Predicting speech intelligibility in conditions with nonlinearly processed noisy speech

The speech-based envelope power spectrum model (sEPSM; [1]) was proposed in order to overcome the limitations of the classical speech transmission index (STI) and speech intelligibility index (SII). The sEPSM applies the signal-to-noise ratio in the envelope domain (SNR$_{env}$), which was demonstrated to successfully predict speech intelligibility in conditions with nonlinearly processed noisy speech, such as processing with spectral subtraction. Moreover, a multiresolution version (mr-sEPSM) was demonstrated to account for speech intelligibility in various conditions with stationary and fluctuating interferers [2]. However, the model fails in the case of phase jitter distortion, in which the spectral structure of speech is affected but the temporal envelope is maintained. This suggests that an across audio-frequency mechanism is required to account for this distortion. It is demonstrated that a measure of the across audio-frequency variance at the output of the modulation-frequency selective process in the model is sufficient to account for the phase jitter distortion. Thus, a joint spectro-temporal modulation analysis, as proposed in [3], does not seem to be required. The results are consistent with concepts from computational auditory scene analysis and further support the hypothesis that the SNR$_{env}$ is a powerful metric for speech intelligibility prediction.