Wavelet Speech Enhancement Based on Time–Scale Adaptation

Mohammed Bahoura [a] and Jean Rouat [b,*]

[a] *Département de mathématiques, d'informatique et de génie Université du Québec à Rimouski, 300 allée des Ursulines, Rimouski, Québec, Canada, G5L 3A1.*

[b] *Département de génie électrique et génie informatique Université de Sherbrooke, 2500 boulevard de l'Université, Sherbrooke, Québec, Canada, J1K 2R1.*

Abstract

We propose a new speech enhancement method based on time and scale adaptation of wavelet thresholds. The time dependency is introduced by approximating the Teager Energy of the wavelet coefficients, while the scale dependency is introduced by extending the principle of level dependent threshold to Wavelet Packet Thresholding.

This technique does not require an explicit estimation of the noise level or of the *apriori* knowledge of the SNR, as is usually needed in most of the popular enhancement methods. Performance of the proposed method is evaluated on speech recorded in real conditions (plane, sawmill, tank, subway, babble, car, exhibition hall, restaurant, street, airport, and train station) and artificially added noise. MEL-scale decomposition based on wavelet packets is also compared to the common wavelet packet scale.

Comparison in terms of Signal-to-Noise Ratio (SNR) is reported for time adaptation and time-scale adaptation thresholding of the wavelet coefficients thresholding. Visual inspection of spectrograms and listening experiments are also used to support the results. Hidden Markov Models Speech recognition experiments are conducted on the AURORA–2 database and show that the proposed method improves the speech recognition rates for low SNRs.

Key words: speech enhancement, wavelet transform, Teager energy operator, speech recognition.

* Corresponding author. Tel: +1-819-821-8000
Email address: Jean.Rouat@usherbrooke.ca (Jean Rouat).

1 Introduction

1.1 Context

New speech-based applications such as automatic speech translation, internet search tools, multimedia teaching and training, and multimodal computer interactions are under development in many public and private research laboratories. A strong limitation to these systems is the inadequacy of processing corrupted or noisy speech.

Many approaches have been studied to increase the robustness of speech processing systems. In the context of speech or speaker recognizers based on a pattern recognition process that separates analysis from recognition, speech enhancement already has a strong potential. Speech enhancement can also be used in coding and with various apparatus such as audio protheses.

When the noise or the Signal-to-Noise Ratio (SNR) is known, contemporary techniques can yield quasi-optimal solutions to the problem of denoising. For example, the algorithm developed by Ephraim and Malah (Ephraim and Malah, 1984, 1985) is one of the most effective. A drawback of these enhancement techniques is the necessity to estimate the noise or the SNR. This can be a strong limitation when recording with non-stationary noise and for situations where the noise can not be estimated (no silence, no speech boundaries). To improve speech enhancement in non-stationary noise, Malah et al. (Malah et al., 1999) propose to control the gain of the update of the estimated noise spectrum during speech presence in a modified Minimum Mean-Square Error Log-Spectral Amplitude (MMSE-LSA) estimator. It is also pertinent to study new techniques that do not require any a priori knowledge of the noise and that can complement the contemporary denoising systems by taking into account speech characteristics. The most effective speech enhancement system would probably combine both approaches: 1) cleaning of the noise when it can be estimated and 2) enhancement by taking into consideration the speech structure when it is not possible to know anything regarding the noise.

In the present paper we propose a new procedure based on a time-scale threshold of wavelet packet coefficients without requirement or knowledge of the noise level. The thresholds are modulated with a nonlinear mask that reflects the spatial and time dominance evolution of speech on noise.

Donoho and Johnstone (Donoho, 1993; Donoho and Johnstone, 1994; Donoho, 1995) proposed a universal wavelet threshold to remove additional white noise. Another approach is also proposed by Johnstone and Silverman (Johnstone and Silverman, 1997) to remove correlated noise, while Vidakovic and Lozoya (Vidakovic and Lozoya, 1998) suggested the time dependence of the

threshold in the context of white additive noise and artificial signals. To our knowledge, all these methods do not succeed in speech enhancement because the thresholding process also removes some speech components.

To prevent the speech quality deterioration during the thresholding process, we propose to adapt the discriminative threshold in the time over scales.

We propose 1) to perform a time adaptation of the thresholds based on the modulation defined with a nonlinear mask that is based on the Teager Energy Operator (Bahoura and Rouat, 2001a) – that we call TA – and 2) to extend the level-dependent threshold (Bahoura and Rouat, 2001b) to the wavelet packet decomposition – that we call TSA or TSA2 depending on the version being used.

1.2 Scope of the paper

During the last decade, wavelet transforms (WT) have been applied to various research areas. Their applications include signal and image denoising, compression, detection, and pattern recognition.

Wavelet shrinkage is a simple denoising technique based on a thresholding of the wavelet coefficients. The estimated threshold is supposed to define the limit between the wavelet coefficients of the noise and those of the target signal. Unfortunately it is not always possible to separate the components corresponding to the target signal from those of noise by a simple thresholding. For noisy speech, energies of unvoiced segments are comparable to those of noise. Applying thresholding uniformly to all wavelet coefficients not only suppresses additional noise but also some speech components like unvoiced ones. Consequently, the perceptive quality of the filtered speech is greatly affected.

Unlike the conventional denoising methods based on the wavelet thresholding, the discriminative threshold in various subbands is time- and spatially-adapted in relation with the speech components when speech dominates the noise. The proposed techniques are tested on noisy speech recorded in real environments and with artificial noise. Three evaluations have been made. The first evaluation is based on a comparison of Signal-to-Noise Ratio increases with the Ephraim and Malah Filter (EMF). It is observed that the proposed Time and Scale Adaptation (TSA) of the wavelet coefficients yields a greater increase of SNR for very noisy speech (-10 dB to 10 dB). The second evaluation is based on listening tests: One with very noisy speech recorded in real environments (like planes, sawmills and tanks) and the other one with the AURORA–2 (Hirsch and Pearce, 2000) database where noise is artificially added to the TI digits. In real environments, the EMF is generally preferred for wide-band white sta-

tionary noise while the TSA is better on more general and realistic noises. On the AURORA–2 database, EMF is preferred to the Times Scale Adaptation that uses a continuous derivative thresholding function (TSA2). The third evaluation uses the HTK Hidden Markov Models kit with TSA2 that is an extension of the original TSA by using a continuous derivative thresholding function. In comparison to EMF, TSA2 improves the speech recognition rates for low SNRs.

The next section summarizes the noise reduction with wavelets, section 3 presents previous speech enhancement work and section 4 describes our method. Sections 5 and 6 present the experimental conditions. Section 7 presents the experiments and results on artificially noisy speech with narrow- and wide-band noises, section 8 evaluates the quality of the methods, while section 9 describes an improved version (TSA2) and gives the listening test results and speech recognition rates on the AURORA–2 *testa* and *testb* sets. Finally, section 10 is the discussion and conclusion.

2 Noise reduction with wavelets

In this section, we present the most popular denoising methods based on the wavelet transform. The techniques dedicated to speech *enhancement* are presented in the next section (section 3).

2.1 Principle

Two basic approaches have been proposed to remove noise with wavelet transforms. The first is based on the singularity information analysis (Mallat and Hwang, 1992), whereas the second is based on the thresholding of the wavelet coefficients (Donoho, 1993).

Mallat *et al.* (Mallat and Hwang, 1992) proved that the modulus maxima of the wavelet coefficients give a complete representation of the signal and they proposed an iterative algorithm to remove noise. In the singularity analysis context, Xu *et al.* (Xu et al., 1994) developed a noise filtration method based on the spatial correlation between the wavelet coefficients over adjacent scales. An improved version is proposed by Pan et al. (Pan et al., 1999). The thresholding method is described in the next subsection.

2.2 Wavelet shrinkage

Donoho and Johnstone proposed their original denoising method (Donoho, 1993; Donoho and Johnstone, 1994), which proceeds by thresholding wavelet coefficients of artificial signals. They attempt to recover a signal $s(t)$ from noisy data $x(t)$ with a Gaussian white noise b_i.

$$x_i = s_i + b_i \qquad i = 1, \ldots, N \tag{1}$$

This algorithm can be summarized in three steps

- Wavelet transform (WT) of the noisy signal,
- Thresholding the resulting wavelet coefficients,
- Transformation back to obtain the cleaned signal.

Donoho and Johnstone (Donoho and Johnstone, 1994; Donoho, 1995) define the soft thresholding function by

$$T_S(\lambda, w_k) = \begin{cases} \text{sgn}(w_k)(|w_k| - \lambda) & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases} \tag{2}$$

where w_k represents the wavelet coefficients.

They proposed a universal threshold λ for the WT :

$$\lambda = \sigma\sqrt{2\log(N)} \tag{3}$$

with $\sigma = MAD/0.6745$, where N is the length of x and σ is the noise level. MAD is the median of the absolute value of the wavelet's coefficients estimated on the first scale. In the context of the WT, Johnstone and Silverman (Johnstone and Silverman, 1997) studied the correlated noise situation and proposed *a level–dependent* threshold

$$\lambda_j = \sigma_j\sqrt{2\log(N)} \tag{4}$$

with $\sigma_j = MAD_j/0.6745$ and MAD_j is the median of the absolute value of the coefficients, estimated on level j. The discriminatory threshold can also be defined by using other criterion such as Minimax and SURE (Stain's Unbiased Risk Estimate) (Donoho and Johnstone, 1995; Zhang and Desai, 1998a).

In the Wavelet Packets Transform (WPT) case, the threshold is defined as:

$$\lambda = \sigma\sqrt{2\log(N\log_2 N)} \tag{5}$$

and is not adapted with the subbands.

According to the results obtained by Vidakovic and Lozoya (Vidakovic and Lozoya, 1998), the time–adaptation of the threshold that takes into consideration the time behavior of the noisy signal constitutes an interesting approach.

To our knowledge, even if the wavelet transform has been extensively combined with other methods to improve the speech quality of corrupted speech, the standard wavelet thresholding has not been successfully applied to speech enhancement. In the next section, we report some of these works and thereafter, our method to enhance speech by spatially and time adapting the thresholds in the context of Wavelet Packet Transforms (WPT).

3 Application to speech enhancement

Even if the classical wavelet thresholding technique cannot be used directly, as the simple threshold can not discriminate efficiently the speech components from those of the noise, the wavelet transforms are successfully combined with other denoising algorithms and can improve the performance of speech enhancement methods. But, these wavelet-base methods generally need an estimation of the noise. They include the Wiener filtering in the wavelet domain (Mahmoudi, 1997), wavelet filter bank for spectral subtraction (Gulzow et al., 1998) or coherence function (Sika and Davidek, 1997; Mahmoudi and Drygajlo, 1998).

3.1 Wavelet thresholding

An algorithm based on wavelet thresholding has been proposed for speech enhancement (Seok and Bae, 1997). To prevent the speech quality deterioration during the thresholding process, the unvoiced regions are first classified and then thresholding is used. Even if the problem is not satisfactory solved (voiced/unvoiced decision necessary), this approach is a potential solution to prevent speech degradation.

3.2 Wiener filtering in the wavelet domain

The wavelet transform based Wiener filtering is a special application of the Wiener filtering. This idea arises from the fact that wavelet transforms tend to uncorrelate data. A multi-microphone system is proposed for speech enhancement (Mahmoudi, 1997). The Wiener filtering performance in the wavelet domain are better than those obtained in the Fourier domain. For example, Cohen (Cohen, 2001) proposes a speech enhancement technique based on a modified Wiener filtering. Another version that combines Wiener and coherence in the wavelet domain has been also proposed (Mahmoudi and Drygajlo, 1998).

3.3 Wavelet filter bank

Most speech enhancement systems are conceived around filter banks. This tendency can be justified by the behavior of the cochlea, that operates as a bank of nonlinear dynamical filters. In addition, it is known that the frequency bands of the cochlear filters are not uniformly distributed. Several transformations (scales) are proposed to take into account the perceptive aspect of hearing (Mel, Bark, etc...). The wavelet transform is used as a bank of filters (not uniformly distributed) to improve performance of the speech enhancement method based on the spectral subtraction (Gulzow et al., 1998). A modified version of the speech enhancement method based on the coherence function is proposed by Sika and Davidek (Sika and Davidek, 1997) where the wavelet transform is also used as a bank of filters. Cohen (Cohen, 2001) proposes to use a Bark scale with WPT. He also uses an estimate of the signal to noise ratio that is closely related to the Ephraim and Malah estimate.

4 New enhancement method

As pointed out previously, the wavelet thresholding techniques have not been successfully applied to speech enhancement. These difficulties are related to the speech signal complexity and to the nature of the noise. To improve *the wavelet thresholding* performance, we propose two approaches (Bahoura and Rouat, 2001b): 1) extend the concept of the *scale–dependent* threshold that was first developed for wavelet transform (WT) to the wavelet packet transform (WPT), 2) adapt in time the thresholds according to a nonlinear function of the wavelet coefficient energies.

The proposed algorithm is the natural continuation of the *time–adapted* thresh-

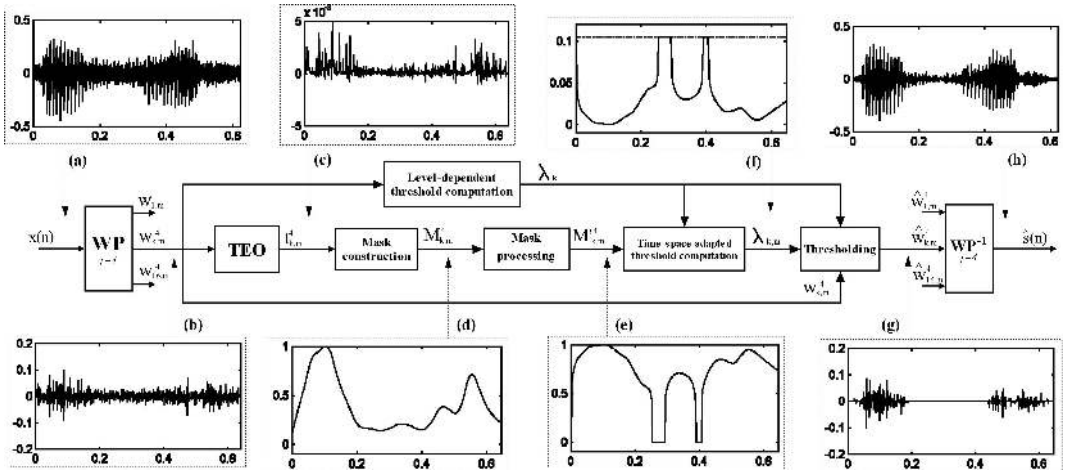

Fig. 1. Speech enhancement diagram using the proposed time–adapted thresholding in the wavelet packet domain

old. Fig. 1 explains schematically this algorithm for a short noisy sentence.

4.1 Wavelet packet analysis

The Wavelet Packet Transform is an extension of the Wavelet Transform. For a given level j, the WPT decomposes the noisy signal $x(n)$ into 2^j subbands corresponding to wavelet coefficient sets $w^j_{k,m}$.

$$w^j_{k,m} = WP\{x(n), j\} \qquad n = 1, \ldots, N \tag{6}$$

By increasing the value of j, the bandwidth of all subbands decreases which improves the scale-adaptation of the discriminatory threshold, specially in the narrow-band noise case. Consequently, the noise is considerably reduced but the quality of the reconstructed speech is affected. In this project, we fix $j = 4$ as a compromise between noise removal and speech intelligibility. Thus, $w^4_{k,m}$ defines the m^{th} coefficient of the k^{th} subband, where $m = 1, ..., N/2^4$ and $k = 1, ..., 2^4$. Fig. 1(b) represents the wavelet coefficient set $w^4_{5,m}$.

4.2 Scale–adapted threshold

The *scale–adapted* threshold is derived from the *scale– dependent* threshold (Equation 4). For a given subband k, the corresponding threshold is defined by:

$$\lambda_k = \sigma_k \sqrt{2\log(N)} \qquad k = 1, \ldots, 16 \tag{7}$$

where $\sigma_k = MAD_k/0.6745$ is the noise level and N is the length of the signal. MAD_k is the median of the absolute value estimated on the subband k.

4.3 Teager Energy Operator

The *time–adapting* approach is introduced by using the Teager Energy Operator (TEO) (Bahoura and Rouat, 2001a) to create a mask. We applied this operator to the resulting wavelet coefficients $w^4_{k,m}$ of each subband k:

$$t^4_{k,m} = [w^4_{k,m}]^2 - w^4_{k,m-1} w^4_{k,m+1} \tag{8}$$

This operation enhances the ability to discriminate speech coefficients from those of noise (Fig. 1(c)).

4.4 Masks Construction

We construct an initial mask for each subband k by smoothing the corresponding TEO coefficients and normalizing (Fig. 1(d)):

$$\mathrm{M}^4_{k,m} = \frac{t^4_{k,l} * h_k(m)}{\max(|t^4_{k,l} * h_k(m)|)} \tag{9}$$

where h_k is an IIR lowpass filter (2^{nd} order) and *max* is the maximum of the smoothed TEO coefficients in the considered subband.

4.5 Time–modulation

For each wavelet's subband k, the corresponding threshold λ_k should be *time–adapted* only for speech–like frames and kept unchanged (that means equal to the maximum universal value λ) for noisy–like ones. By doing so, the cleaning of noise will be maximal when noise is dominant in the wavelet's subband for the speech frame under consideration. The speech dominance is interpreted as an observation of a significant contrast between peaks and valleys of the mask M^4_k, while its absence is observed with a weaker contrast. To distinguish these frames, we define a parameter S^4_k named *offset*, that estimates the valley's level. It is given by the abscissa of the maximum of the amplitude distribution H of the corresponding mask $M^4_{k,m}$, and is estimated over the analyzed frame:

$$S^4_k = \mathrm{abscissa}[H(M^4_{k,m})] \tag{10}$$

If S_k^4 is below the discriminatory value of 0.35 (determined experimentally to discriminate speech from silence), it is assumed that speech is dominant in the k^{th} wavelet's subband (for the current frame), then the threshold is modulated. Otherwise it remains unchanged and the denoising will be at its maximum for all the frame duration. Therefore, for a fixed wavelet's scale k, when speech is dominant in a frame, the threshold is modulated on a very short-time scale (for each coefficient).

4.6 Mask processing for the time–adapting threshold

The modulated threshold must be adapted to the speech waveform independently of its absolute time energy evolution. In this case, the difference between local maxima must be reduced. We proceed by suppressing the *offset* and by normalizing the mask, before applying a root power function of $\frac{1}{8}$. This value is a compromise between noise removal and speech distortion.

$$M'^4_{k,m} = \begin{cases} \left[\frac{|M^4_{k,m}| - S^4_k}{max_m(|M^4_{k,m}| - S^4_k)}\right]^{\frac{1}{8}} & \text{if } S^4_k < 0.35 \\ 0 & \text{if } S^4_k \geq 0.35 \end{cases} \tag{11}$$

$M'^4_{k,m}$ is shown in Fig. 1(e) for $k = 5$.

4.7 Time–scale adapted threshold (TSA)

For each wavelet's subband k, the time–scale adapted threshold is obtained by adapting the corresponding threshold in the time domain:

$$\lambda_{k,m} = \lambda_k(1 - \alpha M'^4_{k,m}) \tag{12}$$

where λ_k is the *scale–dependent* threshold (Equation 7) and α an adjustment parameter ($\alpha = 1$).

Fig. 1(f) represents the *scale–dependent* threshold λ_k (dashed line) and the resulting time adapted threshold $\lambda_{k,m}$ (continuous line) for the wavelet's subband $k = 5$.

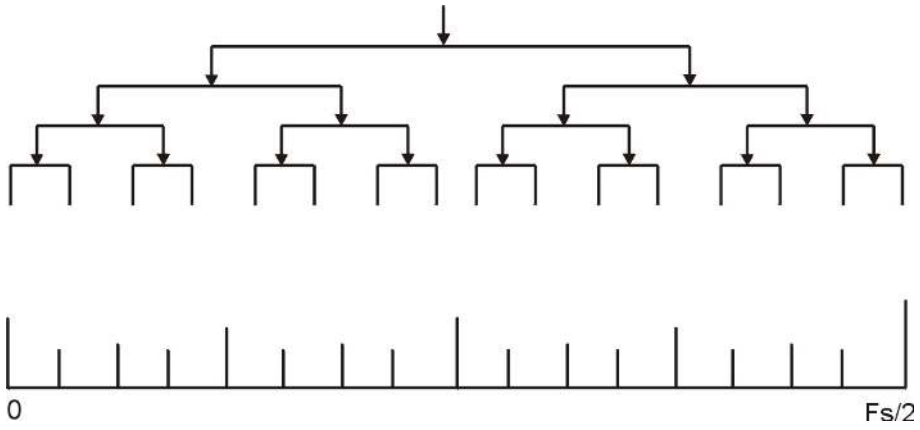

Fig. 2. 16-subband wavelet packet tree

4.8 Thresholding process

The soft thresholding (Equation 2) is then applied to the wavelet packet coefficients (Fig. 1(g))

$$\widehat{w}^4_{k,m} = T_S(\lambda_{k,m}, w^4_{k,m}) \tag{13}$$

where $\lambda_{k,m}$ is the *time–scale* adapted threshold.

4.9 Inverse transformation

The enhanced signal (Fig. 1(h)) is synthesized with the inverse transformation WP^{-1} of the processed wavelet coefficients

$$\widehat{s}_n = WP^{-1}\{\widehat{w}^4_{k,m}, j\} \tag{14}$$

5 MEL-scale multirate filterbank

In our previous work (Bahoura and Rouat, 2001a,b), the speech signal was enhanced by using a bank of 16 wavelet filters, uniformly distributed in the frequency domain (Fig. 2). In this paper, we also extend the previous enhancement approaches to non-uniformly distributed filterbanks (pseudo MEL-scales).

Wavelet filterbanks have been proposed for speech recognition (Jabloun et al., 1999) and speaker identification (Sarikaya et al., 1998). These applications use respectively 21 subbands (Fig. 3) and 24 subbands (Fig. 4).

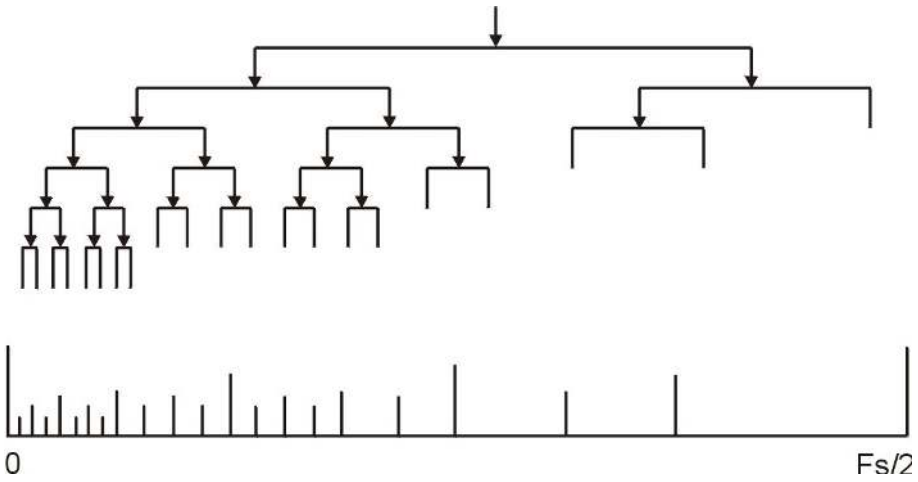

Fig. 3. 21-subband wavelet packet tree

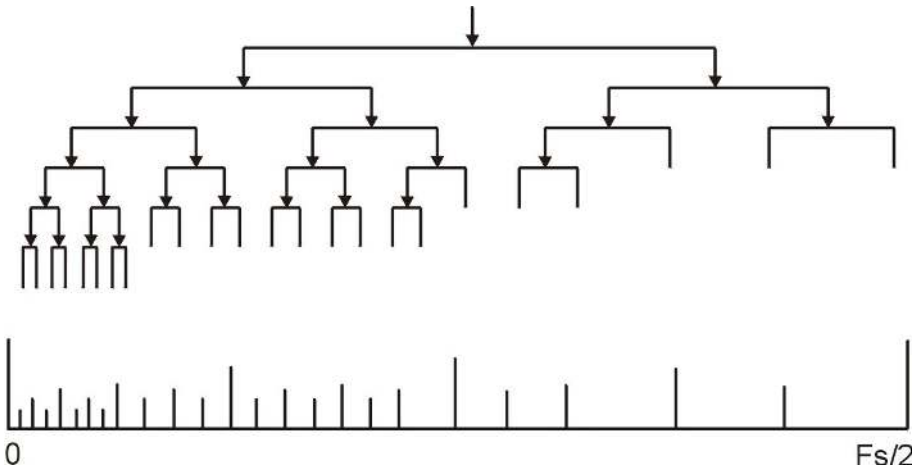

Fig. 4. 24-subband wavelet packet tree

To evaluate the impact of the filterbank, we test three other filterbanks (MEL1, MEL2 and MEL3) based on Daubechies wavelets. MEL1 and MEL2 are based respectively on 21 and 24 filters (according to the decomposition trees illustrated in Fig. 3 and Fig. 4 respectively). The last filterbank MEL3 is obtained by dividing the low frequencies of MEL2 (Fig. 5) and comprises 32 subbands.

6 Signal to Noise evaluations

We define the evaluation measures that will be used. They are based on the estimation of SNR. A signal test $x(n)$ is created by combining a clean speech sentence $s(n)$ and a noise $b(n)$.

$$x(n) = s(n) + b(n) \tag{15}$$

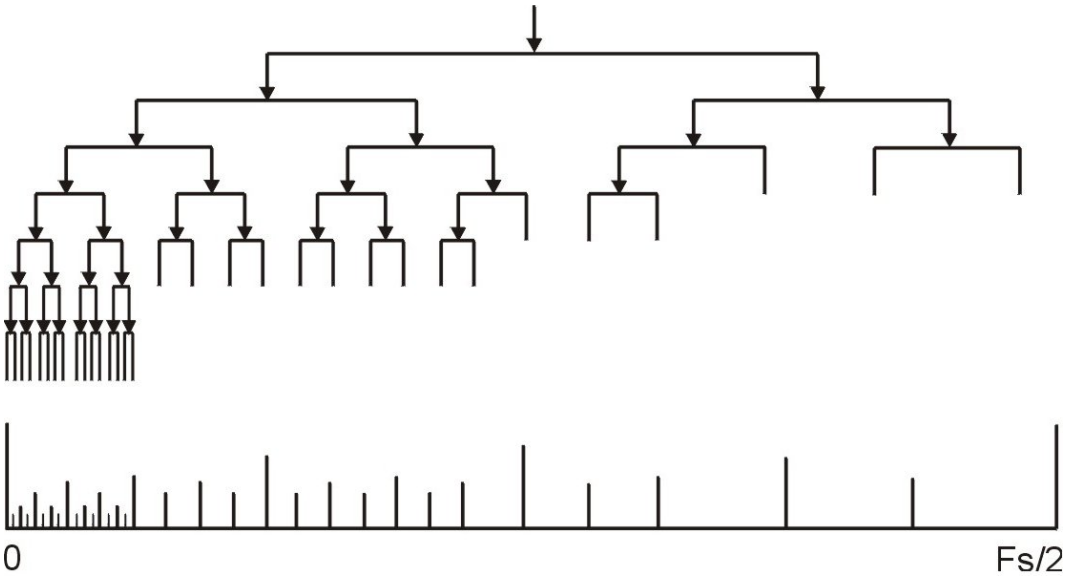

Fig. 5. 32-subband wavelet packet tree

6.1 Unprocessed noisy Speech-to-Noise Ratio

The SNR of the unprocessed noisy speech is defined as the ratio of the clean signal power to the noise power.

$$\mathrm{SNR}_U = 10 \log \frac{\sum_{n=1}^{N} s(n)^2}{\sum_{n=1}^{N} b(n)^2} \tag{16}$$

where N is the length of the sentence expressed in number of samples.

6.2 Processed speech Signal to Noise Ratio

As the enhancement can amplify or attenuate the signal, and for a homogeneous evaluation and comparison between the enhancement methods, we scale the enhanced signal to the same dynamic range as the clean speech. It is accomplished by normalizing the enhanced sound $\hat{s}(n)$ to the clean sound $s(n)$. The resulting scaled signal $\tilde{s}(n)$ is defined as:

$$\tilde{s}(n) = \hat{s}(n) \frac{\max(|s(n)|)}{\max(|\hat{s}(n)|)} \tag{17}$$

As described in (Deller et al., 1993), the efficiency of the enhancement method is defined by the SNR of the enhanced speech and is computed as:

$$\mathrm{SNR}_P = 10 \log \frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} (\tilde{s}(n) - s(n))^2} \tag{18}$$

The denominator is the difference between the clean original signal and the enhanced scaled signal. A small difference characterizes a good match between the two signals.

7 Experiments and results on artificially created noisy speech

We recall that the wavelet thresholding method has been initially proposed to remove additive white noise (Donoho, 1993; Donoho and Johnstone, 1994). In this section we use speech signals also corrupted by narrow–band noises that are not white. Experiments show that our thresholding approach is also efficient for that kind of noise.

The proposed approach is tested and evaluated using speech corrupted with white noise, fan, car noisesand speech recorded in real environments.In this section, we present the performance of the proposed method for clean speech corrupted by additional noise at various SNRs. The results on real environments are reported in section 8.

The size of the analysis frame has been set equal to the length of the speech file, while one estimate of the noise (at the beginning of the sentence) has been used for the Ephraim and Malah algorithm (EMF).

7.1 Time adaptation of the threshold (TA)

We apply the time-adapted thresholding technique (TA) to white wide–band noise and to narrow–band noise.

7.1.1 White noise

The speech sentence from the TIMIT (Garofolo et al., 1993) database has been corrupted with white noise with various Signal-to-Noise Ratios (SNR). The speech signals are sampled at 8 kHz. Results are reported on table 1 and an example is given in Fig. 6.

Table 1
SNR tests for white noise corrupted speech; Time Adaptation only

SNR (dB)	TA (dB)	TAMEL1 (dB)	TAMEL2 (dB)	TAMEL3 (dB)	EMF (dB)
-10	0.99	-0.61	-0.61	1.11	-0.56
-5	3.23	2.16	2.01	2.44	2.66
0	6.36	5.77	6.21	5.40	4.99
5	9.36	9.26	9.09	9.18	7.14
10	12.17	11.66	11.55	12.16	10.49

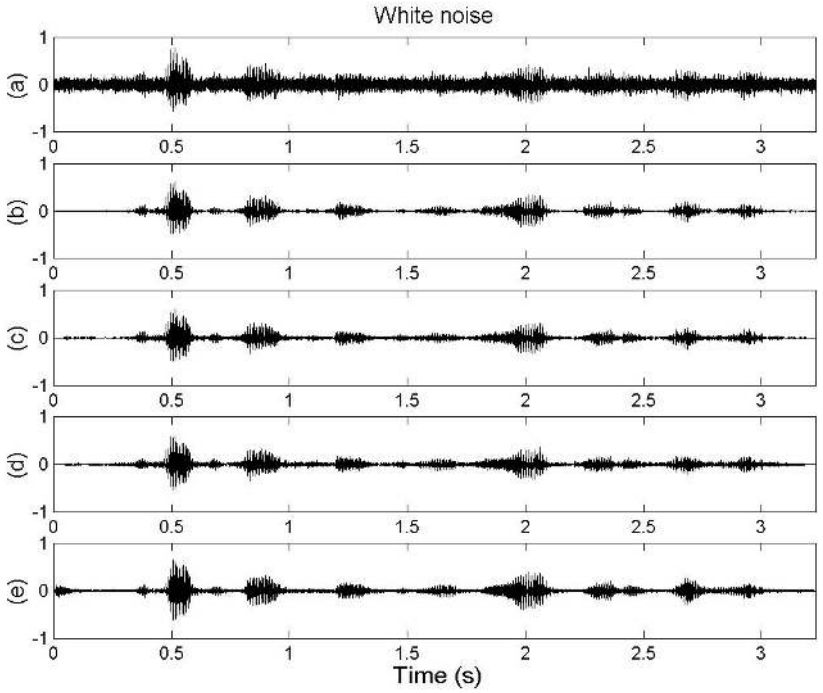

Fig. 6. a) Speech corrupted with white noise (SNR=0dB), enhancement results using b) TA filtering, c) TAMEL1, d) TAMEL3 and e) EMF filtering.

The first column of table 1 gives the SNR expressed in dB. The other columns give the SNR for five enhancement techniques. TA is the time adapted thresholding technique as illustrated in Fig. 1 (no adaptation depending on the wavelet's subbands) and uses a 16-subband wavelet packet (Fig. 2) decomposition. TAMEL1, TAMEL2 and TAMEL3 are an extended version of TA with no adaptation according to the subband and correspond respectively to 21 subband (Fig. 3), 24 subband (Fig. 4) and 32 subband MEL wavelet packet decompositions (Fig. 5). EMF is the Ephraim and Malah Filter (Ephraim and Malah, 1984, 1985).

It is observed that, for white noise, the proposed TA and TAMEL methods are well suited to remove very strong noise with an initial SNR ranging from -10 dB to +10 dB.

Table 2
SNR for speech corrupted by fan noise; Time Adaptation only.

SNR (dB)	TA (dB)	TAMEL1 (dB)	TAMEL2 (dB)	TAMEL3 (dB)	EMF (dB)
-10	-7.33	-7.35	-7.26	-7.59	1.00
-5	-3.03	-3.10	-.07	-3.34	3.79
0	1.11	1.40	1.43	1.28	6.30
5	6.21	6.18	6.23	6.63	8.63
10	10.65	10.55	10.58	10.81	11.02

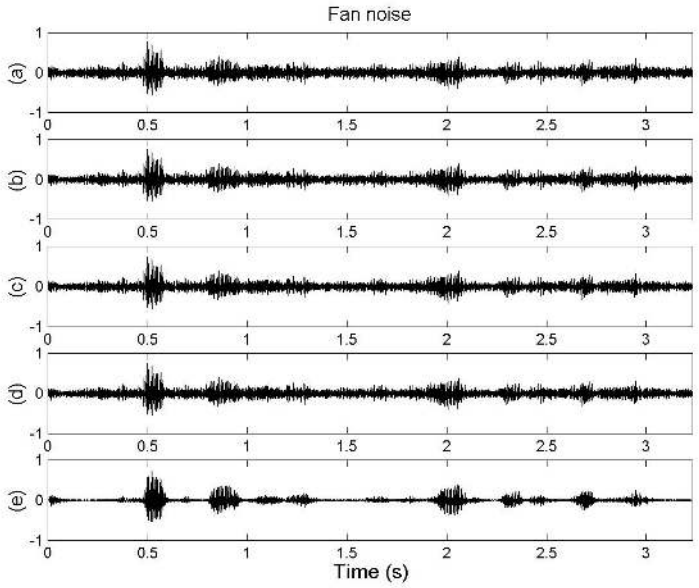

Fig. 7. a) Speech corrupted with fan noise (SNR=0dB), enhancement results using b) TA, c) TAMEL1, d) TAMEL3 and e) EMF.

7.1.2 Fan and car noises

Fan noise (Fig. 7) and car noise (Fig. 8) are band-limited and stationary noise. It is observed by visual inspection of Fig. 7, 8, tables 2 and 3 that the enhancement by thresholding is not adequate for that kind of noise. In fact, in comparison to the thresholding, the EMF is always better even if the threshold is adapted in time.

Such results are predictable as the noises are band-limited and the standard threshold that we temporally adapt is not optimal for each subband and is estimated by using the first detail (high frequencies). Therefore, a standard threshold cannot be used to discriminate the signal coefficients from that of the noise as it is not suitable for each subband. A spatial adaptation of the threshold for each subband is necessary.

Table 3
SNR for speech corrupted by car noise; Time Adaptation only.

SNR (dB)	TA (dB)	TAMEL1 (dB)	TAMEL2 (dB)	TAMEL3 (dB)	EMF (dB)
-10	-7.34	-7.44	-7.41	-7.46	6.79
-5	-3.24	-3.37	-3.34	-3.49	10.35
0	1.34	1.12	1.15	1.00	12.95
5	6.37	5.93	5.95	5.77	14.53
10	10.49	10.88	10.89	10.73	15.77

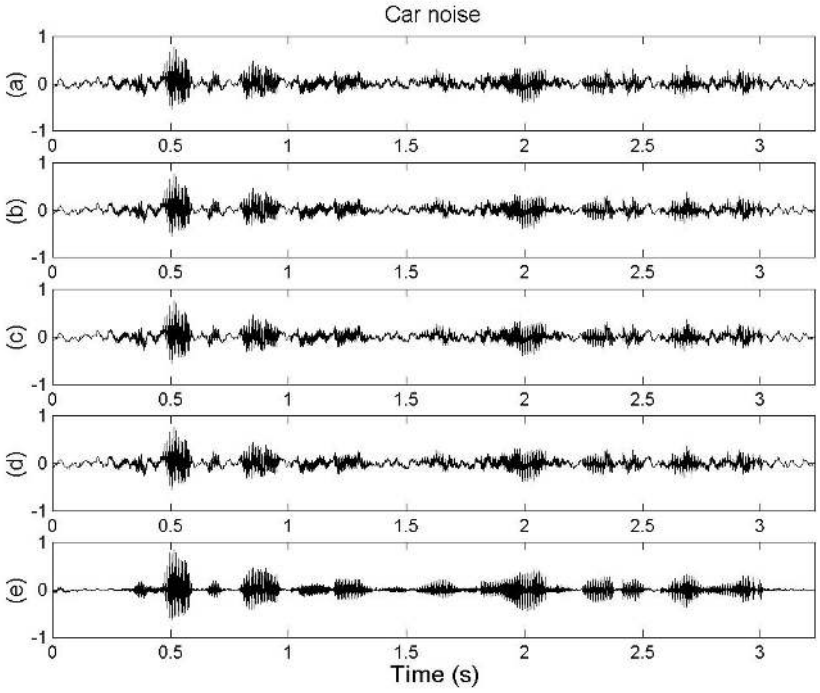

Fig. 8. a) Speech corrupted with car noise (SNR=0dB), enhancement results using b) TA filtering, c) TAMEL1, d) TAMEL3 and e) EMF.

7.2 Scale and Time adaptation of the threshold (TSA)

To extend the usefulness of techniques based on the thresholding of wavelet packet coefficients to the suppression of various kinds of noise, we propose to combine a spatial adaptation of the discriminative threshold with the time adaptation (TA). We denote TSA this new Time and Scale Adapted threshold technique.

The approach is simple and allows the extension of the principle of the time adapted threshold depending on the level to reduce the noise in the wavelet packet domain.

Table 4
SNR tests for white noise corrupted speech; Time and Scale Adaptation.

SNR (dB)	TSA (dB)	TSAMEL1 (dB)	TSAMEL2 (dB)	TSAMEL3 (dB)	EMF (dB)
-10	1.00	-0.74	-0.73	0.93	-0.56
-5	3.11	2.25	2.06	2.40	2.66
0	6.17	5.59	5.94	5.22	4.99
5	8.85	8.78	8.55	8.92	7.14
10	11.10	10.78	10.69	11.65	10.49

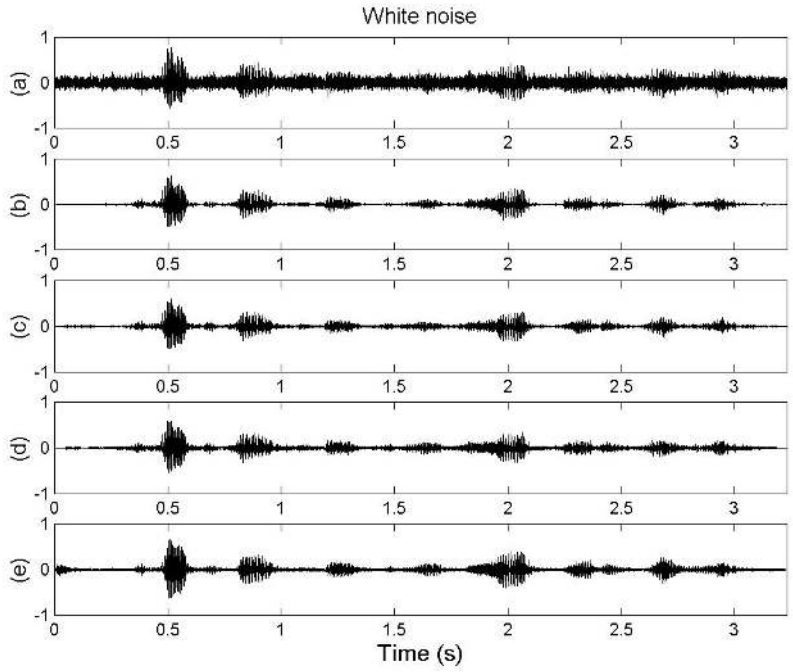

Fig. 9. a) Speech corrupted with white noise (SNR=0dB), enhancement results using b) TSA filtering, c) TSAMEL1, d) TSAMEL3 and e) EMF

The time and scale threshold (TSA) experimental results are reported in next subsections.

7.2.1 white noise

Table 4 summaries the performance of TSA when a white noise is being used. A comparison with table 1 shows that the TSA yields also higher performance than the EMF, but is slightly less robust than TA. From table 4, it is also observed that TSAMEL3 is better than TSA for the highest SNR. An example of the filtering is given in Fig. 9.

Table 5
SNR tests for speech corrupted with fan noise; Time and Scale Adaptation.

SNR (dB)	TSA (dB)	TSAMEL1 (dB)	TSAMEL2 (dB)	TSAMEL3 (dB)	EMF (dB)
-10	0.19	0.60	0.66	1.90	1.00
-5	2.91	3.91	3.99	3.86	3.79
0	4.36	6.02	5.98	6.61	6.30
5	7.10	8.31	8.26	8.40	8.63
10	10.16	10.33	10.29	10.62	11.02

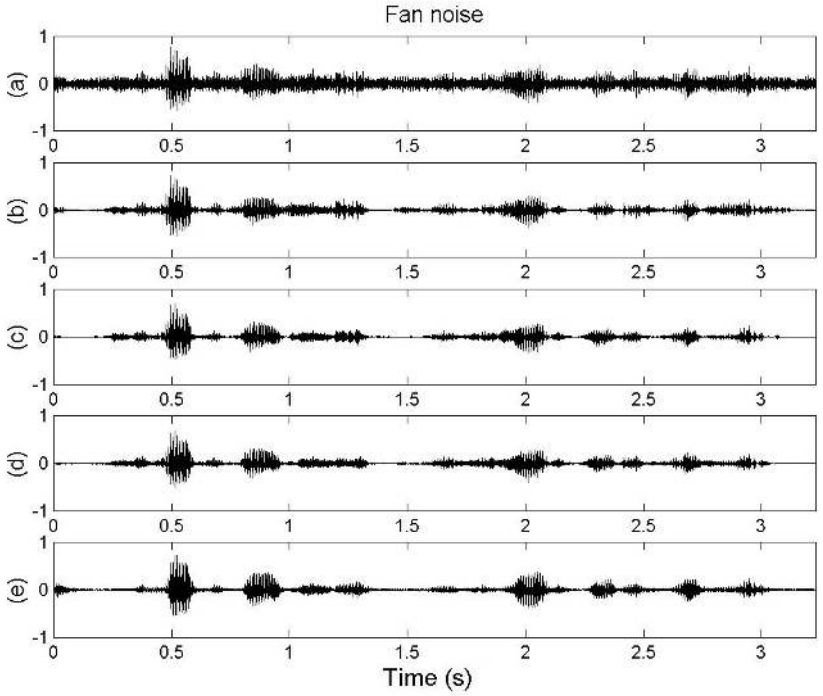

Fig. 10. a) Speech corrupted with fan noise (SNR=0dB), enhancement results using b) TSA filtering, c) TSAMEL1, d) TSAMEL3 and e) EMF.

7.2.2 Fan and car noise

TSA is being performed on the fan and car noises. Table 5 and Fig. 10 are given for the fan noise while Table 6 and Fig. 11 are for the car noise.

In comparison with table 2, table 5 shows a significant improvement of the SNR when TSA is used with fan noise. For SNR less than or equal to 0 dB, the TSAMEL3 yields higher SNR than the reference method (EMF).

Table 6 reports the system results when noise recorded in a Volvo car has been added to the signal. It is also observed that TSAMEL3 gives the best results for initial SNR less or equal to 0 dB, while TAMEL3 gave the worst results as reported in table 3.

Table 6
SNR tests for speech corrupted with car noise; Time and Scale Adaptation.

SNR (dB)	TSA (dB)	TSAMEL1 (dB)	TSAMEL2 (dB)	TSAMEL3 (dB)	EMF (dB)
-10	10.23	10.95	10.94	11.25	6.79
-5	10.96	12.24	12.24	12.72	10.35
0	11.33	12.78	12.79	13.57	12.95
5	11.46	13.08	13.08	14.18	14.53
10	11.62	13.81	13.82	14.49	15.77

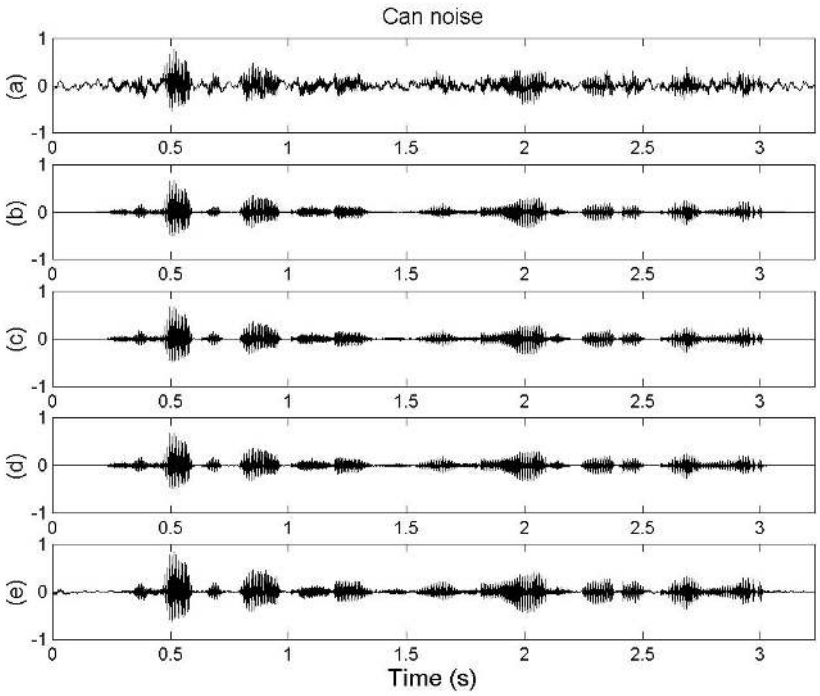

Fig. 11. a) Speech corrupted with car noise (SNR=0dB), enhancement results using b) TSA filtering, c) TSAMEL1, d) TSAMEL3 and e) EMF

This time, the MEL–based approaches yield the highest increase in SNR. The EMF is better for SNR higher than 0dB.

8 Experiments and results on naturally noisy speech

In the previous section we have reported quantitative performance based on SNR of artificially created noisy sentences. We propose here a more qualitative evaluation based on noisy speech recorded in real environments.

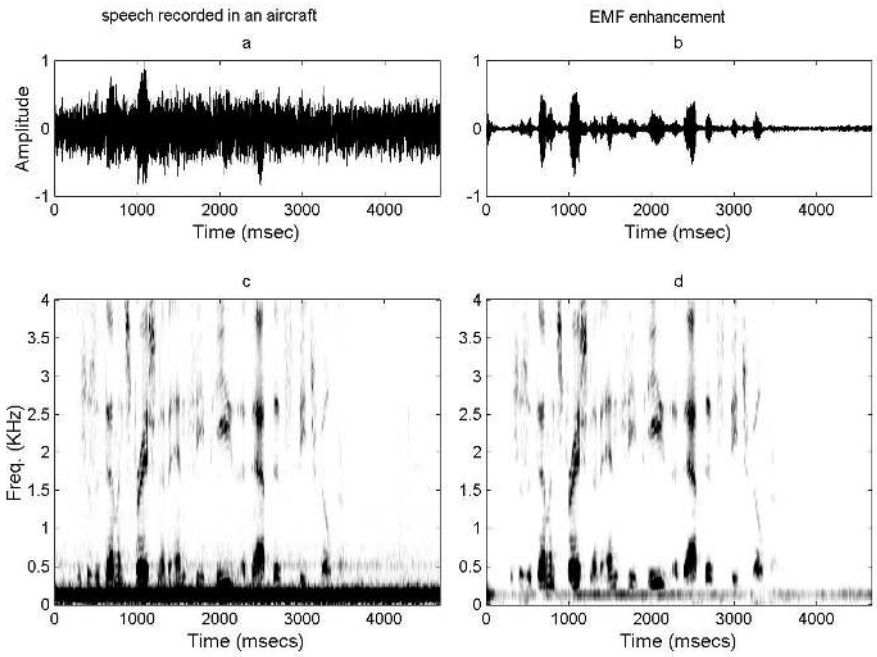

Fig. 12. a) noisy speech recorded in an aircraft, b) EMF enhancement, their spectral representations respectively in c) and d).

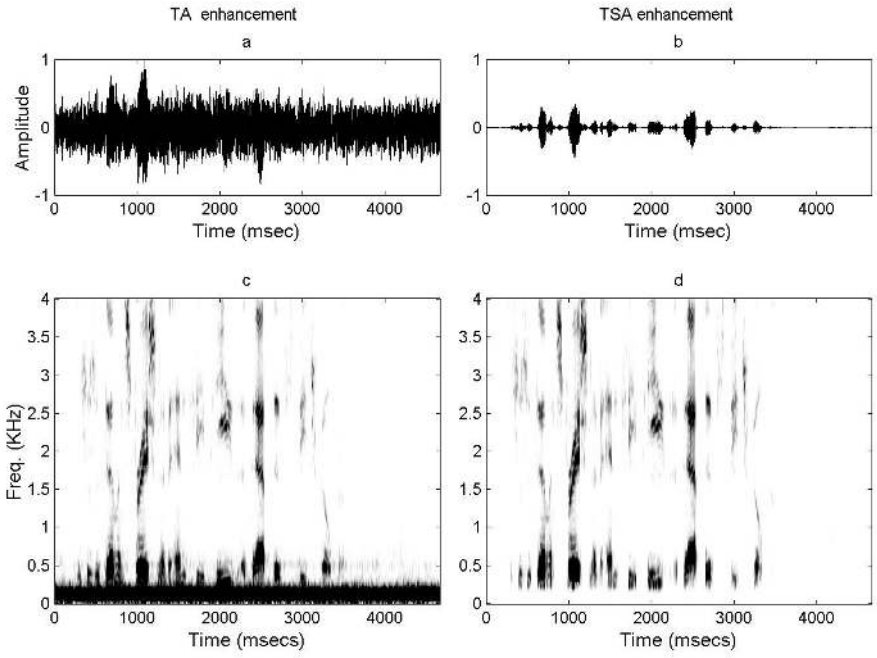

Fig. 13. Enhancement of speech recorded in an aircraft using a) TA, b) TSA, their spectral representations respectively in c) and d).

8.1 Narrow–band noise

We recall that the wavelet thresholding method has been initially proposed to remove additive white noise. In this section we use a speech signal recorded

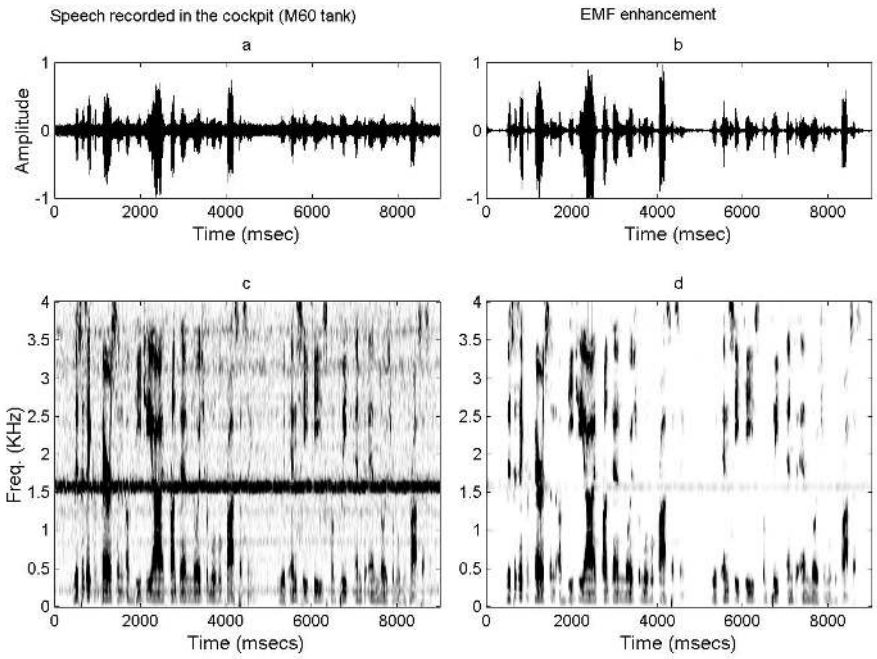

Fig. 14. a) noisy speech recorded in the cockpit of a M60 tank, b) EMF enhancement, their spectral representations respectively in c) and d).

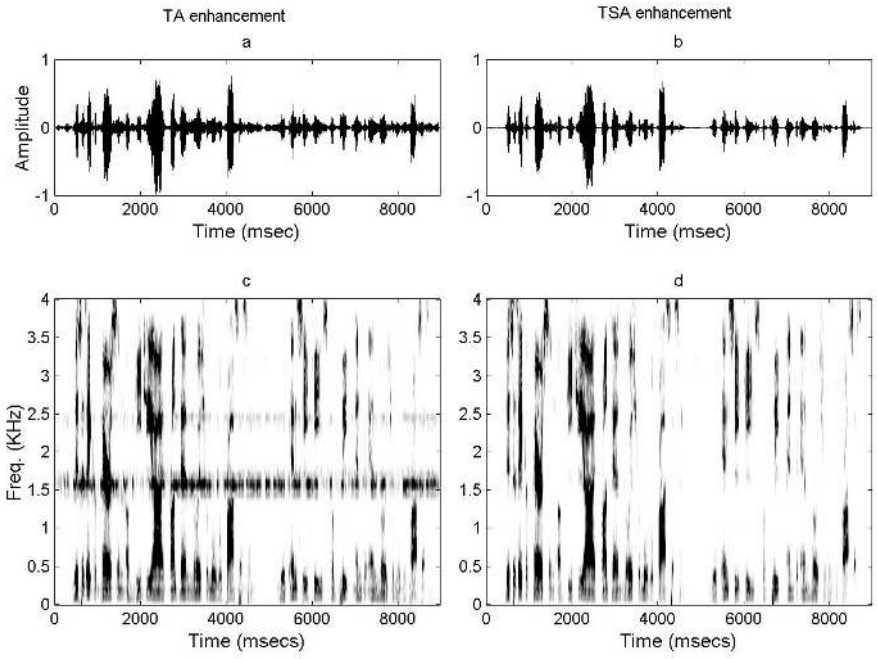

Fig. 15. Enhancement of speech recorded in the cockpit of a M60 tank using a) TA, b) TSA, their spectral representations respectively in c) and d).

in a DC9 jet-aircraft (Figs. 12 and 13). The noise is relatively narrow–band (Fig. 12-a and c) and is far from being white. TSAMEL3 provides the best auditory preference with less echo than TSA and TSAMEL1. EMF removes

less noise but generates less artifacts (echo and musical noise).

We also tested the system on a speech signal recorded in a M60 tank. It is corrupted by a relatively stationary noise (Fig. 14 and 15). TA does not remove the noise and worse enhances the noise components centered around 1600 Hz and 2500 Hz (Fig. 15 a–c). The EMF does not remove entirely the noise (Fig. 14 b–d) while TSA does (Fig. 15 b–d). Also, the perception of the EMF is not as good than the TSA. The perceptual difference between TSA, TSAMEL1 and TSAMEL3 is not obvious. TSA seems to be preferred to TSAMEL3 in terms of speech quality.

The usefulness of the *level–dependent* thresholding is emphasized in these examples. In fact, the universal threshold of the WPT is inefficient to remove the band-limited noise like the Time-Adapted Threshold (TA) that is also inefficient (Fig. 15-a). However, the noise is greatly reduced using the *level–dependent* thresholding (Fig. 13-b). The *Time–Scale* Adapted Threshold (TSA) prevents the speech quality deterioration during the thresholding process (Fig. 13 b-d and Fig. 15 b-d).

8.2 Wide–band noise

Noisy speech was recorded in a sawmill with an omnidirectional microphone (Fig. 16-a). The universal threshold method reduces the noise considerably but it is accompanied by speech quality degradation. Our previous solution (TA) (Bahoura and Rouat, 2001a) is very efficient to remove this kind of noise (Fig. 17-a,c). The results obtained by the new approach (TSA) are also quite efficient (Fig. 17-b,d), in comparison to the Ephraim and Malah Filter (Fig. 16-b,d). There is no obvious difference between TA and TSA. Both are very effective. EMF is less efficient and yields a stronger noise with a somewhat better auditory perception.

9 Experiments and results on the AURORA-2 database

In this section we report comparison results on the AURORA–2 database. Listening tests and speech recognition rates are given.

The original Time Scale Adaption (TSA) algorithm is evaluated and compared to an improved version (TSA2) that uses a continuous derivative thresholding function as proposed by Zhang and Desai (Zhang and Desai, 1998a,b). In fact, common hard or soft–thresholding functions introduce nonlinear transformations of the signal spectrum that can, depending on the application, greatly

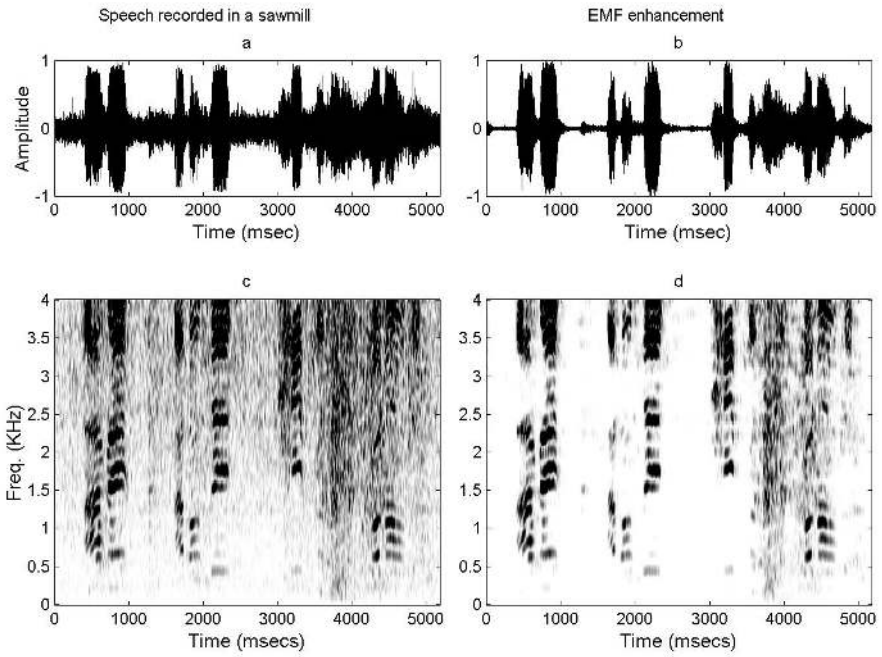

Fig. 16. a) noisy speech recorded in a sawmill, b) EMF enhancement, their spectral representations respectively in c) and d).

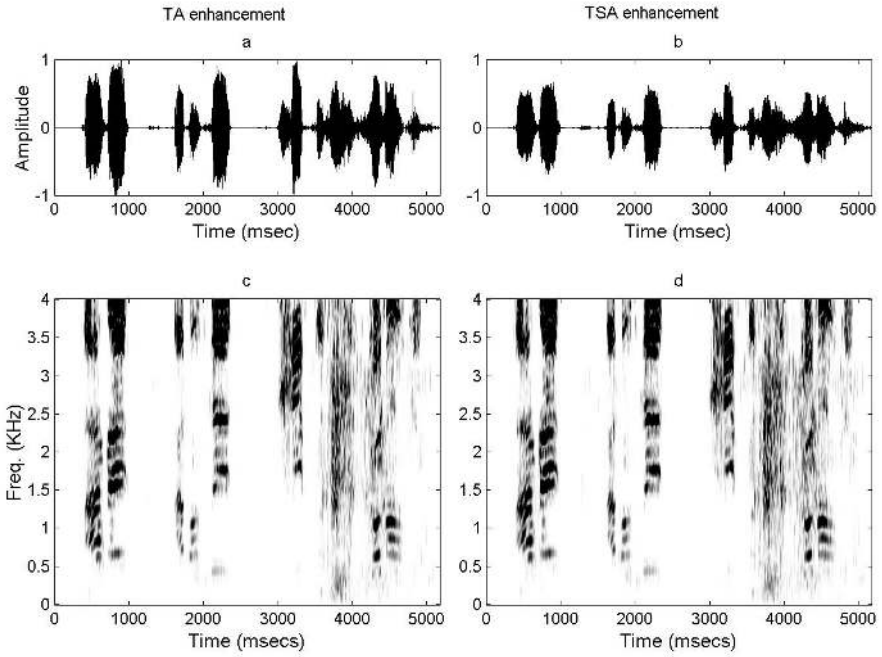

Fig. 17. Enhancement of speech recorded in a sawmill using a) TA, b) TSA, their spectral representations respectively in c) and d).

affect subsequent signal processing. These nonlinear transformations of the spectrum are difficult to detect with listening experiments but can strongly reduce speech recognizer performance.

A new type of shrinkage functions has been developed by Zhang and Desai (Zhang and Desai, 1998a,b). They have continuous derivatives and are defined as follows:

$$\eta_n(\lambda, w_k) = \begin{cases} w_k + \lambda - \frac{\lambda}{2n+1} & \text{if } w_k < -\lambda \\ \frac{1}{(2n+1)\lambda^{2n}} w_k^{2n+1} & \text{if } |w_k| \leq \lambda \\ w_k - \lambda + \frac{\lambda}{2n+1} & \text{if } w_k > \lambda \end{cases} \tag{19}$$

where n is a positive integer. Note that the limit of $\eta_n(\lambda, w_k)$ when $n \to \infty$ is just the commonly used soft-thresholding function $T_S(\lambda, w_k)$. In practice, the authors uses $n = 1$ and $n = 3$. In this work we use $n = 1$.

In the remaining part of the paper, we denote by TSA2 the TSA algorithm where $T_S(\lambda, w_k)$ from equation 2 has been replaced with $\eta_n(\lambda, w_k)$ from equation 19.

9.1 Listening tests

The listening tests are achieved using a convivial tool developed by the audio compression laboratory at Université de Sherbrooke. This tool proposes the AB test, where the listener must choose the best between two signals or decide if he is indifferent. For this experiment, the listening criteria is based on the perceived quality of the enhanced signals. Five French speaking listeners were asked to compare pairs of sentences randomly extracted from *testa* and *testb* AURORA–2 subsets. For each kind of noise, listening tests have been made for each SNRs between [-5dB, 10dB]. The number of pairs is the same for each SNR (the number of files has been balanced). Ninety six sentences have been processed by TSA, TSA2 and EMF.

The AB test shows that the EMF is most of the time preferred to TSA and to TSA2 at low SNRs [-5dB, 10dB]. It is also observed that TSA and TSA2 are indifferently chosen (a great confusion between TSA and TSA2 exists and indicates that there is no preference for one or another). Table 7 is an example of the listening ratings obtained between TSA2 and EMF. Listeners are frequently indifferent to the enhancement methods. When they can decide, they prefer most of the time the EMF enhancement method.

In their recent work, Chen and Wang (Chen and Wang, 2004) used the Mean Opinion Score (MOS) test to evaluate their speech enhancement method in comparison to our TA method and the EMF filter on the AURORA–2 database. Their results show that TA is preferred to EMF for various real environments including airport, car, restaurant, and street. The difference be-

Table 7
Preference ratings (AB test) between TSA2 and EMF for additive noises [-5db,10dB] on *testa* and *testb* of the AURORA–2.

	TSA2 (%)	**EM** (%)	**indifferent** (%)
subway	20	40	40
babble	20	30	50
car	10	50	40
exhibition	10	60	30
restaurant	10	40	50
street	30	40	30
airport	30	30	40
train	30	30	40

tween our results (EMF most of the time preferred to TSA or TSA2) and their results (TA superior to EMF) might be due to *i)* the definition of the quality criteria and to *ii)* the difference between TA and TSA. In our listening experiments we did emphasize on the signal quality instead on the speech intelligibility (as our listeners are French speaking and not English speaking, intelligibility has not been evaluated). Furthermore, each scale in TSA and TSA2 is adapted independently, while TA uses the same time adapted threshold for each scale yielding a continuous change from scales to scales. These differences might explain the greater TA quality when compared to EMF.

In the next subsection we show that TSA2 has a greater potential in speech recognition (we recall here that no training or knowledge is required with our method) for low SNRs [-5dB,10dB].

9.2 Speech recognition

The proposed speech enhancement methods are also evaluated using the Hidden Markov HTK (Young et al., 2000) speech recognition system on *testa* and *testb* sets of the AURORA–2 database. HTK training and recognition have been made with the scripts provided on the AURORA–2 CDs for the full *testa* and *testb* sets. Training is made on the unprocessed clean sentences and recognition is on enhanced noisy sentences. The word (digit) recognition rates have been computed. Preliminary speech recognition experiments that we made have shown that TA and TSA introduce distorsions that strongly degrade the recognizer performance (yielding recognition scores much lower than those obtained with EMF).

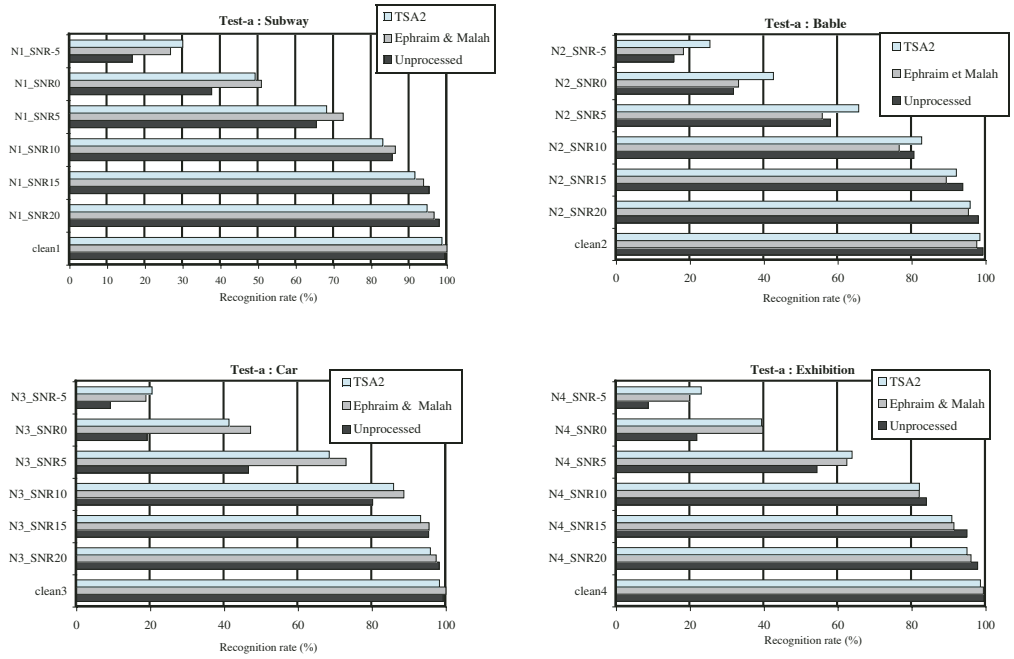

Fig. 18. Speech recognition rates on AURORA–2 *testa* set using the HTK software package for TSA2, EMF and without enhancement (unprocessed). TSA2 is always superior to EMF for the babble noise and better or similar to EMF for the exhibition noise with SNR between -5dB and +10dB. On subway and car noises, TSA2 is superior to EMF only for -5dB SNRs.

Unprocessed and enhanced speech with TSA2 and EMF results are reported on figures 18 and 19. We observe that performance is greatly improved with TSA2 for low SNR situations [-5dB,10dB]. On the *testb* set TSA2 yields the best results compared to EMF. EMF is better for quasi–sationary noises from the *testa* set (like car noise) and for higher SNRs. The proposed method (TSA2) gives the best recognition rates in real environments (like babble noise, restaurant) where conventional enhancement methods are generally very limited (these noises being less stationary). Therefore TSA2 can be considered as being a complementary technique to other speech enhancement methods (like EMF).

10 Conclusion

10.1 Additive noise on initially clean speech

When the increase in SNR is used as criteria, it has been observed that for artificial white noise TA is superior to TSA, which is also better than EMF.

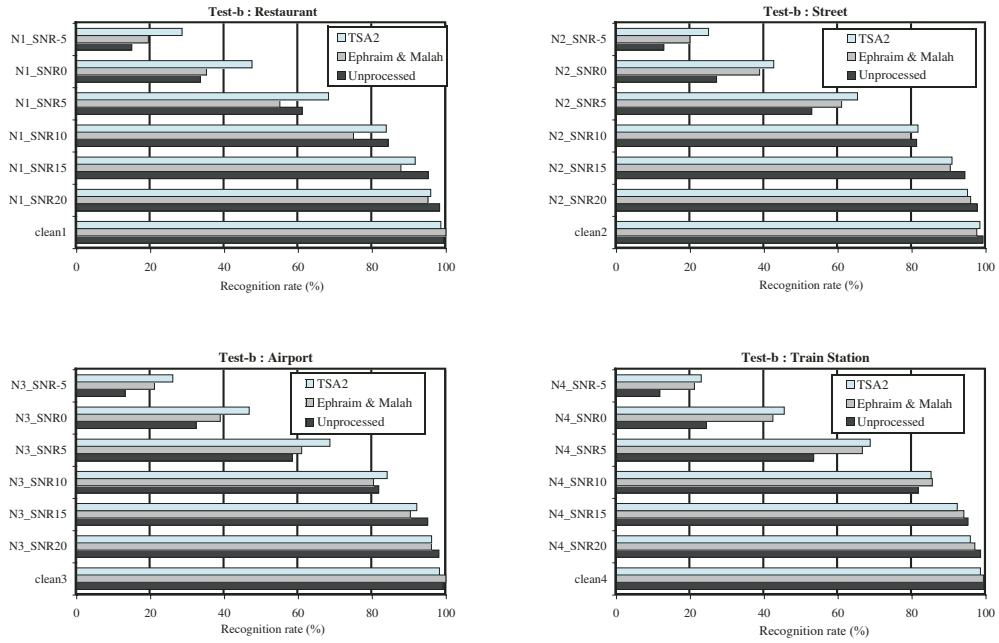

Fig. 19. Speech recognition rates on AURORA–2 *testb* set using the HTK software package for TSA2, EMF and without enhancement (unprocessed). TSA2 is always superior to EMF for all type of noise when the SNRs are lower than 10dB.

The usefulness of a MEL-scale decomposition is not obvious for white wide–band noise. With the Time–Adapted threshold (TA), performance is better when not using the MEL–scale. The same remark is valid for the Time–Scale Adaptation technique (TSA).

Otherwise, and for the kind of band-limited noises that we used, the MEL–scale improves the enhancement performance. With the fan noise, TSAMEL3 provides the best increase in SNR while TA is the worst. With the car noise, TSAMEL3 provides the greatest increase just after the EMF, while TA is the worst. The car noise is very low frequency and the TSAMEL3 uses a bank of wavelets with an important resolution in low-frequencies (32 subbands). It is observed that the EMF is better for $SNR_u \geq 5\ dB$.

10.2 Speech recorded in natural environments

When the speech is recorded in natural environments, it is observed that TA is usually sufficient for wide–band noise (sawmill) while it is absolutely inefficient with narrow–band noises (car and airplane). In that situation, TSAMEL3 usually provides the best increase (32 bands with strong resolution in low frequencies).

The performance of the techniques strongly depends on the nature of the noise and on the applications. From a perceptive point of view, TA, TSA and TSA2 are superior to conventional wavelet shrinkage techniques but not as good than EMF where the noise can be estimated before enhancement. For speech recognition applications in very noisy environments, TSA2 is superior to EMF for a wide range of noise and does not need knowledge of the noise. The Time–Scale Adaptation TSA2 can be used as a complementary technique to other speech enhancement methods as it is efficient on a different kind of noise and does not need *apriori* knowledge of the environment.

11 Acknowledgment

We acknowledge Philippe Boigné for the experiments with HTK and TSA2, Roch Lefebvre from the speech compression group for the AB test software, Arkady Bron for his code of Ephraim and Malah Filter and Douglas O'Shaughnessy for proof reading.

References

Bahoura, M., Rouat, J., January 2001b. Wavelet speech enhancement using the Teager energy operator. IEEE Signal Processing Letters 8, 10–12.

Bahoura, M., Rouat, J., September 3-7 2001a. New approach for wavelet speech enhancement. In: Eurospeech 2001. Aalborg, Denmark, pp. 1937–1940.

Chen, S. H., Wang, J. F., 2004. Speech enhancement using perceptual wavelet packet decomposition and teager energy operator. J. VLSI Signal Process. Syst. 36 (2-3), 125–139.

Cohen, I., September 3-7 2001. Enhancement of speech using bark-scaled wavelet packet decomposition. In: Eurospeech 2001. Aalborg, Denmark, pp. 1933–1936.

Deller, J. R., Proakis, J. G., Hansen, J. H. L., 1993. Discrete-Time Processing of Speech Signals. MacMillan, New York.

Donoho, D., 1993. Nonlinear wavelet methods for recovering signals, images, and densities from indirect and noisy data. Proceedings of Symposia in Applied Mathematics 47, 173–205.

Donoho, D., May 1995. De-noising by soft-thresholding. IEEE Trans. Inform. Theory 41, 613–627.

Donoho, D., Johnstone, I., 1994. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, 425–455.

Donoho, D., Johnstone, I., 1995. Adapting to unknow smoothness via wavelet shrinkage. J. Amer. Stat. Assoc., 1200–1224.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean square error short time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Processing 32, 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean square error log spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Processing 33, 443–445.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., 1993. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. NTIS.

Gulzow, T., Engelsberg, A., Heute, U., 1998. Comparison of a discrete wavelet transformation and nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement. Signal Processing 64, 5–19.

Hirsch, H., Pearce, D., September 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proceedings of the ASR2000 - Automatic Speech Recognition: Challenges for the Next Millennium. Paris, France, pp. 181–188.

Jabloun, F., Cetin, A., Erzin, E., October 1999. Teager energy based feature parameters for speech recognition in car noise. IEEE Signal Processing Letters 6, 259–261.

Johnstone, I., Silverman, B., 1997. Wavelet threshold estimators for data with correlated noise. J. Roy. Statist. Soc. B 59, 319–351.

Mahmoudi, D., September 1997. A microphone array for speech enhancement using multiresolution wavelet transform. In: Proc. Of Eurospeech'97. Rhodes, Greece, pp. 339–342.

Mahmoudi, D., Drygajlo, A., 1998. Combined wiener and coherence filtering in wavelet domain for microphone array speech enhancement. In: ICASSP. Seattle, USA, pp. 385 –388.

Malah, D., Cox, R. V., Accardi, A. J., 1999. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. Vol. 2. p. 789.

Mallat, S., Hwang, W., March 1992. Singularity detection and processing with wavelets. IEEE Trans. Inform. Theory 38, 617–643.

Pan, Q., Zhang, L., Dai, G., Zhang, H., December 1999. Two denoising methods by wavelet transform. IEEE Trans. Signal Processing 47, 3401–3406.

Sarikaya, R., Pellom, B., Hansen, J., June 1998. Wavelet packet transform features with application to speaker identification. In: NORSIG-98 IEEE Norsic Signal Processing Symposium. Vigso, Denmark, pp. 81–84.

Seok, J., Bae, K., April 1997. Speech enhancement with reduction of noise components in the wavelet domain. In: ICASSP 97. Munich, Germany, pp. 1223–1326.

Sika, J., Davidek, V., Spetember 1997. Multi-channel noise reduction using wavelet filter bank. In: EuroSpeech'97. Rhodes, Greece, pp. 2595 – 2598.

Vidakovic, B., Lozoya, C., September 1998. On time-dependant wavelet denoising. IEEE Trans. Signal Processing 46, 2549–2554.

Xu, Y., Weaver, J., Healy, D., Lu, J., November 1994. Wavelet transform domain filters: A spatially selective noise filtration technique. IEEE Trans. Image Processing 3, 747–758.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., July 2000. The HTK Book (for HTK Version 3.0). Microsoft Corporation, Ch. The Fundamentals of HTK, pp. 2–13.

Zhang, X.-P., Desai, M. D., 1998a. Adaptive denoising based on sure risk. Signal Processing Letters, IEEE 5 (10), 265, 1070-9908.

Zhang, X.-P., Desai, M. D., May 1998b. Nonlinear adaptive noise suppression based on wavelet transform. In: Proceedings of ICASSP'98. Vol. 3. pp. 1589–1592.