A Wavelet Based Noise Reduction Algorithm for Speech Signal Corrupted by Coloured Noise

Vladimir Braquet, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
braquet.vladimir@yahoo.fr, takao.kobayashi@ip.titech.ac.jp

Abstract

In this paper, we present a node dependent wavelet thresholding approach in order to remove strongly coloured noises from speech signals. The noise power in each node is first estimated using a recursive method. Given the voiced or unvoiced nature of the frame, the signal is expanded onto a predefined best basis. Then a infinitely smooth soft threshold is applied depending on each node of the decomposition tree. Finally the estimated clean signal is reconstructed. Experimental results on a Japanese database, for various coloured noises, demonstrate the effectiveness of the proposed method, even at low SNR. Compared with the common level dependent method, this algorithm provides better denoising results.

Keywords: Coloured noise, node dependent wavelet denoising, noise estimation, best bases

1. Introduction

Removing noises from corrupted speech is an important task. It is often used as a pre-processing stage in speech communication systems. It has been shown [1] that a wavelet transform followed by level dependent soft thresholding of noisy coefficients leads to a nearly optimal non-linear estimator of clean signal. Some examples can be seen in [2, 3].

In [1], the common wavelet tree is utilized as the basis tree. For white noise, it may be more efficient to use speech adapted bases, like best bases. However for strongly coloured noises, it leads to poor denoising results because the behaviour of the noise in different nodes of the same level may differ completely. Our goal is therefore to reach high denoising performances for coloured noise using bases more efficient than the common wavelet tree.

A new algorithm, which will be referred to as a node dependent algorithm, is introduced in this paper. First, predefined best bases on which each frame of the signal is expanded are introduced. Then, a noise estimation in each node of the basis is explained. Finally, a node dependent thresholding method using an optimal infinitely smooth ramp and the Stein’s Unbiased Risk Estimation (SURE) [1, 4] is proposed.

We will present experimental results for denoising that show effectiveness of the proposed algorithm when applied to a Japanese database degraded by various noises.

2. Predefined Best Bases

The wavelet packet transform [5], given an N-point frame \( \{ x[i] \}_{i \in [1..N]} \), is defined using a wavelet packet base \( B = \{ \{ e_{d,r,j}[k] \}_{k \in [1..N]} \}_{d,r,j} \), which can be represented as a tree, where

- \( d \) is the level in the tree, \( 0 \leq d \leq d_{\text{max}} \)
- \( r \) is the rank in the level \( d, 0 \leq r \leq 2^d - 1 \)
- \( j \) is the temporal index in the node \( [d,r], 1 \leq j \leq N2^{-d} \)

The transformed coefficients are given by

\[
X[d,r,j] = < x[e_{d,r,j}], \quad x = \sum_{i=1}^{N} x[i] e_{d,r,j}[i].
\]  

Once the quadrature mirror filters (8-point Daubechies filters in this work) used for the definition of the bases have been fixed, the problem is the choice of the tree of decomposition. It can be chosen by minimizing some costs [6]. The most common basis is the wavelet tree. Another possibility is to compute for each frame a best basis which concentrates the signal in a restrained number of non-zero coefficients.

The former is not specifically adapted to the speech, but, since it is independent from the noisy frame, it does not concentrate the noise at low SNR. In contrast, the latter is the best adaptation for each frame, however at low SNR the tree will also depend on the noise.

Here, we use a new kind of tree, independent of the noise but depending on the speech. In order to define these trees, we use the following cost. Given a predetermined rate \( 0 < C < 1 \), the cost of a basis \( B \) with respect to a signal \( x \) is defined from the sorted coefficients \( X[\phi(v)] \) (\( \phi \) such as \( \forall v, 1 \leq v < N, |X[\phi(v)]| \geq |X[\phi(v+1)]| \)) by

\[
C_B(B) = \min_k \left( \sum_{v=1}^{k} |X[\phi(v)]|^2 \right) \geq C \sum_{i=1}^{N} |x[i]|^2. \]  

This means that the cost will be the minimal needed number of transformed coefficients to reach a given rate \( C \) of the whole signal energy. By using fast algorithms [6], it is possible to find a so called nearly best basis. This cost was chosen because it leads to bases which concentrate the signal well and because several bases can be defined by changing the rate \( C \).

The next step is to find a cost over a pool of \( M \) samples \( P = \{ \{ x_j[i] \}_{i \in [1..N]} \}_{j \in [1..M]} \). Using \( C_B \), we simply define

\[
C_P(B) = \sum_{j=1}^{M} C_B(B). \]  

The best basis is therefore computed by minimizing \( C_P(B) \) over all the bases. The number of possible trees \( N_{\text{tree}}(d_{\text{max}}) \) increases exponentially with \( d_{\text{max}} \), indeed \( N_{\text{tree}}(d_{\text{max}}) = 1 + (N_{\text{tree}}(d_{\text{max}} - 1))^2 \). Therefore \( d_{\text{max}} \) has to be limited, and we will use here \( d_{\text{max}} = 4 \).

Furthermore, in order to obtain more efficient bases, we define different bases over voiced and unvoiced parts of speech, for male and female speakers, respectively.
3. Noise Estimation

It is necessary to adequately estimate the noise $n$ from the corrupted speech $x$ to attain higher denoising performances. A common method, using the median of the noisy coefficients in the first node of the wavelet tree, is not adapted for coloured noises. Other methods have been developed in [7, 8, 9]. These studies present algorithms that make use of the short time Fourier transform. In this paper, we apply them to the wavelet packet framework.

There are two approaches in these algorithms. The first one, quantile-based [7], assumes that in each frequency band, the voice is rarely present, thus the 50% quantile of a sufficiently long buffer of the noisy speech power gives a reliable estimation of the noise power in this band. This is no more true in the wavelet packet framework because the bases are designed in order to keep the signal concentrated, hence in few nodes the voice is always present.

The second approach is to use a recursive estimation [8, 9]. In each frequency band $w$, at frame $k$, the noise estimation $\hat{\sigma}_n^2(w,k)$ is the linear combination of the previous estimation $\hat{\sigma}_n^2(w,k-1)$ and the current noisy signal power in the band $\sigma_n^2(w,k)$. Once adapted to the node dependent wavelet packet framework, this leads to

$$
\hat{\sigma}_n^2([d,r],k) = \alpha[d,r]\sigma_n^2([d,r],k-1) + (1-\alpha[d,r])\sigma_n^2([d,r],k)
$$

(4)

where $\sigma_n^2([d,r],k) = \frac{1}{N^2} \sum_{i=1}^{N^2} |x_i([d,r],j)|^2$.

The smoothing parameter $0 \leq \alpha[d,r] \leq 1$ controls the rate of the noise estimation: if $\alpha[d,r] \approx 1$ then $\hat{\sigma}_n^2([d,r],k) \approx \sigma_n^2([d,r],k)$, and thus the noise estimation does not change. On the contrary, if $\alpha[d,r] \approx 0$ then $\hat{\sigma}_n^2([d,r],k) \approx \sigma_n^2([d,r],k)$ and $\alpha[d,r]$ is the noise estimation followed by the noisy speech power. Therefore the goal is to have $\alpha[d,r] \approx 1$ for presence of speech and $\alpha[d,r] \approx 0$ for absence of speech.

In order to fulfill this requirement, we follow [9], but the threshold which discriminates between presence of speech and absence of speech is now node dependent $T_{d,r}$, because it improves significantly the performances. Finally we have

$$
\alpha[d,r] = \frac{1}{1 + \exp(-\frac{\hat{\sigma}_n^2([d,r],k)}{\sigma_n^2([d,r],k) - T_{d,r}})}
$$

(5)

where $\hat{\sigma}_n^2([d,r],k)$ is the mean of the noise estimation in a given previous frame of number $a$ and $a > 0$.

Therefore, in the frame $k$, for presence of speech $\frac{\hat{\sigma}_n^2([d,r],k)}{\sigma_n^2([d,r],k)} > T_{d,r}$ and $\alpha[d,r] \approx 1$, and for absence of speech $\frac{\hat{\sigma}_n^2([d,r],k)}{\sigma_n^2([d,r],k)} \approx 1 < T_{d,r}$, so $\alpha[d,r] \approx 0$. Finally the noise estimation is filtered by a first order filter to smooth the estimation. The values $T_{d,r}$ have been selected in order to minimize the error at low SNR while maintaining reasonable error at high SNR.

4. Node Dependent Thresholding

In this section, a method for denoising the corrupted speech in each frame is presented. Given the gender of the speaker and the voiced or unvoiced nature of the frame, the basis tree $B = \{e_{a,r,j}\}_{a,r,j}$ is selected from the predefined best bases. For simplicity we will omit the frame index $k$, for instance $\sigma_n^2([d,r],k) = \sigma_n^2([d,r])$.

4.1. Node dependent algorithm

Here, we present an estimator for denoising the noisy coefficients and our assumptions. Let us assume that

**H1**: The noise $n$ is additive, so the corrupted signal $x$ is given by the clean signal $s$ as $x[i] = s[i] + n[i]$ and its transform becomes $X[d,r,j] = S[d,r,j] + N[d,r,j]$.

**H2**: The noise is centred and stationary. Thus $E[N[d,r,j]] = 0$ and $E[N[d,r,j]^2]$ is independent from the time index $j$. This means $\forall j, E[N[d,r,j]^2] = \sigma_n^2[d,r]$.

A general thresholding estimator of the clean signal is a function $f_{TH}$ such

$$
\hat{S}[d,r,j] = f_{TH}(X[d,r,j], t(\{X[d,r,j]_{d,r,j}\}d,r,j))
$$

(6)

where $t$ is the threshold. An example of such $f_{TH}$ is the usual soft thresholding, $t$ can be given by SureShrink [4].

In general case, it is assumed that each scale $2^d$ of phenomena is independent from the others, so we are led to

**H3**: The levels are independent.

$$
\hat{S}[d,r,j] = f_{TH}(X[d,r,j], t(\{X[d,r,j]_{d,r,j}\}d,r,j), P_a(n)))
$$

(7)

where the threshold depends only on the level $d$ and $P_a(n) = \sum_{j} <n|e_{d,r,j} > e_{d,r,j}$ is the projection of the noise $n$ onto the level $d$.

For wavelet transform [1], in each level except the last one, there is only one node $[d,1]$, so $P_a(n) = \sum_{j} <n|e_{d,1,j} > e_{d,1,j}$ and $\forall d, 0 < d < d_{max} \Rightarrow \sigma_n^2[d,r] = \sigma_n^2[d,1] = \sigma_n^2[d]$. Since the two nodes of the last level $[d_{max},0]$ and $[d_{max},1]$ are centred on close frequencies, no significant error will be created by the assumption

$$
\forall d, \sigma_n^2[d,r] = \sigma_n^2[d].
$$

(8)

The assumptions H3 and (8) mean that we use a global variance in each level. This does not hold if the basis tree is changed: if the noise is strongly coloured and if in a level the nodes are not close, the behaviour of the noise in these nodes may differ considerably. Since the predefined best trees may be different from the wavelet one, another assumption is made.

**H3'**: The nodes are independent.

$$
\hat{S}[d,r,j] = f_{TH}(X[d,r,j], t(\{X[d,r,j]_{d,r,j}\}d,r,j), P_a(n)))
$$

(9)

where the threshold depends only on the node $[d,r]$ and $P_a(n) = \sum_{j} <n|e_{d,r,j} > e_{d,r,j}$ is the projection of the noise $n$ onto the node $[d,r]$.

Finally let us assume that

**H4**: Within each node, the noise $N[d,r,j]$ is a multivariate normal distribution $N(\mu_{d,r,j}, \Lambda_{d,r,j})$.

The assumption H2 means that $\forall j, \mu_{d,r,j}(j) = 0$ and $\Lambda_{d,r,j}(j,j) = \sigma_n^2[d,r]$. Since the noise within the node is a centred multivariate normal distribution, it is completely characterized by $\Lambda_{d,r,j}$.

Given H1, H2, H3' and H4, our estimator is therefore

$$
\hat{S}[d,r,j] = f_{TH}(X[d,r,j], t(\{X[d,r,j]_{d,r,j}\}d,r,j), \Lambda_{d,r,j}))
$$

(10)

where the coefficients $\Lambda_{d,r,j}(j,j) = \sigma_n^2[d,r]$ are estimated by the noise estimation algorithm described in 4.

4.2. Thresholding method

The remaining problems are how to estimate the threshold $t$ in each node and how to find $f_{TH}$ which attenuates the coefficients. The former is described in 4.2.2, the latter in 4.2.1.
4.2.1. Shape of the thresholding

From preliminary experimental results, it appeared that an infinitely smooth soft thresholding yields better results. In order to define this soft thresholding at $t > 0$, given that a ramp is the second primitive of the Dirac delta function $\delta(x)$, $r_t(x) = \int_{-\infty}^{x} \int_{-\infty}^{0} \delta(t) dv du$, we use a “smooth” delta function. Let the test function $\Phi_{t, \epsilon, p_i, K_i}(x)$ whose support is the open interval $D = (t(1 + h), t(1 + h))$ with $h < t$. In $D$,

$$
\Phi_{t, \epsilon, p_i, K_i}(x) = \frac{(x - t - h\epsilon_i)^2}{(t(1 + h) - x)(x - (t(1 - h)))} e^{-t, p_i, K_i} \int_{-\infty}^{1} (\exp(-\frac{w}{w-1}))^n dw,
$$

where $|\epsilon_i| < 1$ controls the location of the maximum because $\Phi_{t, \epsilon, p_i, K_i}(t + h\epsilon_i) = \max_{\epsilon_i} \Phi_{t, \epsilon, p_i, K_i}(x)$. $p_i > 0$ controls how this function has a narrow tap, $K_i > 0$ is the $L_1$-norm. Examples of such function are shown in Fig. 1.

Let $F_{d, r} = \{\epsilon_i, p_i, K_i\}$, $F_{d, r}^m = \{-\epsilon_i, p_i, K_i\}$. Then $d_{t, F}(x) = \sum_{i}(\Phi_{t, \epsilon, p_i, K_i}(x) + \Phi_{t, \epsilon, p_i, K_i}(x))$ with $2\sum_{i} K_i = 1$ is then our “smooth” delta function and the function $r_{t, F}(x) = \int_{-\infty}^{1} d_{t, F}(v) dv du$ follows these properties which seems natural for a smooth ramp.

- (P1) $r_{t, F} \in C^\infty (R)$, $r_{t, F} > 0$, $r'_{t, F} > 0$
- (P2) $r_{t, F}(x) = 0$, $x \leq t(1 - h)$
- (P3) $r_{t, F}(x, t) = x - t$, $x \geq t(1 + h)$.

Thus $r_{t, F}(x)$ is an infinitely smooth soft ramp, $h \approx 0$ leading to the usual soft thresholding at $t$.

In this work, two possibilities are taken in account: $r_{t, (0, p, 0.5)}$ and $r_{t, (0.98, p, 0.5)}$. With different values of $p$, we have $0 \leq r_{t, F}(t) \leq \frac{h}{2}$ which are the limits of the functions following P1, P2 and P3. Examples of these two types of function are shown in Fig. 2. Finally $f_{rH}(x, t) = sign(x)r_{t, F}(x)$.

4.2.2. Threshold estimation

The final task is to find a reliable value of the threshold.

A commonly used approach is the universal thresholding. Since the noise $N[d, r, j]$ is a multivariate normal distribution $\mathcal{N}(\mu_{d, r}, \Lambda_{d, r})$ with $\forall j, \mu_{d, r}(j) = 0$ and $\Lambda_{d, r}(j,j) = \sigma_n^2[d, r]$

$$
\lim_{N \to \infty} P(\max_{1 \leq j \leq N^{2-d}} \frac{|N[d, r, j]|}{\sqrt{2\log(N^{2-d})}} > 1) = 0
$$

all the noise will be removed, with a high probability, using a threshold $t_{univ} = \sigma_n[d, r]\sqrt{2\log(N^{2-d})}$. A drawback of this approach is that this threshold is independent from the clean signal. If the signal is not sufficiently concentrated in a limited number of high coefficients, poor denoising results will be obtained. For example, the unvoiced part of the speech which usually leads to small coefficients.

Another approach is the SURE method [1, 4]: let $g_{T}(x) = f_{rH}(x, T) - x$ then $S[d, r, j] = X[d, r, j] + g_{T}(X[d, r, j])$. Since $\Lambda_{d, r}(j, j) = \sigma_n^2[d, r]$, the Stein’s unbiased risk estimation of $R_{d, r}(T) = E[\sum_{j=1}^{N^{2-d}} |S[d, r, j]|^2]$ is given by

$$
R_{d, r}(T) = N^{2-d}\sigma_n^2[d, r] + \sum_{j=1}^{N^{2-d}} |g_{T}(X[d, r, j])|^2 + 2\sigma_n^2[d, r]\sum_{j=1}^{N^{2-d}} \frac{\partial g_{T}}{\partial X}(X[d, r, j]).
$$

$t_{SURE}$ is given by $t_{SURE} = \min_{T \leq T_{univ}} R_{d, r}(T)$. It is easy to compute the differential of $g_{T}$ due to the definition of the smooth ramps.

In this paper a hybrid scheme [4] is adopted, where the universal threshold $t_{univ}$ is used for extreme sparsity of the coefficients. In the other case, $t_{SURE}$ is applied. Adapted to the node dependent framework it leads to the test $H[d, r, j]$: $\sigma_n^2[d, r]\sum_{j=1}^{N^{2-d}} |X[d, r, j]|^2 > 1 + \sqrt{N^{2-d}\log_{2}(N^{2-d})^{1/2}}$.

- If $H[d, r]$ is true, we use $t_{SURE}$
- if $H[d, r]$ is false, we use $t_{univ}$.
5. Results

To evaluate the proposed algorithm, we carried out denoising experiments with various noises and SNR conditions. The speech database included 10 male and 10 female Japanese speaker utterances, three sentences for each speaker, taken from the CREST database. Each sentence was divided into 32 ms frames, the sampling rate was 16 kHz. Various noises having different behaviours were selected: concentrated in low frequencies for pink noise, F16 cockpit noise, night ambient record, car noise, babble noise; concentrated in high frequencies for violet noise; spread on all frequencies for white noise.

For the definition of the predefined best bases, it appears that the best choice for the cost was to choose $C = 95\%$. Thus we have a $95\%$ predefined best basis for female voiced sounds, female unvoiced sounds, male voiced sounds and male unvoiced sounds. We used an IFAS-based algorithm for voiced and unvoiced classification of speech.

For the noise estimation algorithm, tests on the database led to $a = 10$ and $\sigma_n$ defined over 10 frames. The results are illustrated in Fig. 3 which shows the average of the relative error $\frac{|\hat{\sigma}^2((d,r),k) - \sigma^2((d,r),k)|}{\sigma^2((d,r),k)}$ above all the nodes, all the noises, all the frames and all the speech signals.

For the smooth ramp, testing the database led us to the following suitable choices ($h = 0.1$ for both case)

- voiced case: $r_{1.0,0.1,0.5}(x)$
- unvoiced case: $r_{2.0,0.98,10,0.5}(x)$

Figure 4 shows the results of the whole denoising algorithm (solid line). For comparison, the results of the usual level dependent wavelet algorithm [1] are also shown (dashed line). We used the same noise estimation algorithm for each case (adapting the program for node / level dependence assumption) and the same voiced / unvoiced algorithm. It shows how the SNR is increased by the denoising algorithms with respect to the input SNR for the males (+ points) and the females (o points). It clearly underlines that the level dependence and the wavelet tree are less efficient than the proposed estimator.

The main contribution of result improving appeared to be the choice between H3 and H3’ following by the shape of the ramp and finally the expansion basis. The computational cost of the algorithm can be decreased if we use the usual ramp (since the risk optimization complexity becomes $O(\mathcal{N}^2 \cdot \log_2(\mathcal{N}^2 \cdot d^d))$) however the results will not be as good as the smooth ramp ones.

6. Conclusions

Since the real environment noises are often coloured, a level dependent algorithm is not suited to remove them properly, therefore we have proposed a node dependent algorithm whose expansion bases are built such as to extract speech characteristics and whose thresholding method uses an infinitely smooth ramp function. From experimental results, it has been shown that the proposed algorithms improves the denoising results.

7. References