MMSE LOG-SPECTRAL AMPLITUDE ESTIMATOR FOR MULTIPLE INTERFERENCES

[1]*Emanuël A.P. Habets,* [2]*Israel Cohen and* [3]*Sharon Gannot*

[1]e.a.p.habets@tue.nl
[1]Dept. of Electrical Engineering, Technische Universiteit Eindhoven, The Netherlands
[2]Dept. of Electrical Engineering, Technion - Israel Institute of Technology, Israel
[3]School of Engineering, Bar-Ilan University, Israel

ABSTRACT

In this paper we present an algorithm for robust speech enhancement based on an Optimal Modified Minimum Mean-Square Error Log-Spectral Amplitude (OM-LSA) estimator for multiple interferences. In the original OM-LSA one interference was taken into account. However, there are many situations where multiple interferences are present. Since the human ear is more sensitive to a small amount of residual non-stationary interference than to a stationary interference we would like to reduce the non-stationary interference signal down to the residual noise level of the stationary interference. Possible applications for the proposed algorithm are joint speech dereverberation and noise reduction, and joint residual echo suppression and noise reduction. Additionally, we present two possible methods to estimate the *a priori* Signal to Noise Ratio of each of the interferences.

1. INTRODUCTION

Spectral enhancement has received a lot of attention in the last three decades, especially for single channel noise reduction. Recently, researchers have started to use these techniques for residual echo suppression [1,2] and speech dereverberation [3]. In practical systems one may encounter more than one interference simultaneously.

In [2] Gustafsson et al. proposed two postfilters for residual echo and noise reduction. The first postfilter is based on the Log Spectral Amplitude estimator [4] and was extended to attenuate multiple interferences, the second postfilter was psychoacoustically motivated.

In this paper we present an Optimal Modified Minimum Mean-Square Error Log-Spectral Amplitude (OM-LSA) estimator for multiple interferences. The OM-LSA spectral gain function, which minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty. In the original OM-LSA, proposed by Cohen [5], one interference was taken into account. There are many applications in which we are dealing with one non-stationary and one stationary interference. Since the human ear is more sensitive to a small amount of residual non-stationary interference than to a stationary interference we would like to reduce the non-stationary interference signal down to the residual noise level of the stationary interference, such that the final residual non-stationary interference will be masked by the residual stationary interference. Possible applications for the proposed algorithm are joint speech dereverberation and noise reduction, and joint residual echo suppression and noise reduction. The OM-LSA spectral gain function is a function of the *a priori* and *a posteriori* Signal to Noise Ratios (SNRs). In this paper we additionally present two possible methods to estimate the *a priori* SNR of each of the interferences.

The outline of this paper is as follows. The problem statement can be found in Section 2. A brief review of the OM-LSA and a modification of the spectral gain function is presented in Section 3. In Section 4 we will present two methods to estimate the *a priori* SNR for each of the interferences. Experimental results and conclusions are presented, respectively, in Section 5 and 6.

2. PROBLEM STATEMENT

Let $x(n)$, $r(n)$ and $d(n)$ denote speech and two uncorrelated additive interference signals, respectively,

$$y(n) = x(n) + r(n) + d(n).$$

It should be noted that in case $r(n)$ and $d(n)$ are statistically independent Gaussian random variables they can be considered as one interference. The variance of the total interference is then equal to the sum of the separate variances. However, in case $r(n)$ and $d(n)$ are, for example, a non-stationary and a stationary interference, and the (maximum) amount of desirable reduction is different, their separation is preferred. The OM-LSA spectral gain function, which depends on both time and frequency, is a function of the *a priori* and *a posteriori* Signal to Noise Ratios, which are denoted by $\xi(k,l)$ and $\gamma(k,l)$, respectively. In this paper time frames are denoted by the index l, and frequency bins are denoted by the index k. We show that one can gain control of the noise reduction level for each interference by associating a separate *a priori* SNR with each interference.

The estimated Short-Time Fourier Transform (STFT) of the clean speech, $\hat{X}(k,l)$, is obtained by applying the spectral gain function, $G_{\text{OM-LSA}}$, to each noisy spectral component:

$$\hat{X}(k,l) = G_{\text{OM-LSA}}(k,l)Y(k,l).$$

The estimated clean speech signal can be obtained using the inverse STFT and a weighted overlap-add method.
In the sequel we assume that an estimate of the Power Spectral Density (PSD) of each interference is available at all times. In many applications, such as speech dereverberation or residual echo suppression, it is reasonable to assume that the PSD of the non-stationary interference can be estimated (c.f. [1, 2, 3]). The PSD of the stationary interference can be estimated, for example, using the Improved Minima Controlled Recursive Averaging (IMCRA) method proposed by Cohen.

3. OM-LSA ESTIMATOR

The Log Spectral Amplitude (LSA) estimator from Ephraim and Malah [4] minimizes

$$E\left\{\left(\log(A(k,l)) - \log(\hat{A}(k,l))\right)^2\right\},$$

where $A(k,l) = |X(k,l)|$ denotes the spectral speech amplitude, and $\hat{A}(k,l)$ its optimal estimator. Assuming statistical independent spectral components, the LSA estimator is defined as

$$\hat{A}(k,l) = \exp\left(E\{\log(A(k,l))|Y(k,l)|\}\right).$$

The LSA gain function is given by

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

where

$$\begin{aligned}
\nu(k,l) &= \frac{\xi(k,l)}{1+\xi(k,l)}\,\gamma(k,l),\\
\frac{1}{\xi(k,l)} &= \frac{1}{\xi_r(k,l)} + \frac{1}{\xi_d(k,l)},\\
\xi_r(k,l) = \frac{\lambda_x(k,l)}{\lambda_r(k,l)}, &\quad \xi_d(k,l) = \frac{\lambda_x(k,l)}{\lambda_d(k,l)},\\
\gamma(k,l) &= \frac{|Y(k,l)|^2}{\lambda_r(k,l)+\lambda_d(k,l)},\\
\lambda_x(k,l) &= E\{|X(k,l)|^2\},
\end{aligned} \tag{1}$$

$\lambda_d(k,l) = E\{|D(k,l)|^2\}$, and $\lambda_r(k,l) = E\{|R(k,l)|^2\}$.

The OM-LSA spectral gain function, which minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty [5]. Given two hypotheses, $H_0(k,l)$ and $H_1(k,l)$, which indicate, respectively, speech absence and presence, we have

$$\begin{aligned}
H_0(k,l) &: Y(k,l) = R(k,l) + D(k,l),\\
H_1(k,l) &: Y(k,l) = X(k,l) + R(k,l) + D(k,l).
\end{aligned}$$

Based on a Gaussian statistical model, the speech presence probability is given by

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}\,(1+\xi(k,l))\exp\left(-\nu(k,l)\right)\right\}^{-1},$$

where $q(k,l)$ is the *a priori* signal absence probability [5].
The OM-LSA gain function is given by,

$$G_{\text{OM-LSA}}(k,l) = \{G_{\text{H}_1}(k,l)\}^{p(k,l)}\,\{G_{\text{H}_0}(k,l)\}^{1-p(k,l)},$$

with $G_{\text{H}_1}(k,l) = G_{\text{LSA}}(k,l)$ and $G_{\text{H}_0}(k,l) = G_{\min}$. The lower-bound constraint for the gain when the signal is absent is denoted by $G_{\min}$, and specifies the maximum amount of noise reduction in noise only frames.
In our case the lower-bound constraint does not result in the desired result because $r(n)$ can still be clearly audible. To alleviate this problem we propose the following modification of G_{H_0}. Our goal is to suppress the non-stationary interference down to the noise floor, given by $G_{\min}\ D(k,l)$. We apply $G_{\text{H}_0}(k,l)$ to those time-frequency frames where the desired signal is assumed to be absent, i.e. hypothesis $\text{H}_0(k,l)$ is assumed to be true, such that

$$\hat{X}(k,l) = G_{\text{H}_0}(k,l)\,(R(k,l) + D(k,l)).$$

The desired solution for $\hat{X}(k,l)$ is

$$\hat{X}(k,l) = G_{\min}(k,l)\ D(k,l).$$

Assuming that the interferences are uncorrelated, minimizing

$$E\left\{|G_{\text{H}_0}(k,l)\,(R(k,l)+D(k,l)) - G_{\min}(k,l)\ D(k,l)|^2\right\}$$

results in the desired solution for $G_{\text{H}_0}(k,l)$,

$$G_{\text{H}_0}(k,l) = G_{\min}\frac{\hat{\lambda}_d(k,l)}{\hat{\lambda}_d(k,l) + \hat{\lambda}_r(k,l)}, \tag{2}$$

where $\hat{\lambda}_d$ and $\hat{\lambda}_r$ are estimates of, respectively, λ_d and λ_r. The *a posteriori* SNRs can directly be estimated given the noisy observation and an estimate of the Power Spectral Density of each interference. The estimation of the *a priori* SNR is slightly more complicated and will be discussed in the next section.

4. A PRIORI SNR ESTIMATOR FOR MULTIPLE INTERFERENCES

Many researchers believe that the main advantage of the LSA estimator is related to the Decision Directed approach, proposed by Ephraim and Malah [4]. In this section we show how the Decision Directed approach can be used for estimating $\xi_r(k,l)$ and $\xi_d(k,l)$. We also present a non-causal recursive estimation procedure for the *a priori* SNRs using the same reasoning as in [6].
The total *a priori* SNR can be calculated using (1). However, in case $r(n)$ and $x(n)$ are close to zero this equation may not be properly defined. To alleviate this problem we propose to calculate $\xi(k,l)$ as follows

$$\xi(k,l) = \begin{cases} \xi_d & 10\log_{10}\left(\frac{\lambda_d(k,l)}{\lambda_r(k,l)}\right) > \beta^{\text{dB}},\\ \dfrac{\xi_d(k,l)\xi_r(k,l)}{\xi_d(k,l)+\xi_r(k,l)} & \text{otherwise,}\end{cases} \tag{3}$$

where the threshold β^{dB} specifies the level difference between $\lambda_d(k,l)$ and $\lambda_r(k,l)$ in dB.

4.1. Decision Directed

The Decision-Directed based estimator is given by

$$\hat{\xi}^{\text{DD}}(k,l) = \max\left\{\mu\frac{\hat{A}^2(k,l-1)}{\lambda(k,l-1)} + (1-\mu)\psi(k,l),\ \xi_{\text{min}}\right\},$$

where $\psi(k,l) = \gamma(k,l) - 1$ is the *instantaneous* SNR, $\lambda(k,l) = \lambda_r(k,l) + \lambda_d(k,l)$, and ξ_{min} is a lower-bound constraint on the *a priori* SNR. The weighting factor μ $(0 \leq \mu \leq 1)$ controls the tradeoff between the amount of noise reduction and distortion (e.g. musical tones). To estimate $\xi_v(k,l)$, where $v \in \{r,d\}$, we propose to use the following expression

$$\hat{\xi}_v^{\text{DD}}(k,l) = \max\left\{\mu\frac{\hat{A}^2(k,l-1)}{\lambda_v(k,l-1)} + (1-\mu)\psi_v(k,l),\ \xi_{\text{min},v}\right\},$$

where

$$\begin{aligned}\psi_v(k,l) &= \frac{\lambda(k,l)}{\lambda_v(k,l)}\psi(k,l),\\ &= \frac{\lambda_r(k,l)+\lambda_d(k,l)}{\lambda_v(k,l)}(\gamma(k,l)-1),\\ &= \frac{|Y(k,l)|^2 - \lambda_r(k,l) - \lambda_d(k,l)}{\lambda_v(k,l)}.\end{aligned}$$

4.2. Non-Casual Decision Directed

In this section we propose a non-causal conditional estimator

$$\xi_v(k,l|l+L) \triangleq \frac{\lambda_x(k,l|l+L)}{\lambda_v(k,l)},$$

where $v \in \{r,d\}$ and $\lambda_x(k,l|l+L) \triangleq \text{E}\{A^2(k,l)|Y(k,[0,\ldots,l+L])\}$, for the *a priori* SNRs given the noisy measurements up to frame $l+L$. The non-causal estimator combines two steps, a "propagation" step and an "update" step, following the rationale of Kalman filtering, to recursively predict and update the estimate for $\lambda_x(k,l)$ as new data arrives. The non-causal estimator also employs future spectral measurements in the process to better predict the spectral variance of the clean speech.

Let $\lambda'_x(k,l|l+L) \triangleq \text{E}\{A^2(k,l)|Y(k,[0,\ldots,l-1,l+1,\ldots,l+L])\}$ denote the conditional spectral variance of $X(k,l)$ given the noisy measurements up to frame $l+L$ excluding the noisy measurement at frame l. Let $\lambda_x(k,l|[l+1,\ldots,l+L]) \triangleq \text{E}\{A^2(k,l)|Y(k,[l+1,\ldots,l+L])\}$ denote the conditional spectral variance of $X(k,l)$ given the subsequent noisy measurements $Y(k,[l+1,\ldots,l+L])$.

The estimate for $\lambda_x(k,l)$ given $\lambda'_x(k,l|l+L)$ and $Y(k,l)$ can be updated by (4), where

$$\hat{\xi}'(k,l|l+L) \triangleq \frac{\hat{\lambda}'_x(k,l|l+L)}{\lambda(k,l-1)}$$

is the *a priori* SNR given the noisy speech components up to frame $l+L$, excluding frame l [6].

The "backward estimation" and "backward-forward propagation" are exactly the same as in [6] and are presented here for completeness. The "backward estimation" is given by

$$\begin{aligned}&\hat{\xi}(k,l|[l+1,\ldots,l+L])\\ &\quad= \begin{cases}\frac{1}{L}\sum_{n=1}^{L}\gamma(k,l+n) - \beta & \text{if non-negative,}\\ 0 & \text{otherwise,}\end{cases}\end{aligned}$$

where β $(\beta \geq 1)$ is the over-subtraction factor. The "backward-forward propagation" is calculated using (5), where α $(0 \leq \alpha \leq 1)$ is related to the stationarity of the random process λ_x, α' $(0 \leq \alpha' \leq 1)$ is associated with the reliability of the estimate $\xi(k,l|[l+1,l+L])$, and $\hat{\xi}(k,l-1|l+L-1)$ is calculated similar to $\xi(k,l)$ in (3) using $\hat{\xi}_d(k,l-1|l+L-1)$ and $\hat{\xi}_r(k,l-1|l+L-1)$. Dividing both sides in (4) by $\lambda_v(k,l)$, and applying a lower-bound constraint $\xi_{\text{min},v}$, results in the "update" step of $\hat{\xi}_v(k,l|l+L)$ as denoted in (6).

5. RESULTS

We compared the segmental SNR and Log Spectral Distance (LSD) of the original OM-LSA, using $\lambda = \lambda_r + \lambda_d$, and the proposed algorithm using the modified gain function. The segmental SNR is defined as the average local SNR over the set of frames where the desired signal is active. The desired signal consists of a speech signal sampled at 8 kHz. We used random white Gaussian noise as a stationary interference (segmental SNR=12 dB), and a second speech fragment as a non-stationary interference. The *a priori* SNRs were calculated using the Decision Directed (DD) approach and the non-causal (NC) estimator. All parameters were chosen equal to those used in [6]. The lower-bound for $\xi_{\text{min},r}^{dB} = -40$ dB and $\xi_{\text{min}}^{dB} = \xi_{\text{min},d}^{dB} = -18$ dB, and β^{dB} was set to 3 dB. In this experiment we used the exact power spectra of both interferences, due to this we can exclude the influence of the estimation of λ_r

$$\hat{\lambda}_x(k,l|l+L) = \text{E}\left\{A^2(k,l)|\lambda'_x(k,l|l+L),Y(k,l)\right\} = \frac{\hat{\xi}'(k,l|l+L)}{1+\hat{\xi}'(k,l|l+L)}\left(\frac{1}{\gamma(k,l)} + \frac{\hat{\xi}'(k,l|l+L)}{1+\hat{\xi}'(k,l|l+L)}\right)|Y(k,l)|^2 \tag{4}$$

$$\hat{\xi}'(k,l|l+L) = \alpha\frac{\hat{A}^2(k,l-1)}{\lambda(k,l-1)} + (1-\alpha)\left[\alpha'\hat{\xi}(k,l-1|l+L-1) + (1-\alpha')\hat{\xi}(k,l|[l+1,\ldots,l+L])\right] \tag{5}$$

$$\hat{\xi}_v(k,l|l+L) = \max\left\{\frac{\hat{\xi}'(k,l|l+L)}{1+\hat{\xi}'(k,l|l+L)}\left(\frac{1}{\gamma(k,l)} + \frac{\hat{\xi}'(k,l|l+L)}{1+\hat{\xi}'(k,l|l+L)}\right)\frac{|Y(k,l)|^2}{\lambda_v(k,l)},\ \xi_{\text{min},v}\right\} \tag{6}$$

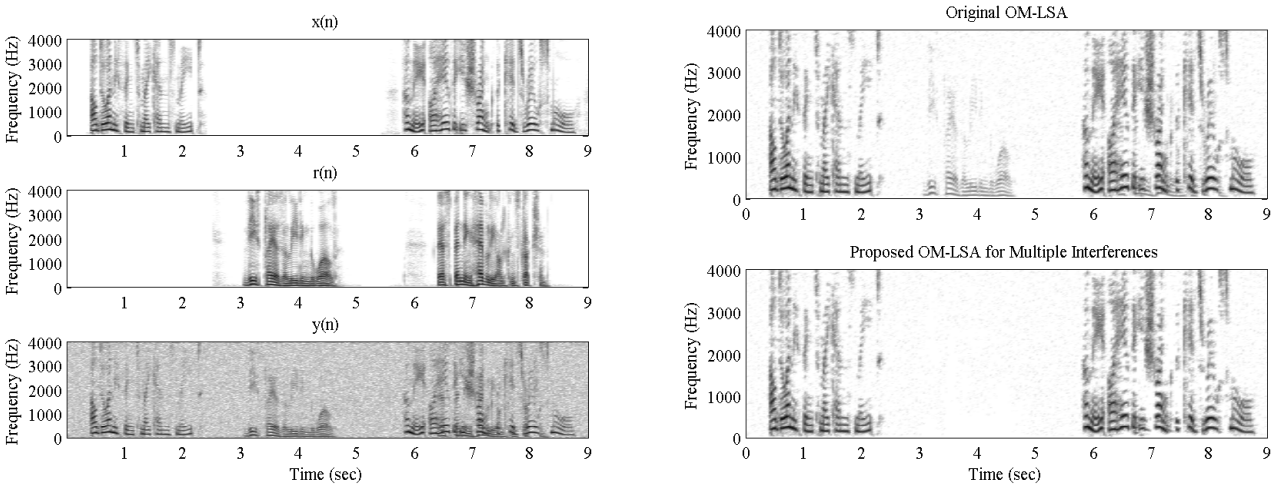

Figure 1: Spectrograms of the desired signal $x(n)$, the non-stationary interference $r(n)$, the microphone signal $y(n)$, and the original and modified OM-LSA using the non-causal *a priori* SNR estimator.

and λ_d. The segmental SNR and LSD results are shown in Table 1. In all cases we can see that the non-causal estimator leads to a larger segmental SNR and smaller LSD. The LSD of the proposed algorithms has decreased significantly compared to the original OM-LSA. The spectrograms of desired signal $x(n)$, the non-stationary interference $r(n)$, the microphone signal $y(n)$, and the original and modified OM-LSA using the non-causal *a priori* SNR estimator, are depicted in Figure 1. From these results we can clearly see that the non-stationary interference was decreased significantly compared to the original OM-LSA.

Method	Segmental SNR (dB)	LSD (dB)
Unprocessed	8.594	2.637
OM-LSA, DD	14.184	0.994
OM-LSA, NC	14.306	0.975
Proposed, DD	14.212	0.733
Proposed, NC	14.429	0.709

Table 1: Segmental SNR and Log Spectral Distance (LSD) results for the OM-LSA and proposed methods.

6. CONCLUSIONS

We have developed an Optimally-Modified Log Spectral Amplitude estimator for multiple interferences. The estimator involves separate a priori SNR estimation of each of the interferences, by using the decision-directed approach, or a recursive non-causal estimator. In some applications (e.g., dereverberation and residual echo suppression) the spectra of both the stationary and non-stationary interferences can be reliably estimated. In such cases, the proposed approach outperforms existing methods, and enables better control of the level of residual non-stationary noise. Experimental results demonstrated that the non-stationary interference can be effectively suppressed down to the residual noise level of the stationary interference.

7. ACKNOWLEDGEMENT

This research is/was partially supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs. The authors express their gratitude to STW for funding.

8. REFERENCES

[1] E. Hänsler and G. Schmidt, "Hands-free telephones - joint control of echo cancellation and postfiltering," *Signal Processing*, vol. 80, pp. 2295–2305, 2000.

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

[3] E.A.P. Habets, "Multi-Channel Speech Dereverberation based on a Statistical Model of Late Reverberation," in *Proc. of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Philadelphia, USA*, March 2005, pp. 173–176.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Speech Audio Processing*, vol. 32, pp. 1109–1121, December 1984.

[5] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," *IEEE Signal Processing Lett.*, vol. 9, no. 4, pp. 113–116, April 2002.

[6] I. Cohen, "Relaxed Statistical Model for Speech Enhancement and A Priori SNR Estimation," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 870–881, September 2005.