ROBUST EARLY ECHO CANCELLATION AND LATE ECHO SUPPRESSION IN THE STFT DOMAIN

Emanuël A. P. Habets[1,2]*, Sharon Gannot*[1] *and Israel Cohen*[2]

[1]Bar-Ilan University, School of Engineering, Ramat-Gan, Israel
[2]Technion - Israel Institute of Technology, Haifa, Israel

ABSTRACT

Acoustic echo arises due to acoustic coupling between the loudspeaker and the microphone of a communication device. Acoustic echo cancellation and suppression techniques are used to reduce the acoustic echo. In this work we propose to first cancel the early echo, which is related to the early part of the echo path, and subsequently suppress the late echo, which is related to the later part of the echo path. The identification of the echo path is carried out in the Short-Time Fourier Transform (STFT) domain, where a trade-off is facilitated between distortion of the near-end speech, residual echo, convergence rate, and robustness to echo path changes. Experimental results demonstrate that the system achieves high echo and noise reduction while maintaining low distortion of the near-end speech. In addition, it is shown that the proposed system is more robust to echo path changes compared to an acoustic canceller alone.

Index Terms— acoustic echo cancellation, acoustic echo suppression, noise suppression, step-size control.

1. INTRODUCTION

Acoustic echo arises due to the acoustic coupling between the loudspeaker and the microphone of a communication device. The acoustic echo, ambient noise and reverberation of the near-end speech decrease the intelligibility of the near-end speech signal. Different techniques have been developed to reduce the acoustic echo, viz., echo cancellation and echo suppression [1]. The echo canceller usually consists of a linear operation while the echo suppressor consists of a non-linear operator. While early developed echo suppression techniques employed hard-decision mechanisms that result in half-duplex communication, echo cancellation techniques provides full-duplex communication. Later developed suppression techniques employed soft-decision mechanisms that provide full-duplex communication. The echo canceller estimates the amplitude and phase of the echo signal. Hence, it is possible to achieve perfect echo cancellation under the assumption that the echo path can be described by a linear system and its impulse response is of finite length. Unfortunately, the echo canceller is sensitive to echo path changes. The acoustic echo suppressors have shown to be more robust to echo path changes [2] and achieve higher echo reduction [1] when compared to echo cancellers. However, during so-called double-talk situation, i.e., when the far-end speaker and the near-end speaker are simultaneously active, the suppressor tends to distort the near-end speech [1].

In many communication devices the cancellation and suppression technique are employed to obtain a satisfactory reduction of the echo. In [2] the authors employed echo cancellation for low frequencies and echo suppression for higher frequencies. In most cases the echo suppressor is used to reduce the residual echo, i.e., echo that is not eliminated by the echo canceller. The residual echo suppression is often part of a post-filter, which is used to increase the quality of the near-end speech by suppressing ambient noise and reverberation of the near-end speech signal.

In [3] Portnoff derived a representation of Linear Time Invariant (LTI) systems in the Short-Term Fourier Transform (STFT) domain. Recently, Avargel and Cohen investigated the identification of such systems in the STFT domain [4, 5]. In this work we first identify the echo path in the STFT domain. We then propose to cancel the early echo, which is related to the early part of the echo path, and to suppress the late echo, which is related to the later part of the echo path. A compromise between low distortion of the near-end speech on the one hand and robustness to echo path changes, high echo reduction, and fast convergence on the other hand can be made depending on the temporal partitioning of the echo path impulse response. To increase the quality of the near-end speech, ambient noise is suppressed as well. The late echo is suppressed down to the residual ambient noise level to prevent dodging of the output signal. Consequently, we can use signals that are available in the post-filter to control the echo canceller. Experimental results demonstrate high echo and noise reduction while maintaining low distortion of the near-end speech. In addition, it is shown that the proposed system is more robust to echo path changes than an acoustic canceller alone.

This paper is organized as follows: In Section 2 the representation and identification of the acoustic echo path in the STFT domain is briefly reviewed. The echo cancellation and suppression system is developed in Section 3. The control of the adaptive filter is discussed in Section 4. Finally, experimental results are presented in Section 5.

2. ACOUSTIC ECHO PATH

In this section the representation and identification of the acoustic echo path in the STFT domain is reviewed.

Let us assume that the echo path $h(n)$ is linear and of finite length Q. The far-end signal $x(n)$ and the echo signal $d(n)$ are then related by

$$d(n) = \sum_{i=0}^{Q-1} h(i)\, x(n-i). \tag{1}$$

The microphone signal $y(n)$ consists of the echo signal $d(n)$, a near-end speech signal $z(n)$, and an ambient noise signal $u(n)$, i.e.,

$$y(n) = d(n) + z(n) + u(n). \tag{2}$$

An estimate of $d(n)$ is required to eliminate the echo that is received by the microphone. This estimate can be obtained by identifying the echo path.

This research was supported by the Israel Science Foundation (grant no. 1085/05).

2.1. Representation in the STFT Domain

In the STFT domain the signal $x(n)$ is given by

$$X(\ell,k) = \sum_{m=-\infty}^{\infty} x(m)\,\tilde{\psi}(m-\ell L)\,\mathrm{e}^{-j\frac{2\pi}{N}k(m-\ell L)}, \qquad (3)$$

where ℓ is the frame index, k is the frequency band index, L is the discrete time shift, and $\tilde{\psi}(m)$ denotes the analysis window of length N. Subsequently we can express $d(n)$ in the STFT domain as [5]

$$D(\ell,k) = \sum_{k'=0}^{N-1} \sum_{\ell'=-\infty}^{\infty} H(\ell',k,k')\,X(\ell-\ell',k'). \qquad (4)$$

The STFT response $H(\ell',k,k')$ is related to impulse response $h(n)$ by

$$H(\ell',k,k') = \big(h(n) * \vartheta(n,k,k')\big)\big|_{n=\ell' L}, \qquad (5)$$

where $*$ denoted convolution with respect to n. The function $\vartheta(n,k,k')$ is related to the analysis window $\tilde{\psi}(m)$ and the synthesis window $\psi(m)$ of length N:

$$\vartheta(n,k,k') \triangleq \mathrm{e}^{j\frac{2\pi}{N}k'n} \sum_{m=-\infty}^{\infty} \tilde{\psi}(m)\psi(m+n)\mathrm{e}^{-j\frac{2\pi}{N}m(k-k')}. \qquad (6)$$

The STFT response $H(\ell',k,k')$ may be interpreted as a response to an impulse $\delta(\ell',k-k')$ in the time-frequency domain. The cross-band filters for $k \neq k'$ are used to cancel the aliasing effect caused by the subsampling. In practice, only a limited number of cross-bands are required to cancel the aliasing effect [5]. It should be noted that the filter $H(\ell',k,k')$ (for fixed k and k') is noncausal. Therefore, the microphone signal $y(n)$ is usually delayed by $\left\lceil \frac{N}{L}-1 \right\rceil L$ samples.

2.2. Identification

Let $\hat{H}(\ell,\ell',k,k')$ denote an adaptive filter of length Q' at frame index ℓ, which estimates $H(\ell',k,k')$. Using (5) it can be shown that $Q' = \left\lceil \frac{Q+N-1}{L} \right\rceil + \left\lceil \frac{N}{L} \right\rceil - 1$. In case $2K+1$ cross-band filters are used the estimated echo signal $\hat{D}(\ell,k)$ is given by[1]

$$\hat{D}(\ell,k) = \sum_{k'=k-K}^{k+K} \sum_{\ell'=0}^{Q'-1} \hat{H}(\ell,\ell',k,k' \bmod N)\,X(\ell-\ell',k' \bmod N). \qquad (7)$$

The $2K+1$ cross-band filters for each frequency band k are concatenated such that

$$\hat{\mathbf{H}}(\ell,\ell',k) = \Big[\hat{H}(\ell,\ell',k,(k-K) \bmod N), \ldots, \hat{H}(\ell,\ell',k,(k+K) \bmod N)\Big]^T. \qquad (8)$$

Given the STFT of the microphone signal $y(n)$, i.e., $Y(\ell,k)$, we compute an error signal $E(\ell,k)$ using

$$E(\ell,k) = Y(\ell,k) - \hat{D}(\ell,k). \qquad (9)$$

Here the adaptive filter coefficients are updated using the Normalized Least Mean Squares (NLMS) algorithm:

$$\hat{\mathbf{H}}(\ell,\ell',k) = \hat{\mathbf{H}}(\ell-1,\ell',k) + \mu(\ell,k)\mathbf{X}(\ell-\ell',k)E^*(\ell,k), \qquad (10)$$

where $\mu(\ell,k)$ denotes the step-size that is discussed in Section 4, $(\cdot)^*$ denotes complex-conjugation, and $\mathbf{X}(\ell,k) = [X(\ell,(k-K) \bmod N), \ldots, X(\ell,(k+K) \bmod N)]^T$.

[1]The expression $k' \bmod N$ ensures the periodicity of the frequency bands.

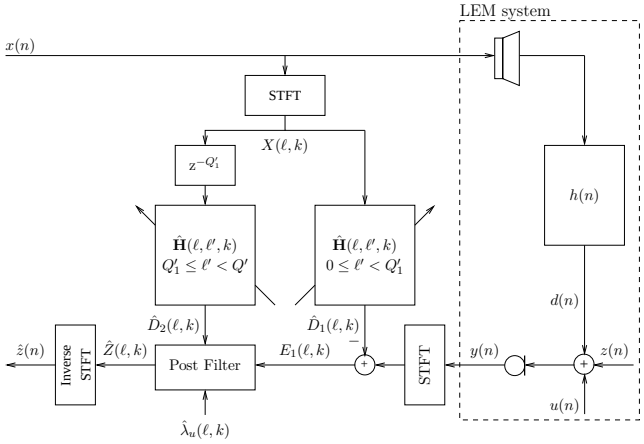

Fig. 1. Proposed acoustic echo cancellation and suppression system.

3. PROPOSED SYSTEM

In the previous section we have shown how the echo signal $D(\ell,k)$ can be estimated in the STFT domain. The main difference between echo cancellation and suppression is determined by the way the knowledge of $\hat{D}(\ell,k)$ is used. Cancellation of the echo can be achieved by subtracting $\hat{D}(\ell,k)$ from the microphone signal $Y(\ell,k)$ (see for example [4]). Alternatively, suppression can be achieved by applying a gain function to $Y(\ell,k)$ (see for example [1]). The gain function is usually related to the *a posteriori* and/or *a priori* (near-end) Signal to Interference Ratio (SIR).

Compared to echo cancellation, echo suppression can lead to considerable distortions of the near-end speech signal during double-talk periods, i.e., when the SIR is low [1]. On the other hand echo suppressors have been shown to be more robust with respect to echo path changes [2] and achieve higher echo reduction [1].

We make the following observations: Firstly, in many practical scenarios the speaker-microphone distance is small. Therefore, the early part of the echo path contains more energy than the later part. Secondly, continuous or sudden movements in the enclosure can significantly affect the echo path. However, the envelope of the later part of the echo path remains almost constant.

In order to achieve a compromise between echo cancellation and suppression we propose the following system. We first cancel part of the echo using the filter coefficients of $\hat{\mathbf{H}}$ that are related to the early part of the echo path, i.e. $\ell' = \{0,1,\ldots,Q_1'-1\}$, where $Q_1' \le Q'$. The estimated early echo signal is denoted by $\hat{D}_1(\ell,k)$. The canceller will increase the SIR of the obtained error signal $E_1(\ell,k)$. Subsequently, we apply a post-filter to $E_1(\ell,k)$. The post-filter suppresses the late echo and ambient noise. The late echo is estimated using the filter coefficients of $\hat{\mathbf{H}}$ that are related to the later part of the echo path, i.e., $\ell' = \{Q_1', \ldots, Q_1'+Q_2'-1\}$, where $Q_2' \le Q'-Q_1'$. The estimated late echo signal is denoted by $\hat{D}_2(\ell,k)$. The spectral variance of the ambient noise is denoted by $\lambda_u(\ell,k)$, and can be estimated using the Improved Minima Controlled Recursive Averaging (IMCRA) approach [6]. The proposed system is depicted in Fig. 1.

It should be noted that the estimates $\hat{D}_1(\ell,k)$ and $\hat{D}_2(\ell,k)$ can be constructed in many ways. For example, we could use the filter coefficients that are related to the early part and low frequencies of the echo path to construct $\hat{D}_1(\ell,k)$, while using all other filter coefficients to construct $\hat{D}_2(\ell,k)$. In this contribution we focus on the temporal partitioning of the echo path.

3.1. Acoustic Echo Cancellation

The echo signal $\hat{D}_1(\ell,k)$ can be calculated using

$$\hat{D}_1(\ell,k) = \sum_{k'=k-K}^{k+K} \sum_{\ell'=0}^{Q_1'-1} \hat{H}(\ell,\ell',k,k' \bmod N)\, X(\ell-\ell',k' \bmod N) \tag{11}$$

where $0 \leq Q_1' \leq Q'$. The resulting error signal that will be further processed by the post-filter is given by

$$E_1(\ell,k) = Y(\ell,k) - \hat{D}_1(\ell,k). \tag{12}$$

3.2. Acoustic Echo and Noise Suppression

The late echo signal can now be estimated using

$$\hat{D}_2(\ell,k) = \sum_{k'=k-K}^{k+K} \sum_{\ell'=Q_1'}^{Q_1'+Q_2'-1} \hat{H}(\ell,\ell',k,k' \bmod N) \cdot X(\ell-\ell',k' \bmod N), \tag{13}$$

where $0 \leq Q_2' \leq Q' - Q_1'$.

Let us define the spectral variance of the late echo signal and the error signal of the first stage as $\lambda_{d_2}(\ell,k) = \mathcal{E}\{|D_2(\ell,k)|^2\}$ and $\lambda_{e_1}(\ell,k) = \mathcal{E}\{|E_1(\ell,k)|^2\}$, respectively. The spectral variances are estimated using:

$$\hat{\lambda}_r(\ell,k) = \beta\, \hat{\lambda}_r(\ell,k) + (1-\beta)\, |R(\ell,k)|^2, \tag{14}$$

where β is a forgetting factor, and $R \in \{D_2, E_1\}$. We define the *a priori* and *a posteriori* SIR as

$$\zeta(\ell,k) = \frac{\lambda_{e_1}(\ell,k)}{\lambda_{d_2}(\ell,k) + \lambda_u(\ell,k)} \tag{15}$$

and

$$\xi(\ell,k) = \frac{|E_1(\ell,k)|^2}{\lambda_{d_2}(\ell,k) + \lambda_u(\ell,k)}, \tag{16}$$

respectively. While the *a posteriori* SIR can be calculated directly we need to estimate the *a priori* SIR $\zeta(\ell,k)$. Here $\zeta(\ell,k)$ is estimated using the Decision-Directed approach (see for example [7] and the references therein).

In the last few decades a large number of gain functions have been developed that can be used to suppress interferences. A comprehensive overview can be found in [7]. Here we have used a gain function that minimizes the Mean Squared Error (MSE) between the log spectral amplitude of the desired near-end signal and its estimate. This gain function is given by

$$G_{\text{LSA}}(\ell,k) = \frac{\zeta(l,k)}{1+\zeta(l,k)} \exp\left(\frac{1}{2}\int_{\gamma(l,k)}^{\infty} \frac{e^{-t}}{t}\, \mathrm{d}t\right), \tag{17}$$

where $\gamma(l,k) = \frac{\zeta(l,k)}{1+\zeta(l,k)}\xi(\ell,k)$. To reduce musical tones and dodging of the output power below the residual ambient noise power a time-varying lower bound is applied to $G_{\text{LSA}}(\ell,k)$. The estimated near-end signal is then given by

$$\hat{Z}(\ell,k) = \min\left\{G(\ell,k), G_{\min}\frac{\hat{\lambda}_u(\ell,k)}{\hat{\lambda}_{d_2}(\ell,k) + \hat{\lambda}_u(\ell,k)}\right\} E_1(\ell,k), \tag{18}$$

where $G_{\min}$ determines the maximum ambient noise suppression.

4. ADAPTATION CONTROL

For a regular NLMS algorithm the step-size is determined by

$$\mu(\ell,k) = \frac{\mu_{\text{NLMS}}}{\epsilon + \mathcal{X}^T(\ell,k)\mathcal{X}(\ell,k)}, \tag{19}$$

where ϵ is an auxiliary parameter that avoids division by zero and

$$\mathcal{X}(\ell,k) = \Big[\mathbf{X}^T(\ell,k), \mathbf{X}^T(\ell-1,k), \ldots, \mathbf{X}^T(\ell-Q'+1,k)\Big]^T. \tag{20}$$

The stability of the NLMS algorithm is governed by a step-size parameter. It is well known that the choice of this parameter reflects a tradeoff between good tracking ability and fast convergence on the one hand and low misadjustment on the other hand. To cope with this conflicting requirement, the step-size needs to be controlled. In [8] Benesty et al. proposed a Non-Parametric Variable Step-Size (NPVSS) NLMS algorithm in the time-domain. In the time-domain the step-size is related to the variance of the ambient noise $u(n)$, the variance of the far-end speech signal $x(n)$, and the variance of the total error signal $e(n)$.

Since the adaptive filter as well as the post-filter are implemented in the STFT domain the available signals can easily be exchanged. According to the same lines we formulate the NPVSS in the STFT domain:

$$\mu_{\text{NPVSS}}(\ell,k) = \frac{1}{\epsilon + \eta(\ell,k) + \mathcal{X}^T(\ell,k)\mathcal{X}(\ell,k)} \cdot \left(1 - \frac{\sqrt{\lambda_u(\ell,k)}}{\epsilon + \sqrt{\lambda_e(\ell,k)}}\right), \tag{21}$$

where $\eta(\ell,k) = \text{constant}\cdot\lambda_x(\ell,k)$ is a regularization parameter, and $\lambda_e(\ell,k)$ is the spectral variance of the error signal $E(\ell,k)$. Finally, the step-size is determined by

$$\mu(\ell,k) = \begin{cases} \mu_{\text{NPVSS}}(\ell,k) & \text{if } \lambda_e(\ell,k) \geq \lambda_u(\ell,k), \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Compared to the frequency independent step-size the frequency dependent step-size can improve the performance of the adaptive filter in terms of convergence, tracking, and misadjustment.

5. EXPERIMENTAL RESULTS

In this section we evaluate the performance of the proposed system using a sample frequency of 16 kHz. The following parameters were used: $N = 1024, L = 0.25N, K = 2, Q' = 10, \beta = 1 - \frac{1}{6Q'}$. In this paper an 'ideal' detector was used to indicate double-talk periods, and the adaptation of the filter was stopped during these periods. The distance between the microphone and the loudspeaker was 10 cm and the distance between the microphone and the near-end speaker was 25 cm. The acoustic impulse responses were generated using an efficient implementation of the image method [9]. The room dimensions were 5 m x 6 m x 4 m (length x width x height) and the reverberation time was approximately 500 ms.

5.1. AEC and AES Performance

We tested the proposed system during single- and double-talk (between 4 and 6.2 seconds). The near-end speech to echo ratio was -16.3 dB. The following signals are depicted in Fig. 2: a) microphone signal $y(n)$, b) signal $\hat{z}_1(n)$ with $Q_1' = 3 \wedge Q_2' = 0$, c) signal $\hat{z}_2(n)$

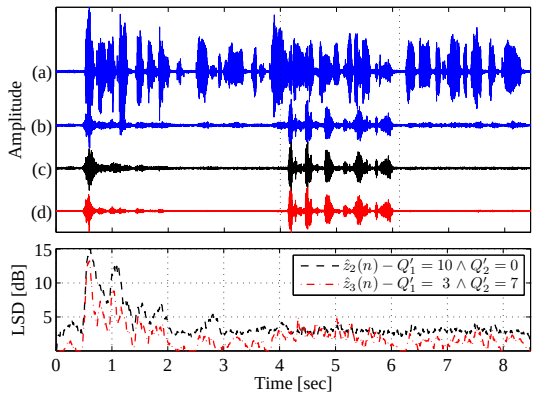

Fig. 2. Top: a) microphone signal, processed signals b) $\hat{z}_1(n)$ - $Q_1' = 3 \wedge Q_2' = 0$, c) $\hat{z}_2(n)$ - $Q_1' = Q' \wedge Q_2' = 0$, d) $\hat{z}_3(n)$ - $Q_1' = 3 \wedge Q_2' = 7$. Bottom: LSD for $\hat{z}_2(n)$ and $\hat{z}_3(n)$ with respect to $z(n)$.

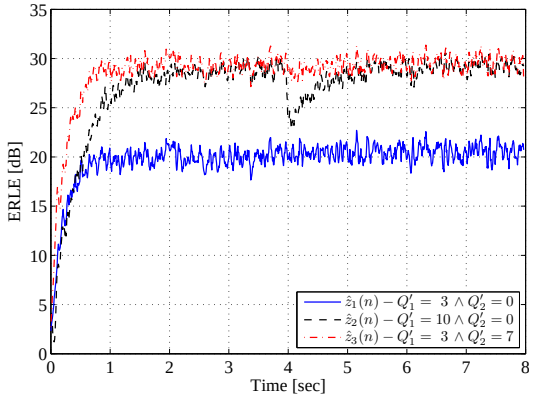

Fig. 3. ERLE for a speech-like noise signal ($f_s = 16$ kHz, SNR=30 dB) of $\hat{z}_1(n)$, $\hat{z}_2(n)$, and $\hat{z}_3(n)$. The echo path changes after 4 seconds.

with $Q_1' = Q' \wedge Q_2' = 0$, d) signal $\hat{z}_3(n)$ with $Q_1' = 3 \wedge Q_2' = 7$. We also depicted the Log Spectral Distance (LSD) [7] between the near-end signal $z(n)$ and $\hat{z}_2(n)$ and $\hat{z}_3(n)$. As seen from the waveforms and the LSD the echo and ambient noise was reduced. Compared to an echo canceller alone ($\hat{z}_2(n)$) the proposed system achieves faster convergence and slightly higher echo reduction without significantly decreasing the speech quality during double-talk.

5.2. Robustness

In order to test the robustness with respect to echo path changes the loudspeaker position was changed after 4 seconds. The position was rotated in the x-y plane by 30°, the microphone position was the center of the rotation. Since the distance between the microphone and the loudspeaker is not affected we expect little changes in the early part of the echo path and larger changes in the later part of the echo path. The near-end signal to ambient noise ratio was 30 dB. In this experiment no noise was reduced by the post-filter. In Fig. 3 the Echo Return Loss Enhancement (ERLE) [4] of the signals $\hat{z}_1(n)$, $\hat{z}_2(n)$, and $\hat{z}_3(n)$ are depicted. In case the complete echo path is cancelled using $Q_1' = Q' \wedge Q_2' = 0$ a clear dip in the ERLE of approximately 7 dB occurs when the echo path changes. However, no dip occurs when the proposed combination of the canceller and the suppressor is used ($Q_1' = 3 \wedge Q_2' = 7$). We emphasis that it is difficult to track the changes in the echo path during double-talk. Since the proposed system is not sensitive to small echo path changes an increased performance is expected during these periods. Subjective listening tests confirmed the robustness to echo path changes when using speech and speech-like signals.

6. CONCLUSION

In this paper a system is proposed for acoustic echo and noise reduction in the STFT. In this system two commonly used techniques, viz., echo cancellation and echo suppression are combined. The early echo that is related to the early part of the echo path is cancelled while the echo that is related to the later part of the echo path is suppressed. The temporal partitioning of the echo path impulse response admits a compromise between low speech distortion on one hand, and robustness to echo path changes, high echo reduction and fast convergence on the other hand. The gain function that controls the suppression minimizes the MSE between the log spectral amplitude of the near-end speech and its estimate. A lower-bound on this function results in a constant residual ambient noise power at the output. In addition, the ambient noise level is used to determine the variable step-size of the adaptive algorithm. Experimental results have demonstrated the robustness with respect to echo path changes while maintaining a low distortion of the near-end speech signal. Future research is required to further explore the impact of temporal partitioning of the echo path impulse response on echo reduction performance under various conditions.

7. REFERENCES

[1] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE Workshop on the Applications of Signal Process. to Audio and Acoust.*, New Paltz, NY, Oct. 2001, pp. 175–178.

[2] C. Faller and J. Chen, "Suppressing acoustic echo in a sampled spectral envelope space," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 13, pp. 1048–1062, Sept. 2005.

[3] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Signal Process.*, vol. 28, no. 1, pp. 55–69, Feb. 1980.

[4] Y. Avargel and I. Cohen, "Performance analysis of cross-band adaptation for subband acoustic echo cancellation," in *Proc. International Workshop on Acoust. Echo and Noise Control (IWAENC'06)*, Paris, France, Sept. 2006.

[5] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[7] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. Mohan Sondhi, and Y. Huang, Eds., chapter 45. Springer, 2007, part H.

[8] J. Benesty, H. Rey, L.R. Vega, and S. Tressens, "A non-parametric vss nlms algorithm," *IEEE Signal Process. Lett.*, vol. 13, pp. 581–584, Oct. 2006.

[9] E.A.P. Habets, "Room Impulse Response (RIR) generator," [Online] Available: http://home.tiscali.nl/ehabets/rir_generator.html, May 2008.