A Statistical Model Based Post-Filtering Algorithm for Residual Echo Suppression

Seung Yeol Lee, Jong Won Shin, Hwan Sik Yun and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University, Seoul, Korea
{sylee, jwshin, hsyun}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

In this paper, we propose a novel residual echo suppression (RES) algorithm constructed in the acoustic echo canceller. In the proposed approach, we introduce a statistical model to detect the signal components of the output signal and the state of signal is classified into four distinct hypothesis depending on the activity of near-end signal and residual echo. For hypothesis testing, the conventional likelihood ratio test is performed to make an optimal decision. The parameters specified in terms of the power spectral densities can be obtained by updating according to the hypothesis testing results and we can obtain the optimal RES filter by adopting the estimated parameters. The experimental results show that the proposed algorithm yields improved performance compared to that of the previous RES technique.

Index Terms: acoustic echo cancellation, residual echo suppression, post-filtering

1. Introduction

In two-way telecommunication, acoustic echo makes a serious conversation problem. To overcome this, acoustic echo cancellers (AEC's) have been developed for a comfortable conversation by reducing the effect of acoustic echo. In many practical applications, however, there still exists some amount of residual echo at the output of AEC filter. The difficulties of AEC are mainly due to the possible mismatch between the actual echo path and the employed adaptive filter structure, slow tracking capability of the adaptation algorithm, the influences which disturb the adaptive filter such as background noise, near-end speech and variations of the acoustic environment. In order to further reduce the residual echo, the residual echo suppression (RES) filter have been applied to AEC. Various RES post-filtering techniques [1]-[3] have been developed to obtain sufficient echo attenuation.

In this paper, we propose a novel RES algorithm based on a statistical model. In the proposed algorithm, the frequency response of the RES filter is determined differently according to the activity of near-end speech and residual echo. For this, all the possible signal conditions are classified into four categories depending on the presence or absence of the near-end speech and the residual echo. Identification of each category can be treated as a hypothesis testing problem and we apply a set of parametric models to perform likelihood ratio test. All the parameters are specified in terms of the power spectral densities (PSD's) of the relevant signals and their estimates are updated depending on the decision made by the hypothesis testing [4]. The optimal RES filter gain is given as a function of the signal-to-noise ratio (SNR) and signal-to-echo ratio (SER) obtained in a decision-directed method [5, 6]. The performance of the proposed algorithm is evaluated through echo return loss enhancement (ERLE) and speech attenuation tests.

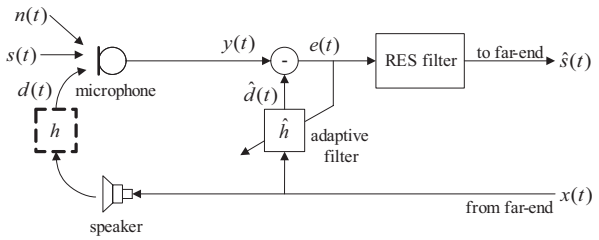


Figure 1: Block diagram of AEC system.

2. Signal detection based on a statistical model

A block diagram of the conventional AEC is shown in Fig. 1 where $x(t)$ represents the far-end signal at time t, h denotes the impulse response of the real acoustic echo path and $\hat{h}$ characterizes the corresponding echo path estimate provided by the adaptive filtering algorithm. Let $y(t)$ be the microphone signal and $e(t)$ be the AEC output signal which is to be transmitted to the far-end. Then,

$$y(t) = d(t) + s(t) + n(t) \quad (1)$$

$$\begin{aligned} e(t) &= (d(t) - \hat{d}(t)) + s(t) + n(t) \\ &= b(t) + s(t) + n(t) \end{aligned} \quad (2)$$

where $d(t)$ is the echo signal, $s(t)$ is the near-end signal, $n(t)$ is the near-end ambient noise, and $\hat{d}(t)$ is the adaptive filter output. In a practical implementation of the AEC, it is almost impossible to make the residual echo signal $b(t)$ completely suppressed due to the inherent modeling mismatch and the lack of adaptability. Therefore, the AEC output $e(t)$ is usually further processed by a RES post-filter.

RES is generally performed in the frequency domain. For this, near-end signal, $s(t)$, the ambient noise, $n(t)$ and the residual echo $b(t)$ in (2) are assumed to be stationary random processes. In addition, we also assume that all the signal components are statistically independent. Based on these assumptions, we can derive the equivalent frequency domain relations as follows:

$$E_k(m) = B_k(m) + S_k(m) + N_k(m) \quad (3)$$

where $E_k(m)$, $B_k(m)$, $S_k(m)$ and $N_k(m)$ represent the frequency domain spectra of $e(t)$, $b(t)$, $s(t)$ and $n(t)$, respectively, computed in the mth frame for kth frequency bin.

Let $\hat{s}(t)$ be the output of the RES post-filter and $\hat{S}_k(m)$ be the corresponding spectrum in the mth frame. Then,

$$\hat{S}_k(m) = F_k(m)E_k(m) \tag{4}$$

in which $F_k(m)$ denotes the gain of the RES post-filter in the mth frame for the kth frequency bin. Finally, $\hat{s}(t)$ is obtained by taking the inverse Fourier transform of $\hat{S}_k(m)$, and then transmitted to the far-end.

Let $\mathbf{E}(m) = [E_1(m), E_2(m), \ldots, E_M(m)]$ denote the spectrum of the AEC output signal at the mth frame, with $E_k(m)$ being the kth spectral component. Given four hypothesis, H_0, H_1, H_2 and H_3, depending on the presence or absence of the active near-end and far-end signal components, it is assumed that

$$\begin{array}{lcll} H_0 & : & \mathbf{E}(m) & = & \mathbf{N}(m) \\ H_1 & : & \mathbf{E}(m) & = & \mathbf{N}(m) + \mathbf{S}(m) \\ H_2 & : & \mathbf{E}(m) & = & \mathbf{N}(m) + \mathbf{B}(m) \\ H_3 & : & \mathbf{E}(m) & = & \mathbf{N}(m) + \mathbf{B}(m) + \mathbf{S}(m) \end{array} \tag{5}$$

where $\mathbf{N}(m) = [N_1(m), N_2(m), \ldots, N_M(m)]$, $\mathbf{B}(m) = [B_1(m), B_2(m), \ldots, B_M(m)]$ and $\mathbf{S}(m) = [S_1(m), S_2(m), \ldots, S_M(m)]$ represent the spectra of the ambient noise, the residual echo and near-end signal, respectively. We also assume that $\mathbf{N}(m)$, $\mathbf{B}(m)$ and $\mathbf{S}(m)$ are characterized by separate zero-mean complex Gaussian distributions and consequently, the following is obtained:

$$\begin{aligned} p(E_k(m)|H_0) &= \frac{1}{\pi\lambda_{n,k}(m)} \exp\left[-\frac{|E_k(m)|^2}{\lambda_{n,k}(m)}\right] \\ p(E_k(m)|H_1) &= \frac{1}{\pi(\lambda_{n,k}(m) + \lambda_{s,k}(m))} \\ &\quad \cdot \exp\left[-\frac{|E_k(m)|^2}{\lambda_{n,k}(m) + \lambda_{s,k}(m)}\right] \\ p(E_k(m)|H_2) &= \frac{1}{\pi(\lambda_{n,k}(m) + \lambda_{b,k}(m))} \\ &\quad \cdot \exp\left[-\frac{|E_k(m)|^2}{\lambda_{n,k}(m) + \lambda_{b,k}(m)}\right] \\ p(E_k(m)|H_3) &= \frac{1}{\pi(\lambda_{n,k}(m) + \lambda_{b,k}(m) + \lambda_{s,k}(m))} \\ &\quad \cdot \exp\left[-\frac{|E_k(m)|^2}{\lambda_{n,k}(m) + \lambda_{b,k}(m) + \lambda_{s,k}(m)}\right] \\ &\quad \text{for } k = 1, 2, \cdots, M \end{aligned} \tag{6}$$

in which $\lambda_{n,k}(m)$, $\lambda_{b,k}(m)$ and $\lambda_{s,k}(m)$ are the variances of the noise, residual echo and near-end signal in the kth frequency bin, respectively.

The problem of hypothesis testing in accordance with the presence of the near-end signal and residual echo is the same as the case considered in conventional noise suppression techniques [5]. When we apply the technique proposed in [4] to the current task of classifying hypotheses, the statistical model-based technique which computes the global state probabilities (GSP's), $p(H_i|\mathbf{E}(m)), i = 0, 1, 2, 3$ is adopted. Applying Bayes rule, it is easily derived that

$$\begin{aligned} p(H_i|\mathbf{E}(m)) &= \frac{p(\mathbf{E}(m)|H_i)p(H_i)}{p(\mathbf{E}(m))} \\ &= \frac{p(\mathbf{E}(m)|H_i)p(H_i)}{\sum_{j=0}^{3} p(\mathbf{E}(m)|H_j)p(H_j)} \,. \end{aligned} \tag{7}$$

Since the spectral component in each frequency bin is assumed to be statistically independent, (7) can be converted to

$$\begin{aligned} p(H_i|\mathbf{E}(m)) &= \frac{p(H_i)\prod_{k=1}^{M} p(E_k(m)|H_i)}{\sum_{j=0}^{3}\left[p(H_j)\prod_{k=1}^{M} p(E_k(m)|H_j)\right]} \\ &= \frac{1}{\sum_{j=0}^{3}\left[\frac{p(H_j)}{p(H_i)}\prod_{k=1}^{M}\Lambda_{ij}(E_k(m))\right]} \end{aligned} \tag{8}$$

in which $\Lambda_{ij}(E_k(m))$ is the likelihood ratio computed in the kth frequency bin such that

$$\Lambda_{ij}(E_k(m)) = \frac{p(E_k(m)|H_j)}{p(E_k(m)|H_i)} \,. \tag{9}$$

According to (6), the likelihood ratios, $\Lambda_{ij}(E_k(m))$ can be written as follows:

$$\begin{aligned} \Lambda_{01}(E_k(m)) &= \frac{1}{1+\xi_k(m)} \exp\left[\frac{\gamma_k(m)\xi_k(m)}{1+\xi_k(m)}\right] \\ \Lambda_{02}(E_k(m)) &= \frac{\zeta_k(m)}{\xi_k(m)+\zeta_k(m)} \exp\left[\frac{\gamma_k(m)\xi_k(m)}{\xi_k(m)+\zeta_k(m)}\right] \\ \Lambda_{03}(E_k(m)) &= \frac{\chi(m,\omega_k)}{\xi(m,\omega_k)(1+\chi(m,\omega_k))} \\ &\quad \times \exp\left[\frac{\psi(m,\omega_k)\xi(m,\omega_k)(1+\zeta(m,\omega_k))}{\zeta(m,\omega_k)(1+\chi(m,\omega_k))}\right] \\ \Lambda_{12}(E_k(m)) &= \frac{(1+\xi_k(m))\zeta_k(m)}{\xi_k(m)+\zeta_k(m)} \\ &\quad \times \exp\left[\frac{\gamma_k(m)\xi_k(m)(1-\zeta_k(m))}{(1+\xi_k(m))(\xi_k(m)+\zeta_k(m))}\right] \\ \Lambda_{13}(E_k(m)) &= \frac{(1+\xi_k(m))\chi_k(m)}{\xi_k(m)(1+\chi_k(m))} \\ &\quad \times \exp\left[\frac{\psi_k(m)\xi_k(m)}{(1+\xi_k(m))\zeta_k(m)(1+\chi_k(m))}\right] \\ \Lambda_{23}(E_k(m)) &= \frac{1}{1+\chi_k(m)} \exp\left[\frac{\psi_k(m)\chi_k(m)}{1+\chi_k(m)}\right] \\ \Lambda_{ij}(E_k(m)) &= 1 \,, \qquad \text{if} \quad i = j \\ \Lambda_{ij}(E_k(m)) &= \frac{1}{\Lambda_{ji}(E_k(m))} \,, \quad \text{if} \quad i > j \end{aligned} \tag{10}$$

where

$$\begin{aligned} \xi_k(m) &\equiv \frac{\lambda_{s,k}(m)}{\lambda_{n,k}(m)} \\ \gamma_k(m) &\equiv \frac{|E_k(m)|^2}{\lambda_{n,k}(m)} \\ \zeta_k(m) &\equiv \frac{\lambda_{s,k}(m)}{\lambda_{b,k}(m)} \\ \delta_k(m) &\equiv \frac{|E_k(m)|^2}{\lambda_{b,k}(m)} \,. \end{aligned} \tag{11}$$

In (11), $\xi_k(m)$, $\gamma_k(m)$, $\zeta_k(m)$ and $\delta_k(m)$ are referred to as the *a priori* signal-to-noise ratio (SNR), *a posteriori* SNR, *a priori* signal-to-echo ratio (SER) and *a posteriori* SER, respectively.

$\chi_k(m)$, $\psi_k(m)$ can be defined as

$$\chi_k(m) = \frac{\xi_k(m)\zeta_k(m)}{\xi_k(m)+\zeta_k(m)}$$
$$\psi_k(m) = \frac{\gamma_k(m)\delta_k(m)}{\gamma_k(m)+\delta_k(m)} \,. \quad (12)$$

Now, what remains is how to make a rule for hypothesis classification. Each GSP $p(H_i|\mathbf{E}(m))$ is calculated in mth frame, and the sum of all GSP's equal to one. If $p(H_i|\mathbf{E}(m))$ exceeds other GSP's, $p(H_j|\mathbf{E}(m)), j \neq i$, decision is made in favor of H_i. We can summarize the tests for detecting each condition as follows:

$$\begin{array}{lcl} H_0 & : & p(H_0|\mathbf{E}(m)) > p(H_j|\mathbf{E}(m)),\ \forall j \neq 0 \\ H_1 & : & p(H_1|\mathbf{E}(m)) > p(H_j|\mathbf{E}(m)),\ \forall j \neq 1 \\ H_2 & : & p(H_2|\mathbf{E}(m)) > p(H_j|\mathbf{E}(m)),\ \forall j \neq 2 \\ H_3 & : & p(H_3|\mathbf{E}(m)) > p(H_j|\mathbf{E}(m)),\ \forall j \neq 3 \,. \end{array} \quad (13)$$

3. Noise and residual echo PSD estimation

A crucial part of the RES operation requires a robust estimation of the PSD's for the relevant signal components. In this section, we describe the procedures for estimating the PSD's of the background noise and residual echo. Let $\hat{\lambda}_{n,k}(m)$ and $\hat{\lambda}_{b,k}(m)$ be the estimates for the PSD's of the background noise and residual echo, respectively. For robustness reasons, $\hat{\lambda}_{n,k}(m)$ should be updated only when H_0 is decided to be true. Updating $\hat{\lambda}_{b,k}(m)$ should be performed by considering the model of the loudspeaker-enclosure-microphone (LEM) system [7]. In a conventional LEM system, the residual echo spectrum is usually given by the product of the far-end signal spectrum, $X_k(m)$ with the frequency response of the system mismatch $H_{\Delta,k}(m)$ such that

$$B_k(m) = X_k(m)H_{\Delta,k}(m) \quad (14)$$

where $H_{\Delta,k}(m) = H_k(m) - \hat{H}_k(m)$, i.e. the difference of frequency response between the actual echo path, $H_k(m)$ and its estimate, $\hat{H}_k(m)$ [7]. By (14), a straightforward way to estimate the PSD of the residual echo is given by

$$\hat{\lambda}_{b,k}(m) = \hat{\lambda}_{x,k}(m)|H_{\Delta,k}(m)|^2 \quad (15)$$

where $\hat{\lambda}_{x,k}(m)$ is the PSD estimate of the far-end signal, $x(t)$. The squared magnitude response of the system mismatch, $|H_{\Delta,k}(m)|^2$, can be updated by means of the decision directed approach given as follows:

$$|H_{\Delta,k}(m)|^2 = \begin{cases} (1-\alpha_h)\dfrac{\hat{\lambda}_{e,k}(m)-\hat{\lambda}_{n,k}(m)}{\hat{\lambda}_{x,k}(m)} \\ \quad + \alpha_h|H_{\Delta,k}(m-1)|^2 \,, \\ \text{if } H_2 \text{ is chosen in } (m-1)\text{th frame} \\ |H_{\Delta,k}(m-1)|^2 \,, \text{ otherwise} \end{cases} \quad (16)$$

in which $0 < \alpha_h < 1$ is an appropriate smoothing parameter, and $\hat{\lambda}_{e,k}(m)$ represents the PSD estimate of the AEC output, $e(t)$. Both $\hat{\lambda}_{x,k}(m)$ and $\hat{\lambda}_{e,k}(m)$ can be easily updated through a first-order recursion with suitable smoothing parameters α_x and α_e.

Once the estimates for the PSD's of the background noise and residual echo are obtained, the next step is to update the $\xi_k(m)$ and $\zeta_k(m)$. Among many possible approaches, we apply the decision directed technique proposed in [5]. Given $\hat{\lambda}_{n,k}(m)$ and $\hat{\lambda}_{b,k}(m)$, the decision directed approach updates the estimated *a priori* SNR, $\hat{\xi}_k(m)$ and and *a priori* SER, $\hat{\zeta}_k(m)$ in the following way:

$$\hat{\xi}_k(m) = (1-\alpha_\xi)u(\hat{\gamma}_k(m)-1) + \alpha_\xi \frac{|F_k(m-1)E_k(m-1)|^2}{\hat{\lambda}_{n,k}(m)} \quad (17)$$

$$\hat{\zeta}_k(m) = (1-\alpha_\zeta)u(\hat{\delta}_k(m)-1) + \alpha_\zeta \frac{|F_k(m-1)E_k(m-1)|^2}{\hat{\lambda}_{b,k}(m)} \quad (18)$$

where $u(\cdot)$ is a unit step function, and $F_k(m-1)$ represents the gain of the RES post-filter computed in the previous frame at kth frequency bin. The estimated *a posteriori* SNR, $\hat{\gamma}_k(m)$ and *a posteriori* SER, $\hat{\delta}_k(m)$ can be obtained from the instantaneous spectrum of the AEC filter output signal, $E_k(m)$ such that

$$\hat{\gamma}_k(m) = \frac{|E_k(m)|^2}{\hat{\lambda}_{n,k}(m)} \quad (19)$$

$$\hat{\delta}_k(m) = \frac{|E_k(m)|^2}{\hat{\lambda}_{b,k}(m)} \,. \quad (20)$$

The optimal gain, $F_k(m)$ of the RES post-filter is described in terms of $\hat{\xi}_k(m)$ and $\hat{\zeta}_k(m)$. It is noted that $F_k(m)$ modifies the magnitude of $E_k(m)$ while retaining its phase. According to the minimum mean squared error criterion, it can be derived that

$$F_k(m) = \frac{\hat{\xi}_k(m)\hat{\zeta}_k(m)}{\hat{\xi}_k(m)\hat{\zeta}_k(m)+\hat{\xi}_k(m)+\hat{\zeta}_k(m)} \quad (21)$$

which is equivalent to the Wiener filtering solution [6].

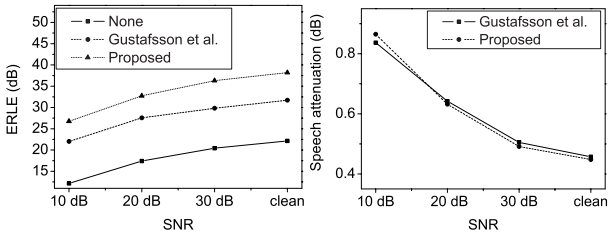
4. Experimental Results

In order to evaluate the performance of the proposed RES algorithm, we conducted computer simulations under various conditions. Twenty sentences were spoken by four speakers and sampled at 16 kHz. For performance assessment, we artificially created twenty data files such that each file was obtained by mixing the far-end signal with the near-end signal. The far-end speech was passed through a filter simulating the acoustic echo path before being mixed. The echo level measured at input microphone was 3.7 dB lower than that of the input speech on average. Two types of noise sources, the babble and vehicular noises from the NOISEX-92 database were added to the clean speech waveforms by varying SNR. To simulate the echo, the LEM system was modeled by a time-invariant FIR filter derived from an analysis of room acoustics. The simulation environment was designed to fit a small office room of a size $4 \times 3 \times 3\, m^3$. In order to estimate the echo, an adaptive filter with the number of filter taps, $L = 512$, was used and the coefficients of the AEC filter were adapted by means of the normalized least mean square (NLMS) algorithm with an adaptive step-size control strategy [2].

The performance of the RES approach was measured in terms of $ERLE(t)$ which is defined by

$$ERLE(t) = 10\log_{10}\left[\frac{E\{y^2(t)\}}{E\{\hat{s}^2(t)\}}\right] \text{ (dB)} \quad (22)$$

with $E\{\cdot\}$ denoting expected value at time t and ERLE denotes the corresponding value averaged over all time duration. For



(a) (b)

Figure 2: Performance of RES algorithms: (a) ERLE score and (b) speech attenuation during double-talk.

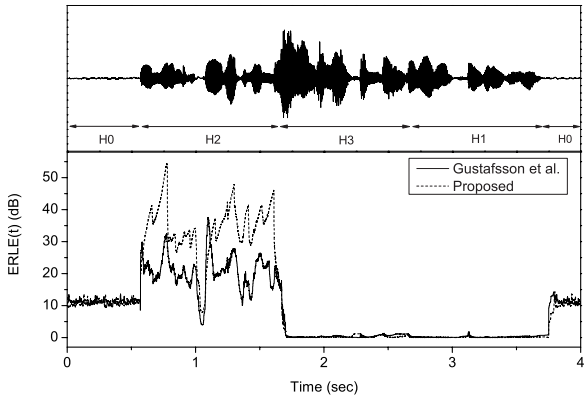


Figure 3: Time variation of $ERLE(t)$.

the purpose of comparison, we also evaluated the performance of the original AEC system without any RES module and the one with the RES algorithm proposed by Gustafsson et al. [3]. The overall results for ERLE are plotted in Fig. 2(a). From these results, we can observe that the ERLE's of the proposed algorithm were higher than those of the previous RES technique in all the tested conditions. Another important factor we should consider in the performance evaluation is the speech attenuation during the double-talk periods. The speech attenuation (SA) is defined as the ratio of the input speech power to the output speech power in the following way:

$$SA = \frac{1}{N}\sum^{N} 10\log_{10}\left(\frac{E[s^2(t)]}{E[\tilde{s}^2(t)]}\right) \qquad (23)$$

where N denotes the total sample number of double talk period and $\tilde{s}(t)$ represents the near-end signal component of output signal. The speech attenuation during the double-talk periods is shown in Fig. 2(b) where we can see that the proposed algorithm resulted in a similar level of attenuation compared to the conventional RES technique.

In Fig.3, an example of $ERLE(t)$ variation over time is given in conjunction with the corresponding waveform. We can observe that the proposed algorithm attenuated the residual echo more efficiently than the conventional RES technique while preserving the near-end signal quite well.

5. Conclusions

In this paper, we have presented a novel RES algorithm based on a statistical model. The principal contribution of this work is a systematic classification of the output signal state according to the existence of activity near-end and residual-echo signals. In order to test each hypothesis, a statistical approach resulting in likelihood ratio tests has been adopted. The PSD estimates of relevant signals are updated depending on the state decided by hypothesis testing and the optimal gain of RES post-filter can be found by adopting estimated parameters. The performance of the proposed approach has been found superior to that of the conventional technique through ERLE and speech attenuation evaluation tests.

6. Acknowledgements

This work was supported in part by ETRI SoC Industry Promotion Center and the Korea Science and Engineering Foundation (KOSEF) grant funded by Korea government (MOST) (No. R01-2007-000-10818-0).

7. References

[1] V. Turbin, A. Gilloire and P. Scalart, "Comparison of three post-filtering algorithms for residual echo reduction," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1997, pp. 307-310.

[2] E. Hänsler and G. Schmidt, "Hands-free telephones - joint control of each cancellation and postfiltering," *Signal Processing*, vol. 80, no. 11, pp. 2295-2305, Nov. 2000.

[3] S. Gustafsson, R. Martin, P. Jax and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 245-256, Jul. 2002.

[4] N. S. Kim and J. -H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108-110, May 2000.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[6] R. Le Bouquin Jeannés, P. Scalart, G. Faucon and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: a survey," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 8, pp. 808-820, Nov. 2001.

[7] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control : A Practical Approach*, Hoboken, NJ : John Wiley & Sons, 2004.