Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator

Israel Cohen, *Member, IEEE*

***Abstract*—In this paper, we present an *optimally modified* log-spectral amplitude estimator, which minimizes the mean-square error of the log-spectra for speech signals under signal presence uncertainty. We propose an estimator for the *a priori* signal-to-noise ratio (SNR), and introduce an efficient estimator for the *a priori* speech absence probability. Speech presence probability is estimated for each frequency bin and each frame by a soft-decision approach, which exploits the strong correlation of speech presence in neighboring frequency bins of consecutive frames. Objective and subjective evaluation confirm superiority in noise suppression and quality of the enhanced speech.**

***Index Terms*—Estimation, spectral analysis, speech enhancement.**

I. Introduction

RECENTLY, the use of a soft-decision gain modification in speech enhancement algorithms has been the object of considerable research. While traditional spectral enhancement techniques estimate the clean speech spectrum under speech presence hypothesis, a modified estimator, which incorporates the *a priori* speech absence probability (SAP), generally yields better performance [1]–[5].

The log-spectral amplitude (LSA) estimator, developed by Ephraim and Malah [6], proved very efficient in reducing musical residual noise phenomena. Its modification under signal presence uncertainty is obtained by multiplying the spectral gain by the conditional speech presence probability, which is estimated for each frequency bin and each frame [4]. Unfortunately, the multiplicative modifier is not optimal [4]. Moreover, the interaction between the estimate for the *a priori* signal-to-noise ratio (SNR) and the estimate for the *a priori* SAP may deteriorate the performance of the speech enhancement system [3], [7], [8].

An alternative approach [5] is to use a small fixed *a priori* SAP, $q = 0.0625$, and a multiplicative modifier, which is based on the *global* conditional SAP in each frame. This modifier is applied to the *a priori* and *a posteriori* SNRs. However, such a modification is inconsistent with the statistical model, and may not be sufficient due to the small value of q and the influence of a few noise-only bins on the global SAP.

In this paper, we present an *optimally modified* LSA (OM-LSA) estimator. We introduce a new estimator for the *a priori* SNR, and propose an efficient estimator for the *a priori* SAP. The spectral gain function is obtained as a weighted geometric mean of the hypothetical gains associated with signal presence and absence. The *a priori* SAP is estimated for each frequency bin and each frame by a soft-decision approach, which exploits the strong correlation of speech presence in neighboring frequency bins of consecutive frames. Objective and subjective evaluation in various environmental conditions show that the proposed modification approach is advantageous, particularly for low input SNRs and nonstationary noise. Excellent noise reduction can be achieved even in the most adverse noise conditions, while avoiding musical residual noise and the attenuation of weak speech components.

Manuscript received July 30, 2001; revised September 18, 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yair Shoham.

The author was with Lamar Signal Processing Ltd., Yokneam Ilit 20692, Israel. He is now with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: icohen@ee.technion.ac.il).

Publisher Item Identifier S 1070-9908(02)05034-4.

II. Optimal Gain Modification

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively. The observed signal $y(n)$ is divided into overlapping frames by the application of a window function and analyzed using the short-time Fourier transform (STFT). In the time-frequency domain we have $Y(k,\ell) = X(k,\ell) + D(k,\ell)$, where k represents the frequency bin index, and ℓ the frame index. Given two hypotheses, $H_0(k,\ell)$ and $H_1(k,\ell)$, which indicate respectively speech absence and presence, and assuming a complex Gaussian distribution of the STFT coefficients for both speech and noise [2], the conditional PDFs of the observed signal are given by

$$p(Y(k,\ell)|H_0(k,\ell)) = \frac{1}{\pi\lambda_d(k,\ell)}\exp\left\{-\frac{|Y(k,\ell)|^2}{\lambda_d(k,\ell)}\right\}$$

$$p(Y(k,\ell)|H_1(k,\ell)) = \frac{1}{\pi(\lambda_x(k,\ell)+\lambda_d(k,\ell))} \cdot \exp\left\{-\frac{|Y(k,\ell)|^2}{\lambda_x(k,\ell)+\lambda_d(k,\ell)}\right\} \quad (1)$$

where $\lambda_x(k,\ell) \triangleq E[|X(k,\ell)|^2|H_1(k,\ell)]$ and $\lambda_d(k,\ell) \triangleq E[|D(k,\ell)|^2]$ denote respectively the variances of speech and noise. Applying Bayes rule, the conditional speech presence probability $p(k,\ell) \triangleq P(H_1(k,\ell)|Y(k,\ell))$ can be written as [2]

$$p(k,\ell) = \left\{1 + \frac{q(k,\ell)}{1-q(k,\ell)}(1+\xi(k,\ell))\exp(-v(k,\ell))\right\}^{-1} \quad (2)$$

where $q(k,\ell) \triangleq P(H_0(k,\ell))$ is the *a priori* probability for speech absence, $\xi(k,\ell) \triangleq \lambda_x(k,\ell)/\lambda_d(k,\ell)$ is the *a priori* SNR, $\gamma(k,\ell) \triangleq |Y(k,\ell)|^2/\lambda_d(k,\ell)$ is the *a posteriori* SNR, and $v(k,\ell) \triangleq \gamma(k,\ell)\xi(k,\ell)/(1+\xi(k,\ell))$.

Let $A = |X|$ denote the spectral speech amplitude, and $\hat{A}$ its optimal estimate. Assuming statistically independent spectral components [6], the LSA estimator is defined by

$$\hat{A}(k, \ell) = \exp\{E[\log A(k, \ell)|Y(k, \ell)]\} \triangleq G(k, \ell)|Y(k, \ell)|. \tag{3}$$

Based on the statistical model

$$\begin{aligned} &E[\log A(k, \ell)|Y(k, \ell)] \\ &\quad = E[\log A(k, \ell)|Y(k, \ell), H_1(k, \ell)]p(k, \ell) \\ &\qquad + E[\log A(k, \ell)|Y(k, \ell), H_0(k, \ell)](1 - p(k, \ell)). \end{aligned} \tag{4}$$

When speech is absent, the gain is constrained to be larger than a threshold $G_{\min}$, which is determined by a subjective criteria for the noise naturalness. Accordingly,

$$\exp\{E[\log A(k, \ell)|Y(k, \ell), H_0(k, \ell)]\} = G_{\min} \cdot |Y(k, \ell)|. \tag{5}$$

When speech is present, the conditional gain function, defined by

$$\begin{aligned} &\exp\{E[\log A(k, \ell)|Y(k, \ell), H_1(k, \ell)]\} \\ &\qquad = G_{H_1}(k, \ell) \cdot |Y(k, \ell)| \end{aligned} \tag{6}$$

is derived in [6] to be

$$G_{H_1}(k, \ell) = \frac{\xi(k, \ell)}{1 + \xi(k, \ell)} \exp\left(\frac{1}{2}\int_{\upsilon(k, \ell)}^{\infty} \frac{e^{-t}}{t}\, dt\right). \tag{7}$$

Substituting (5) and (6) into (3), the gain function for the OM-LSA estimator is obtained by

$$G(k, \ell) = \{G_{H_1}(k, \ell)\}^{p(k, \ell)} \cdot G_{\min}^{1 - p(k, \ell)}. \tag{8}$$

It is worthwhile mentioning that trying to optimally modify the spectral gain function for the LSA estimator without taking into account a lower bound threshold ($G_{\min}$) results in a non-multiplicative modification, which fails to provide a meaningful improvement over using G_{H_1} alone [4], [6].

III. *A Priori* SNR Estimation

In this section we address the problem of the *a priori* SNR estimation under speech presence uncertainty.

The decision-directed approach, proposed by Ephraim and Malah [2], provides a useful estimation method for the *a priori* SNR. Accordingly, if speech presence is assumed ($q(k, \ell) \equiv 0$), then the expression

$$\begin{aligned} \Xi(k, \ell) = \alpha G^2(k, \ell - 1)\gamma(k, \ell - 1) \\ + (1 - \alpha)\max\{\gamma(k, \ell) - 1, 0\} \end{aligned} \tag{9}$$

can be substituted for the *a priori* SNR, where α is a weighting factor that controls the tradeoff between noise reduction and speech distortion [2], [9]. Under speech presence uncertainty, this expression estimates a *nonconditional a priori* SNR $\eta(k, \ell) \triangleq E[|X(k, \ell)|^2]/\lambda_d(k, \ell)$, and therefore the estimate for the *a priori* SNR $\xi(k, \ell)$ should be given by $\Xi(k, \ell)/(1 - q(k, \ell))$ [2], [4]. However, the division by $1 - q(k, \ell)$ may introduce interaction between the estimated $q(k, \ell)$ and the *a priori* SNR, generally deteriorating the performance of the speech enhancement system [3], [7], [8].

TABLE I
VALUES OF PARAMETERS USED FOR THE ESTIMATION OF THE *A PRIORI* SPEECH ABSENCE PROBABILITY

$\beta = 0.7$	$\zeta_{min} = -10\text{dB}$	$\zeta_{p\min} = 0\text{dB}$
$w_{local} = 1$	$\zeta_{max} = -5\text{dB}$	$\zeta_{p\max} = 10\text{dB}$
$w_{global} = 15$	$q_{max} = 0.95$	h_λ: Hanning windows

In [3], [10], it was suggested to simply estimate the *a priori* SNR by $\Xi(k, \ell)$, rather than $\Xi(k, \ell)/(1 - q(k, \ell))$, even though the latter better approximates an unbiased estimate for $\xi(k, \ell)$. Here, we propose the following estimator:

$$\begin{aligned} \hat{\xi}(k, \ell) = \alpha G_{H_1}^2(k, \ell - 1)\gamma(k, \ell - 1) \\ + (1 - \alpha)\max\{\gamma(k, \ell) - 1, 0\}. \end{aligned} \tag{10}$$

The use of $G_{H_1}(k, \ell - 1)$, instead of $G(k, \ell - 1)$, boosts the gain up when speech is present, which provides a compensation for not dividing by $1 - q(k, \ell)$. By definition, if $H_1(k, \ell)$ is true, then the spectral gain $G(k, \ell)$ should degenerate to $G_{H_1}(k, \ell)$, and the *a priori* SNR estimate should coincide with $\Xi(k, \ell)$. On the contrary, if $H_0(k, \ell)$ is true, then $G(k, \ell)$ should decrease to $G_{\min}$, or equivalently the *a priori* SNR estimate should be as small as possible. This is satisfied more favorably by the proposed $\hat{\xi}$ rather than by $\Xi/(1 - \hat{q})$ [8].

IV. *A Priori* SAP Estimation

In this section we derive an efficient estimator $\hat{q}(k, \ell)$ for the *a priori* SAP. This estimator uses a soft-decision approach to compute three parameters based on the time-frequency distribution of the estimated *a priori* SNR, $\hat{\xi}(k, \ell)$. The parameters exploit the strong correlation of speech presence in neighboring frequency bins of consecutive frames.

Let $\zeta(k, \ell)$ be a recursive average of the *a priori* SNR with a time constant β

$$\zeta(k, \ell) = \beta\zeta(k, \ell - 1) + (1 - \beta)\hat{\xi}(k, \ell - 1). \tag{11}$$

By applying *local* and *global* averaging windows in the frequency domain, we obtain, respectively, local and global averages of the *a priori* SNR:

$$\zeta_\lambda(k, \ell) = \sum_{i=-w_\lambda}^{w_\lambda} h_\lambda(i)\zeta(k - i, \ell) \tag{12}$$

where the subscript λ designates either "local" or "global," and h_λ is a normalized window of size $2w_\lambda + 1$. We define two parameters, $P_{local}(k, \ell)$ and $P_{global}(k, \ell)$, which represent the relation between the above averages and the likelihood of speech in the kth frequency bin of the ℓth frame. These parameters are given by

$$P_\lambda(k, \ell) = \begin{cases} 0, & \text{if } \zeta_\lambda(k, \ell) \le \zeta_{\min} \\ 1, & \text{if } \zeta_\lambda(k, \ell) \ge \zeta_{\max} \\ \dfrac{\log(\zeta_\lambda(k, \ell)/\zeta_{min})}{\log(\zeta_{max}/\zeta_{min})}, & \text{otherwise} \end{cases} \tag{13}$$

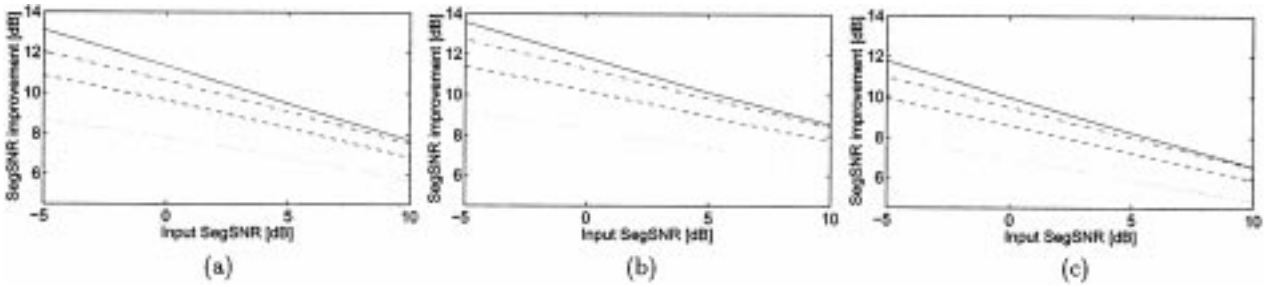

Fig. 1. Comparison of speech estimators, OM-LSA (solid), MM-LSA (dashdot), LSA (dashed), and STSA (dotted), for various noise types and levels. Average segmental SNR improvement for: (a) white Gaussian noise, (b) car interior noise, and (c) F16 cockpit noise.

where $\zeta_{\min}$ and $\zeta_{\max}$ are empirical constants, maximized to attenuate noise while maintaining weak speech components.

In order to further attenuate noise in noise-only frames, we define a third parameter, $P_{frame}(\ell)$, which is based on the speech energy in neighboring frames. An averaging of $\zeta(k, \ell)$ in the frequency domain (possibly over a certain frequency band) yields

$$\zeta_{frame}(\ell) = \underset{1\le k\le N/2+1}{\text{mean}} \{\zeta(k, \ell)\}. \tag{14}$$

To prevent clipping of speech onsets or weak components, speech is assumed whenever $\zeta_{frame}(\cdot)$ increases. Clipping of weak speech tails is prevented by delaying the transition from H_1 to H_0, and allowing for a certain decrease in the value of ζ_{frame}. A pseudocode for the computation of P_{frame} is given by,

If $\zeta_{frame}(\ell) > \zeta_{\min}$ **then**
  If $\zeta_{frame}(\ell) > \zeta_{frame}(\ell-1)$ **then**
    $P_{frame}(\ell) = 1$
    $\zeta_{peak}(\ell) = \min\{\max[\zeta_{frame}(\ell), \zeta_{p\min}], \zeta_{p\max}\}$
  Else
    $P_{frame}(\ell) = \mu(\ell)$
Else
  $P_{frame}(\ell) = 0$

where

$$\mu(\ell) \triangleq \begin{cases} 0, & \text{if } \zeta_{frame}(\ell) \le \zeta_{peak}(\ell)\cdot\zeta_{\min} \\ 1, & \text{if } \zeta_{frame}(\ell) \ge \zeta_{peak}(\ell)\cdot\zeta_{\max} \\ \dfrac{\log(\zeta_{frame}(\ell)/\zeta_{peak}(\ell)/\zeta_{\min})}{\log(\zeta_{\max}/\zeta_{\min})}, & \text{otherwise} \end{cases} \tag{15}$$

represents a soft transition from "speech" to "noise," ζ_{peak} is a confined peak value of ζ_{frame}, and $\zeta_{p\min}$ and $\zeta_{p\max}$ are empirical constants that determine the delay of the transition.

The proposed estimate for the *a priori* SAP is obtained by

$$\hat{q}(k, \ell) = 1 - P_{local}(k, \ell)\cdot P_{global}(k, \ell)\cdot P_{frame}(\ell). \tag{16}$$

Accordingly, $\hat{q}(k, \ell)$ is larger if either previous frames, or recent neighboring frequency bins, do not contain speech. When $\hat{q}(k, \ell) \to 1$, the conditional speech presence probability $p(k, \ell) \to 0$ by (2), and consequently the gain function reduces to $G_{\min}$ by (8). Therefore, to reduce the possibility of speech distortion we restrict $\hat{q}(k, \ell)$ to be smaller than a threshold $q_{\max}$ ($q_{\max} < 1$).

V. Performance Evaluation and Discussion

The OM-LSA estimator is combined with the proposed *a priori* SNR and SAP estimators, and compared to the LSA [6], *short-time spectral amplitude* (STSA) [2], and *multiplicatively modified* LSA (MM-LSA) [4] estimators. The evaluation consists of an objective segmental SNR measure, a subjective study of speech spectrograms and informal listening tests. Three different noise types, taken from Noisex92 database, are used: white Gaussian noise, car noise, and F16 cockpit noise. The performance results are averaged out using six different utterances, taken from the TIMIT database. Half of the utterances are from male speakers and half are from female speakers.

The speech signals, sampled at 16 kHz, are degraded by the various noise types with segmental SNRs in the range $[-5, 10]$ dB. The STFT is implemented with Hamming windows of 512 samples length (32 ms) and 128 samples frame update step. The *a priori* SNR is estimated using the decision-directed approach with $\alpha = 0.92$, where the proposed algorithm employs the new estimator (10). The spectral gain is restricted to a minimum of -20 dB, and the noise statistics is assumed to be known (recursively smoothed periodogram with a forgetting factor set to 0.9). Values of parameters used for the estimation of the *a priori* SAP are summarized in Table I.

Fig. 1 shows the average segmental SNR improvement obtained for various noise types and at various noise levels. The OM-LSA estimator, combined with the proposed *a priori* SNR and SAP estimators, achieves the best results under all noise conditions. The advantage is more significant for low SNRs. The segmental SNR measure takes into account both residual noise and speech distortion. It lacks indication about the structure of the residual noise. A subjective comparison, conducted using speech spectrograms and validated by informal listening tests, confirms the improvement obtained by the proposed method [8]. In contrast to other methods, where abrupt bursts of noise generally produce high *a posteriori* SNRs and high spectral gains, resulting in musical noise phenomena, the proposed method attenuates noise by identifying noise-only regions ($\hat{q} \to q_{\max}$) and reducing the gain correspondingly to $G_{\min}$. Yet, it avoids the attenuation of weak speech components by letting $\hat{q}$ descend to zero in speech regions.

ACKNOWLEDGMENT

The author thanks Prof. D. Malah and B. Berdugo for valuable discussions.

REFERENCES

[1] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 137–145, Apr. 1980.

[2] 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, pp. 1109–1121, Dec. 1984.

[3] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence," *Signal Process.*, vol. 75, pp. 151–159, 1999.

[4] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing 1999*, pp. 789–792.

[5] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Lett.*, vol. 7, pp. 108–110, May 2000.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, Apr. 1985.

[7] R. Martin, I. Wittke, and P. Jax, "Optimized estimation of spectral parameters for the coding of noisy speech," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing 2000*, pp. 1479–1482.

[8] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Oct. 2001.

[9] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 345–349, Apr. 1994.

[10] I. Cohen, "On speech enhancement under signal presence uncertainty," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing 2001*, Salt Lake City, UT, May 2001, pp. 167–170.