Spectrogram SNR and Spectrogram Displays

Jan. 18, 2014/1/26
J. MacAuslan

The problem: We are often presented with a waveform or spectrogram for which it is helpful to suppress details in noise-dominated sections of time (in the waveform) or of time-frequency ("T-F", in the spectrogram). See Figure 1, which uses the standard SpeechMark® toolbox display, 'sgram'.

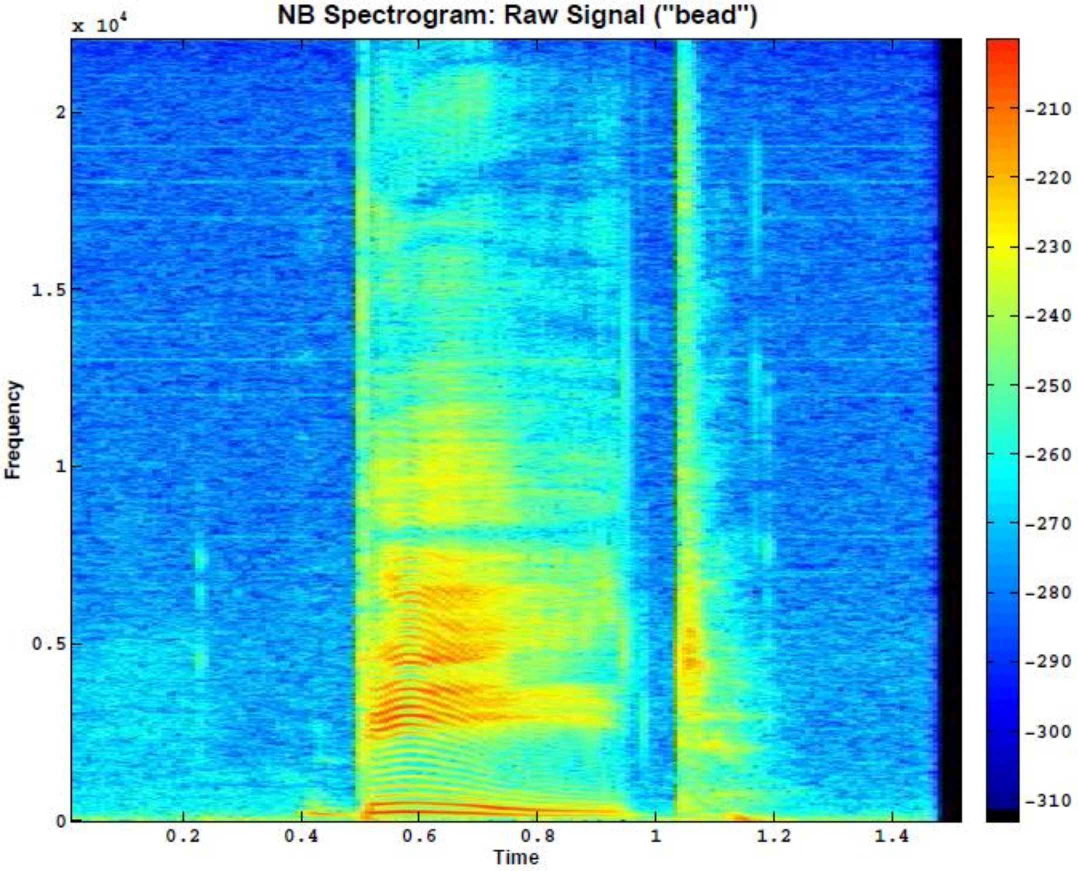

Figure 1. Standard narrow-band spectrogram of the word *bead* spoken in moderate noise. Female speaker, 44.1 kHz sampling. Notice the horizontal stripes of higher power on 1-kHz multiples, especially extending across all times at 12, 13, 14, and 17, 18, and 19 kHz, as well as at other frequencies for briefer segments of time. Apart from these stripes, the noise floor appears to be ~ -275 dB. The frequent fluctuations as much as 40 dB below this level merely introduce variability that represents visual clutter, not meaningful measurements of the speech, nor even of other environmental or speaker sounds.

In keeping with the knowledge-based focus of SpeechMark, we are particularly interested in solutions based on broad principles rather than ones that must be determined in a subtle, complicated, or *ad hoc* fashion, whether by the user or by SpeechMark. That is, we will look to solutions based on a combination of:

- mathematical properties,
- statistical estimators that are insensitive to probability-distribution variation (so-called *robust* estimators), and
- knowledge of physical acoustics, speech physiology, etc.

Such principled solutions are amenable to packaging as “black boxes” into larger systems, because the principles apply broadly. Users of the larger systems can therefore understand and trust the principles’ validity despite the packaging.

A solution: In many cases, there is a principled solution to this problem: Determine the signal/noise ratio (SNR), and use a threshold θ = unity or somewhat higher to distinguish noise-dominated from information-carrying samples. For convenience, we may also speak (here) of the noise as the “background”, at least when it appears with constant amplitude and spectrum, essentially the conditions of statistical stationarity. The signal or non-noise component(s) of the waveform we may refer to as the “foreground”.

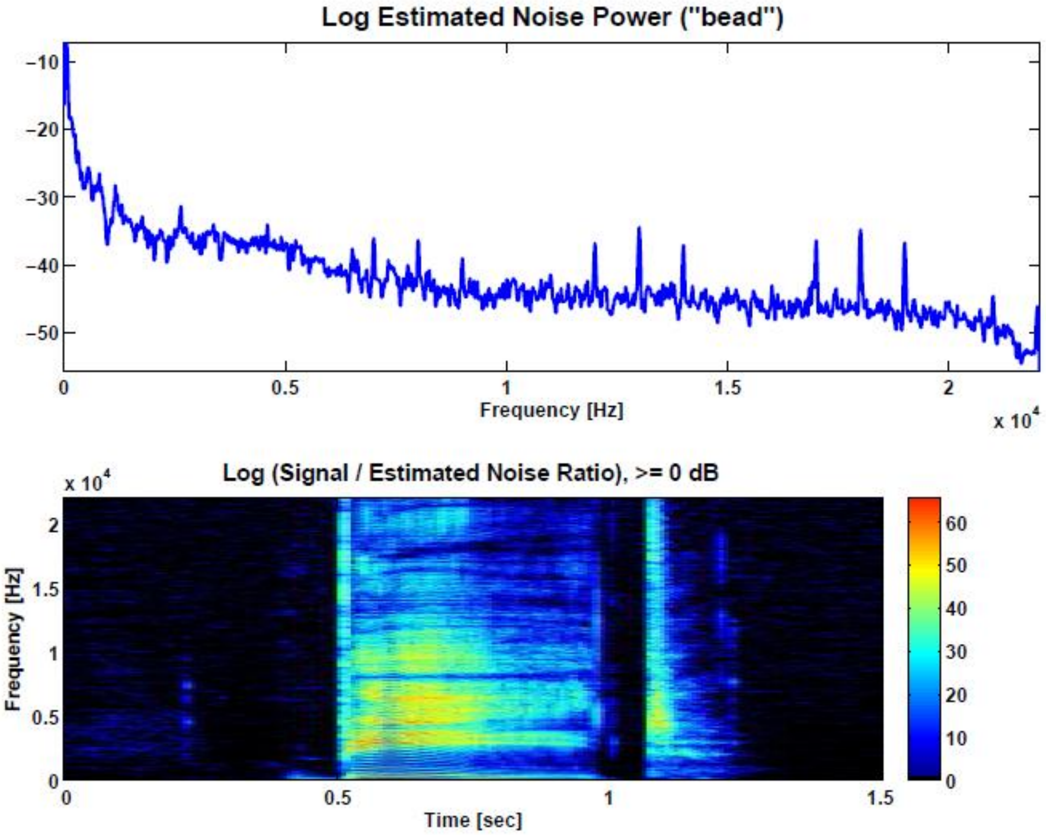

Figure 2. (a: *top*) Spectrum of stationary component of the spectrogram; (b: *bottom*) “foreground-to-background ratio” or “signal-to-noise ratio” spectrogram. At each frequency separately, the algorithm identifies the “background noise” as the spectrogram’s minimum power value, taken over all times. In the present case, this spectrum correctly includes the substantial peaks, ~10 dB, at 12, 13, 14, and 17, 18, and 19 kHz, as well as some others for which we can see evidence in Figure 1 (e.g., 8 kHz). As a result, the spectrogram of Figure 1 can be divided by this spectrum at every moment in time to produce a T-F representation of the foreground-to-background or signal-to-noise ratio: The noise peaks are virtually absent from this image (< 3 dB, appearing here as very dark blue). When fluctuations below unity (0 dB) are raised to unity, the result is a “SNR spectrogram” that strongly deemphasizes both the magnitude of the stationary background and its variability in time, but it retains and even emphasizes the structure of the foreground or signal. In this case, the SNR displays 70 dB of meaningful foreground or signal dynamic range. It does not include the original spectrogram’s bottom 40 dB, which is caused only by temporal fluctuations in the stationary background.

Computationally and acoustically, this might mean performing spectral subtraction on the level of noise. Visually, it might mean removing all fluctuations below this threshold (i.e., *raising* the amplitude to the level of θ), in order to avoid distracting the eye with these fluctuations in a spectrogram image.[1]

In typical spectrogram displays, removing the fluctuations would mean spreading the colors — dark blue through bright red, here — over only the dynamic range of the signal, rather than over the much broader range of noise fluctuations and signal. In Figure 1, for instance, these fluctuations appear to be centered on ~ -275 dB (the "noise floor"), and therefore account for the bottom 40 dB, out of a total range of 110 dB: More than one-third of the color spread is "wasted", only describing these irrelevant fluctuations.

But how to determine this noise-floor level in a principled, automatic fashion?

(a) If the noise floor of the waveform is already known, whether in time, frequency, or T-F, then we may use the criterion of SNR ≥ θ to suppress noise details.

(b) If it is not known in advance, we may estimate the noise spectrum from the statistically stationary or background part of the spectrogram; we would use, at each frequency, the minimum power across all times [as in the SpeechMark functions 'estnoisesig', 'estnoisesig_std', and 'estnoisesgram']. Alternatively, we could estimate the T-F noise floor as being 30 bits (90 dB) down from its peak, on the assumption of a 15-bit digitized waveform. In general, we would probably use the maximum of these two estimates at each time, frequency, or T-F. The result is shown in Figure 2 [from the function 'sgram_snr', which uses 'estnoisesgram'[2]].

Using the minimum power over all times is again a principle-based estimate. The power of unrelated sources is additive, so this rule properly identifies the background component at a given frequency, provided that the unrelated foreground components are silent (at that frequency) at least occasionally within the time interval available. That is, this principle quite properly associates the absence of any foreground signal with a power minimum.

We can now reconstruct a more helpful spectrogram of the waveform itself. Note that the SNR spectrogram (Figure 2b) is constructed by dividing the original spectrogram by the noise level at each frequency separately. It is then rendered visually more helpful by clipping from below at θ to suppress downward noise fluctuations: the bottom 40 dB of Figure 1. Reconstructing the spectrogram merely reverses this process, though crucially keeping the thresholding. That is, at each frequency, the noise level is multiplied by the SNR spectrogram *after* the latter has been clipped at θ. (Note that this clipping, too, is principle-based: Unity has special significance for SNR.)

The result appears in Figure 3. This shows the signal component(s) as well as the original frequency dependence of the noise spectrum, but it does not contain the temporal fluctuations of the noise, at least in the lower-amplitude direction. (In the higher-amplitude direction, noise only accounts for ~6 dB.)

As a result, some features are now much more apparent: e.g., the sustained power at 12.8, 14.9, and 16.7 kHz from 0.6 to 0.9 sec (perhaps high-order formants?); the final power at 900 Hz and 1.3 sec (final exhalation?); and the narrow linear feature at 3500 Hz and 1.25 sec.

[1] Following this with subtraction of θ produces exactly the same result as spectral subtraction. The two techniques have identical information. However, performing the subtraction makes any subsequent visual presentation slightly more difficult than raising the spectrogram floor.

[2] In fact, the typical noise distribution has a mode at zero amplitude, rendering a naïve computation of the minimum-power level severely biased downward; 'estnoisesgram' uses estimators that deal with this complication.

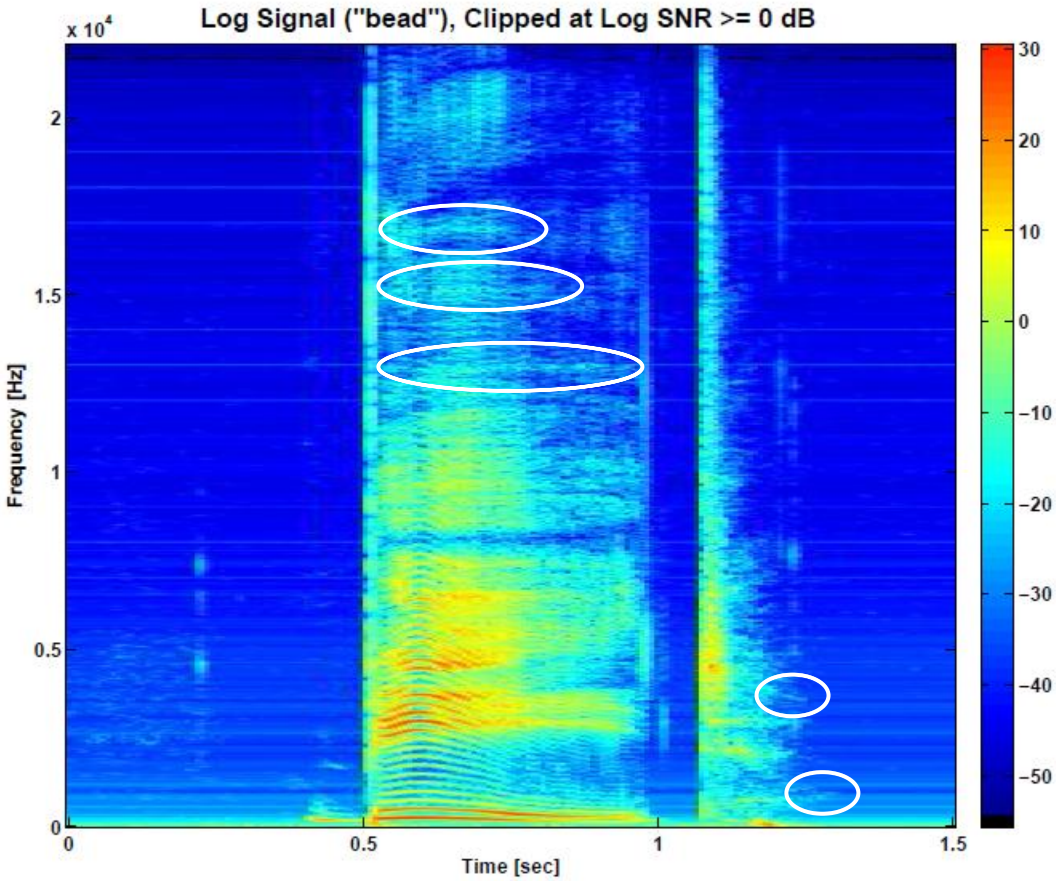

Figure 3. Enhanced spectrogram image based on automatic determination of the stationary background. The image is constructed directly from Figure 2, i.e., suppressing fluctuations below unity (0 dB), as shown in 2b, and multiplying the result by the spectrum of 2a. This properly reinserts the background peaks, e.g., 12-14 kHz, as in the original spectrogram; but it strongly reduces the background fluctuations, as in the SNR spectrogram.[3] This better allows the visual appearance to emphasize the foreground content of the waveform. For example, the three encircled features at 0.6-0.9 sec, the one at 1.25 sec, and the one at 1.3 sec are much easier to notice than in Figure 1.

Extensions: A few extensions are straightforward. First, the resolution of the frequency bands of the spectrogram affects the precision of the SNR determination, but the underlying principles are still valid regardless of resolution. If only the amplitude contour of the signal is available, then its square is a “spectrogram” with a single frequency band. The noise “spectrum” becomes a scalar, the total noise power. The SNR computation, thresholding, and reconstruction of the spectrogram (squared amplitude), proceed as before. The noise power has the same complications as in the multiple-frequency case — a mode at zero, requiring careful debiasing — but the computation remains valid.

A slightly more sophisticated variant of this would estimate the noise spectrum and select a single value, such as its minimum (a very cautious choice), to set the frequency-independent threshold. This would be a crude variant of the SNR processing developed here, but, as we shall see, noticeably better than no thresholding at all.

[3] The color scale has a different offset than in Figure 1, but the dynamic range is properly marked.

In fact, the SpeechMark toolbox uses exactly this variant for visualization. It was used even in the original spectrogram, Figure 1. Among all debiased noise estimates in the spectrum of Figure 2a, the minimum over all frequencies — the value at 22 kHz in this case — was selected, and all smaller spectrogram components were raised to this single, low floor. Figure 4 shows the result of *omitting* this operation: The dynamic range is set by the actual minimum value across all non-zero spectrogram components (image pixels), rather than by the debiased estimate. This results in ~25 dB more "waste" of the color spread, so even some components that were visible in Figure 1, such as the feature at 1600 Hz, 0.45 sec (a formant inside the closed mouth?), are now obscured.

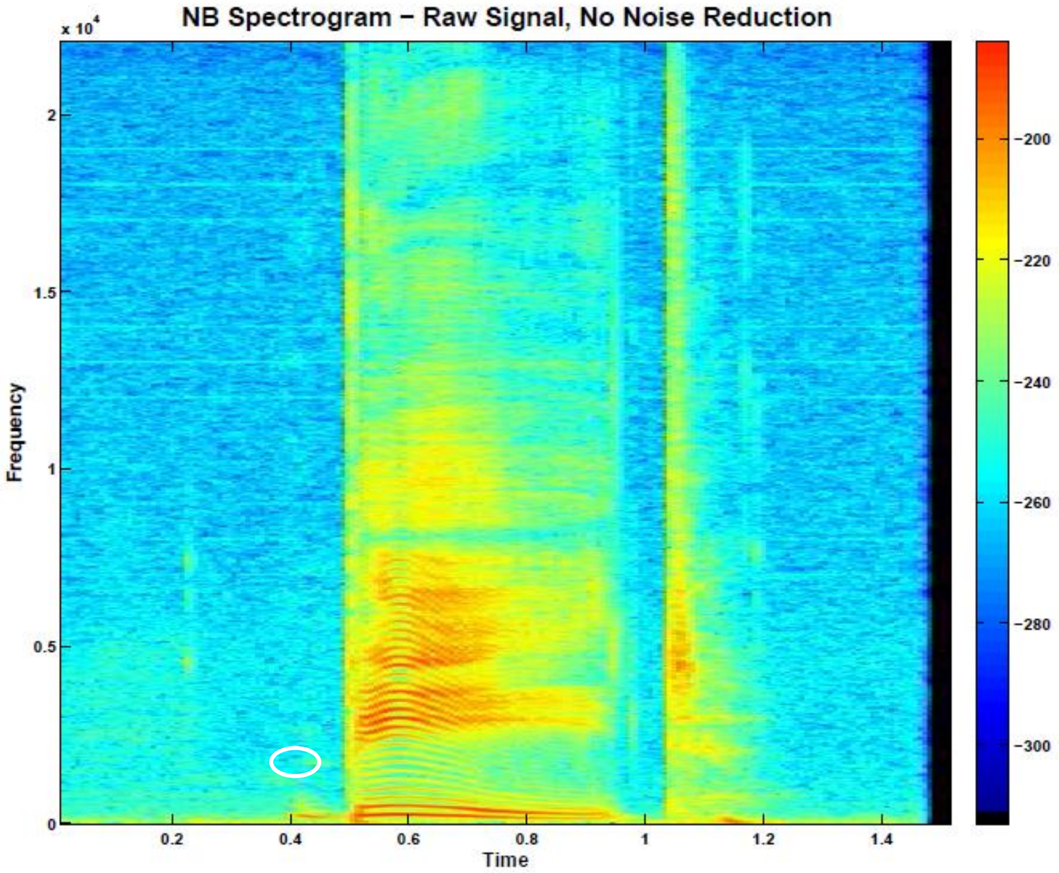

Figure 4. **Spectrogram of raw signal, no noise handling.** Without even the minimal noise-background estimation and suppression used in Figure 1, this spectrogram display obscures still more waveform features, such as that marked at 0.45 sec, because the dynamic range has (unhelpfully) expanded to ~135 dB.

The final extension deals with non-stationary noise backgrounds. As usual, if this is slowly varying, then it is appropriate to estimate it and compute the SNR spectrogram over short intervals, perhaps blending the estimates across adjacent intervals to improve their accuracy.

Conversely, if the background is rapidly varying, it is appropriate to regard the result as a background itself, a highly fluctuating one. However, the estimates, and the resulting SNR computations, will likely be less useful. In particular, any background estimates will likely understate the background power. The result will be to set any estimate-based threshold to a particularly low value. In this sense, the SNR technique here is conservative, allowing subsequent processing to make use of further information if available, or to use some higher, more aggressive threshold if desired.