Wavelet denoising of speech using neural networks for threshold selection

C.A. Medina, A. Alcaim and J.A. Apolinário, Jnr.

A new wavelet-based denoising scheme for speech enhancement applications is proposed. The new approach uses a neural network to estimate the threshold to be applied to the detail coefficients after wavelet decomposition. The proposed method is compared to classical wavelet-based denoising techniques in the presence of additive white Gaussian noise. Results are presented in terms of SNR gain as well as error rate performance in a speaker verification system. Especially in terms of SNR improvement, the proposed scheme has outperformed other wavelet-based speech enhancement techniques.

Introduction: Speech enhancement techniques are important tools to reduce the degradations introduced by ambient noise and to improve the performance of voice communication systems. Several approaches have been proposed and the most recent techniques are those related to wavelet denoising. They are based on thresholding and shrinking wavelet coefficients of noisy signals. In wavelet denoising, a critical step is the computation of the threshold, which represents a rough estimate limit between the signal and the noise wavelet components. Several schemes have been reported in the literature, which include soft and hard thresholding, and fixed or level-dependent thresholds [1–3]. After a brief review of these schemes, we propose in this Letter a new method to estimate the threshold. This method is level-dependent and based on the use of a neural network structure.

Wavelet-based speech enhancement: The wavelet-based algorithm proposed by Donoho and Johnstone [1–3] attempts to recover a signal $s = \{s_i\}$ from noisy data $d_i = s_i + \sigma w_i$, $i = 1, \ldots, N$, where σ is the noise standard deviation and $\{w_i\}$ are independent and identically distributed Gaussian random variables with zero mean and unit variance.

The algorithm has the following three steps:

Step 1: Apply a J-level wavelet decomposition to the noisy signal.
Step 2: Apply a thresholding nonlinearity to the detail coefficients, in order to shrink the wavelet coefficients of the noisy signal. The threshold rule can be either soft or hard. Hard-threshold keeps unchanged only large observations, is sensitive to small variations in the data, and yields less smooth fits. In this Letter, we use the soft-threshold function defined by

$$\hat{\alpha}_{jk} = \eta_S(\beta_{jk}, t) = \begin{cases} \mathrm{sgn}(\beta_{jk})(|\beta_{jk}| - t), & |\beta_{jk}| \geq t \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where β_{jk} is the kth detail coefficient of the noisy signal at level j, $j = 1, \ldots, J$, $k = 1, \ldots, n_j$, $n_j = N/2^{J-j+1}$; and $\mathrm{sgn}(x)$ is $+1$ if x is positive and -1 otherwise. Note that α_{jk} is the kth detail coefficient of the clean signal at level j.
Step 3: Invert the wavelet transform to obtain the denoised signal $\hat{s} = \{\hat{s}_i\}$.

The thresholding function depends on a parameter t which can be level-dependent or level-independent. Classical techniques to compute the threshold are the VisuShrink and the SureShrink [1–3]. When used with soft-thresholding, the VisuShrink method usually provides 'noise free' reconstructions. Donoho and Johnstone [1] proposed a universal threshold estimate given by $\hat{t} = \hat{\sigma}\sqrt{(2\log N)}$, where N is the total number of coefficients and $\hat{\sigma}$ is a rough estimate of the noise level. This estimate is given by $\hat{\sigma} = m/0.6745$, where m is the median absolute deviation (MAD) of the detail coefficients at the highest resolution level ($j = J$). In this Letter, we use a level-dependent threshold, with m being replaced by m_j, the MAD of the detail coefficients at level j.

The SureShrink algorithm [1–3] uses the Stein's method of unbiased risk estimation (SURE). The kth detail coefficient of the wavelet transform of the noisy signal at level j is defined by $\beta_{jk} = \alpha_{jk} + \sigma_j \xi_{jk}$, where $k = 1, \ldots, n_j$, $\sigma_j = m_j/0.6745$ are estimates of the noise level and ξ_{jk} are independent random Gaussian variables with zero mean and unit variance.

The mean square risk of $\hat{\alpha}$ at level j is defined as

$$\mathcal{R}_j(\hat{s}, s) = \sum_{k=1}^{n_j} E[(\hat{\alpha}_{jk} - \alpha_{jk})^2] \tag{2}$$

The threshold t_j, at level j, must be chosen so as to minimise (2). In practice, we do not know α_{jk}. However, we can choose a t_j that minimises an unbiased risk estimator, $\hat{\mathcal{R}}_j(\hat{s}, s)$, which is a function of σ_j and β_{jk}, $k = 1, \ldots, n_j$. For the case of soft-thresholding, it can be shown that the minimisation of $\hat{\mathcal{R}}_j(\hat{s}, s)$ leads to

$$\hat{t}_j = \min_{t \geq 0} \sum_{k=1}^{n_j} (2\sigma_j^2 + t^2 - \beta_{jk}^2) I\{|\beta_{jk}| \geq t\} \tag{3}$$

where $I(x)$ is such that if x is a logical variable assuming the values of *true* or *false*, then $I(x) = 1$ if x is *true*, and $I(x) = 0$ otherwise.

Threshold selection based on neural networks: In this Section we propose a novel neural network (NN) based approach to obtain the level-dependent threshold. In our experiments we have considered a backpropagation NN with one hidden layer. All neurons are activated by a log-sigmoid function, with maximum value equal to 1. The network is designed so that its output is an estimate of the threshold $\hat{t}_j$ that minimises the mean square error between $\hat{t}_j$ and an ideal threshold T_j. The inputs are the median absolute deviation and the variance of the detail coefficients at each decomposition level, which means that one NN is used at each level.

In the training stage, the NN output is compared with an ideal threshold, i.e. the one that minimises the risk function expressed by (2). During the training procedure, the ideal threshold at level j is set equal to the magnitude of the lth detail coefficient $|\beta_{jl}|$ that minimises the risk function. This means that $T_j = |\beta_{jl}|$, where

$$l = \min_{1 \leq l \leq N} \sum_{k=1}^{n_j} [\eta_S(\beta_{jk}, |\beta_{jl}|) - \alpha_{jk}]^2 \tag{4}$$

Performance evaluation: In this Section we present the simulation results of the proposed scheme as well as a comparison with three well-known algorithms for speech enhancement: the classical power spectral subtraction (PSS) [4] and the wavelet denoising methods, namely, VisuShrink (VS) and SureShrink (SS) [1–3]. The original speech signal was sampled at a frequency of 8 kHz, and segmented into 32 ms frames using Hamming windows. A five-level ($J = 5$) wavelet decomposition was obtained with 'Daubechies 10' mother wavelet [5], which is shown to preserve perceptual information better than other Daubechies wavelets [6]. Performance evaluation was carried out with two objective measures:

1 SNR gain—G_{SNR}, the SNR gain, is defined as the mean of the differences (measured over each speech frame) between the SNR values, in dB, before and after the application of the enhancement technique. In our experiments, the neural networks were trained with seven minutes of speech (uttered by seven speakers) corrupted by additive Gaussian noise (SNR = −5 dB). The database for tests consisted of 45 minutes of speech obtained from nine different speakers (different from those used to train the neural networks), each one uttering five minutes of speech. To avoid time aliasing, a 75% overlapping of speech frames was used in these experiments.
2 Equal error rate (EER) of an automatic speaker verification (ASV) system—We have considered a text-independent ASV system that employs 15 mel-cepstral features, a Gaussian mixture model (GMM) with 32 mixtures, and a universal background model [7]. The utterances were taken from a database composed by 60 male speakers, ten of them forming the background. A 50% overlapping of speech frames was used. From each speaker we have taken two minutes for training. The training signal was corrupted by additive white Gaussian noise at SNR = 5 dB and preprocessed with the same speech enhancement method used to enhance the test signal. Tests were applied to 25 s segments of speech. The EER was measured over 23 400 false tests and 600 true tests.

A first experiment was carried out to determine the number of neurons to be used in the hidden layer of the proposed scheme. Varying this number from two to 32, we have found that approximately the same result is obtained in all cases. For this reason, two neurons were used in all simulations involving the proposed algorithm.

Table 1 shows the results of the SNR improvement of the proposed algorithm (NN) compared with other classical methods—power spectral subtraction (PSS), VisuShrink (VS), and SureShrink (SS). As shown in Table 1, the neural network outperforms other speech enhancement techniques in terms of SNR improvement. Although a formal subjective evaluation was not carried out, it was observed from informal listening tests that the proposed method does not present the uncomfortable musical noise in the enhanced speech that is perceived when the spectral subtraction technique is used. Moreover, the proposed scheme presents less residual background noise than the VS and SS approaches. Fig. 1 shows examples of the speech enhancement results of a signal corrupted by white Gaussian noise at SNR = 0 dB.

Table 1: SNR gain (dB) for signal corrupted with additive Gaussian noise

SNR	PSS	SS	VS	NN
−5	7.36	2.53	7.32	11.16
0	6.83	1.74	5.37	10.14
5	6.21	0.43	2.99	8.25
10	5.46	−1.42	0.24	5.63

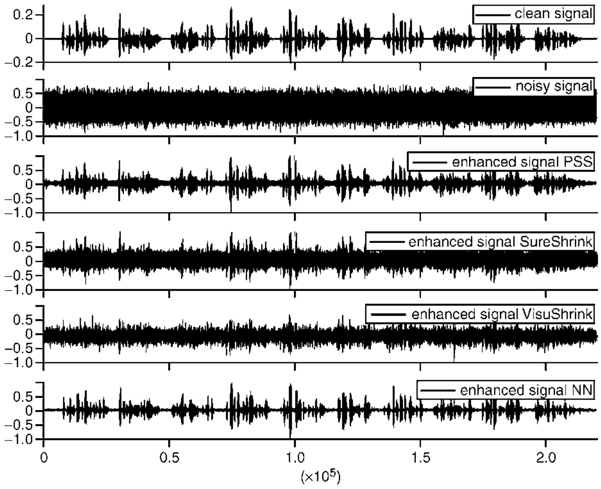

Fig. 1 *Enhanced signal corrupted with white Gaussian noise at SNR = 0 dB*

In Table 2, the equal error rate (EER) for the speaker verification system is presented. The speech signal is corrupted by additive white Gaussian noise at SNR values ranging from −5 to 10 dB. Table 2 shows that, for the ASV task, the wavelet-based methods perform better than the PSS scheme. In addition, the performance of the proposed algorithm is comparable to the VS wavelet denoising scheme and outperforms the SS wavelet denoising technique for very noisy environments.

Table 2: EER (%) in automatic speaker verification system for signal corrupted with additive Gaussian noise

SNR	PSS	SS	VS	NN
−5	36.27	34.61	33.78	33.61
0	24.46	16.31	16.47	16.31
5	15.97	6.32	6.32	6.32
10	13.81	8.15	7.82	7.99

Conclusions: We have proposed a wavelet denoising method for speech enhancement in which the threshold selection is based on neural networks. Our approach shows a small improvement over the classical VisuShrink (VS) and SureShrink (SS) thresholding schemes and does not contain the musical noise of the power spectral subtraction (PSS) technique. For a speaker verification task, the proposed method is superior to the PSS and comparable to the VS scheme. Moreover, for very noisy environments, it yields an error rate lower than the SS approach.

© IEE 2003 *3 September 2003*
Electronics Letters Online No: 20031200
DOI: 10.1049/el:20031200

C.A. Medina and J.A. Apolinário, Jnr. (*Department of Electrical Engineering, IME, Rio de Janeiro, RJ 22290-270, Brazil*)

A. Alcaim (*Center for Telecommunications Studies of the Catholic University-CETUC-PUC/Rio, Rio de Janeiro, RJ 22453-900, Brazil*)

E-mail: alcaim@cetuc.puc-rio.br

References

1 DONOHO, D.L., and JOHNSTONE, I.M.: ‘Threshold selection for wavelet shrinkage of noisy data’. Proc. 16th Annual Conf. of the IEEE Engineering in Medicine and biology society, Maryland, USA, 1994, pp. 24a–25a

2 DONOHO, D.L.: ‘De-noising by soft-thresholding’, *IEEE Trans. Inf. Theory*, 1995, **41**, (3), pp. 613–627

3 DONOHO, D.L., and JOHNSTONE, I.M.: ‘Adapting to unknown smoothness via wavelet shrinkage’, *J. Am. Stat. Assoc.*, 1995, **90**, (432), pp. 1200–1224

4 BOLL, S.F.: ‘Suppression of acoustic noise in speech using spectral subtraction’, *IEEE Trans. Acoustics Speech Signal Process.*, 1979, **27**, pp. 113–120

5 DAUBECHIES, I.: ‘Orthonormal bases of compactly supported wavelets’, *Commun. Pure Appl. Math.*, 1988, **41**, pp. 909–996

6 AGBINYA, J.I.: ‘Discrete wavelet transform techniques in speech processing’. Proc. IEEE TENCON—Digital Signal Processing Applications, Perth, WA, Australia, 1996, pp. 514–519

7 REYNOLDS, D.A., QUARTIERI, T.F., and DUNN, R.B.: ‘Speaker verification using adapted Gaussian mixture models’, *Digit. Signal Process.*, 2000, **10**, pp. 19–41