Wavelet-Based Denoising of Speech

Arkady Bron[1]**, Shalom Raz, David Malah**
Department of Electrical Engineering, Technion, Haifa, Israel

1. Introduction

Wavelet bases are widely used for estimating signals embedded in noise. The *wavelet shrinkage* method [3] transforms the noisy data into the wavelet-domain, applies soft or hard thresholding to the resulting coefficients, and subsequently transforms the modified wavelet-domain coefficients back into the original space. Saito [6] proposed to use an information-theoretic criterion, the *Minimum Description Length* (MDL), for noise removal. It has been observed [2, 3, 6] that denoising with the conventional *wavelet transform* (WT) and *wavelet packet decomposition* (WPD) may exhibit artifacts, such as pseudo-Gibbs phenomena in the neighborhood of discontinuities. These artifacts were related to the lack of *shift-invariance* of the applied transforms, and their reduction was achieved by averaging over several translations. This averaging procedure was termed in [3] *Cycle-Spinning*. Cohen, et al. [2] presented an extension of WPD into shift-invariant WPD (SIWPD). Moreover, they formulated the MDL criterion as an additive information cost function [1] and presented an adaptive translation-invariant denoising algorithm.

The main purpose of this work was to improve the performance of existing wavelet-based denoising algorithms when applied to speech signals, and to study the consequences of SIWPD on the resulting artifacts.

2. Denoising in the wavelet domain

Suppose we have noisy data $\mathbf{y} = \{y_i\}_{i=0}^{N-1}$ $(N = 2^J)$, where $y_i = f_i + e_i$, $i = 0, ..., N-1$, $\mathbf{f} = \{f_i\}_{i=0}^{N-1}$ is an unknown real-valued signal that we would like to recover, and $\mathbf{e} = \{e_i\}_{i=0}^{N-1}$ is a stationary additive white Gaussian noise (AWGN) with zero mean and a presumably known power spectral density σ^2. The vector $\mathbf{w} = \{w_{\ell,n,k}\}$ of wavelet expansion coefficients of the noisy data $\mathbf{y}$ is defined by $\mathbf{w} = W\mathbf{y}$, where ℓ is the resolution level index, n is the oscillation index, k is the time-domain position index and W is the finite orthonormal wavelet transform matrix. The orthonormality of W yields the following reconstruction formula: $\mathbf{y} = W^T\mathbf{w}$. Under the underlying noise model, noise contaminates all wavelet coefficients equally. Since $\mathbf{e}$ is assumed to represent WGN, its orthogonal transform, $\mathbf{z} = W\mathbf{e} = \{z_{\ell,n,k}\}$, is also WGN, $w_{\ell,n,k} = \theta_{\ell,n,k} + z_{\ell,n,k}$, where $\Theta = W\mathbf{f} = \{\theta_{\ell,n,k}\}$ is the vector representing the unknown wavelet transform coefficients of the noiseless data $\mathbf{f}$. Denoising in the wavelet domain is based on the principle of *selective wavelet reconstruction*: Given $\mathbf{w}$ we need to determine a subset Γ of indexes (ℓ, n, k) of wavelet coefficients in $\mathbf{w}$ that will be kept and possibly modified, usually by applying appropriate gain values $g_{\ell,n,k}$. Estimation of Θ is then obtained by $\widehat{\Theta} = \{g_{\ell,n,k} \cdot w_{\ell,n,k}\}_{(\ell,n,k)\in\Gamma}$, and the denoised signal is obtained by $\widehat{\mathbf{f}} = W^T\widehat{\Theta}$. [1]

3. Speech denoising algorithm development

Assuming that the speech and noise signals are independent, we conclude that $E||\mathbf{w}||_{2,N}^2 = E||\Theta||_{2,N}^2 + E||\mathbf{z}||_{2,N}^2 = ||\Theta||_{2,N}^2 + N\sigma^2$, where $||\mathbf{x}||_{2,N}^2 = \sum_{i=0}^{N-1} x_i^2$, and E denotes the expectation operator.

Among all linear estimators $\widehat{\Theta} = G(\mathbf{w}, \sigma) \cdot \mathbf{w}$, the *Wiener Gain Estimator* is optimal in the *mean squared error* sense:

$$G_w^*(\Theta, \sigma) = \frac{||\Theta||_{2,N}^2}{||\Theta||_{2,N}^2 + E||\mathbf{z}||_{2,N}^2}. \tag{1}$$

Since in practice $||\Theta||_{2,N}^2$ is unknown $G_w^*(\Theta, \sigma)$ is an *ideal gain*; it is replaced in (1) by its estimate $\eta(||\mathbf{w}||_{2,N}^2, E||\mathbf{z}||_{2,N}^2)$, yielding $G_w(\mathbf{w}, \sigma)$, where η is a general operator that is used for estimating $||\Theta||_{2,N}^2$. Soft thresholding [3] can be used to define η. Given $\mathbf{y}$, the variance σ^2 characterizing the AWGN must be estimated first. We based the selection of the various components of the denoising algorithm on simulation results. The algorithm performance was evaluated both subjectively, by listening, and objectively by the often used measures: SNR, *Segmental SNR* (SEGSNR) and *Log-Spectral Distance* (LSD) [1]. The noisy signals were created by adding WGN at 10 dB SNR.

In our simulations, we used WPD to decompose the noisy signals, and applied various thresholding methods, as well as the Wiener estimator, for speech denoising. *SureShrink* [3] and Wiener estimators were found to be clearly better than other examined alternatives [1]. However, it was found that the SureShrink estimator generates spike-like artifacts due to the thresholding operation it uses. These artifacts are common to both soft and hard thresholding processes. Moreover, speech enhanced by thresholding-based algorithms sounds muffled, because most high frequency speech coefficients are zeroed out [1]. In addition we observed that full subband WPD-based denoising leads to somewhat better results compared to adaptive WPD, constructed with an additive information cost function [4]. Full subband decomposition is attractive as it represents an efficient non-adaptive tree structure.

Additional tests have shown that the use of SIWPD typically does not improve the enhancement results. The improvement obtained by the Cycle Spinning denoising approach probably stems from signal averaging over different time-shifts which smoothes the signal. It suppresses the Gibbs-like artifacts but also blurs sharp signal transitions.

[1] Currently with Rafael Ltd., P.O.Box 2250(39), Haifa 31021, Israel

3.1. Mother-wavelet: phase linearity and design

An important aspect of a wavelet-based algorithm is the selection of the underlying mother wavelet. From our simulations we concluded that symmetry of the mother-wavelet is of secondary importance. We particularly examined the importance of frequency localization and have done so by defining a *generalized* Meyer mother-wavelet that is given by a *modified* Quadrature Mirror Filter (QMF) $m(\omega)$:

$$m(\omega) = \begin{cases} 1, & |\omega| \le \frac{\pi}{2}(1-r), \\ \cos\left[\frac{\pi}{2} \cdot \nu\left(\frac{|\omega| - \frac{\pi}{2}(1-r)}{\pi r}\right)\right], & \frac{\pi}{2}(1-r) \le |\omega| \le \frac{\pi}{2}(1+r), \\ 0, & |\omega| \ge \frac{\pi}{2}(1+r), \end{cases} \tag{2}$$

where $\nu(x)$ is an *auxiliary function* ($x \in [0,1]$), and r is the *roll-off* parameter [1]. The choice of $r = 1/3$ corresponds to the standard Meyer mother-wavelet [1]. A comparison of the enhancement results obtained with the Meyer mother-wavelet vs. other common mother-wavelets has shown a clear advantage of the first due to its improved frequency localization. This property corresponds to a wider time-support of the basis functions and thus helps to suppress time-localized artifacts. Moreover, when using either the Wiener gain estimator or *SureShrink* thresholding, using the above Meyer mother-wavelet with a small roll-off, the quality of the enhanced speech improves, resulting also in better SNR and LSD values. However, threshold methods, including *SureShrink* were not judged to provide an enhanced speech quality that is as good as the quality obtained by the Wiener gain estimator.

3.2. Decision-directed a priori SNR estimation

Obviously, when speech is corrupted by stationary colored noise, the noise variance has to be estimated for each decomposition band. Such an estimation is facilitated by the decision-directed (D-D) *a priori* SNR estimation [5] . In order to utilize it we have to segment the speech signal into frames. Segmentation of the speech signal results in better tracking of temporal transitions and also decreases the processing time-delay. The Wiener gain function (eqn. (1)), can be rewritten as follows:

$$G^{*}_{\ell,n}(\Theta_{\ell,n}(j), \mathbf{z}_{\ell,n}(j)) = \frac{\xi_{\ell,n}(j)}{\xi_{\ell,n}(j)+1}, \quad \xi_{\ell,n}(j) = \frac{||\Theta_{\ell,n}(j)||^2_{2,d}}{E||\mathbf{z}_{\ell,n}(j)||^2_{2,d}} \tag{3}$$

where j is the index of the analysis frame ($j = 1, 2, ...M$), $\Theta_{\ell,n}(j)$ and $\mathbf{z}_{\ell,n}(j)$ are the clean speech and the noise process wavelet coefficients in the band indexed by the pair (ℓ, n), $\xi_{\ell,n}(j)$ is the *a priori* SNR of the wavelet coefficients in that band, and d is the number of the coefficients in the band ($d = 2^{(\ell+J)}$). Again, by substituting in eqn. (3) an estimate for $\xi_{\ell,n}(j)$, which we denote by $\widehat{\xi}_{\ell,n}(j)$, we obtain the practical gain function $G_{\ell,n}(\mathbf{w}_{\ell,n}(j), \mathbf{z}_{\ell,n}(j))$. So, we can use now $g_{\ell,n,k} = G_{\ell,n}$ for all $k \in [0; d-1]$. As in the Ephraim-Malah Log-Spectral Amplitude (LSA) algorithm [5], the estimation of the *a priori* SNR is performed via the D-D approach [1]:

$$\widehat{\xi}_{\ell,n}(j) = \alpha \frac{||\widehat{\Theta}_{\ell,n}(j-1)||^2_{2,d}}{E||\mathbf{z}_{\ell,n}(j-1)||^2_{2,d}} + (1-\alpha)\max(\gamma_{\ell,n}(j) - 1, 0), \;\; j = 2, 3, ...M, \tag{4}$$

where $\gamma_{\ell,n}(j) = \frac{||\mathbf{w}_{\ell,n}(j)||^2_{2,d}}{E||\mathbf{z}_{\ell,n}(j)||^2_{2,d}}$ is the *a posteriori* SNR and α is a smoothing parameter. The initial condition for computing (4) is $\widehat{\xi}_{\ell,n}(1) = \alpha + (1-\alpha)\max(\gamma_{\ell,n}(1) - 1, 0)$.

The best results, in terms of SNR and the quality of the enhanced speech, were obtained by using a Hann window (frame size is 256 samples, with 50% overlap), full subband decomposition (FSD), $L = \log_2 N = J$ (the lowest allowed decomposition level, where N is the length of the analysis frame), a generalized Meyer mother-wavelet (64 taps, 10% roll-off), and $\alpha = 0.9$. Utilization of the D-D *a priori* SNR estimation generally requires tracking the *a priori* SNR at the terminal tree nodes. Thus for a fixed (non-adaptive) tree structure (such as FSD) the task is simplified since the indices of terminal nodes remain invariant from frame to frame. In contrast, tracking the *a priori* SNR when terminal node indices change from frame to frame, as would be the case with an adaptive WPD, is a more complex task.

Local Trigonometric Decompositions (LTD) are joint time-frequency representations that perform an adaptive time-axis segmentation. We have found that the conclusions stated above, with regard to WPD-based speech denoising, hold also for LTD-based denoising of speech (cf. [1]).

4. Ideal denoising

When using the Wiener estimator, the D-D *a priori* SNR estimation can be applied to any of the discussed decompositions, as well as to other spectral transforms like the DFT and DCT, to effectively smooth frame to frame gain fluctuations. These fluctuation are caused by fluctuations in the estimated *squared spectral amplitude* of the noise process. In all examined cases, the D-D approach improved the quality of enhanced speech and resulted in higher SNR values. Obviously, if we could use the exact values (unknown, in practice) of the squared spectral components ($|Z_k|^2$ for a DFT) of the noise process in each frame, instead of estimated expected values ($E|Z_k|^2$, in the case of a DFT), we would expect better performance of the denoising algorithms. We refer to a denoising process based on the exact values of the squared spectral amplitudes of the noise signal as "*ideal denoising*". Results of a comparison between WPD-based ideal denoising and a DFT-based one have shown that the WPD-based ideal denoising attains a consistently higher SNR [1] (see below). The

reasons could be:
1) While performing denoising with N samples per time frame, we obtain N complex-valued spectral (DFT) coefficients. WPD (or any real-valued orthogonal transform) yields N independent real-valued coefficients. Consequently, DFT-based denoising has at its disposal only half ($\sim N/2$) the number of amplitude coefficients.
2) While performing DFT-based denoising, if $|Y_k|^2 > |Z_k|^2$ we do not modify the phase of the noisy signal, taking it as the optimal estimate of the clean speech phase [5]. Whenever $|Y_k|^2 \leq |Z_k|^2$ we estimate that the speech component in the k-th bin is zero (amplitude and phase). On the other hand, when the *ideal denoising* is based on a real-valued transform, one can achieve an exact phase reconstruction of the clean speech signal (expressed via the sign of the real-valued transform coefficients) [1]. This is an intrinsic advantage of using real-valued transforms for denoising, as compared to DFT-based denoising.

A fair comparison is achieved by zero padding the signal frames in DFT-based ideal denoising (i.e., using DFT_{2N}), thus improving its frequency resolution, and subsequently slightly improving the global SNR of enhanced speech. Simulation results show that WPD-based ideal denoising outperforms DFT_{2N}-based ideal denoising by $0.69 \div 1.12$ [dB] in global SNR, in the particular simulations [1].

5. Comparative performance analysis

Results of simulations under practical conditions have indicated that the DFT-based Wiener estimator attained the best values of global SNR, segmental SNR and LSD [1]. The notable differences is the level and type of the residual background noise. All of the algorithms, LSA being the exception, introduce a colored background noise that was found to be disturbing to the listener. The DFT-based Wiener estimator is characterized by the lowest level of residual noise, and is better then the proposed wavelet-based algorithm. The LSA estimator is characterized by a higher level of background noise, than DFT and WPD-based Wiener estimators, but, advantageously, the background noise is almost white. It's important to note that the Wiener estimator, by definition, minimizes the mean squared error (MSE) in estimating an unknown signal f, while the LSA estimator minimizes the MSE in estimating the log-spectral amplitude of the unknown signal. Hence, it's expected that the DFT-based Wiener estimator achieves higher SNR than the LSA estimator. Thus, the comparisons done in this work show that despite the advantages of WPD-based algorithms under *ideal denoising* conditions, in practice the DFT-based denoising algorithms obtain better results. We suggest that the reasons for that are:
1) Using estimated expected values of the squared spectral amplitudes of the noise, we can't expect to exactly reconstruct the clean speech phase.
2) As shown in [1], if the additive noise is WGN, then, with the exception of the DC component, the variance of its squared spectral amplitudes, resulting from a real-valued orthogonal transform, is twice the variance of those obtained via the DFT. This leads to higher frame to frame gain fluctuations, thus reducing the resulting global and segmental SNR values. The gain fluctuations also result in colored residual background noise.

6. Conclusion

We have developed and examined speech denoising algorithms based on WPD and LTD. We have shown that artifacts introduced by wavelet-based denoising algorithms [2, 3, 6], when applied to speech enhancement, can be suppressed by sharply defining the spectral support of the mother-wavelet. The improved frequency localization of the basis functions correspondingly improves the speech denoising performance. It also has been shown that shift-invariance achieved by SIWPD does not contribute to artifacts suppression and does not guarantee an improved denoising performance.

Denoising based on presumed knowledge of the squared spectral amplitude of the noise is referred to as *ideal denoising*. We obtained in this case that WPD-based denoising performs better than DFT-based one. In practice, state of the art DFT-based speech denoising algorithms perform somewhat better than the proposed wavelet-based speech denoising algorithm, for which we offered some possible explanations.

In spite of this result, wavelet-based speech enhancement can be advantageous in certain circumstances. For example, WPD-based denoising can be easily incorporated into a WPD-based speech coding system. Also, LTD can be used as a time-segmentation tool. Thus, the LTD-based denoising algorithm can be conveniently incorporated in speech analysis systems that require adaptive time-segmentation. Moreover, the LTD-based speech denoising algorithm can be used in conjunction with the Shift-Invariant Adaptive Polarity Local Trigonometric Decomposition (SIAP-LTD) [1]. Its shift-invariance property is potentially important for recognition applications.

7. References

[1] A. Bron, "Wavelet-Based Denoising of Speech", M.Sc. Research Thesis, Technion - Israel Institute of Technology, Haifa, Israel, July 2000. http://www-sipl.technion.ac.il/Arkady_Bron_thesis.ps

[2] I. Cohen, S. Raz and D. Malah, "Orthonormal shift-invariant wavelet packet decomposition and representation", Signal Processing, Vol. 57, No. 3, Mar. 1997, pp. 251–270.

[3] R. R. Coifman, D. L. Donoho, "Translation-invariant de-noising", in: A. Antoniadis and G. Oppenheim, ed., *Wavelet and Statistics*, Lecture Notes in Statistics, Springer-Verlag, 1995, pp. 125–150.

[4] R. R. Coifman, M. V. Wickerhauser, "Entropy-based algorithms for best basis selection", IEEE Trans. Inform. Theory, Vol. 38, No. 2, Mar. 1992, pp. 713–718.

[5] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", IEEE Trans. on Acoust., Speech and Signal Processing, Vol. ASSP-33, No. 2, April 1985, pp. 443–445.

[6] N. Saito, "Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion", Proc. SPIE, Vol. 2242, 1994, pp. 224–235.