FaNT and the calculation of the signal-to-noise-ratio (SNR)

FaNT supports speech detection and frequency weighting in the calculation of the signal-to-noise ratio. The motivation behind these features is explained below.

As noted below by Daniel Ellis and George Doddington, there is considerable ambiguity in the research literature about how SNR is calculated. To prevent such ambiguity, we recommend that researchers using FaNT mention their use of FaNT (and what FaNT options were used). FaNT was used to create the popular Aurora 2 (Hirsch and Pearce, 2000) and Aurora 4 data sets. Thus, many researchers are already familiar with SNR levels calculated using FaNT.

Speech detection in FaNT

FaNT performs speech detection using the ITU-T P.56 Speech Voltmeter from the ITU Software Tools Library (G.191 Annex A). Section 9 of (Neto, 1999) explains the design and operation of the Speech Voltmeter. FaNT should not be used to lower the SNR of speech that is already noisy, because the speech detection in the Speech Voltmeter is not designed for noisy speech. Also note that the correct audio sampling rate must be specified to FaNT for the Speech Voltmeter to operate reliably.

Frequency weighting in FaNT

FaNT supports frequency weighting during SNR calculation using G.712 frequency weighting or A-weighting. It is also possible to use no frequency weighting, but please read the discussion of frequency weighting below by George Doddington. (The ITU Software Tools Library manual suggests bandlimiting the Speech Voltmeter's input signal to 300-3400 Hz for telephony band signals and 100-7000 Hz for wideband signals. FaNT does not follow this suggestion. However, Simão Ferraz De Campos Neto, the author of the Speech Voltmeter, told us by email that he believes this suggestion is only meant to increase the comparability of different measurements, and is not based on an inherent limitation of the Speech Voltmeter algorithm.)

Why use speech detection in SNR calculation? (by Daniel P.W. Ellis)

SNR (Signal-to-noise ratio) is a standard measure of the amount of background noise present in a speech (or other) signal. It is defined as the ratio of signal intensity to noise intensity, expressed in decibels, e.g.

```
SNR_dB = 20.log10(S_rms / N_rms)
```

where S_rms is the root-mean square of the speech signal (without any noise present) i.e. sqrt(1/N*sum(s[n]^2)), and N_rms is the root-mean square level of the noise without speech. This is equal to:

```
SNR_dB = 10.log10(S_e / N_e)
```

where S_e is the total energy of the speech i.e. sum(s[n]^2) etc.

The difficulty of this measure comes from the highly nonuniform nature of the speech. Consider an utterance of 1 second duration; it has a certain energy E. We can construct a noise-corrupted version at a given SNR by finding some noise sample (say white noise, or a recording of ambience in a

moving car) of the same duration, and scaling its level to obtain the desired SNR according to the above equations, then adding the two together.

If we then consider a second version of the speech example with 1 second of silence (zero-valued samples) added to make its total duration 2 seconds, its total energy is unchanged. However, to make a 2 second sample of the noise that has the same total energy as the 1 second example, we would need to reduce its amplitude by about 30% so that sum(n[n]^2) is the same when twice as many values of n are involved. This is a real problem: the actual level of noise added to achieve a given global SNR depends strongly on the amount of padding added to (or, in general, silence present in) the speech example. Much confusion has resulted from SNR levels quoted in papers that have fallen foul of this ambiguity.

Frequency weighting

FaNT can apply a G.712 or A-weighting filter to the speech and noise signals during SNR calcluation.

Why use frequency weighting in SNR calculation? (by George Doddington)

ASR [automatic speech recognition] degradation will be far greater for white noise than for low frequency noise at the same SNR value [if the SNR calculation does not use frequency weighting]. This is true for human as well as ASR performance.

A common method for minimizing this variability is to weight the signal and noise energy as a function of frequency. There is a standard "A weighting" to achieve this, which can be approximated reasonably by measuring the energy in the first-order difference signal. Of course, without this frequency weighting, it becomes possible to show really good noise robustness numbers -- just use a noise signal with a lot of very low frequency energy! But for the scientifically inclined, frequency weighting is de rigueur.

Bibliography

Hans-Günter Hirsch and David Pearce, "The AURORA Experimental Framework For The Performance Evaluation of Speech Recognition Systems Under Noisy Conditions". ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium, September 18-20, 2000. Paris, France. Available online here or here.

Simão Ferraz De Campos Neto, "The ITU-T Software Tool Library". International Journal of Speech Technology, Vol. 2, Number 4, May 1999. Available online through SpringerLink.

ITU Software Tools Library (G.191 Annex A). Available online through the ITU.