A Noise Estimation Method Based on Speech Presence Probability and Spectral Sparseness

Chao Li, and Wenju Liu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

{cli,lwj}@nlpr.ia.ac.cn

Abstract

This paper addresses the problem of noise power spectrum estimation. Existing noise estimation methods cannot perform quite reliably when noise level increasing abruptly (e.g., narrowband noise bursts). To overcome this problem, we improve the time-recursive averaging algorithm based on speech presence probability (SPP), by exploiting the sparseness of speech spectrum. Firstly, we utilize the SPP estimation method based on fixed *prior*s to achieve low SPP estimates at time-frequency bins where speech is absent. Furthermore, a spectral sparseness measure is proposed to adjust the SPP estimates. Experiments show the proposed method can update the noise estimates faster than state-of-the-art approaches in both stationary and nonstationary noise.

Index Terms: noise estimation, spectral sparseness, speech presence probability, speech enhancement

1. Introduction

Reliable noise tracking is critical for the performance of speech-enhancement algorithms under low signal-to-noise (SNR) conditions and nonstationary noise environments. Noise estimation is a particularly challenging task in adverse environments.

Traditional noise estimation methods based on voice activity detectors (VAD) algorithm [1] and update the noise spectrum during the silent segments of signal. Obviously, such a method cannot work satisfactorily in nonstationary noise environments. The minimum statistics (MS) method [2] achieves biased estimate of the noise spectrum by tracking the minimum of noisy speech over a finite window, but it cannot follow the changing noise immediately but with a delay of minimum-search window. The minima controlled recursive averaging (MCRA) [3] algorithm and the improved MCRA (IMCRA) [4] update the noise estimate by tracking the noise-only regions of the noisy speech spectrum based on the speech presence probability (SPP). However, the noise spectrum estimated by both methods lags by at most twice that window length when the noise spectrum increases abruptly. To reduce the noise estimation delay, another MCRA (MCRA-2) [5] uses the continuous spectral minima tracking method [6] for the local minimum, instead of the MS approach. However, it fails to differentiate between an increase in noise floor and an increase of speech power.

To overcome these problems, in this paper we propose a novel noise estimate method using the time-recursive averaging algorithm. Firstly, the SPP estimation method based on fixed *prior*s is introduced and adjusted for our application. This method avoids the estimation of *a priori* SNR and *a priori* SPP, while achieves lower SPP estimates at time-frequency bins where speech is absent than traditional methods. Secondly, the spectral sparseness of speech is taken into account, and an effective measure for spectral sparseness is proposed to adjust the SPP estimates. The adjustment procedure enables the elimination of the increasing SPP estimates caused by narrowband noise bursts in time-frequency bins where speech is absent. Finally, the noise power spectrum is updated using adjusted smoothing factor. The experimental results verify the effectiveness of the proposed method compared with the state-of-the-art methods.

The rest of this paper is organized as follows: Section 2 presents the proposed noise estimation method. Section 3 describes the comparison experiments and discusses the results. Section 4 concludes this paper.

2. Proposed Noise Estimation Method

In short-time Fourier transform (STFT) domain, we assume the observation $Y(k,l)$ is an additive mixture of clean speech $X(k,l)$ and the noise $D(k,l)$. Here k is the frequency bin index, l is the time frame index. Given two hypotheses, $H_0(k,l)$ and $H_1(k,l)$, which indicate speech absence and presence respectively, noise power spectrum $\sigma_d^2(k,l)$ can be estimated, respectively [3]:

$$\begin{aligned} H_0(k,l) &: \hat{\sigma}_d^2(k,l) = \alpha\hat{\sigma}_d^2(k,l-1) + (1-\alpha)|Y(k,l)|^2, \\ H_1(k,l) &: \hat{\sigma}_d^2(k,l) = \hat{\sigma}_d^2(k,l-1), \end{aligned} \tag{1}$$

where $0 \leq \alpha \leq 1$ is a smoothing factor. (1) is based on the principle that the noise estimate is updated whenever speech is absent, otherwise it is kept constant. The optimum estimate, in terms of minimum mean-square error (MMSE), is given by:

$$\begin{aligned} \hat{\sigma}_d^2(k,l) = E\left\{\sigma_d^2(k,l)\,|H_0\right\} P(H_0\,|Y(k,l)\,) \\ +E\left\{\sigma_d^2(k,l)\,|H_1\right\} P(H_1\,|Y(k,l)\,). \end{aligned} \tag{2}$$

Based on the two hypotheses stated in (1), we can express $\hat{\sigma}_d^2(k,l)$ as [4]:

$$\hat{\sigma}_d^2(k,l) = \alpha_d(k,l)\hat{\sigma}_d^2(k,l-1) + [1-\alpha_d(k,l)]\,|Y(k,l)|^2, \tag{3}$$

where

$$\alpha_d(k,l) \triangleq \alpha + (1-\alpha)p(k,l), \tag{4}$$

where $p(k,l) \triangleq P(H_1\,|Y(k,l)\,)$ denotes the *a posteriori* probability of speech presence.

2.1. Estimation of *a posteriori* SPP

The approach taken to estimate *a posteriori* SPP in each time-frequency bin is similar to the method proposed in [7]. We assume that the STFT coefficients, for both speech and noise, are

complex Gaussian variables, and apply Bayes rule for the conditional SPP, one obtains

$$p(k,l) = \frac{\Lambda(k,l)}{1+\Lambda(k,l)}, \tag{5}$$

where Λ_k is the generalized likelihood ratio (GLR) defined by

$$\Lambda(k,l) = \frac{q(k,l)}{1-q(k,l)} \frac{p\left(Y\left(k,l\right)|H_1\right)}{p\left(Y\left(k,l\right)|H_0\right)}, \tag{6}$$

where, $q\left(k,l\right) = P\left(H_1\left(k,l\right)\right)$ is the *a priori* SPP, and we assume that the speech and noise are equally likely and use the fixed *a priori* SPP $q = 0.5$.

In [7] an *a posteriori* SPP estimation method is proposed based on fixed *a priori* SNR. The authors consider the *a priori* SNR should reflect the SNR that is expected if speech was present. An optimal fixed *a priori* SNR is found in [7] to minimize the total probability of error.

Therefore, $p(k,l)$ is a function of the *a posteriori* SNR, which can be calculated by,

$$\gamma\left(k,l\right) = |Y(k,l)|^2 / \hat{\sigma}_d^2(k,l-1). \tag{7}$$

In order to reduce random fluctuations in $p(k,l)$, which may result in *musical noise* in enhanced signal, the smoothed observation over a time-frequency region is calculated by

$$\bar{\gamma}\left(k,l\right) = \frac{1}{\left(2\Delta_k+1\right)\left(\Delta_l+1\right)} \sum_{j=l-\Delta_l}^{l} \sum_{i=k-\Delta_k}^{k+\Delta_k} \gamma\left(i,j\right), \tag{8}$$

where Δ_k and Δ_l are the range of time frame and frequency bin for smoothing, respectively.

As the STFT coefficients are complex Gaussian distribution, the resulting values $\bar{\gamma}\left(k,l\right)$ are approximately chi-squared distributed. Thus, the GLR can be expressed by [7]:

$$\Lambda(k,l) = \frac{q}{1-q}\left(\frac{1}{1+\xi}\right)^{r/2} \exp\left(\frac{\xi\bar{\gamma}(k,l)}{1+\xi}\frac{r}{2}\right), \tag{9}$$

where r is the degree of freedom.

In order to ensure a low false-alarm rate and preserve the fine structure of speech, the combination of two initial SPPs $\bar{p}_{\mathrm{local}}\left(k,l\right)$ and $\bar{p}_{\mathrm{global}}\left(k,l\right)$ is applied, based on averaging windows of different size. Thus, the final SPP is achieved as a multiplication:

$$\bar{p}\left(k,l\right) = \bar{p}_{\mathrm{global}}\left(k,l\right) \cdot \bar{p}_{\mathrm{local}}\left(k,l\right). \tag{10}$$

The values of parameters used in implementation of SPP estimation are summarized in Tab.1

Table 1: *Parameter values for SPP estimate*.

	Δ_k	Δ_l	r	ξ
local	1	2	10.7	8 dB
global	4	2	27.8	3 dB

This method of *a posteriori* SPP estimate overcomes the necessity for adaptively tracking the *a priori* SPP and the *a priori* SNR, which enables a decoupling of the estimation of the *a posteriori* SPP and the estimation of clean-speech coefficients. Furthermore, this method provides low *a posteriori* SPP estimates at time-frequency bins where speech is absent, which can benefit the update of noise power estimate. However, this method cannot distinguish narrowband noise bursts from speech onsets, especially under nonstationary noise environments and low SNR conditions.

2.2. Computing the spectral sparseness

Why cannot the presented approach differentiate between narrowband noise burst and an increase of speech power? Because the smoothed estimation of *a posteriori* SNR in (8) is sensitive to rising spectral amplitudes, it does not only respond to speech onsets, but also to noise bursts that are not tracked by the noise power estimation algorithm. Therefore, noise bursts will cause rising SPP estimates, and noise power spectrum cannot be updated in these frequency bins, even if the speech is absence. To overcome this problem, we propose the spectral sparseness measure to adjust the SPP estimates.

Firstly, in order to achieve reasonable spectral sparseness, the SPP estimates are normalized to 0 or 1,

$$\tilde{p}\left(k,l\right) = \begin{cases} 0 & \text{if } \bar{p}\left(k,l\right) < p_{\mathrm{th}}\left(l\right), \\ 1 & \text{if } \bar{p}\left(k,l\right) \geq p_{\mathrm{th}}\left(l\right), \end{cases} \tag{11}$$

where $p_{\mathrm{th}}\left(l\right)$ is the time-dependent threshold, which depends on the full-band SNR, $SNR(l)$, as follows:

$$p_{\mathrm{th}}\left(l\right) = \begin{cases} 0.7 & \text{if } SNR(l) < 0\mathrm{dB}, \\ 0.3 & \text{if } SNR(l) \geq 10\mathrm{dB}, \\ 0.7 - 0.04SNR(l) & \text{else}. \end{cases} \tag{12}$$

The allband SNR of the current frame can be calculated by:

$$SNR(l) = 10\log_{10}\left(\sum\nolimits_k |Y\left(k,l\right)|^2 \Big/ \sum\nolimits_k \hat{\sigma}_d^2(k,l-1)\right). \tag{13}$$

Then, with the normalized SPP estimates, we use a sparseness measure based on the relationship between the L_1 norm and the L_2 norm:

$$Q\left(l\right) = \frac{\sqrt{N} - \left(\sum_k \tilde{p}\left(k,l\right)\right) / \sqrt{\sum_k \tilde{p}^2\left(k,l\right)}}{\sqrt{N}-1}, \tag{14}$$

where N is the size of discrete Forurier transform (DFT). Particularly, $Q\left(l\right) = 1$ if and only if $\tilde{p}\left(k,l\right)$ contains only a single non-zero component, and $Q\left(l\right) = 0$ if and only if all components of $\tilde{p}\left(k,l\right)$ are equal to 1, interpolating smoothly between the two extremes.

As in [8], we create an ideal binary SPP from the clean speech signal that contains ones at all time-frequency bins, where the energy is no less than 50 dB below the maximum bin energy. Then, by using (14) we obtain the distribution discipline of speech spectral sparseness: above 90% of clean speech frames have lower sparseness measure than 0.71.

Finally, in order to eliminate the increasing SPP estimates caused by noise narrowband bursts, in time-frequency bins where speech is absent, SPP estimates are adjusted according to spectral sparseness as:

$$\hat{p}\left(k,l\right) = \begin{cases} \bar{p}\left(k,l\right) & Q\left(l\right) \leq 0.71, \\ 0 & Q\left(l\right) > 0.71. \end{cases} \tag{15}$$

The adjustment procedure means that if the current frame signal shows distinct sparseness, all bins are considered as noise, even if some bins have large SPP estimates.

2.3. Update of noise spectrum estimate

Substituting $\hat{p}\left(k,l\right)$ into (4), we compute the time-frequency dependent smoothing factor as follows:

$$\alpha_d(k,l) = \alpha + (1-\alpha)\hat{p}(k,l), \tag{16}$$

where $\alpha = 0.85$ is a fundamental smoothing factor. Note that $\alpha_d(k,l)$ takes values in the range of $\alpha \leq \alpha_d(k,l) \leq 1$.

Finally, the noise power spectrum estimate is updated as (3).

Hence, the overall algorithm can be summarized as follows:

1. Compute the *a posteriori* SNR $\gamma(k,l)$, and achieve the smoothed values $\bar{\gamma}(k,l)$ using (8).
2. Compute *a posteriori* SPP $\bar{p}(k,l)$ using (9)(5)(10).
3. Classify the frequency bins into speech present or absent using (12)(11), according to the allband SNR in (13).
4. Compute the spectral sparseness $Q(l)$ using (14), and adjust SPP estimates $\hat{p}(k,l)$ using (15).
5. Compute the smoothing factor $\alpha_d(k,l)$ using (16), and update noise power spectrum using (3).

3. Experimental Results

To evaluate the performance of the proposed approach, we select the state-of-the-art approaches for comparison, including MS [2], IMCRA [4] and MCRA-2 [5]. Furthermore, for the evaluation we implement all noise estimation approaches in the log-MMSE algorithm [4].

For evaluation, the improvement of the segmental SNR (segSNRI) and log-likelihood ratio (LLR) (marked in [0, 2]) [9] are adopted as the objective measures to denote noise reduction and speech distortion respectively. High speech quality is denoted by high values of the segSNRI, and low value of the LLR.

In our experiments, we utilize two stationary noise types taken from NOISEX92 database [10], i.e., white Gaussian noise (WGN) and babble noise. Additionally, two versions of nonstationary noise are simulated by adding a number of narrowband noise bursts around several fixed frequencies. The duration of every narrowband noise burst is longer than 1 s, and the increasing level of noise power is 15 dB. Therefore, the four available noise types are stationary WGN, stationary babble noise, nonstationary WGN, and nonstationary babble noise.

We processed 30 utterances selected from TIMIT database [11] (15 male, 15 female). The utterances are corrupted by four different noise types with input segmental SNR (SegSNR) of 0 dB. The signal is sampled at 16 KHz, hamming windowed using a 32-ms window and 50% overlap. Tab.2 and Tab.3 show the comparison results of various noise estimation methods in stationary noise and nonstationary noise, respectively.

Tab.2 indicates that the proposed method has a similar performance with IMCRA method in stationary noise, by obtaining relatively higher SegSNRI values and lower LLR values compared to the other methods. Simultaneously, Tab.3 indicates that the proposed method yields the highest SegSNRI values, while yields similar or lower LLR values than IMCRA and MCRA-2 methods, in nonstationary noise. Additionally, the comparison results in Tab.2 and Tab.3 suggest that MS method exhibits lower LLR and lower SegSNRI, because it always obtains underestimation of noise power spectrum. This outcome is confirmed by visual inspection of spectrograms and periodograms of noise estimations by various methods in nonstationary WGN (see Fig.1 and Fig.2 on the last page).

Fig.1 indicates that MS and IMCRA methods exhibit consistent underestimation of noise power spectrum, due to their inadaptability to increasing noise level. This undesired behavior is overcome by MCRA-2 to some extent. However, severe overestimation of noise power spectrum is produced in low frequency bins, due to its incapability to differentiate between an increase in noise floor and an increase of speech power. Another drawback of all comparison algorithms is the notable noise tracking delay. The proposed method can track narrowband noise burst quickly, while rejecting speech component into noise estimate. Subjective tests also reveal that the proposed method performs better intelligibility and comfort quality.

Fig.2 depicts the periodogram comparisons in the frequency of 1 KHz ($k = 32$). When noise level increases 15 dB abruptly at $t = 2.8$ s, although SPP estimate increases to 1, but the spectral sparseness still higher than the threshold 0.71, implying that speech is absent in the current frame. Therefore, the smoothing factor yields the lower value 0.85, which enables a fast tacking speed. The proposed method, however, does not utilize any minimal-tracking algorithm, and therefore it incurs a notable delay when noise level decreases abruptly ($t = 5.5$ s) and stays at that level.

Table 2: *Comparison results of various noise estimation methods in terms of SegSNRI (dB) and LLR, in stationary WGN and babble noise.*

	Stationary WGN		Stationary babble noise	
Method	SegSNRI	LLR	SegSNRI	LLR
MS	3.67	1.10	1.92	0.46
IMCRA	6.04	0.95	3.89	0.58
MCRA-2	4.49	0.98	3.70	0.77
Proposed	5.65	1.01	3.97	0.60

Table 3: *Comparison results of various noise estimation methods in terms of SegSNRI (dB) and LLR, in nonstationary WGN and babble noise.*

	Nonstationary WGN		Nonstationary babble noise	
Method	SegSNRI	LLR	SegSNRI	LLR
MS	2.65	1.10	1.67	0.51
IMCRA	3.56	1.03	3.04	0.66
MCRA-2	2.87	1.06	2.80	0.84
Proposed	4.35	1.03	3.06	0.67

4. Conclusions

In this paper, a novel noise estimation method was presented for speech enhancement in nonstationary noise environments, which based on two essential algorithms: SPP estimation approach based on fixed *prior*s, SPP adjustment approach based on spectral-sparseness measure. These approaches enable a fast tacking when noise level increases abruptly. Experiment results indicate that the proposed approach can achieve lower or similar LLR and higher SegSNRI comparing to existing methods, when integrated into a speech enhancement system.

5. Acknowledgements

This work was supported in part by the National Nature Science Foundation of China (No.60675026, No.60121302, No.90820011), and the National Grand Fundamental Research 973 Program of China (No.2004CB318105).

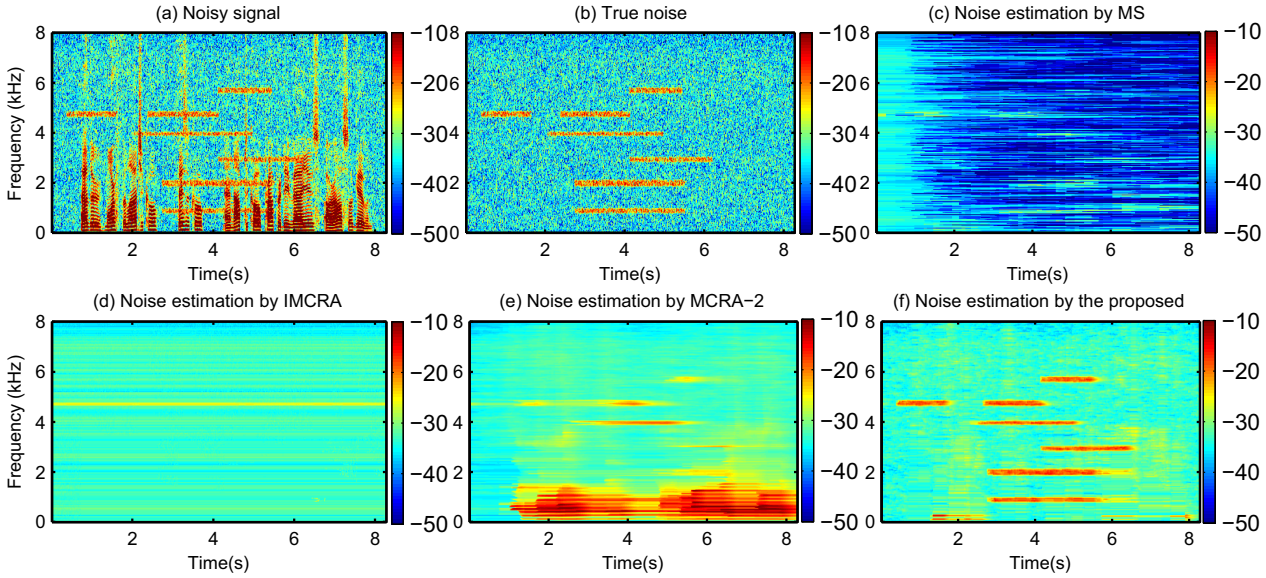

Figure 1: *Spectrograms of noisy signal (a) and true noise (b), and the noise estimates by MS (c), IMCRA (d), MCRA-2 (e) and the proposed (f).*

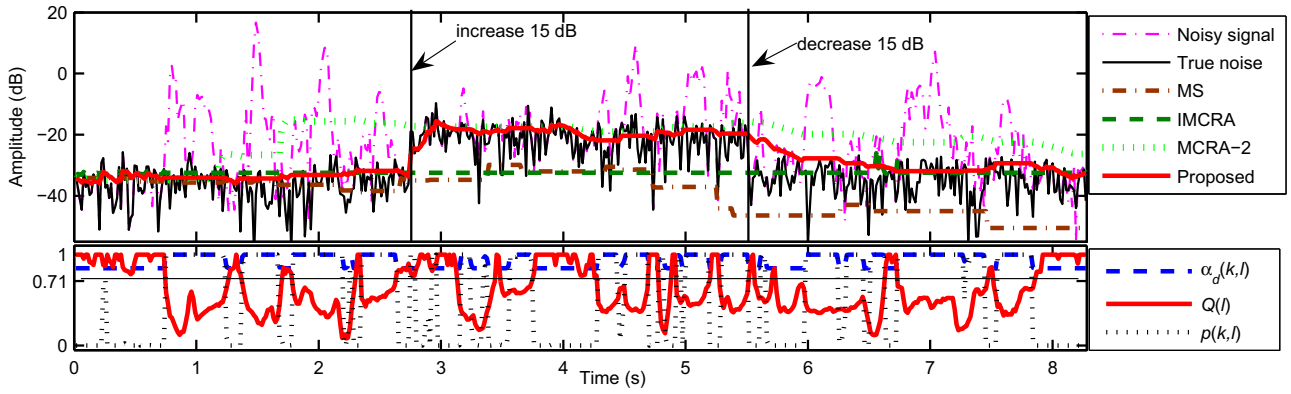

Figure 2: *Periodogram comparisons (top), smoothing factor, sparseness measure and SPP estimate (bottom), for frequency 1 KHz ($k = 32$).*

6. References

[1] Sohn, J., Kim, N., "Statistical model-based voice activity detection". *IEEE Signal Processing Letter*, 6(1):1-3. 1999.

[2] Martin, R., "Noise power spectral density estimation based on optimal smoothing and minimum statistics". *IEEE Transacation on Speech Audio Processing*, 9(5): 504-512, 2001.

[3] I. Cohen, "Noise estimation by minima controlled recursive averaging for robust speech enhancement", *IEEE signal Processing Letter*, 9(1): 12-15, 2002.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controller recursive averaging", *IEEE Transaction on Speech Audio Processing*, 11(5): 466-475, 2003.

[5] S. Rangachari, and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments", *Speech Communication*, 48: 220-231, 2006.

[6] Doblinger, G., "Computationally efficient speech enhancement by spectral minima tracking in subbands". *Proc. Eurospeech* 2: 1513-1516, 1995.

[7] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors", *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5): 910-919, 2008.

[8] 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 Transaction on Audio, Speech, and Language Processing*, 15(6), pp. 1741-1752, 2007.

[9] Y. Hu, and P. C. Loizou, "Evaluation of objuective quality measures for speech enhancement ", *IEEE transactions on Speech Audio Processing*, 16: 229-238, 2008.

[10] A. Varga, H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to studythe ePect of additive noise on speech recognition systems", *Speech Communication*, 12(3): 247-251, 1993.

[11] J.S. Garofolo, "DARPA TIMIT: an acoustic phonetic continuous speech database", *National Institude of Standards and Technology (NIST)*, 1988.