Combined Detection and Estimation Based on Mean-Square Error Log-Spectral Amplitude for Speech Enhancement

van Khanh Mai, Dominique Pastor, Abdeldjalil Aissa El Bey, Raphaël Le Bidan

Combined Detection and Estimation Based on Mean-Square Error Log-Spectral Amplitude for Speech Enhancement

Van-Khanh MAI, Dominique PASTOR, Abdeldjalil AÏSSA-EL-BEY, Raphaël LE BIDAN

IMT-Atlantique, UMR CNRS 6285 Lab-STICC
Technopôle Brest Iroise, CS 83818, 29238 Brest, France
{van.mai, dominique.pastor, abdeldjalil.aissaelbey,
raphael.lebidan}@telecom-bretagne.eu

Résumé – Nous présentons une nouvelle méthode qui combine la détection et l'estimation du signal de parole pour en améliorer le débruitage et la perception après débruitage. La méthode proposée associe une fonction de type Neyman-Pearson pour détecter la présence ou l'absence de la parole dans chaque case du plan temps-fréquence à un estimateur bayésien dédié à l'estimation de l'amplitude spectrale à court terme. Basée sur l'erreur quadratique du logarithme de l'amplitude spectrale à court terme, notre méthode permet de réduire davantage le bruit de fond sans introduire de distorsion du signal. Les résultats montrent que cette approche combinée améliore non seulement le débruitage en lui-même, mais également l'intelligibilité du signal de parole débruité, en comparaison avec d'autres méthodes de référence.

Abstract – In this paper, we address the problem of simultaneously detecting and estimating the speech short-time spectral amplitude (STSA) through combinations of the Neyman-Pearson and Bayesian approaches for speech enhancement. The main idea is that a non-continuous cost function, which depends on the absence/presence of speech, is applied to optimal joint detection and estimation for improving performance of mean square error estimators. Furthermore, our proposed method based on the square error of the logarithmic STSA makes it possible for us to reduce much more the background noise without introducing severe signal distortion. Preliminary experimental results demonstrate the advantage of our method in terms of speech quality and intelligibility.

1 Introduction

Optimal algorithms used to remove or reduce background noise are frequently preferred in speech enhancement. By assuming a particular statistical distribution on the signal of interest and by processing observation in the short time Fourier transform (STFT) domain, some optimal estimators have been proposed in [1–3]. Nevertheless, in these algorithms, speech is supposed to be present in every time-frequency bin, which degrades performance. Hence, some studies try to estimate the speech short-time spectral amplitude (STSA) under signal presence uncertainty for improving quality of speech [1, 4], which provides much more attenuation because the gain function is multiplied by the speech presence probability. A similar approach applied to log-spectral amplitude (LSA) is proposed in [5], but the resulting method cannot yield better performance than standard LSA [6]. Recent research efforts have considered combined detection and estimation for improving performance [7]. This method uses a non-continuous cost function based on the square error of the magnitude spectra, which is not enough subjectively meaningful [6].

In this paper, we propose a new log spectral amplitude estimator based on joint detection and estimation theory. By defining the cost function on the log-spectral amplitude error, we determine a gain function in the form of a generalized binary mask, which enables improved speech intelligibility [8].

The paper is organized as follows. The proposed algorithm is introduced in Section 2. Then, experimental results are presented in Section 3. Finally, Section 4 concludes this paper.

2 Joint detection and estimation

In speech enhancement applications, noisy speech is often segmented, windowed and transformed by STFT. The corrupted speech in the time-frequency domain is expressed by $Y_{mk} = S_{mk} + X_{mk}$, where m and k are the time frame and frequency-bin indices and S_{mk} and X_{mk} are the STFT coefficients of the clean speech signal and noise, respectively. These STFT coefficients are assumed to have complex Gaussian distribution with zero-mean [1]. For simplicity, the m, k indices will be omitted in the sequel unless required for clarification, and the estimated signals are pointed by a wide hat symbol. The noisy coefficients are rewritten in polar form as $Re^{\phi_Y} = Ae^{\phi_S} + Ne^{\phi_X}$, where $\{R, A, N\}$ and $\{\phi_Y, \phi_S, \phi_X\}$ denote the amplitude and phase of observed signal, clean speech and noise correspondingly. In addition, speech and noise are assumed to be independent so that the noisy spectral power is $\mathbf{E}(R^2) = \mathbf{E}(A^2) + \mathbf{E}(N^2) = \sigma_S^2 + \sigma_X^2$, where the spectral speech and noise power are denoted by $\mathbf{E}(A^2) = \sigma_S^2$, $\mathbf{E}(N^2) = \sigma_X^2$. The *a priori* signal-to-noise ratio (SNR) ξ and the *a posteriori* SNR γ are also defined as follows $\xi = \sigma_S^2/\sigma_X^2$, $\gamma = R^2/\sigma_X^2$.

The main strategy of joint detection and estimation methods is that a detector is first applied to each time-frequency bin for detecting the presence of speech. Then, an estimator is used to retrieve the signal of interest. In the classical two-states model for the presence/absence of speech, the observed signal is usually given by

$$\begin{aligned} H_1 &: Y = S + X \\ H_0 &: Y = X, \end{aligned} \tag{1}$$

where H_1 and H_0 denote the speech presence and speech absence hypotheses in each time-frequency bin, respectively. As in [1], we suppose that :

$$f_Y(y|H_0) = \frac{1}{\pi\sigma_X^2}\exp\left(-\frac{|y|^2}{\sigma_X^2}\right), \tag{2}$$

$$f_Y(y|H_1) = \frac{1}{\pi\sigma_X^2(1+\xi)}\exp\left(-\frac{|y|^2}{\sigma_X^2(1+\xi)}\right), \tag{3}$$

for any complex value y and where $f_Y(\cdot|H_i)$ is the probability density function (pdf) of Y under hypothesis H_i, $i \in \{0,1\}$. Generally, Bayesian estimators rely on Bayes risks that are constructed via a cost function $\mathbf{C}(\widehat{A}, A)$ where A is the true amplitude and $\widehat{A}$ is its estimate.

With the introduction of a decision D on the presence or absence of speech, the cost function must reflect that the situation is different depending on the decision made. To this end, we follow [9]. Specifically, if the decision is that no speech is present ($D = H_0$), the estimate will be 0 and the value of the cost function is then $\mathbf{C}(A) = \mathbf{C}(0, A)$. However, under hypothesis H_1, such an estimate is incorrect since some speech signal is present with non-zero amplitude. On the other hand, still under H_1, no decision error is made if $D = H_1$ and the cost value is $\mathbf{C}(\widehat{A}, A)$. The decision D being a random variable taking value in $\{H_0, H_1\}$ with conditional probabilities $\mathbb{P}(D = H_0|Y)$ and $\mathbb{P}(D = H_1|Y)$, the estimation cost under H_1 becomes the weighted sum $\mathbf{C}(\widehat{A}, A)\mathbb{P}(D = H_1|Y) + \mathbf{C}(0, A)\mathbb{P}(D = H_0|Y)$. For simple notation, let $\mathbb{P}(H_i|Y)$ denote $\mathbb{P}(D = H_i|Y)$ and the average Bayes risk $\mathbf{R}$ under H_1 is then defined by :

$$\begin{aligned} &\mathbf{R}\left(\widehat{A}, \mathbb{P}(H_0|Y), \mathbb{P}(H_1|Y)\right) = \\ &\mathbf{E}_1\left[\mathbf{C}(\widehat{A}, A)\mathbb{P}(H_1|Y) + \mathbf{C}(A)\mathbb{P}(H_0|Y)\right], \end{aligned} \tag{4}$$

where $\mathbf{E}_1$ stands for the expectation under H_1. Such an approach is similar to the ideal binary mask [8].

Let the probability of a false alarm be $\mathbb{P}(D = H_1|H_0)$. The joint detection and estimation method then results in the following constrained minimization problem

$$\begin{aligned} \min_{\widehat{A}, \mathbb{P}(H_0|Y), \mathbb{P}(H_1|Y)} \quad & \mathbf{R}\left(\widehat{A}, \mathbb{P}(H_0|Y), \mathbb{P}(H_1|Y)\right) \\ \text{subject to} : \quad & \mathbb{P}(D = H_1|H_0) \leq \alpha \end{aligned} \tag{5}$$

This problem is addressed and solved in [9, theorem 1]. The obtained result is simply formulated as

$$\widehat{A} = \arg\min_a \mathbf{E}_1\left[\mathbf{C}(a, A)\right] \text{ if } D = H_1 \tag{6}$$

$$\frac{f_Y(y|H_1)}{f_Y(y|H_0)}\left[\mathbb{C}(Y) - \mathbb{C}(\widehat{A}, Y)\right] \underset{D=H_0}{\overset{D=H_1}{\gtrless}} \tau, \tag{7}$$

where $\mathbb{C}(Y) = \mathbf{E}_1[\mathbf{C}(A)|Y]$ and $\mathbb{C}(\widehat{A}, Y) = \mathbf{E}_1[\mathbf{C}(\widehat{A}, A)|Y]$. In addition, τ is calculated by imposing $\mathbb{P}(D = H_0|H_1) = \alpha$.

In speech enhancement, LSA error is known to be more meaningful from the subjective point of view [6]. This motivates us to propose in the two following two novel detectors based on two different cost function models.

2.1 Optimum joint method (OJLSA)

Since the cost function should take advantage of the presence/absence speech hypotheses and use the LSA error for speech enhancement, our first cost function is chosen as

$$\mathbf{C}(A) = \left(\log(A) - \log(\varepsilon)\right)^2 \tag{8}$$

$$\mathbf{C}(\widehat{A}, A) = \left(\log(\widehat{A}) - \log(A)\right)^2 \tag{9}$$

where ε ($0 < \varepsilon \leq A$) is a fixed constant that allows us to obtain a monotonic cost function. Therefore, the Bayesian estimator under H_1 hypothesis is as follows

$$\log(\widehat{A}) = \int_0^\infty \log(a) f_A(a|Y, H_1) da, \tag{10}$$

and can be simply written by

$$\widehat{A} = G_{\text{LSA}}(\xi, \gamma) R, \tag{11}$$

where $G_{\text{LSA}}(\xi, \gamma)$ is the gain function of two variables ξ, γ proposed in [2]. The value of $\mathbb{C}(Y)$ is calculated by

$$\begin{aligned} \mathbb{C}(Y) = \mathbf{E}_1[\mathbf{C}(A)|Y] &= \int_0^\infty \mathbf{C}(a) f_A(a|Y, H_1) da \\ &= \int_0^\infty \left(\log(a) - \log(\varepsilon)\right)^2 f_A(a|Y, H_1) da \end{aligned} \tag{12}$$

and similarly, the cost value with the optimal estimate $\widehat{A}$ is

$$\begin{aligned} \mathbb{C}(\widehat{A}, Y) &= \mathbf{E}_1[\mathbf{C}(\widehat{A}, A)|Y] \\ &= \int_0^\infty \mathbf{C}(\widehat{A}, a) f_A(a|Y, H_1) dA \\ &= \int_0^\infty \left(\log(a) - \log(\widehat{A})\right)^2 f_A(a|Y, H_1) da, \\ &= \mathbb{C}(Y) - \left(\log(\widehat{A}) - \log(\varepsilon)\right)^2. \end{aligned} \tag{13}$$

Using (2) and (3) we have the likelihood ratio as

$$\frac{f_Y(y|H_1)}{f_Y(y|H_0)} = \frac{\exp\left(\frac{\gamma\xi}{1+\xi}\right)}{1+\xi} = \frac{\exp(\nu)}{1+\xi}, \tag{14}$$

where $\nu = \frac{\gamma\xi}{1+\xi}$. Using the results of (11), (13) and (14), we obtain from (6) and (7) that the proposed estimator simplifies to

$$\widehat{A} = G_{\text{OJLSA}}(\xi, \gamma, R) R, \tag{15}$$

where the spectral gain function G_{OJLSA}

$$G_{\text{OJLSA}}(\xi,\gamma,R)=\begin{cases} G_{\text{LSA}}(\xi,\gamma) & \text{if } \mathcal{D}_{\text{OJ}}(R)\geq\tau,\\ 0 & \text{otherwise}\end{cases} \quad (16)$$

is a function of three parameters : the *a priori* SNR ξ, the *a posteriori* SNR γ and the spectral amplitude R, and where the detector is

$$\mathcal{D}_{\text{OJ}}(R)=\frac{\exp(\nu)}{1+\xi}\left(\log(G_{\text{LSA}}(\xi,\gamma)R)-\log(\varepsilon)\right)^2. \quad (17)$$

2.2 Sub-optimum joint method (SJLSA)

For ensuring non-decreasingness of the cost function, eliminating the fixed constant ε and taking advantage at the same time from the performance of the LSA approach, an alternative cost function can be defined as

$$\mathbf{C}(A)=\left(\log(A+1)\right)^2 \quad (18)$$

$$\mathbf{C}(\widehat{A},A)=\Big(\log(\widehat{A}+1)-\log(A+1)\Big)^2 \quad (19)$$

The choice of $\mathbf{C}(A)$ using (18) is suitable for penalizing the decision in terms of LSA. The cost function is monotonically increasing and equals zero when the true amplitude is zero. This choice of $\mathbf{C}(\widehat{A},A)$ is adapted to these constraints.

Following the same demarch as in the previous subsection, the corresponding estimator under hypothesis H_1 is

$$\log(\widehat{A}+1)=\int_0^\infty \log(a+1)f_A(a|Y,H_1)da. \quad (20)$$

Thus, the cost value of $\mathbb{C}(Y)$ is given by

$$\mathbb{C}(Y)=\int_0^\infty \left(\log(a+1)\right)^2 f_A(a|Y,H_1)da. \quad (21)$$

and $\mathbb{C}(\widehat{A},Y)$ is therefore,

$$\begin{aligned}\mathbb{C}(\widehat{A},Y)&=\mathbf{E}_1[\mathbf{C}(\widehat{A},A|Y)]\\ &=\int_0^\infty \mathbf{C}(\widehat{A},a)f_A(a|Y,H_1)da\\ &=\mathbb{C}(Y)-\Big(\log(\widehat{A}+1)\Big)^2.\end{aligned} \quad (22)$$

The estimator of (20) is similar to that of (10). The latter will thus be used to approximate the former since the integral in (20) is hardly tractable. Hence, we obtain a sub-optimal spectral gain function

$$G_{\text{SJLSA}}(\xi,\gamma,R)=\begin{cases} G_{\text{LSA}}(\xi,\gamma) & \text{if } \mathcal{D}_{\text{SJ}}(R)\geq\tau,\\ 0 & \text{otherwise,}\end{cases} \quad (23)$$

where $\mathcal{D}_{\text{SJ}}(R)$ is calculated similarly to $\mathcal{D}_{\text{OJ}}(R)$ and is thus given by :

$$\mathcal{D}_{\text{SJ}}(R)=\frac{\exp(\nu)}{1+\xi}\left(\log(G_{\text{LSA}}(\xi,\gamma)R+1)\right)^2 \quad (24)$$

The detectors $\mathcal{D}_{\text{OJ}}(R)$ (17) and $\mathcal{D}_{\text{SJ}}(R)$ (24) are slightly different. Both are monotonic increasing and depend on the LSA estimators. In turn, the OJLSA and SJLSA estimators depend on the detectors. This twofold dependency is expected to improve the performance of the two detectors and estimators. However, in contrast to OJLSA, SJLSA does not introduce any auxiliary parameter ε, which should be beneficial.

3 Experimental results

We assessed our proposed methods OJLSA and SJLSA on the NOIZEUS database [6]. This database contains IEEE sentences corrupted by noise from the AURORA database and synthetic noise as white noise and auto-regressive noise (AR), at four levels, namely 0, 5, 10 and 15 dB. In our experiments, speech signals are sampled at 8 kHz, segmented into frames of 256 samples each, transformed by STFT with 50% overlapped Hamming windows. Thresholds are calculated by fixing the false alarm probability to 0.05 for all noise levels. The parameter ε in the method OJSLA is chosen as $\varepsilon=\beta R$, where $\beta=-25$dB. The noise power spectrum is estimated by method proposed in [10] :

For assessing speech quality and preliminary speech intelligibility after denoising, objective quality and intelligibility criteria have been used. Speech quality is firstly measured by segmental SNR (SSNR) and perceptual evaluation of speech quality (PESQ). The SSNR values were set in the range $[-10, 35$ dB$]$ to bypass the use of a silence/speech detector [6]. PESQ is recommended by ITU-T and aims at predicting the subjective mean opinion score (MOS). Secondly, intelligibility of speech was initially evaluated by the short-time objective intelligibility measure (STOI), which highly correlates with intelligibility measured by listening tests. Roughly speaking, STOI measures the mean correlation between clean and enhanced speech coefficients [11]. A logistic function is applied to map the STOI measure onto intelligibility scores :

$$f(\text{STOI})=\frac{100}{1+\exp(a\times\text{STOI}+b)}, \quad (25)$$

where, for fitting with IEEE sentences, $a=-17.4906$ and $b=9.6921$ [11].

The average results for different noise types and SNR values are given in Table 1, where the best scores are highlighted using boldface notation. The proposed methods are compared to the original LSA and to the optimal LSA under speech presence uncertainty (OMLSA) proposed in [5]. On the one hand, in terms of PESQ measurements, the proposed methods perform better than LSA and OMLSA for quasi-stationary (white, AR, car) noises and yield similar performance to the best LSA method for non-stationary noises (airport, babble). For the SSNR criterion, OJLSA and SJLSA achieve better than or similar performance measurements as the OMLSA method. On the other hand, in terms of STOI, SJLSA outperforms the other methods for all noise types, and especially at low SNR levels.

4 Conclusion

In this paper, we have proposed a joint detection and estimation method based on log-spectral amplitude estimation for speech enhancement. The key idea is to take into account the presence and absence of speech in each time-frequency bin. Optimal hard detectors are then derived to improve quality of speech in noisy environments. The experimental results have shown the

TABLE 1 – Performance evaluation with three criteria : PESQ, STOI, SSNR

Noise	Method	PESQ				SSNR				STOI(%)			
		0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB	0dB	5dB	10dB	15dB
White	LSA	2.07	2.46	2.79	3.07	0.61	2.87	5.26	7.57	84.97	96.65	99.11	99.67
	OMLSA	1.96	2.36	2.71	2.94	**1.69**	**3.66**	**5.86**	8.03	66.78	91.57	98.26	99.43
	OJLSA	**2.14**	**2.53**	**2.87**	**3.19**	1.23	3.47	5.78	8.00	**86.07**	97.04	99.20	**99.70**
	SJLSA	2.13	2.52	**2.87**	3.18	1.23	3.48	5.79	**8.04**	85.89	**97.09**	**99.22**	99.69
AR	LSA	2.07	2.48	2.81	3.09	-0.15	2.11	4.55	7.05	76.75	94.22	98.77	99.64
	OMLSA	1.80	2.26	2.62	2.89	**0.82**	**2.81**	5.07	7.50	43.54	82.57	96.96	99.35
	OJLSA	**2.13**	**2.52**	**2.88**	**3.21**	0.43	2.70	5.06	7.52	**78.24**	95.16	98.92	99.66
	SJLSA	2.12	2.51	2.87	**3.21**	0.43	2.72	**5.08**	**7.55**	78.13	**95.22**	**98.94**	**99.67**
Car	LSA	2.07	2.46	2.80	3.06	-0.21	2.07	4.42	6.88	83.92	97.26	99.42	99.80
	OMLSA	1.75	2.24	2.63	2.92	**0.75**	**2.84**	4.99	7.35	56.17	91.44	98.88	99.74
	OJLSA	**2.09**	**2.48**	**2.85**	**3.16**	.42	2.66	5.00	7.40	85.75	97.52	99.46	**99.81**
	SJLSA	2.08	2.47	**2.85**	**3.16**	0.45	2.70	**5.07**	**7.45**	**85.84**	**97.57**	**99.49**	**99.81**
Airport	LSA	**2.16**	**2.51**	2.85	3.15	-0.07	2.08	4.55	7.02	88.80	98.00	99.58	**99.86**
	OMLSA	1.86	2.31	2.72	3.05	**0.93**	**2.77**	**5.17**	7.51	68.69	94.43	99.34	99.83
	OJLSA	2.13	2.50	2.87	**3.24**	0.37	2.51	5.03	7.45	88.83	98.03	**99.59**	99.85
	SJLSA	2.13	2.49	**2.88**	3.23	0.54	2.66	5.11	**7.52**	**89.55**	**98.19**	**99.59**	**99.86**
Babble	LSA	**2.08**	**2.47**	2.79	3.11	-0.56	1.66	4.30	6.77	80.97	96.93	98.50	**99.83**
	OMLSA	1.75	2.22	2.63	2.99	**0.49**	**2.44**	**4.98**	**7.30**	52.75	92.47	99.21	99.80
	OJLSA	2.07	2.44	2.80	3.17	-0.19	2.05	4.72	7.14	81.99	97.07	99.52	**99.83**
	SJLSA	2.04	2.43	**2.81**	**3.17**	0.01	2.22	4.82	7.23	**82.98**	**97.40**	**99.54**	**99.83**

relevance of the approach. Whenever speech is detected as absent, OJLSA and SJLSA set the STSA to zero. This is questionable. Additional performance improvement may be possible by adding small background noise instead. Future work will conduct a performance evaluation by associating the noise reduction method with a noise estimation method. A detector that operates regardless of any prior on the signal distribution [12] could also be considered to further improve robustness.

Références

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

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

[3] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[4] N. S. Kim and J. H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Process. Lett.*, vol. 7, no. 5, pp. 108–110, 2000.

[5] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, 2002.

[6] P.C. Loizou, *Speech Enhancement : Theory and Practice*, CRC press, 2013.

[7] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, 2012.

[8] Y. Hu and P. C. Loizou, "Techniques for estimating the ideal binary mask," in *Proc. 11th Int. Workshop Acoust. Echo Noise Control*, 2008, pp. 154–157.

[9] G. V. Moustakides, "Optimum joint detection and estimation," in *Proc. IEEE Int. Inf. Theory*, 2011, pp. 2984–2988.

[10] V. K. Mai, D. Pastor, A. Aïssa-El-Bey, and R. Le-Bidan, "Robust estimation of non-stationary noise power spectrum for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 670–682, 2015.

[11] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[12] D. Pastor and Q. T. Nguyen, "Random distortion testing and optimality of thresholding tests," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4161–4171, 2013.