REDUCING AUDIO NOISE USING SPECTROGRAM RANDOM TEXTURES

Ramin Samadani

Imaging Technology Department

Abstract

This paper discusses audio enhancement when a strong, additive noise is present only during a known or easily detected period of moderate length (of around one second). The signals may contain *intelligible* components such as speech or music, and may also contain desired, but *unintelligible*, background components such as rivers or waterfalls. A first estimate synthesizes the unintelligible components from the noise-free neighboring spectrogram. A second estimate recovers the intelligible components using spectral attenuation. The two estimates are combined using ideas from statistical process control. Tests with audio containing digital camera zoom motor noise, and with simulations, validate the approach.

1. PROBLEM STATEMENT

Recovery is desired of an audio signal with strong additive noise that occurs only in a finite time interval. This situation may occur in communications interference or, the motivation for this work, in digital cameras while the zoom motor runs. Assume the signal consists of *intelligible* components such as speech or music that are sensitive to distortion, and *unintelligible* components such as rivers, waterfalls, etc., that allow higher distortion, and may therefore be synthesized. In the time domain,

$$x(t) = s(t) + \eta(t) = s_I(t) + s_U(t) + \eta(t), \qquad (1)$$

where the noisy signal $x(t)$ is the sum of signal $s(t)$ and noise $\eta(t)$. The signal $s(t)$ is composed of two components, $s_I(t)$, the intelligible component, and $s_U(t)$, the unintelligible component. The quantity $\eta(t)$ is assumed to be non-zero only when $t_1 \leq t \leq t_2$.

Discretizing Equation 1 and applying the discrete short-time Fourier transform results in the situation shown in Figure 1, and corresponding to the following equation:

$$\begin{aligned} X(n,k) &= S(n,k) + \mathcal{N}(n,k) \\ &= S_I(n,k) + S_U(n,k) + \mathcal{N}(n,k). \qquad (2) \end{aligned}$$

Here, the noisy signal $x(t)$ is transformed to the discrete-time, short-time transform $X(n,k)$, with time index, k, and spectral index, n. There are corresponding spectral quantities for all of the terms in Equation 1. Component $S_I(n,k)$ represents the intelligible signal, and $S_U(n,k)$ represents the unintelligible signal. The time-limited noise to be reduced, $\mathcal{N}(n,k)$ is non-zero only for times k such that $k_1 \leq k \leq k_2$.

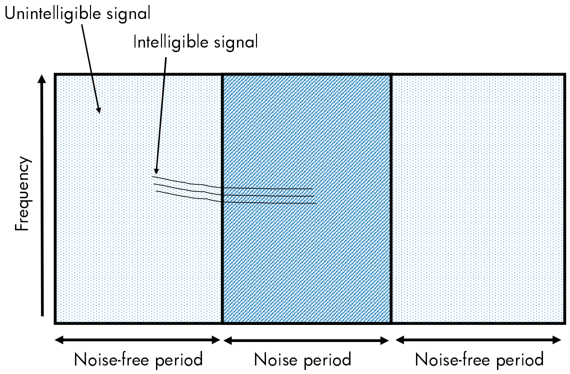

Fig. 1. Schematic of the problem in the short-time spectral domain.

The solution below uses the information *before* the noise period to synthesize, during the noise period, the unintelligible signals. The information during the noise period is used in a spectral attenuation to estimate the intelligible signal. The solution described below provides, without complex auditory segmentation, a separation, $S = S_I + S_U$, based on the randomness of the spectrogram in the noise free region $0 \leq k < k_1$, and also based on the average signal and noise energy in the same noise-free region.

2. DESCRIPTION OF THE METHOD

Figure 2 shows the processing steps, together with example spectrograms. The top left has the spectrogram for about three seconds of input audio — brighter, warmer colors representing more energy. Piano notes occur throughout, and loud zoom noise occurs in the middle. Traditional noise reduction [1] for Equation 2 estimates noise characteristics and applies spectral subtraction [2] or attenuation [3]. This is a component of the current solution, but for low SNRs, it leaves noticeable residuals seen in the spectrogram on the bottom of Figure 2. The new solution additionally uses the spectrogram, $|X(n,k)|$, of the noise free period $0 \leq k < k_1$ to synthesize the unintelligible background during the noise period, $k_1 \leq k \leq k_2$. The top middle of Figure 2 shows the synthesized spectrogram. From the figure, one sees this spectrogram does not reproduce the intelligible component (the piano notes). The final result combines the synthesized estimate and the traditional estimate. The residual zoom noise is much reduced, and

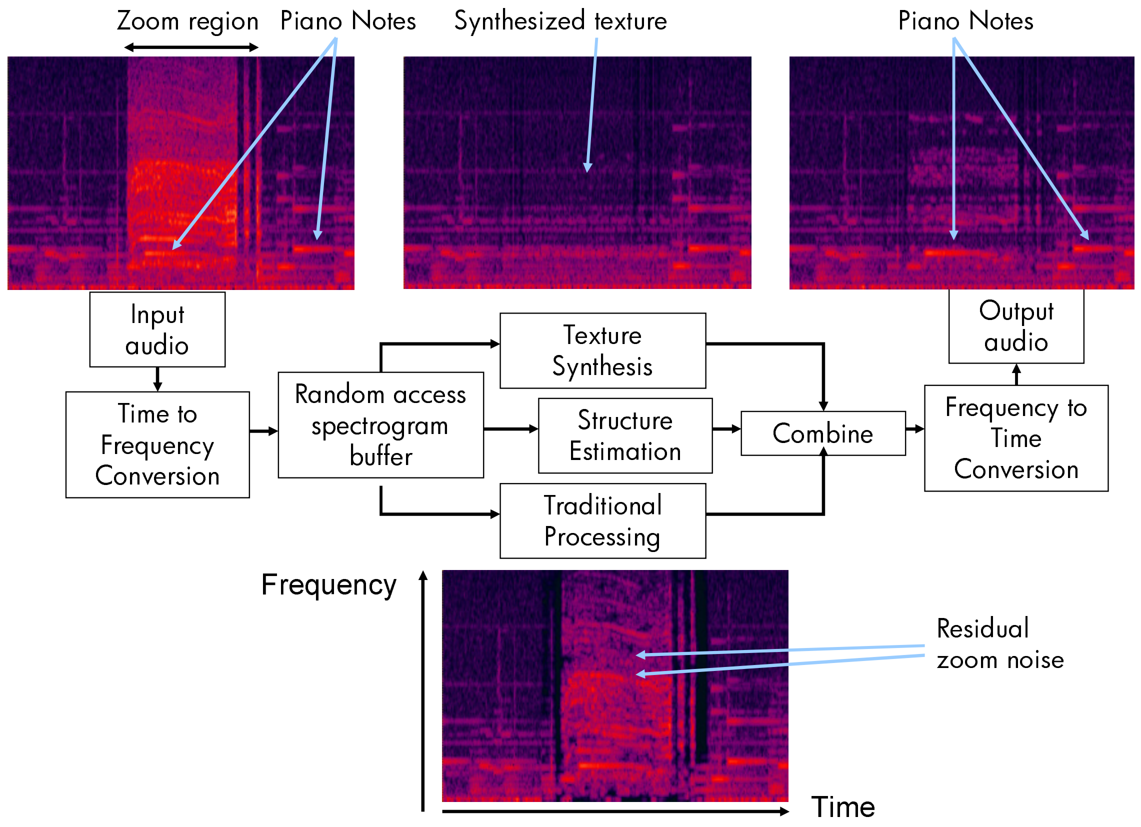

Fig. 2. Audio processing diagram together with example spectrograms for audio with piano and zoom noise.

the piano notes during the noise period are evident, in the final spectrogram shown on the top right of the figure. The final estimate $\hat{S}(n,k)$, is a frequency-weighted combination,

$$\hat{S}(n,k) = \alpha(n)\hat{S}_1(n,k) + (1-\alpha(n))\hat{S}_2(n,k). \tag{3}$$

Here, $\hat{S}_1(n,k)$ corresponds to the synthesized *spectrogram random texture* estimate, and $\hat{S}_2(n,k)$ corresponds to the spectral attenuation estimate. Only magnitudes are modified.

Figure 3 refers to the derivation of $\hat{S}_1(n,k)$. To compute $\hat{S}_1(n,k)$, the short-time energy, $E(k_a)$, of the signal $X(n,k)$ is calculated for each $k_a \in [0...k_1-1]$. Assuming pauses in the intelligible signal, $S_I(n,k)$, the lower values of $E(k)$ occur when only $S_U(n,k)$ is present. Thus, the spectrogram time slices (vertical slice for a given k_a) with the smaller $E(k_a)$, in the period before the noise period, are pseudo-randomly sampled to generate $k_2 - k_1 + 1$ synthetic spectrogram time slices for the noise period. In this way, $\hat{S}_1(n,k)$ approximates the unintelligible signal.

The *Structure Estimation* block of Figure 2 computes weight $\alpha(n)$ for each frequency n. Intuitively, $\alpha(n)$ are set so that $\hat{S}_1(n,k)$ dominates for frequency bins with mostly unintelligible signal, and $\hat{S}_2(n,k)$ dominates when it is predicted that intelligible frequencies will occur during the noise period. The quantity $\alpha(n)$ in Equation 3 is set using two considerations. A first factor is the randomness of the expected signal (corresponding to unintelligibility) within that spectral bin. A second factor is the amount of signal energy within a spectral bin during the noise-free period compared to the amount of noise energy expected in that spectral bin during the noise-period.

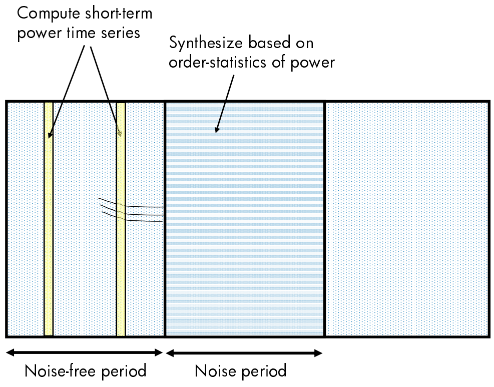
Compute short-term power time series
Synthesize based on order-statistics of power
Noise-free period
Noise period

Fig. 3. Computing short-time power from the spectrogram.

The randomness of the signal is determined in the spectral domain. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are Normal RVs, corresponding to the real and imaginary components of the Fourier transform, their joint probability density function (pdf) is given by,

$$f(x,y) = \frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2}. \tag{4}$$

Then the magnitude, $\boldsymbol{r} = \sqrt{\boldsymbol{x}^2 + \boldsymbol{y}^2}$, has a Rayleigh [4] pdf

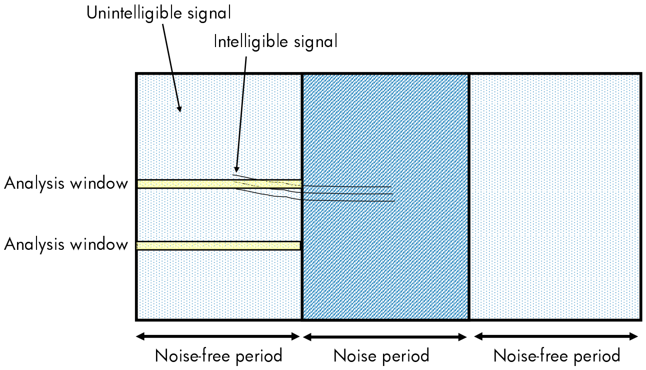

Fig. 4. Analysis of randomness for each spectral band.

given by

$$f(r) = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2} u(r). \tag{5}$$

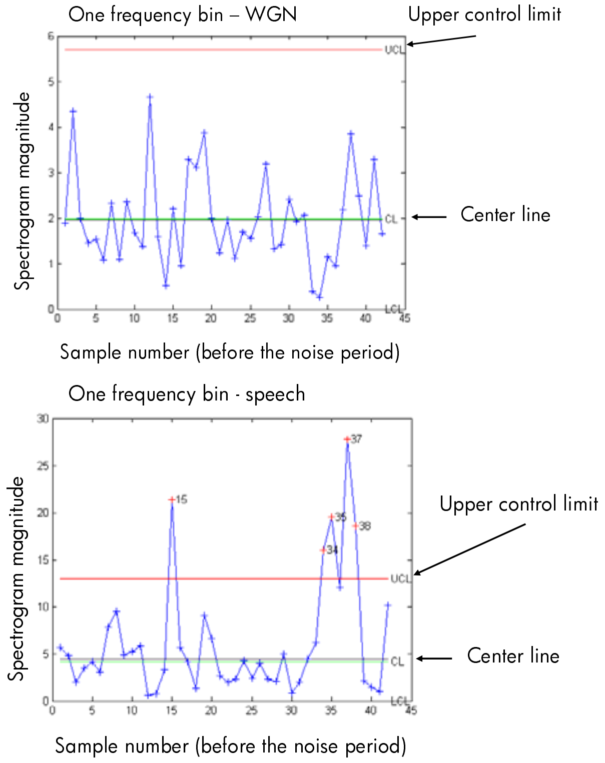

Fig. 5. Control charts using Rayleigh random variables for magnitude of spectra for the example of white Gaussian noise input on top, and speech at the bottom.

A one-sided control chart [5] was derived, with the Rayleigh distribution of Equation 5 used for the random variables in the *spectrogram frequency slice* (horizontal slice for each n, examples shown in Figure 4), for $0 \leq k_a < k_1$. Figure 5 shows control charts, for a single, mid-frequency spectral band, for two different input signals. The top shows the control chart when the input is white Gaussian random noise. In this case, the variable is in control. The bottom shows the control chart when the input is speech. In this case, it is seen that the variable is out of control (there are points outside the upper control limit).

The frequency slice is assumed unintelligible background when it remains within the control limits. In this case, $\alpha(n) = 1$. When intelligible components are otherwise detected, $\alpha(n)$ is set to

$$\alpha(n) = 1 - \frac{P_s(n)}{P_s(n) + P_\eta(n)}. \tag{6}$$

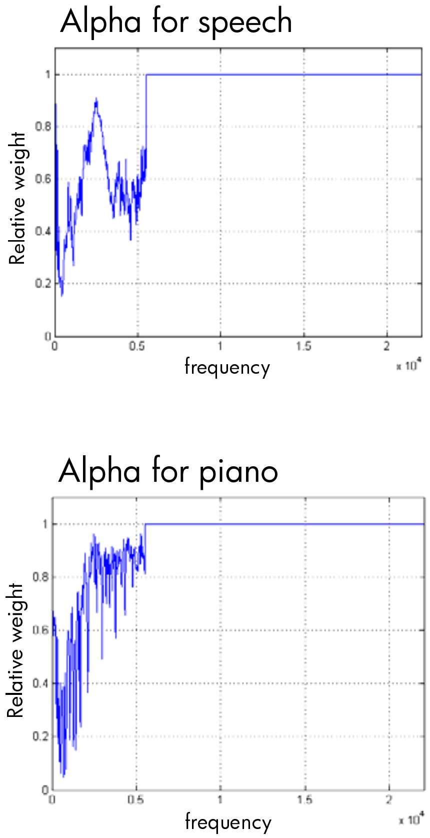

Fig. 6. The quantity $\alpha(n)$ for two different input signals. The top shows a speech example, and the bottom shows a piano example.

Here, $P_s(n)$ refers to the signal power for spectral band n, during the noise-free period, and $P_\eta(n)$ corresponds to a precalculated average noise power (during the noise period) for the same spectral band. The quantities P_s and P_n are normalized so that each of their sums, over spectral index, is one. This normalization is used to avoid, since the SNRs are very low, having degenerate $\alpha(n) \approx 1$. A final modification, useful for the case where the signal is mostly random, is made. If the proportion of spectral bands in control is near one, replace the in control band alphas with $\alpha = 1$.

Figure 6 shows two examples of the calculated $\alpha(n)$. The top of the figure shows an example where the input signal is speech, and the bottom of the figure shows an example where the input signal is piano. In the figure, because the audio signals were resampled, $\alpha(n) = 1$ for the higher frequency spectral bands.

3. EXPERIMENTAL RESULTS

The results were first tested with zoom motor noise samples captured by digital camera video. Since the noise is nonstationary, as seen in the top left spectrogram of Figure 2, a time-dependent spectral subtraction was used for the traditional processing. Training was done with 12 zoom noise samples captured during quiet. Testing was done with independent samples. Informal, subjective evaluations found improved results over traditional processing. Particularly, when mostly unstructured signals occurred, the noise appeared largely eliminated. The results with structured signals varied, but the results were still improved over the traditional processing.

Representative results of simulations used to calculate SNRs are presented. A fifth-order autoregressive model was fit to the zoom noise data, reproducing its spectral characteristics. Simulated noise with this AR model, one second in length, with power adjusted so SNR $= -5$ db, was added to the middle of three second audio samples. Since the noise for these simulations is stationary, the attenuation method [3] was applied as the traditional processing block. Then, the SNRs for the new versus the standard method were compared.

AUDIO	SNR		SNR $\|x - ky\|$	
	new	standard	new	standard
speech	4.0	4.3	4.9	4.3
piano	6.9	6.5	7.2	6.5
sin	21.9	10.8	21.9	11.1
sweep	.86	12.6	.90	12.6
WGN	4.5	6.1	4.5	6.2

Table 1. SNR and scaled SNR for five simulations.

Table 1 shows results of the simulations for various input signals. The table shows two different comparisons. The first comparison, shown in the second and third columns, is standard SNR. The second comparison, shown in the fourth and fifth columns, is *scaled* SNR, where the estimate is allowed to have a scale factor, before comparing to the original. Studying the table shows that both the new and standard methods improve objective SNR values. From the examples, it is seen that signals which are constant in frequency (piano notes, sin waves) do the best, in terms of SNR, with the new method. In the important case of speech, the new method has slightly lower SNR, but slightly higher scaled SNR. This may be due to the particular way the $\alpha(n)$ are calculated, causing a slight, but overall gain change in the restored signal. Subjectively, even in the case of speech, the new method seems better than the standard method. The example with white Gaussian noise (WGN), the restored SNR for the new method is low because the signal is being synthesized, but the subjective result is dramatically better. Finally, an artificially generated frequency sweep has the worst SNR results possible with the new method because it erroneously predicts the frequency bands that the signal will have during the noise period.

Both the new and standard methods offer objective improvements in SNR. The difference in improvements between the two methods varies depending on the input signals. This is because the traditional methods minimize objective error criteria whereas the new method synthesizes a signal that may be perceptually similar to the original, but is not faithful to the original signal. Nevertheless, informal listening to the simulations preferred the new method.

4. CONCLUSIONS

The approach taken here makes the *easy problems* easy by using information before the known noise period. Many of the results seem subjectively equal or better than standard processing. The new method requires an additional memory buffer to store a length of spectrogram. If latency is allowed, non-causal estimates are possible by using the information after the noise period as well as the information before the noise period.

5. REFERENCES

[1] G. Davis, *Noise Reduction in Speech Applications*, CRC Press, 2002.

[2] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 2, pp. 113–120, Apr. 1979.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[4] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 3rd edition, 1991.

[5] *Statistical Quality Control Handbook, AT&T,* Western Electric, 2nd edition, 1958.