LOW-COMPLEXITY NOISE POWER SPECTRAL DENSITY ESTIMATION FOR HARSH AUTOMOBILE ENVIRONMENTS

Christin Baasch†, Vasudev Kandade Rajan†
†Digital Signal Processing and System Theory, Kiel University, Germany
{chr,b,vakr,gus}@tf.uni-kiel.de

Mohamed Krini*, Gerhard Schmidt‡
*paragon AG, Delbrueck, Germany
{Mohamed.Krini}@paragon.ag

ABSTRACT

In this paper a simple yet robust noise power spectral density estimation algorithm is presented. The motivation for the algorithm developed here is the harsh noise environment present in automobiles along with the need to keep the complexity low for real-time implementations. The scheme is based on a multiplicative estimator in which multiple increment and decrement time-constants are utilized. The time-constants are chosen based on noise-only and speech-like situations. Further, by observing the long-term “trend” of the noisy input spectrum, suitable time-constants are chosen which reduces the tracking delay significantly. The trend factor is measured taking into account the dynamics of speech. Evaluation of the proposed algorithm and comparisons with state of the art systems will be performed in the context of varying automobile noises.

1. INTRODUCTION

Estimating the background noise of a microphone signal is an extensively studied topic with many practical applications, such as hands-free systems, mobile phones and many other devices. In automotive environments speech enhancement systems have to deal with low signal-to-noise ratio (SNR) conditions, and different kinds of non-stationary noise. Along with it, the computational complexity of the various algorithms in the speech enhancement system has to be kept low and is a serious constraint. Several approaches exist for estimating the power spectral density (PSD) of noise. The minimum statistics approach for example tracks the minimum of a smoothed noisy input spectrum within a window [1]. In this approach the noisy input is smoothed with an optimal smoothing parameter and a minimum is searched within a time window. A similar approach but in combination with a voice activity detector (VAD) is presented in [2]. Exploitation of subband limited-rank structure of speech in noisy inputs using its DFT coefficients is explored in [3]. The method presented in [4] computes the probability of speech (SPP) and uses this probability to estimate the noise PSD by combining it with a minimum mean square error (MMSE) based noise estimator.

The MMSE based noise estimator is presented in [5]. An approach that utilizes the dynamics of speech to detect speech pauses is presented in [6]. Here the broadband, low-pass and high-pass power of the noisy input signal is tracked and speech pauses are detected by adaptively tracking the minima in the power envelope. A data-driven approach to track sudden changes in the background noise is presented in [7]. Here an offline gain table is computed from a database which is used for estimating the noise PSD. In this paper an extremely simple approach which is called as the multiplicative time-constant based approach is applied similar to the one shown in [8]. This algorithm has the advantage of very low computational complexity. The basic idea is to choose a multiplicative constant depending on a specific situation. By using a long-term observation of the trend of the input signal along with several threshold parameters decisions on specific situations are taken. Finally a weighted combination of the pre-estimate and the input spectrum is calculated to obtain the final noise estimate. The remainder of the paper is organized as follows. The noisy signal model is introduced in Sec. 2. The trend factor computation, followed by the situation based choosing of multiplicative time-constants are discussed in Sec. 3. Results of the proposed algorithm are presented in Sec. 4.

2. NOISY SIGNAL MODEL AND FRAMEWORK

The noisy speech signal in the discrete-time domain is modeled here as \( y(n) = s(n) + b(n) \) where \( n \) is the discrete-time index, \( y(n) \) is the signal recorded by a microphone, \( s(n) \) is the clean speech signal and \( b(n) \) is the noise component. The processing of the signals is performed in the subband domain. An short-time-Fourier-transform (STFT) based analysis-synthesis filterbank is used to transform the signal into its subbands and back to time-domain. The output of the analysis filterbank is the short-term spectrum of the input signal \( Y(\mu, k) \) where \( \mu \) is the subband index and \( k \) is the frame index. The estimated background noise \( \hat{B}(\mu, k) \) is used by a noise suppression filter like the Wiener filter to obtain an estimate of the clean speech.
3. PROPOSED NOISE PSD ESTIMATION

3.1. Tracking the Noise Spectrum

Noise present in the input spectrum can be estimated by accurately tracking the segments of the spectrum in which speech is absent. The behavior of this spectrum is dependent on the environment in which the microphone is placed. In an automobile environment there are many factors that contribute to the non-stationarity of the noise spectrum. Generally for such environments the noise spectrum can be described as non-flat with a low-pass characteristic dominated below 500 Hz. Apart from this, low-pass characteristics, changes in speed, opening and closing of windows, passing cars, etc. cause the noise floor to vary with time. A close look at one frequency bin of the noise spectrum reveals the following properties:

1. Instantaneous power can vary a large extent from the mean power even during steady conditions.
2. A steady increase or a steady decrease of power is observed during certain situations (e.g. during acceleration).

A simple estimator which can be used to track these magnitude changes for each frequency bin is shown in Eq. (1)

\[
\hat{B}(\mu, k) = \begin{cases} 
\hat{B}(\mu, k-1) \Delta_{inc}, & \text{if } Y(\mu, k) > \hat{B}(\mu, k-1), \\
\hat{B}(\mu, k-1) \Delta_{dec}, & \text{else},
\end{cases}
\]

where \(\hat{B}(\mu, k)\) is the estimated background noise. This estimator follows a smoothed input \(Y(\mu, k)\) based on previous noise estimate. The speed at which it tracks the noise floor is controlled by the increment constant \(\Delta_{inc}\) and the decrement constant \(\Delta_{dec}\). The advantage of this algorithm is its low computational complexity. Such an estimator can be made to work with careful parametrization of increment and decrement constants combined with a highly smoothed input. From the observations presented about the noise this estimator would fail for the following two reasons

- low time-constants will lag in tracking the noise power, and
- high time-constants will estimate speech as noise.

Starting from this simple estimator the proposed noise estimation algorithm tries to find a balance by keeping the computational complexity low and offering fast, accurate tracking. The principle behind the new estimator is to choose the “right” multiplicative constant given a specific situation. Such a situation can be a speech passage, a consistent background noise, increasing background noise, decreasing background noise, etc. A measure called as the “trend” is computed which indicates if the long-term direction of the input signal is going up or down. The incremental and decremental time-constants along with the trend are applied together in Eq. (9).

3.2. Input Spectrum Smoothing

The tracking of the noise estimator is dependent on the smoothed input signal \(Y(\mu, k)\). The input spectrum is smoothed using a first order IIR filter

\[
Y(\mu, k) = \gamma_{smth} |Y(\mu, k)| + (1 - \gamma_{smth}) Y(\mu, k-1),
\]

where \(\gamma_{smth}\) is the smoothing constant. The smoothing constant must be chosen in such a way that it retains fine variations of the input spectrum as well as eliminate the high variation of the instantaneous spectrum. Optionally additional frequency-domain smoothing can be applied. A value of 300 dB per second is chosen here. \(^1\)

3.3. Trend: Long-term Activity Measurement

One of the difficulties for noise estimators in non-stationary environments is differentiating between a speech part of the spectrum and an actual change in the spectral floor. This problem can be partially overcome by measuring the duration for a power increase. If the increase is due to a speech source then the power will drop down after the utterance of a syllable, whereas if the power continues to stay up for a longer duration, then it is an indication of an increased background noise. It is these dynamics of the input spectrum that the trend factor measures in the algorithm. By observing the direction of the trend, going up or down, the spectral floor changes can be tracked by avoiding to track speech-like parts of the spectrum. The decision about current state of the frame is made by comparing if the estimated noise of the previous frame is smaller than the smoothed input spectrum of the current frame and a set of values are obtained. A positive value indicates that the direction is going up, and a negative value indicates that the direction is going down

\[
A_{curr}(\mu, k) = \begin{cases} 
1, & \text{if } Y(\mu, k) > \hat{B}(\mu, k-1), \\
-4, & \text{else},
\end{cases}
\]

where \(\hat{B}(\mu, k-1)\) is the estimated noise of the previous frame. The values 1 and -4 are chosen empirically. The trend is smoothed along both the time and the frequency axis. A zero-phase forward-backward filter is used for smoothing along the frequency axis. Smoothing along the frequency ensures isolated peaks caused by non-speech like activities are suppressed. Smoothing is applied by using

\[
\hat{A}_{trnd}(\mu, k) = \gamma_{trnd-fq} A_{curr}(\mu, k) + (1 - \gamma_{trnd-fq}) \hat{A}_{trnd}(\mu - 1, k),
\]

for \(\mu = 1, \ldots, N_{SBB}\) and similarly backward smoothing is applied. Both the frequency smoothing constants \(\gamma_{trnd-fq}\) are

\(^1\)In order to be independent of the sample rate and frameshift all time-constants are denoted in dB per second.
chosen to be at about 35 dB per Hertz. The time-smoothed trend factor $\overline{A}_{\text{trnd}}(\mu, k)$ again is given by an IIR filter

$$
\overline{A}_{\text{trnd}}(\mu, k) = \gamma_{\text{trnd-tm}} \overline{A}_{\text{trnd}}(\mu, k) + (1 - \gamma_{\text{trnd-tm}}) \overline{A}_{\text{trnd}}(\mu, k - 1),
$$

where $\gamma_{\text{trnd-tm}}$ is the smoothing constant chosen to be at about 15 dB per second. The behavior of the double-smoothed trend factor $\overline{A}_{\text{trnd}}(\mu, k)$ can be summarized as follows. The trend factor is a long-term indicator of the power level of the input spectrum. During speech parts the trend factor temporarily goes up but comes down quickly. When the true background noise increases then the trend goes up and stays there until the noise estimate catches up. Similar behavior is seen for a decreasing background noise power. This trend measure is used to further “push” the noise estimate in the desired direction. The trend is compared to an upward threshold and a downward threshold. When either of these thresholds are reached then the respective time-constant to be used later is chosen as shown in Eq. (6)

$$
\Delta_{\text{trnd}}(\mu, k) = \begin{cases} 
\Delta_{\text{trnd-up}}, & \text{if } \overline{A}_{\text{trnd}}(\mu, k) > T_{\text{trnd-up}}, \\
\Delta_{\text{trnd-down}}, & \text{if } \overline{A}_{\text{trnd}}(\mu, k) < T_{\text{trnd-down}}, \\
1, & \text{else}.
\end{cases}
$$

The values of $\Delta_{\text{trnd-up}}$ and $\Delta_{\text{trnd-down}}$ are chosen to be at 20 and −20 dB per second.

### 3.4. Situation Based Selection of Tracking Constants

The tracking of the noise estimation has to be performed for two cases, one, when the smoothed input is greater than the estimated noise, and, two, when it is smaller.

#### 3.4.1. Incrementation of the Noise Estimate

The input spectrum can be greater than the estimated noise due to three reasons. First, when there is speech activity, second when the previous noise estimate has dipped too low and has to rise up, and third when there is a continuous increase in the true background noise. The first case is handled by checking if the level of $\overline{Y}(\mu, k)$ is greater than a certain SNR threshold $T_{\text{snr}}$, in which case the chosen incremental constant $\Delta_{\text{inc-fast}}$ has to be very slow because speech should not be tracked. For the second case the incremental constant is set to $\Delta_{\text{inc-noise}}$ which means that this is a case of normal rise and fall during tracking. For the case of a continuous increase in the true background noise the estimate must catch up with it as fast as possible. For this a counter $k_{\text{cut}}(\mu, k)$ is utilized. The counter counts the duration for which the input spectrum has stayed above the estimated noise. If this counter reaches a threshold $K_{\text{inc-max}}$ then the $\Delta_{\text{inc-fast}}$ is chosen. The counter is incremented by 1 every time $\overline{Y}(\mu, k)$ is greater than $\hat{B}(\mu, k - 1)$ and reset to 0 otherwise. Equation (7) captures these conditions

$$
\Delta_{\text{inc}}(\mu, k) = \begin{cases} 
\Delta_{\text{inc-fast}}, & \text{if } k_{\text{cut}}(\mu, k) > K_{\text{inc-max}}, \\
\Delta_{\text{speech}}, & \text{else if } \overline{Y}(\mu, k) > \hat{B}(\mu, k - 1) T_{\text{snr}}, \\
\Delta_{\text{inc-noise}}, & \text{else}.
\end{cases}
$$

The value for the fast increment $\Delta_{\text{inc-fast}}$ is chosen to be at about 40 dB per second. For the speech case $\Delta_{\text{speech}}$ has to very slow and is chosen to be at about 0.5 dB per second, and finally the $\Delta_{\text{inc-noise}}$ is chosen to be at about 6 dB per second.

#### 3.4.2. Decrementation of the Noise Estimate

The choice of a decrementing constant does not have to be as explicit as the in the incrementing case. This is because of lesser ambiguity in when $\overline{Y}(\mu, k)$ is less than $\hat{B}(\mu, k - 1)$. Here the noise estimator chooses a decremental constant $\Delta_{\text{dec}}$ by default. The value for falling edge is chosen to be at about −20 dB per second. For a subband $\mu$ only one of the above two stated conditions is chosen. From either of the two conditions a final multiplicative constant is determined

$$
\Delta_{\text{final}}(\mu, k) = \begin{cases} 
\Delta_{\text{inc}}(\mu, k), & \text{if } \overline{Y}(\mu, k) > \hat{B}(\mu, k - 1), \\
\Delta_{\text{dec}}, & \text{else}.
\end{cases}
$$

### 3.5. Combining Noise Estimate and Input Spectrum

The input spectrum consists of only background noise when no speech-like activity is present. During this time the best estimate is to set the noise estimate equal to the input spectrum. When the estimated noise is lower than the input spectrum, the noise estimate and the input spectrum are combined with a certain weight. The weights are computed according
to Eq (10). A pre-estimate $\hat{B}(\mu, k)$ is obtained for computing the weights. The pre-estimate is used in combination with the input spectrum. It is obtained by multiplying the input spectrum with the multiplicative constant $\Delta_{\text{final}}(\mu, k)$ and the trend constant $\Delta_{\text{Trend}}(\mu, k)$

$$\hat{B}_{\text{pre}}(\mu, k) = \Delta_{\text{final}}(\mu, k) \Delta_{\text{Trend}}(\mu, k) \hat{B}(\mu, k - 1). \quad (9)$$

The weighting factor for combining the input spectrum and the pre-estimate is given by

$$W_{\hat{B}}(\mu, k) = \min \left\{ 1, \left( \frac{\hat{B}_{\text{pre}}(\mu, k)}{\hat{Y}(\mu, k)} \right)^2 \right\}. \quad (10)$$

The final noise estimate is computed by applying this weighting factor

$$\hat{B}(\mu, k) = W_{\hat{B}}(\mu, k) \hat{Y}(\mu, k) + (1 - W_{\hat{B}}(\mu, k)) \hat{B}_{\text{pre}}(\mu, k). \quad (11)$$

During the first few frames of the noise estimation algorithm, the input spectrum itself is directly chosen as the noise estimate for faster convergence. The plot in Fig. 1 shows the working of the noise estimation algorithm for subband $\mu = 34$.

4. RESULTS AND CONCLUSION

The proposed algorithm was evaluated under different automobile noise conditions to test the performance in realistic noise situations encountered while driving. Noise recordings were performed under different speeds, with the air-condition system turned on, opening/closing of a window, accelerating to a high speed, breaking to a low speed, etc where the mics were fitted on the seat belt and Harward Sentences [9] were used for speech. Fig. 2 shows the log error distance plot of the proposed noise estimation algorithm as compared to SPP and minimum statistics methods as these were the best among the 5 estimation schemes that were evaluated in our tests for different SNRs [10]. The errors for the proposed scheme occur mainly when the rising spectrum has to be followed. The SPP follows the noise well but also follows some speech segments thereby distorting parts of speech. The minimum statistics was not able to follow the rising spectrum parts but has minimum speech distortion. The proposed algorithm also performed better in terms of segmental SNR and overall SNR improvement. In this paper a simple approach to noise estimation has been presented based on switching of the time-constants. The advantages are its low complexity, low tracking delay, easy implementation, and a robust performance in various conditions crucial in harsh automobile environments.

Fig. 2. Log error plot of the proposed algorithm as compared to SPP [4] and MS [1] under various automobile noise conditions.
5. REFERENCES


