International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:9, 2007

Spectral Entropy Employment in Speech Enhancement based on Wavelet Packet

Talbi Mourad, Salhi Lotfi, Chérif Adnen

***Abstract*—**In this work, we are interested in developing a speech denoising tool by using a discrete wavelet packet transform (DWPT). This speech denoising tool will be employed for applications of recognition, coding and synthesis. For noise reduction, instead of applying the classical thresholding technique, some wavelet packet nodes are set to zero and the others are thresholded. To estimate the non stationary noise level, we employ the spectral entropy. A comparison of our proposed technique to classical denoising methods based on thresholding and spectral subtraction is made in order to evaluate our approach. The experimental implementation uses speech signals corrupted by two sorts of noise, white and Volvo noises. The obtained results from listening tests show that our proposed technique is better than spectral subtraction. The obtained results from SNR computation show the superiority of our technique when compared to the classical thresholding method using the modified hard thresholding function based on $\mu - law$ algorithm.

***Keywords*—**Enhancement, spectral subtraction, SNR, discrete wavelet packet transform, spectral entropy Histogram

I. Introduction

IN many applications of speech signal processing, we have to process speech in an undesirable background noise presence. Many speech enhancement techniques have been suggested such as spectral subtraction [1] and wavelet thresholding [2]. These techniques necessitate the estimation of statistical noise information. Donoho and Johnstone introduced an estimation technique based on wavelet for statistical information of additive white Gaussian noise. But in many real cases, the noise that corrupts the speech signal is a background noise. White noise is useful as a conceptual entity; however it rarely occurs in real environment. The majority of noises captured by a microphone are colored and their spectres aren't whites. As an example of colored noise, we can mention pink noise which has a low pass nature. We can approximate the noises generated by an air conditioner, a car engine as pink noises. To handle a colored noise, a level dependent thresholding has been proposed by Johnstone and Silverman[3]. However, a great interest to real environmental noises is in non stationary noises. The statistical properties of this type of noises change over time. The car and air conditioner noises aren't perfectly stationary [4]. To handle background noise, Sungwook & al [5] have proposed a node dependent thresholding technique as a level dependent thresholding extension.

Authors are with the Signal Processing Laboratory - Science Faculty of Tunis, 1060 Tunis, Tunisia. e-mail: mouradtalbi1969@yahoo.fr

The noise level estimation is based on spectral entropy. In addition, the classical estimation technique based on MAD supposes that the noise distribution is Gaussian. However, in practice this supposition is not forever valid. That's way Sungwook has proposed a noise estimation technique based on spectral entropy employing intensity histogram of wavelet packet coefficients for each node on adapted wavelet packet tree. In this work, we try to employ their technique for thresholding a part of wavelet packet nodes and the others are set to zero. In this paper, we compare our proposed technique to the wavelet packet thresholding method using a modified hard thresholding function based on $\mu - law$ algorithm, and spectral subtraction based denoising method. The comparison is based on two criteria, one is objective by computing SNR and the other is subjective by making listening tests. The speech signals are corrupted by two types of noises: a white noise and a car noise.

II. Wavelet Techniques

The wavelet based denoising technique employs the decomposition concept in adaptive base of wavelets. This technique is efficient in denoising of signals corrupted by an additive noise [6].

A. Thresholding based denoising technique

A noise reduction technique was developed by Donoho[7]. This technique uses the wavelet coefficients contraction and its principle consists in three steps:

- Applying discrete wavelet transform to noisy signal:

$$W \cdot x = W \cdot s + W \cdot b \quad (1)$$

where x , s , b and W are respectively the noisy speech signal, the clean speech signal , the noise signal and the matrix associated to the discrete wavelet transform.

- Thresholding the obtained wavelet coefficients.
- Reconstructing the desired signal by applying the inverse transform to the thresholded coefficients.

In general, two sorts of thresholding are employed, Hard and Soft thresholding which are expressed by [8, 9]:

$$THR_{Hard} = \begin{cases} x & if \quad |x| > \lambda \\ 0 & if \quad |x| \le \lambda \end{cases} \quad (2)$$

$$THR_{Soft} = \begin{cases} sign(x)(|x| - \lambda) & if \quad |x| > \lambda \\ 0 & if \quad |x| \le \lambda \end{cases} \tag{3}$$

Equations (1) and (2) designate successively the expressions of hard and soft thresholding functions, where λ is the employed threshold.

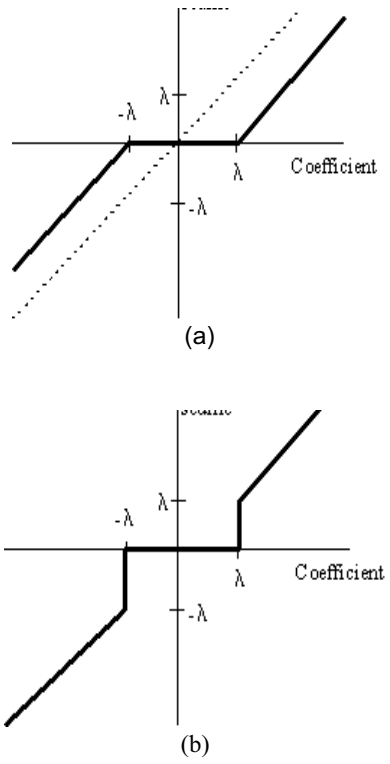


Fig. 1. (a) Soft thresholding. (b) Hard thresholding.

Hard thresholding maintains the scale of the signal but introduces ringing and artefacts after reconstruction due to a discontinuity in the wavelet coefficients. However, soft thresholding eliminates this discontinuity resulting in smoother signals but slightly decreases the magnitude of the reconstructed signal [10]. Different modifications of soft and hard thresholding functions have been done. For example, Sheikhzadeh & al [11] have proposed an exponential function to attenuate, in a non linear manner, the coefficients which are less than the threshold value in order to eliminate abrupt changes. Other thresholding functions could be selected such as $\mu - law$ function [5]:

$$S(x,\lambda) = \begin{cases} x \quad si \quad |x| > \lambda \\ \lambda \left(\dfrac{\left[(1+\mu)^{|x/\lambda|} - 1 \right]}{\mu} sign(x) \right) \quad si \quad |x| < \lambda \end{cases} \tag{4}$$

Equation (3) represents the expression of the modified hard thresholding with a parameter μ and the employed threshold λ. In the reference [5], μ is chosen to be equal to 255. There are different types of threshold, the constant or the variable ones. A global threshold was proposed in [7, 12]. It was defined as follow:

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

With N as the length of the noisy signal and $\hat{\sigma}$ as the estimate of the noise standard deviation, given by:

$$\hat{\sigma} = MAD / 0.6745 \tag{6}$$

where the MAD represents the absolute median estimated on the first scale.

The threshold given by (4) is generally used in denoising signals corrupted by a Gaussian white noise. In practical situations, one is frequently encountered with colored rather than white noises. To handle a colored noise, [3] proposes a level dependent threshold, defined by:

$$\lambda_j = \hat{\sigma}_j \sqrt{2\log N_j} \tag{7}$$

Where N_j represents the number of samples in scale j and $\hat{\sigma}_j$ is noise level at the scale j. Its expression is: $\hat{\sigma}_j = MAD_j / 0.6745$ with MAD_j represents the absolute median estimated on the scale j.

B. The discrete wavelet packet

The wavelet packets provide a great margin of signal analysis. In fact, if the wavelet analysis decomposes the approximation signals [13, 14], the wavelet packets permit to decompose the details too.

An approximation subspace V_j is divided to a lower resolution subspace V_{j+1} and a subspace of details, W_{j+1}. This is realized by decomposing the orthogonal basis, $\{\varphi_j(t - 2^j n)\}_{n\in Z}$ of V_j to two orthogonal bases: $\{\varphi_{j+1}(t-2^{j+1}n)\}_{n\in Z}$ of V_{j+1} and $\{\psi_{j+1}(t-2^{j+1}n)\}_{n\in Z}$ of W_{j+1}. Where φ_j and ψ_j are respectively the dilated or contracted versions of scale function φ and mother wavelet ψ. The decompositions of φ_{j+1} and ψ_{j+1} in the first base are determined by the pair of conjugate mirror filters $h(n)$ and $g(n)$. In the case of wavelet packets, this decomposition is applied to the detail subspace W_j. In this new representation, each node corresponds to a subspace W_j^p where j the depth is, and p is the number

of the nodes to the left of this particular node at the same depth and we have:
$W_j^p = W_{j+1}^{2p} \oplus W_{j+1}^{2p+1}$ (7) Where $\oplus$ is the direct sum of the two detail subspaces. The relations of splitting [15] define the wavelet packet orthogonal bases of the subspaces W_{j+1}^{2p} and W_{j+1}^{2p+1}. These relations are defined as follow:

$$\psi_{j+1}^{2p}(t) = \sum_n h(n)\psi_j^p(t - 2^j n) \quad (8)$$

$$\psi_{j+1}^{2p+1}(t) = \sum_n g(n)\psi_j^p(t - 2^j n) \quad (9)$$

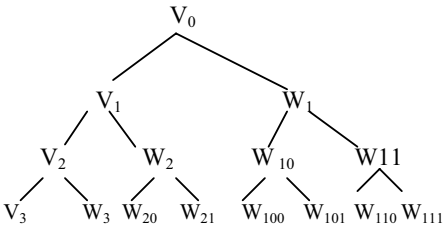


Fig. 2. A subspace binary tree of wavelet packet.

The authors in [5] propose to employ a threshold dependent of node for denoising speech signals corrupted by colored noise. This threshold is:

$$\lambda_{j,k} = \hat{\sigma}_{j,k}\sqrt{2\log N_{j,k}} \quad (10)$$

Where $\hat{\sigma}_{j,k}$ is the noise level, defined as follow:

$$\hat{\sigma}_{j,k} = MAD_{j,k} / 0.6745 \quad (11)$$

and $MAD_{j,k}$ represents the absolute median estimated at the scale j and subband k in the best wavelet packet tree. $N_{j,k}$ represents the length of the node (j,k). The use of $MAD_{j,k}$ is based on the fact that noise distribution is supposed to be Gaussian. In fact, this assumption is not always true in real situations. To solve this problem, [5] proposed to use spectral entropy using Histogram of intensity to estimate $\hat{\sigma}_{j,k}$.

III. The Used Technique

In this work, we use the discrete wavelet packet transform (DWPT) for speech signal enhancement. Instead of thresholding the wavelet packet nodes, a part of them are set to zero and the others are thresholded by employing a soft thresholding function. Setting coefficients to zero is based on the fact that the speech signal may be classified as voiced, unvoiced sounds [16] and silent segments. The voiced sounds are quasi-periodic in the time domain and harmonically structured. In frequency domain, these sounds are generally localized in bands that are less than 1 kHz. However, the energy of the unvoiced sounds is usually concentrated in high frequencies ($\geq 3kHz$). If we want to have a distinction between voiced and unvoiced sounds, we must benefit from pieces of information contained in these frequency bands where the voiced or unvoiced sound is dominant. It is well known that the power of the speech signal is contained around the frequency of the first formant. For many vowels of male and female voices, the statistic results indicate approximately that the frequency of the first formant doesn't exceed 1 kHz and is superior to 100Hz. In addition, the fundamental frequency of a normal voice is localized between 80 and 500Hz.

In this paper, the used speech signals are sampled at 16 kHz and the employed wavelet packet tree is shown by the Fig.3.

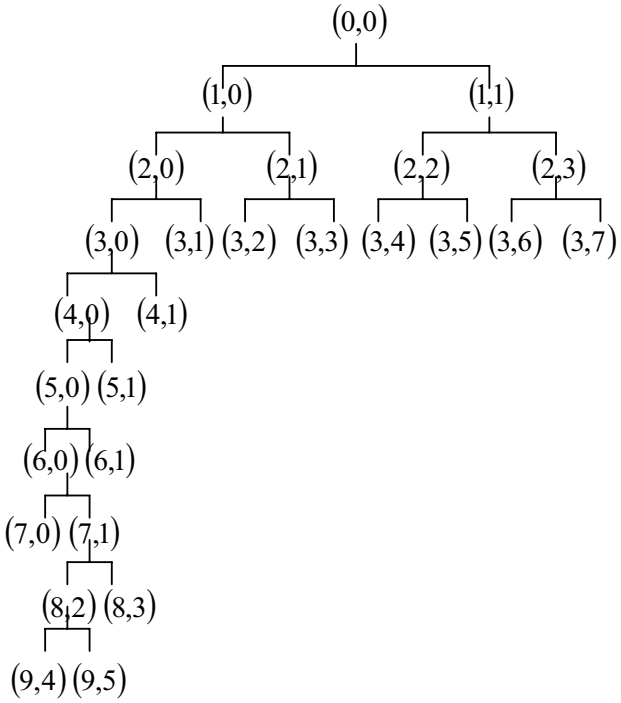


Fig. 3. The used wavelet packet tree.

The nodes $(7,0)$ and $(9,4)$ are set to zero because they occupy the subbands that are less than 80 kHz so they are considered to be the nodes associated to the noise. Since the node $(7,0)$ is considered to be the node associated to the noise, so it is used for estimating the noise level in node $(6,1)$. After thresholding the latter using a node dependent thresholding given by (eq.10) and the soft thresholding, we split it into the nodes $(8,2)$ and $(8,3)$. Then we split the node $(7,1)$ into $(9,4)$ and $(9,5)$. The same thing is done for the nodes $(9,4)$ and $(9,5)$: the node $(9,4)$ is also used for estimating the noise level in the node $(9,5)$ before setting the former to zero. The remained terminal nodes are also thresholded using the soft thresholding.

In practical situations, the noise doesn't have a Gaussian distribution, that's why the noise estimation in the remained nodes is based on spectral Entropy using histogram of intensity summarized as follow [5]:

Step 1. Estimate spectral pdf through histogram of wavelet packet coefficients for each node. The Histogram is composed of B bins.

Step 2. Compute the normalized spectral entropy:

$$ENT(n) = -\sum_{b=1}^{B} P \cdot \log_B(P) \quad (12)$$

where $n = 1, 2, ..., N_0$, with N_0 as the number of the best nodes. The probability P was defined as:

$$P = \frac{N_0 \text{ . of wavelet packet coefficients in bin } b}{\text{Node size in adapted wavelet packet tree}}$$

Step 3. Estimate the spectral magnitude intensity by histogram and the standard deviation of noise for node dependent wavelet thresholding.

A. Experimental evaluations

In this paper, an Arabic speech database was disposed. It contains twenty sentences sampled at 16 kHz, ten pronounced by a male voice and the others by a female's. These sentences are corrupted by a white noise and a Volvo car noise separately. The SNR, before enhancement, named RSBi, varies from -5 to 15dB.

TABLE I FEMALE VOICE CORRUPTED BY VOLVO NOISE.

RSBi	DWPT based denoising technique by using a $\mu - law$ thresholding function	DWPT based denoising technique by using spectral entropy	Spectral subtraction based denoising technique
-5	0.50	1.76	7.99
0	5.27	6.00	10.87
5	9.52	10.69	12.84
10	12.75	13.55	14.81
15	14.43	17.97	15.97

TABLE II MALE VOICE CORRUPTED BY VOLVO NOISE

RSBi	DWPT based denoising technique by using a $\mu - law$ thresholding function	DWPT based denoising technique by using spectral entropy	Spectral subtraction based denoising technique
-5	0.36	1.92	4.45
0	4.83	5.67	6.82
5	8.44	9.41	10.35
10	10.67	12.04	14.37
15	11.69	13.36	16.72

a. SNR after enhancement

Tables 1to 4 report the obtained values of the SNR after enhancement. The obtained results are given in these tables.

TABLE III MALE VOICE CORRUPTED BY WHITE NOISE.

RSBi	DWPT based denoising technique by using a $\mu - law$ thresholding function	DWPT based denoising technique by using spectral entropy	Spectral subtraction based denoising technique
-5	-2.46	-0.27	6.96
0	2.37	3.08	10.04
5	6.60	8.11	13.15
10	9.64	11.47	15.51
15	11.29	13.36	17.14

TABLE IV FEMALE VOICE CORRUPTED BY WHITE NOISE.

RSBi	DWPT based denoising technique by using a $\mu - law$ thresholding function	DWPT based denoising technique by using spectral entropy	Spectral subtraction based denoising technique
-5	-2.37	-1.17	8.76
0	2.59	3.20	11.96
5	7.21	8.95	14.15
10	11.19	13.18	15.34
15	13.56	18.24	16.42

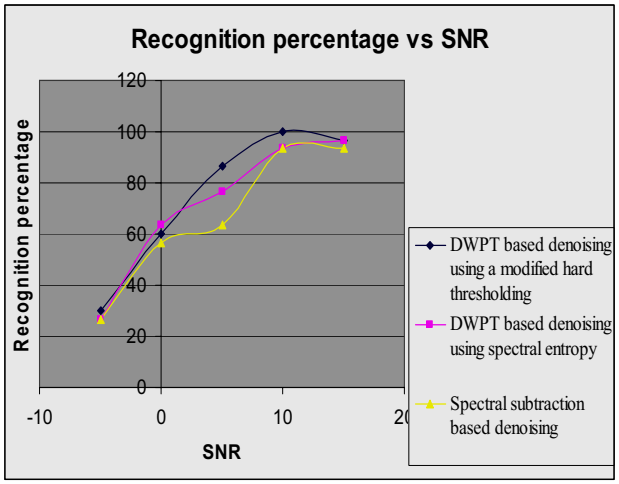
The four tables show that the three denoising techniques improve the signal to noise ratio. The results show also that the spectral subtraction based method is better than the two thresholding based denoising methods. We also remark that results obtained by applying the DWPT based denoising technique using the spectral entropy are better than those obtained from the DWPT based denoising method using $\mu - law$ thresholding function.

b. Listening tests

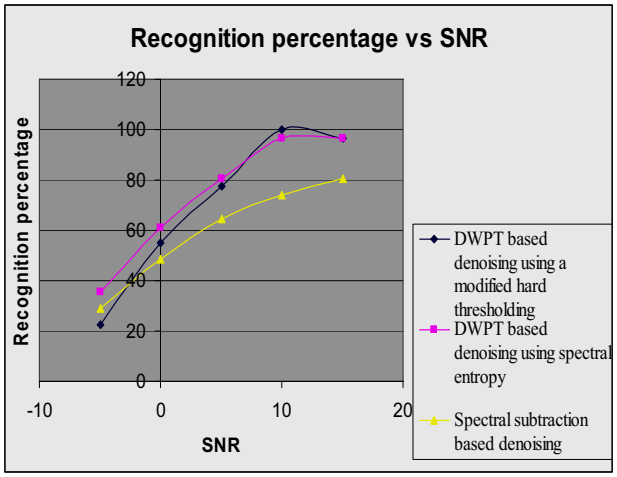
By making the listening tests, we compute the percentage of the recognized words by using the following formula:

$$\frac{\text{number of the recognized words}}{\text{total number of the employed words}} \times 100$$

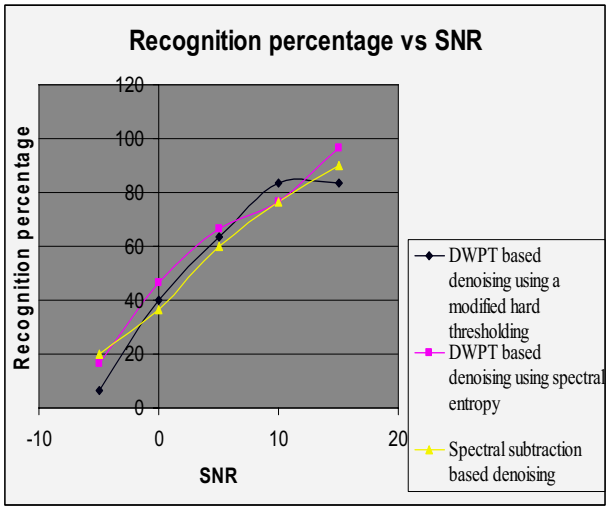
Fig, 2-a, 2-b and 2-c illustrate the recognized words percentage vs SNR for the three techniques.



(a) Case of female voice corrupted by car noise.



(b) Case of male voice corrupted by car noise.



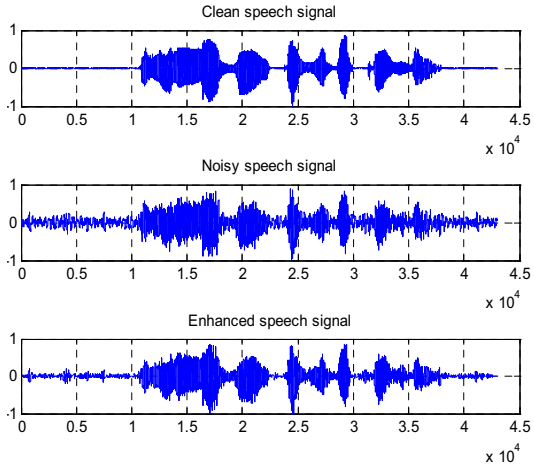
(c) Case of male voice corrupted by White noise.

Fig, 2. (a), (b), (c) Words recognition percentage vs SNR.

The obtained curves from listening tests show that the percentage of recognized words decreases significantly when the SNR varies from 15dB to –5dB. These curves show also that the obtained values of the percentage from the denoising technique based on DWPT using spectral entropy are better than those obtained from spectral subtraction. Thus, when speaking of intelligibility, the DWPT based denoising technique using spectral entropy is better than the spectral subtraction based one. The curves also show that when SNR is lower, the denoising technique using spectral entropy is better than that using $\mu - law$ thresholding function, and we have the opposite when SNR is higher.

c. Speech signal representation:

The figures 3, 4 and 5 are examples of speech denoising by using the DWPT based denoising technique using spectral entropy, and figures 6and 7are examples of speech enhancement by employing the DWPT based denoising method using $\mu - law$ thresholding function. The figure 8 is an example of speech denoising method based on spectral subtraction.



Fig, 3. Female voice corrupted by Volvo noise (SNR=5dB).

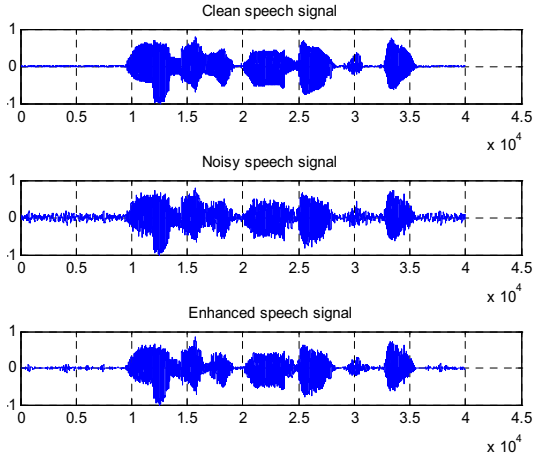
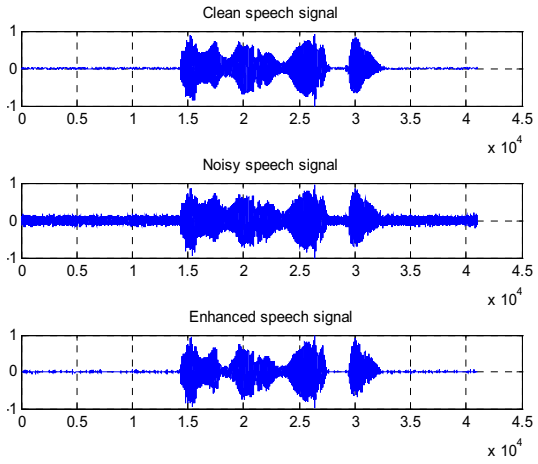


Fig. 4. Male voice corrupted by Volvo noise (SNR=10dB).



Fig, 5. Male voice corrupted by White noise (SNR=10dB).

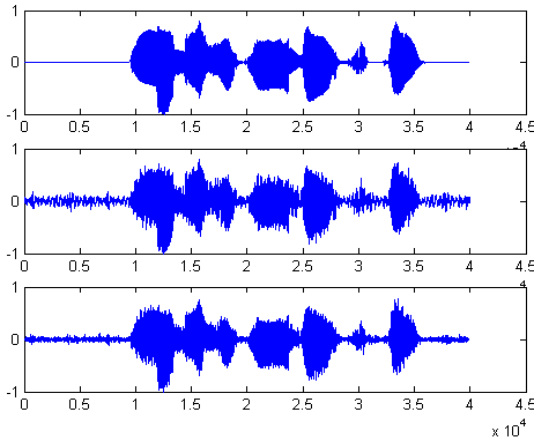
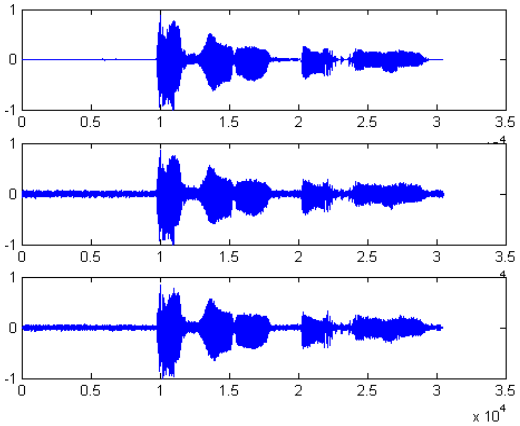
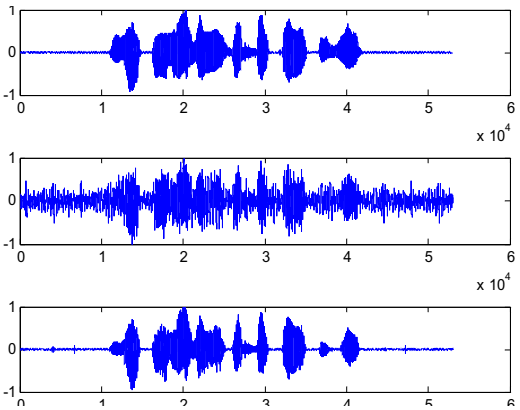


Fig. 6. Male voice corrupted by Volvo noise (SNR=10dB).



Fig, 7. Female voice corrupted by White noise (SNR=10dB).



Fig, 8. male voice corrupted by Volvo noise (SNR=0dB).

These figures illustrate the time representations of the clean, the noisy and the enhanced speech signals. Figures 3 to 7 show the superiority of the denoising technique based on DWPT using spectral entropy when compared with the denoising technique based on DWPT using the modified hard thresholding function based on $\mu - law$ algorithm. Our proposed technique is very efficient in speech denoising. In fact a more important amount of noise was suppressed from non speech and speech segments while preserving the majority of the speech signal. We remark that there is a little difference between the enhanced speech signal and the clean one. Thus there isn't any sever distortion of the denoised signal, when using our proposed technique. The figure number 8 shows that spectral subtraction reduces a great part of noise but introduces distortions on the enhanced speech signal.

The figures 9, 10, and 11 represent respectively the spectrograms of clean speech, noisy speech and enhanced speech obtained by applying our proposed denoising technique. The noise which corrupting the speech signal, is The F16 cockipt with the SNR is equals to 5dB.

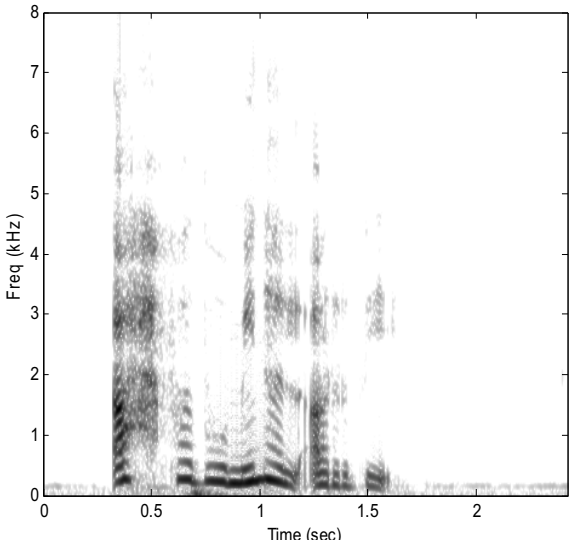


Fig. 9. Spectrogram of the clean speech.

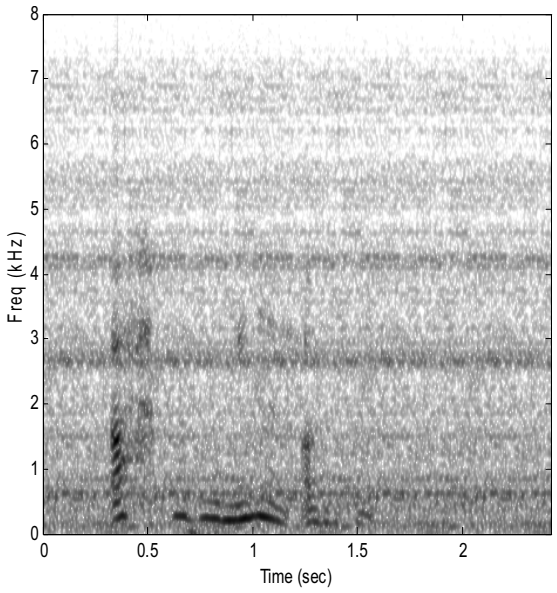


Fig. 10. Spectrogram of the noisy speech.

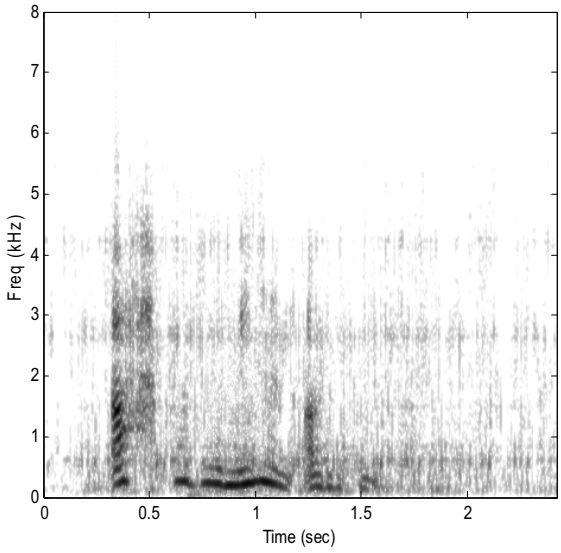


Fig. 11. Spectrogram of the enhanced speech.

This figures show a certain similarity between the clean speech spectrogram and the enhanced speech spectrogram and there is a big difference between the latter and the spectrogram of the noisy speech signal.

IV. Conclusion

In this study, a denoising method is developed under Matlab and compared with conventional thresholding technique using $\mu - law$ thresholding function and spectral subtraction based denoising method. It employs the discrete wavelet packet transform and the spectral entropy to estimate the noise level. The results obtained from the computation of the signal to noise ratio and listening tests show that the two techniques based on DWPT and spectral subtraction based denoising one ameliorate the SNR. When speaking about intelligibility of the enhanced speech signal, the two DWPT based denoising techniques are better than the spectral subtraction based denoising technique. The results obtained from listening tests show also that when SNR is lower the denoising technique using spectral entropy is better than that using $\mu - law$ thresholding function, and we have the opposite when SNR is higher.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," IEEE Trans. On Acoust. Speech Signal Processing, vol. 32, no. 6, pp. 1109-1121, 1984.
[2] D.L. Donoho, "Denoising by soft thresholding," IEEE Trans on Information Theory, vol. 41, no. 3, pp. 613-627, 1995.
[3] I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise", J. Roy. Statist. Soc. B, vol. 59, pp. 319-351, 1997.
[4] X. Huang, A. Acero, H. Hon, "Spoken Language Processing," Prentice Hall, p. 474, 2001.
[5] Sungwook Chang, Y. Kwon, Sung-il Yang, and I-jae Kim. "Speech enhancement for non-stationary noise environment by adaptive wavelet packet" IEEE Tans. pp. 561-564, 2000.
[6] S.S. Chen, Basis Pursuit, Phd Thesis, Standford University, November 1995.
[7] D. Donoho and I. M. Johnstone. ''Ideal spatial Adaptation via Wavelet Shrinkage'' Biometrika, 41. pp. 425-455, 1994.
[8] Waleed H. Abdulla. "HMM-based techniques for speech segments extraction". ISSN 1058-9244/02/S8.00 © 2002-IOS Press.
[9] Mohammed BAHOURA and Jean ROUAT "Wavelet noise reduction: application to speech enhancement". CiteSeer, 2000.
[10] [V. Balakrishnan, Nash Borges and Luke Parchment. "Wavelet denoising and speech enhancement," Spring 2006.
[11] H. Sheikhzadeh and H. Reza Abutalebi. ''An improved wavelet-based speech enhancement system''. Eurospeech, 2001.
[12] D. Donoho, I. M. Johnstone, G. Kerkyacharian et D.Picard. "Wavelet Shrinkage: Asymptotia" Journal of the Royel Statistical Society, Serie B,57, pp. 3019-3069, 1995.
[13] Jong Won Seok and Keun Sung Bae. "Speech enhancement with reduction of noise components in the wavelet domain," 0-8186-7919-0/97 S10.00 © 1997 IEEE, pp. 1323-1326.
[14] Pham Van Tuan and Gernot Kubin ."DWT-Based classification of Acoustic-Phonetic Classes and Phonetic Units," International Conference on Spoken Language Processing (Interspeech-ICSLP) -2004.
[15] S. Mallat, A wavelet tour of signal processing. Academic Press, San Diego, USA (1998).
[16] E. Jafer, A.E.Mahdi, "Wavelet-based Voiced/Unvoiced classification algorithm", 4th EURASIP Conf., Vol.2, pp.667-672, Croatia,2003.

Talbi Mourad obtained his engineering degree in 1995 and master at 2003 from the Engineering Faculty of Tunis in automatic and signal processing. He is a researcher member of the Signal Processing Laboratory and preparing his doctorate thesis.

Lotfi Salhi was born the 05-16-1974 at Thala, (center west of Tunisia). He obtained the Bac exp sciences in 1993. He received his bachelor in physics in 1999 from the sciences faculty of Sfax. He has been recruited in computer engineer post in a Tuniso- American company for two years. He received the diploma of Master degree in automatic and signal processing (ATS) in 2004 from the national school of engineers of Tunis (ENIT) with concentration in digital speech signal processing (cochlear filter).

Cherif Adnene Cherif Adnene obtained his engineering diploma in 1988 from the Engineering Faculty of Tunis, and his Ph.D. in electrical engineering and electronics in 1997. Actually he is a professor at the Science Faculty of Tunis, responsible for the Signal Processing Laboratory. He participated in several research and cooperation projects, and is the author of more than 60 international communications and publications.