Robust Noise Estimation Using Minimum Correction with Harmonicity Control

Xuejing Sun, Kuan-Chieh Yen, and Rogerio Alves

Cambridge Silicon Radio
Auburn Hills, MI, USA
{xuejing.sun,kuan.yen,rogerio.alves}@csr.com

Abstract
In this paper a new noise spectrum estimation algorithm is described for single-channel acoustic noise suppression systems. To achieve fast convergence during abrupt change of noise floor, the proposed algorithm uses a minimum correction module to adjust an adaptive noise estimator. The minimum search duration is controlled by a harmonicity module for improved noise tracking under continuous voicing condition. Objective test results show that the proposed algorithm consistently outperforms competitive noise estimation methods.

Index Terms: speech enhancement, noise estimation, minimum correction, harmonicity

1. Introduction
In most speech enhancement algorithms, accurately tracking the noise spectrum is the pre-requisite of the subsequent operations and has a major impact on the overall quality of the system. Unfortunately, obtaining reliable estimates has shown to be highly difficult in practice due to diverse environments. Furthermore, battery powered devices, such as Bluetooth headsets, require highly efficient algorithms with a small memory footprint, which precludes the use of many sophisticated algorithms.

Numerous noise spectrum estimation algorithms have been proposed in the past [1]-[7], among which Minimum Statistics (MS) based methods have received considerable interest due to their capability to track non-stationary noise and relatively low complexity [1]. MS and related algorithms normally track minima on a smoothed spectrum within a finite search window and use the minimum value itself as the noise estimate [1] or to calculate speech presence probability to control recursive noise estimation [2][3]. The minimum-search window duration has a crucial impact on noise estimation. A short window allows a faster response to noise variation but may be prone to false classification of speech as noise when continuous phonation is longer than the window length. On the other hand, a long window results in slow noise adaptation. In practice, the window length is often empirically introduced. The details of minimum correction and harmonicity control modules are described in Sections 3 and 4, respectively. Experimental results are presented in Section 5 and a conclusion is made in Section 6.

2. Framework description
The proposed system employs a DFT filter bank and tracks the power of background noise in each subband. The input signal and the noise power estimate in each subband are denoted by $D_k(l)$ and $P_k(l)$ respectively, where $k$ is the subband index and $l$ is the frame index. As shown in Figure 1, $P_k(l)$ is obtained after the time domain input signal $d(n)$ passes through the analysis filter bank and then the adaptive noise estimation module. In parallel, a Minimum Correction module controls the noise estimation, which itself is controlled by a Harmonicity module.

![Figure 1: Block diagram of the noise spectrum estimation process.](image-url)

The core noise estimation process is realized through a recursive noise estimator, adaptively controlled by Speech Absence Probability (SAP), similar to [2] [3]:

$$P_k(l) = P_k(l-1) + a q_k(l) [D_k(l)^2 - P_k(l-1)],$$  \hspace{1cm} (1)

The parameter $\alpha \ (0 < \alpha < 1)$ is a smoothing coefficient and $q_k(l)$ is the SAP based on a Gaussian statistical model [3]:

$$q_k(l) = \begin{cases} R_k(l) \exp(-R_k(l)), & |D_k(l)|^2 > P_k(l-1) \\ 1, & \text{otherwise} \end{cases}$$  \hspace{1cm} (2)

$R_k(l)$ is the energy ratio between the noisy speech signal and the previously estimated noise, which is defined as:

$$R_k(l) = \frac{\sum_{j=k-l+1}^{j=k} |D_j(l)|^2}{\sum_{j=k-l+1}^{j=k} P_j(l-1)}.$$  \hspace{1cm} (3)
where \( b(k) \) is a predefined neighborhood range value for subband \( k \).

Using a SAP model allows smooth transitions to be tracked but prevents any dramatic power increase, presumably due to the presence of speech, from affecting the noise estimate. Note that \( q_{l}(i) \) is set to 1 when \( |D_{l}(i)|^{2} \) is smaller than \( P_{l}(i-1) \) in order to adapt at full speed. Eq. (3) represents a smoothing process along the frequency axis, which improves the robustness of the SAP model. This, however, may smear the spectral variation of underlying speech signal. Thus, \( b(k) \) typically is set to 1 for low frequencies (e.g. below 2 kHz) where frequency resolution is more critical, and is set to 2 or 3 for frequencies above this threshold.

3. Minimum correction

The SAP controlled recursive noise estimator in Eq. (1) works the best for quasi-stationary noise. When the noise floor rises abruptly, however, the energy-based SAP approaches zero, which results in a very slow re-convergence rate. To overcome this limitation, a novel application of minimum statistics is proposed. In contrast to previous approaches where the minima values are used either directly as the noise estimate [1] or to calculate the SAP as in [2][3], the minima values are used to correct the output of the recursive noise estimator \( P_{l}(i) \) only when it significantly deviates from the local minimum. It is worth noting that compared with the Minima Controlled Recursive Averaging (MCRA) [2][3] algorithm, the proposed minimum correction module is a less intrusive control mechanism, which affects the recursive noise estimate only under certain ‘outlier’ conditions. For typical noisy conditions, the SAP controlled estimator, based on the Gaussian distribution assumption, remains intact.

Since minimum statistics are used indirectly rather than directly as the noise estimates, the requirement for its frequency resolution can be relaxed. Specifically, instead of performing minima tracking in each subband, the frequency bands are combined into several groups and the minimum average power is tracked for each group. Frequency grouping smoothes out unwanted spectral fluctuation and reduces the overall system complexity. For example, with 8 kHz sampling rate, signals are generally split into 64 subbands for noise estimation and gain application. However, these subbands are typically split into two equal groups, i.e. [0, 2] and [2, 4] kHz for minimum tracking. Thus the number of minima need to be tracked is reduced from 64 to 2.

Minimum statistics are used in the following way to correct the noise estimate. Let \( \overline{P}_{g}(i) \) and \( \overline{S}_{g}(i) \) be the average noise power estimate and the average input signal power for group \( g \) at frame \( l \), respectively. Additionally, let \( \overline{S}_{\text{min},g}(l) \) be the minimum value of \( \overline{S}_{g}(i) \) within a finite window \( L \):

\[
\overline{S}_{\text{min},g}(l) = \min \{ \overline{S}_{g}(i) \}, \quad l = 1, 2, ..., L.
\]

To extend the effective search length with minimum complexity, another variable is used to store the minimum value which is updated every \( L \) frames. Finally, a correction factor \( C_{g}(l) \) is derived as

\[
C_{g}(l) = \frac{\overline{S}_{\text{min},g}(l)}{\overline{P}_{g}(l)}.
\]

Then the noise estimate \( P_{g}(i) \) in group \( g \) (denoted by \( P_{k,g}(l) \)) is adjusted based on the correction factor as below:

\[
P_{k,g}(l) = \begin{cases} \frac{P_{g}(i)C_{g}(l)}{C_{\text{min},g}} & C_{\text{g}}(l) \leq C_{\text{min},g} \\ P_{g}(i)C_{g}(l) & C_{\text{g}}(l) > C_{\text{min},g} \end{cases}
\]

where \( C_{\text{min},g} \) and \( C_{\text{max},g} \) denote the predefined thresholds with typical values of 0.17 and 6.0, respectively.

When the estimated noise power is only slightly higher than the minimum power, as indicated by \( C_{\text{min},g} < C_{\text{g}}(l) \leq 1 \), no adjustment to \( P_{k,g}(l) \) is required as this is fully expected. However, if the minimum value is significantly smaller than the noise estimate as in the first condition ( \( C_{\text{g}}(l) \leq C_{\text{min},g} \) ), it may indicate the noise estimate mistakenly adapts to the speech level or there is a sudden drop in the noise floor. Therefore, the noise estimate needs to be adjusted to a smaller value. To avoid over-adjustment caused by outliers in minimum statistics, the adjustment is limited by \( C_{\text{max},g} \) so that the corrected noise estimate is still higher than the minimum value. When the estimated noise power is lower than the corresponding minimum value, i.e. \( 1 < C_{\text{g}}(l) \leq C_{\text{min},g} \), simple upward correction is applied as there might be a sudden rise of noise floor and the noise estimate is lagging behind. Note that when the minimum value is significantly higher than the noise estimate ( \( C_{\text{g}}(l) > C_{\text{min},g} \) ), special treatment is applied as shown in the last condition of Eq. (6): the estimated noise spectrum is reset to a flat spectrum for each group according to the minimum \( \overline{S}_{\text{min},g}(l) \).

This is because multiplication of a single correction factor, as in the third condition, may result in slow re-convergence if there is a substantial spectrum mismatch between old and new noise floors. Resetting to a flat spectrum generally yields quicker convergence and fewer artifacts for highly non-stationary noise.

\[\text{Figure 2: Average noise power estimate for signal with a sudden rise of noise floor: (a) waveform, (b) average noise power estimate with (solid) and without (dotted) minimum correction.}\]

Figure 2 demonstrates the impact of minimum correction on the noise estimate when there is an abrupt and significant jump of noise energy level. Figure 2(a) is the input waveform with the time axis in samples and 2(b) shows the noise estimate with (solid) and without (dotted) minimum correction, and the time axis is in frames where each frame is 16 ms in duration with 50 percent of overlap. The curves are
created by averaging the noise power values across all frequency bins. It can be seen that the two noise estimates match well for the low noise level but diverge after the noise floor rises sharply. The adaptive noise estimator moves upwards slowly but remains well below the true noise level until the end. The minimum-corrected noise estimator adapts to the true noise level in less than 100 frames (i.e., just before the 400th frame).

4. Harmonicity control

As described in the Introduction, defining an optimal window length for minimum statistics is by no means a trivial task. This paper proposes to start with a short fixed window length, and then use speech harmonicity (periodicity) to dynamically lengthen the effective span of the search window. In contrast to using harmonicity directly as a voice activity detector (VAD), the proposed structure uses harmonicity estimation as a secondary procedure to control the search window. This avoids the need of designing a sophisticated voicing detection algorithm or a complex combination logic of several VADs.

There are many ways to determine the harmonicity of a signal. In this case, Average Magnitude Difference Function (AMDF) is used due to its simplicity and reasonable performance. For a short-term signal $x_m[n]$ ending at $m$, AMDF can be defined as

$$\text{AMDF}_m[\tau] = \frac{1}{N} \sum_{n=m-N+1}^{m} |x_m[n] - x_m[n-\tau]| \quad (7)$$

where $\tau$ is the lag value subject to $\tau_{\text{min}} \leq \tau < \tau_{\text{max}}$ and $N$ is number of samples involved for each $\tau$. $N$ typically should be twice as long as the maximum expected pitch period.

A representation of harmonicity based on AMDF can be derived from the ratio of its minimum and maximum:

$$H = 1 - \frac{\min_{\tau} \text{AMDF}_m[\tau]}{\max_{\tau} \text{AMDF}_m[\tau]} \quad (8)$$

With this definition, a value close to 1 indicates voiced speech whereas 0 indicates unvoiced speech or noise. Typically 0.6 is used as the threshold for classification.

Harmonicity offers some unique yet important information previously unavailable to adaptive noise estimation and minimum statistics, which exploit mostly energy variation patterns. Instead of using it to adjust the recursive estimator (Eq. (1)) directly, here the harmonicity is used to control the effective length of the minimum search window. Specifically, when a frame is classified as voiced by the harmonicity function, the corresponding power value $\Sigma_g(i)$ is skipped and not entered into the minimum search queue (Eq. (4)). Consequently when there are prolonged periods of voiced speech without pause, the search queue still has the pre-voicing minimum values representing the true noise floor. This is equivalent to lengthening the effective window span without physically increasing the search length. As a result, the default search duration can be set relatively short for fast noise adaptation, e.g. 0.25 seconds.

It can be seen that the above harmonicity estimation employs the most basic implementation, which reflects the design goal of maintaining low complexity yet achieving the desired impact. The harmonicity module targets only long voiced segments. Unvoiced speech segments, which normally are short, are passed through to the minimum correction queue. The reliability of the harmonicity module in the proposed algorithm is not as crucial as that in the conventional usage. This is due to the fact that the harmonicity module has to make continuous wrong decisions over many frames in order to “flush” the pre-voicing minimum out of the search queue. Furthermore, the harmonicity controller is designed particularly to help noise estimation at high SNR conditions, under which a failure of minimum correction (false classification of speech as noise), due to duration limitation, causes more serious speech degradation perceptually. The reliability of harmonicity typically decreases with the decrease of SNR level. Fortunately, its impact on noise estimation is also naturally reduced under the current scheme.

A straight calculation of harmonicity can be computationally expensive for resource-limited devices such as a Bluetooth headset. However, its efficiency can be dramatically improved with some optimizations. First, instead of using the conventional wide pitch range (e.g. [60, 500] Hz) for time lag $\tau$, we only use one octave range for harmonicity, e.g. [60, 120] Hz. Second, at 8 kHz sampling rate, the input signal $x_m[n]$ can be decimated by a factor of two without causing noticeable performance degradation. Finally, a portion of intermediate results derived from Eq. (7) can be saved and reused for the next frame due to the overlapping frame structure. Details of these optimizations and implementation specifics are beyond the scope of this paper.

![Figure 3: Average noise power estimate for signal with long phonation (a) waveform (b) minima corrected adaptive noise estimate with (solid) and without (dotted) harmonicity control](image)

Figure 3 illustrates the noise estimation results for a long phonation scenario with and without harmonicity control. Similar to Figure 2, average noise power values are plotted. Figure 3(b) shows the corresponding noise estimates using minimum corrected noise estimation (dotted line) and the same estimator with harmonicity control (solid line), respectively. It can be observed that without harmonicity control, the noise estimate jumps significantly higher in the middle of a long speech segment (after about 300 frames) while the harmonicity-controlled noise estimate remains low throughout the whole utterance.

5. Experimental results

To evaluate the effectiveness of the proposed minimum correction and harmonicity control, the ITU standard - Perceptual Evaluation of Speech Quality (PESQ) [8] is used as an objective measure. The test materials are from NOIZEUS [9], which contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different
types of real-world noise (babble, car, exhibition hall, restaurant, street, airport, train station, and train) at four different SNRs (15dB, 10dB, 5dB, and 0dB). For each SNR condition, there are 240 utterances sampled at 8 kHz. To compare the proposed algorithm to other approaches, the Matlab implementation of several popular noise estimation algorithms provided in [10] is used without any modifications. Its modular structure allows the noise estimator to be replaced and everything else is kept equal for a fair comparison. The proposed system employs a 128-point square-root Hanning window with 50 percent overlap. All the adjustable system parameter values are kept the same for all testing conditions. For different noise estimation algorithms, a plain Wiener filter function is used to generate the instantaneous gain values \((G_{\alpha}(l))\), which are subsequently smoothed to reduce musical noise [11]:

\[
G_{\alpha}(l) = G_{\alpha}(l-1) + \alpha_{\alpha}(k)(G_{\alpha}(l) - G_{\alpha}(l-1)),
\]

\[
G_{\alpha}(l) = G_{\alpha}(l-1) + \alpha_{\alpha}(k)(G_{\alpha}(l) - G_{\alpha}(l-1)),
\]

where \(\alpha_{\alpha}(k)\) and \(\alpha_{\alpha}(k)\) are constants between 0 and 1, and \(G_{\alpha}(l)\) is a pre-estimate of \(G_{\alpha}(l)\) based on the latest gain estimate and the instantaneous gain. \(G_{\alpha}(l)\) is the final gain to be applied to the noisy signals.

Table 1 shows the PESQ MOS-LQO score improvement over the unprocessed noisy signals for various SNR conditions. The results demonstrate that the proposed noise estimation algorithm consistently achieves better performance over other algorithms.

Table 1. PESQ MOS-LQO score improvement over unprocessed noisy signal.

<table>
<thead>
<tr>
<th>SNR</th>
<th>Conn_freq</th>
<th>Doblinger</th>
<th>Hirsch</th>
<th>Imcra</th>
<th>Martin</th>
<th>Mcra</th>
<th>Mera</th>
<th>Mcra2</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>15dB</td>
<td>0.41</td>
<td>0.26</td>
<td>0.29</td>
<td>0.32</td>
<td>0.18</td>
<td>0.38</td>
<td>0.35</td>
<td>0.24</td>
<td>0.45</td>
</tr>
<tr>
<td>10dB</td>
<td>0.31</td>
<td>0.19</td>
<td>0.14</td>
<td>0.23</td>
<td>0.09</td>
<td>0.28</td>
<td>0.25</td>
<td>0.22</td>
<td>0.34</td>
</tr>
<tr>
<td>5dB</td>
<td>0.19</td>
<td>0.13</td>
<td>0.15</td>
<td>0.28</td>
<td>0.18</td>
<td>0.18</td>
<td>0.17</td>
<td>0.20</td>
<td>0.34</td>
</tr>
<tr>
<td>0dB</td>
<td>0.08</td>
<td>0.07</td>
<td>0.05</td>
<td>0.10</td>
<td>0.09</td>
<td>0.10</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
</tbody>
</table>

- Conn_freq: connected frequency regions [4]
- Doblinger: continuous spectral minimum tracking [5]
- Imcra: improved minimum controlled recursive average [3]
- Martin: Martin's minimum tracking [1]
- Mera: minimum controlled recursive average [2]
- Mcra2: variant of minimum controlled recursive average [7]

Since PESQ was developed mainly for speech coding and transmission applications, it may not be adequate to address the quality of noise floor which is crucial in evaluating speech enhancement algorithms[9]. Therefore, in addition to PESQ, a composite objective measure proposed by Hu and Loizou [9] is also used, which is derived by running regression analysis on the formal subjective evaluation results over the same NOIZEUS database. It has been shown to yield better correlation with subjective measures on the NOIZEUS database over other objective measures including PESQ. Since the same database is used here, such metric could potentially provide better prediction of subjective quality. As shown in Table 2, the composite score further validates the effectiveness of the proposed approach.

Table 2. Composite score improvement over unprocessed noisy signal.

<table>
<thead>
<tr>
<th>SNR</th>
<th>Conn_freq</th>
<th>Doblinger</th>
<th>Hirsch</th>
<th>Imcra</th>
<th>Martin</th>
<th>Mera</th>
<th>Mera2</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>15dB</td>
<td>0.18</td>
<td>0.09</td>
<td>0.12</td>
<td>0.09</td>
<td>0.08</td>
<td>0.17</td>
<td>0.13</td>
<td>0.24</td>
</tr>
<tr>
<td>10dB</td>
<td>0.17</td>
<td>0.08</td>
<td>0.11</td>
<td>0.09</td>
<td>0.06</td>
<td>0.16</td>
<td>0.11</td>
<td>0.22</td>
</tr>
<tr>
<td>5dB</td>
<td>0.13</td>
<td>0.06</td>
<td>0.11</td>
<td>0.09</td>
<td>0.02</td>
<td>0.13</td>
<td>0.07</td>
<td>0.22</td>
</tr>
<tr>
<td>0dB</td>
<td>0.03</td>
<td>0.02</td>
<td>0.08</td>
<td>0.03</td>
<td>0.02</td>
<td>0.07</td>
<td>0.03</td>
<td>0.11</td>
</tr>
</tbody>
</table>

6. Conclusions

A new noise spectrum estimation algorithm has been described. Specifically, a SAP weighted recursive noise estimator is adjusted by a minimum correction module, whose search duration is controlled by harmonicity. The advantage of such a structure is demonstrated with examples of a sudden jump of noise floor and long continuous phonation. Objective test results reveal that the proposed algorithm yields the highest PESQ and Composite score among several comparable noise estimation algorithms.

7. Acknowledgements

The authors wish to thank Gary Spittle for helpful comments and colleagues at CSR for their support.

8. References