LETTER

Noise Spectrum Estimation Based on SNR Discrepancy for Speech Enhancement

Atanu SAHA[†a)], ***Nonmember and*** **Tetsuya SHIMAMURA**[†b)], ***Member***

SUMMARY This letter proposes a noise spectrum estimation algorithm for speech enhancement. The algorithm incorporates the speech presence probability, which is calculated from SNR (signal-to-noise ratio) discrepancy. The discrepancy is measured based on the estimation of the *a priori* and *a posteriori* SNR. The proposed algorithm is found to be effective in rapidly switched noise environments. This is confirmed by the experimental results which indicate that the proposed algorithm when integrated in a speech enhancement scheme performs better than conventional noise estimation algorithms.

key words: *speech enhancement, noise estimation, SNR discrepancy, speech presence probability*

1. Introduction

A crucial aspect of the speech enhancement algorithms is estimation of the noise spectrum. The noise spectrum estimation can have a major impact on the quality of the signal processed by the enhancement algorithms. More specifically, underestimation of the noise spectrum causes annoying musical noise, whereas overestimation causes speech distortion, which may in turn impair the intelligibility.

Several methods have been proposed for noise estimation in the last few decades. The most widely used approach in noise estimation is the minimum statistics (MS) algorithm [1] that tracks the minimum values of the noisy speech through a finite window. Depending on the length of this window, the MS algorithm becomes robust to speech onsets, but the finite window length causes the estimation delay that is the major drawback of this algorithm. An alternative approach, which tracks the spectral minima continuously without requiring a finite window (referred to here as CSMT), was also proposed in [2]. Although this approach searches the minimum values of the noisy speech continuously, it cannot distinguish between a rise in noise power and a rise in speech power.

Recently, Cohen *et. al.* [3] proposed a method, known as minima controlled recursive averaging (MCRA), that combines the MS algorithm with the the control of the time constant for the noise update. An improved version of MCRA, known as IMCRA, was also proposed in [4]. The IMCRA algorithm was also based on spectral minima, obtained using the MS algorithm. However, the main problem of these two algorithms is that they cannot avoid the latency of the MS owing to the utilization of the finite window. On the other hand, for highly nonstationary noise, a variant of the MCRA algorithm (referred to here as HNN) was also proposed [5]. Although this method outperforms the IMCRA, it may not be fast enough to track noise spectrum when there is a sudden change in the noise level.

The objective of this letter is to propose a noise estimation algorithm for speech enhancement so as to take care the aforementioned drawbacks. The algorithm incorporates the speech presence probability (SPP), which employs a discrepancy measure between the estimation of the *a priori* SNR (signal-to-noise ratio) and *a posteriori* SNR.

The organization of the letter is as follows. Section 2 describes the proposed method with its principle and implementation. Section 3 shows the experimental results, whereas Sect. 4 concludes the letter.

2. Proposed Noise Estimation Method

In this section, the proposed noise estimation method is described on the basis of the principle and implementation.

2.1 Principle of the Proposed Method

Let $y(n) = s(n) + d(n)$ denote the noisy speech in the time domain, where $s(n)$ is the clean speech and $d(n)$ is the additive noise. Assuming the additive noise is uncorrelated with clean speech, the spectral component of the noisy speech in the frequency domain is given by

$$Y(\lambda, k) = S(\lambda, k) + D(\lambda, k) \tag{1}$$

where $Y(\lambda, k)$, $S(\lambda, k)$ and $D(\lambda, k)$ denote the discrete Fourier transform coefficients of the noisy speech, clean speech and noise respectively, for the k-th frequency bin at frame λ.

The proposed noise estimation method is formulated based on the principle of a detection theory framework. Generally, the clean speech and the noise are assumed to be present in the noisy speech. In reality, however, the clean speech contains many pauses while the noise may be continuously present. The noisy speech can thus be described as a detection problem using two possible hypotheses; one that indicates the speech absence is as $H_0^k : Y(\lambda, k) = D(\lambda, k)$, and another that indicates the speech presence is as $H_1^k : Y(\lambda, k) = S(\lambda, k) + D(\lambda, k)$.

Let $\sigma_d^2(\lambda, k) = E[|D(\lambda, k)|^2]$ denote the variance of the

Manuscript received July 22, 2013.

Manuscript revised October 23, 2013.

[†]The authors are with the Graduate School of Science and Engineering, Saitama University, Saitama-shi, 338–8570 Japan.

a) E-mail: saha@sie.ics.saitama-u.ac.jp

b) E-mail: shima@sie.ics.saitama-u.ac.jp

DOI: 10.1587/transinf.E97.D.373

noise, where $E[.]$ is an expectation operator. The estimate of the noise spectrum is then given by

$$\begin{aligned}\hat{\sigma}_d^2(\lambda,k) &= E\left[\sigma_d^2(\lambda,k)|Y(\lambda,k)\right]\\ &= E\left[\sigma_d^2(\lambda,k)|H_0^k\right]P(H_0^k|Y(\lambda,k))\\ &\quad +E\left[\sigma_d^2(\lambda,k)|H_1^k\right]P(H_1^k|Y(\lambda,k)) \qquad (2)\end{aligned}$$

where $P(H_1^k|Y(\lambda,k))$ and $P(H_0^k|Y(\lambda,k))$ denote respectively the conditional probability of speech presence and absence. Applying a recursive smoothing operation during the periods of speech absence based on the method proposed in [6], the estimate of the noise spectrum is obtained from (2) as

$$\begin{aligned}\hat{\sigma}_d^2(\lambda,k) &= \left[\alpha\hat{\sigma}_d^2(\lambda-1,k)+(1-\alpha)|Y(\lambda,k)|^2\right](1-p(\lambda,k))\\ &\quad +\hat{\sigma}_d^2(\lambda-1,k)p(\lambda,k)\\ &= \alpha_d(\lambda,k)\hat{\sigma}_d^2(\lambda-1,k)+[1-\alpha_d(\lambda,k)]|Y(\lambda,k)|^2 \quad (3)\end{aligned}$$

where $p(\lambda,k) \triangleq P(H_1^k|Y(\lambda,k))$ is the probability of speech presence and $\alpha_d(\lambda,k)$ is defined as

$$\alpha_d(\lambda,k) = \alpha + (1-\alpha)p(\lambda,k) \qquad (4)$$

where α is a constant. The preceding equation of the noise spectrum estimation stated in (3) is the generalized form of the three algorithms, which are MCRA, IMCRA, and HNN. The main difference of the three algorithms is to use different methods to compute $p(\lambda,k)$ needed in (4) to estimate the smoothing factor $\alpha_d(\lambda,k)$. In MCRA and HNN, $p(\lambda,k)$ is calculated using the following recursion:

$$p(\lambda,k) = \alpha_p p(\lambda-1,k) + (1-\alpha_p)I(\lambda,k) \qquad (5)$$

where α_p is a smoothing constant, and $I(\lambda,k)$ is an indicator function.

Note that the MCRA algorithm utilizes the MS approach to calculate $I(\lambda,k)$, whereas the HNN algorithm utilizes the CSMT approach to calculate $I(\lambda,k)$. In both the methods, however, $I(\lambda,k)$ is decided as binary (either 1 or 0) based on the speech activity, and thereby $p(\lambda,k)$ in (5) depends on the previous frame $p(\lambda-1,k)$. Hence $p(\lambda,k)$ will not respond fast enough to abruptly changes of the noise. Moreover, in MCRA, the values of $p(\lambda,k)$ are for the most part binary despite the recursion in (5). This is because $p(\lambda,k)$ is based on spectral minima (obtained using the MS algorithm), which may remain constant within a finite window. As a result, $\alpha_d(\lambda,k)$ may take binary values, either $\alpha_d(\lambda,k) = \alpha$ or $\alpha_d(\lambda,k) = 1$, and the estimated noise spectrum will follow the spectral minima, similar to the MS algorithm. The similar problem may also arise in IMCRA algorithm, since the spectral minima obtained using the MS approach are used to compute $p(\lambda,k)$. These problems inspire us to propose a noise spectrum estimation technique so as to incorporate the SPP $p(\lambda,k)$ in (4). Note that the incorporation of $p(\lambda,k)$ in (4) eliminates the dependency of $p(\lambda,k)$ on the previous frame $p(\lambda-1,k)$, since it reduces the computation of the step stated in (5). The SPP $p(\lambda,k)$ computed from the method proposed in [7] is based on the principle that $p(\lambda,k)$ achieves probabilities close to zero for speech absence and close to one for speech presence, and as a result the smoothing parameter $\alpha_d(\lambda,k)$ will be updated continuously to abruptly changes of the noise power.

2.2 Implementation of the Proposed Method

The SPP, which requires the computation of the *a priori* speech absence probability (SAP), provides an estimate of the probability of speech being present at particular frequency bins. The generalized form of the SPP is given by

$$p(\lambda,k) = \frac{1-\rho(\lambda,k)}{1-\rho(\lambda,k)+\rho(\lambda,k)(1+\xi(\lambda,k))e^{-\nu(\lambda,k)}} \qquad (6)$$

where $\rho(\lambda,k)$ is the *a priori* SAP, $\xi(\lambda,k)$ is the *a priori* SNR, and $\nu(\lambda,k) = \xi(\lambda,k)\gamma(\lambda,k)/(1+\xi(\lambda,k))$ in which $\gamma(\lambda,k)$ is called the *a posteriori* SNR.

The SPP $p(\lambda,k)$ in (6) is based on computing the *a priori* SAP $\rho(\lambda,k)$. The motivation behind the computation of $\rho(\lambda,k)$ is to get a higher estimated value in the noise-dominant regions, whereas the estimated value should be lower in the speech-dominant regions. This is achieved by the following binary decision rule:

$$\rho(\lambda,k) = \begin{cases} 1-\beta(\lambda,k)\zeta(\lambda,k) & \text{under } H_0^k\\ \delta(\lambda,k) & \text{under } H_1^k\end{cases} \qquad (7)$$

where $\zeta(\lambda,k)$ is called the SNR discrepancy, $\beta(\lambda,k)$ is the subtraction factor, and $\delta(\lambda,k)$ is determined from the SNR discrepancy measure $\zeta(\lambda,k)$. The classification of H_0^k and H_1^k is done by the following comparison:

$$P_s(\lambda,k) \underset{H_1^k}{\overset{H_0^k}{\lesseqgtr}} \sigma \qquad (8)$$

where σ is a threshold, and $P_s(\lambda,k)$, which can be recognized as a SNR, denotes the ratio of noisy speech power spectrum to its local minimum, that is,

$$P_s(\lambda,k) = \frac{|Y(\lambda,k)|^2}{P_{min}(\lambda,k)} \qquad (9)$$

where $P_{min}(\lambda,k)$, which corresponds to the minimum of the noisy speech periodogram, is calculated by the method in [2] as

```
if $P_{min}(\lambda-1,k) < |Y(\lambda,k)|^2$
    $P_{min}(\lambda,k) = \gamma P_{min}(\lambda-1,k) + \frac{1-\gamma}{1-\beta}(|Y(\lambda,k)|^2$
    $-\beta|Y(\lambda-1,k)|^2)$        (10)
else
    $P_{min}(\lambda,k) = |Y(\lambda,k)|^2$        (11)
end
```

where γ and β are constants. Note that a different rule has been used to calculate $P_s(\lambda,k)$ in [7]. In this letter, however, we use a nonlinear minimum tracking algorithm because the parameter σ does not appear to be a sensitive parameter.

In (7), the SNR discrepancy measure $\zeta(\lambda, k)$, which is related to the estimation of the *a posteriori* and *a priori SNR*, is defined as

$$\zeta(\lambda, k) = \min\left(\left|\frac{|\gamma(\lambda, k)| - |\xi(\lambda, k)|}{|\gamma(\lambda, k)|}\right|, 1\right) \tag{12}$$

where $\xi(\lambda, k)$ is the *a priori* SNR that is calculated by the *decision-directed* approach [8], and $\gamma(\lambda, k) = |Y(\lambda, k)|^2/\hat{\sigma}_d^2(\lambda - 1, k)$. The preceding discrepancy measure $\zeta(\lambda, k)$ is assessed in a scheme that is related to the background noise. A large discrepancy value is obtained during the situations where speech is absent, whereas a small value is obtained in the situations where speech is present. The discrepancy measure is therefore interpreted as being directly proportional to the estimated noise power level. This provides the concept of the derivative $(1 - \zeta(\lambda, k))$ in (7) for deriving the *a priori* SAP.

The subtraction factor $\beta(\lambda, k)$ in (7) is determined by the following recursive equation:

$$\beta(\lambda, k) = \alpha_\beta \beta(\lambda - 1, k) + (1 - \alpha_\beta)\hat{\beta}(\lambda, k) \tag{13}$$

where $\alpha_\beta \in [0, 1)$ is a smoothing factor and $\hat{\beta}(\lambda, k)$ is calculated as the ratio of the *a priori* SNR to the *a posteriori* SNR, that is, $\hat{\beta}(\lambda, k) = \min(|\xi(\lambda, k|)/|\gamma(\lambda, k)|, 1)$. The preceding subtraction factor $\beta(\lambda, k)$ yields a small value during the noise-only regions, and thus, a large value of $\rho(\lambda, k)$ is obtained. This confirms a low SPP during the situations where speech is not present.

On the other hand, the term $\delta(\lambda, k)$ is calculated as

```
**if** $\delta(\lambda - 1, k) \leq 1 - \zeta(\lambda, k)$
    $\delta(\lambda, k) = \alpha_\delta \delta(\lambda - 1, k) + (1 - \alpha_\delta)(1 - \zeta(\lambda, k))$    (14)
**else**
    $\delta(\lambda, k) = 1 - \zeta(\lambda, k)$    (15)
**end**
```

where $\alpha_\delta \in [0, 1)$ is a smoothing constant. The weak speech components are preserved in the preceding conditional form (15) by allowing $\delta(\lambda, k)$ to adapt slowly according to the smoothing constant α_δ.

3. Experimental Results

The experimental results include two different noise types, namely single noise and multiple noise. In the single noise case, sentences are degraded at 5 dB SNR by either babble noise or train noise. In the multiple noise case, two noisy sets of stimuli are used. The first noisy set consisted of a sentence degraded by babble noise at 15 dB SNR followed by the same sentence degraded by train noise at 0 dB SNR, whereas the second noisy set consisted of a sentence degraded by train noise at 0 dB SNR followed by the same sentence degraded by babble noise at 15 dB SNR. All sentences (30 sentences) used in the experiments are taken from the NOIZEUS [9] speech corpus. Half of them are from male speaker and half of them are from female speaker.

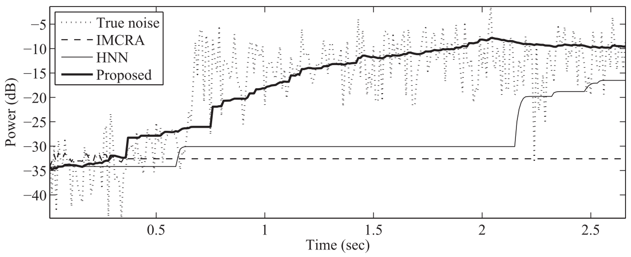

Fig. 1 Tracking performance (for $f = 1050$ Hz) of noise estimators for suddenly increasing noise level.

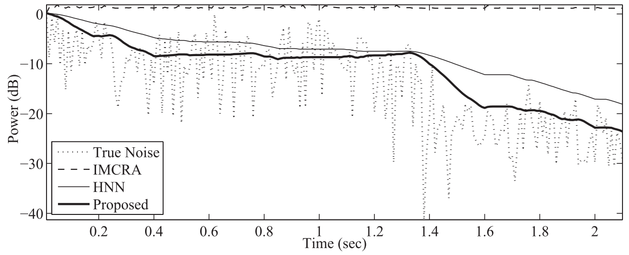

Fig. 2 Tracking performance (for $f = 850$ Hz) of noise estimators for suddenly decreasing noise level.

The sentences used in the experiments are sampled at 8 KHz. A 20-msec analysis Hamming window is used with 50% overlap between frames. The performnace of the proposed method is compared with that of the methods in IMCRA and HNN. The typical parameter selection is the same as that in those approaches. The following values are used in our implementation: $\alpha = 0.85, \sigma = 5, \gamma = 0.998, \beta = 0.8$, $\alpha_\beta = 0.20$, and $\alpha_\delta = 0.80$. The above parameters are optimized based on the NOIZEUS speech corpus. A large number of simulations using a large number of data are thus conducted for optimizing the above parameters. Note that the value of α, β and γ in our implementation is the same as in [5]. The evaluation is conducted with regard to the following aspects: tracking performance of the noise estimators, and improvement in the speech quality by integrating the noise estimators into a speech enhancement algorithm.

3.1 Tracking Performance

Firstly, we evaluate the tracking performance of all noise estimators for multiple noise case with Figs. 1 and 2. Figure 1 shows the sudden increase of noise power level in which babble noise ($t < 0.6$ sec) at 15 dB SNR is added to a sentence followed by the train noise ($t > 0.6$ sec) at 0 dB SNR. Figure 2, on the other hand, shows the sudden decrease of noise power level in which initially train noise ($t < 1.4$ sec) at 0 dB SNR is added to a sentence followed by the babble noise ($t > 1.4$ sec) at 15 dB SNR. For comparative purposes, the true noise spectrum is also superimposed in both figures.

As mentioned earlier, the computation of the SPP $p(\lambda, k)$ in IMCRA depends on the spectral minima, which is obtained by tracking the minimum of noisy speech over a search window spanning L frames. Unfortunately, this has

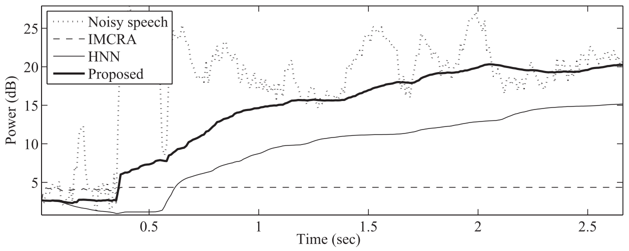

Fig. 3 Tracking performance (for whole frequency components) of noise estimators for suddenly increasing noise level.

few drawbacks. First, the minimum is sensitive to outliers. Second, the update of minimum can take at most 2L frames. The update of noise estimation in IMCRA is therefore restricted to the minimum tracking, which may lag by as many as 2L frames. Hence, the IMCRA does not result in any significant update of the noise estimation. This can be seen in Figs. 1 and 2.

The update of the noise estimation in HNN is based on a rough detection of speech region. Depending on this detection, the indicator function $I(\lambda, k)$ stated in (5) is considered as binary (either 1 or 0). The SPP $p(\lambda, k)$ is thus updated according to the previous analysis frame. This makes the problem when there is a sudden change in the noise level. As is evident from Figs. 1 and 2, where the noise power is suddenly increasing and decreasing respectively, the update of the noise estimation cannot follow the noise power. This is caused by the previous analysis frame which is already analyzed as a frame of speech being present, and therefore the noise estimation is not updated.

On the other hand, the SPP $p(\lambda, k)$ in the proposed method is updated based on the computation of the SAP $\rho(\lambda, k)$ in each frequency bin. This makes the proposed method to be much faster than that of the other algorithms. As can be seen from Fig. 1, at $t \approx 0.6$ sec, the noise power level suddenly increases. It takes about 1.5 sec for the HNN algorithm to track the rising noise level, whereas it only requires roughly 0.1 sec for the proposed method to track the noise. Note that Figs. 1 and 2 use a single frequency component. We therefore further evaluate the results which use the total power of noisy speech across whole frequency components. The same phenomenon is observed that is shown in Fig. 3.

3.2 Speech Quality Performance

Secondly, in order to evaluate the speech quality, the noise estimators are integrated into a power spectral subtraction method proposed in [10]. The spectral floor used for power spectral subtraction is 0.02. Two objective measures are used to evaluate the performance of the noise estimation algorithms: segmental SNR (segSNR) and weighted spectral slope (WSS) measure [11].

The segSNR and WSS results are shown in Table 1. Relatively larger segSNR values are obtained by our proposed method. Smaller WSS values are also obtained by our proposed method in both noise cases. Although the results obtained by the proposed method are found better in both noise cases, a paired t-test confirms that the proposed method performs significantly ($p < 0.05$) better than the other algorithms in the multiple noise case. The most statistical significant ($p = 0.007$) results are obtained by our proposed method in comparison with the IMCRA approach. These observations are also found to be consistent with the informal subjective listening test, where the listeners overwhelmingly have reported that the proposed method, particularly for the multiple noise, produces less musical noise than that the conventional methods do.

Table 1 Results of the noise estimation algorithms in terms of segSNR and WSS values.

Method	Single noise		Multiple noise	
	SegSNR	WSS	SegSNR	WSS
IMCRA	−0.28	68.45	−2.05	81.93
HNN	0.11	66.69	−1.25	78.74
Proposed	0.43	57.85	−0.73	61.45

4. Conclusion

In this letter, we have proposed a noise spectrum estimation algorithm for speech enhancement. Unlike other algorithms, the proposed algorithm is updated based on the SPP that provides an estimate of the probability of speech being present at particular frequency bins. The update of the noise spectrum estimation in the proposed method is therefore much faster, particularly for highly nonstationary noise environments, than that of the other algorithms. This is confirmed by the experimental results which indicate that the proposed noise estimation algorithm when integrated in a noise reduction scheme performs well over the conventional noise estimation algorithms.

References

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process., vol.9, no.5, pp.504–512, 2001.

[2] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," Proc. 4th European Conf. Speech Commun. Technology, pp.1513–1516, 1995.

[3] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," IEEE Signal Process. Lett., vol.9, no.1, pp.12–15, 2002.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," IEEE Trans. Speech Audio Process., vol.11, no.5, pp.466–475, 2003.

[5] S. Rangachari and P. Loizou, "A noise estimation algorithm for highly nonstationary environments," Speech Commun., vol.48, pp.220–231, 2006.

[6] 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.

[7] A. Saha and T. Shimamura, "Speech enhancement by incorporating speech presence probability based on SNR discrepancy," J. Signal Process., vol.16, no.1, pp.57–65, 2012.

[8] 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.ASSP-32, no.6, pp.1109–1121, 1984.

[9] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio Speech Language Process., vol.16, no.1, pp.229–238, 2008.

[10] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process., pp.208–211, 1979.

[11] J.R. Deller Jr., J.G. Proakis, and J.H.L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan, New York, 1993.