ON SPEECH ENHANCEMENT UNDER SIGNAL PRESENCE UNCERTAINTY

Israel Cohen

Lamar Signal Processing Ltd., P.O.Box 273, Yokneam Ilit 20692, Israel
icohen@lamar.co.il; http://www.AndreaElectronics.com

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. The spectral gain function is obtained as a weighted geometric mean of the hypothetical gains associated with signal presence and absence. The exponential weight of each hypothetical gain is its corresponding probability, conditioned on the observed signal. We introduce an efficient estimation approach for the *a priori* signal absence probability in each frequency bin, 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.

1. 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 signal presence hypothesis, a modified estimator, which incorporates the a priori probability of speech absence, generally yields better performance [1]–[8].

The Log-Spectral Amplitude (LSA) estimator , was developed by Ephraim and Malah [1] to minimize the mean-square error of the log-spectra. Although showed superior performance in eliminating noise, its modification under signal presence uncertainty was believed "unworthy". That assertion was re-emphasized by Malah *et al.* [2], who proposed a *multiplicatively-modified*, rather than *optimally-modified*, LSA estimator.

A central issue in [2] is a method for estimating the a priori probability of speech absence, q, for each frequency bin in each frame. The authors were puzzled, however, by the fact that their method was not advantageous over using a fixed $q = 0.5$. Furthermore, the interaction between q and the a priori signal-to-noise ratio (SNR) adversely affected the total gain for noise-only bins, and resulted in a "musical residual noise" with an unnatural structure [3].

An alternative approach [4] is to use a small fixed q ($q = 0.0625$) and a multiplicative modifier, which is based on the *global* conditioned speech absence probability in each frame. This modifier is applied to the a priori and a posteriori SNRs. Not only such a modification is inconsistent with the statistical model, but also insignificant due to the small value of q and the influence of a few noise-only bins on the global speech absence probability.

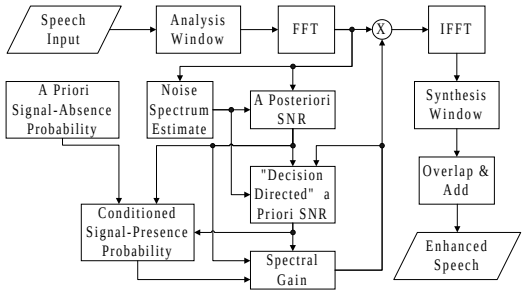

Fig. 1. A block diagram of speech enhancement configuration under signal presence uncertainty.

In this paper, we present an *optimally-modified* LSA estimator. The optimal spectral gain function is obtained as a weighted geometric mean of the hypothetical gains associated with the signal presence uncertainty. The exponential weight of each hypothetical gain is its corresponding probability, conditioned on the observed signal. To derive the conditioned signal presence and absence probabilities, we introduce an efficient estimator for the *a priori* signal absence probability, based on the time-frequency distribution of the a priori SNR. The estimation is implemented for each frequency bin in each frame through a soft-decision approach, which exploits the strong correlation of speech presence in neighboring frequency bins of consecutive frames.

Objective and subjective evaluation of the optimally-modified LSA estimator is performed under various environmental conditions. It is shown that the proposed modification approach is superior, particularly for low input SNRs and non-stationary noise. A real time test demonstrates that a 25 dB noise reduction can be achieved even in the most adverse noise conditions, while avoiding musical residual noise and the attenuation of weak speech components.

2. OPTIMAL GAIN MODIFICATION

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively, where n is a discrete-time index. The observed signal $y(n)$, given by $y(n) = x(n) + d(n)$, is divided into overlapping frames by the application of a window function and analyzed using the short-time Fourier transform (STFT). Specifically,

$$Y(k,\ell) = \sum_{n=0}^{M-1} h_a(n) \cdot y[n + \ell(M - M_o)] \cdot \exp(-\tfrac{j2\pi nk}{M}) \quad (1)$$

where k is the frequency bin index, ℓ is the time frame index, h_a is an analysis window of size M (*e.g.,* Hanning window), and M_o is the number of overlapping samples in consecutive frames. Given two hypotheses, $H_0(k,\ell)$ and $H_1(k,\ell)$, which indicate respectively speech absence and presence in the kth frequency bin of the ℓth frame, we have

$$\begin{aligned} H_0(k,\ell) &: Y(k,\ell) = D(k,\ell) \\ H_1(k,\ell) &: Y(k,\ell) = X(k,\ell) + D(k,\ell) \end{aligned} \tag{2}$$

where $X(k,\ell)$ and $D(k,\ell)$ represent the STFT of the clean and noise signals, respectively. Assuming a complex Gaussian distribution of the STFT coefficients for both speech and noise [5], the conditional PDFs of the observed signal are given by

$$\begin{aligned} p(Y(k,\ell)|H_0(k,\ell)) &= \frac{1}{\pi\lambda_d(k,\ell)} \exp\left\{ -\frac{|Y(k,\ell)|^2}{\lambda_d(k,\ell)} \right\} \\ p(Y(k,\ell)|H_1(k,\ell)) &= \frac{1}{\pi(\lambda_x(k,\ell)+\lambda_d(k,\ell))} \\ &\quad \cdot \exp\left\{ -\frac{|Y(k,\ell)|^2}{\lambda_x(k,\ell)+\lambda_d(k,\ell)} \right\} \end{aligned} \tag{3}$$

where $\lambda_x(k,\ell) = E[|X(k,\ell)|^2]$ and $\lambda_d(k,\ell) = E[|D(k,\ell)|^2]$ denote respectively the variances of speech and noise. Applying Bayes rule for the conditional signal presence probability, one obtains

$$P(H_1(k,\ell)|Y(k,\ell)) = \frac{\Lambda(k,\ell)}{1+\Lambda(k,\ell)} \triangleq p(k,\ell) \tag{4}$$

where $\Lambda(k,\ell)$ is the generalized likelihood ratio defined by

$$\Lambda(k,\ell) = \frac{1-q(k,\ell)}{q(k,\ell)} \cdot \frac{p(Y(k,\ell)|H_1(k,\ell))}{p(Y(k,\ell)|H_0(k,\ell))} \tag{5}$$

and $q(k,\ell) \triangleq P(H_0(k,\ell))$ is *a priori* probability for speech absence. Let $\eta(k,\ell)$ and $\gamma(k,\ell)$ denote the *a priori* and *a posteriori* signal-to-noise ratios [6, 5],

$$\eta(k,\ell) \triangleq \frac{\lambda_x(k,\ell)}{\lambda_d(k,\ell)}; \quad \gamma(k,\ell) \triangleq \frac{|Y(k,\ell)|^2}{\lambda_d(k,\ell)}, \tag{6}$$

substituting (3) and (5) into (4), we have

$$p(k,\ell) = \left\{ 1 + \frac{q(k,\ell)}{1-q(k,\ell)} (1+\eta(k,\ell)) \exp(-\upsilon(k,\ell)) \right\}^{-1} \tag{7}$$

where

$$\upsilon(k,\ell) \triangleq \frac{\gamma(k,\ell)\eta(k,\ell)}{1+\eta(k,\ell)}. \tag{8}$$

An estimate of the clean speech spectrum is obtained by applying a specific gain function to each spectral component of the noisy speech signal: $\hat{X}(k,\ell) = G(k,\ell)Y(k,\ell)$. Among various existing speech enhancement methods, which can be represented by different spectral gain functions, we choose the LSA estimator [1] due to its superiority in reducing musical noise phenomena. The LSA estimator minimizes

$$E\{(\log A(k,\ell) - \log \hat{A}(k,\ell))^2\}$$

where $A(k,\ell) = |X(k,\ell)|$ denotes the spectral speech amplitude, and $\hat{A}(k,\ell)$ its optimal estimate. Accordingly [1],

$$\hat{A}(k,\ell) = \exp\left\{ E\left[\log A(k,\ell)|Y(k,\ell)\right] \right\}. \tag{9}$$

Based on the statistical model,

$$\begin{aligned} &E\left[\log A(k,\ell)|Y(k,\ell)\right] \\ &\quad = E\left[\log A(k,\ell)|Y(k,\ell), H_1(k,\ell)\right] p(k,\ell) \\ &\quad + E\left[\log A(k,\ell)|Y(k,\ell), H_0(k,\ell)\right] (1-p(k,\ell)). \end{aligned} \tag{10}$$

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 [9, 8]. Accordingly,

$$\exp\left\{ E\left[\log A(k,\ell)|Y(k,\ell), H_0(k,\ell)\right] \right\} = G_{min} \cdot |Y(k,\ell)|. \tag{11}$$

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

$$\exp\left\{ E\left[\log A(k,\ell)|Y(k,\ell), H_1(k,\ell)\right] \right\} = G_{H_1}(k,\ell) \cdot |Y(k,\ell)|, \tag{12}$$

is derived in [1] to be

$$G_{H_1}(k,\ell) = \frac{\eta(k,\ell)}{1+\eta(k,\ell)} exp\left(\frac{1}{2} \int_{\upsilon(k,\ell)}^{\infty} \frac{e^{-t}}{t} dt \right). \tag{13}$$

Hence the gain function for the optimally-modified LSA estimator is obtained by

$$G(k,\ell) = \{G_{H_1}(k,\ell)\}^{p(k,\ell)} \cdot G_{min}^{1-p(k,\ell)}. \tag{14}$$

A block diagram of the speech enhancement configuration under signal presence uncertainty is described in Fig. 1.

Notice that in [2] the spectral gain function is not bounded by a lower threshold, which leads to a meaningless gain modification and a false conclusion that "it is unworthy to incorporate signal presence uncertainty in the LSA estimator" [1, 2]. Moreover, according to [2] the *a priori* SNR should be conditioned on the presence of speech, *i.e.* η should be replaced in (7), (8) and (13) with $\eta/(1-q)$, where η is obtained using a decision-directed estimation approach [5, 1]. However, the derivation of the a priori SNR estimator with a decision-directed approach *already assumes presence of speech*. Clearly, the a priori SNR, estimated by

$$\hat{\eta}(k,\ell) = \alpha \frac{|\hat{X}(k,\ell-1)|^2}{\lambda_d(k,\ell-1)} + (1-\alpha) \max\left\{ \gamma(k,\ell) - 1, 0 \right\} \tag{15}$$

is conditioned on the $H_1(k,\ell)$ hypothesis. This misconception might explain some of the puzzling results mentioned in [2, 3, 4].

3. A PRIORI PROBABILITY FOR SIGNAL ABSENCE ESTIMATE

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

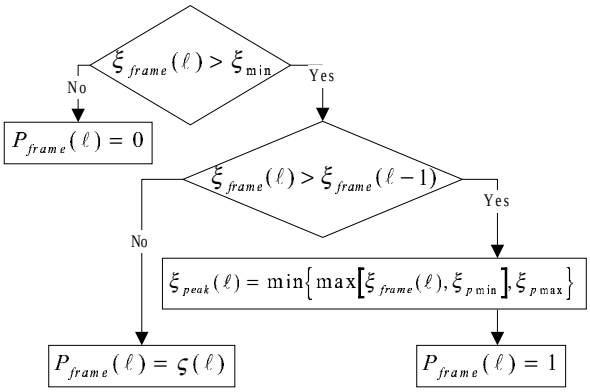

Fig. 2. A block diagram for computing P_{frame} (a parameter representing the likelihood of speech in a given frame).

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

$$\xi(k,\ell) = \beta\xi(k,\ell-1) + (1-\beta)\eta(k,\ell-1)\,. \qquad (16)$$

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

$$\xi_\lambda(k,\ell) = \sum_{i=-w_\lambda}^{w_\lambda} h_\lambda(i)\xi(k-i,\ell) \qquad (17)$$

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

$$P_\lambda(k,\ell) = \begin{cases} 0, & \text{if } \xi_\lambda(k,\ell) \le \xi_{min} \\ 1, & \text{if } \xi_\lambda(k,\ell) \ge \xi_{max} \\ \frac{\log(\xi_\lambda(k,\ell)/\xi_{min})}{\log(\xi_{max}/\xi_{min})}, & \text{otherwise.} \end{cases} \qquad (18)$$

where ξ_{min} and ξ_{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 $\xi(k,\ell)$ in the frequency domain (possibly over a certain frequency band) yields

$$\xi_{frame}(\ell) = \underset{1\le k\le M/2+1}{\text{mean}} \{\xi(k,\ell)\}\,. \qquad (19)$$

To prevent clipping of speech startings or weak components, speech is assumed whenever $\xi_{frame}(\cdot)$ increases. Moreover, the transition from H_1 to H_0 is delayed, which reduces the misdetection of weak speech tails, by allowing for a certain decrease in the value of ξ_{frame}. Fig. 2 describes a block diagram for computing $P_{frame}(\ell)$, where

$$\zeta(\ell) \triangleq \begin{cases} 0, & \text{if } \xi_{frame}(\ell) \le \xi_{peak}(\ell)\cdot\xi_{min} \\ 1, & \text{if } \xi_{frame}(\ell) \ge \xi_{peak}(\ell)\cdot\xi_{max} \\ \frac{\log(\xi_{frame}(\ell)/\xi_{peak}(\ell)/\xi_{min})}{\log(\xi_{max}/\xi_{min})}, & \text{otherwise}\,, \end{cases} \qquad (20)$$

$M = 512$	$w_{local} = 1$	$\xi_{p\min} = 0$dB
$M_o = 384$	$w_{global} = 15$	$\xi_{p\max} = 10$dB
$\alpha = 0.92$	$\xi_{min} = -10$dB	$G_{min} = -25$dB
$\beta = 0.7$	$\xi_{max} = -5$dB	$q_{max} = 0.95$

Table 1. Values of parameters used in the implementation of the optimally-modified LSA estimator.

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

The proposed estimate for the *a priori* probability for speech absence is obtained by

$$\hat{q}(k,\ell) = 1 - P_{local}(k,\ell)\cdot P_{global}(k,\ell)\cdot P_{frame}(\ell)\,. \qquad (21)$$

Accordingly, $\hat{q}(k,\ell)$ is larger if either previous frames, or recent neighboring frequency bins, do not contain speech.

When $q(k,\ell) \to 1$, the conditioned signal presence probability $p(k,\ell) \to 0$ by (7), and consequently the gain function reduces to G_{min} by (14). Therefore, to reduce the possibility of speech distortion we restrict $\hat{q}(k,\ell)$ to be smaller than a threshold q_{max} ($q_{max} < 1$).

4. PERFORMANCE EVALUATION AND DISCUSSION

The optimally-modified LSA estimator was compared to the multiplicatively-modified LSA estimator [2] and to the original STSA and LSA estimators [5, 1]. A clean speech sentence "Draw every outer line first, then fill in the interior," spoken by a female and sampled at 16 kHz, was degraded by non-stationary car noise and white Gaussian noise with global SNR in the range $[-5, 10]$ dB. The evaluation consisted an objective segmental SNR measure [10], a subjective study of speech spectrograms and informal listening tests.

Table 1 summarizes the values of parameters used in the implementation of the proposed algorithm. We chose Hanning windows in (1) and (17). The noise power spectrum, $\lambda_d(k,\ell)$, was estimated from the noisy signals using the *Minima Controlled Recursive Averaging* (MCRA) approach [11].

The segmental SNR improvements obtained by the above-mentioned estimators are compared in Table 2. This measure takes into account both residual noise and speech distortion. The proposed estimator achieves the best results under all noise conditions. Its superiority is more significant for low input SNRs and non-stationary noise.

Since the segmental SNR lacks indication about the structure of the residual noise, a subjective comparison was conducted using speech spectrograms and validated by informal listening tests. Example of speech spectrograms obtained by the proposed algorithm are shown in Fig. 3. In contrast to other methods, where high *a posteriori* SNR produces high spectral gain resulting in a random appearance of tone-like noise (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.

	Car Noise				White Gaussian Noise			
Noisy Speech SNR (dB)	-5.0	0.0	5.0	10.0	-5.0	0.0	5.0	10.0
Noisy Speech Segmental SNR (dB)	-13.21	-8.21	-3.21	1.78	-14.47	-9.47	-4.47	0.53
Enhancement Method	Segmental SNR Improvement (dB)				Segmental SNR Improvement (dB)			
STSA, No Modification ($q \equiv 0$) [5]	8.38	7.72	7.25	6.87	8.81	8.26	7.57	6.68
LSA, No Modification ($q \equiv 0$) [1]	10.55	9.41	8.73	8.36	11.11	10.28	9.24	7.94
Multiplicatively-Modified LSA [2]	13.22	11.30	10.22	9.87	13.94	12.58	11.03	9.13
Optimally-Modified LSA	**13.99**	**11.81**	**10.72**	**10.16**	**14.39**	**12.88**	**11.16**	**9.16**

Table 2. Improvements in segmental SNR with the optimally-modified LSA estimator, compared to other estimators, for various environmental conditions.

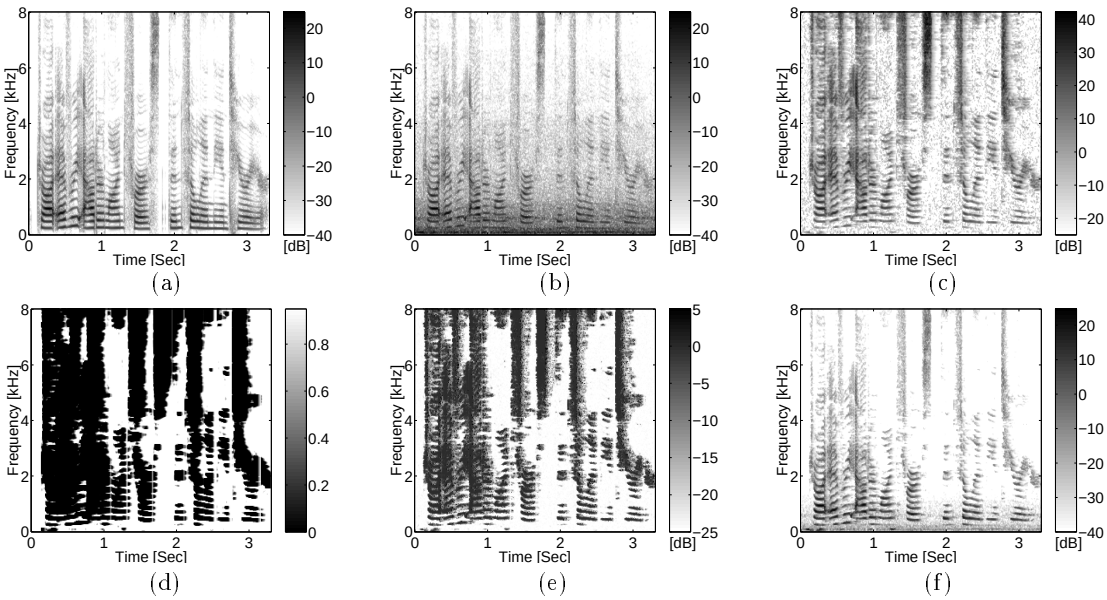

Fig. 3. Speech spectrograms: (a) Original clean speech signal: "Draw every outer line first, then fill in the interior."; (b) Noisy speech in the case of additive car noise (SNR= 0dB, $\text{SNR}_{seg} = -8.21$dB); (c) Decision-directed a priori SNR estimate ($\hat{\eta}$); (d) A priori probability for signal absence estimate ($\hat{q}$); (e) Modified spectral gain; (f) Speech enhanced with the proposed method (SNR_{seg} Improvement= 11.81dB).

5. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," IEEE Trans. ASSP, vol. ASSP-33, pp. 443–445, 1985.

[2] D. Malah, R. V. Cox and A. J. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," ICASSP, 1999, pp. 789–792.

[3] R. Martin, I. Wittke and P. Jax, "Optimized Estimation of Spectral Parameters for the Coding of Noisy Speech," ICASSP, 2000, pp. 1479–1482.

[4] N. S Kim and J.-H. Chang, "Spectral Enhancement Based on Global Soft Decision," IEEE SP Let., vol. 7, pp. 108–110, 2000.

[5] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," IEEE Trans. ASSP, vol. ASSP-32, pp. 1109–1121, 1984.

[6] R. J. McAulay and M. L. Malpass "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," IEEE Trans. on ASSP, vol. ASSP-28, pp. 137–145, 1980.

[7] J. Sohn, N. S Kim and W. Sung, "A Statistical Model-Based Voice Activity Detector," IEEE SP Let., vol. 6, pp. 1–3, 1999.

[8] J. Yang, "Frequency Domain Noise Suppression Approaches in Mobile Telephone Systems," ICASSP, 1993, pp. 363–366.

[9] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," IEEE Trans. Speech and Audio Proc., vol. 2, pp. 345–349, 1994.

[10] S. Quackenbush, T. Barnwell and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[11] I. Cohen and B. Berdugo, "Spectral Enhancement by Tracking Speech Presence Probability in Subbands," in *Proc. IEEE Workshop on Hands Free Speech Communication*, Kyoto, Japan, April 2001.