BLOCK THRESHOLDING ALGORITHM FOR ENHANCEMENT OF AN AUDIO SIGNAL CORRUPTED BY NOISE

V. Harini, B. Sindhu, G. Sasi kumar & Habibulla Khan

Abstract

Removing noise from audio signals requires a non-diagonal processing of time-frequency coefficients to avoid producing “Musical noise”. State of the art algorithms perform a parameterized filtering of spectrogram coefficients with empirically fixed parameters. A block thresholding estimation procedure is introduced, which adjusts all parameters adaptively to signal property by minimizing a Stein estimation of the risk.

Keywords: Audio denoising, Ephraim and Malah, Block thresholding, power subtraction.

1. Introduction

Audio signals are often contaminated by background environment noise and buzzing or humming noise from audio equipments. Audio denoising aims at attenuating the noise while retaining the underlying signals. Applications such as music and speech restoration are numerous. This paper introduces a new non-diagonal audio denoising algorithm through adaptive time-frequency block thresholding[1]. Block thresholding has been introduced by Cai and Silverman in mathematical statistics to improve the asymptotic decay of diagonal thresholding estimators. For audio time-frequency denoising, we show that block thresholding regularizes the estimate and is thus effective in musical noise reduction. Block parameters are automatically adjusted by minimizing a Stein estimator of the risk, which is calculated analytically from the noisy signal values.

2. State of the Art

2.1. Time-Frequency Audio Denoising

Time-frequency audio denoising procedures compute a short-time Fourier transform or a wavelet transform or a wavelet packet transform of the noisy signal, and processes the resulting coefficients to attenuate the noise. These representations reveal the time-frequency signal structures that can be discriminated from the noise. We concentrate on the coefficient processing as opposed to the choice of representations. Numerical experiments are performed with short-time Fourier transforms that are most commonly used in audio processing. The audio signal ‘f’ is contaminated by a noise that is often modeled as a zero-mean Gaussian process independent of ‘f’:

\[ y[n] = f[n] + \varepsilon[n], n = 0,1, \ldots, N-1. \]

A time-frequency transform decomposes the audio signal over a family of time-frequency atoms \( \{ g_{l,k} \} \) where \( l \) and \( k \) are the time and frequency (or scale) localization indices. The resulting coefficients shall be written

\[ Y[l, k] = \langle y, g_{l,k} \rangle = \sum_{n=0}^{N-1} y[n] g_{l,k}^* [n] \]

Where \( * \) denotes the conjugate. These transforms define a complete and often redundant signal representation. In this paper we shall suppose that these time-frequency atoms define a tight frame, which means that there exists \( A > 0 \) such that

\[ \| y \|^2 = \frac{1}{A} \sum_{l,k} | \langle y, g_{l,k} \rangle |^2 \]

This implies a simple reconstruction formula

\[ y[n] = \frac{1}{A} \sum_{l,k} Y[l, k] g_{l,k} [n] \]
The constant $A$ is a redundancy factor and if $A=1$ then a tight frame is an orthogonal basis. A tight frame behaves like a union of orthogonal bases. A frame representation provides an energy control. The redundancy implies that a signal ‘$f$’ has a non unique way to be reconstructed from a tight frame representation:

$$f[n] = (1/A) \sum_{l,k} C[l,k] g_{l,k}[n]$$

but all such reconstructions satisfy

$$\|f\|^2 = \frac{1}{A} \sum_{l,k} |c[l,k]|^2$$  \hspace{1cm} (2)

with an equality if, $C[l,k] = \langle f, g_{l,k} \rangle, \forall l,k$.

Short-time Fourier atoms can be written:

$$g_{l,k}[n] = w[n- lu] \exp(i2\pi kn/K).$$

Where $\omega(n)$ is a time window of support size $k$, which is shifted with a step $u \leq k$. ‘$l$’ and ‘$k$’ are respectively the integer time and frequency indexes with $0 \leq l \leq N/u$ and $0 \leq k \leq K$. In this paper, $\omega(n)$ is the square root of a Hanning window and $u=k/2$ so one can verify that the resulting window Fourier atoms define a tight frame with $A=2$. A denoising algorithm modifies time-frequency coefficients by multiplying each of them by an attenuation factor $a[l,k]$ to attenuate the noise component. The resulting “denoised” signal estimator is

$$\hat{f}[n] = \frac{1}{A} \sum_{l,k} \hat{F}[l,k] g_{l,k}[n] = \frac{1}{A} \sum_{l,k} a[l,k] Y[l,k] g_{l,k}[n]$$ \hspace{1cm} (3)

Time-frequency denoising algorithms differ through the calculation of the attenuation factors $a[l,k]$. The noise coefficient Variance

$$\sigma^2[l,k] = E[\langle \epsilon, g_{l,k} \rangle^2]$$

is supposed to be known or estimated with methods such as [1], [2]. If the noise is stationary, which is often the case, then the noise variance does not depend upon time:

$$\sigma^2[l,k] = \sigma^2[k].$$

2.2. Non diagonal Estimation

To reduce musical noise as well as the estimation risk, several authors have proposed to estimate the a priori SNR $\xi[l,k]$ with a time-frequency regularization of the a posteriori SNR $\gamma[l,k]$. Non diagonal estimators clearly outperform diagonal estimators but depend upon regularization filtering parameters. Large regularization filters reduce the noise energy but introduce more signal distortion. It is desirable that filter parameters are adjusted depending upon the nature of audio signals. In practice, however, they are selected empirically. Moreover, the attenuation rules and the a priori SNR estimators that are derived with a Bayesian approach [4], model audio signals with Gaussian, Gamma or Laplacian processes. Although such models are often appropriate for speech, they do not take into account the complexity of other audio signals such as music that include strong attacks.

3. TIME-FREQUENCY BLOCK THRESHOLDING

3.1. Block Thresholding Algorithm

A time-frequency block thresholding estimator regularizes power subtraction estimation by calculating a single attenuation factor over time-frequency blocks. The time-frequency plane $\{l,k\}$ is segmented in $I$ blocks $B_i$, whose shape may be chosen arbitrarily. The signal estimator is calculated from the noisy data $y$ with a constant attenuation factor $a_i$ over each block $B_i$.

$$\hat{f}[n] - \sum_{i=1}^{I} \sum_{(l,k) \in B_i} a_i Y[l,k] g_{l,k}[n]$$ \hspace{1cm} (4)

To understand how to compute each $a_i$, one relates the risk ‘$r$’ to the frame energy conservation (2) and obtains

$$r = E[\|f - \hat{f}\|^2]$$

$$\leq \frac{1}{A} \sum_{i=1}^{I} \sum_{(l,k) \in B_i} E[|a_i Y[l,k] - F[l,k]|^2]$$ \hspace{1cm} (5)

Since $Y[l,k]= F[l,k] + \epsilon[l,k]$ one can verify that the upper bound of (5) is minimized by choosing
Where $\xi_i = F_i^2 / \sigma_i^2$ is the average a priori SNR in $i$. It is calculated from
\[
F_i^2 = \frac{1}{B_i^r} \sum_{(l,k) \in B_i} |F[l,k]|^2 \quad \text{and}
\]
\[
\sigma_i^2 = \frac{1}{B_i^r} \sum_{(l,k) \in B_i} \sigma^2[l,k].
\]
Which are the average signal energy and noise energy in $B_i$ and $B_i^r$ is the number of coefficients $(l, k) \in B_i$. The resulting oracle block risk $r_{bo}$ satisfies
\[
r_{bo} \leq A \cdot R_{bo}, \quad \text{where}
\]
\[
R_{bo} = \sum_{i=1}^I \frac{F_i^2 \sigma_i^2}{F_i^2 + \sigma_i^2}.
\]
The oracle block attenuation coefficients $a_i$, in (6) cannot be calculated because the a priori SNR $\xi_i$ is unknown. Cai and Silverman [4] introduced block thresholding estimators $B_i$ that estimate the SNR over each by averaging the noisy signal energy
\[
\hat{\xi}_i = \frac{\bar{Y}_i^2}{\sigma_i^2} - 1 \quad (7)
\]
Where $\bar{Y}_i^2 = \frac{1}{B_i^r} \sum_{(l,k) \in B_i} |Y[l,k]|^2$

Observe that if $\sigma[l,k]=\sigma$, for all $(l,k) \in B_i$, then (7) is an unbiased estimator of $\xi_i$. The resulting attenuation factor $a_i$ is a block thresholding estimator can thus be interpreted as a non diagonal estimator derived from averaged SNR estimations over blocks. Each attenuation factor is calculated from all coefficients in each block, which regularizes the time-frequency coefficient estimation. Computed with a power subtraction estimator
\[
a_i = (1 - \frac{\lambda}{\xi_i + 1}). \quad (8)
\]

### 3.2. Block Thresholding Risk & Choice of $\lambda$

An upper bound of the risk of the block thresholding estimator is computed by analyzing separately the bias and variance terms. Observe that the upper bound of the oracle risk in $r_{bo}$ with blocks is always larger than that of the oracle risk in $r_c$ without blocks, because the former is obtained through the same minimization but with less parameters as attenuation factors remain constant over each block. A direct calculation shows that [see (9)].
\[
R_{bo} - R_c = \sum_{i=1}^I \sum_{(l,k) \in B_i} \frac{\xi_i^2[l,k] \sigma_i^2 - \sigma_i^2[l,k] + (\bar{F}_i^2 - |F[l,k]|^2)}{(\xi_i + 1)(\xi_i[l,k] + 1)} \geq 0 \quad (9)
\]

### 3.3. Adaptive Block Thresholding

A block thresholding segments the time frequency plane in disjoint rectangular blocks of length $L$, in time and width $W$, in frequency. In the following by “block size” we mean a choice of block shapes and sizes among a collection of possibilities. The adaptive block thresholding chooses the sizes by minimizing an estimate of the risk. The risk ‘r’ cannot be calculated since $f$ is unknown, but it can be estimated with a Stein risk estimate. Best block sizes are computed by minimizing this estimated risk. We saw in (5) that the block thresholding risk satisfies
\[
r = E[\| f - \hat{f}^2 \|] \leq \frac{1}{A} \sum_{i=1}^I \sum_{(l,k) \in B_i} E[|a_i Y[l,k] - F[l,k]|^2] \quad (10)
\]
Since $Y[l,k]=F[l,k]+e[l,k]$ and has a zero mean, $F[l,k]$ is the mean of $Y[l,k]$ .To estimate the block thresholding risk Cai uses the Stein estimator of the risk when computing the mean of a random vector, which is given by Stein theorem [3]. A block thresholding algorithm regularizes the time-frequency estimation as compared to a diagonal thresholding, but it outputs a time frequency estimation with some block structures.
Fig. 1: Partition of macro blocks into blocks of different sizes.

3.4. Non diagonal Wiener Post processing and Masking Noise
This first estimation is used as an input to compute a Wiener time-frequency estimation that takes advantage of the time-frequency regularization provided by the block thresholding estimation. Retaining a low-amplitude noise is sometimes desirable to mask artifacts generated by an estimation procedure [2]. Following [2], one can retain a masking noise by setting a floor value to the attenuation factor:

$$\tilde{a}_M[l,k] = \max(\tilde{a}[l,k], a_o)$$

(11)

Where $0 < a_o << 1$ minimum attenuation factor of the noise.

4. RESULTS
The experiments presented below can also performed for various types of audio signals: Audio signal Babble is musical excerpt that contains relatively quick notes. Short-time Fourier transform with half-overlapping windows were used in the experiments.

Fig 2: Original signal
5. CONCLUSION
Non diagonal time-frequency estimators are more effective than diagonal estimators to remove noise from audio signals because they introduce less musical noise. These non diagonal estimators are derived from a time-frequency SNR estimation performed with parameterized filters applied to time frequency coefficients. This paper introduces an adaptive audio block thresholding algorithm that adapts all parameters to the time-frequency regularity of the audio signal. The adaptation is performed by minimizing a Stein unbiased risk estimator calculated from the data. The resulting algorithm is robust to variations of signal structures such as short transients and long harmonics.

6. REFERENCES