MUSICAL NOISE LIMITS TO SPEECH ENHANCEMENT

Brady Laska
Research in Motion Ltd., 295 Phillip St. Waterloo, ON, N2L 3W8 blaska@rim.com

1. INTRODUCTION

Musical noise is a term used to describe short-duration narrowband artifacts present in speech processed by spectral modification noise suppression systems. The phenomenon is most easily understood within the context of the spectral subtraction algorithm [1]. Fig. 1 presents a block diagram of magnitude spectral subtraction. The noisy signal $z[n]$, consisting of the clean speech signal $x[n]$ and the additive noise $v[n]$, is divided into overlapping blocks, then transformed to the frequency domain via the fast Fourier transform (FFT). An estimate of the noise signal FFT magnitude is subtracted from the noisy speech magnitude, and the result is re-combined with the noisy FFT phase to reconstruct the enhanced signal in the time domain. Due to stochastic fluctuations, the actual noise magnitude in a given FFT bin will differ from its estimate. When the true value is lower than the estimate, the noise at that frequency is completely removed; when it is higher, some residual noise will remain. This successive elimination and under-suppression produces isolated peaks in the time-frequency noise representation. This is illustrated in Fig. 2: the top plot shows the spectra of two successive noise-only frames input to a spectral-subtraction system along with the estimated noise spectrum (thick solid); the bottom plot shows the enhanced output spectrum with clearly visible isolated spectral peaks. When converted back to the time domain, these peaks become short-lived tones that randomly vary in frequency. This type of on-off tonal switching is the most well-known manifestation; however any significant random modulation in the noise spectrum will create musical-noise type artifacts. Depending on bandwidth of each FFT bin, the sound of the artifacts may range from a tinkling bell to beeping tones to flowing water. These artifacts are so distracting and un-natural sounding that listeners generally prefer the original noisy signal to the processed speech with musical noise.

The human auditory system is tuned to recognize spectral change. When presented with a tone burst, the discharge rate of auditory nerve fibers will rapidly rise, and then gradually decay to the background level, even as the tone persists [2]. This response to spectral change is beneficial for extracting short-lived consonants and speech onsets in low signal-to-noise ratio (SNR) environments. It may also explain why musical noise is so objectionable: While listeners can adjust to a steady-state noise, a randomly modulated noise constantly reminds the listener of its presence. To achieve a natural sounding output that preserves the character of the residual noise, all noise-only spectral components must receive the same attenuation regardless of their absolute magnitude.

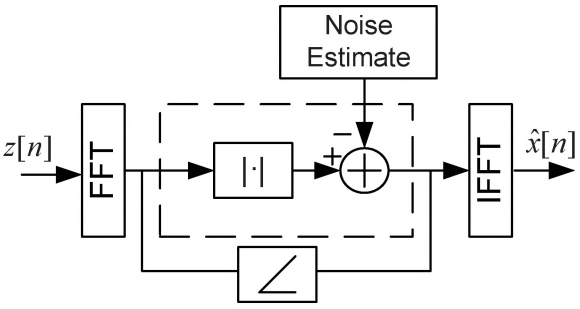

Figure 1. Block diagram of spectral subtraction speech enhancement.

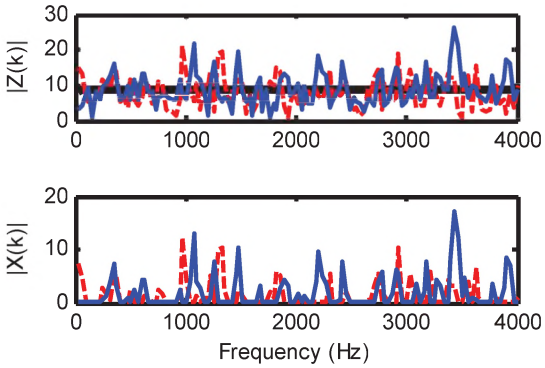

Figure 2. Example of spectral-subtraction processing of noise.

The damaging impact of musical noise on perceived speech quality is well known, and much research has been undertaken to mitigate its impact; however little has been done to understand musical noise on a theoretical basis. In this work we use statistical noise models to demonstrate how musical noise arises in speech enhancement, to characterize the sensitivity of different noise suppression algorithms to artifacts, and to understand the limits to artifact-free noise suppression.

2. GAIN FUNCTION SENSITIVITY

Fig. 3 plots the gain functions of the Wiener filter and the power and magnitude spectral subtraction speech enhancement algorithms as a function of *a posteriori* SNR, $\gamma = 20\log_{10}(|Z(n)| / |V(n)|)$ dB. The plotted attenuation is limited to -25 dB, though the theoretical gain at 0 dB (noise-only) is $-\infty$ dB. At non-zero SNRs, corresponding to intervals where speech is present, the slope of the gain functions is low, so SNR errors have a minor impact. In contrast, around 0 dB the gain functions are very steep, which means that small errors in SNR estimation are amplified to produce large gain variations. Since speech can mask its presence, musical noise is most noticeable in noise-only regions. Thus, the impact of SNR errors is greatest at SNRs where gain fluctuations are most

noticeable. The Wiener gain is the minimum mean-squared error (MMSE) optimal linear estimator of the speech FFT coefficients; however, computing the gain requires knowledge of the SNR, therefore the gain is only optimal insofar as the SNR estimate is exact. However, noise is a random signal; therefore there will always be some uncertainty in the SNR estimate.

2.1 Noise Estimation Error

The FFT coefficients of acoustic noise signals are commonly modeled as complex Gaussian random variables, leading to Rayleigh distributed spectral amplitudes [3]. The Rayleigh distribution is a single parameter distribution with mean and cumulative density function (cdf) given by: $\mu = \sigma\sqrt{\pi/2}$ and $F_X(x) = 1 - e^{-x^2/2\sigma^2}$.

Since the noise is mixed with the desired speech signal, we can only obtain smoothed estimates of the noise statistics. In a stationary environment, a noise estimator will converge to the mean of the noise amplitude distribution. To illustrate the effect of noise fluctuations on a given speech enhancement system, we define the noise to expected-noise ratio as NENR = $20\log_{10}(|V(n)|/\mu)$ dB. During noise-only periods, the NENR is the SNR seen by the speech enhancement system as a result of noise fluctuations. Since 1-$F_X(x)$ is the probability that the random variable X will exceed x, this can be used to give the probability of a given NENR. Fig. 4 plots the gain error as a function of NENR, as well as the probability of the NENR, for Rayleigh distributed noise with parameter $\sigma = 1$.

3. DISCUSSION

Fig. 4 provides some additional insight into the musical noise performance of the enhancement algorithms. For example, there is a 25% probability that the NENR will exceed 2.5 dB. At this NENR the gain error is about 12 dB for magnitude spectral subtraction, 17 dB for the Wiener filter and over 21 dB for power spectral subtraction. This high probability of large fluctuations is why the power spectral subtraction algorithm exhibits the highest levels of musical noise.

Fig. 4 can also be used to explain the fundamental limitations of spectral modification enhancement. The true noise (and NENR) distribution cannot be controlled by the system designer; only the gain function can be adjusted. Classical approaches to musical noise reduction involve a combination of over-subtraction and spectral flooring.

Over-subtraction artificially inflates the noise estimate, which reduces the NENR, thereby shifting the gain functions to the right and reducing the probability of large gain errors. Spectral flooring puts a limit on the amount of attenuation that can be applied, reducing the dynamic range of the possible gain. This has the effect of shifting the gain curves down, so the error for a given NENR is reduced and the maximum gain error is constrained. While these approaches offer some musical noise control, spectral flooring limits the noise attenuation and over-subtraction increases the probability that a low-level speech component will be attenuated or removed.

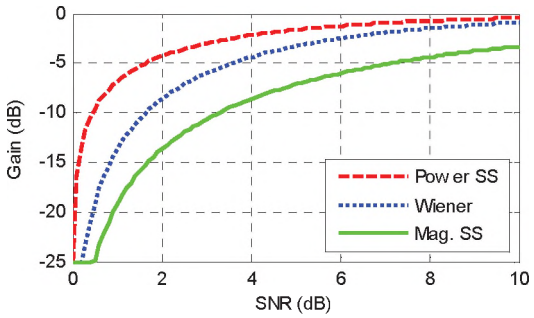

Figure 3. Spectral amplitude enhancement gain functions.

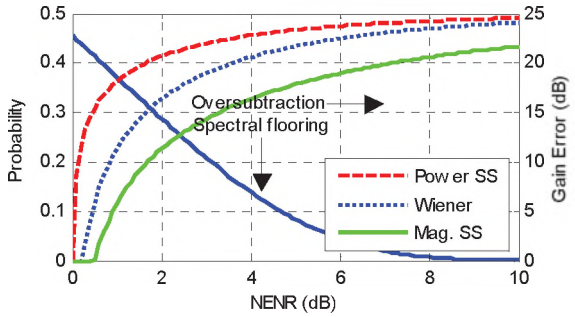

Figure 4. Gain error resulting from noise fluctuations.

4. SUMMARY

Stochastic variations of the noise signal from its expected value prevent us from obtaining exact SNR estimates. In most spectral modification speech enhancement systems, small SNR errors during noise-only periods can produce large fluctuations in the applied gain which modifies the character of the residual noise and resulting in musical noise artifacts. The need to prevent the emergence of these highly objectionable artifacts limits the amount of noise attenuation that can be applied without severely distorting the speech signal.

5. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics Speech Signal Process.*, vol. 27, pp. 113 –120, Apr. 1979.

[2] B. Delgutte, "Auditory neural processing of speech," in *The handbook of phonetic sciences*, pp. 507 – 538, Oxford: Blackwell, 1997.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoustics Speech Signal Process., vol. 32, no. 6, pp. 1109 – 1121, 1984.