SIGNAL PROCESSING VII: Theories and Applications
M. Holt, C. Cowan, P. Grant, W. Sandham (Eds.)

Spectral Subtraction Based on Minimum Statistics

Rainer MARTIN

Institute for Communication Systems and Data Processing (IND), Aachen University of Technology, Templergraben 55, 52056 Aachen, Germany, Phone: +49 241 806984, Fax: +49 241 8888186, E-Mail: martin@ind.rwth-aachen.de

Abstract. This contribution presents and analyses an algorithm for the enhancement of noisy speech signals by means of spectral subtraction. In contrast to the standard spectral subtraction algorithm the proposed method does not need a speech activity detector nor histograms to learn signal statistics. The algorithm is capable to track non stationary noise signals and compares favorably with standard spectral subtraction methods in terms of performance and computational complexity. Our noise estimation method is based on the observation that a noise power estimate can be obtained using minimum values of a smoothed power estimate of the noisy speech signal. Thus, the use of minimum statistics eliminates the problem of speech activity detection. The proposed method is conceptually simple and well suited for real time implementations. In this paper we derive an unbiased noise power estimator based on minimum statistics and discuss its statistical properties and its performance in the context of spectral subtraction.

1. Introduction

Spectral subtraction is a well known speech enhancement technique which has by now developed numerous facets [1]. At the heart of a spectral subtraction algorithm is a noise power estimator and a subtraction rule which translates the subband SNR into a spectral weighting factor, such that subbands with low SNR are attenuated and subbands with high SNR are not modified. Both the noise power estimator and the subtraction rule have significant impact on the audible residual noise [2]. The basic spectral subtraction algorithm which requires only one microphone employs a speech activity detector to update noise statistics. Therefore, tracking of varying noise levels might be slow and confined to periods of no speech activity.

In this paper we address the problem of noise power estimation and develop an algorithm which essentially eliminates the need for explicit speech pause detection without an substantial increase in computational complexity. While the conventional approach to spectral subtraction employs a speech activity detector we here use the minimum of the subband noise power within a finite window to estimate the noise floor. The algorithm is based on the observation that a short time subband power estimate of a noisy speech signal exhibits distinct peaks and valleys (see Figure 1). While the peaks correspond to speech activity the valleys of the smoothed noise estimate can be used to obtain an estimate of subband noise power. To obtain reliable noise power estimates the data window for the minimum search must be large enough to bridge any peak of speech activity.

The remainder of this paper is organized as follows. In section 2 we will present the spectral subtraction algorithm and the noise power estimation method. In section 3 we will discuss the statistical properties of minimum subband noise power estimates. Section 4 will present experimental results.

2. Description of Algorithm

A block diagram of the basic spectral subtraction method is shown in Figure 2. The algorithm appropriately modifies the short time spectral magnitude of the disturbed speech signal such that the synthesized signal is perceptually as close as possible to the undisturbed speech signal. The optimal weighting of spectral magnitudes is computed using a noise power estimate and a subtraction rule.

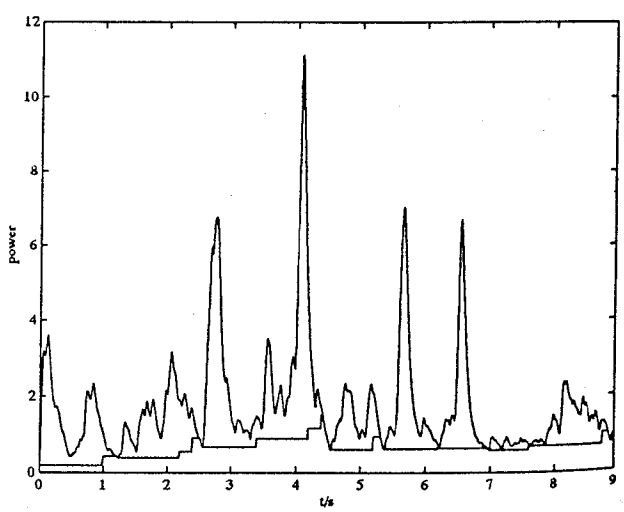

Figure 1: Short time subband power and estimated noise floor of noisy speech signal (f_s=8kHz, $W_{DFT} = 256$, subband k=8)

2.1 Spectral Analysis/Synthesis

We assume that the bandlimited and sampled disturbed signal $x(i)$ is a sum of a zero mean speech signal $s(i)$ and a zero mean noise signal $n(i)$, $x(i) = s(i) + n(i)$, where i denotes the time index. We further assume that $s(i)$ and $n(i)$ are statistically independent, hence $E\{x^2(i)\} = E\{s^2(i)\} + E\{n^2(i)\}$. Spectral processing is based on a DFT filter bank with W_{DFT} subbands and with decimation/interpolation ratio R [3]. The phase of the disturbed signal is not modified. We denote the data window by $h(i)$ and the DFT of the windowed disturbed signal $x(i)$ by

$$X(\lambda, k) = \sum_{\mu=0}^{W_{DFT}-1} x(\lambda R - \mu) \cdot h(\mu) \cdot exp\left(-j \frac{2\pi\mu k}{W_{DFT}}\right) \quad (1)$$

λ and k refer to the decimated time index and the DFT frequency bins $\Omega_k = 2\pi k/W_{DFT}$, $k \in 0, 1, ..., W_{DFT} - 1$, respectively. Typically we use a DFT length of $W_{DFT} = 256$ and decimation ratio $R = 64$. The improved subband signals are converted back to the time domain using an inverse DFT. The synthesized improved speech signal is denoted by $y(i)$, the corresponding spectral magnitude by $|Y(\lambda, k)|$.

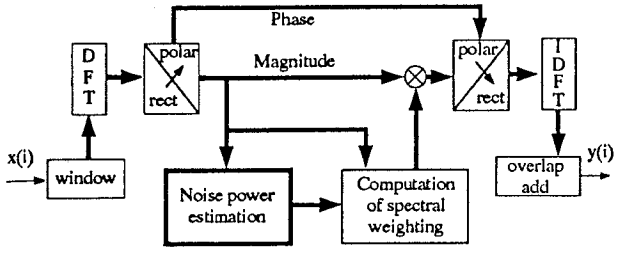

Figure 2: Block diagram of spectral processing

2.2 Subtraction Rule

Let $P_n(\lambda, k)$ and $\overline{|X(\lambda, k)|^2}$ denote the estimated subband noise power and short time signal power, respectively. To obtain the short time signal power subsequent magnitude squared input spectra are smoothed with a first order recursive network ($\gamma \lesssim 0.9$)

$$\overline{|X(\lambda, k)|^2} = \gamma \cdot \overline{|X(\lambda - 1, k)|^2} + (1 - \gamma) \cdot |X(\lambda, k)|^2 \quad (2)$$

Following the proposal of Berouti et. al. [4] we subtract spectral magnitudes with an oversubtraction factor $osub(\lambda, k)$ and a limitation of the maximum subtraction by a spectral floor constant $subf$ ($0.01 \leq subf \leq 0.05$)

$$|Y(\lambda, k)| = \begin{cases} subf \sqrt{P_n(\lambda, k)} & if\ |X(\lambda, k)| \cdot Q(\lambda, k) \leq subf \sqrt{P_n(\lambda, k)} \\ |X(\lambda, k)| \cdot Q(\lambda, k) & else \end{cases}$$

$$\text{where } Q(\lambda, k) = \left(1 - \sqrt{osub(\lambda, k) \frac{P_n(\lambda, k)}{|X(\lambda, k)|^2}}\right) \quad (3)$$

While a large oversubtraction factor $osub(\lambda, k)$ essentially eliminates residual spectral peaks ('musical noise') it also affects speech quality such that some of the low energy phonemes are suppressed. To limit this undesirable effect the oversubtraction factor is computed as a function of the subband signal-to-noise ratio $SNR_x(\lambda, k)$ and the frequency bin k, i.e. $osub(\lambda, k) = f(\lambda, k, SNR_x(\lambda, k))$. In general we use less oversubtraction for high SNR conditions and for high frequencies than for low SNR conditions and for low frequencies.

2.3 Subband Noise Power and SNR Estimation

We first compute the short time subband signal power $P_x(\lambda, k)$ using recursively smoothed periodograms. The update recursion is given by eq.(4). The smoothing constant is typically set to values between $\alpha = 0.9...0.95$.

$$P_x(\lambda, k) = \alpha \cdot P_x(\lambda - 1, k) + (1 - \alpha) \cdot |X(\lambda, k)|^2 \quad (4)$$

The noise power estimate $P_n(\lambda, k)$ is obtained as a weighted minimum of the short time power estimate $P_x(\lambda, k)$ within a window of D subband power samples [5], i.e.

$$P_n(\lambda, k) = omin \cdot P_{min}(\lambda, k) \quad (5)$$

$P_{min}(\lambda, k)$ is the estimated minimum power and $omin$ is a factor to compensate the bias of the minimum estimate. In section 3 we show that $omin$ depends only on known algorithmic parameters.

For reasons of computational complexity and delay the data window of length D is decomposed into W windows of length M such that $M \cdot W = D$. For a sampling rate of f_s=8 kHz and a decimation ratio R=64 typical window parameters are M=25 and W=4, thus D=100 corresponding to a time window of $((D-1) \cdot R + W_{DFT})/f_s = 0.824s$.

To determine the minimum of M consecutive subband power samples at time $\lambda = \lambda_1$ we initialize a variable $P_{Mact}(\lambda = \lambda_1, k)$ to the first of the M samples $P_{Mact}(\lambda = \lambda_1, k) = P_x(\lambda = \lambda_1, k)$. The minimum of the M samples, $P_{Mmin}(\lambda, k)$, is then found by a samplewise comparison of the actual minimum $P_{Mact}(\lambda, k)$ with the short time power $P_x(\lambda, k)$. Whenever M samples have been read, i.e. $\lambda = \lambda_1 + M - 1$ we store the minimum power of the last M samples $P_{Mmin}(\lambda = \lambda_1 + M - 1, k) = P_{Mact}(\lambda = \lambda_1 + M - 1, k)$ and the search for the minimum begins over again. The minimum power of the length D window is now easily obtained as the minimum of the last W minimum power estimates $P_{Mmin}(\lambda = \lambda_1 + qM - 1, k)$ with $q = 1, 0, ..., (2 - W)$. The decomposition of the length D window into W subwindows has the advantage that a new minimum estimate is available after already M samples without an substantial increase in compare operations.

If the actual subband power $P_x(\lambda, k)$ is smaller than the estimated minimum noise power $P_{min}(\lambda, k)$ the noise power is updated immediately independent of window adjustment: $P_{min}(\lambda, k) = \min(P_x(\lambda, k), P_{min}(\lambda, k))$. Thus in case of decreasing noise power we achieve a fast update of the minimum power estimate. In case of increasing noise power the update of noise estimates is delayed by $D + M$ samples. Finally, to control the oversubtraction factor $osub(\lambda, k)$ we compute the SNR in each subband

$$SNR_x(\lambda, k) = 10 \cdot log\left(\frac{P_x(\lambda, k) - \min(P_n(\lambda, k), P_x(\lambda, k))}{P_n(\lambda, k)}\right) \quad (6)$$

Figure 1 plots the short time power estimate and the estimated noise floor for a noisy speech sample. The window length $D = M \cdot W$ must be large enough to bridge any peak of speech activity, but short enough to follow non stationary noise variations. Experiments with different speakers and modulated noise signals have shown that window lengths of approximately 0.8s - 1.4s give good results.

3. Statistical Properties of the Minimum Power Estimate

In this section we derive bias and variance of the minimum estimate with the aim to develop an unbiased noise power estimator and to evaluate its statistical efficiency. As a result we will be able to compute the overestimation factor $omin$. To facilitate the analytical evaluation of minimum estimates we assume that the noise process $n(i)$ is stationary and that no speech is present, i.e.

$x(i) = n(i)$. $S_{xx}(\Omega)$ will denote the power spectral density of signal $x(i)$.

We also assume that the computation of the short time subband power estimate is done by means of non recursive smoothing of K successive magnitude squared spectra, i.e.

$$P_x(\lambda, k) = \frac{1}{K} \sum_{m=0}^{K-1} |X(\lambda - m, k)|^2 \qquad (7)$$

If successive spectra $X(\lambda, k)$ are independent the subband power estimate $P_x(\lambda, k)$ is asymptotically chi-square distributed with mean $\sigma^2(k) = S_{xx}(\Omega_k = 2\pi k/W_{DFT}) \cdot \sum_{m=0}^{W_{DFT}-1} h^2(m)$ and $N = K$ degrees of freedom for DC and Nyquist frequency bins and $N = 2K$ degrees of freedom for all other frequency bins

$$f_{P_x}(y, k) = \frac{1}{\left(\sqrt{\frac{2\sigma^2(k)}{N}}\right)^N \Gamma(N/2)} \cdot y^{N/2-1} \cdot e^{-yN/2\sigma^2(k)} \cdot U(y) \qquad (8)$$

$\Gamma()$ and $U()$ denote the Gamma function and the unit step function, respectively. Since the data segments from which the periodogram is computed usually overlap the approximating chi-square distribution has less than N degrees of freedom. Following the approach by Welch [6] the equivalent degrees of freedom for overlapping data segments are estimated by fitting a chi-square distribution with same mean and variance as the smoothed periodogram in eq. (7). For non recursive (eq. (7)) and for recursive smoothing (eq. (4)) the equivalent degrees of freedom are given by N_{nrec} and N_{rec}, respectively

$$N_{nrec} \approx \frac{2K}{a(K)} \qquad N_{rec} \approx 2\frac{1+\alpha}{(1-\alpha)b(\alpha)} \qquad (9)$$

where $a(K)$ and $b(\alpha)$ are a functions of K or α, the data window $h(i)$, and the frame rate R. Typical values are given for a Hamming window of length 256 in the Appendix. Thus, regardless of what kind of smoothing is employed the noise power estimate in eq. (5) can be modelled as the minimum of D (approximately) chi-square distributed power estimates $P_x(\lambda, k)$.

3.1 Minimum of uncorrelated power estimates

In this section we assume that the minimum power estimate $P_{min}(\lambda, k)$ is based on D independent power estimates $P_x(\lambda, k)$. This is clearly not the case if we use successive smoothed estimates $P_x(\lambda, k)$ but with non recursive smoothing and a suitable decimation we will be able to approximate this condition. The density of the minimum of D independent power estimates is given by [7]

$$f_{P_{min}}(y) = D \cdot (1 - F_{P_x}(y))^{D-1} \cdot f_{P_x}(y) \qquad (10)$$

where $F_{P_x}(y)$ denotes the distribution function of the chi-square density which becomes after repeated partial integrations of eq. (8)

$$F_{P_x}(y) = 1 - e^{-yN/2\sigma^2} \cdot \sum_{m=0}^{N/2-1} \frac{1}{m!} \cdot \left(\frac{yN}{2\sigma^2}\right)^m \cdot U(y) \qquad (11)$$

We use eq. (10) to compute mean and variance of the minimum power estimate. Note that mean and variance of the minimum estimate are proportional to $\sigma^2(k)$ and $\sigma^4(k)$, respectively [8]. It follows that the bias of the minimum subband power estimate is proportional to the noise power $\sigma^2(k)$ and that the bias can be compensated by multiplying the minimum estimate with the inverse of the mean computed for $\sigma^2(k) = 1$

$$omin = \frac{1}{E\{P_{min}\}_{|\sigma^2(k)=1}} \qquad (12)$$

The upper graph in Figure 3 plots the mean $E\{P_{min}\}$ versus

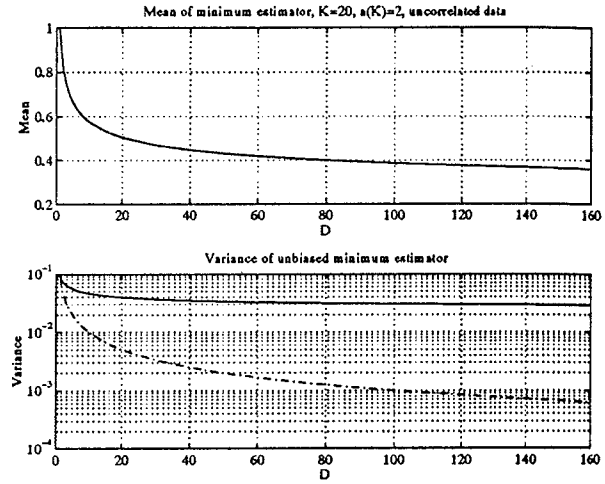

Figure 3: Mean (upper graph) and variance (lower graph, solid line) of minimum power estimate computed for uncorrelated data, $\sigma^2(k) = 1$, K=20, and D=1–160 (Dashed line gives minimum variance achieved by non recursive smoothing over $K \cdot D$ segments).

D for $\sigma^2(k) = 1$, K=20, and a(K)=2. To evaluate the statistical efficiency we compare the variance of the unbiased estimate $P_n(\lambda, k) = omin \cdot P_{min}(\lambda, k)$ with the variance of a periodogram smoothed over $K \cdot D$ data segments ('conventional' noise estimation). In this comparison the minimum subband power estimate and the conventional smoothed periodogram use the same amount of data. The lower graph in Figure 3 plots the variance for $\sigma^2(k) = 1$. It is immediately obvious that the variance of the minimum estimate is much larger than the variance of the smoothed periodogram. Since we typically use D=80–140 the minimum power estimation method is certainly not attractive for uncorrelated data. We now turn to case of correlated data and show that there the situation is quite different.

3.2 Minimum of correlated power estimates

Clearly, successive values of $P_x(\lambda, k)$ are correlated. In the case of correlated data there is, however, no closed form solution for the probability density of the minimum or for its mean and variance available. We therefore generated data of variance $\sigma^2(k) = 1$, computed the smoothed periodogram (eq. (4)), and evaluated the mean and the variance of the minimum estimate. Figure 4 presents these simulation results for recursive smoothing with $\alpha = 0.95$. Again, the dashed line in the lower graph gives the variance for non recursive smoothing of D successive values of $P_x(\lambda, k)$. In the region of interest, i.e. D=80–140, we now have only a small deviation of the variance of the unbiased minimum estimator with respect to the variance achieved by a conventional estimator.

During speech activity we have noise power information only within the narrow valleys of the short time power of the speech signal. Thus, the effective window length D is much shorter than for speech pause. The reduced window length and a possible distortion of the noise power estimate by low energy phonemes imply to use a smaller overestimation factor omin during speech activity. However, as the experiments show, an overestimation factor adjusted according to Figure 4 gave good results also during speech activity.

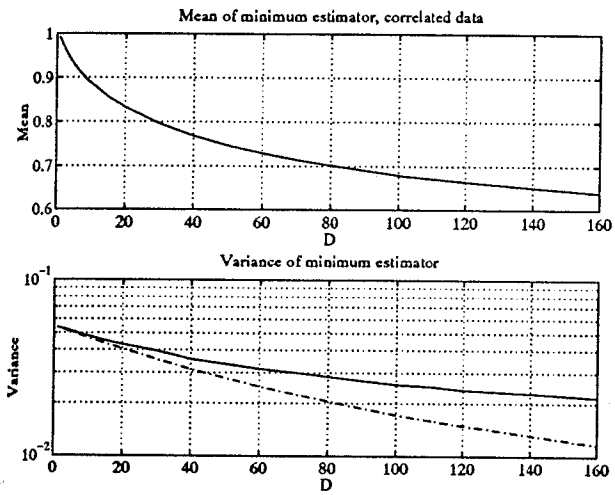

Figure 4: Mean (upper graph) and variance (lower graph, solid line) of minimum power estimate computed for correlated data with $\sigma^2(k) = 1$, $\alpha = 0.95$, and D=1–160 (Dashed line gives variance for non recursive smoothing of D short time power estimates).

4. Experimental Results

The proposed algorithm was compared to a 'conventional' spectral subtraction algorithm which uses the same subtraction rule and approximately the same amount of data to estimate the subband noise power. The 'conventional' spectral subtraction was equipped with an ideal speech activity detector (i.e. manually segmented speech files) and a noise power estimator equivalent to eq. (4). The speech material consisted of 8 phonetically balanced German speech samples recorded by two male and two female speakers. After adding car noise at two different levels the segmental SNR during speech activity was 6dB and 0dB.

After the first experiments it became clear that the most crucial compromise is that of power smoothing (controlled by α) versus window length for minimum search (controlled by D). As it is evident from Figures 3 and 4 smoothing is a statistically more efficient procedure than the minimum search. With too much smoothing, however, the valleys of power (see Figure 1) are not pronounced enough to warrant reliable noise estimates. We therefore use an additional first order network with a pole at 0.9 to smooth the minimum estimates. With this additional measure and the other parameters set to $D = 100$, $\alpha = 0.95$, $\gamma = 0.9$, $\omega_{min} = 1.5$ the speech quality and noise suppression was almost identical to the 'conventional' spectral subtraction. Occasionally, however, a phoneme is somewhat attenuated. The litmus test for our noise estimation method is of course non stationary noise and high speech activity. To investigate the performance under non stationary noise conditions the noise signal was multiplied with a increasing or decreasing ramp function prior to adding it to the clean speech signal. In these experiments the SNR of the speech samples thus varied between 6dB and 0dB and as expected, the performance of the proposed method was indeed superior to the performance of the 'conventional' spectral subtraction.

5. Conclusion

The proposed minimum subtraction method eliminates the need for a speech activity detector by exploiting the short time characteristics of speech signals. For stationary noise the performance of our method is very close to the performance of spectral subtraction with conventional noise power estimation and ideal speech activity detection. For non stationary noise the method is at a clear advantage. The theoretical analysis shows that for typical parameter settings, the variance of the minimum estimator is less than twice as large as the variance of the conventional noise power estimator. However, more experiments with different speaking situations and different languages are necessary to improve the trade-off between the smoothing constant and the window length for minimum search.

Appendix: Equivalent Degrees of Freedom (see [6])

K	a(K)	N_{nrec}	α	b(α)	N_{rec}
10	1.98	10.1	0.7	1.75	6.48
20	2.04	19.6	0.9	1.98	18.2
100	2.09	95.6	0.95	2.04	38.2

Table 1: $a(K)$, N_{nrec}, $b(\alpha)$ and N_{rec} for various values of K and α computed for a Hamming window with $W_{DFT} = 256$ and R=64.

References

[1] J. Deller, J. Proakis, and J. Hansen: "Discrete-Time Processing of Speech Signals", Macmillan, 1993.

[2] P. Vary: "Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits -", Signal Processing 8, Elsevier Science Publishers B.V. (North-Holland), pp. 387-400, 1985.

[3] R. Crochiere and L. Rabiner: "Multirate Digital Signal Processing", Prentice Hall, 1983.

[4] M. Berouti, R. Schwartz, and J. Makhoul: "Enhancement of Speech Corrupted by Acoustic Noise", Proc. IEEE Conf. ASSP, pp. 208-211, April 1979.

[5] R. Martin: "An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals", Proc. EUROSPEECH '93, pp. 1093-1096, Berlin, September 21-23, 1993.

[6] P. Welch: "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms", IEEE Trans. Audio and Electroacoustics, Vol. 15, No. 2, June 1967.

[7] A. Papoulis: "Probability, Random Variables, and Stochastic Processes", 2nd ed., McGraw-Hill, 1984.

[8] H. David: "Order statistics", 2nd ed., Wiley, New York, 1981.