Enhanced Minimum Statistics Technique Incorporating Soft Decision For Noise Suppression

Yun-Sik Park, Ji-Hyun Song, Jae-Hun Choi, Joon-Hyuk Chang

School of Electronic Engineering
Inha University
Incheon, Korea

{yspark,jhsong,jchoi}@dsp.inha.ac.kr, changjh@inha.ac.kr

Abstract

In this paper, we propose a novel approach to noise power estimation for robust noise suppression in noisy environments. From investigation of the state-of-the-art techniques for noise power estimation, it is discovered that the previously known methods are accurate mostly either during speech absence or speech presence but none of it works well in both situations. Our approach combines minimum statistics (MS) and soft decision (SD) techniques based on probability of speech absence. The performance of the proposed approach is evaluated by a quantitative comparison method and subjective test under various noise environments and found to yield better results compared with conventional MS and SD-based schemes.

Index Terms: Minimum Statistics, Soft Decision

1. Introduction

In practice, most current speech coding systems are confronted with various background noises that degrade their performance. The easiest way to alleviate speech quality degradation is to employ a noise suppression technique in which the background noise is reduced before being encoded by the speech coding system. In general, noise suppression techniques consider relevant procedures such as the selection of an adequate statistical model of speech [1], [2], noise suppression gain computation [3]-[6], and noise power estimation [7]-[9]. With the wide dissemination of mobile speech communication, the noise power estimation approach has received attention due to the nonstationary characteristics of the noise signal and the difficulty of estimating the background noise during speech activity.

One of the successful noise estimation techniques is minimum statistics (MS) [7], [10], which obtains the noise estimate as the minima values of a smoothed power estimate of the noisy signal. The MS method was originally motivated by the observation that the speech and the disturbing noise are usually statistically independent and the power of a noisy speech signal frequently reduces to the power level of the noise signal.

Soft decision (SD), the well-known noise power estimation technique, can be applied as a preprocessing module of speech processing systems [2], [8], [11], [12]. In the SD technique, the long-term smoothed power spectrum of the background noise depending on the probability of speech absence is adopted. The speech absence probability (SAP) is derived from a likelihood ratio test (LRT) by employing the decision-directed (DD) method for the estimation of the unknown parameters.

In this paper, we propose the probabilistically combined noise power estimation (PCNPE) approach, a novel noise power estimation technique that simultaneously applies both the MS and SD methods. The PCNPE combines the noise power estimates provided by MS and SD modules depending on the SAP, which is a byproduct of a common noise suppression technique [8]. The performance of the proposed algorithm is evaluated by an objective method and subjective quality test and is demonstrated to be better than that of the conventional MS and SD methods.

2. Minimum statistics and soft decision

In this section, we briefly review the MS and SD techniques used for noise power estimation in noise suppression. Interested readers are referred to [8], [10], and [11] for details of the two approaches.

A. Minimum Statistics

First, let \( x \) and \( d \) denote the speech and the noise signal, respectively. Applying a discrete Fourier transform to the noisy speech signal \( y \), we have in the time-frequency domain

\[
Y(n, k) = X(n, k) + D(n, k)
\]

where \( n \) is the time frame index and \( k \) is the frequency-bin index. The MS approach was originally motivated by the minimum signal power estimate during speech pauses, since the signal power is instantaneously reduced to the noise power level during nonspeech or within brief periods in words and syllables. For that reason, it is possible to obtain an accurate noise power estimate by tracking the minimum power of noisy speech, where the bias is usually be compensated [10]. To search the minimum of the local energy, the following recursively smoothed periodogram \( P(n, k) \) is considered:

\[
P(n, k) = \alpha P(n - 1, k) + (1 - \alpha)|Y(n, k)|^2
\]

where \( P(n, k) \) is obtained using the smoothing parameter \( \alpha \). Since the fixed smoothing parameter \( \alpha \) is generally chosen to be very close to 1, this widens the peaks of speech activity of the smoothed power spectrum estimate. As a result, the estimated minimum noise power, which are deduced from these smoothed periodograms, can be misleading as inaccurate noise power estimates. To overcome these drawbacks, a time varying optimal smoothing parameter \( \alpha_{opt} \) is given as follows:

\[
\alpha_{opt} = \frac{1}{1 + (P(n - 1, k)/\lambda_d(n, k) - 1)^2}
\]

where \( \lambda_d(n, k) = E[|D(n, k)|^2] \) is the noise power spectral density and \( E[\cdot] \) denotes the expectation operator. Actually,
the estimated noise power based the MS approach is obtained by searching for minima within a finite-window length $L_{MS}$ of the smoothed periodogram $P(n, k)$ as follow:

$$\lambda_{\text{min}, d}(n, k) = \min \left[ P(n, k), P(n-1, k), \ldots, P(n-L_{MS}, k) \right]$$

(4)

where $\min[\cdot]$ is a minimum value operator. Since the minimum power estimate obtained through the time-varying smoothing parameter is smaller than the mean value, the MS approach requires a bias compensation for the unbiased noise power estimate as given below [10]:

$$\hat{\lambda}_{d}^{\text{MS}}(n, k) = B_{\text{min}}(n, k) \lambda_{\text{min}, d}(n, k)$$

(5)

where $\hat{\lambda}_{d}^{\text{MS}}(n, k)$ is the unbiased noise power estimate and $B_{\text{min}}(n, k)$ represents the bias compensation factor.

B. Soft Decision

Given two hypotheses, $H_0$ and $H_1$, which respectively indicate speech absence and presence compared with (1), it is assumed that

$$H_0 : \text{speech absent : } Y(n, k) = D(n, k)$$

$$H_1 : \text{speech present : } Y(n, k) = X(n, k) + D(n, k).$$

(6)

Under the assumption that the disturbing noise $D(n, k)$ and the speech $X(n, k)$ are statistically independent, $X(n, k)$ and $D(n, k)$ are characterized by zero-mean complex Gaussian distributions such that [4]:

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

(7)

$$p(Y(n, k)|H_1) = \frac{1}{\pi(\lambda_s(n, k) + \lambda_d(n, k))} \exp\left[-\frac{|Y(n, k)|^2}{\lambda_s(n, k) + \lambda_d(n, k)}\right]$$

(8)

where $\lambda_s(n, k)$ and $\lambda_d(n, k)$ are the clean speech and noise power in the $k$th frequency bin, respectively. The speech absence probability (SAP) $p(H_0|Y(n, k))$ for each frequency band is derived from Bayes’ rule such that [6]:

$$p(H_0|Y(n, k)) = \frac{p(Y(n, k)|H_0)p(H_0)}{p(Y(n, k)|H_0)p(H_0) + p(Y(n, k)|H_1)p(H_1)}$$

$$= \frac{1}{1 + q \Lambda(Y(n, k))}$$

(9)

where $q = p(H_1)/p(H_0)$ and $p(H_0)$ represents the a priori probability of speech absence. Substituting (7) and (8) into (9), the likelihood ratio $\Lambda(Y(n, k))$ can be computed as follows [12]:

$$\Lambda(Y(n, k)) = \frac{p(Y(n, k)|H_1)}{p(Y(n, k)|H_0)}$$

$$= \frac{1}{1 + \xi(n, k)} \exp\left[\frac{\gamma(n, k) \xi(n, k)}{1 + \xi(n, k)}\right]$$

(10)

where the a posteriori signal-to-noise ratio (SNR) $\gamma(n, k)$ and the a priori SNR $\xi(n, k)$ are defined by [4]

$$\gamma(n, k) \equiv \frac{|Y(n, k)|^2}{\lambda_d(n, k)}$$

(11)

$$\xi(n, k) \equiv \frac{\lambda_s(n, k)}{\lambda_d(n, k)}$$

(12)

Actually, the SD method adopts a long-term smoothed noise power estimate $\hat{\lambda}_{d}^{\text{SD}}(n, k)$, which is given by [11]:

$$\hat{\lambda}_{d}^{\text{SD}}(n+1, k) = \xi_0 \hat{\lambda}_{d}^{\text{SD}}(n, k) + (1 - \xi_0) E[D(n, k)^2|Y(n, k)]$$

(13)

where the smoothing parameter $0 < \xi_0 < 1$ is set to 0.99. The SAP in (9) is applied to derive $E[D(n, k)^2|Y(n, k)]$, as specified in [2] and given by

$$E[D(n, k)^2|Y(n, k), H_0] = |Y(n, k)|^2$$

(14)

$$E[D(n, k)^2|Y(n, k), H_1] = \left(\frac{\xi(n, k)}{1 + \xi(n, k)}\right) \hat{\lambda}_{d}^{\text{SD}}(n, k) + \left(\frac{1}{1 + \xi(n, k)}\right)^2 |Y(n, k)|^2.$$ 

(16)

3. Probabilistically combined noise power estimation approach

Thus far, we have focused on two major approaches of noise power estimation suitable for a real-time noise update. A major drawback of both these approaches is that they cannot fully take all conditions of speech activity into consideration. This means that, during speech activity, the SD technique does not accurately estimate the noise power spectra, since the long-term smoothed estimation by (13) for short non-speech intervals is considered. In Fig. 1, we give an example of the estimated noise power contours by the SD and MS methods in conjunction with the input clean speech waveform and noisy speech waveform. Specifically, from Fig. 1(a), we can see that the long-term smoothed estimation of the SD scheme is not adequate for speech active periods, because the recursive long-term averaging should stop updating and maintain the same values of the corresponding subbands in the last speech inactive frame. But, the MS approach takes into account the very instantaneous silence between words and syllables and finds the minimal value within a finite time-window. For this reason, the background noise, which exists in speech periods, can be more effectively updated by the MS technique as displayed in Fig. 1(a).

On the contrary, when there is no speech ($p(Y(n, k)|H_0) = 1$) the SD approach continually tracks the noise power spectra $|D(n, k)|^2$ of noise using (13) and (15) since $Y(n, k)$ directly corresponds to $D(n, k)$ during nonspeech periods as specified in (6). Again, from Fig. 1(a), we can see that the noise power spectra estimated by the SD scheme along the time axis are more close to the actual noise spectra compared to the MS method during speech pauses. The disadvantage of the MS scheme during non-speech periods is attributable to the fact that the MS approach estimates the noise spectra through the non-linear operation such as min $[\cdot]$ which accounts for a shape of the sudden rising and falling noise contour stemming from picking the minimal value within a sliding window.

Therefore, in order to attain high estimation accuracy with the SD method during nonspeech periods while maintaining the good performance of the MS approach in active speech regions,
we propose a new noise power estimation technique, the probabilistically combined noise estimation (PCNPE) algorithm, which combines the MS estimate with the SD estimate depending on the SAP computed in the SD scheme. Let $\hat{\lambda}^\text{MS}$ be the noise estimate provided by the MS algorithm, and $\hat{\lambda}^\text{SD}$ the noise estimate obtained based on the SD technique. The PCNPE method then combines the two estimates as follows:

$$\hat{\lambda}^\text{PCNPE}(n, k) = (1 - \rho(H_0|Y(n, k)))\hat{\lambda}^\text{MS}(n, k) + \rho(H_0|Y(n, k))\hat{\lambda}^\text{SD}(n, k).$$

(17)

The PCNPE is finally addressed as $\hat{\lambda}^\text{PCNPE}$ replaces $\hat{\lambda}^\text{MS}$ within the active speech periods on each frequency bin (i.e., $P(H_0|Y(n, k)) \approx 0$) and $\hat{\lambda}^\text{SD}$ in the case of the speech absence. Also, in the case of transition periods from speech to silence ($0 < P(H_0|Y(n, k)) < 1$), $\hat{\lambda}^\text{PCNPE}$ replaces the smoothed version between $\hat{\lambda}^\text{MS}$ and $\hat{\lambda}^\text{SD}$ as depicted in Fig. 1d. As an application of the presented technique, we consider a noise suppression algorithm based on a minimum mean-square error (MMSE) as follows [4]:

$$\hat{X}(n, k) = G(\hat{\xi}(n, k), \hat{\gamma}(n, k))Y(n, k)$$

(18)

where $\hat{X}(n, k)$ is the estimated clean speech spectrum and $G(\cdot)$ denotes the noise suppression gain. Specifically, the MMSE estimator-based gain is represented by

$$G(\hat{\xi}(n, k), \hat{\gamma}(n, k)) = \frac{\sqrt{\nu(n, k)} \exp\left(-\frac{\nu(n, k)}{2}\right)}{2\hat{\gamma}(n, k)}$$

(19)

$$= \left[1 + \frac{\nu(n, k)}{2}\right]I_0\left(\frac{\nu(n, k)}{2}\right) + \nu(n, k)I_1\left(\frac{\nu(n, k)}{2}\right)$$

4. Experimental results

The proposed noise power estimation technique was adopted for a noise suppression application and was evaluated with a quantitative comparison and subjective quality test experiment under various noise conditions. Ten test phrases, where five were spoken by a male speaker and the other were spoken by a female, were used as the experimental data. Each phrase consisted of two different meaningful sentences and lasted 8 sec. The noise power estimation was performed for each frame of 10 ms duration with a sampling rate of 8 kHz. Three types of noise sources, white, babble, and F16 cockpit noises from the NOISEX-92 database were added to the clean speech waveform at SNRs of 5, 10 and 15 dB.

At first, noise estimation accuracy was evaluated frame based on the normalized relative estimation error in various background noise environments. For this, the relative estimation error $\varepsilon_n$ is defined by [13]

$$\varepsilon_n = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{k} \left[\hat{\lambda}_n(n, k) - \lambda_d(n, k)\right]^2$$

(23)

where $\lambda_d(n, k)$ is the actual noise power estimate obtained by the noise waveform [13], $\hat{\lambda}_n(n, k)$ is the noise power estimated by the tested method and $N$ is the number of frames in the analyzed signal. For the MS, we see that a suitable window is typically 0.6-1.4 see through our experiments with different speakers and noise conditions [7], [13]. Accordingly, the finite-window length $L_{\text{MS}}$ was set to 100 (i.e., 1 sec). Also, $\alpha_{\text{opt}}$ was limited to a maximum value $\alpha_{\text{max}}^{\text{opt}}$, e.g., $\alpha_{\text{max}}^{\text{opt}} = 0.96$ as specified in [10]. Table I presents the results of the relative estimation error for the evaluated noise estimation methods under the
given noise conditions. The proposed PCNPE scheme achieves a consistently higher improvement in the relative estimation error over the MS and SD approaches when the babble and F16 noises were used. In the case of the white noise, the proposed method outperformed the other approaches except for low SNR (= 5 dB). This exception is attributed to the imperfection of the SAP at the adverse white noise condition.

Secondly in order to evaluate the subjective quality of the proposed scheme, we carried out a set of informal listening tests. Opinion scores were recorded by twenty listeners and averaged to yield final mean opinion score (MOS) results. Twenty listeners (16 male and 4 female) whose ages ranged from 20 to 32 participated in the experiment. Eight of them were students specialized in speech processing while the others were non-specialists. All the scores from the listeners were then averaged to yield the average test results. The test results are summarized in Table II in which a higher value is preferred. In addition, results of the corresponding hypothesis test against references (MS and SD) are classified into three categories: (1) better than (B), (2) not worse than (NW) and worse than (W) [14]. Table II illustrates that the proposed PCNPE approach outperformed or at least was comparable to the conventional MS and SD methods under the given noise conditions as the hypothesis test confirms again. Performance improvement was found greater for the babble noise case at all SNRs. These results confirm that the combined use of the MS and the SD approaches is an effective approach for noise suppression.

5. Conclusions

In this paper, we have analyzed the conventional MS and SD approaches applied to noise power estimated for noise suppression in various noise environments. Based on the analysis, we have presented the PCNPE algorithm, a novel noise power estimation algorithm that combines the MS estimate with the SD estimate in accordance with the SAP used in the SD approach. On the basis of the relative estimation error and a number of MOS evaluation tests, the performance of the proposed algorithm was found to be superior to that of the conventional MS and SD approaches, not only in active speech periods but also during non-speech periods.

6. Acknowledgements

This work was partly supported by ETRI SoC Industry Promotion Center and This work was partly supported by the IT R&D program of MKE/ITTA [2008-F-045-01].

7. References


