Audio Engineering Society

Convention Paper

Presented at the 123rd Convention
2007 October 5–8 New York, NY, USA

Suppression of Musical Noise Artifacts in Audio Noise Reduction by Adaptive 2D Filtering

Alexey Lukin[1] and Jeremy Todd[2]

[1] Dept. of Computational Mathematics and Cybernetics,
Moscow State University, Moscow, Russia
lukin@graphics.cs.msu.ru

[2] iZotope, Inc., Cambridge MA 02139, US
jeremy@izotope.com

ABSTRACT

Spectral attenuation algorithms for audio noise reduction often generate annoying musical noise artifacts. Most existing methods for suppression of musical noise employ a combination of instantaneous and time-smoothed spectral estimates for calculation of spectral gains. In this paper, a 2D approach to the filtering of a time-frequency spectrum is proposed, based on a recently developed Non-Local Means image denoising algorithm. The proposed algorithm demonstrates efficient reduction of musical noise, without creating "noise echoes" artifacts inherent in time-smoothing methods.

1. INTRODUCTION

In the spectral attenuation (or spectral subtraction) methods for reduction of stationary noises, short-time spectral estimates are used to adaptively calculate suppression gains for time-frequency spectral coefficients of a noisy audio signal [1]. Due to statistical variance of short-time spectral estimates, calculated gains can contain random oscillations leading to spurious time-frequency bursts of energy in the processed signal known as musical noise artifacts. They can be quite annoying and sometimes even more objectionable than the original noise.

This paper suggests using a novel 2D method of smoothing of time-frequency transform coefficients to reduce the musical noise. It is the development of a recently proposed Non-Local Means algorithm for image processing [2]. The key property of this algorithm is an implicit search for 2-dimensional feature patterns in the image and the use of these patterns to aid the process of denoising.

The rest of the paper is organized as follows. In section 2, we review existing methods of musical noise reduction. In section 3, we summarize the idea of a Non-Local Means method for image denosing and propose its hybridization with the DFT thresholding method. In section 4, we describe application of this 2D adaptive smoothing to time-frequency coefficients during spectral subtraction. In section 5, the results are discussed, and section 6 concludes the paper.

2. EXISTING ALGORITHMS

Simple methods for reducing musical noise artifacts include:

1. Overestimation of a noise power spectrum density, which suppresses more noise, but also more low-level signal components, leading to a signal distortion.

2. Imposing a lower limit on suppression gains, which leaves some part of noise unsuppressed and masks the musical noise.

3. Restricting the speed of change of suppression gains in time, which is similar to the introduction of attack and release times in multiband gates, leading to suppression of transients and introduction of "noise echoes" artifacts in the processed signal.

Beyond the simple methods, the commonly used method of Emphaim/Malah [3] uses "instantaneous" spectral estimates (called "a-posteriori") and time-smoothed spectral estimates (called "a-priori") to calculate suppression gains in a MMSE (minimum mean-square error) way, under certain assumptions on distribution of signal and noise. The idea of Ephraim-Malah method is to use time-smoothed SNR estimates at low SNR levels to reduce musical noise, and use instantaneous SNR estimates at high SNR levels to prevent smearing or suppression of target signal.

This method's shortcoming includes the presence of "noisy echoes": segments of weakly suppressed noise that follow target signals. They happen because at low SNR levels after transient signals, "a-priori" SNR estimates, used for suppression, are high due to simple 1st order recursive time averaging.

The algorithm described by Whipple in [4] utilizes a simple 2D analysis of magnitude spectrograms to find energy bursts that are localized in time and frequency. Such bursts are replaced with zero energy to suppress the musical noise.

Another similar approach for processing of a 2D spectrogram is described by Goh et al. in [5]. The local variance of coefficients is used for detection of musical noise, and a median filter is used to repair regions detected as musical noise.

In the work of Lin and Gourban [6], a non-adaptive 2D smoothing of a magnitude spectrogram is used to detect speech/noise regions by applying a magnitude threshold. The spectrogram for regions that are classified as noise is time-smoothed with a box filter. The speech regions are processed by the Epraim-Malah method.

The algorithm by Soon and Koh [7] uses a 2D Fourier transform applied to a matrix of time-domain STFT (short-time Fourier transform) windows, which is equivalent to applying a 1-dimensional DFT to every row of a complex spectrogram. This allows one to effectively analyze time correlations of STFT coefficients, but the frequency correlation of spectrograms is not exploited effectively. To alleviate this problem, the application of algorithm [4] is suggested as a post-processing step.

We propose an adaptive algorithm that accounts for 2D patterns in a time-frequency magnitude spectrogram and effectively suppresses musical noise using a recently-developed Non-Local Means approach for image processing [2]. Unlike many prior art approaches, this is not a decision-based algorithm, which makes it less sensitive to a possibly inaccurate detection of the noise level.

3. A NON-LOCAL MEANS ALGORITHM

3.1. Non-linear image denoising

Many algorithms for image denoising are based on adaptive smoothing by means of pixel averaging.

$$y_{i,j} = \sum_{(k,m)\in\Omega} x_{i+k,j+m} W(i,j,k,m) \qquad (1)$$

Pixels from some local window Ω around the currently processed pixel $x_{i,j}$ are compared to the current pixel using geometric (by position) and photometric (by value) distance [8].

$$W(i,j,k,m) \approx \exp\left(-\frac{(x_{i,j}-x_{i+k,j+m})^2}{h^2}\right) \cdot \exp\left(-\frac{k^2+m^2}{\rho^2}\right) \qquad (2)$$

Those pixels which are closer to the current one are averaged with a higher weight $W(i,j,k,m)$. Comparison of pixel values prevents blurring of image details.

3.2. The Non-Local Means algorithm

Recently, a Non-Local Means (NLM) algorithm for image denoising has been introduced [2]. Instead of comparing values of single pixels, the NLM algorithm compares the content of image patches surrounding these pixels.

$$W(i,j,k,m) \approx \exp\left(-\frac{\left\|v(x_{i,j})-v(x_{i+k,j+m})\right\|^2}{h^2}\right) \qquad (3)$$

Here $v(x)$ is a vector of pixel values from a geometric neighborhood of pixel x, which is usually defined as a square block centered at the pixel x. The range Ω for (k, m) in the NLM algorithm can be as large as a whole image, hence the name "non-local".

In this way, only pixels whose surrounding patches have a similar structure are averaged. This preserves image structure, fine repeating textures, and patterns significantly better than with previous image denoising methods.

The original NLM algorithm is very computationally expensive, but some optimizations are discussed in [9] and [8] allowing it to be used for real-time audio processing.

4. ADAPTIVE 2D SPECTROGRAM SMOOTHING

Two-dimensional magnitude spectrograms are images with prominent structure: repeating horizontal lines of instrument harmonics, vertical onsets of transients, and frequency-modulated harmonics of speech and vocals. It is clear that one-dimensional recursive smoothing of spectrograms will blur many such details, especially for non-stationary audio content. That's why a more precise, 2D adaptive method of smoothing is required.

4.1. Applying NLM algorithm to spectrograms

We propose to use a Non-Local Means algorithm for the smoothing of a spectrogram. It will be able to perform edge-directional smoothing and use repeating harmonic patterns as a guide toward adaptive averaging of spectrogram blocks.

The input data to the NLM algorithm is a 2D array of signal-to-noise ratios, i.e. real non-negative STFT magnitudes rated to noise thresholds for every STFT bin (if the noise is assumed white, the noise threshold can be set to a constant). Noise thresholds are assumed known; they can be learned from a noise-only section of the audio signal. A real-time estimation of noise thresholds is also possible, but is not a topic of this paper.

It should be noted that NLM algorithm has been designed to work with white noise. The noise in magnitude spectrogram $X[f,t]$ (reflecting a variance of short-term spectral estimates) is non-white because of correlation of spectral data in frequency due to STFT windowing and in time due to overlapping of STFT windows.

However this noise is a low-pass filtered white noise. Indeed, every column of the magnitude spectrogram is a low-pass filtered spectrum of white noise: $X[f,t] = Z[f,t] * H[f]$, where $Z[f,t]$ is the spectrum of non-windowed white noise, which is white along the frequency axis, and $H[f]$ is the frequency response of a zero-phase low-pass weighting window used for STFT. Similarly, every row of the magnitude spectrogram is a low-pass filtered white noise because downsampling of $X[f,t]$ in time produces spectra of uncorrelated portions of white noise.

This whiteness of a noisy spectrogram across the range of quefrencies (frequencies in a spectrogram space) that also contains the signal energy makes the application of the NLM algorithm possible.

The noise threshold h in the NLM algorithm's formula (3) defines the strength of desired spectrogram smoothing. It should be noted that even strong smoothing by the NLM algorithm leaves major structures in 2D image intact, but eliminates more small structures (beyond musical noise).

The resulting smoothed map of signal-to-noise ratios is suitable for use in spectral subtraction. Since the smoothing has reduced variations in the SNR map, the gain variations in spectral subtraction, leading to musical noise, will also be reduced.

4.2. Hybrid DFT thresholding + NLM smoothing algorithm

NLM is a novel and high-quality algorithm for image denoising, but still it has few specific artifacts. In [10], it is proposed to combine NLM denoising with a DFT thresholding method (DFTT), which is similar to the spectral subtraction, to achieve better image denoising performance.

For spectrogram smoothing, DFTT has several appealing advantages. The 2D discrete Fourier transform is able to compactly localize energy of repeating waves of arbitrary direction in a 2D signal. For the case of application to magnitude spectrograms, DFT is going to aid with a compact localization of harmonics, which are quasi-periodic signals that may not be parallel to the time-frequency axis in case of pitch modulation.

As described in [10], the DFTT algorithm subdivides the 2D image into overlapping blocks, applies a weighting window to each block, performs the 2D DFT transform, and applies gain reduction similar to that used in spectral subtraction algorithms. After gain reduction, the DFT is inverted, and windowing is applied again before pasting the reconstructed block in the resulting image.

For smoothing of spectrograms, we have added a DFTT stage after the NLM algorithm. The input data to DFTT are both noisy spectrogram and the one pre-processed by NLM algorithm. The second one is used for SNR estimation in the DFTT suppression rule, while the first one is undergoing the analysis/modification/synthesis cycle, as suggested in [10]. The noise spectrum in the DFTT is assumed to be white.

4.3. Implementation details

In our implementation, we have used the following parameters. Analysis and synthesis filter banks are based on a STFT with 50 ms long Hann windows that have a 75% overlap.

The NLM algorithm is using 8x8 blocks for pattern matching, and the search range is +/–8 bins along the frequency axis and [–16…+4] blocks along the time axis. A non-symmetrical search range is used to reduce the overall algorithm processing latency and favor post-echoes to pre-echoes of noise. The pasted block size is 4x4 bins, for the sake of optimization.

The DFTT algorithm is using 32x16 blocks, where 32 is the number of bins along frequency axis. We are using blocks elongated along frequency axis to more efficiently account for the harmonic structure of the spectrum in the DFTT algorithm. The analysis/synthesis hop of the DFT is 8 and 4 bins correspondingly. A 2D Hann window is used both for analysis and synthesis.

Particular noise thresholds are not given here, because they essentially depend on a user preference to the amount of reduction of musical noise. There is a trade-off between the amount of musical noise reduction and the suppression of minor details in the desired signal.

5. RESULTS

We have performed several experiments with a test sample compiled from a diverse audio material, including speech from the "Orator" database (2 male and 2 female utterances were taken) and various types of music, including transient content. We have used artificially generated additive white noise at different SNR levels. Below we present analysis of spectrograms, lis-

tening observations and PSNR figures. Corresponding audio samples and more detailed spectrograms can be found at this paper's web page [11].

Figures 1-6 show magnitude spectrograms of a 4-second fragment of the test signal containing music with vocals at SNR = 15 dB. We have selected a fragment containing a transient sibilant in vocals (around the middle of the spectrogram) and sharp transients of drums (the right part of the spectrogram).

Figure 3 presents the result of a simple spectral subtraction with per-bin attenuation, without any smoothing of a spectrogram. A musical noise is clearly visible as multiple unsuppressed "dots" of noise and very objectionable in listening tests.

In figure 4, the result of the Ephraim-Malah algorithm is presented. The musical noise has been reduced. But as a result of one-dimensional recursive smoothing, there are areas of low suppression after transient events, visible as "noisy tails". The length of these tails can be controlled by the recursive filtering coefficient, but reduction of time smoothing leads to increase in musical noise. Another drawback of the Ephraim-Malah algorithm is the excessive suppression of transients at low SNR levels (visible as reduced brightness of transients in figure 4, e.g. around 9 sec). This has resulted in reduction of overall PSNR level, compared to a simple spectral subtraction. However the overall sound is more pleasing due to reduction of a musical noise.

Figures 5 and 6 display spectrograms after NLM and NLM+DFTT algorithms. The musical noise is effectively suppressed in both time and frequency directions, and there are no noisy areas after transients. The transients after the NLM algorithm are somewhat suppressed, similarly to the Ephraim-Malah algorithm. However the NLM+DFTT algorithm has reduced suppression of transients while still preserving good suppression of musical noise.

Table 1 presents PSNR improvement after applying different noise reduction algorithms to test files with artificial noise at different SNR levels. In each case, the noise threshold level has been manually tuned to maximize PSNR.

These PSNR measurements show that simple methods for musical noise reduction often result in worse PSNR figures than a regular spectral subtraction, due to excessive suppression of parts of a signal. However listening tests usually demonstrate the preference of having the musical noise reduced. At the same time, the proposed algorithm for musical noise reduction achieves some improvement of PSNR, compared to a regular spectral subtraction.

Method \ **SNR**	**25 dB**	**15 dB**	**5 dB**
Simple subtraction	4.44	6.69	9.74
Ephraim-Malah	3.96	5.98	9.46
NLM smoothing	4.37	6.56	9.61
NLM+DFTT	**4.53**	**6. 79**	**9.98**

Table 1. Improvement in PSNR after noise reduction

The PSNR figures of our algorithm slightly depend on the amount of musical noise reduction. A side effect of the excessive musical noise reduction with our algorithm is modulation of noise by a signal and suppression of details of the desired signal.

Our implementation of the algorithm runs marginally slower than real time on a 3 GHz P4 workstation for a mono 44.1 kHz audio signal. It is possible to significantly reduce the computational complexity by using larger analysis/synthesis hops in the DFTT algorithm and larger pasted block size in the NLM algorithm, at the expense of a slight quality reduction. Also, the algorithm allows easy parallelization for multi-core processors.

6. CONCLUSION

Many simple algorithms for musical noise reduction result in deterioration of PSNR compared to a regular spectral subtraction, due to excessive suppression of desired signal details.

The proposed algorithm using adaptive 2D spectrogram smoothing achieves effective reduction of musical noise artifacts with minimal damage to the target signal. The algorithm is based on a novel Non-Local Means image denoising method combined with a DFT thresholding method, which are applied to a 2D magnitude spectrogram. Good results in listening tests are supported by inspection of spectrograms and PSNR measurements.

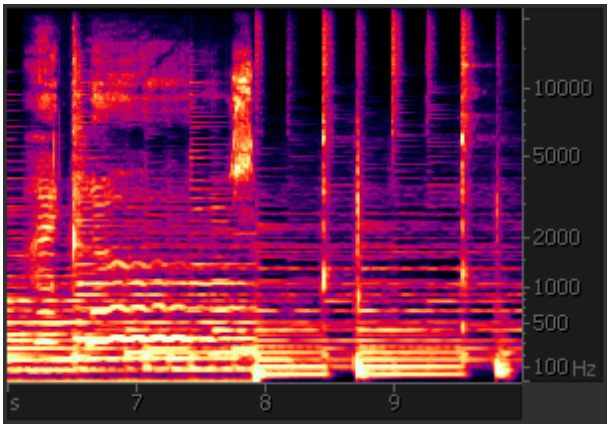

Figure 1. Clean fragment of the test signal

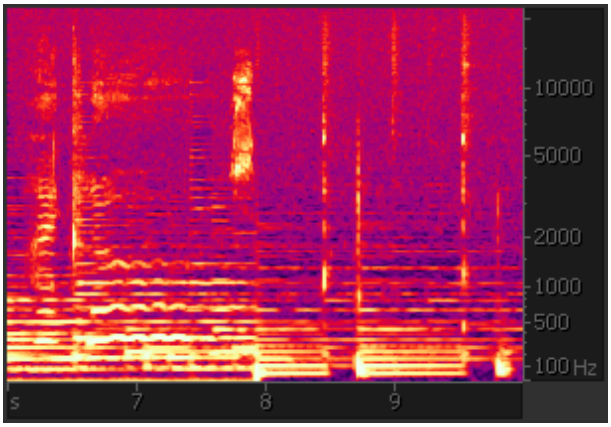

Figure 2. Noisy signal, SNR = 15 dB.

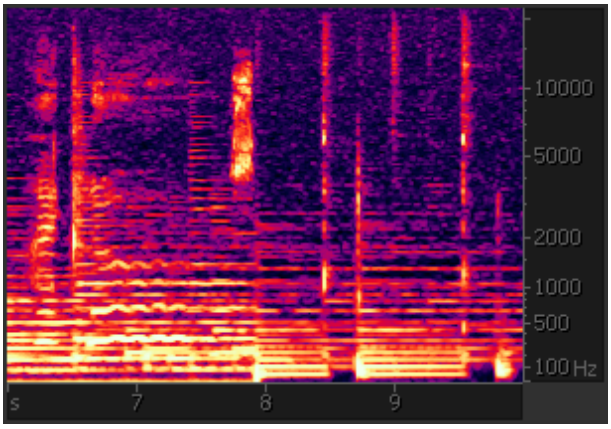

Figure 3. Simple spectral subtraction

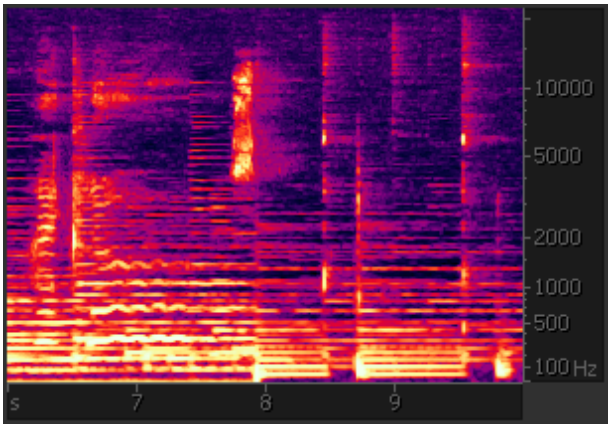

Figure 4. Ephraim-Malah spectral subtraction

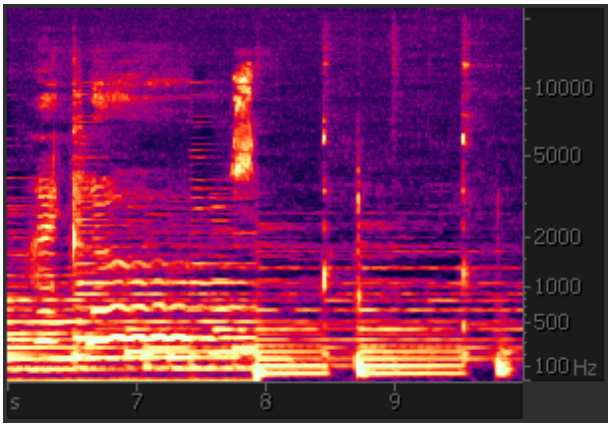

Figure 5. NLM smoothing of spectrogram

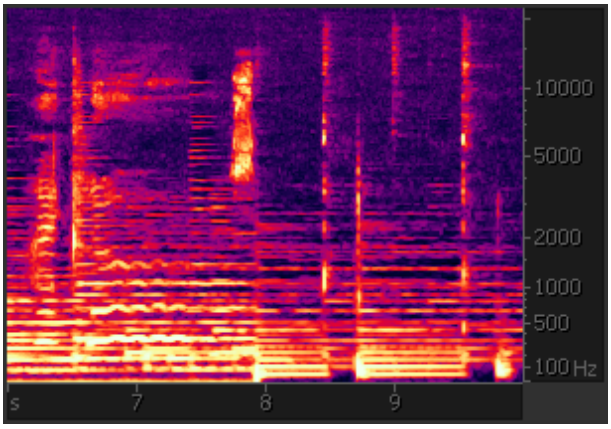

Figure 6. NLM+DFTT smoothing of spectrogram

7. REFERENCES

[1] J. Thiemann “Acoustic Noise Suppression for Speech Signals Using Auditory Masking Effects” // Ph.D. thesis, Department of Electrical & Computer Engineering, McGill University, Montreal, Canada, July 2001.

[2] A. Buades, B. Coll, J. Morel “Image Denoising By Non-Local Averaging” // Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, vol. 2, pp. 25-28, March 18-23, 2005.

[3] Y. Ephraim and D. Malah “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator" // IEEE Trans. Acoustics, Speech, Signal Processing, vol. 32, pp. 1109-1121, Dec. 1984.

[4] G. Whipple “Low residual noise speech enhancement utilizing time-frequency filtering” // Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I/5-I/8, Apr. 1994.

[5] Z. Goh, K.C. Tan, and B.T.G. Tan “Postprocessing Method for Suppressing Musical Noise Generated by Spectral Subtraction” // IEEE Transactions on Speech and Audio Processing, vol. 6, no. 3, pp. 287-292, May 1998.

[6] Z. Lin, R. Gourban “Musical noise reduction in speech using two-dimensional spectrogram enhancement” // Proceedings of the 2nd IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications, pp. 61-64, Sept. 2003.

[7] I.Y. Soon, S.N. Koh “Speech Enhancement Using 2-D Fourier Transform” // IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp. 717-724, Nov. 2003.

[8] A. Lukin “A Multiresolution Approach for Improving Quality of Image Denoising Algorithms” // IEEE International Conference On Acoustics, Speech, and Signal Processing (ICASSP-2006), Toulouse, France, 2006.

[9] A. Buades “Image and film denoising by non-local means” // Ph.D. thesis, Universitat de les Illes Balears, Spain, 2006.

[10] S. Putilin, A. Lukin “Modifications of a Non-Local Means Denoising for Video” (in Russian) // Proceedings of Grahicon’2007 International Conference on Computer Graphics, Moscow, Russia, June 2007.

[11] Demo web-page with audio examples: http://www.izotope.com/tech/aes_suppr