A DOUBLE-TALK DETECTION ALGORITHM USING A PSYCHOACOUSTIC AUDITORY MODEL

Thien-An Vu[1], Heping Ding[2], and Martin Bouchard[1]

[1]School of Information Technology and Engineering, University of Ottawa
800 King Edward Avenue, Ottawa, Ontario, K1N 6N5, Canada
[2]Acoustic and Signal Processing Group, IMS, National Research Council
1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada
emails: tvu052@uottawa.ca, heping.ding@nrc-cnrc.gc.ca, bouchard@site.uottawa.ca

ABSTRACT

Successful adaptive echo cancellation in telecommunications depends on a control device called a double-talk (DT) detector. DT refers to the situation when signals from both ends of an echo cancellation system are simultaneously active. In the presence of a DT condition, the role of a DT detector is to prevent divergence of the adaptive filter in an echo cancellation system. This paper presents a novel double-talk detection (DTD) algorithm using a psychoacoustic auditory model. The model exploits the frequency masking properties of the human auditory system. It performs an analysis of the far-end signal and removes spectral components below a perceptual threshold, to create spectral holes without affecting the perceptual quality of the signal. A DT condition can be detected by monitoring the energy level in the created holes. Simulations with real speech data and comparisons with other DTD algorithms are presented to show the performance of the proposed algorithm.

1. INTRODUCTION

An echo canceller removes undesired echoes in a full-duplex telecommunications system. The cancellation is done by modeling the echo path with an adaptive filter and subtracting the echo estimate from the signal received at the near end, as depicted in **Figure 1.** The signals $x(n)$ and $v(n)$ represent the far-end and near-end speeches, respectively. The signals $y(n)$ and $\hat{y}(n)$ represent the echo generated by the actual echo path with impulse response $\underline{h}$ and the echo estimate made by the adaptive filter, respectively. The signal $e(n)$ denotes the residual error, which is transmitted to the far-end and is used to update the coefficient vector $\hat{\underline{h}}$ of the adaptive filter. $w(n)$ represents additive background noise.

When $v(n)$ is zero and the background noise $w(n)$ at the near-end is insignificant, the adaptive filter can converge to a good estimate of the echo path and can largely cancel the echo. However, when both $x(n)$ and $v(n)$ are not zero, i.e. a DT situation, $v(n)$ which acts as an uncorrelated noise to the adaptive algorithm can cause the adaptive filter to diverge and allow the undesired echo to pass through to the far-end. A common solution to prevent the divergence of the adaptive filter is to slow down or completely stop the filter adaptation in the presence of a near-end speech. This is determined by a DT detector.

The masking properties of the human auditory system have been widely exploited in many areas of research such as audio coding, digital watermarking, and speech enhancement. Masking refers to a psychoacoustic phenomenon where one sound is rendered inaudible because of the presence of another. One of the most important masking properties is the simultaneous (frequency) masking which occurs when two separate sounds are close enough in frequency; the stronger one covering up the other. In this paper, the frequency masking property of the human auditory system is used, based on a psychoacoustic auditory model, to create spectral holes in frames of the signal $x(n)$. By monitoring the energy of the signal $d(n)$ in the created holes, the presence of $v(n)$, i.e. a DT condition, can be detected.

This paper is organized as follows. Section 2 reviews some basics of a generic DTD algorithm and describes two typical DTD algorithms, which are later used for performance comparison against the proposed algorithm. Section 3 describes the proposed DTD scheme using a psychoacoustic auditory model proposed in [7]. Using real speech data, Section 4 evaluates the proposed DTD algorithm and compares it with other typical DTD algorithms, and Section 5 concludes the paper.

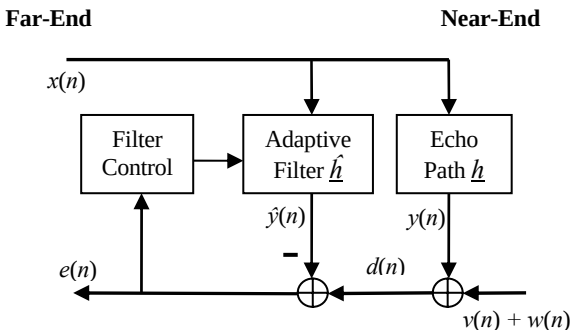

Figure 1: Echo Cancellation Configuration

2. BASICS OF A GENERIC DTD ALGORITHM

2.1 Basics

A common basic operation for most DTD algorithms involves the computation of a detection variable from the available signals such as the near-end $d(n)$, the far-end $x(n)$, the residual error $e(n)$, and/or the estimated filter weights $\underline{\hat{h}}$. The detection variable is then compared with a certain threshold. Depending on whether the detection variable is above or below the threshold, a decision is made on whether a DT condition is present or not. If the condition is declared, the filter adaptation is stopped or slowed down for a minimum period of hold time. When a non-DT condition lasts continuously for a period longer than the hold time, the filter can resume the adaptation until the next double-talk condition occurs. The hold time is necessary to suppress detection dropouts, because of the noisy behavior of the detection variable [5].

A special case that some DTD algorithms are struggling with is when there is a sudden change in the echo path, for example in an acoustic environment. This can often be falsely detected as a DT condition. This is a case where the adaptive filter really needs to adapt to the change in the echo path, and it is not desired for the adaptation to be turned off by the false alarm. Furthermore, background noise at the near-end should not be detected as a DT condition.

There are many DTD algorithms existing in the literature. They can mostly be classified into energy-based or correlation-based techniques. In the following subsection, we review the Geigel algorithm, which is an energy-based DTD algorithm, and the normalized cross correlation algorithm, which is a correlation-based DTD algorithm. The performance of the proposed algorithm will be compared against these two DTD algorithms in Section 4.

2.2 Geigel Algorithm

The Geigel algorithm [4] compares the magnitude of the near-end signal $d(n)$ with the maximum magnitude of the N most recent samples of the far-end signal $x(n)$, where N is the adaptive filter length. N past samples are used because of the possible end delay of $x(n)$ through the echo path. The echo path typically dampens the signal $x(n)$, and as a result the magnitude of the signal $d(n)$ containing only the echo $y(n)$ will be smaller than that containing both $y(n)$ and $v(n)$. The Geigel algorithm computes its detection variable as

$$\xi = \frac{|d(n)|}{\max\{|x(n)|,\ldots,|x(n-N+1)|\}} \qquad (1)$$

If ξ is larger than a threshold T, DT is declared, otherwise it is not. The choice for T needs to be made with care, and will strongly affect the performance of the detector. For line echo cancellers, T is set to 0.5 because the hybrid attenuation is assumed to be 6dB. For acoustic echo cancellers, the background noise level and/or the echo path can be time varying. Therefore, it is not easy to decide a proper value for T. In particular, for the time-varying echo path, the Geigel algorithm can falsely regard a change of the echo path as a DT situation. As a result, the adaptive filter stops updating the coefficients when the coefficients update is actually needed.

2.3 Normalized Cross-Correlation Algorithm

A DTD algorithm was proposed in [1] based on a normalized cross-correlation (NCC) measurement between the signal vector $\underline{x}(n) = [x(n) \quad x(n-1) \quad \ldots \quad x(n-N+1)]^T$ and the near-end signal $d(n)$. The algorithm normalizes the detection variable such that it is equal to 1 when the near-end speech $v(n)$ is zero, and less than 1 when $v(n)$ is not zero, i.e. DT condition. Therefore, the threshold T can be selected between 0 and 1 and is independent of the input signal. The detection variable is defined as following:

$$\xi = \sqrt{\underline{r}_{xd}^T(\sigma_d^2 R_x)^{-1}\underline{r}_{xd}} \qquad (2)$$

where σ_d^2 is the variance of $d(n)$, R_X is the autocorrelation matrix of the vector $\underline{x}(n)$, and $\underline{r}_{xd}$ is the cross-correlation vector between vector $\underline{x}(n)$ and signal $d(n)$. When the detection variable $\xi < T$, a DT condition is declared; when $\xi \geq T$, DT is not present. In practice, equation (2) is computationally expensive. For computational simplicity, (2) can be simplified by assuming that $\underline{\hat{h}} \approx \underline{h} = R_x^{-1}\underline{r}_{xd}$ when the adaptive filter has converged. Therefore, the detection variable in (2) can be written as:

$$\xi = \sqrt{\underline{r}_{xd}^T\underline{\hat{h}}\sigma_d^{-2}} \qquad (3)$$

Also, σ_d^2 and $\underline{r}_{xd}$ can be practically estimated by averaging over a window of W samples, for example as:

$$\underline{r}_{xd} \approx \frac{1}{W}\sum_{k=0}^{W-1}\underline{x}(n-k)d(n-k) \qquad (4)$$

3. PROPOSED DTD ALGORITHM

3.1 Overview

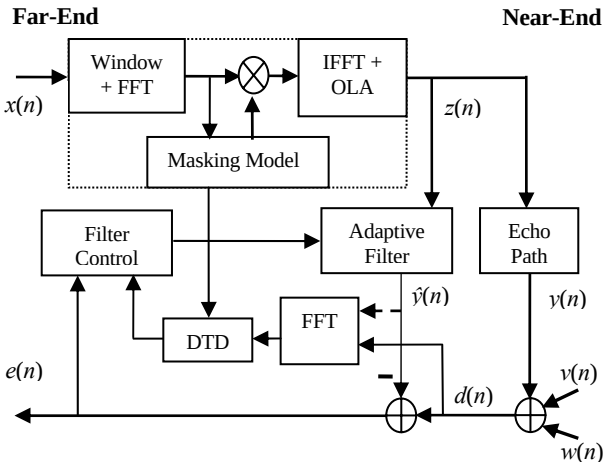

Figure 2: Echo Cancellation with the Proposed DTD Algorithm

The configuration diagram of an echo canceller with the proposed DT detector is shown in **Figure 2**. The far-end samples

$x(n)$ are segmented into overlapping frames. Each frame is transformed into the frequency domain, and then fed into a psychoacoustic auditory model to determine a masking threshold. A binary masking template is generated by comparing the power spectrum of a frame with its corresponding masking threshold. The masking template is set with a value of zero at frequency components below the masking threshold, and a value of one elsewhere. Spectral holes are created by multiplying the complex spectrum of a frame with its corresponding masking template at each frequency component. The masked signal $z(n)$ is generated by transforming the masked spectrum of frames back to the time domain and adding back together overlapped sections between consecutive frames. Note that the masking template should be stored for use in locating the spectral holes in a frame of the near-end signal $d(n)$. At the input port of the near end side, a DT condition can be detected by monitoring the energy level at the locations of the created holes, for the presence of the near-end speech $v(n)$.

3.2 Spectral Hole Insertions Using a Psychoacoustic Auditory Model

This section briefly describes the process of creating spectral holes based on the frequency masking properties of the human auditory system. The psychoacoustic auditory model proposed by Johnston in [7] is used in determining a masking threshold for an input signal.

1. Segmenting the signal $x(n)$ into overlapping frames of length N, with one frame being $\underline{x}(n) = [x(n) \;\; x(n-1) \;\; \ldots \;\; x(n-N+1)]^T$.
2. Computing a $2N$-point Fast Fourier Transform (FFT) on the input frame $\underline{x}(n)$ that has been appended with N zeroes and previously weighted by a Hanning window of length N. The power spectrum is calculated as $P_x(k) = |X(k)|^2$, where $X(k) = \mathrm{FFT}_{2N}\{\underline{x}(n)\}$.
3. Mapping the power spectrum from the frequency domain into a critical band (Bark) scale, by adding up energies in critical bands as:
$$P_b(z) = \sum_{k=f_l}^{f_h} P_x(k)$$
where f_h and f_l are, respectively, the high and low frequencies in the critical band z, and $z = 1, 2, \ldots, Z_t$ (the total number of critical bands for speech with 8 kHz sampling rate is $Z_t = 18$; see [8] for more details).
4. Convolving critical band energies $P_b(z)$ with a spreading function defined as :
$$B(z) = 15.91+7.5(z + 0.474)-17.5\sqrt{1 + (z + 0.474)^2}$$
where $z = |j-i|$, i is the Bark index of the masked component and j is the Bark index of the masking component. This step basically takes into account the masking between different critical bands to give $P_{bs}(z) = P_b(z) * B(z)$.
5. Computing a relative threshold offset based on whether the spectral flatness measure (SFM) of the frame is noise-like or tone-like. The relative threshold offset is defined as $T_{offset}(z) = \alpha(14.5 + z) + (1-\alpha)5.5$, where
$$\alpha = \min\left(\frac{SFM_{db}}{-60_{db}}, 1\right) \qquad SFM = \left\{\frac{GeometricMean}{ArithmeticMean}\right\}^{\frac{1}{Z_t}}$$
6. Computing a raw masking threshold by subtracting the threshold offset from the spread power spectrum as:
$$T_{raw}(z) = 10^{(\log_{10}(P_{bs}(z)) - 0.1T_{offset}(z))}$$
7. Normalizing the $T_{raw}(z)$ to take into account different number of frequency components in critical bands. The normalized threshold is defined as:
$$T_{norm}(z) = \frac{T_{raw}(z)}{P_z}$$
where P_z is the number of frequency components in critical band z ($z = 1, 2, \ldots, Z_t$).
8. Calculating a final masking threshold $T(z)$, taking into account the absolute threshold of hearing $T_{abs}(z)$ (see [8] for details), as $T(z) = \max[T_{norm}(z), T_{abs}(z)]$.
9. Mapping the final masking threshold from the Bark scale back to the linear frequency scale. A binary masking template is then created by comparing the spectral components with the final masking threshold. Components with energy above the threshold are retained; otherwise they are set to zero. This masking template will also be used to locate spectral holes for detecting the near-end signal. **Figure 3a** shows in log scale the spectrum of an input frame of $\underline{x}(n)$ with its masking threshold, and **Figure 3b** shows the masked or holed spectrum with the generated masking template, in linear scale.
10. Computing an inverse FFT_{2N} of the masked spectrum to get a signal frame of $\underline{z}(n)$ in the time domain, with the overlapped portions of consecutive frames properly added together.

3.3 Detection of Near-end Signal at Spectral Holes

The near-end signal $d(n)$, being the sum of the echo $y(n)$, the speech $v(n)$ and the additive noise $w(n)$, is segmented into frames in the same way as $x(n)$ previously. The masking template generated in Step 8 of Section 3.2 is stored to keep track of the spectral holes locations for a corresponding frame of $d(n)$. It is important to properly synchronize a masking template pulled out of the storage with a $d(n)$ frame properly, so that the created spectral holes in $d(n)$ align with those of the template. The alignment can be done by monitoring the magnitude peak of the adaptive filter coefficients, or by a signal cross-correlation method between $x(n)$ and $d(n)$. The detection of $v(n)$, i.e. a DT condition, is performed by comparing the spectral level of the $d(n)$ frame at the spectral holes with a threshold T just above the monitored noise floor level. If the spectral level is greater than T, then a DT condition is declared. The detection variable of the proposed algorithm is defined as $\xi = \{P_D(k)\} > T$, for any $k \in \{\textit{hole indices set}\}$, where $P_D(k) = |D(k)|^2$, and $D(k) = \mathrm{FFT}_{2N}\{d(n)\}$.

3.4 Smearing Effect at Spectral Holes

The smearing effect refers to a leakage or spilling of energy into the spectral holes. Since the proposed algorithm depends on the integrity of spectral holes for detection of DT, any excessive level of the smearing effect at the spectral

holes can cause the algorithm to falsely detect the condition as a DT. The following discusses briefly the causes of the smearing effect on the spectral holes in the frequency masking process. Also a remedy to the problem is discussed.

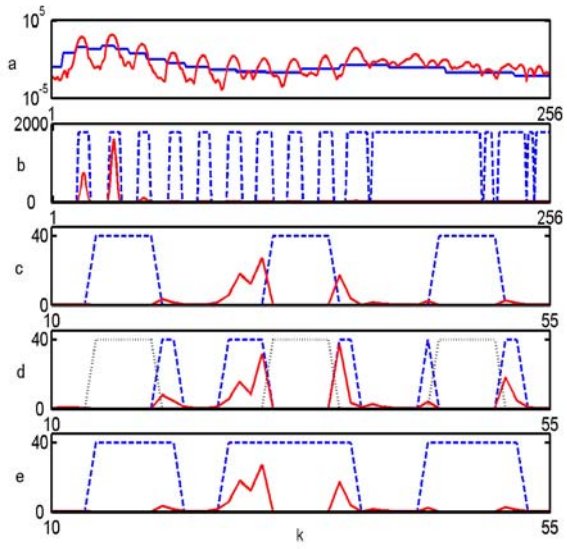

Figure 3: Spectral holes and smearing effect on the same frame: a) spectrum and masking threshold of $x(n)$ in log scale, b) masking template and holed spectrum c) zoomed masking template and smearing effect at spectral holes of $y(n)$, d) smearing template and smearing estimates at spectral holes of $\hat{y}(n)$, e) modified masking template as the union of the masking template and the smearing template.

First, spectral holes determined by the masking template can change from frame to frame, therefore the overlap-add process as described in Step 10 of Section 3.2 may add non-zero spectral components from the previous frame to the spectral holes in the current frame. Second, because of the filtering by the echo path $\underline{h}$, the echo $y(n)$ is a weighted sum of signals originating from consecutive frames with different hole locations. As a result, some "would be" holes of a frame of $y(n)$ are being filled up by signal components from previous frames.

As explained above, the smearing effect on spectral holes is certainly unavoidable. So the detection of $v(n)$, i.e. a DT condition, should be done only at intact locations of spectral holes. Therefore, it is very important to identify the positions of spectral holes where the level of smearing is excessively high, so that the DT detection can be avoided at these places. A remedy to the problem is to use the output $\hat{y}(n)$ of the adaptive filter to identify these places, because it is a reasonable estimate of $y(n)$ once the adaptive filter has converged. A smearing template is created by comparing the level of $\hat{y}(n)$ at positions of the spectral holes with a threshold. Components with a level above the threshold are excluded from the spectral holes in the DT detection process. To demonstrate the idea, **Figure 3c-e** shows a zoom-in section of the same signal frame as in **Figure 3a-b**. **Figure 3c** illustrates the smearing effect at some spectral holes of the echo $y(n)$. **Figure 3d** shows the corresponding spectral holes in the echo estimate $\hat{y}(n)$ and a smearing template which identifies the spectral components at the holes where the smearing level is high. Note that the misadjustment level for the echo estimate shown in **Figure 3d** is at around -10db. **Figure 3e** shows the spectral holes of $y(n)$ and the modified masking template which is a union of the masking and the smearing templates.

4. SIMULATION RESULTS

4.1 Evaluation Methodology

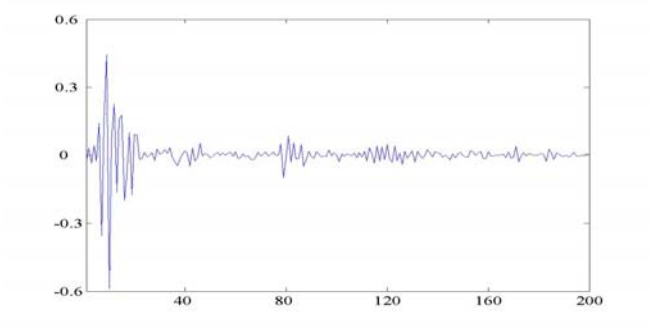

Figure 4: Echo Path IR (first 200 of 1200 samples)

The proposed algorithm is evaluated based on the method proposed in [3]. The algorithm is compared against the Geigel algorithm (1) and the NCC algorithm (3) and (4), using the probability of miss P_m as a measure versus other variable parameters such as the Near-end $v(n)$ to Echo $y(n)$ Ratio (NER), the Echo $y(n)$ to Near-end Noise $w(n)$ Ratio (ENNR), and probability of false alarm P_f. The P_m is defined as the percentage of samples where a true DT condition is not detected. The P_f is defined as the percentage of samples where a DT is falsely claimed. The tests for this paper were conducted using a real impulse response from a small conference room with 1200 samples and is shown in **Figure 4.** The far-end and near-end signals were taken from the Harvard sentences speech database, at 8 kHz sampling rate. The far-end signal from a female talker was about 10 s long. The 6 DT signals from 3 different female and 3 different male talkers were about 2 s long. The noise $w(n)$ was a white noise at different levels with respect to the echo level. For all simulations, the DT detector was turned on from sample number 20000, to allow the NLMS adaptive filter to converge roughly in the first 2.5 s. The adaptive filter had 1024 weights and the step size for the NLMS was set at 1.0 for the first 2.5 s and then reduced to 0.1 when the DT detector was turned on. For the proposed algorithm, a FFT length of 256 was used on a frame length of 128 samples, with a 50% overlap between frames. A short frame length is preferred as it reduces the delay of the system, and that of the DTD decision. Note that informal subjective tests for the psychoacoustic auditory model used in the proposed algorithm showed a minimal degradation of the output (i.e. partially masked) speech signals. The Geigel algorithm used a window length of the $N = 1024$ most recent samples, whereas the NCC algorithm used a window of size $W = 500$ to estimate the correlation measures. The DTD hold time was set at 30ms for all algorithms. The following describes the procedure in evaluating and comparing the DTD algorithms:

1. Select one of the parameters P_f, NER or ENNR as a variable and set the others to fixed values.
2. With DT signal $v(n) = 0$, select a threshold T for each algorithm that results in the selected P_f.
3. For each value of the variable selected in Step 1:
   - Select one of the six DT signals
   - Select six random positions within the last 7.5 s of the far-end speech
   - Compute P_m
4. Average the P_m obtained in Step 3 over all 36 conditions.
5. Plot the averaged P_m as a function of the variable.

4.2 Results and Discussion

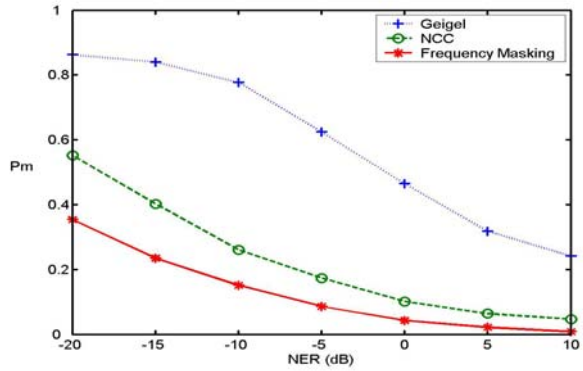

Figure 5: P_m versus NER for ENNR = 30 dB and P_f = 0.1

Figure 5 shows a plot of P_m as a function of the NER for a range from -20dB to 10dB, a fixed ENNR = 30 dB and P_f = 0.1. For all algorithms, it can be seen that as the power of the near-end speech $v(n)$ increases compared to the far-end signal $x(n)$, P_m decreases. The Geigel algorithm performs much worse than the other two algorithms, while the proposed algorithm performs the best for the whole range of NER.

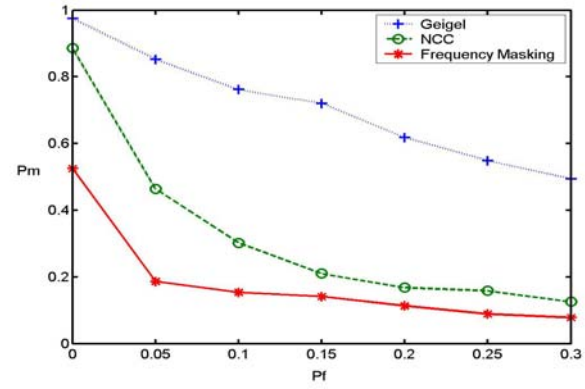

Figure 6: P_m versus P_f for NER = -10 dB and ENNR = 30 dB

Figure 6 shows a plot of P_m as a function of P_f in a range from 0 to 0.3, with NER = -10 dB and ENNR = 30 dB. From the plot, it can be seen that there is a trade off between P_m and P_f. Setting a DT detector at a low P_m would increase P_f. Again, the proposed algorithm performs the best for the whole range of P_f.

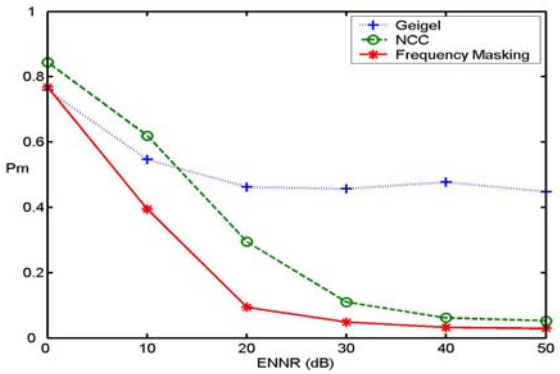

Figure 7: P_m versus ENNR for NER = 0 dB and P_f = 0.1

Figure 7 shows a plot of P_m as a function of the ENNR for a range from 0 dB to 50 dB, with NER = 0 dB and P_f = 0.1. As the power of near-end noise $w(n)$ increases, P_f increases. Thus, the threshold T has to be adjusted to keep P_f = 0.1. This increases P_m as a result. From the plot, all 3 algorithms do not perform very well when the echo level, which is the same as the DT signal level, is only 0-10dB greater than the near-end noise level. However, the proposed algorithm still performs better than the other algorithms overall.

5. CONCLUSIONS

This paper presents a novel DTD algorithm using a psychoacoustic auditory model and investigates the feasibility of the idea. Simulation results show the superior performance of the proposed method compared with the Geigel and the normalized cross correlation DTD algorithms.

6. ACKNOWLEDGEMENT

The impulse response in the simulations was provided by the DSP lab of Carleton University, chaired by Prof. Rafik Goubran.

REFERENCES

[1] J. Benesty, D.R. Morgan, and J.H. Cho, "A new class of doubletalk detectors based on cross-correlation," IEEE Trans. Speech Audio Processing, Vol. 8, pp. 168-172, March 2000.

[2] J. Benesty, T. Gansler, D.R. Morgan, M.M. Sondhi, and S.L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer, 2001.

[3] J.H. Cho, D.R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancellation," IEEE Trans. Speech Audio Processing, Vol. 7, pp. 718-724, Nov. 1999.

[4] D.L. Duttweiller, "A twelve-channel digital echo canceller," IEEE Trans. Comm., Vol. 26, pp. 647-653, May 1978.

[5] S.L. Gay, and J. Benesty, *Acoustic Signal Processing for Telecommunication*, Kluwer Academic Publishers, Boston, 2000.

[6] S. Haykin, *Adaptive Filter Theory*, 3rd Edition, Prentice Hall, Upper Saddle River, N.J., 1996.

[7] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," IEEE Journal on Selected Areas in Communications, Vol. 6, No. 2, February 1988.

[8] E. Zwicker, *Psychoacoustics: Facts and Models*, 2nd Edition, Springer, New York, 1999