PERSISTENT EMPIRICAL WIENER ESTIMATION WITH ADAPTIVE THRESHOLD SELECTION FOR AUDIO DENOISING

Kai Siedenburg
Austrian Research Institute for Artificial Intelligence (OFAI)
kai.siedenburg@ofai.at

ABSTRACT

Exploiting the persistence properties of signals leads to significant improvements in audio denoising. This contribution derives a novel denoising operator based on neighborhood smoothed, Wiener filter like shrinkage. Relations to the sparse denoising approach via thresholding are drawn. Further, a rationale for adapting the threshold level to a performance criterion is developed. Using a simple but efficient estimator of the noise level, the introduced operators with adaptive thresholds are demonstrated to act as attractive alternatives to the state of the art in audio denoising.

1. INTRODUCTION

Recovering an audio signal from disruptive noise - audio *denoising* - is a classical problem in signal processing. It has numerous applications in telecommunication and speech/music restoration while encompassing various subproblems, mostly specified by the type of noise and the signal model under usage [1]. The non parametric model with observations $y = f + e$, where f denotes the signal and e the additive noise, is a setting that has attracted particular attention because of its mathematical simplicity and generality.

Nowadays, the latter model is often equipped with the additional assumption of *sparsity*, which refers to the circumstance that many natural signals can be expanded (using a suited dictionary Φ) with only few non zero coefficients. The initial ℓ_1-norm approach [2, 3] models the sparse coefficients of the denoised signal as the minimizer of

$$\|y - \Phi c\|^2 + 2\lambda\|c\|_1 \tag{1}$$

It has been refined considerably and recent developments include wavelet-based denoising [4], and generalizations to mixed norm regularization and multilayered decompositions [5]. Moreover, efficient neighborhood-based heuristics have been proposed, see [6], to further account for signal structures while keeping the non parametric signal model. This approach leads to estimates which are sparse and structured, but not necessarily optimal when considered with respect to the estimation risk, i.e. its expected quadratic error.

On the other hand, it is inherent to the sparse denoising approach (1) that the threshold λ, regulating the sparsity of the solution (the larger the sparser), must be determined separately, since no signal model was assumed. This implies that the threshold λ must often be tuned by hand in order to optimize the respective performance criterion (e.g. SNR [1] or perceptual judgements). Although some suggestions towards adaptive selection of the threshold λ have been made in the wavelet literature, cf. [7, 8], this question appears to be untouched for audio denoising using time frequency dictionaries.

The purpose of this this paper thus is two-fold. Concerning the operator design, Section 2 derives a novel audio denoising operator, the *persistent empirical Wiener* estimate, which fuses recent developments in the field of structured sparsity with the properties of empirical Wiener filtering. Regarding the aspect of threshold selection, Section 3 proposes a rationale for adaptive threshold selection according to a given performance criterion [2] and compares it to other common threshold choices. It turns out that a plain linear model depending on the level of the noise achieves minor performance differences compared to the optimal thresholds. Section 4 then proposes a simple method for estimating this noise level in case it is unknown. Finally, Section 5 demonstrates that the proposed operators perform competitively compared to the state of the art, while being much more computationally efficient and robust to minor perturbation of the noise level.

[1] The *signal to noise ratio* of signals $f, \hat{f} \in \mathbb{R}^p$ is defined by $10\log_{10}\left(\frac{\sum_{n=1}^{p}|f(n)|^2}{\sum_{n=1}^{p}|f(n)-\hat{f}(n)|^2}\right)$.

[2] In this article we only work with the SNR. Various alternative objective measures have been proposed, see e.g. [9], in order to better model human perception. However, to our experience, these are either highly correlated with the regular SNR, or not robust which sometimes produces *very* counterintuitive results. Future work hence must evaluate the same algorithms w.r.t. listening tests using human subjects on concrete denoising stimuli.

2. PERSISTENT TIME-FREQUENCY SHRINKAGE

2.1 The Signal Model

Let us assume we observe a signal $f \in \mathbb{R}^p$ which is deteriorated by Gaussian white noise $e \sim \mathcal{N}(0, \sigma^2 I)$, I denoting the unit matrix in $\mathbb{R}^p$. That is, we consider the standard additive noise model

$$y(n) = f(n) + e(n), \quad n = 1, \ldots, p$$

where it is assumed that $f = \Phi c$ with sparse synthesis coefficients c and time-frequency transform $\Phi : \mathbb{R}^p \to \mathbb{R}^p$. The noise level σ is assumed to be known (while section 4 provides estimates $\hat{\sigma}$ for the contrary situation).

In this paper, we work with the *Modified Discrete Cosine Transform* (MDCT), which is a particularly influential lapped orthogonal time frequency transform, see e.g. [8]. For indices $\gamma = (l, k)$, l denoting time and k frequency, its atoms $\phi_{l,k}$ (constituting the columns of Φ) can be expressed by

$$\phi_{l,k}(n) = g_l(n)\sqrt{\frac{2}{L}} \cos\Big[\frac{\pi}{L}\big(k + \frac{1}{2}\big)(n + n_l)\Big].$$

Here, $n_l = (L+1)/2 - lL$, $k = 0, \ldots, L-1$, and L is the half of the window length as well as the hop size. The window g is usually chosen as a sinusoid:

$$g_l(n) = \sin\Big[\frac{\pi}{2L}\big(n - lL + \frac{L}{2}\big)\Big].$$

On the global signal (not on local frames), the MDCT is an orthonormal basis, i.e. $\Phi^*\Phi = I$, where Φ^* stands for the adjoint analysis operation $c^* = \Phi^* f = (\langle f, \phi_\gamma \rangle)_{\gamma \in \Gamma} \in \mathbb{R}^p$, i.e. c^* denotes the clean analysis coefficients of the signal f.

The orthonormality of the MDCT has many convenient mathematical consequences. One of them imposes that white noise transforms into white noise: $z := \Phi^* e \sim \mathcal{N}(0, \sigma^2 I)$, which will simplify many of the following discussions.

2.2 Denoising by Diagonal Estimation

Using time frequency dictionaries, a natural first approach to denoising is to process the analysis coefficients y^* of the observation y such that the overall estimation risk is minimized, as excercised in classical diagonal estimation.

2.2.1 The Empirical Wiener

Let $D : \mathbb{R}^p \to \mathbb{R}^p$, $D = \mathrm{diag}(d_1, \ldots, d_p)$ be a diagonal operator with non-negative $d_\gamma \geq 0$. The diagonal estimate $\hat{f} = \Phi D y^*$ of f then simply is a reconstruction of d_γ-weighted analysis coefficients. The estimation quality is measured by the squared ℓ_2 risk

$$R(f, \hat{f}) = \mathbb{E}[\|f - \hat{f}\|^2] = \mathbb{E}\big[\sum_{n=1}^{p} |f(n) - \hat{f}(n)|^2\big]$$

where $\mathbb{E}$ denotes the expectation operator. We are interested in choices of D which minimize this risk. Due to the orthogonality of Φ, the situation can be directly transferred to the analysis domain where a straight forward computation shows that

$$R(f, \hat{f}) = \sum_{\gamma \in \Gamma} d_\gamma^2(\sigma^2 + c_\gamma^{*\,2}) - 2 d_\gamma c_\gamma^{*\,2} + c_\gamma^{*\,2}$$

Minimizing this expression component wise yields

$$d_\gamma^{or} = \frac{c_\gamma^{*\,2}}{\sigma^2 + c_\gamma^{*\,2}} = 1 - \frac{1}{1 + \sigma^{-2} c_\gamma^{*\,2}} \qquad (2)$$

Since these optimal *oracle* shrinkage weights d_γ^{or} depend on the *a priori* signal to noise ratio (SNR) $\xi_\gamma := c_\gamma^{*\,2}/\sigma^2$, they are unknown. However, a simple calculation shows that

$$\hat{\xi}_\gamma = \sigma^{-2} y_\gamma^{*\,2} - 1 \qquad (3)$$

is an unbiased estimator of exactly this a priori SNR, $\mathbb{E}[\hat{\xi}_\gamma] = \xi_\gamma$. It is hence reasonable to work with $\hat{\xi}$ instead, cf. [10]. Since $d_\gamma \geq 0$, combining (2) and (3) yields the shrinkage weights $d_\gamma = (1 - \sigma^2/|y_\gamma^*|^2)_+$ where as usual $(a)_+ = \max(a, 0)$ and $\frac{1}{0} := \infty$. Due to the resemblance of Wiener filters, they have been dubbed *empirical Wiener attenuation* (EW), cf. [8, 11, 12]. The corresponding empirical Wiener diagonal estimation, shrinking each coefficient towards zero if not completely discarding it, then writes component wise as $\mathbb{S}^{EW}(z_\gamma) := z_\gamma d_\gamma$.

2.2.2 Relations to Other Operators

Let us note that the empirical Wiener, also discussed as *nonnegative garrote* [13], which aims for minimal risk is closely related to the so called Basis Pursuit Denoising [2] or *LASSO* (L) [3], targeting at a sparse signal expansion in the presence of noise. The latter is defined as the minimizer of expression (1). In case of Φ being orthonormal, the minimizer of this problem is simply given by the well-known soft thresholding operation $\mathbb{S}_\lambda^L := \mathbb{S}_\lambda^{ST} = \mathrm{sign}(\cdot)(|\cdot| - \lambda)_+$. From a more general point of view, the soft thresholding operator can be rephrased by

$$\mathbb{S}_\lambda^\alpha(y^*) := y^* \left(1 - \left[\frac{\lambda}{\|y^*\|_\star}\right]^\alpha\right)_+ \qquad (4)$$

for $\alpha = 1$ and $\|\cdot\|_\star = |\cdot|$. The same framework now encompasses the empirical Wiener: $\mathbb{S}^{EW} = \mathbb{S}_\sigma^2$, i.e. it can be viewed as a soft-threshold operation with threshold σ and increased exponent $\alpha = 2$. In particular, this implies that for orthogonal bases, the Lasso $\mathbb{S}_\sigma = \mathbb{S}_\sigma^1$ and empirical Wiener $\mathbb{S}^{EW} = \mathbb{S}_\sigma^2$ shrinkage impose the same significant map of coefficients, i.e. select the same model for any given coefficients y^*.

Moreover, the generalized shrinkage (4) recovers the influential (positive-part) James-Stein shrinkage estimate with $\lambda = \sigma(p-2)$, $\|\cdot\|_\star = \|\cdot\|_2$, and $y^* = m^{-1}\sum_{i=1}^m x_i$, where $x_i \sim \mathcal{N}(\mu, \sigma^2 I)$ i.i.d. for $i = 1, \ldots, m$. James and Stein [14] proved that for dimensions $p \geq 3$ this shrinkage (even without the strictly non-negative part) yields smaller ℓ_2 risk than the arithmetic mean (i.e. the maximum likelihood estimate) in the estimation of the expected value $\mu \in \mathbb{R}^p$. Later on in the context of wavelet denoising, Cai generalized the approach to account for grouping structures in the coefficients by introducing a *block-James-Stein estimate* which estimates attenuation factors for disjoint blocks of coefficients separately [4]. This principle of non-diagonal estimation was exploited successfully in [10] for audio denoising. The authors showed that their block-James-Stein based algorithm improved the state of the art with respect to signal to noise ratio (SNR) and perception, as it managed to eliminate the severe artifacts of musical noise which are natural consequences of diagonal denoising.

2.3 From Diagonal to Persistent Estimation

In between the extremes of processing each coefficient independently and acting on disjoint blocks of coefficients, there lies a neighborhood based approach which considers sliding windows of mutual dependency. It was introduced in [6] from the stance of structured sparse signal approximation, termed Windowed Group Lasso (WGL), and evaluated for denoising tasks in [15]. Here, we derive a related operator, the *persistent empirical Wiener*[3] which combines neighborhood persistence and empirical Wiener shrinkage. In light of (4), this operator then will only differ from WGL in the subtle switch from $\alpha = 1$ to $\alpha = 2$.

2.3.1 A Persistent Empirical Wiener Estimator

Let us consider a regularized version of the SNR estimate (3), which is based on local smoothing of the coefficients y^*. This persistent SNR estimate is of the form

$$\xi_\gamma^* := \sigma^{-2}\sum_{\gamma'\in\Gamma} w_{\gamma\gamma'}|y_{\gamma'}^*|^2 \; - 1 \tag{5}$$

where for each index γ, $w_\gamma = (w_{\gamma\gamma'})_{\gamma'\in\Gamma}$ is a sequence of non-negative and normalized neighborhood weights such that $\sum_{\gamma'\in\Gamma} w_{\gamma\gamma'} = 1$ and $\forall\gamma \in \Gamma : w_{\gamma\gamma} > 0$. The corresponding *persistent empirical Wiener* (PEW) writes coordinate-wise as

$$\mathbb{S}(y_\gamma^*) := y_\gamma^* \left(1 - \frac{\sigma^2}{\sum_{\gamma'\in\Gamma} w_{\gamma\gamma'}|y_{\gamma'}^*|^2}\right)_+ \tag{6}$$

with $\mathbb{S}(y^*) := (\mathbb{S}(y_\gamma^*))_{\gamma\in\Gamma}$. Of course, by setting

$$\|y^*\|_\star = \sqrt{\sum_{\gamma'\in\Gamma} w_{\gamma\gamma'}|y_{\gamma'}^*|^2} \tag{7}$$

we obtain $\mathbb{S}^{PEW} = \mathbb{S}_\sigma^2$ in the terminology of (4). For neighborhood weights with support $\operatorname{supp}(w_\gamma) = \{\gamma\}$ as the singleton, PEW directly coincides with EW. For neighborhood weights which induce a disjoint partition of the index set Γ, PEW coincides (up to the dimensionality factor) with the block-James-Stein shrinkage estimator mentioned above. For the practical purposes of introducing continuous persistence to the estimation, it is natural to choose the weights $w_{\gamma\gamma'} = w_{\gamma-\gamma'}$ as sliding windows. Then the sum in (7) is nothing else than a moving average and can moreover be computed via fast convolution.

2.3.2 Comparing Risks

To compare EW and PEW, let us consider the squared error risk of the corresponding persistent and non persistent SNR estimators ξ_γ^* and $\hat{\xi}_\gamma$. For any estimator $\tilde{\xi}$ of an underlying true ξ, we naturally have the bias-variance decomposition of the risk

$$R(\tilde{\xi}, \xi) = \mathbb{E}[(\tilde{\xi} - \xi)^2] = \mathrm{Var}(\tilde{\xi}) + \mathrm{Bias}(\tilde{\xi})^2.$$

Since $\hat{\xi}$ is unbiased, one can check that

$$R(\hat{\xi}_\gamma, \xi_\gamma) = 2 + 4\sigma^{-2}{c_\gamma^*}^2 \tag{8}$$

In the persistent case, we similarly obtain

$$\mathrm{Var}(\xi_\gamma^*) = 2\sum_{\gamma'\in\Gamma} w_{\gamma\gamma'}^2 + 4\sigma^{-2}\sum_{\gamma'\in\Gamma} w_{\gamma\gamma'}^2 {c_{\gamma'}^*}^2 \tag{9}$$

$$\mathrm{Bias}(\xi_\gamma^*)^2 = \sigma^{-4}\Big(\sum_{\gamma'\in\Gamma} w_{\gamma\gamma'}{c_{\gamma'}^*}^2 - {c_\gamma^*}^2\Big)^2. \tag{10}$$

Consequently, we can explicitly compare the component wise risk of the persistent and non-persistent SNR estimators, given a clean signal f, neighborhood weights w and noise level σ. In case of $R(\hat{\xi}_\gamma, \xi_\gamma) \geq R(\xi_\gamma^*, \xi_\gamma)$ the persistent estimator should be preferable, at least from the stance of reducing estimation risk.

Since we did not assume any stochastic model of f, we must consider concrete numerical examples. Our first toy example features a well-known audio signal containing a simple phrase played by a glockenspiel. Here, we use a MDCT basis with window length of 512 samples at 44.1 kHz sampling rate. Figure 1 shows its time frequency representation. The bottom graphic shows the corresponding map of coefficients for which the above risk inequality holds, i.e. the coefficients which exhibit greater diagonal than persistent risk. This example assumes a noise level of $\sigma = 0.1$ and a neighborhood extending five coefficients in time only. Apparently, there are few coefficients for which the risk

[3] This operator was already used in [15] under the name WGL-James-Stein but not further motivated which is one aim of the current paper.

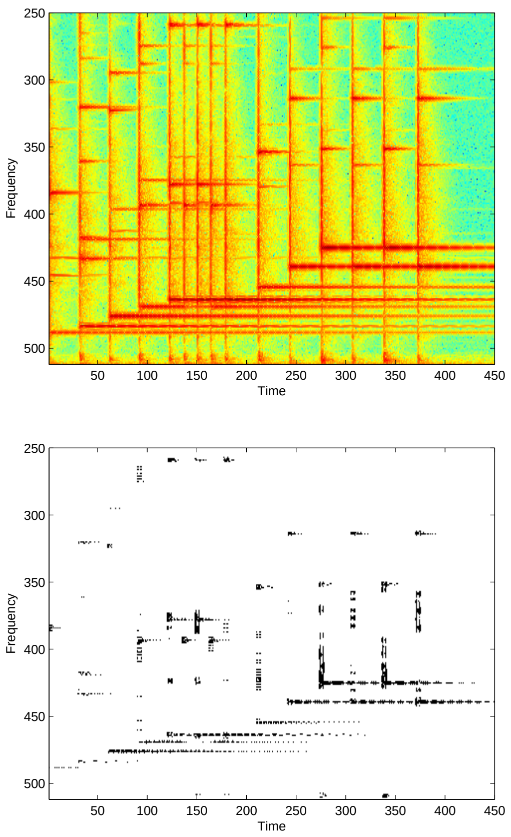

Figure 1. Top: time-frequency representation of a glockenspiel signal. Bottom: risk map of the same signal. Coefficients for which the risk of persistent SNR estimation is larger than that of non-persistent estimation are shown in black. Here, the neighborhoods extend 5 coefficients in time and the noise level was fixed with $\sigma = 0.1$.

of SNR estimation is not reduced by neighborhood persistence. Naturally, these appear close to the high energy coefficients, i.e. in areas where the energy variance is very high and, correspondingly, the bias with respect to the chosen neighborhood is very large. From a more global point of view, we can also consider the risk summed over all components $\sum_\gamma R(\xi^*_\gamma, \xi_\gamma)$ with respect to different noise levels and neighborhoods. Again using the glockenspiel signal, Figure 2 depicts the interdependence of the accumulated risk and one particular feature of the neighborhoods, namely their overall length in time which varies from 1 to 65 coefficients. The example suggests that for low noise levels non-persistent estimation (i.e. with neighborhood size 1) has the least cumulative risk. With increasing σ, larger neighborhoods achieve smaller risks; from this perspective a neighborhood encompassing 65 coefficients seems to be preferable as soon as $\sigma > 1$. In applications, however, this would lead to suboptimal results due to smearing and pre-echos, and it turns out that neighborhood sizes of 3 to 12 coefficients work well, see [15].

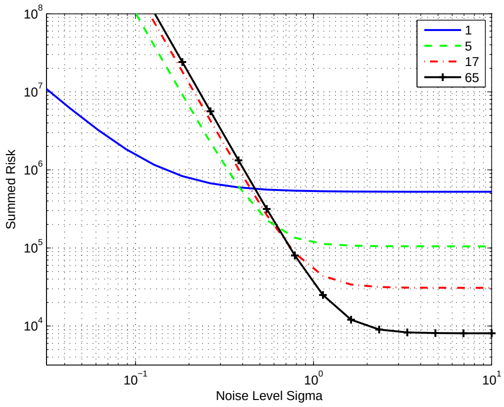

Figure 2. Summed Risk of the SNR estimation for different noise levels σ and neighborhood sizes. The latter extend over time only and encompass 1, 5, 17, and 65 coefficients with the 1st, 3rd, 9th, and 33rd as center coefficients.

3. ADAPTIVE THRESHOLD SELECTION

Let us now consider the setting of generalized shrinkage as exposed in (4) with $\alpha \in \{1, 2\}$ where $\| \cdot \|_\star$ either stands for the neighborhood weighted norm (7) or simply the absolute value $|\cdot|$. This yields the shrinkage operators Lasso (L), Windowed Group Lasso (WGL), empirical Wiener (EW) and persistent empirical Wiener (PEW). We are now interested in the question of how to chose the threshold λ in (4) for these operators. Therefore, we will briefly review common choices from the literature while their order of appearance corresponds to increasing stages of adaptivity. The second subsection proposes a simple novel method for adapting λ according to an empirical quality criterion.

3.1 Common Threshold Choices

3.1.1 Noise Level σ

A fully non adaptive and easiest possible threshold choice would be the noise level σ. For a zero signal $f \equiv 0$ this would imply to retain around one third of the overall noise, which of course seems very unsatisfying. Still, $\lambda = \sigma$ appears as a natural choice for the (P)EW as outlined above.

3.1.2 Universal

The so called *universal* threshold is given by

$$\lambda = \sigma\sqrt{2\ln(p)}.$$

Donoho and Johnstone showed in the context of wavelet shrinkage [16] that in conjunction with hard and soft thresholding (Lasso) it yields a risk which is very close to that of an oracle projector, i.e. nearly is a decision theoretic optimum. Slowly increasing with the signal length (where e.g. $\sqrt{2\ln(44100)} \approx 4.6$) the universal threshold often gives rise to estimates which are too sparse.

3.1.3 SURE

The *Stein unbiased risk estimate (SURE)* (not to be confused with the JS-estimate from above) provides a tool for adapting the threshold to the actual data such that the estimation risk is minimized. Stein's initial contribution [17] was to show that given a multivariate normal observation $y \sim \mathcal{N}(\mu, I)$, $\mu \in \mathbb{R}^p$, and a *nearly* arbitrary estimator $\hat{\mu} = y + g(y)$, with $g : \mathbb{R}^p \to \mathbb{R}^p$ being weakly differentiable, the risk of $\hat{\mu}$ can be estimated unbiasedly. It is given by

$$\mathbb{E}\|\hat{\mu} - \mu\|^2 = p + \mathbb{E}[\|g(y)\|^2 + 2\nabla g(y)] \qquad (11)$$

where $\nabla g = \sum_{i=1}^{L} \frac{\partial}{\partial_i} g_i$, cf. [7].

In our scenario we observe $\Phi^* y = y^* \sim \mathcal{N}(c^*, \sigma^2 I)$ with $c^* = \Phi^* f \in \mathbb{R}^p$. The soft threshold operator can be written component wise in this form by $\mathbb{S}_\lambda(z) = z + g_\lambda(z)$ with

$$g_\lambda(z) = \begin{cases} -\operatorname{sgn}(z)\lambda & \text{if} \quad |z| < \lambda \\ -z & \text{if} \quad |z| \leq \lambda \end{cases}$$

Using Stein's unbiased risk estimate (11), this implies that the risk of the Lasso is estimated by

$$\begin{aligned} & SURE(y^*, \mathbb{S}_\lambda^L(y^*)) \\ &= p - 2\#\{\gamma : |y_\gamma^*| \leq \lambda\} + \sum_{\gamma=1}^{p} \min(|y_\gamma^*|, \lambda)^2 \end{aligned}$$

and thus after rescaling the SURE threshold reads as

$$\lambda = \sigma \ \underset{\nu}{\operatorname{argmin}} \ SURE(y^*, \mathbb{S}_\nu^L(y^*)).$$

For the EW case, pursuing the same argument with

$$g_\lambda(z) = \begin{cases} -z & \text{if} \quad |z| < \lambda \\ -\lambda^2/z & \text{if} \quad |z| \leq \lambda \end{cases}$$

leads to

$$\begin{aligned} & SURE(y^*, \mathbb{S}_\lambda^{EW}(y^*)) \\ &= p - 2\#\{\gamma : |y_\gamma^*| \leq \lambda\} + \sum_{\gamma \leq |y_\gamma^*|} |y_\gamma^*|^2 \end{aligned}$$

where the optimal threshold is again chosen as the minimizer. In cases when the signal is *very* sparse, the SURE threshold is often replaced by the universal, see [7]. However, according to the experience gathered in our scenario while regarding natural audio signals, this does not seem happen too often.

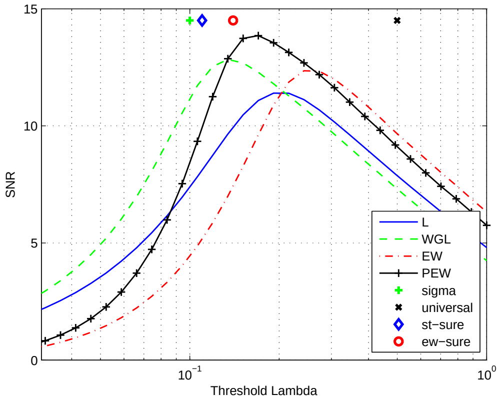

Figure 3. Denoising performance in SNR dB for different thresholds λ and x-coordinates of 4 adaptive threshold suggestions. The underlying signal is the glockenspiel with additive Gaussian white noise, $\sigma = 0.1$, leading to an basic noise level 0 SNR dB. The neighborhoods extend 5 coefficients in time.

3.2 Criterion-Based Threshold Prediction

Again considering our favorite glockenspiel signal at a noise level of $\sigma = 0.1$, Figure 3 presents the evolution of the denoising performance of the four studied operators over the threshold parameter λ while also displaying as x-coordinate the four discussed thresholds. While the SNR curves peak at different thresholds for the different operators, the proposed thresholds do not seem to coincide with any of these minimizers. As expected, the universal is very large, and the noise level σ too small. Even the SURE thresholds do not seem to maximize the SNR, based on the ℓ_2 distance of f and $\hat{f}$ in practice.

Notwithstanding, the relative positions of the SNR maximizers and the noise level σ are anything but arbitrary. Figure 4 shows a scatter plot[4] of the SNR-optimal thresholds λ of ten different speech signals at 8 kHz sampling rate, each at 10 different noise levels. Music signals at 44.1 kHz feature the same dependencies. The graphic clearly suggests a simple linear relation $\lambda = \beta\sigma$ for a scalar β which depends on the shrinkage operator. That means, if we are interested in maximizing the SNR criterion, we can simply learn the coefficients β via linear regression. These coefficients are displayed in Table 1 for the 8 and 44.1 kHz signals, resp.

The approach, as well as the thresholds discussed in the foregoing section, was evaluated on a set of speech signals[5] at 8 kHZ sampling frequency. The experi-

[4] The log-log scale, transforming linear relations to affine linear relations, is chosen here to better account for the logarithmic granularity of the noise levels σ and thresholds λ in the setup of the simulation.

[5] Again the results on speech were qualitatively very similar to

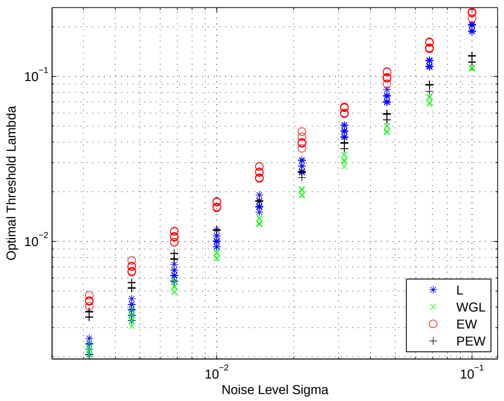

Figure 4. Optimal thresholds λ for the shrinkage operators L, WGL, EW, PEW of 10 speech signals at 10 different noise levels in log-log-scale.

β	L	WGL	EW	PEW
Speech	1.8180	1.0795	2.2955	1.2899
Music	2.5191	1.3675	2.9031	1.4979

Table 1. Regression coefficients β for different shrinkage operators computed on speech signals at 8 kHz sampling frequency and music signals at 44.1 kHZ.

ment was conducted by testing with 10 signals not included in the learning set at 10 different noise levels σ. For each condition, we collected the differences of the thresholds λ from above as well as the linear estimates $\lambda = \beta\sigma$ to the respective experimental optimum $\lambda_\star$ (i.e. the λ_i which maximizes SNR in the respective condition). Figure 5 presents the mean of these errors $|\lambda_\star - \lambda|$ over the 10×10 test conditions.

Obviously, the universal threshold is worst for all four operators. L and EW SURE are moderately far away from the optimal thresholds of the persistent operators, the same holds for σ, surprisingly. In particular, the linear model $\lambda = \beta\sigma$ tailored to the specific operators seems to work well. For each operator it yields the smallest mean distances (compared to all other tested thresholds) to the respective optimal thresholds.

4. NOISE LEVEL ESTIMATION

In most real life cases, the noise level σ is unknown and has to be estimated. This is a task of major relevance which has attracted much research attention, see e.g. [8, 18] and references therein. Here, it must suffice to briefly discuss a very simple but efficient heuristic which performs reasonably well for the white noise case focussed on in this article. It relies on the basic observation that the energy of most natural classes of audio signals concentrates in the low frequencies (assuming that sampling rates greater or equal than, say, 8 kHz are used). It hence uses a number of high frequency bins to estimate the overall noise level. Consequently, the greater the frequency range to be taken into account, the greater is the probability of such a white noise level estimator to be systematically biased by the clean signals' energy. On the other hand, including more frequency bins reduces the estimations variance and hence improves its average accuracy.

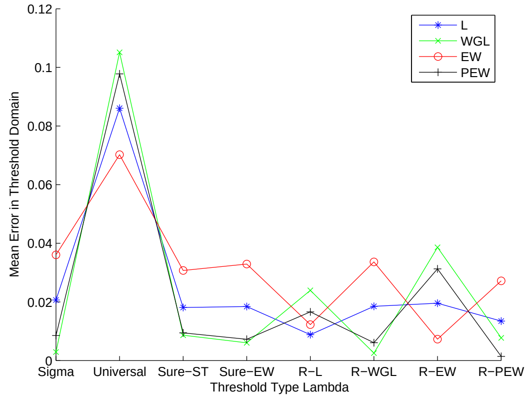

Figure 5. Threshold prediction results. The x-axis features the different proposed thresholds. The y-axis corresponds to the mean distance from the optimal threshold (w.r.t. to an operator) to the respective threshold on the x-axis.

Besides this fundamental bias-variance trade-off, we strive for robustness to outliers (corresponding to sparse but energetic coefficients of the true signal). Thus, instead of directly estimating the noise standard deviation via $\hat{\sigma} = \sqrt{\frac{1}{p-1}\sum_\gamma (y^*_\gamma - \bar{y^*})^2}$ (where $\bar{y^*}$ denotes the arithmetic mean which is not robust) a slight detour via the distribution of $|y^*|$ can be taken: As the component wise error $z_\gamma \sim \mathcal{N}(0, \sigma^2)$ follows the standard normal, its absolute value $|z_\gamma|$ is distributed according to the *half-normal* distribution with parameter θ. It has the median $M = \frac{\mathrm{erf}^{-1}(1/2)\sqrt{\pi}}{\theta} = \frac{0.8453}{\theta}$, cf. [19], while *erf* denotes the Gauss error function. The half normal is bound to the original normal distribution by $\theta = \sigma^{-1}\sqrt{\frac{\pi}{2}}$. Given an appropriate choice of a frequency range, we can hence use the empirical median $\hat{M}$ of the $|y^*_\gamma|$ to obtain a robust estimate of the noise standard deviation by setting [6] $\hat{\sigma} = \sqrt{\frac{\pi}{2}}\,\frac{\hat{M}}{0.8453}$.

Figure 6 compares the mean squared error of the regular empirical variance estimator and the median based one with two different selected frequency ranges each

the music signals. Hence we only present the former.

[6] The corresponding recommendation $\hat{\sigma} = \hat{M}/0.6745$ can be found in [8, p. 565] where it is without derivation applied in the context of wavelet thresholding.

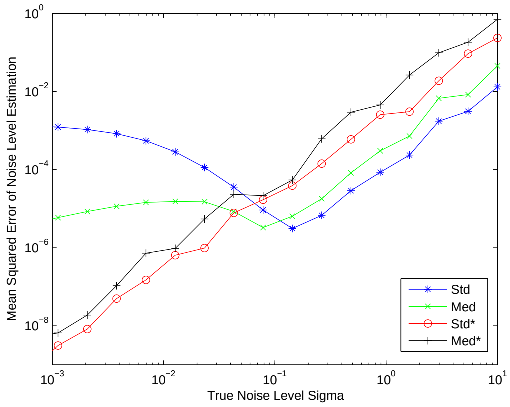

Figure 6. Mean squared error of different noise level estimators over true noise levels. The estimators are i) the empirical standard deviation of the upper half of the time-frequency coefficients ("Std"), ii) the median based estimator of the upper half of the coefficients ("Med"), iii) the empirical standard deviation on the highest frequency only ("Std*"), iv) the median based estimator on the highest frequency only ("Med*").

over 10 different signals for $0.001 \leq \sigma \leq 10$ with 10 experimental trials each. The first range reaches from the Nyquist frequency N to $N/2$. The second just contains the highest frequency bin (as commonly done in the wavelet denoising, see e.g. [7]). The graphic depicts the results for the case of 8kHz speech signals, they were surprisingly similar to the results obtained with 44.1 kHz signals, however. The median-based estimator with range $N - N/2$ here appears as a good compromise between the precision of the empirical standard deviation with the same range for large σ and the good performance of the estimators only using the highest bin. It is consequently employed in the following experiments.

5. COMPARISON OF ALGORITHMS

A final case study was conducted to compare and evaluate the developed approaches. For each operator using its respective coefficients β as displayed in Table1, we compare the SNR denoising performance with the respective optimal threshold of each operator. Both, the true and estimated noise level σ was used. Additionally to the 4 operators considered so far (here equipped with a neighborhood extending in time encompassing 9 coefficients) the results of Yu et al.'s block thresholding algorithm (BT) [10] were included which also requires the noise level σ as input variable. The experiments were conducted on a female and male speech signal at 8kHz with noise levels of 10 and 0dB, resp., as well as a classical and jazz music signal at 44.1 kHz also with 10 and 0dB resp. noise level, each of two seconds length. To avoid random variations in SNR dependent on the specific realization of the noise process, SNR results were averaged over 5 trials.

SNR	L	WGL	EW	PEW	BT
Optimal	14.6	15.1	15.0	15.9	16.7
True	13.6	14.8	14.5	15.9	16.6
Estimated	13.4	14.5	14.2	15.7	10.1

SNR	L	WGL	EW	PEW	BT
Optimal	7.4	8.2	7.8	9.0	9.5
True	7.3	8.2	7.7	9.0	9.4
Estimated	7.3	8.2	7.7	9.0	0.0

Table 2. SNR denoising results for different thresholds and operators for speech signals at 8 kHz. Top: male speech at 10 dB SNR. Bottom: female Speech at 0 dB SNR. Considered thresholds are the SNR-optimal one, the adapted $\beta\sigma$ as well as $\beta\hat{\sigma}$.

SNR	L	WGL	EW	PEW	BT
Optimal	16.0	17.5	16.8	18.8	19.0
True	13.8	16.5	15.5	18.5	19.0
Estimated	13.9	16.5	15.6	18.6	19.0

SNR	L	WGL	EW	PEW	BT
Optimal	9.1	10.9	9.7	11.9	11.9
True	8.3	10.5	9.3	11.8	11.8
Estimated	8.3	10.6	9.3	11.8	11.9

Table 3. SNR denoising results for different thresholds and operators for music signals at 44.1 kHz. Top: classical orchestral excerpt at 10 dB SNR. Bottom: Jazz quintet at 0 dB SNR. Considered thresholds are the SNR-optimal one, the adapted $\beta\sigma$ as well as $\beta\hat{\sigma}$.

Table 2 presents the results for the speech case, Table 3 for music. In both cases, the threshold adaption yields satisfactory results. The differences for true and the estimated σ seem to be negligible for the operators L, WGL, EW, PEW in most cases. However, the performance of BT seems to sensitively depend on the exact noise level, at least in the 8 kHz sampling rate case. PEW appears to be surprisingly robust to threshold deviations, as performance differences from the two adaptive cases to the optimal threshold does not exceed 0.3 dB in these examples. Moreover, the performance of the PEW comes very close to that of BT, even with adaptive, i.e. somewhat sub-optimal thresholds. In terms of perception, the persistent operators WGL and PEW even seem preferable, cf. [15]. Note that in the current implementations, PEW is around 6 times faster than BT. Overall, the PEW hence appears as an efficient alternative to block thresholding.

6. CONCLUSION

This contribution investigated a novel denoising operator, the *persistent empirical Wiener*, which features a flexible, neighborhood-based grouping structure of the coefficients. Its denoising properties with respect to the component wise estimation risk were compared to the regular empirical Wiener estimate and relations to operators as the Lasso and the block-James-Stein shrinkage were discussed. Moreover, we gave a simple but efficient heuristic for adapting the threshold λ to a given empirical maximization criterion and a set of audio signals. Initial tests showed that the coefficients β of the linear model $\lambda = \beta\sigma$, learned for maximizing denoising performance in SNR, seem to be quite robust and only yield minor differences compared to the respective empirical optima, even with an estimated noise level $\hat{\sigma}$.

Future work concerns the theoretical study of the proposed operator, generalizations to colored noise, as well as the evaluation of the proposed threshold selection principle with respect to perceptual criteria.

Acknowledgments

The author thanks Monika Dörfler and Matthieu Kowalski for fruitful discussions and advice.

This work was supported by the Austrian Science Fund (FWF) : T384-N13 and the WWTF project Audio-Miner (MA09-024).

7. REFERENCES

[1] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*. John Wiley and Sons, 1996.

[2] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal of Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Statistical Methodology),*, vol. 58, no. 1, pp. 267–288, 1996.

[4] T. T. Cai, "Adaptive wavelet estimation: A block thresholding and oracle inequality approach," *The Annals of Statistics*, vol. 27, no. 3, pp. 898–924, 1999.

[5] M. Kowalski, "Sparse regression using mixed norms," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 303 – 324, 2009a.

[6] M. Kowalski and B. Torresani, "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients," *Signal, Image and Video Processing*, vol. 3, no. 3, pp. 251–264, 2008a.

[7] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the american statistical association*, pp. 1200–1224, 1995.

[8] S. Mallat, *A Wavelet Tour of Signal Processing - The Sparse Way*, 3rd ed. Academic Press, 2009.

[9] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.

[10] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1830–1839, 2008.

[11] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[12] S. Ghael, A. Sayeed, and R. Baraniuk, "Improved wavelet denoising via empirical wiener filtering," in *Proceedings of SPIE*, vol. 3169. San Diego, CA, 1997, pp. 389–399.

[13] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, pp. 373–384, 1995.

[14] W. James and C. Stein, "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1961, pp. 361–379.

[15] K. Siedenburg and M. Dörfler, "Audio denoising by generalized time-frequency thresholding," in *Proceedings of the AES 45th Conference on Applications of Time-Frequency Processing, Helsinki, Finland, March1-4*, 2012.

[16] D. Donoho and J. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[17] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.

[18] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transaction on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[19] Wolfram Mathematica Online Documentation. (2012, April) Half normal distribution. [Online]. Available: http://reference.wolfram.com/mathematica/ref/