LENS-MOTOR NOISE SUPPRESSION FOR DIGITAL CAMERAS

Avihay Barazany, Royi Levy and Yekutiel Avargel

Signal & Image Processing Lab, Department of Electrical Engineering,
Technion - Israel Institute of Technology
Technion City, Haifa 3200, Israel
{savihay, levy.royi}@gmail.com, kutiav@tx.technion.ac.il

ABSTRACT

In this paper, we formulate a speech enhancement problem under multiple hypotheses, assuming an indicator for the lens motor noise presence is available in the time domain. This approach is largely based on an existing speech enhancement algorithm [3], and was adapted to specifically handle lens motor noises. Hypothetical presence of speech together with an estimation of the stationary noise and the lens motor noise, are considered in the calculation of the desired signal spectral coefficients. 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 a substantial suppression of the motor noise with a minimal degrade to the perceived quality of the desired signal.

Index Terms— Speech enhancement, transient noise, noise suppression, spectral analysis

1. INTRODUCTION

Today, digital still cameras are widely used for video and audio recordings. When activating the lens-motor during these recordings (for zooming in or out a specific object), the noise generated by the motor may be recorded by the camera's microphone together with the desired audio signal. This noise, which is highly non-stationary, may be extremely annoying and significantly degrade the perceived quality and intelligibility of the desired signal. To solve this problem, many digital-cameras manufacturers disable the option of activating the lens motor during audio recordings, while others alternatively attempt to reduce the noise by acoustically isolating the lens motor.

In this paper, we introduce a modified algorithm, based on an existing speech enhancement algorithm [3], which fit to deal with lens-motor noise. The problem is formulated as a speech enhancement problem under multiple hypotheses, by using cost parameters to 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 assuming an indication of motor-noise presence in the time domain. We correspondingly construct a reliable noise detector in the time-frequency domain, and use training recording sets to derive an estimator for the lens-motor noise.

Experimental results with real speech signals demonstrate the advantage of the proposed approach. A substantial suppression of the motor noise is achieved without degrading the perceived quality of the desired signal. A significant improvement in performance is attained over conventional speech enhancement techniques, when applied to this problem.

This paper is organized as followed:

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 the speech signal, background stationary noise, and zoom motor (non-stationary) noise, 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 digital cameras, an indicator for the lens motor noise is available using the camera control board. In this case, *a priori* information based on a recording set may yield a reliable detector for the lens motor noise.

However, false detection of transient noise components when signal components are present may significantly degrade the speech quality and intelligibility. Furthermore, miss 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 $\eta_1^{\ell k}$ and as a noise component under $\eta_0^{\ell k}$[1]. Let C_{10} denote the false-alarm cost with relation to the noise transient, similarly, let C_{01} denote the miss detection cost. Let $d(x,y) \triangleq \left(\log|x| - \log|y|\right)^2$ 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 $\eta_j^{\ell k}$:

$$\begin{aligned}\hat{A}_{\ell k} = \arg\min_{\hat{A}} \Big\{ & C_{1j}\, p\left(H_1^{\ell k} \mid \eta_j^{\ell k}, Y_{\ell k}\right) \\ & \times E\left[d\left(X_{\ell k}, \hat{A}\right) \mid Y_{\ell k}, H_1^{\ell k} \right] \\ & + C_{0j}\, p\left(H_0^{\ell k} \mid \eta_j^{\ell k}, Y_{\ell k}\right) d\left(G_{\min} R_{\ell k}, \hat{A}\right) \Big\}\end{aligned} \qquad (2)$$

where the costs of perfect detection C_{11} and C_{00} 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 [3]. The cost parameters control the tradeoff between speech distortion, consequent upon false detection of noise transients, and residual transient noise, resulting from miss 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 $H_{1t}^{\ell k}$. The hypothesis $H_0^{\ell k}$ includes the cases $H_{0s}^{\ell k}$, $H_{0t}^{\ell k}$ and also $H_{1t}^{\ell k}$ in case the noise component is more dominant then the speech component. 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 $H_1^{\ell k}$ 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 $\xi_{\ell k}$ and $\gamma_{\ell k}$ denote the *a prio*ri and *a posteriori* SNRs, respectively[2]. 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 $\eta_j^{\ell k}$:

$$\hat{X}_{lk} = G_{\eta_j}(\xi_{lk}, \gamma_{lk}) Y_{lk} \qquad (3)$$

while $G_{\eta_j}(\xi_{lk}, \gamma_{lk})$ denote the gain function [#Ref].

In case the lens motor is inactive (according to the camera's indicator) the estimator (3) reduces to the OM-LSA estimator [5].

3.2 A priori and a posteriori SNR estimation

The spectrum of the background noise $\lambda_{s,\ell k} \triangleq E\left\{ \left| D_{\ell k}^s \right|^2 \right\}$ can be estimated using the Minima-Controlled recursive averaging algorithm [2]. The spectrum estimation will not be updated during the lens motor operation according to the camera's indicator. The *a priori* signal-to-noise ratio estimation $\hat{\xi}_{\ell k} = \hat{\lambda}_{x,\ell k} / \left(\hat{\lambda}_{s,\ell k} + \hat{\lambda}_{t,\ell k} \right)$, where $\lambda_{x,\ell k} \triangleq E\left\{ \left| X_{\ell k} \right|^2 \right\}$, is practically estimated using the decision-directed approach [1], with the modification of the additional lens motor noise to the stationary noise:

$$\begin{aligned}\hat{\lambda}_{x,lk} = \max \Big\{ & \alpha G_{LSA}{}^2\left(\hat{\xi}_{l-1,k}, \gamma_{l-1,k}\right) \left| Y_{l-1,k} \right|^2 \\ & + (1-\alpha)\left(\left| Y_{l,k} \right|^2 - \hat{\lambda}_s - \hat{\lambda}_t \right),\ \lambda_{\min} \Big\}\end{aligned} \qquad (4)$$

$$\lambda_{\min} = \hat{\xi}_{l,k} \hat{\lambda}_{s,lk}$$

Training recording sets is used to derive an *a priori* estimator, λ_0, for the lens-motor noise. Given that lens motor camera's indicator is on, this estimator is updated using the followed updating rule:

$$\begin{aligned}\tilde{H}_0 : \hat{\lambda}_t(\ell,k) = {} & \alpha\lambda_0(\ell,k) + (1-\alpha)\Big\{ \beta \hat{\lambda}_t(\ell-1,k) \\ & + (1-\beta)\left[\left| Y(\ell,k) \right|^2 - \hat{\lambda}_s(\ell,k) \right] \Big\} \\ \tilde{H}_1 : \hat{\lambda}_t(\ell,k) = {} & \alpha\lambda_0(\ell,k) + (1-\alpha)\hat{\lambda}_t(\ell-1,k)\end{aligned} \qquad (5)$$

[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.

If the motor noise is more dominant than the speech signal, it will be classified as $\tilde{H}_0$, Otherwise, $\tilde{H}_1$. This classification can be executed in several methods. We have used the following: The frequencies above 4KHz (low speech energy in these frequencies) are classified as $\tilde{H}_0$. The same goes for high amplitude harmonies above an empiric threshold with a low speech-presence probability. The rest of the spectrum will be classified as $\tilde{H}_1$.

The parameters α and β are constant smoothing factors [0..1] that control whether to rely further more on the *a priori* estimator, λ_0, or on input signal estimation.

In order to suppress the stationary noise when speech is absent, the OM-LSA algorithm uses a constant attenuation function G_f. In this problem, where both stationary and transient noises exist, the attenuation function should attenuate both noises in the same way the OM-LSA attenuate the stationary noise. Hence, minimizing the next equation:

$$\arg\min_{G_{\min}} \left\{ E\left[G_{\min}\left(\lambda_{s,lk}+\lambda_{t,lk}\right) - G_f \lambda_{s,lk} \right] \right\} \qquad (6)$$

Yields the desired solution:

$$G_{\min} = G_f \frac{\lambda_{s,lk}}{\lambda_{s,lk}+\lambda_{t,lk}} \qquad (7)$$

4. EXPERIMENTAL RESULTS

In this section, we demonstrate the application of the proposed algorithm to speech enhancement on a speech signal noised by background and lens motor noise of a camera. The signals are sampled at 11,025Hz and degraded by a stationary background noise with 15dB SNR and with a lens motor noise at 8dB SNR. The STFT is applied to the noisy signal with Hanning windows of 32 msec length and 75% overlap.

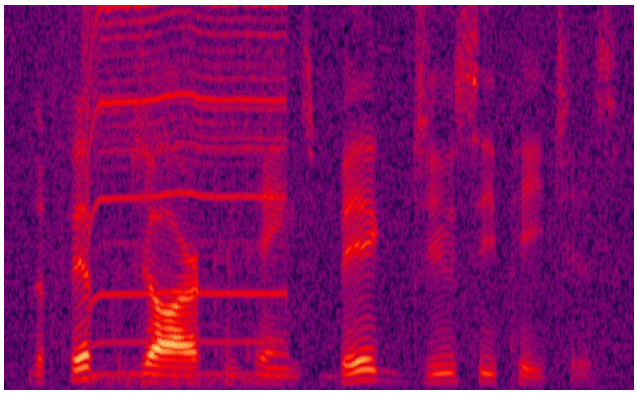

Fig. 1. Input signal (include motor noise, SNR=8 dB).

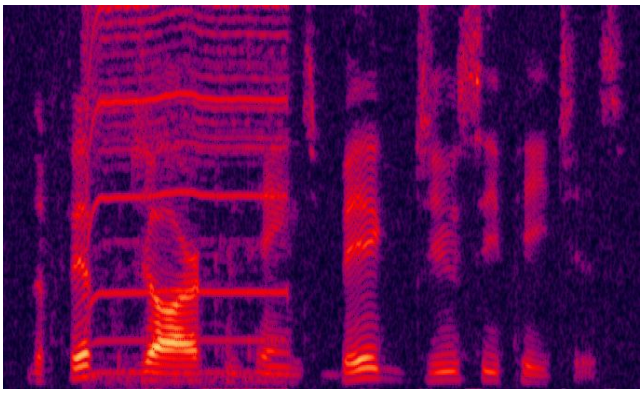

Fig. 2. Speech enhanced by using OM-LSA estimator

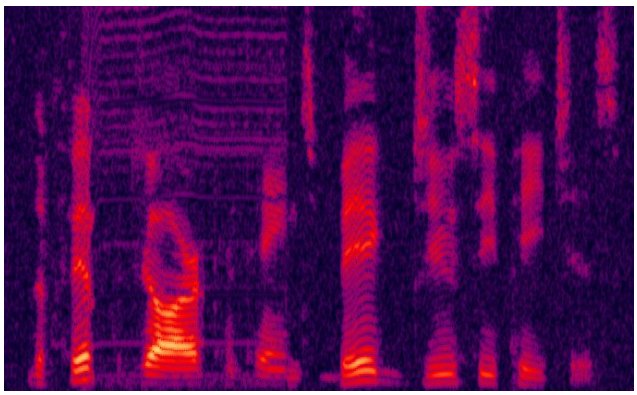

Fig. 3. Speech enhanced by using the proposed algorithm, $G_f = 15[\text{dB}]$

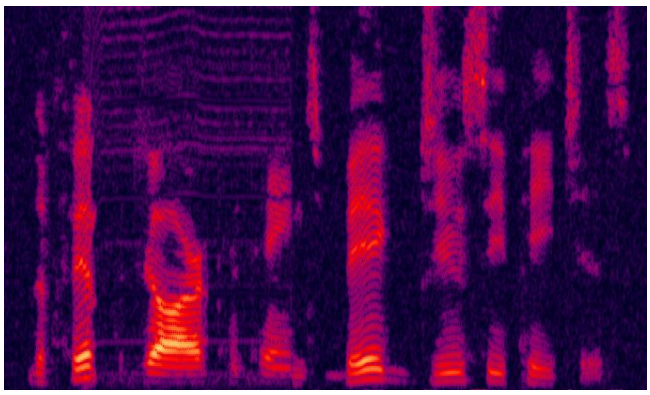

Fig. 4. Speech enhanced by using the proposed algorithm, $G_f = 20[\text{dB}]$

Figure 1 is a spectrogram of the input signal as explained above. Figure 2 demonstrates the input signal enhanced by using the OM-LSA algorithm. It can be seen that using the OM-LSA algorithm, the stationary noise is attenuated, but the lens motor noise is left untouched. Figure 3 and figure 4 demonstrate the same input signal enhanced by using our approach, with difference in the constant attenuation level. It can be seen that the lens motor noise is significantly attenuated, while the stationary noise is attenuated in the same levels like in the OM-LSA results. The results are valid for partial zoom as well.

5. CONCLUSIONS

We have introduced an approach for a single-channel speech enhancement in a non-stationary noise environment where a reliable detector for the lens motor noise, and a recording set of that noise 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 lens motor noise components.

6. ACKNOWLEDGMENT

The work on this algorithm was performed in the Signal and Image Processing laboratory, Dept. of EE, Technion IIT.
The authors are grateful to Nimrod Peleg and to Orly Wigderson from the Control and Robotics laboratory for their help which allowed us to carry out this project.

7. REFERENCES

[1] Cohen and B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments", *Signal Processing*, Vol. 81, No. 11, pp. 2403-2418 , Nov. 2001.

[2] Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement", *Signal Processing*, Vol. 9, Issue 1, pp. 12 – 15, Jan 2002.

[3] Abramson and I. Cohen," Enhancement of Speech Signals Under Multiple Hypotheses Using an Indicator for Transient Noise Presence " *Proc. 31th IEEE Internat.*

[4] A., Abramson, I. Cohen, "Simultaneous Detection and Estimation Approach for Speech Enhancement", *Audio, Speech, and Language Processing, IEEE Transactions* on Vol. 15, Issue 8, pp. 2348 – 2359 , Nov. 2007.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator". *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443.445, April 1985.