A blind signal-to-noise ratio estimator for high noise speech recordings

Charles Mercier, Roch Lefebvre

Speech and Audio Research Group, Université de Sherbrooke, Sherbrooke (QC), J1K 2R1, CANADA
Charles.Mercier@USherbrooke.ca, Roch.Lefebvre@USherbrooke.ca

Abstract

Blind estimation of the signal-to-noise ratio in noisy speech recordings is useful to enhance the performance of many speech processing algorithms. Most current techniques are efficient in low noise environments only, justifying the need for a high noise estimator, such as the one presented here. A pitch tracker robust in high noise was developed and is used to create a two-dimensional representation of the audio input. Signal-to-noise ratio estimation is then performed using an image processing algorithm, effectively combining the short-term and long-term properties of speech. The proposed technique is shown to perform accurately even in high noise situations.

Index Terms: signal-to-noise ratio estimation, noisy speech, pitch tracking, speech signal processing

1. Introduction

The estimation of signal-to-noise ratio (SNR) from a single mixture of signal and noise is a topic of interest in many areas of signal processing. It is often an important part of algorithms that focus on noise reduction [1], voice activity detection [2] or automatic speech recognition [3] to name a few. Such systems frequently rely on an accurate estimation of SNR or an explicit noise model in order to adapt different parameters that will ensure optimal performances.

A broad range of techniques have been developed to achieve good results for low to moderate SNR (typically at SNR values above 0 dB). A standard approach formally detailed by the National Institute of Standards and Technology (NIST) sets a good reference point for SNR estimation systems. [4] It considers the amplitude distribution of short sections of signal, trying to fit those to Gaussian distributions, each of which corresponding to either the noise or the signal of interest. Analyzing the energy of the different distributions allows one to obtain a relatively accurate estimation of local SNR across the signal.

More sophisticated methods have also been developed to achieve higher precision. Current state-of-the-art approaches are based on the adaptive smoothing of the SNR maximum likelihood estimation in the frequency or cepstral domain. [5]

A common drawback of those approaches in SNR estimation is that they concentrate on providing highly accurate results for low-noise signals while neglecting the very noisy signals (SNR below 0 dB). To our knowledge, the performance in such adverse conditions is seldom investigated and rarely satisfactory. The proposed technique will specifically target this kind of high noise environment, where a good SNR estimation is required.

The basic concepts used in the proposed approach will first be presented in Section 2.1 with the global architecture of the system. The two main processing modules will then be detailed respectively in Sections 2.2 and 2.3. Performance evaluation will be presented in Section 3, followed by a discussion in Section 4. Section 5 will conclude on the technique developed and discuss further improvements.

2. Proposed system

2.1. Basic principles

Since the proposed system is intended to specifically target high noise speech recordings, it is logical to use a combination of noise-resistant speech properties. In the case of additive noise, we make the hypothesis that short-term autocorrelation and long-term periodicity of voiced sections will be preserved to some extent, even in the noisy mixture.

The noisy signal must then be arranged in such a way that one of these two properties will compensate for situations where the other would likely fail. A simple, yet effective way to achieve this is to locally segment the noisy speech signal and rearrange it to create a two-dimensional matrix where each line contains a single period of the underlying clean speech. A technique to estimate this period in high levels of noise is proposed in Subsection 2.2.

Analyzing the matrix along the horizontal axis provides information related to the short-term correlation while the vertical axis relates to the long-term periodicity found in speech, which is a function of its pitch. It is then possible to use this "segmented speech matrix" (SSM) to compute the SNR with a two-dimensional approach normally used in image processing. Details of this technique are presented in Subsection 2.3. Figure 1 illustrates the high-level architecture of the proposed system, where x is the noisy speech signal and p is the length of the fundamental period.

![Figure 1: High-level schematic diagram of proposed system.](image-url)

2.2. Fundamental period analysis

In order to correctly align consecutive periods of the voiced speech signal on each line of the SSM, the fundamental period of the speech signal buried in noise must first be computed. Pitch tracking is in itself a very rich area of research, where numerous methods have emerged over the past decades. However, while some very efficient algorithms exist for clean speech signals, few appear to have been proposed specifically for high noise recordings. In fact, it appears in different studies [6],[7] that pitch tracking performances dramatically...
deteriorate when SNR drops below 0 dB for the most common techniques in the literature. A novel approach directly focusing on low SNR speech recordings was therefore developed and is presented here.

2.2.1. Fundamental period computation

If we first make the hypothesis that the noise corrupting the clean signal is locally stationary, we can rightly presume that its variance would not change over a given analysis interval. Furthermore, if the noise is purely additive and uncorrelated to the speech, considering both to be random variables, the variance of the noise signal can be approximated as the sum of the noise variance and the clean signal variance, according to the Bienaymé equality. [8] This means that the variance of non-consecutive samples of the noisy signal will only be modulated by the contribution of the clean signal, the noise part acting as a simple bias.

This leads to the idea of using the variance of samples \( x_{k+n} \), where \( k \) is an offset in signal \( x \), \( p \) is a candidate fundamental period and \( n \) is an integer, to compute a score value \( S \). Formally, we apply Equations (1) to (3) to evaluate the local fundamental period candidate around the sample positioned at a given point \( C \) in signal \( x \). Variable \( M \) is the number of periods considered on each side of point \( C \) and \( S \) is the score value computed. Typical values of \( M \) range between 1 and 5.

\[
\text{mean}_{p,k} = \frac{1}{2M+1} \sum_{n=-M}^{M} x_{k+n} \quad (1)
\]

\[
\text{var}_{p,k} = \frac{1}{2M+1} \sum_{m=-M}^{M} (x_{k+n} - \text{mean}_{p,k})^2 \quad (2)
\]

\[
S_p = \frac{1}{p} \sum_{k=C}^{C+p-1} \text{var}_{p,k} \quad (3)
\]

Studying the progression of the score value \( S \) over a predefined range of \( p \) corresponding to the usual fundamental frequencies of human speech allows us to determine the fundamental period (FP) of the underlying speech signal using Equation (4). In Equations (4) and (8) to (12), we use the form \( f \) to represent the approximation of function \( f \).

\[
FP = \text{argmin}_p(S_p) \quad (4)
\]

Intuitively, it makes sense that \( S_p \) in Equation (3) should be minimized if the length \( p \) of consecutive blocks equals the period of the clean signal. In the case of a noise-free synthetic periodic signal, \( S_p \) would be 0 for the correct period length. Even with quasi-periodic noisy signals such as those being studied here, this method tends to produce mostly accurate results, as will be shown in Section 3. Rearranging the analysis interval in a matrix of \( 2M+1 \) lines and \( FP \) columns, where sample \( C \) is the first element of the central line, gives us the desired SSM, which will be processed to determine the SNR.

2.2.2. Possible noise-dependent adaptation

Accurate estimation of the fundamental period is crucial for proper SNR estimation in the proposed algorithm and must therefore be robust to typical forms of noise. In the presence of noise with strong periodicity, such as car noise, the algorithm would be likely to track the fundamental period of the noise instead of the desired speech fundamental period. An easy way to adapt the system to be able to cope with such situations is to apply a high-pass (HP) filter before computing the score values of Section 2.2.1. Considering that speech contains strong harmonics at higher frequencies than, for example, the car noise, correctly adjusting the cutoff frequency of the high-pass filter allows the algorithm to track speech periodicity rather than noise periodicity.

2.3. Two-dimensional SNR estimation

The SSM created in Section 2.2.1 using the fundamental period analysis can now be used in order to estimate SNR. As detailed in Equations (5) to (11), it is possible to use an estimation of the clean signal zero-lag autocorrelation \( \rho_x(0,0) \) to estimate the SNR of the noisy signal. This idea was introduced in [9] for blindly estimating the SNR in images. The method is adapted here by considering the two-dimensional SSM as an image. The mean value of the SSM is subtracted prior to the autocorrelation computations.

First, we define the two-dimensional autocorrelation in Equation (5). It is important to note that the autocorrelation applied here is circular rather than zero-padded. This is justified by the periodic nature of the signal creating the SSM and was shown to provide better results.

\[
\rho_x(t_1, t_2) = \frac{1}{L} \sum_{l} x_{l-t_1} x_{l-t_2} \quad (5)
\]

Here, \( x \) is the noisy signal mixture rendered two-dimensional by its disposition in the SSM. The SNR (in decibels) for discrete signals with a mean value of 0 is:

\[
\text{SNR}(dB) = 10 \log_{10} \left( \frac{\sum_k s_k^2}{\sum_k n_k^2} \right) \quad (6)
\]

where \( s \) denotes the clean speech signal and \( n \) the noise part, both unavailable from the mixture \( x \). Combining Equations (5) and (6) allows for another useful form of the SNR equation:

\[
\text{SNR}(dB) = 10 \log_{10} \left( \frac{\phi_x(0,0)}{\rho_n(0,0)} \right) \quad (7)
\]

Since the underlying signals \( s \) and \( n \) are not readily available, we need to adapt Equation (7) to obtain an SNR approximation using only parameters that can be evaluated. As was detailed in [9], the value of \( \rho_n \) can be expressed as \( \phi_s - \phi_x \) if \( s \) and \( n \) are uncorrelated, leading to the SNR estimation formula in Equation (8).

\[
\text{SNR}(dB) = 10 \log_{10} \left( \frac{\phi_x(0,0)}{\phi_s(0,0)} \right) \quad (8)
\]

As can be seen, the only missing parameter in Equation (8) for an accurate estimation of SNR is \( \phi_s(0,0) \). The particular form of the SSM allows us to combine the short-term (ST) and long-term (LT) properties of speech to obtain this estimation. A simple linear interpolation of the adjacent horizontal and vertical autocorrelation coefficients is mixed using coefficient \( \rho \) in Equation (11) so that both properties will be taken into account. It has been determined experimentally that setting \( \rho \) to 0.5 provides satisfying results, but this value can be increased to better cope with highly periodic types of noise.

\[
\phi_{SY}(0,0) = 2\phi_x(1,0) - \phi_x(2,0) \quad (9)
\]

\[
\phi_{TY}(0,0) = 2\phi_x(0,1) - \phi_x(0,2) \quad (10)
\]

\[
\phi_s(0,0) = \rho\phi_{LT}(0,0) + (1 - \rho)\phi_{SY}(0,0) \quad (11)
\]
Using Equation (11) in Equation (8) gives the desired blind SNR estimator, which is rendered noise-robust by the combination of short-term and long-term speech signal information. Considering the very low level of correlation in noise-only sections, it is possible that Equation (8) produces an invalid result trying to evaluate the logarithm of a negative value. Such occasional occurrences can simply be discarded, since they most likely do not contain a significant portion of the signal of interest.

3. Performance evaluation

3.1. Fundamental period tracking

To illustrate the efficiency of the fundamental period tracking of Section 2.2.1, a test signal consisting of a stationary mixture of 4 sinusoids was used and is shown in Figure 2. Note that for all results presented in this section, a sampling frequency of 16 kHz was used.

![Figure 2: Test signal used for evaluating the performance of the fundamental period tracking algorithm.](image)

The fundamental period of the test signal was fixed at a constant value of 100 samples, which is a typical value found in human speech. The test signal used is also similar to common voiced sections of speech, providing a valuable insight on how the system will perform in real-life scenarios. The car noise is an extract from the NOISEX-92 database. [10]

Table 1. Standard deviation (σ) values of fundamental period tracking error (ε) under various conditions.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>No HP filter</th>
<th>HP filter</th>
<th>No HP filter</th>
<th>HP filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.0027</td>
<td>0.0030</td>
<td>0.0077</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0.0070</td>
<td>0.0073</td>
<td>0.0830</td>
<td>0.0004</td>
</tr>
<tr>
<td>-5</td>
<td>0.0265</td>
<td>0.0459</td>
<td>0.2073</td>
<td>0.0027</td>
</tr>
<tr>
<td>-10</td>
<td>0.1855</td>
<td>0.2190</td>
<td>0.2560</td>
<td>0.0092</td>
</tr>
<tr>
<td>-15</td>
<td>0.2923</td>
<td>0.3072</td>
<td>0.2608</td>
<td>0.0581</td>
</tr>
<tr>
<td>-20</td>
<td>0.3271</td>
<td>0.3367</td>
<td>0.2598</td>
<td>0.1439</td>
</tr>
</tbody>
</table>

3.2. Signal-to-noise ratio estimation

The SNR estimation performance analysis was done using real speech data consisting of 166 phonetically balanced sentences uttered by 4 male and 4 female speakers, totaling over 10 minutes of speech, extracted from the NTT-AT database. [11] SNR estimation was performed once per 25 milliseconds. The SNR values were computed using Equation (8), with a new SSM created once per 25 milliseconds.

![Figure 4: Correlation between clean signal energy and SNR estimation error for stationary white noise at -5 dB SNR, using no HP, M = 2 and p = 0.5.](image)
Figure 4 shows the distribution of the SNR estimation errors for white noise at an SNR of -5 dB, relative to the energy of the local clean signal. This illustrates that the algorithm performs better during high energy sections of speech, where periodic voiced sections are typically concentrated. Performances degrade for low energy sections, where the absence of periodicity negates the benefits of the two-dimensional representation of speech used.

The estimation error statistics are shown in Figure 5, both for the case of white noise and car noise. We notice that performances under white noise remain accurate down to approximately -10 dB with the standard deviation progressively rising as SNR drops. Performance under car noise conditions on the other hand exhibits nearly constant standard deviation and an average error value gradually rising through the entire SNR range studied.

The score function developed in Section 2.2.1 also proved to be very noise-resistant and effective in finding the fundamental period of a signal hidden in noise. The applications of this technique could easily extend to the domain of pitch tracking in high noise situations. The effectiveness and low complexity of this approach make it interesting to complement current systems, helping to cope with highly degraded input signals.

More sophisticated techniques could improve the accuracy of the SNR estimation, for example by adding a voice-activity detector. This would prevent the system from considering SNR estimations during noise-only sections of the signal, contributing to biasing the results.

5. Conclusion

A novel algorithm for blind SNR estimation in a single mixture of speech and noise was presented. The system applies a new pitch tracking algorithm coupled with an SNR estimation step derived from image processing. The algorithm provides good performance for SNR levels as low as -10 dB. The proposed pitch tracking algorithm was shown to be very robust in noisy conditions and could be applied in different applications such as voice recognition in high noise recordings. The system is being used as part of a speech intelligibility enhancement application currently in development.

6. Acknowledgements

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada for funding this research.

7. References