant improvement was observed in the various tests as tabulated in Table 7.1. An improvement of 5-16 dBm0 is observed in the first five tests, and no degradation is observed in any of the remaining tests. Based on these results, it is concluded that the proposed download test helps improve the overall performance of the system.

7.5. CONCLUSIONS

The key to a good two-path method performance lies in the definitions of the download test. In this section, a novel download test which is a direct measure of the adaptive filter's convergence is proposed. Significant improvement in the overall performance of the system was observed with the aid of the proposed novel download test. Further, no deterioration

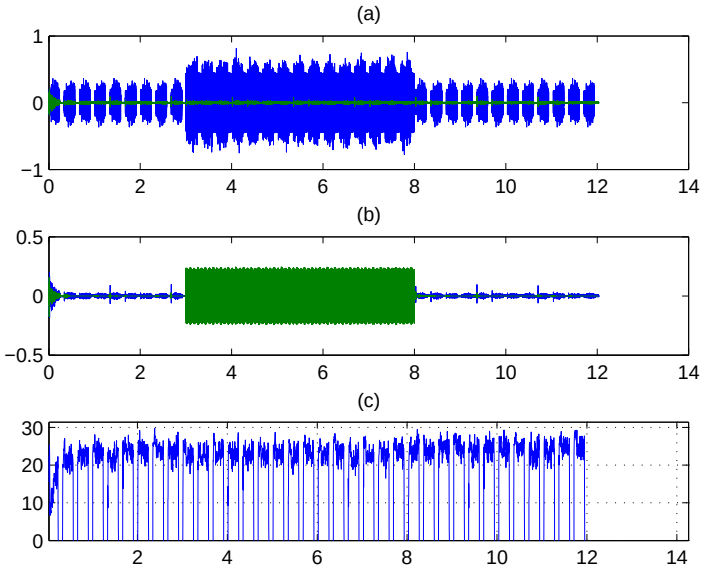

Figure 7.5: A 12-second composite source signal is used as the far-end excitation, near-end tone is introduced at the microphone from 3 to 8 seconds.

in the performance was observed during and after periods of double-talk and no undesirable artifacts were detected in listening.

Table 7.1: Various ITU G.168 tests with and without the proposed download test.

ITU-G.168 tests.	With out the new proposed download test.	With the new proposed download test.
Test 2: Convergence Requirement: -65dBm0	NLP enabled:-71.9 dBm0	NLP enabled:-Infty dBm0
Test 2b: Re-convergence Requirement: -65dBm0	NLP enabled:-65.5 dBm0	NLP enabled: -75.2 dBm0
Test 2c: Convergence in the presence of background noise. Requirement: -60dBm0	NLP enabled: -63.7 dBm0	NLP enabled: -68.1 dBm0
Test 3a: Double-talk convergence with low cancelled end levels. Requirement: -65dBm0	NLP enabled: -67.7 dBm0	NLP enabled: -83.2 dBm0
Test 3b: Double-talk tests. Requirement: -40dBm0	During Double -talk: -46.4 dBm0	During Double -talk: -57.7 dBm0
Test 3c Double-talk tests.	No divergence at all, no undesirable artifacts due to double-talk.	No divergen ce at all, no undesirable artifacts due to double-talk.
Test 4: Leak rate test. Requirement: -40dBm0	NLP disabled: -53.4 dBm0	NLP disabled: -54.2 dBm0
Test 6: Narrow band test. After immediate application of tones, with NLP disabled and adaptation frozen. Requirement: --40dBm0	-54.09 dBm0.	-56.04dBm0.

8. SUMMARY OF CONTRIBUTIONS

- Double-talk detection for acoustic/line echo cancellation.
  - Developed three different techniques for double-talk detection.

    Designed a novel frame-level double-talk detector using novel RTRL based frequency domain speech detectors.

    Designed a new hybrid frame-level double-talk detector using novel RTRL based frequency domain speech detectors and a cross-correlation measure.

    Designed a novel sample by sample normalized, cross correlation based double-talk detector that outperforms the best existing algorithms. This also has a low added complexity when implemented with any adaptive algorithms such as FRLS, NLMS, etc.
  - Implemented a frequency domain double-talk detector based on the microphone and AEC residual cross correlation. Extended the idea of this proposed double-talk detector to the multi-channel case.
- Echo-path change detector for acoustic/line echo cancellation.
  - Formulated an optimal echo-path change detector that meets the needs of an optimal echo-path change detector and detects any echo-path variations. This further aids the double-talk detector in detecting some of it's false alarms to improve the overall performance of the system.
- Acoustic/Line echo canceller.
  - Realized a robust fast recursive least squares (FRLS) based echo-canceller for acoustic case. Combined it with the proposed techniques for double-talk and echo-path change detection to solve the problems that arise in an AEC in a better way.
  - Implemented a sub-band based, two-path echo-canceller using normalized least mean square (NLMS) adaptive algorithm for line / network echo cancellation. Designed a novel download test that improved the overall performance of the two-path system.

BIBLIOGRAPHY

[1] A. C. Surendran, S. Sukittanon, and J. Platt, "Logistic discriminative speech detectors using posterior snr," in *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, pp. 625–628.

[2] M. Asif, J. W. Stokes, J. C. Platt, A. Surendran, and S. L. Grant, "Doubletalk detection using real time recurrent learning," in *International Workshop on Acoustic Echoes and Noise Control*, Paris, France, September 2006.

[3] J. Benesty, D. R. Morgan, and J. H. 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.

[4] M. A. Iqbal, J. W. Stokes, and S. L. Grant, "A new class of double-talk detectors," in *Proceedings of the 2007 IEEE International Conference on Multi-media and Expo*, Beijing, China, July 2007.

[5] M. A. Iqbal, S. L. Grant, and J. Stokes, "A frequency domain double-talk detector based on cross-correlation and extension to multi-channel case," in *Submitted: Signal Processing Journal, Elsevier*, 2007.

[6] M. A. Iqbal and S. L. Grant, "A novel normalized cross-correlation based echo-path change detector," in *Proceedings of the 2007 IEEE Region 5 Conference*, Fayetteville, Arkansas, April 2007.

[7] M. A. Iqbal, S. L. Grant, and J. Stokes, "Simple and efficient solutions to aec problems," *Submitted: IEEE Transactions on Speech and Audio Processing*, April 2007.

[8] M. A. Iqbal and S. L. Grant, "Novel and efficient download test for two-path echo canceller," in *Proceedings of the 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, New York, October 2007.

[9] J.Benesty, T. Gansler, D. Morgan, M. Sondhi, and S. Gay, *Advances in Network and Acoustic Echo Cancellation*. New York: Springer, Inc., 2001.

[10] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice Hall, 1991.

[11] R. J. Williams and D. Zipser, "Experimental analysis of real-time recurrent learning algorithm," in *Connection Science,Vol 1, No 1*, 1989, pp. 87–111.

[12] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press.

[13] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," in *Signal Processing 81*, 2001, pp. 2403–2418.

[14] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of the 7th European Signal Processing Conference*, Edinburgh, Scotland, September 1994, pp. 1182–1185.

[15] D. Pearce and H. Hirsch, "The aurora experimental framework," in *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, October 2000, pp. 16–20.

[16] 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.

[17] H. Ye and B.-X. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Transactions on Communications*, vol. 39, pp. 1542–1545, November 1991.

[18] R. Cutler, "The distributed meetings system," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, China, April 2003, pp. 756–759.

[19] B. Porat, "Second-order equvalence of rectangular and exponential windows in least-squares estimation of autoregressive processes," *IEEE Transactions on Acoustics,Speech and Signal Processing*, 1985.

[20] Y. Hua, "Adaptive filter theory and applications," Ph.D. dissertation, South-East University, Tiangsu, China, March 1989.

[21] P. Eneroth, J. Benesty, T. Gansler, and S. Gay, "Comparision of different adaptive algorithms for stereophonic acoustic echo cancellation," in *Proceedings of EUSIPCO*, 2000, pp. 1835–1837.

[22] T. Gansler and J. Benesty, "A frequency-domain doubletalk detector based on a normalized cross-correlation vector," *Signal Processing*, vol. 81, pp. 1783–1787, August 2001.

[23] S. Gay and J. (Eds), *Acoustic Signal Processing for Telecommunication.* Boston: Kluwer Academic Publishers, 2000.

[24] D. Mansour and J. A.H. Gray, "Unconstrained frequency-domain adaptive filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-30, pp. 726–734, October 1982.

[25] J. Benesty and T. Gansler, "A multi-channel acoustic echo canceller doubletalk detector based on a normalized cross-correlation matrix," *Eur. Trans. Telecomm*, vol. 13, pp. 95–101, March-April 2002.

VITA

Mohammad Asif Iqbal was born on September 11, 1980 in Hyderabad, India. He received his B.S. degree in Electronics and Communication Engineering from Osmania University, Hyderabad in July 2001. He received his M.S. and Ph.D. degrees in Electrical Engineering from the University of Missouri-Rolla in December 2003 and August 2007, respectively. He worked as a Research Intern at Microsoft Research during his Ph.D. program under CPT. His primary responsibilities there included researching on various double-talk detection techniques. He designed a novel double-talk detector that out performed the best existing algorithms, this also reduced the echo transmitted during the voice communications over the video/audio conferencing devices. Four of his algorithms are also patented and he has almost a dozen publications in the related field. His current research interests include acoustic/line echo cancellation, double-talk detection, adaptive filtering techniques and full-surface communication.