Contents

10

Spectral Subtraction

CHAPTER 9 PRESENTED several different approaches to noise suppression based on how the noise spectrum or envelope modulation differed from that of speech. This chapter concentrates on spectral subtraction. The basic idea is to estimate the noise spectrum, and then subtract it from the noisy speech spectrum to get an improved estimate of the original speech spectrum. The estimated speech spectrum is then used to reconstruct an enhanced signal waveform.

The speech is assumed to be degraded by additive noise, as shown in Figure 10-1. In an ideal world, we would estimate the noise signal and subtract it from the noisy speech to reconstruct the clean speech signal. But separating the noise from the noisy speech in this way requires access to the original noise signal, and the noise signal is generally not available. All that is available is the noisy speech. The best that can be done is to estimate the statistics of the noise from observations of the noisy speech, and to then try to produce an enhanced signal that has statistics closer to the original clean speech signal. This process is shown in the block diagram of Figure 10-2 (reprinted from Chapter 9), where the estimated noise parameters are used to adjust the gain in each of the analysis frequency bands. One can think of spectral subtraction as starting with a distribution of noisy speech magnitude samples measured over some time interval, and then adaptively adjusting the gain in each frequency band so that the distribution of the processed magnitude samples more closely represents that of the clean speech.

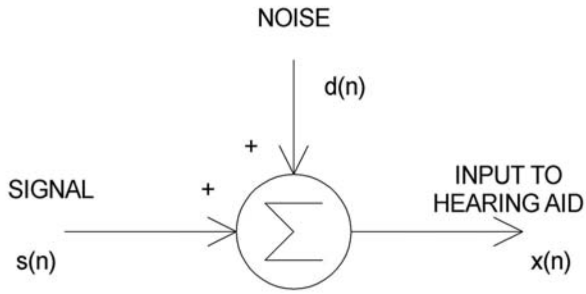

Figure 10-1. Block diagram illustrating additive noise being combined with the speech signal.

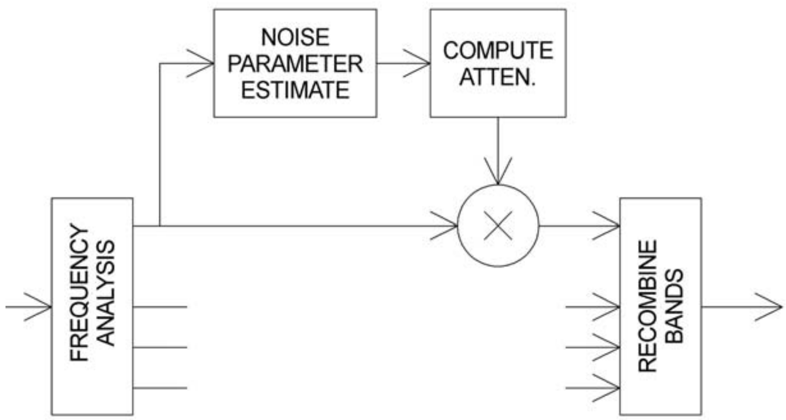

Figure 10-2. Block diagram of a generic multichannel single-microphone noise-suppression system.

The chapter begins with a section on noise estimation, as the accuracy of the spectral subtraction depends on the accuracy of the estimated noise properties. One of the earliest approaches to noise suppression was the Wiener filter, which is described next. The Wiener filter assumes that the speech and noise are stationary (signal statistics such as the signal power do not change over time), which is not the case in the real world. The adaptive Wiener filter is intended to deal with slowly fluctuating speech and noise, and is closely related to spectral subtraction. Spectral subtraction is discussed next, and several different forms of the basic signal-processing strategy are presented. The chapter concludes with a summary of perceptual experiments evaluating the effectiveness of spectral subtraction in improving the intelligibility and quality of noisy speech.

Noise Estimation

Spectral subtraction requires an accurate estimate of the noise power in each frequency band. Two general approaches to noise estimation can be found in the literature. Early work concentrated on identifying signal segments containing voiced speech, termed a voice activity detector (VAD). The noise power estimate is held constant during the voiced speech segments and updated during the nonspeech segments. The reader is referred to Marzinzik and Kollmeier (2002) for a review of this approach. Recent research has concentrated on continuously updating the noise power estimate without using a VAD. These newer approaches are attractive for hearing aids as they do not require the extra computations needed for a VAD. Three noise-estimation approaches that do not use a VAD are discussed in this section.

All of the noise-estimation schemes, those that use a VAD and those that do not, are based on the assumption that the noise is stationary, that is, the noise statistics do not change over time. The noise power estimate is updated when conditions are favorable for doing so, and the estimate is held with minimal change during those signal intervals that are more likely to be speech. The stationarity assumption is thus an assumption that the noise power estimate will be valid even during those time intervals where it is not being updated. If the noise level is fluctuating rapidly, this stationarity assumption will be violated and the error in the estimated noise power will reduce the effectiveness of the spectral subtraction algorithm.

VALLEY DETECTION

The simplest approach to noise power estimation is to use valley detection in each frequency band. In valley detection the estimated signal level is decreased rapidly when the signal level decreases but increased slowly when the signal level increases. The valley detector is thus the opposite of the peak detector given by *Eq. (8.1)*. The valley detector output is given by:

$$
\begin{aligned}
&if\ |x(n)| \le v(n-1)\\
&\quad v(n) = \alpha v(n-1) + (1-\alpha)|x(n)|\\
&else \qquad\qquad\qquad\qquad\qquad\qquad \textit{Equation (10.1)}\\
&\quad v(n) = (1+\frac{1}{\beta})v(n-1)\\
&end
\end{aligned}
$$

where the input signal is $x(n)$ and the valley detector output is $v(n)$. The valley detector uses a fast attack time constant α when tracking decreases in the signal level and a slow release time constant β when tracking increases. In a multichannel system, a valley detector would be implemented in each frequency band.

An example of valley detection is presented in Figure 10-3 for a segment of speech. The noise is multitalker babble at a SNR of 20 dB. The noise alone is present for the first 250 msec of the signal, after which the speech is present. The figure gives the output of a frequency band centered at 728 Hz and having a bandwidth of about 340 Hz. The light gray curve is the signal envelope given by computing the signal power in 1-msec segments. The output of a peak detector having an attack time of 5 msec and a release time of 70 msec is given by the dashed black line, and the output of a valley detector having an attack time of 50 msec and a release time of 500 msec is given by the solid black line. The valley detector output rises slowly during the initial noise-only portion of the signal. It then drops slightly during the first pause in the speech, rises again during the following voiced segment, and proceeds to fall and rise in synchrony with the speech during the remainder of the utterance. Longer attack and release times for the valley detector would smooth out these fluctuations; for example, Arslan et al. (1995) recommend adjusting the valley detector to increase at a rate of 3 dB per second for increasing signal levels and to decrease at a rate of 12 dB per second for decreasing signal levels. But even with the longer time constants, the same behavior of rising during voiced speech and falling during speech pauses would be present.

MINIMA STATISTICS

A more elegant approach to estimating the noise power is to test each sample of the signal envelope in each frequency band. The signal sample is used to update the estimated noise power if the signal sample has a

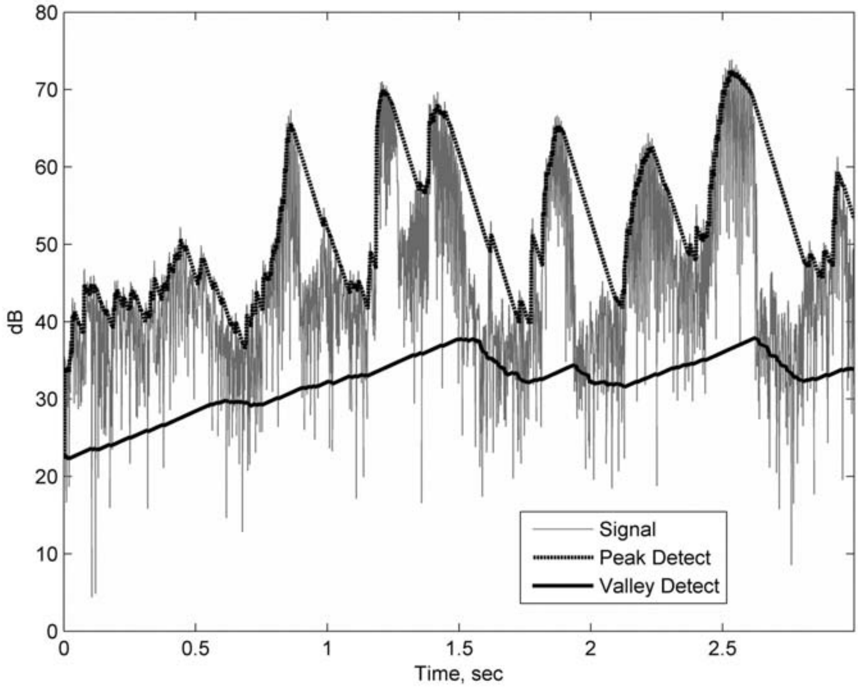

Figure 10-3. Peak detection (attack time 5 msec, release time 70 msec) and valley detection (attack time 50 msec, release time 500 msec) for a segment of speech in additive low-pass filtered noise at an SNR of 20 dB. Results for the 728-Hz band are shown.

higher probability of being noise than speech. A reasonable assumption is that the speech samples in dB are drawn from a Gaussian probability density function, and that the noise samples in dB are drawn from a Gaussian probability density function having a lower mean. Figure 10-4 gives a histogram of samples in dB taken every millisecond from the end of the noisy speech signal plotted in Figure 10-3 for the 728-Hz frequency band. An approximate fit of a pair of Gaussian probability density functions to this histogram is shown in Figure 10-5. The noise probability density function has a mean of 40 dB SPL and a standard deviation of 8 dB, and the speech probability density function has a mean of 60 dB SPL and a standard deviation of 7 dB. The ratio of the amplitudes of the two probability density functions represents the relative occurrence of the noise and speech samples in noisy speech.

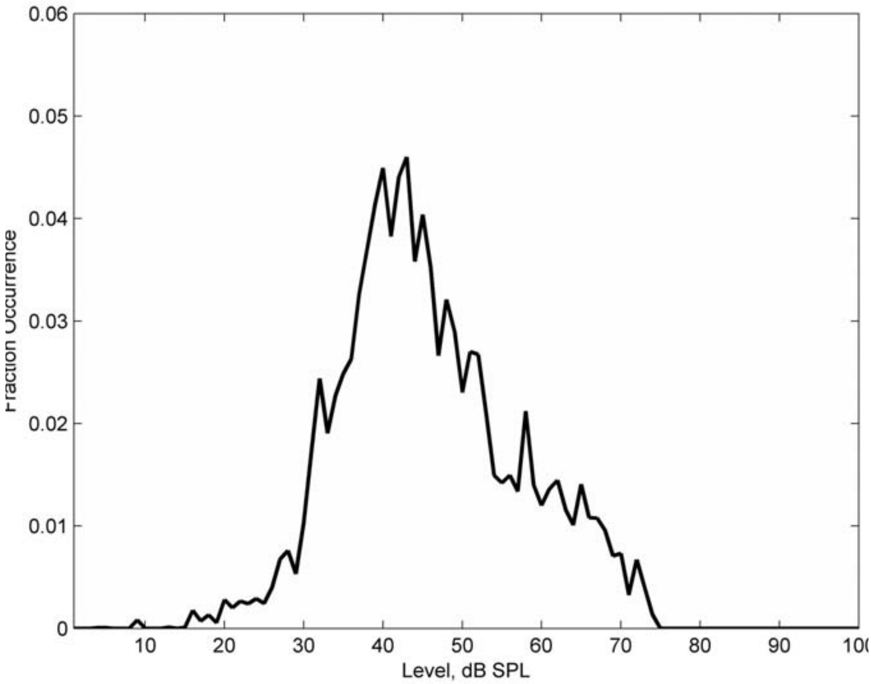

Figure 10-4. Log-level histogram of the noisy speech signal in the 728-Hz band at the end of the segment shown in Figure 10-3.

The noise power estimate is updated using an adaptive averaging operation (Cohen & Berdugo, 2002; Doblinger, 1995; Martin, 2001). Let $|N(k,m)|$ be the noise magnitude estimate for frequency band k and processing block m, and let $|X(k,m)|$ be the incoming noisy signal magnitude. The noise update is given by:

$$|N(k,m)| = \mu(k,m)|N(k,m-1)| + [1-\mu(k,m)]|X(k,m)|$$

Equation (10.2)

The averaging time constant $\mu(k,m)$ depends on the amplitude of the signal envelope sample $|X(k,m)|$. A fast averaging time is used for low-intensity samples that have a high probability of being noise, and a slow averaging time is used for high-intensity samples that have a high prob-

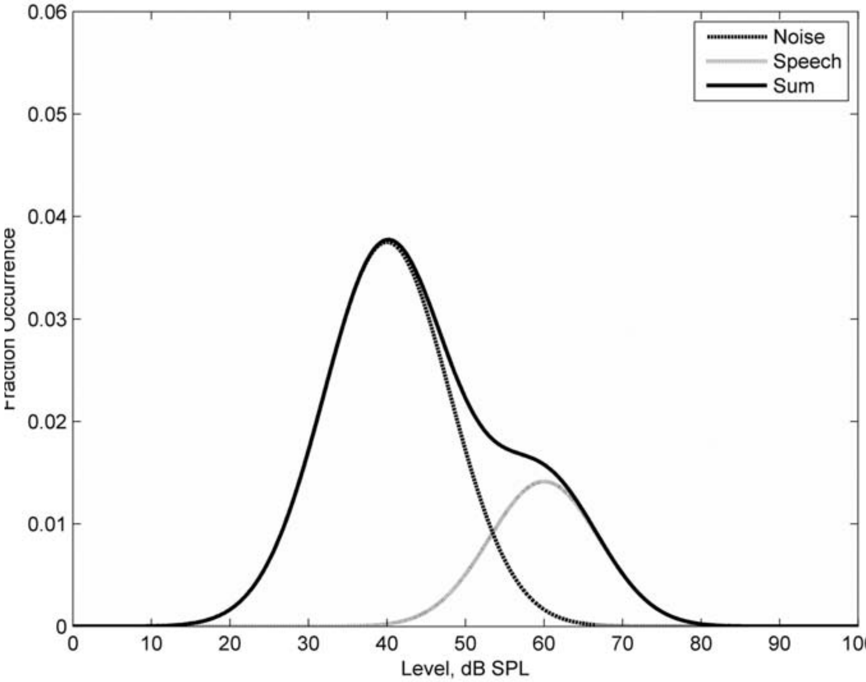

Figure 10-5. Gaussian probability density functions approximating the distribution of signal levels for speech and noise in the histogram plotted in Figure 10-4. The noise distribution has a mean of 40 dB SPL and a standard deviation of 8 dB. The speech distribution has a mean of 60 dB SPL and a standard deviation of 7 dB.

ability of being speech. Intermediate time constant values are used for intermediate signal envelope amplitudes.

A simplified version of this approach is the algorithm described by Hirsch and Ehrlicher (1995). Let $|N(k,m)|$ again be the noise magnitude estimate for frequency band k and processing block m, and let $|X(k,m)|$ be the incoming noisy signal magnitude. Their algorithm is given by:

$$\begin{aligned}
&\textit{if } |X(k,m)| > b|N(k,m-1)| \\
&\quad |N(k,m)| = |N(k,m-1)| \\
&\textit{else} \\
&\quad |N(k,m)| = a|N(k,m-1)| + (1-a)|X(k,m)| \\
&\textit{end}
\end{aligned} \qquad \textit{Equation (10.3)}$$

This algorithm replaces the adaptive averaging time constant used in *Eq. (10.2)* with just two values: a short time constant for samples assumed to be noise or an infinitely long time constant for samples assumed to be speech. The selection of which time constant depends on the signal level compared to an adaptive threshold. The algorithm averages the incoming signal into the noise estimate if the signal power is close to or less than that of the existing noise estimate, and holds the new noise estimate to the previous value if the signal power exceeds the noise estimate by the preset threshold b.

The noise level estimate produced by the Hirsh and Ehrlicher (1995) algorithm is plotted in Figure 10-6 for the same noisy speech segment plotted in Figure 10-3. The low-pass filter time constant in *Eq. (10.3)* was chosen to be 200 msec, and the threshold value b was set to 2. The estimated noise level quickly adapts to the actual noise level during the initial noise-only portion of the noisy speech signal. The noise estimate is then held constant during the more intense speech portions of the signal, but continues to adapt during the less intense portions of the signal where the babble dominates. The result is a nearly constant noise level estimate despite the fluctuations in the speech amplitude.

HISTOGRAM

It is clear from Figures 10-4 and 10-5 that the histogram of signal levels plotted in dB contains useful information about the noise amplitude. The histogram bin containing the greatest number of samples is a good estimate of the noise level if there are more noise samples than speech samples. As the speech and noise levels change over time, the histogram must also track these level changes. The histogram should therefore be "leaky," with the importance of old signal samples slowly decaying and new samples given the greatest weight. Let $H(k,m)$ be the current histogram of the signal levels in dB in frequency band k processing block m. The histogram contains L bins, each encompassing a different signal level in dB, so an individual bin would be denoted by $h(k,m,l)$. Each bin contains the relative count of the number of occurrences of that level in the signal. The noisy signal level $|X(k,m)|$ in dB is then assigned to the closest histogram bin; call this bin index i. The algorithm for updating the histogram is then:

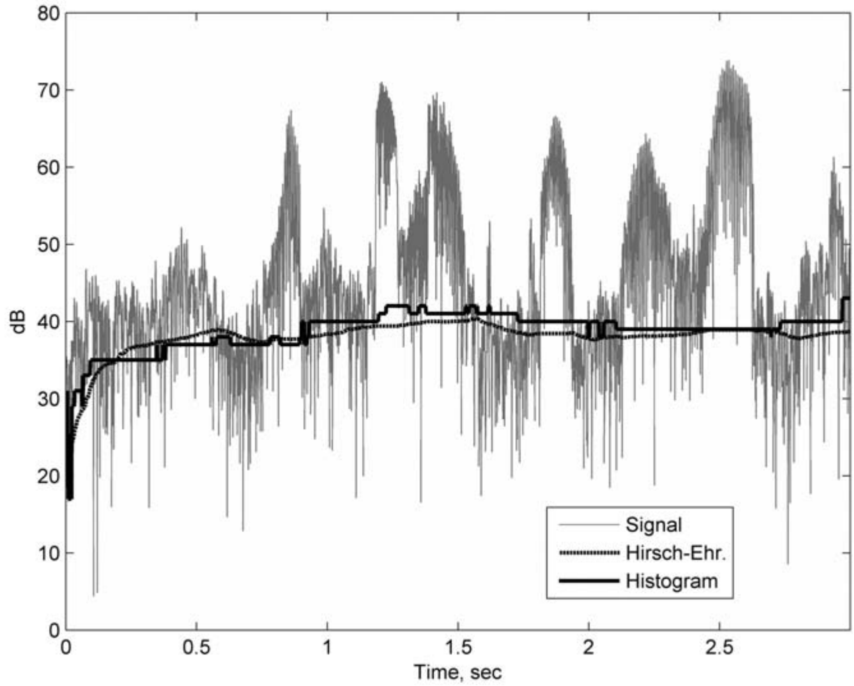

Figure 10-6. Noise detection using the Hirsch-Ehrlicher (1995) algorithm and a log-level histogram approach for a segment of speech in additive low-pass filtered noise at a SNR of 20 dB. The 728-Hz band is shown.

$$\begin{aligned} h(k,m,l) &= \eta h(k,m-1,l) \text{ for } 1 \le l \le L \\ h(k,m,i) &= h(k,m,i) + (1-\eta) \end{aligned} \qquad \textit{Equation (10.4)}$$

where $\eta < 1$. The contents of all of the histogram bins are decayed with a long time constant, thus reducing the influence of old samples, and then the histogram bin corresponding to the current signal envelope sample is incremented. The plot of Figure 10-4 is actually the histogram updated using *Eq. (10.4)* with a time constant of 1 sec and plotted at the end of the 3 seconds of noisy speech

To use the histogram for noise estimation, one must identify the peak of the noise distribution given that only the distribution of the noisy speech is available. Several different approaches have been proposed (McAulay & Malpass, 1980; Stahl et al., 2000; Van Compernolle, 1987).

A simple method is to compute the mean of the time-varying histogram computed using *Eq. (10.4)* for each processing block, and then find the peak at or below the mean. If only noise is present, the distribution of the signal levels in dB can be approximated by a single Gaussian distribution, and the peak will be near the mean of the histogram. Selecting the peak of the histogram will therefore find the peak of the underlying distribution and will give the mean noise level $\left|N(k,m)\right|$. If speech is present, the distribution will be similar to the sum of two Gaussians, as shown in Figure 10-5, and the mean of the histogram will lie above the peak of the noise distribution. Searching below the mean for the peak will again return the mean noise level.

The noise level estimate produced by this histogram approach is plotted in Figure 10-6 along with the estimate produce by the Hirsch-Ehrlicher (1995) algorithm. The histogram bins are 1-dB wide and no smoothing has been applied to the noise estimate, so the histogram noise level is quantized in 1-dB steps. There appears to be very little difference between the histogram estimate and that produced by the Hirsch-Ehrlicher algorithm for this signal. The estimate rises quickly to the actual noise level during the initial noise-only portion of the noisy speech signal, and then holds a relatively constant noise estimate despite the fluctuations of the speech.

Wiener Filter

The Wiener filter (Wiener, 1949) is one of the oldest techniques for suppressing noise in a noisy signal. The Wiener filter requires separate estimates of the speech and noise powers. The filter design also assumes that the speech and noise are stationary. In practical terms, this assumption means that the Wiener filter can be designed for the average speech and noise, but cannot take into account the speech or noise fluctuations.

The basic idea for the Wiener filter is illustrated in the block diagram of Figure 10-7. The speech signal $s(n)$ is corrupted by additive noise $d(n)$ to form the input signal $x(n) = s(n) + d(n)$. This noisy signal is then filtered by the Wiener filter $G(f)$ to give the output $y(n)$. The criterion for the design of the Wiener filter is to minimize the mean-squared error between the clean input and the filtered output:

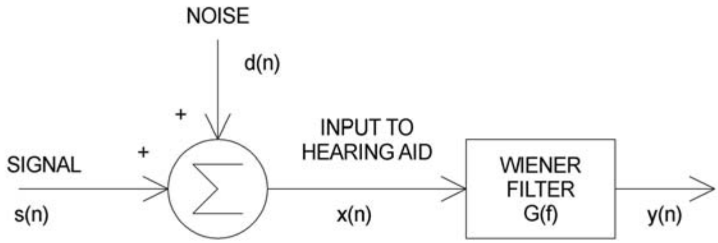

Figure 10-7. Block diagram showing a Wiener filter used to process the input of a hearing aid.

$$\varepsilon = \sum_n \left[s(n) - y(n)\right]^2 \quad \text{Equation (10.5)}$$

The Wiener filter thus filters the noisy speech *x(n)* to form the closest possible match to the clean speech *s(n)*. The solution to the filter design problem in the continuous frequency domain is:

$$G(f) = \frac{\left|S(f)\right|^2}{\left|S(f)\right|^2 + \left|D(f)\right|^2} \quad \text{Equation (10.6)}$$

where *S(f)* is the long-term average speech spectrum and *D(f)* is the long-term average noise spectrum. In practice the speech and noise spectra are not available, so the Wiener filter is usually approximated from the noisy signal spectrum *X(f)* and the estimated noise spectrum *N(f)*:

$$G(f) \approx \frac{\left|X(f)\right|^2 - \left|N(f)\right|^2}{\left|X(f)\right|^2} \quad \text{Equation (10.7)}$$

The Wiener filter maximizes the signal-to-noise ratio (SNR) of the noisy speech averaged over the entire signal. However, the minimum mean-squared error criterion used in the Wiener filter design does not directly involve any aspects of auditory perception. The assumption in applying the Wiener filter to speech is that the mathematical optimization will also lead to improvements in speech perception. The Wiener filter can be effective if the noise is stationary and concentrated in narrow frequency regions or in regions that do not contain much speech energy (Lim, 1986). In most problems of interest, however, the speech and noise both fluctuate in amplitude and the noise spectrum substantially overlaps the speech spectrum. Signal processing for these conditions requires a Wiener filter that varies over time to reflect the effect of the noise on different speech sounds (Levitt et al., 1993). The time-varying Wiener filter is discussed in the next section as it can be considered to be a form of spectral subtraction.

Spectral Subtraction

Spectral subtraction refers to a family of related noise-suppression algorithms. In these algorithms the estimated noise spectrum (or a function of the noise spectrum) is subtracted from the noisy speech spectrum to produce an estimate of the clean speech spectrum. Early forms of spectrum subtraction, including adaptive Wiener filters, magnitude spectrum subtraction, and power spectrum subtraction are described in the next section. These forms of spectral subtraction had problems, in particular, residual noise components termed "musical noise" comprising sinusoids at random frequencies that appear and then disappear in the processed speech (Boll, 1979; Cappé, 1994) as the amount of signal attenuation fluctuates. Improvements in spectral subtraction, described next, have concentrated on improving the noise estimation and on reducing the musical noise in the processed speech.

CLASSICAL APPROACHES

Spectral subtraction works on the short-time spectrum of the noisy speech, as shown in Figure 10-2. The incoming signal is divided into blocks. Each block is windowed and the short-time FFT is computed from the windowed data sequence. The magnitude spectrum of the noise is estimated, as described above, and the noise-suppression gain as a function of frequency is computed from the magnitude spectrum of the noisy signal and the estimated noise magnitude spectrum. The gain function $G(f)$ is applied to the noisy spectrum to give the estimated clean speech spectrum, and the modified spectrum is inverse transformed to give the time waveform of the enhanced signal. In general the signal blocks have an overlap of 50% and the processed waveform is synthesized using the overlap-add procedure. The enhanced signal has a modified envelope, but retains the noisy phase.

The Wiener filter given by *Eq. (10.7)* is based on the long-term average estimated noise and noisy speech spectra. An adaptive Wiener filter can be created by replacing the long-term average noisy speech spectrum *X(f)* with the short-term spectrum *X(k,m)* and the long-term noise estimate *N(f)* with an estimate *N(k,m)* that is allowed to track changes in the noise level. The resultant gain $G_W(k,m)$ for frequency band *k* and block *m* can be written as:

$$G_W(k,m) = \frac{|X(k,m)|^2 - |N(k,m)|^2}{|X(k,m)|^2} = 1 - \frac{|N(k,m)|^2}{|X(k,m)|^2} \quad \textit{Equation (10.8)}$$

Another approach is power spectral subtraction (Lim & Oppenheim, 1979). Assume that the noise and speech are uncorrelated. The noisy speech power spectrum can then be represented as the sum of the speech power spectrum and the noise power spectrum, $|X(k,m)|^2 = |S(k,m)|^2 + |D(k,m)|^2$. Using the estimated noise spectrum to replace the actual noise spectrum leads to

$$|\hat{S}(k,m)|^2 \approx |X(k,m)|^2 - |N(k,m)|^2 \quad \textit{Equation (10.9)}$$

The gain needed to get the noisy speech power spectrum to match the clean speech power spectrum is then

$$G_P^2(k,m) = \frac{|\hat{S}(k,m)|^2}{|X(k,m|^2} = 1 - \frac{|N(k,m)|^2}{|X(k,m)|^2} \quad \textit{Equation (10.10)}$$

The gain for power spectral subtraction is thus the square root of the Wiener filter gain:

$$G_P(k,m) = \sqrt{1 - \frac{|N(k,m)|^2}{|X(k,m)|^2}} \quad \textit{Equation (10.11)}$$

Another approach is magnitude spectral subtraction (Boll, 1979). The gain for magnitude spectral subtraction is given by:

$$G_M(k,m) = 1 - \frac{|N(k,m)|}{|X(k,m)|} \quad \textit{Equation (10.12)}$$

In all of the spectral subtraction gain equations *Eq. (10.8)*, *Eq. (10.11)*, and *Eq. (10.12)*, it is possible for the estimated noise level to be greater than the short-term spectrum for the signal block. When this occurs, the spectral subtraction would compute a negative gain, which is not physically meaningful. All of the algorithms therefore replace a negative gain with zero.

The gain for these three spectral subtraction approaches is plotted in Figure 10-8 as a function of the SNR. For low noise levels (high SNR), the algorithms all give gains near 0 dB. If the noise level is low, there is no need to attenuate the signal. As the noise level increases the signal attenuation also increases, but the algorithms differ in the degree to which the noisy signal is attenuated as a function of SNR. The magnitude spectral subtraction provides the greatest attenuation, whereas the power spectral subtraction provides the least.

GENERAL EQUATION

The spectral subtraction given by *Eqs. 10.8*, *10.11*, and *10.12* can be combined into a single general formulation (Berouti et al., 1979; Virag, 1999):

$$G(k,m) = \begin{cases} \left[1 - \alpha \dfrac{|N(k,m)|^{\gamma}}{|X(k,m)|^{\gamma}}\right]^{1/\delta} & \textit{if } \dfrac{|N(k,m)|^{\gamma}}{|X(k,m)|^{\gamma}} < \dfrac{1}{\alpha+\beta} \\ \left[\beta \dfrac{|N(k,m)|^{\gamma}}{|X(k,m)|^{\gamma}}\right]^{1/\delta} & \textit{otherwise} \end{cases}$$

Equation (10.13)

All of the classical spectral subtraction algorithms described above use $\alpha = 1$ and $\beta = 0$ when expressed in this formulation. The Wiener filter is

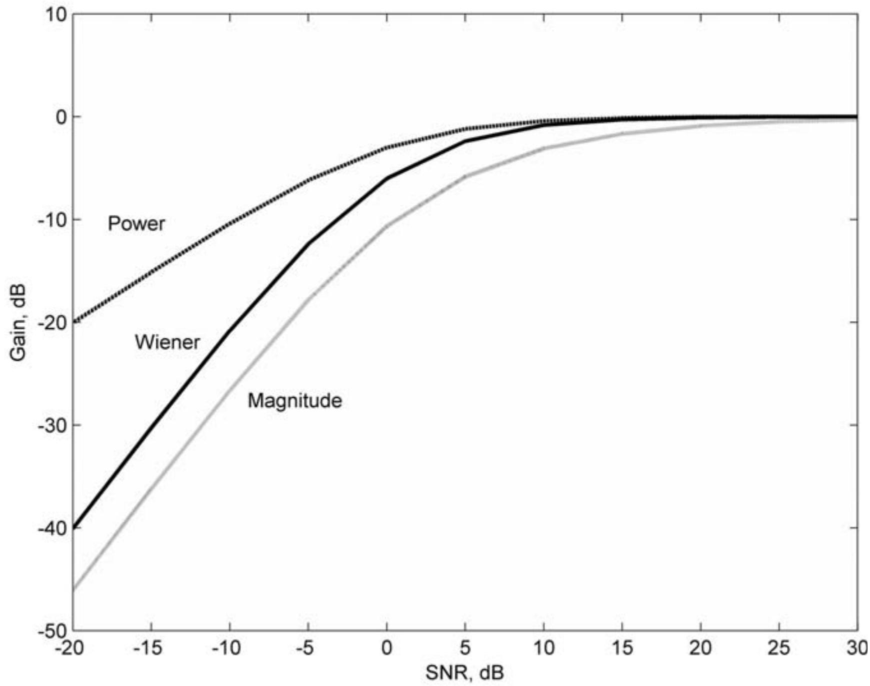

Figure 10-8. Gain as a function of the signal-to-noise ratio (SNR) for different versions of spectral subtraction.

given by $\{\gamma = 2, \delta = 1\}$, power spectral subtraction by $\{\gamma = 2, \delta = 2\}$, and magnitude spectral subtraction by $\{\gamma = 1, \delta = 1\}$.

The general formulation adds two additional factors. One is an oversubtraction factor $\alpha \geq 1$ and the other is the attenuation floor $0 \leq \beta < 1$. The oversubtraction factor α increases the signal attenuation at poor SNRs beyond the amount shown in Figure 10-8, which reduces the amplitude of the background noise when no speech is present. The attenuation floor β limits the maximum attenuation that the algorithm produces. Increasing β increases the background noise level but reduces the gain fluctuations that can contribute to musical noise at low signal levels. Setting β within the range of -10 to -20 dB re: 1 can provide audible noise suppression while reducing the amount of musical noise. A further variation is to make the factor α time-dependent. Virag (1999) has proposed making α a function of the estimated SNR; α is large at poor SNRs when speech is absent, thus increasing the background noise suppression, and α is reduced at good SNRs when speech is present to reduce the amount of distortion introduced by the spectral subtraction. In her implementation α ranged from 6 at poor SNRs to 1 at very good SNRs.

An alternative technique to reduce musical noise is to average the spectral subtraction gains across signal blocks. Gustafsson et al. (2001) propose using an adaptive averaging time constant for the suppression gains. If the spectrum in the current signal block is similar to that of the previous block, a long averaging time constant is used on the assumption that the signal is noise; noise spectra will be very similar from one block to the next. If the spectrum in the current block differs significantly from that in the previous block, a short averaging time constant is used to allow the suppression gain to change rapidly in response to the presence of speech.

NONLINEAR EXPANSION

An approach related to spectral subtraction is to compute a suppression gain based on the signal and estimated noise powers. One such function that has been successfully used (Clarkson & Bahgat, 1991; Eger et al., 1984; Tsoukalas et al., 1997) is given by:

$$G(k,m) = \frac{1}{1+\nu\left[\frac{|N(k,m)|}{|X(k,m)|}\right]^{\gamma}} \quad \textit{Equation (10.14)}$$

The family of gain curves given by *Eq. (10.14)* is plotted in Figure 10-9. The curves are drawn for different combinations of γ and ν. The value of ωcontrols the maximum attenuation provided by the expansion when only noise is present; increasing ν increases the maximum attenuation. The value of γ controls the rate at which the gain changes with SNR; increasing γ reduces the SNR range in dB needed to go from minimal attenuation to nearly full attenuation. For curve (i) $\gamma = 1$ and $\nu = \Delta 10$, (ii) $\gamma = 2$ and $= \sqrt{}10$, (iii) $\gamma = 1$ and $\nu = 10$, (iv) $\gamma = 2$ and $\nu = 10$, (v) $\gamma = 2$ and $\nu = 100$.

This family of gain functions has the advantage that the gain remains constant at poor SNRs, whereas the gain curves plotted in Figure 10-8 continue to decrease forever as the SNR decreases. If a signal segment is just noise, the SNR will be negative and the gain from *Eq. (10.14)*

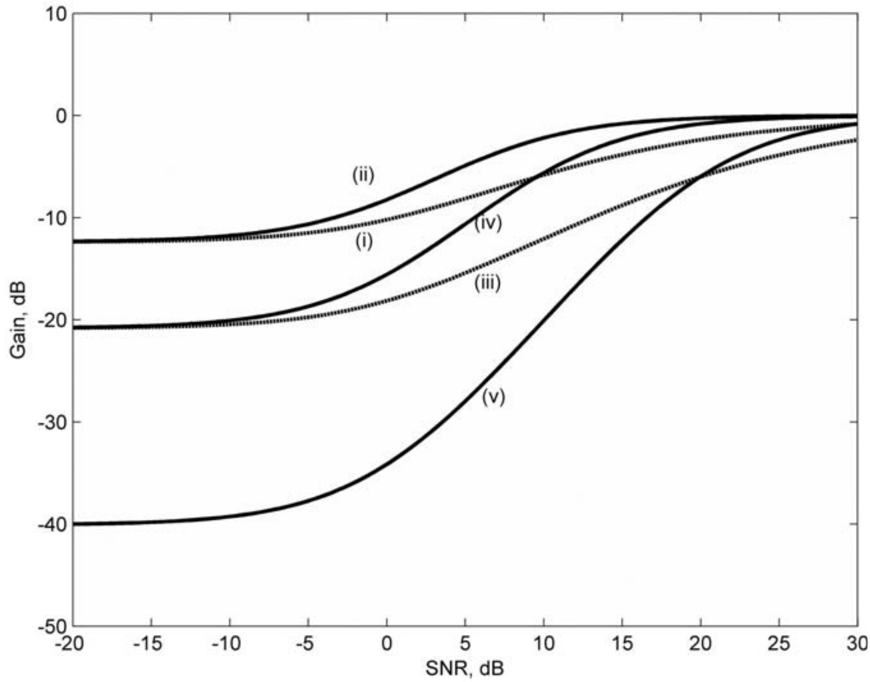

Figure 10-9. Gain versus the instantaneous SNR for the nonlinear expansion approach. The curves are for different settings of the gain rule given by *Eq. (10.14)* and are described in the text.

will barely fluctuate. This constant gain for signals dominated by noise will minimize the musical noise and gain-modulation artifacts in the processed signal.

EPHRAIM-MALAH ALGORITHM

The noise suppression gain given by *Eq. (10.13)* depends on the instantaneous SNR given by the ratio of noisy signal $|X(k,m)|$ in block m to the estimated average noise power $|N(k,m)|$. If the SNR is good $|X(k,m)| >> |N(k,m)|$, and fluctuations in the signal level will only cause small changes in the gain when expressed in dB re: 1. However, as the SNR approaches 0 dB, we get $|X(k,m)|/|N(k,m)| \approx 1$. If $\alpha = 1$, as in the classical spectral subtraction algorithms, the gain will be approximately 0. Small changes in $|X(k,m)|$ will then cause large changes in the gain expressed in dB. These large relative gain fluctuations are one of the causes of the musical noise artifacts in spectral subtraction (Cappé, 1994; Marzinzik, 2000).

An alternative approach to spectral subtraction is to assume that the noise samples come from a Gaussian probability distribution and that the speech samples come from a second independent Gaussian distribution. The gain is set close to 0 dB if a given sample appears to come from the speech distribution, and the gain is reduced if the sample appears to come from the noise distribution (Ephraim & Malah, 1984; McAulay & Malpass, 1980; Wolfe & Godsill, 2001). The resultant suppression gain rule depends on the SNR, and lies between the power spectral subtraction and Wiener filter gain rules plotted in Figure 10-8.

Where these statistical approaches differ from classical spectral subtraction is in how the SNR is calculated. The estimated SNR is a combination of two terms, the a priori SNR and the a posteriori SNR. Assume that the clean speech spectrum $S(k,m)$ and the noise spectrum $D(k,m)$ are known. The a priori SNR is then given by:

$$\varepsilon_{priori}(k,m) = \frac{E[|S(k,m)|^2]}{E[|D(k,m)|^2]} \quad \textit{Equation (10.15)}$$

where E[ ] is the expectation operation, typically implemented as the long-term average. The a posteriori SNR estimate uses the noisy signal measurements, and it is given by:

$$\varepsilon_{post}(k,m) = \frac{|X(k,m)|^2 - |N(k,m)|^2}{|N(k,m)|^2} = \frac{|X(k,m)|^2}{|N(k,m)|^2} - 1$$

Equation (10.16)

The SNR estimate used in the Ephraim and Malah (1984) algorithm is a linear combination of the a priori and the a posteriori SNR estimates (Cohen, 2004; Hendriks et al., 2005). In practice, the a priori SNR is not known, so it is approximated by using the speech signal estimate and the noise power estimate from the previous block:

$$\varepsilon(k,m) = \alpha \frac{\left|\hat{S}(k,m-1)\right|^2}{\left|N(k,m-1)\right|^2} + (1-\alpha)\,max\left[\frac{\left|X(k,m)\right|^2}{\left|N(k,m)\right|^2} - 1, 0\right]$$

Equation (10.17)

The averaging factor α, $0 < \alpha < 1$ controls the tradeoff between noise reduction and signal distortion. A value of α close to 1 smoothes the SNR estimate, thus reducing the processing artifacts but reducing the ability of the estimate to track changes in the signal level. A small value of α gives a rapid response to changes in the signal level but will cause more processing artifacts.

At a good SNR, the SNR calculation is close to that used in classical spectral subtraction and depends primarily on the instantaneous signal level and the estimated average noise level. At poor SNRs, however, the SNR calculation averages the instantaneous SNR value over many signal blocks, thus smoothing the SNR estimate and the resultant suppression gain. The smoothed gain has reduced fluctuations compared to classical spectral subtraction. The effect of the averaging on the estimated SNR is plotted in Figure 10-10. During the noise-only portion of the signal the averaged SNR is much smoother than the instantaneous SNR. Once the speech starts, however, the Ephraim and Malah (1984) algorithm reduces the amount of averaging and the estimated SNR is very close to the instantaneous SNR used in classical spectral subtraction.

In practice, the Ephraim and Malah (1984) algorithm has been found to give reduced musical noise artifacts when compared to classical spectral subtraction (Cappé, 1994; Marzinzik, 2000). But it is also possible to use the Ephraim and Malah approach to estimate the SNR without implementing the rest of their algorithm or their gain-reduction rule. Combining the SNR estimation procedure of *Eq. (10.17)* with the Wiener filter given by *Eq. (10.8)* also results in a reduction in musical noise (Hendriks et al., 2005), strongly suggesting that the benefit of the Ephraim and Malah approach is in the smoothed SNR estimate and not in the theory used to derive the noise suppression.

AUDITORY MASKING

Up to this point, the spectral subtraction algorithms have been based on mathematical properties of the speech and noise, and have ignored the fact that a human will be listening to the processed signal. The goal of

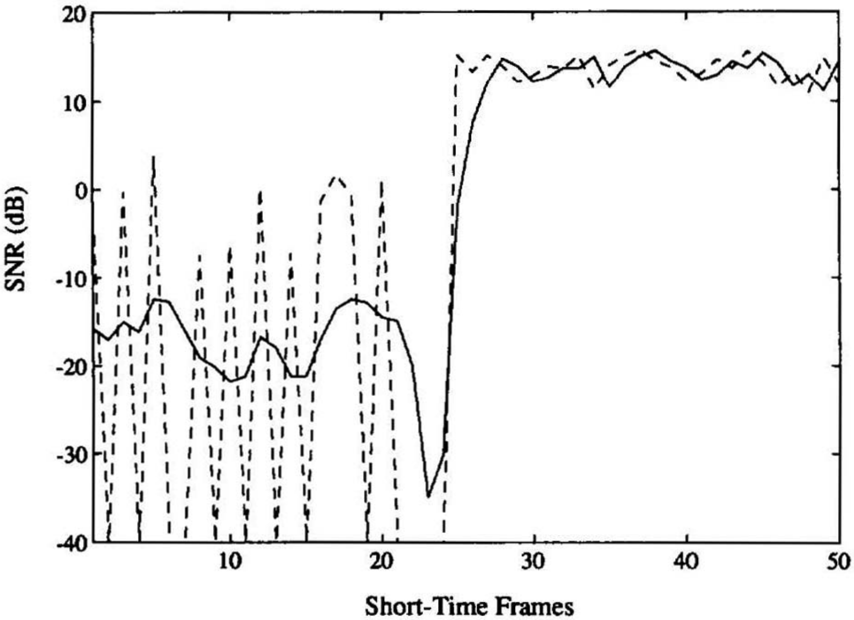

Figure 10-10. Estimated SNRs in successive signal blocks (denoted as short-time frames) for the Ephraim and Malah (1984) noise suppression algorithm. The dashed line gives the SNR computed from the instantaneous signal envelope samples, and the solid line gives the averaged SNR computed using the algorithm. For the first 25 blocks the signal contains only noise. For the next 25 blocks speech is present at a SNR of 15 dB (from Cappé, 1994).

the processing is to render the noise inaudible. One aspect of spectral subtraction, and also the Ephraim and Malah (1984) algorithm, is that a large amount of suppression is provided by the gain rule at poor SNRs. But gain reductions that push the noise level below the audible threshold are actually counterproductive. The changes in the suppression gain from one signal block to the next modulate the envelope of the signal, and this modulation can generate audible processing artifacts. The more the gain varies from one block to the next the greater the envelope modulation, so limiting the maximum amount of signal attenuation to that which just renders the noise inaudible will limit the modulation depth imposed by the processing and reduce the probability of generating audible artifacts such as musical noise.

Limiting the maximum signal attenuation requires implementing a model of noise audibility. The spectral subtraction attenuation is computed for the frequency band, and the reduced noise level is compared to the threshold of audibility in that band. If the attenuation is greater than that needed to place the noise at the audible threshold, the attenuation is reduced so that the noise level matches the threshold (Azirani et al., 1995; Tsoukalas et al., 1997; Virag, 1999).

To compute the model of audibility, the signal spectrum is first grouped into auditory analysis bands, such as critical bands (Zwicker & Terhardt, 1980) or equivalent rectangular bandwidths (ERBs) (Moore & Glasberg, 1983). Auditory masking for the noisy speech signal is then computed using the masking model of Johnston (1988), which takes into account the spread of excitation across auditory filter bands and whether the excitation is tonal or noiselike in nature. For hearing-impaired listeners, the maximum of the auditory masking function or the impaired auditory threshold is then taken as the noise threshold (Natarajan et al., 2005).

A FUNDAMENTAL COMPROMISE

Spectral subtraction involves multiplying the noisy speech signal $X(k,m)$ by a gain factor $G(k,m)$. The processed output is given by:

$$Y(k,m) = G(k,m)\left[S(k,m) + D(k,m)\right] \quad \textit{Equation (10.18)}$$

The reduced noise level is given by $G(k,m)D(k,m)$, and the speech distortion is given by the change in the speech amplitude $[1-G(k,m)]S(k,m)$. If $G(k,m)$ is close to 1, there will be minimal modification to the speech signal along with only a small reduction in the noise level. If $G(k,m) \approx 0$, the speech signal will be substantially modified, but there will be a large reduction in the noise level. Thus, no matter what procedure is used to estimate the noise level or compute the suppression gain, there still remains a tradeoff between the amount of noise suppression and the amount of signal distortion. For any spectral subtraction algorithm, increasing the amount of noise suppression will increase the amount of speech distortion. Thus, designing a signal processing algorithm and adjusting the parameters will always involve a compromise between noise suppression and distortion, and the effects of the algorithm on intelligibility and sound quality will depend on the choices made.

Algorithm Effectiveness

Jamieson et al. (1995) looked at the effects of an adaptive Wiener filter on different kinds of noise and different SNRs. The noises included white Gaussian noise, white noise low-pass filtered at 1 kHz, and multitalker babble. Both normal-hearing and hearing-impaired listeners took part in the experiments. An SRT test using a closed set of spondees showed significantly improved intelligibility for the Wiener filter for both the normal-hearing and hearing-impaired subjects listening in low-pass filtered noise, and no benefit for the white noise or babble. No statistically significant intelligibility benefit was observed for the Wiener filter for any noise condition or group of listeners for isolated vowel-consonant-vowel (VCV) test materials. Preference testing used continuous discourse (female talker), and both the normal-hearing and hearing-impaired listeners preferred the processed material at positive SNRs, had no preference at a SNR of 0 dB, and preferred the unprocessed material at negative SNR values.

A study of spectral subtraction was reported by Levitt et al. (1993). The test stimuli were nonsense syllables in a background of multitalker babble at an SNR of 15 dB. The speech was pre-emphasized with a 6-dB/octave slope. In one version of the processing magnitude spectral subtraction was applied to the entire signal spectrum. In the second version of the processing the spectral subtraction was applied to the speech spectrum below 2800 Hz and the spectrum at higher frequencies was left unprocessed. The split-band processing gave slightly higher speech recognition scores than the full-band processing, but neither version of spectral subtraction was more intelligible than the unprocessed stimuli. Both versions of spectral subtraction were preferred to the unprocessed materials about 70 percent of the time.

Marzinzik (2000) investigated the spectral subtraction scheme proposed by Ephraim and Malah (1984). The test stimuli were sentences in three kinds of noise: industry noise recorded on a factory floor, cafeteria babble, and stationary speech-shaped noise. Intelligibility for the normal-hearing subjects in cafeteria babble showed no significant difference between the processed and unprocessed stimuli. Intelligibility for the hearing-impaired subjects showed a slight nonsignificant reduction for the processed stimuli in comparison with the unprocessed stimuli for cafeteria babble and speech-shaped noise. Preference judgments for the normal-hearing subjects in cafeteria babble showed a slight preference for the unprocessed over the processed stimuli at both −5 and +5 dB SNRs, although the differences in overall impression were smaller at the higher SNR. The hearing-impaired subjects in industry noise showed no preference at an SNR of 0 dB, and a significant preference for the processed stimuli at an SNR of 10 dB. The hearing-impaired subjects showed no significant preference in the cafeteria babble at either an SNR of 0 or 10 dB. For the speech-shaped noise, the hearing-impaired subjects showed a slight preference for the unprocessed stimuli at an SNR of 0 dB and a slight preference for the processed stimuli at an SNR of 10 dB.

Arehart et al. (2003) evaluated both intelligibility and speech quality for the Tsoukalas et al. (1997) spectral-subtraction approach using both normal-hearing and hearing-impaired listeners. Intelligibility was measured using monosyllables in communication channel (white stationary) noise and highway (low-frequency emphasis, nonstationary) noise. The processed stimuli had a small but significant improvement in intelligibility for the communication noise but not for the highway noise for both groups of subjects. Overall quality judgments using sentence materials for the normal-hearing subjects showed a small improvement for the processed stimuli in the communication noise but not for the highway noise. The hearing-impaired subjects gave higher overall quality ratings to the processed stimuli in all conditions, but the preferences for the processed materials were greater at an SNR of 5 dB than for 0 dB, and were greater for the communication noise than for the highway noise. Further experiments with the same spectral-subtraction algorithm (Natarajan et al., 2005) show a small but not significant benefit in intelligibility for the processed stimuli for both groups of listeners. Both the normal-hearing and hearing-impaired subjects gave the processed speech higher overall quality ratings; however, the benefit of the processing was smaller for crowd noise than for communication channel noise.

A recent study by Hu and Loizou (2006) compared sound quality judgments for 13 speech-enhancement algorithms using normal-hearing listeners. They looked at sentences in four types of noise at 5-dB and 10-dB SNRs. Overall quality judgments appeared to be highest for the Ephraim and Malah (1984) approach and for related algorithms. Judgments were also high for a multiband spectral subtraction approach that smoothed the spectrum across FFT frequency bins before computing the noise suppression gain and then applied power spectral subtraction using an oversubtraction factor that depended on the estimated SNR (Kamath & Loizou, 2002).

Concluding Remarks

This chapter has covered many different approaches to spectral subtraction and related noise-suppression algorithms. Some of the algorithms, such as Wiener filtering (1949) and the Ephraim and Malah (1984) algorithm, are based on theoretical derivations intended to mathematically optimize the separation of the speech and noise. Other algorithms have been designed on more of an ad hoc basis, with the signal processing adjusted to sound as good as possible to the person designing it. The theoretical approach ignores the ear, whereas the ad hoc approach implicitly includes the ear as the algorithm parameters are adjusted based on the results of informal ("Sounds good to me") listening tests. What is interesting is that both approaches to algorithm design have produced comparable results; small improvements in quality and possible improvements in intelligibility for stationary background noise. Ignoring the ear appears to greatly reduce the potential benefit of the theoretically optimum processing, whereas mathematically unsophisticated approaches can be effective if the intuitive design process is effective in including the ear. One can hope that in the future combining mathematical rigor with auditory models will lead to improved spectral subtraction algorithms, but there will always remain the tradeoff between the amount of noise suppression and the amount of signal distortion introduced by the processing.

There still are problems that may limit the effectiveness of spectral subtraction. One problem is nonstationary noise. The basic approach to estimating the noise is to form the estimate during pauses in speech, and then to hold the estimate constant during speech sounds. The assumption is that the noise power will not vary during the speech interval. If the noise fluctuates, the estimate will be in error and the spectral subtraction gain will also be in error. Some types of noise, such as communications channel noise, are nearly stationary and spectral subtraction offers measurable benefits. Other types of noise, such as babble or traffic noise, fluctuate and spectral subtraction may be of only limited benefit.

A second concern is that spectral subtraction affects the magnitude of the signal but not its phase. The envelope is modified to more closely resemble that of the clean speech, but the speech is then reconstructed using the noisy phase. No matter how closely the envelope of the reconstructed signal matches the clean speech, there will still be residual effects of the noise in the phase that have not been removed by the processing.

A further concern is that the spectral enhancement gain changes across processing block boundaries as the estimated SNR changes. Smoothing the gain reduces the fluctuations and thus reduces the audible modulation and processing artifacts such as musical noise introduced by the processing. So there is a processing tradeoff—long smoothing time constants reduce the musical noise, but interfere with the ability of the algorithm to response to rapid changes in the speech level. Short time constants increase the musical noise but also allow the processing to react to the fast onset of a speech sound. Minimizing artifacts while still being able to react to sudden speech onsets is a difficult balancing act that requires as much art as science at the present time.

References

Arehart, K. H., Hansen, J. H. L, Gallant, S., & Kalstein, L. (2003). Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners. *Speech Communication*, *40*, 575–592.

Arslan, L., McCree, A., & Viswanathan, V. (1995). New methods for adaptive noise suppression. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, *I* (pp. 812–I:815), Detroit, May 9–12, 1995.

Azirani, A. A., Le Bouquin, R., & Faucon, J.-G. (1995) Optimizing speech enhancement by exploiting masking properties of the human ear. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, I* (pp. 800–I:803), Detroit, May 9–12, 1995.

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 208–211), April 2–4, 1979, Washington, D.C.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP-27*, 113–120.

Cappé, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, *2*, 345–349.

Clarkson, P. M., & Bahgat, S. F. (1991). Envelope expansion methods for speech enhancement. *Journal of the Acoustical Society of America*, *89*, 1378–1382.

Cohen, I. (2004). On the decision-directed estimation approach of Ephraim and Malah. *Proceedinngs of the IEEE International Conference on Acoustics, Speech, and Signal Processing, I* (pp. 293–I:296), Montreal, May 17–21, 2004.

Cohen, I., & Berdugo, B. (2002). Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Processing Letters*, *9*, 12–15.

Doblinger, G. (1995). Computationally efficient speech enhancement by spectral minima tracking in subbands. *Proceedings of the 4th European Conference on Speech Communication and Technology*, Vol. 2 (pp. 1513–1516), Madrid, Sept 18–21, 1995.

Eger, T. E., Su, J. C., & Varner, W. (1984). A nonlinear processing technique for speech enhancement. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 18A.1.1–18A.1.4), San Diego, March 19–21, 1984.

Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *ASSP-32*, 1109–1121.

Gustafsson, H., Nordholm, S. E., & Claesson, I. (2001). Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Transactions on Speech and Audio Processing*, *9*, 799–807.

Hendriks, R. C., Heusdens, R., & Jensen, J. (2005). Improved decision directed approach for speech enhancement using an adaptive time segmentation. *Proceedings of Interspeech* 2005, 2101–2104.

Hirsch, H. G., & Ehrlicher, C. (1995). Noise estimation techniques for robust speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, I* (pp. 153–I:156), Detroit, May 9–12, 1995.

Hu, Y., & Loizou, P. (2006). Subjective comparison of speech enhancement algorithms. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, May 15–19, 2006.

Jamieson, D. G., Brennan, R. L., & Cornelisse, L. E. (1995). Evaluation of a speech enhancement strategy with normal-hearing and hearing-impaired listeners. *Ear and Hearing*, *16*, 274–286.

Johnston, J. D. (1988). Transform coding of audio signals using perceptual noise criteria. *IEEE Journal of Selected Areas in Communication*, *6*, 314–323.

Kamath, S. D., & Loizou, P. C. (2002). A multi-band subtraction method for enhancing speech corrupted by colored noise. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, IV* (p. 4164), Orlando, FL, May 13–17, 2002. (Available at: http://www.utdallas.edu/~loizou/speech/mbss2002.pdf).

Levitt, H., Bakke, M., Kates, J., Neuman, A., & Weiss, M. (1993). Signal processing for hearing impairment. *Scandinavian Audiology* (Suppl. *38*), 7–19.

Lim, J. S. (1986). Speech enhancement. *Proceedings of the IEEE International Conferece on Acoustics, Speech, and Signal Processing* (pp. 3135–3142), Tokyo, April 7–11, 1986.

Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, *67*, 1586–1604.

McAulay, R. J., & Malpass, M. L. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *ASSP-28*, 137-145.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, *9*, 504–512.

Marzinzik, M. (2000). *Noise reduction schemes for digital hearing aids and their use for the hearing impaired.* Doctoral Dissertation, Universität Oldenburg.

Marzinzik, M., & Kollmeier, B. (2002). Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing*, *10*, 109–118.

Moore, B. C. J., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, *74*, 750–753.

Natarajan, A., Hansen, J. H. L., Arehart, K. H., & Rossi-Katz, J. (2005). An auditory masked threshold based noise suppression algorithm GMMSE-AMT [ERB] for listeners. *EURASIP Journal on Applied Signal Processing, 2005,* 2938–2953.

Stahl, V., Fischer, A., & Bippus, R. (2000). Quantile based noise estimation for spectral subtraction and wiener filtering. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, III* (pp. 1875–III:1878), Istanbul, June 5–9, 2000.

Tsoukalas, D. E., Mourjopoulos, J. N., & Kokkinakis, G. (1997). Speech enhancement based on audible noise suppression. *IEEE Transactions on Speech and Audio Processing*, *5*, 497-514.

Van Compernolle, D. (1987). Increased noise immunity in large vocabulary speech recognition with the aid of spectral subtraction. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 1143–1146), Dallas, TX, April 1987.

Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing*, *7*, 126–137.

Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series*, Cambridge, MA: MIT Press (formerly Technology Press).

Wolfe, P. J., & Godsill, S. J. (2001). Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement. *Proceedings of the 11th IEEE Workshop on Statistical Signal Processing* (pp. 496–499), Singapore, August 2001.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America*, *68*, 1523–1525.