ENHANCEMENT OF SPEECH SIGNALS UNDER MULTIPLE HYPOTHESES USING AN INDICATOR FOR TRANSIENT NOISE PRESENCE

Ari Abramson and Israel Cohen

Department of Electrical Engineering, Technion - Israel Institute of Technology
Technion City, Haifa 3200, Israel
{aari@tx,icohen@ee}.technion.ac.il

ABSTRACT

In this paper, we formulate a speech enhancement problem under multiple hypotheses, assuming an indicator or detector for the transient noise presence is available in the short-time Fourier transform (STFT) domain. Hypothetical presence of speech or transient noise is considered in the observed spectral coefficients, and cost parameters control the trade-off between speech distortion and residual transient noise. An optimal estimator, which minimizes the mean-square error of the log-spectral amplitude, is derived, while taking into account the probability of erroneous detection. Experimental results demonstrate the improved performance in transient noise suppression, compared to using the optimally-modified log-spectral amplitude estimator.

Index Terms— Speech enhancement, acoustic signal detection, transient noise, acoustic noise

1. INTRODUCTION

Enhancement of speech signals is of great interest in many voice communication systems, whenever the source signal is corrupted by noise. In a highly non-stationary noise environments, noise transients may be extremely annoying and significantly degrade the perceived quality and performances of subsequent coding or speech recognition systems. Existing speech enhancement algorithms, *e.g.*, [1, 2], are generally inadequate for eliminating non-stationary noise components.

In some applications, an indicator for the transient noise activity may be available, *e.g.*, a siren noise in an emergency car, lens-motor noise of a digital video camera or a keyboard typing noise in a computer-based communication system. The transient spectral variances can be estimated in such cases from training signals. However, applying a standard estimator to the spectral coefficients may result in removal of critical speech components in case of falsely detecting the speech components, or under-suppression of transient noise in case of missing to detect the noise transients.

This research was supported by the Israel Science Foundation (grant no. 1085/05) and by the European Commission's IST program under project Memories.

In this paper, we formulate a speech enhancement problem under multiple hypotheses, assuming some indicator or detector for the presence of noise transients in the STFT domain is available. Cost parameters control the trade-off between speech distortion and residual transient noise. We derive an optimal signal estimator that employs the available detector and show that the resulting estimator generalizes the optimally-modified log-spectral amplitude (OM-LSA) estimator [2]. Experimental results demonstrate the improved performance obtained by the proposed algorithm, compared to using the OM-LSA.

This paper is organized as follows. In Section 2 we formulate the problem of spectral enhancement under multiple hypotheses. In Section 3 we derive the optimal estimator. In Section 4 we provide some experimental results and conclude in Section 5.

2. PROBLEM FORMULATION

Let $x(n)$, $d^s(n)$ and $d^t(n)$ denote speech and two uncorrelated additive interference signals, respectively, and let $y(n) = x(n) + d^s(n) + d^t(n)$ be the observed signal. We assume that $d^s(n)$ is a quasi-stationary background noise while $d^t(n)$ is a highly non-stationary transient signal. The speech signal and the transient noise are not always present in the STFT domain, so we have four hypotheses for the noisy coefficients:

$$\begin{aligned} H_{1s}^{\ell k} &: Y_{\ell k} = X_{\ell k} + D_{\ell k}^{s}\,, \\ H_{1t}^{\ell k} &: Y_{\ell k} = X_{\ell k} + D_{\ell k}^{s} + D_{\ell k}^{t}\,, \\ H_{0s}^{\ell k} &: Y_{\ell k} = D_{\ell k}^{s}\,, \\ H_{0t}^{\ell k} &: Y_{\ell k} = D_{\ell k}^{s} + D_{\ell k}^{t}\,, \end{aligned} \tag{1}$$

where ℓ denotes the time frame index and k denotes the frequency-bin index.

In many speech enhancement applications, an indicator for the transient source may be available, *e.g.*, siren noise in an emergency car, keyboard typing in computer-based communication system and a lens-motor noise in a digital video

camera. In such cases, *a priori* information based on a training phase may yield a reliable detector for the transient noise. However, false detection of transient noise components when signal components are present may significantly degrade the speech quality and intelligibility. Furthermore, missed detection of transient noise components may result in a residual transient noise, which is perceptually annoying.

Let $\eta_j^{\ell k}$, $j \in \{0,1\}$ denote the detector decision in the time-frequency bin (ℓ,k), *i.e.*, a transient component is classified as a speech component under η_1 and as a noise component under η_0[1]. Let C_{10} denote the false-alarm cost with relation to the noise transient, *i.e.*, cost of making a decision η_0 when a noise transient is inactive or is not dominant w.r.t the speech component, and let the missed detection cost C_{01} be defined similarly. Let $d(x,y) \triangleq (\log|x| - \log|y|)^2$ denote the squared log-amplitude distortion function, let $A_{\ell k} \triangleq |X_{\ell k}|$ and let $R_{\ell k} \triangleq |Y_{\ell k}|$. Considering a realistic detector, we introduce the following criterion for the estimation of the speech expansion coefficient under the decision $\eta_j^{\ell k}$:

$$\begin{aligned}\hat{A}_{\ell k} = \arg\min_{\hat{A}} \Big\{ & C_{1j} p\left(H_{1s}^{\ell k} \cup H_{1t}^{\ell k} \,|\, \eta_j^{\ell k}, Y_{\ell k}\right) \\ & \times E\left[d\left(X_{\ell k}, \hat{A}\right) \,|\, Y_{\ell k}, H_{1s}^{\ell k} \cup H_{1t}^{\ell k} \right] \\ & + C_{0j} p\left(H_{0t}^{\ell k} \cup H_{0s}^{\ell k} \,|\, \eta_j^{\ell k}, Y_{\ell k}\right) d\left(G_{\min} R_{\ell k}, \hat{A}\right) \Big\} \end{aligned} \quad (2)$$

where the costs of perfect detection C_{00} and C_{11} are normalized to one. That is, under speech presence we aim at minimizing the MSE of the LSA. Otherwise, a constant attenuation $G_{\min} << 1$ is imposed for maintaining naturalness of the residual noise [2]. The cost parameters control the trade-off between speech distortion, consequent upon false detection of noise transients, and residual transient noise, resulting from missed detection of transient noise components.

3. OPTIMAL ESTIMATION UNDER A GIVEN DETECTION

In this section we derive an optimal estimator for the speech signal under multiple hypotheses.

3.1. Spectral Estimation

We first reduce the problem into two basic hypotheses, $H_1^{\ell k}$ and $H_0^{\ell k}$. Under $H_1^{\ell k}$, the speech component is assumed present and more dominant than the noise component. This hypothesis includes $H_{1s}^{\ell k}$ as well as $H_{1t}^{\ell k}$ given that $|X_{\ell k}| \geq \beta |D_{\ell k}^t|$, where $\beta > 0$ is a predefined threshold parameter. The hypothesis $H_0^{\ell k}$ includes the cases $H_{0s}^{\ell k}$, $H_{0t}^{\ell k}$ and also $H_{1t}^{\ell k}$ with $|X_{\ell k}| < \beta |D_{\ell k}^t|$. Under $H_1^{\ell k}$ we estimate the speech in the MMSE-LSA sense, and under $H_0^{\ell k}$ we impose a constant attenuation to the noisy component. Note that ideally under $H_{1t}^{\ell k}$ an estimate for the speech component would be desired. However, if the noise transient is much more dominant we would better apply the constant low attenuation to the noisy component to avoid a strong residual noisy transient.

Let $p_{ij} \triangleq p\left(\eta_j^{\ell k} \,|\, H_i^{\ell k}\right)$. We are interested in detecting the interfering transient noise so p_{01} is the probability of a false alarm and p_{10} is the probability of missed detection. We assume that given any transient in the noisy coefficients, the detection error probability is independent of the observation and the signal-to-noise ratio (SNR). Therefore, $p\left(\eta_j^{\ell k} \,|\, H_i^{\ell k}, Y_{\ell k}\right) = p_{ij}$ and

$$p\left(H_i^{\ell k} \,|\, \eta_j^{\ell k}, Y_{\ell k}\right) = p_{ij}\, p\left(H_i^{\ell k} \,|\, Y_{\ell k}\right) / p\left(\eta_j^{\ell k} \,|\, Y_{\ell k}\right) . \quad (3)$$

This assumption can be easily relaxed by employing a time-frequency dependent probability $p_{ij}^{\ell k}$. Considering the two basic hypotheses and substituting (3) into (2) we obtain

$$\begin{aligned}\hat{A}_{\ell k} = \arg\min_{\hat{A}} \Big\{ & p_{1j} C_{1j} p\left(H_1^{\ell k} \,|\, Y_{\ell k}\right) \\ & \times \int d\left(X_{\ell k}, \hat{A}\right) p\left(X_{\ell k} \,|\, Y_{\ell k}, H_1^{\ell k}\right) dX_{\ell k} \\ & + \; p_{0j} C_{0j} p\left(H_0^{\ell k} \,|\, Y_{\ell k}\right) d\left(G_{\min} R_{\ell k}, \hat{A}\right) \Big\} , \end{aligned} \quad (4)$$

which yields

$$\begin{aligned}\log \hat{A}_{\ell k} \left[p_{1j} C_{1j} p\left(H_1^{\ell k} \,|\, Y_{\ell k}\right) + p_{0j} C_{0j} p\left(H_0^{\ell k} \,|\, Y_{\ell k}\right) \right] = \\ p_{1j} C_{1j} p\left(H_1^{\ell k} \,|\, Y_{\ell k}\right) E\left\{ \log|X_{\ell k}| \,|\, Y_{\ell k}, H_1^{\ell k} \right\} \\ + p_{0j} C_{0j} p\left(H_0^{\ell k} \,|\, Y_{\ell k}\right) \log\left(G_{\min} R_{\ell k}\right) . \end{aligned} \quad (5)$$

Let $\xi_{\ell k}$ and $\gamma_{\ell k}$ denote the *a priori* and *a posteriori* SNRs, respectively[2], let $\upsilon_{\ell k} \triangleq \xi_{\ell k}\gamma_{\ell k}/(1+\xi_{\ell k})$ and let

$$\begin{aligned}\Lambda\left(\xi_{\ell k}, \gamma_{\ell k}\right) &\triangleq \frac{p\left(H_1^{\ell k}\right)}{p\left(H_0^{\ell k}\right)} \frac{p\left(Y_{\ell k} \,|\, H_1^{\ell k}\right)}{p\left(Y_{\ell k} \,|\, H_0^{\ell k}\right)} \\ &= \frac{p\left(H_1^{\ell k}\right)}{p\left(H_0^{\ell k}\right)} \frac{e^{\upsilon_{\ell k}}}{1+\xi_{\ell k}} \end{aligned} \quad (6)$$

denote the generalized likelihood ratio [1]. Accordingly, $p\left(H_1^{\ell k} \,|\, Y_{\ell k}\right) = \Lambda\left(\xi_{\ell k}, \gamma_{\ell k}\right) / \left(1 + \Lambda\left(\xi_{\ell k}, \gamma_{\ell k}\right)\right)$. Let $\phi_j\left(\xi_{\ell k}, \gamma_{\ell k}\right) = p_{1j} C_{1j} \Lambda\left(\xi_{\ell k}, \gamma_{\ell k}\right) + p_{0j} C_{0j}$ and let

$$G_{LSA}\left(\xi, \gamma\right) \triangleq \frac{\xi}{1+\xi} \exp\left(\frac{1}{2} \int_{\vartheta}^{\infty} \frac{e^{-t}}{t} dt \right) \quad (7)$$

denote the LSA gain function [3]. Then, combining the magnitude estimate $\hat{A}_{\ell k}$ with the phase of the noisy spectral coefficient $Y_{\ell k}$ we obtain an optimal estimate under the decision $\eta_j^{\ell k}$:

$$\begin{aligned}\hat{X}_{\ell k} &= \left[G_{\min}^{p_{0j} C_{0j}} G_{LSA}\left(\xi_{\ell k}, \gamma_{\ell k}\right)^{p_{1j} C_{1j} \Lambda} \right]^{\phi_j^{-1}} Y_{\ell k} \\ &\triangleq G_{\eta_j}\left(\xi_{\ell k}, \gamma_{\ell k}\right) Y_{\ell k} , \end{aligned} \quad (8)$$

[1] Note that the detector is used for discriminating between transient speech components and transient noise components, and therefore not employed when transients are absent.

[2] Note that the noise variance depends on whether a transient component is present or not. This will be specified in the next subsection.

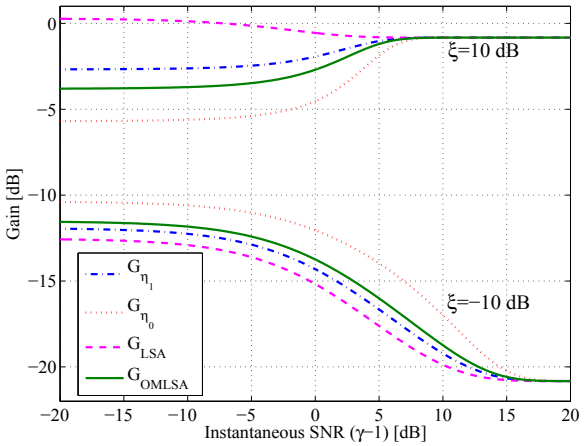

Fig. 1. Gain curves for $p(H_1) = 0.8$, $C_{01} = 5$, $C_{10} = 3$, $G_{\min} = -15$ [dB] and false-detection and missed-detection probabilities of $p_{01} = p_{10} = 0.1$.

where Λ and ϕ_j hold for $\Lambda(\xi_{\ell k}, \gamma_{\ell k})$ and $\phi_j(\xi_{\ell k}, \gamma_{\ell k})$, respectively.

In case of a decision η_1 (*i.e.*, transient component is classified as speech), the missed-detection cost C_{01} as well the probabilities p_{01} and p_{11} control the trade-off between the attenuation associated with the hypothesis H_1 and the constant attenuation under speech absence, $G_{\min}$. Under a decision η_0, the trade-off is controlled by the false-alarm cost and the probabilities p_{00} and p_{10}.

Note that in case $p_{0j} = p_{1j}$ and $C_{0j} = C_{1j}$ for $j \in \{0, 1\}$, the estimator (8) reduces to the OM-LSA estimator [2] under any of the detector decisions, since in that case the decision made by the detector does not contribute any statistical information.

Figure 1 shows attenuation curves as a function of the *instantaneous* SNR, $\gamma - 1$, for different *a priori* SNRs. The detection-dependent gains G_{η_0} (dashed-dotted line) and G_{η_1} (dotted line) are compared to the LSA gain (dashed line) and the OM-LSA gain (solid line) [3, 2]. It shows that the cost parameters with the error probabilities of the detector shape the attenuation curve under any of the decisions made by the detector to compensate for any erroneous detection.

3.2. A priori and a posteriori SNR estimation

The spectrum of the background noise, $\lambda_{s,\ell k} \triangleq E\left\{|D^s_{\ell k}|^2\right\}$, can be estimated by using the minima-controlled recursive averaging algorithm [4]. The *a priori* signal-to-stationary noise ratio $\xi^s_{\ell k} \triangleq \lambda_{x,\ell k}/\lambda_{s,\ell k}$, where $\lambda_{x,\ell k} \triangleq E\left\{|X_{\ell k}|^2\right\}$, is practically estimated using the decision-directed approach [1, 2]. Given that a transient noise is present, the transient noise spectrum may be estimated from a training phase. Therefore, under η_0 we may estimate the *a priori* and *a posteriori* SNRs by using $\hat{\lambda}_{s,\ell k} + \hat{\lambda}_{t,\ell k}$ as the estimate for the noise spectrum [5], where $\lambda_{t,\ell k}$ is defined similarly to $\lambda_{s,\ell k}$. However, in case of an erroneous detection, this approach may significantly distort the speech component, since both the *a priori* and *a posteriori* SNRs would be much smaller than their desired values. Therefore, we propose to smooth the noisy spectra

$$\zeta_{\ell k} = \mu\zeta_{\ell-1,k} + (1-\mu)\,|Y_{\ell k}|^2\,, \tag{9}$$

with $0 < \mu < 1$. Accordingly, under a decision $\eta_0^{\ell k}$ we update the estimates such that

$$\begin{aligned} \eta_1^{\ell k} &: \hat{\xi}_{\ell k} = \hat{\xi}^s_{\ell k}\,, \quad \hat{\gamma}_{\ell k} = \hat{\gamma}^s_{\ell k}\,, \\ \eta_0^{\ell k} &: \hat{\xi}_{\ell k} = \hat{\xi}^s_{\ell k}\frac{\hat{\lambda}^s_{d,\ell k}}{\zeta_{\ell k}}\,, \quad \hat{\gamma}_{\ell k} = \hat{\gamma}^s_{\ell k}\frac{\hat{\lambda}^s_{d,\ell k}}{\zeta_{\ell k}}\,. \end{aligned} \tag{10}$$

As a result, the outcome of falsely detecting transient noise is less destructive since $\zeta_{\ell k}$ would be much smaller than $\lambda_{s,\ell k} + \lambda_{t,\ell k}$. However, in case of a perfect detection, $\zeta_{\ell k}$ is a reliable estimator for the noise spectrum given that μ is sufficiently small. In addition, under the existence of a high energy transient component we would like to further attenuate the noisy component to the level of the residual background noise. Therefore, under $\eta_0^{\ell k}$ we update $\tilde{G}_{\min} = G_{\min}\sqrt{\hat{\lambda}_{s,\ell k}/\zeta_{\ell k}}$.

4. EXPERIMENTAL RESULTS

In this section, we demonstrate the application of the proposed algorithm to speech enhancement in a typical office communication system, based on the DUET Conference Speakerphone of Phoenix Audio Technologies. The background office noise is slowly-varying while possible keyboard typing interference may exist. Since the keyboard signal is available to the computer, a reliable detector for the transient-like keyboard noise is assumed to be available based on a training phase but still, erroneous detections are reasonable. The speech signals are sampled at 16 kHz and degraded by a stationary background noise with 15 dB SNR and a keyboard typing noise such that the total SNR is 0.8 dB. The STFT is applied to the noisy signal with Hamming windows of 32 msec length and 75% overlap. The transient noise detector is assumed to have an error probability of 10% and the missed-detection and false-detection costs are set to 1.2. The weighting factor for the noisy spectra is $\mu = 0.5$.

Figure 2 demonstrates the spectrograms and waveforms of a signal enhanced by using the proposed algorithm, compared to using the OM-LSA algorithm. It can be seen that using our approach, the transient noise is significantly attenuated, while the OM-LSA is unable to eliminate the keyboard transients.

The objective evaluation includes three quality measures: segmental SNR (SegSNR), log-spectral distortion (LSD) and perceptual evaluation of speech quality (PESQ) score. The results are summarized in Table 1. It can be seen that the

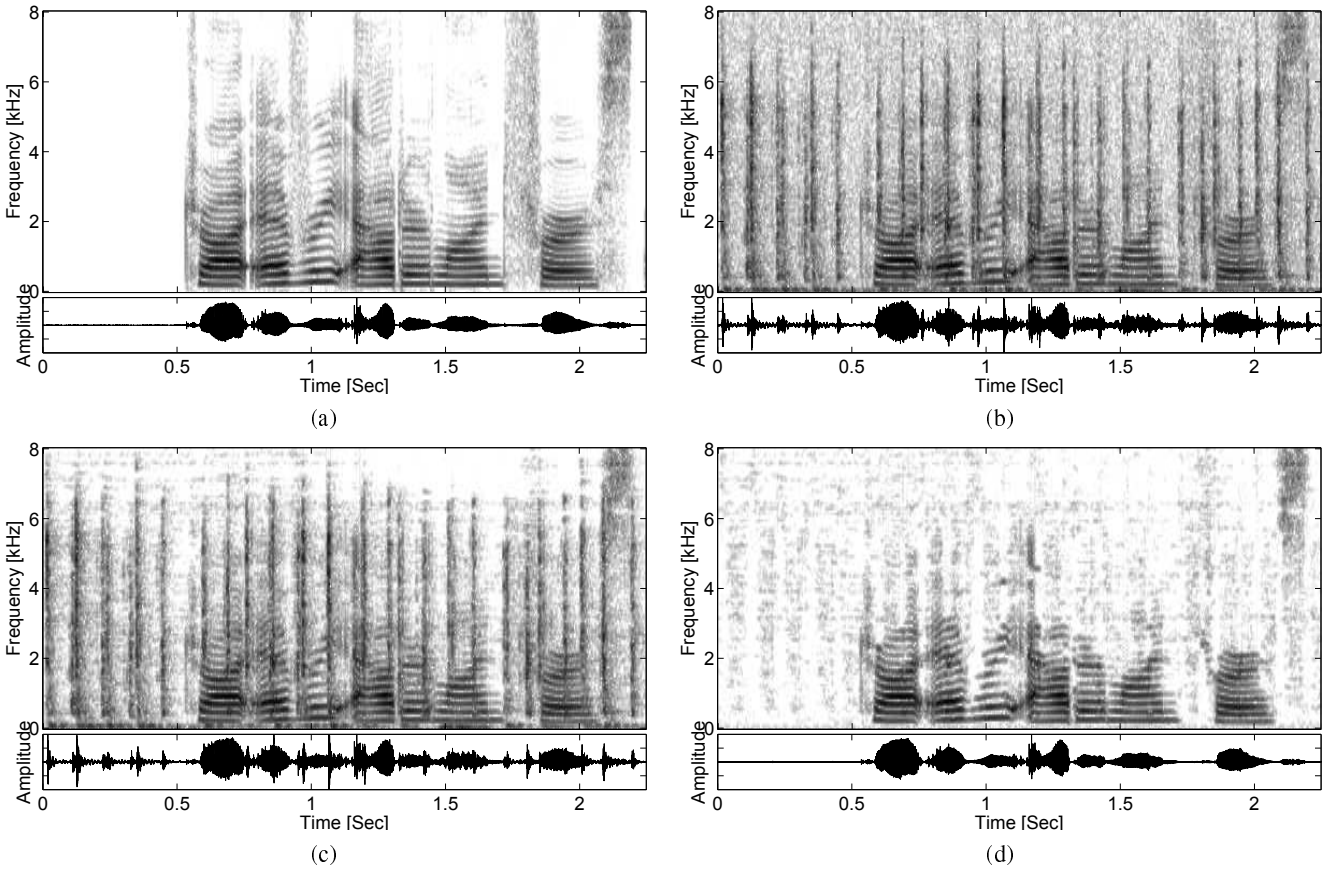

Fig. 2. Speech spectrograms and waveforms. (a) Clean signal ("Try any other line first"); (b) noisy signal (office noise including keyboard typing noise, SNR=0.8 dB); (c) speech enhanced by using the OM-LSA estimator; (d) speech enhanced by using the proposed algorithm.

Table 1. Segmental SNR and Log Spectral Distortion Obtained Using the OM-LSA and the Proposed Algorithm.

Method	SegSNR [dB]	LSD [dB]	PESQ
Noisy speech	-2.23	7.69	1.07
OM-LSA	-1.31	6.77	0.97
Proposed Alg.	5.41	1.67	2.87

proposed detection and estimation approach significantly improves speech quality compared to using the OM-LSA algorithm. Informal listening tests confirm that the annoying keyboard typing noise is dramatically reduced and the speech quality is significantly improved.

5. CONCLUSIONS

We have introduced a new approach for a single-channel speech enhancement in a highly non-stationary noise environment where a reliable detector for interfering transients is available. The speech expansion coefficients are estimated under multiple-hypotheses in the MMSE-LSA sense while considering possible erroneous detection. The proposed algorithm generalizes the OM-LSA estimator and enables greater suppression of transient noise components.

6. ACKNOWLEDGMENT

The authors thank Phoenix Audio Technologies for providing audio equipment and for their helpful technical support.

7. REFERENCES

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

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

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 466–475, Sept. 2003.

[5] E. Habets, I. Cohen, and S. Gannot, "MMSE log-spectral amplitude estimator for multiple interferences," in *Proc. Int. Workshop on Acoust. Echo and Noise Control., IWAENC-06*, Paris, France, Sept. 2006.