An Open-Loop Doubletalk Detector Using Power Spectrum Estimation

FREDRIC LINDSTROM*, MATTIAS DAHL**, INGVAR CLAESSON**

*Research and Development
Konftel
S-906 51, Box 208, Umea
SWEDEN

**Department of Signal Processing
Blekinge Institute of Technology
S-37225, Ronneby
SWEDEN

Abstract: Doubletalk detection is commonly used in acoustic echo cancellation units. In situations of doubletalk the adaptive filter in the acoustic echo canceler might diverge, thereof the need for doubletalk detection. In this paper, an open-loop doubletalk detector based on power spectrum estimation is presented. The detector explores the signal characteristic of speech signals to define a frequency distance measure. The present signals are evaluated by this distance measure, and the measure is compared to a preset threshold. Doubletalk is declared whenever the measure exceeds the threshold. An objective evaluation technique is used to compare the proposed detector with two classic open-loop detectors, the Geigel detector and the open-loop correlation detector. It is shown that the proposed method outperforms the other two in the given evaluation technique.

Key-Words: - Doubletalk detection, DTD, Acoustic Echo Cancellation, AEC, Hands-free

1 Introduction

Hands-free operation can be a desirable feature in many products, e.g. car phones, desktop phones, videoconference systems, conference phones, etc. A hands-free phone call takes part between the participants located in the same room/car as the hands-free phone, the near-end talkers, and participants at a remote location, the far-end talkers. In a hands-free system it is possible that acoustic echoes arise, i.e. speech originating from the far-end talkers, that is reproduced by the loudspeaker, picked up by the microphone, and thereafter transmitted back to the far-end talkers. Acoustic echoes are in general considered very annoying. Several solutions to the acoustic echo problem have been proposed. One type of solutions are those based on the system identification scheme [1], [2]. A common solution, conforming to the system identification scheme, is often denoted an Acoustic Echo Canceler (AEC) [3]. Other acoustic echo cancelers using adaptive methods can also be categorized as acoustic echo cancelers. In the sequel, an AEC will denote the type of solution defined in [3]. The most algorithms for the implementation of an AEC are based on the Least-Mean-Square(LMS) algorithm, thanks to the LMS algorithm's robustness [4]. The performance of an AEC is linked to the estimation of certain parameters, such as speech activity, acoustic coupling etc [5]. The detection of speech activity parameters is crucial for most AEC systems, in particular doubletalk detection. Doubletalk occurs when the near-end speech is of such a level that the adaptation of the AEC should be stopped. Several doubletalk detectors have been proposed, e.g. the Giegel detector [6], cross-correlation and coherence based detectors [7]-[9], detectors making use of parallel filters [10], and detectors using power comparison or cepstral techniques [5]. This paper proposes a DoubleTalk Detector (DTD) based on power spectrum estimation. Doubletalk is detected by comparing a frequency domain distance measure with a preset threshold.

2 AEC and Doubletalk Detection

An AEC and its environment are depicted in figure 1. The present signals are: the far-end signal, $x(k)$, the acoustic echo, $a(k)$, the near-end speech signal, $s(k)$, the near-end background noise, $n(k)$, the microphone signal, $m(k)$, the estimated echo, $\hat{a}(k)$, and the error signal, $e(k)$, i.e. the near-end line-out signal, where k is the sample index. The acoustic echo is generarated through filtering the far-end signal with the Loudspeaker-Enclosure-Microphone (LEM) system, i.e. the combined influence from the loudspeaker, the room, and the microphone, see figure 1. The LEM system is often modeled as a linear system [5], [11], it is assumed that the nonlinear part of the LEM can be modeled as a part of $n(k)$. Further, it is assumed that the power of the background noise $n(k)$ is at a low level as compared to $a(k)$, a sudden large increase of the background noise is modeled as a part of $s(k)$. These assumptions will be considered valid.

The AEC consist of an Adaptive Filter (AF) and a Adaptive Control Mechanism. The purpose of the AEC, is to adapt the AF in such a manner, that its transfer function is as similar as possible to that of the LEM, in some given measure. Since $e(k) = m(k) - \hat{a}(k)$, and $m(k) = a(k) + s(k) + n(k)$, the effect will be that the acoustic echo $a(k)$ is in large removed from the near-end line-out signal $e(k)$, providing that the transfer function of the AF is sufficiently close to that of the LEM, i.e. that $\hat{a}(k) \approx a(k)$. In the AEC, the signal $e(k)$ is used as feed-back input to the ACM, examples of algorithms possible for the implementation of the ACM are: the Normalized Least Mean Squares (NLMS), the Recursive Least Squares (RLS), and the Affine Projection Algorithm (APA) [1]. If a near-end speech signal $s(k)$ exists the AF might diverge, and thus an increased portion of the acoustic echo will be transferred back to the far-end talkers. A near-end speech detector is thus desirable in order to stop the adaptation of the AF. If the near-end signal $x(k)$ is not present, there is no acoustic echo $a(k)$ and the detection of a high near-end speech signal is easy, since $m(k) \approx s(k)$ in such a situation. It is therefore the detection of doubletalk that is of interest, i.e. the detection of simultaneous existence of the $x(k)$ and $s(k)$ signals.

Closed-loop DTDs are defined as those detectors whose performances depend on the adjustment of the AF, while open-loop DTDs are those whose performances are independent of the AF [5]. From figure 1 it can be seen, that DTDs using only signals $x(k)$ and $m(k)$ are open-loop, while DTDs making use of signal $\hat{a}(k)$ or $e(k)$ are likely to be closed-loop detectors, since $\hat{a}(k)$ and $e(k)$ are dependent on the AF. Closed-loop detectors are considered to outperform open-loop detectors in situations where the AF is well adapted [5]. However, the adaptive filter of a hands-free phone is not always well adapted, e.g. during initial adaptation or after a sudden spatial movement of the phone. In these situations, the open-loop detector might outperform the closed-loop detector. Thus, improved open-loop detectors or alternatively combined open-loop/closed-loop methods are still attractive.

The proposed DTD is a single statistic binary detection DTD, i.e. a detector that either declares doubletalk or not doubletalk, and which:

1. Produces a single detection statistic, $\xi(k)$
2. Declares doubletalk if $\xi(k) > T$, where T is a preset threshold
3. If doubletalk is declared for sample $k = k_0$ continues to declare doubletalk for the next T_{hold} samples no matter the value of $\xi(k)$

Most DTDs follow the characterization given above [11]. Two classic open-loop DTDs that conform to this characterization are the open-loop cross-correlation detector and the Geigel detector. The definitions of these two detectors given below are as in [5] and [11], respectively. The detection statistic for the Geigel detector $\xi_g(k)$ is defined as

$$\xi_g(k) = \frac{|m(k)|}{\max\{|x(k)|, \cdots, |x(k - N_g)|\}}, \qquad (1)$$

where N_g is a positive integer constant. The detection statistic for the open-loop cross-correlation detector $\xi_c(k)$ is defined as

$$\xi_c(k) = \left[\max_{l \epsilon [0, \cdots, L_c]} \frac{|\sum_{i=0}^{N_c - 1} x(k-i-l) m(k-i)|}{\sum_{n=0}^{N_c - 1} |x(k-i-l) m(k-i)|} \right]^{-1}, \qquad (2)$$

where N_c and L_c are two positive integer constants.

3 The Proposed DTD

The proposed DTD compares a Power Spectrum Estimate (PSE) of the far-end speech signal $x(k)$ and the

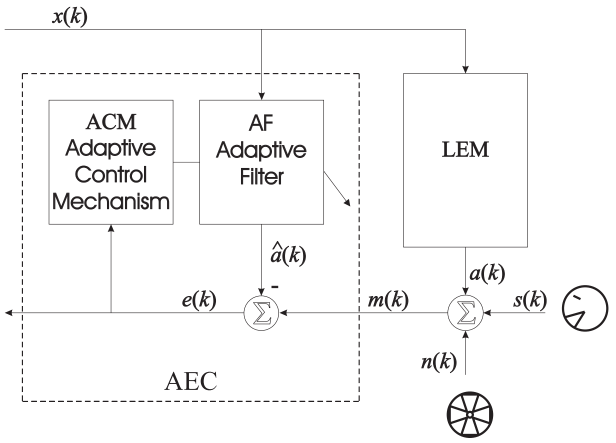

Figure 1: The AEC and its environment.

microphone signal $m(k)$. The main idea of the proposed DTD, is that a frequency bin in the microphone PSE cannot have a significantly larger magnitude than the corresponding far-end PSE bin, unless the near-end speech signal $s(k)$ is present. A sound signal is damped as it travels through open-air or is reflected by a wall. During sufficiently short durations of time the LEM system can be approximated by a Finite Impulse Response (FIR) filter, [5], [11]. This motivates the assumption that the LEM-system does not introduce too extreme amplification, delay or non-linear effects. These conditions imply, that if the PSE of the microphone signal contains bins with significantly larger magnitude than the magnitude of the corresponding bins in the PSE of the far-end speech signal, it is likely that near-end speech is present.

As speech is far from a flat spectrum signal [12], some of the frequency bins in the far-end speech PSE are likely to have magnitudes that are significantly less than the mean magnitude of the other bins. An example of a PSE of a far-end speech signal is shown in the top plot in figure 2, where bin 24, 25, and 29 can be considered as low energy bins. Assume that the near-end speech signal has low energy as compared to the acoustic echo. Then the near-end speech signal cannot effect the microphone signal PSE, except possibly for the bins where the acoustic echo has its lowest energy.

Thus, doubletalk can be detected by comparing the PSE of the far-end speech signal and the microphone signal for the bins corresponding to the lowest magnitude bins in the far-end speech PSE. This concept is visualized in the lower two plots in figure 2, where the presence of a near-end speech signal can be noted by comparing the microphone PSEs and far-end speech PSE for the ”low” energy bins 24, 25, and 29.

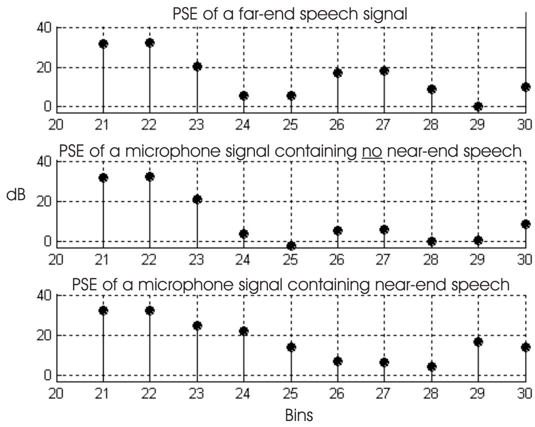

Figure 2: Bins 21-30 of PSEs for a 512 sample segment of a far-end speech signal (top plot), the corresponding microphone signal segment when no speech is present (middle plot), and the corresponding microphone segment when near-end speech is present (bottom plot). Scale is in dB, where 0dB is the magnitude of the smallest bin in the far-end speech PSE.

The proposed DTD is defined mathematically through the computation of its detection statistic, $\xi_p(k)$, which is given by

$$\xi_p(k) = \sum_{l=0}^{N_p-1} w_l(k)(\mathrm{PSE}_{mm,l}(k) - \mathrm{PSE}_{xx,l}(k)) \quad (3)$$

where $w_l(k)$ is a weight function and $\mathrm{PSE}_{mm,l}(k)$ and $\mathrm{PSE}_{xx,l}(k)$ are PSEs of the signals $m(k)$ and $x(k)$, where l is the frequency bin index. In this paper, the modified periodogram is used for PSE. The PSE of the signals $m(k)$ and $x(k)$ are thus given by

$$\mathrm{PSE}_{mm,l}(k) = \Big| \sum_{i=0}^{N_p-1} m(k-N_p+1+i) g_i e^{-2j\pi li/N_p} \Big|^2 \quad (4)$$

$$\mathrm{PSE}_{xx,l}(k) = \Big| \sum_{i=0}^{N_p-1} x(k-N_p+1+i) g_i e^{-2j\pi li/N_p} \Big|^2 \quad (5)$$

where N_p is a positive integer and $\mathbf{g} = [g_0, \cdots, g_{N_p-1}]$ is a window function of length N_p. The length of the power spectrum estimates, N_p, should be sufficiently long to provide satisfying estimation quality. However, detection delay increases with higher values of N_p, so there is a tradeoff between detection delay and estimation quality.

The weight function $w_l(k)$, used in equation (3), should reflect the principle given above, i.e. that bins with a ”low” magnitude in the PSE of the far-end signal $x(k)$ should be favored in the calculation of the detection statistic $\xi_p(k)$. One way to obtain such a weight function is to define $w_l(k)$ by

$$w_l(k) = \left\{ \begin{array}{ll} 1 & \text{if } \mathrm{PSE}_{xx,l}(k) \leq T_p \\ 0 & \text{if } \mathrm{PSE}_{xx,l}(k) > T_p, \end{array} \right. \tag{6}$$

where T_p is some preset threshold.

The calculation of $\xi_p(k)$ contains two power spectrum estimates, which are quite computational demanding as compared to the calculations performed in equations (1) and (2). To reduce the computational load, the detection statistic $\xi_p(k)$ could be computed only every M_p sample. The intermediate $M_p - 1$ samples are then defined as $\xi_p(M_p i + j)$= $\xi_p(M_p i)$ with $i = [0, 1, 2, \cdots]$ and $j = [1, \cdots, M_p - 1]$. High values of M_p implies a lower computational load. However, with increased values for M_p, there is an increased detection delay.

The proposed DTD, as given in equations (3)-(6), requires that the power spectrum of the near-end speech signal $x(k)$ contains sufficiently many frequency bins that fulfill the condition (6) to make a detection according to equation (3) possible. This implies that N_p cannot be set too large, due to the non-stationarity of speech signals. If the near-end speech signal $s(k)$ is present, but its spectrum has no or only a small part of its energy located in the bins used in equation (3) the near-end speech might pass undetected. Thus, it is required that when the near-end speech signal $s(k)$ is present, its spectrum contains significant energy in the bins corresponding to those weights, $w_l(k)$, that are equal to one, see equation (6). In order to determine if these requirements can be sufficiently met to make the proposed DTD attractive, an objective evaluation was performed.

4 Evaluation and Result

An objective technique for evaluating doubletalk detectors was proposed in [11]. The technique presented here is mainly the same, for motivations of methods and the settings of certain variables see [11]. The evaluation is based on Receiver Operating Characteristics (ROC). The characteristics used are probability of a false alarm, P_f, i.e. declaring doubletalk when doubletalk is not present, and probability of miss, P_m, i.e. not declaring doubletalk when doubletalk in fact is present. The procedure is as follows: for a number of specific preset P_f values one computes the value of P_m for a number of different levels of the Near-end to Far-end speech power Ratio (NFR). This measure is defined as

$$\mathrm{NFR} = \frac{\sigma_s}{\sigma_x}, \tag{7}$$

where σ_s and σ_x are the variance of the near-end signal, $s(k)$, and far-end speech signal, $x(k)$, respectively. Thus, one plot of P_m vs. NFR is obtained for each specified value of P_f. From these plots visual inspection is then used to make judgments about the DTD. Because the effect of a pause of the adaptation of the filter in the AEC during inactive far-end speech is minimal, a miss or a false alarm is counted only during the active portion of the far-end speech. A speech activity detector $K[z(k)]$, where $z(k) = x(k)$ or $z(k) = s(k)$, is defined as:

$$K[z(k)] = \left\{ \begin{array}{ll} 1 & \text{if } \overline{z}(k) > Z_l \\ 0 & \text{if } \overline{z}(k) \leq Z_l \end{array} \right. , \tag{8}$$

where Z_l is a preset constant and $\overline{z}(k)$ is an average of the absolute value of $z(k)$ given by $\overline{z}(k) = (1 - e^{-1/Z_m})\overline{z}(k-1) + e^{-1/Z_m}|z(k)|$, where Z_m is a preset constant. The constant Z_l defines for what level speech activity should be declared and the constant Z_m defines the ”memory” in the average estimator.

The characteristics P_f and P_m are obtained through a method where several single realization characteristics are used. A single realization means that the characteristics are obtained by using *one* specific set of signals $x(k)$, $n(k)$, $s(k)$ and *one* specific LEM system. The single realization parameters corresponding to P_f and P_m are denoted P_{fs} and P_{ms}, respectively. They are defined through

$$P_{fs} = \frac{\sum_k \phi(k) K[x(k)]}{\sum_k K[x(k)]} \tag{9}$$

$$P_{ms} = \frac{\sum_k \phi(k) K[x(k)] K[s(k)]}{\sum_k K[x(k)] K[s(k)]}, \tag{10}$$

where L_k is the total number of samples of the signal $x(k)$ and $\phi(k)$ is the doubletalk detector output, i.e.

$$\phi(k) = \left\{ \begin{array}{ll} 1 & \text{if } \xi(k) > T \\ 0 & \text{if } \xi(k) < T \end{array} \right. , \tag{11}$$

To obtain the plot of P_m vs. NFR for a given P_f, the following method is proposed in [11], introduced signals are defined below:

1. Select a P_f value, set $x(k) = x_t(k)$, $n(k) = n_t(k)$, and choose one representative room impulse response as the LEM system

2. Set $s(k) = 0$ and compute $m(k)$
3. Select a threshold T and compute $\phi(k)$
4. Compute P_{fs}
5. Repeat step 3 and 4 over a range of values for T
6. Select the value of T which corresponding P_{fs} value is closest to the chosen P_f value
7. Select an NFR value
8. Set $s(k) = v_{t,i}(k)$ with i=1, and compute $m(k)$
9. Compute $\phi(k)$ using the value of T chosen in 6
10. Compute P_{ms}
11. Repeat step 8-10 for $i = [2, \cdots, 16]$
12. Set P_m to the average of the 16 P_{ms} values
13. Repeat step 7-12 over a range of NFR values
14. Plot P_m as a function of NFR

In the above $x_t(k)$ consists of a male voice speech signal with continuous speech with a duration of 5s. The 16 near-end signals $v_{t,i}(k)$ are generated by choosing two male sentences and two female sentences of a duration of approximately 2s each, and from each of these creating 4 signals with a duration of 5s, by adding each sentence into a silent signal of length 5s at different random positions. The background noise, $n_t(k)$, is a flat spectrum bandlimited signal, with its variance set to -30dB as compared to the variance of $x_t(k)$. The communication bandwidth of all speech signals is $[0Hz, 8000Hz]$.

The Giegel, the open-loop correlation, and the proposed detector, were evaluated using the technique described above. The sample rate was 16kHz, and T_{hold} was set to 30ms for all three detectors. The constants in the speech activity detector were set to $Z_l = -30$dB and $Z_m = 512$. The parameters in the Giegel and the open-loop correlation detectors were set $N_g = 1000$, $N_c = 300$, $L_c = 30$. These settings were obtained from a large set of tried settings, and were chosen as being the optimal ones of the values in the tried set. The decision of optimality was made by visual inspection of the P_m vs. NFR plots. The settings for the proposed DTD were $N_p = 512$, $M_p = 256$, $T_p = 0.0001$, and **g** was the Hanning window function. These settings are not optimal, but were considered sufficient for demonstrating the virtues of the proposed DTD. In [11] P_f values in the range [0, 0.3] were considered acceptable for AEC systems. In this paper, as well as in [11], simulations were performed for P_f=0.1 and P_f=0.3.

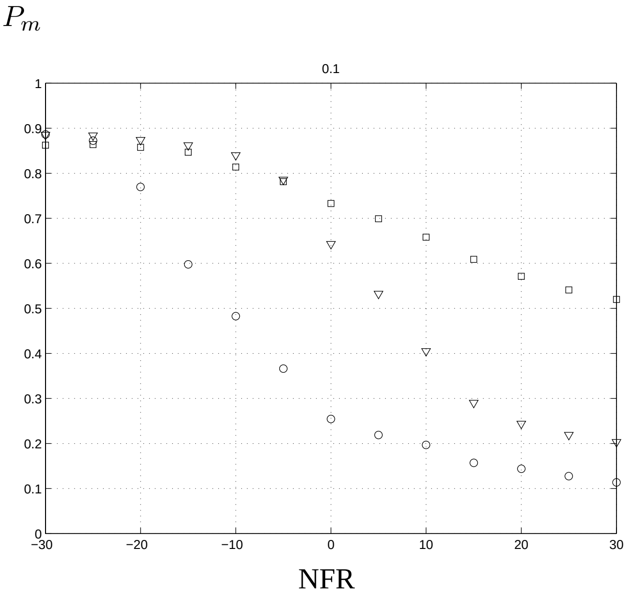

Figure 3: P_m vs. NFR for $P_f = 0.1$. Proposed DTD: *circles*, Giegel DTD: *triangles*, Open-loop correlation DTD: *squares*.

Figure 3 show a plot of P_m vs. NFR for the three detectors for $P_f = 0.1$. Figure 4 show the same for $P_f = 0.3$. From these two plots it can be seen that the probability of a miss in the detection for the proposed DTD is less or approximately the same as compared to the other two methods. For NFR values in the range -20dB to 30dB the difference is significant for both $P_f = 0.1$ and $P_f = 0.3$. The result of the evaluation is thus that the proposed DTD outperforms the other two methods. The increased performance is mainly given by the weighting performed in equation (6). Through this weighting specific frequency bands with low NFR is excluded in the detection of the near-end signal.

5 Conclusions

In this paper an open-loop DTD was presented. The proposed DTD uses the power spectrum estimates of the far-end speech signal and the microphone signal to produce a frequency domain measure. The measure is compared with a preset threshold, and doubletalk is declared if the measure exceeds the threshold. The proposed DTD was compared with two classic open-loop

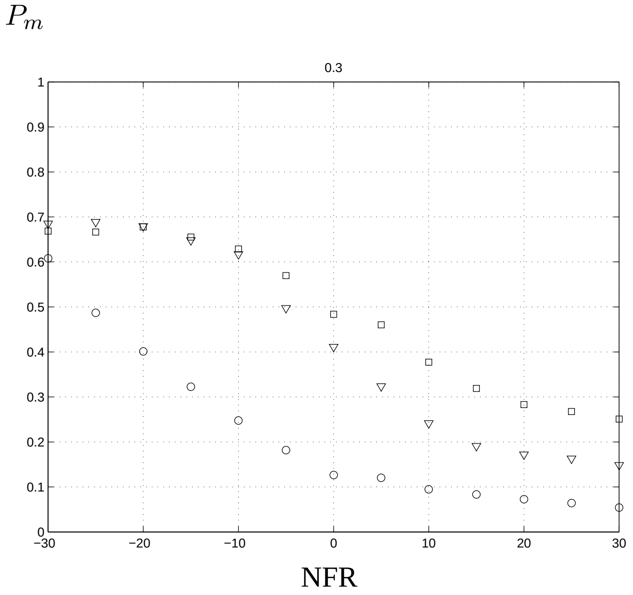

Figure 4: P_m vs. NFR for $P_f = 0.3$. Proposed DTD: *circles*, Giegel DTD: *triangles*, Open-loop correlation DTD: *squares*.

DTDs by an objective evaluating technique. The evaluation showed that the proposed DTD outperforms the other two. The proposed DTD is thus an interesting candidate for the doubletalk detection in future AEC implementations.

References

[1] S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice-Hall, 2002.

[2] B. Widrow, S. D. Stearns, *Adaptive signal processing*, Prentice-Hall, 1985.

[3] M. M. Sondhi, ”An adaptive echo canceler”, *Bell Syst. Tech. J.*, vol. 46, pp. 497-510, March 1967.

[4] F. Lindstrom, M. Dahl and I. Claesson, ”An LMS Based Algorithm for Reduced Finite Precision Effects”, *Proc. of WSEAS ICECS’02*, Singapore, December 2002.

[5] A. Mader, H. Puder, G. U. Schmidt, ”Step-size control for acoustic cancellation filters - an overview”, *Signal Processing*, vol. 80, pp. 1697-1719, 2000.

[6] D.L. Duttweiler, ”A twelve-channel digital echo canceler”, *IEEE Trans. on Commun.*, vol. COM-26, pp. 647-653, May 1978.

[7] H. Ye, B. X. Wu, ”A new double talk detection based on the orthognality theorem”, *IEEE Trans. on Commun.*, vol. 39, pp. 1542-1545, November 1991.

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

[9] T. Gansler, M Hansson, C.-J. Ivarsson, G. Salomonsson, ”A double-talk detector based on coherence”, *IEEE Trans. on Commun.*, vol. 44, pp. 1421-1427, November 1996.

[10] F. Lindstrom, M. Dahl and I. Claesson, ”Delayed Filter Update - An Acoustic Echo Canceler Structure for Improved Doubletalk Detection Handling”, *WSEAS Transactions on Communications*, Issue 4, October 2003.

[11] J. H. Cho, D. R. Morgan, J. Benesty, ”An objective technique for evaluating doubletalk detectors in acoustic echo cancelers”, *IEEE Trans. on Speech and Audio Process.*, vol. 7, pp. 718-724, November 1999.

[12] J. Deller, J. Hansen, J. Proakis, *Discrete-time processing of speech signals*, IEEE Press, 2003.