THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.22, No.2E 2003. 6. pp. 87~92

Speech Enhancement Using Level Adapted Wavelet Packet with Adaptive Noise Estimation

Sungwook Chang*, Younghun Kwon**, Sung-Il Jung*, Sung-Il Yang*, Kunsang Lee**
*School of Electrical and Computer Engineering, Hanyang University
**Department of Physics, Hanyang University
(Received March 6 2003; accepted July 24 2003)

Abstract

In this paper, a new speech enhancement method using level adapted wavelet packet is presented. First, we propose a level adapted wavelet packet to alleviate a drawback of the conventional node adapted one in noisy environment. Next, we suggest an adaptive noise estimation method at each node on level adapted wavelet packet tree. Then, for more accurate noise component subtraction, we propose a new estimation method of spectral subtraction weight. Finally, we present a modified spectral subtraction method. The proposed method is evaluated on various noise conditions: speech babble noise, F-16 cockpit noise, factory noise, pink noise, and Volvo car interior noise. For an objective evaluation, the SNR test was performed. Also, spectrogram test and a very simple listening test as a subjective evaluation were performed.

Keywords: *Level adapted wavelet packet, Spectral subtraction weight, Adaptive noise estimation*

I. Introduction

The environmental robustness of practical speech processing applications is definitely very important. To enhance the robustness, various approaches have been proposed. One of the approaches is adaptive wavelet packet based method. However, an additive noise makes a serious interference with a proper adaptation of the conventional adapted node based wavelet packet. To alleviate the drawback of the conventional adapted node based wavelet packet, we suggest a level adapted wavelet packet (LAWP). An adaptation for spectral bandwidth of speech can be achieved with the proposed LAWP at each frame.

Unfortunately, a great deal of real environmental noise is non-stationary. Even the noises generated by a computer fan, an air conditioner, or an automobile engine are not perfectly stationary[1]. Thus, the recent majority of the speech enhancement methods are focused on the accurate estimation of non-stationary or colored noise[2,3]. To resolve the problem about the non-stationary characteristic of environmental noise, we propose an adaptive noise estimation method using probability distribution function (pdf) at each node on wavelet packet tree. The method is very useful to chase the time varying characteristic of the non-stationary noise.

Finally, we present a modified version of the conventional spectral subtraction method[4] with the proposed noise estimation method. In the modified spectral subtraction method, a new adaptive spectral subtraction weight is used. The adaptive spectral subtraction weight provides a time varying quantity of noise at each frame.

Corresponding author: Sungwook Chang (schang@ihanyang.ac.kr)
School of Electrical and Computer Engineering, Hanyang University
17, Haengdang-dong, Seongdong-gu, Seoul, 133-791, Korea

II. Level Adapted Wavelet Packet

Adaptive wavelet packet[5] divides the time-frequency plane into elementary atoms that are best adapted to approximate a particular signal. This property is very useful to analyze a time varying characteristics of speech. Thus, adaptive wavelet packet (node adapted one) based speech enhancement is one of the widely used methods [2,3]. However, the conventional node adapted wavelet packet has a crucial drawback to analyze a noisy speech. That is, a background noise makes serious interference with a proper adaptation of the conventional node adapted wavelet packet. Moreover, if the background noise has a non-stationary characteristic, it is almost impossible to carry a proper adaptation out. For that reason, we propose an entropy based level adapted wavelet packet (LAWP) to alleviate the drawback of the conventional node adapted one in colored or non-stationary noise environment.

At first, we define an entropy for each node on the wavelet packet tree,

$$H_l(n) = \sum_{i=0}^{NS(l)-1} p(i)\log\frac{1}{p(i)} \tag{1}$$

$$p(i) = |x(i)|^2 / \|x\|^2 \tag{2}$$

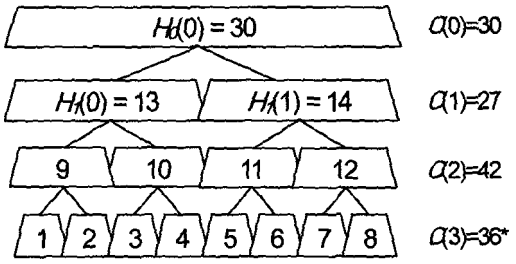
where $x(i)$ is the ith wavelet packet coefficient at a node on wavelet packet tree, $NS(l)$ is the size of corresponding node at lth level on the tree, and $p(i)$ may be interpreted as a probability distribution function (pdf) for the sample space N(natural number space)[5]. Finally, $H_l(n)$ is the entropy at the nth node in the lth level. The $H_l(n)$ is the same measure to perform the node adaptation in the conventional adaptive wavelet packet.

Next, we define a level information cost (LIC) using the entropy $H_l(n)$,

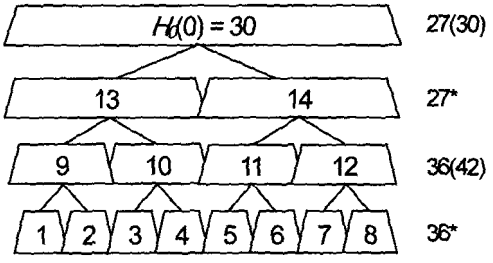
$$C(l) = \sum_{n=0}^{NN(l)-1} H_l(n) \tag{3}$$

where $NN(l)$ is the number of nodes of wavelet packet tree and $C(l)$ denotes an information cost at lth level, respectively. The LIC may be represents a rough spectral bandwidth of speech. The proposed LIC is not optimal measure but asymptotic one to pursue the best wavelet basis for a speech. Nevertheless, the LIC can be more reliable measure than the entropy $H_l(n)$ which is used for node adaptation in non-stationary noisy environment or low SNR case. It is because the LIC can be less influenced by the interference of noise at the expense of the optimality.

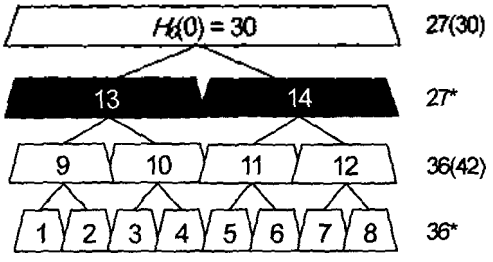
Finally, we propose a level adaptation procedure. It is very similar to the conventional node adaptation procedure proposed by Wickerhauser. We use the bottom-up procedure [5] with minimal level information cost for each level in wavelet packet library tree to pursue the asymptotic best basis. To illustrate the level adaptation procedure, consider the following example expansion into a three level wavelet packet library tree as shown in Fig. 1. We have placed numbers representing node information costs $H_l(n)$ at nth node in lth level inside the nodes of the tree. And we find



(a) Step 1. Mark bottom level



(b) Step 2. Mark all levels of lower cost



Step 3. Retain topmost marked level

Figure 1. The bottom-up procedure for level adaptation.

the LIC C(l) by Eq. (3) for each level as shown in the right column of Fig. 1. Now, we start by marking bottom level on the wavelet packet tree, as indicated by the asterisks in Fig. 1(a). The marked LIC is an initial value which we will try to reduce. Whenever an upper level has lower LIC than the current level, we mark the upper level with an asterisk. If the current level has lower LIC, we do not mark the upper level, but we assign the lower LIC of the current level to the upper one as shown in Fig. 1(b). Finally, after all the levels have been examined, we take the topmost marked level. The best level displayed as shaded blocks in Fig. 1(c).

III. Adaptive Noise Estimation Method

Most of the noise estimation methods are performed by detection of speech pauses to evaluate segments of pure noise. However, in practical environments this is a difficult task, especially if the background noise is not stationary or the SNR is low. To solve the problem, Hirsch et. al. proposed a noise estimation method with a simple first order recursive system[6],

$$\left|\hat{N}_k(n)\right| = \alpha \cdot \left|\hat{N}_{k-1}(n)\right| + (1-\alpha)\cdot\left|X_k(n)\right| \tag{4}$$

where $|X_k(n)|$ denotes the spectral magnitude and $|\hat{N}_k(n)|$ is an estimation for the noise magnitude at subband n in kth frame. However, the method is not yet enough to represent the non-stationary characteristics of noise. It is because the time-varying characteristic of spectral distribution of non-stationary noise is not considered. That is, the weighting parameter α is fixed over all the nodes (subbands) and all the frames. Fortunately, the entropy based adaptive wavelet packet provides a pdf for each node (subband). The pdf sequence at each node shows a degree of energy compaction. Also, it represents the time-varying characteristic of spectral distribution over all nodes at each frame. Thus, the proposed adaptive noise estimation method uses the pdf sequence with the LWAP to alleviate the drawback of the Hirsch et. al.'s method.

For the purpose, we find a geometric mean $\gamma_k(n)$ of the ratio between the ith pdf at the nth node in kth frame and the one at the corresponding node in the (k-1)th frame,

$$\gamma_k(n) = \left(\prod_{i=0}^{NS-1}\left(p_k(i)/p_{k-1}(i)\right)\right)^{\frac{1}{NS}} \tag{5}$$

where $p_k(i)$ is the ith pdf at current node (nth node for $\gamma_k(n)$) in the kth frame and $p_{k-1}(i)$ is the one at the corresponding node in the (k-1)th frame. NS is the node size at the adapted level. Here, every node size is equal in the adapted level since all subbands in the LAWP are uniformly distributed. And $\gamma_k(n)$ is normalized by the sum of $\gamma_k(n)$s for all the nodes in the adapted level. Thus, the $\gamma_k(n)$ may be interpreted as a geometric pdf variation over the current and previous frame.

Finally, we define an adaptive noise estimation method using the geometric mean of pdf ratio,

$$\left|\hat{N}_k(n)\right| = (1-\gamma_k(n))\cdot\left|\hat{N}_{k-1}(n)\right| + \gamma_k(n)\cdot\left|X_k(n)\right| \tag{6}$$

where $|\hat{N}_k(n)|$ is the spectrum magnitude of estimated noise at the nth node in kth frame. And $|X_k(n)|$ denotes the average spectrum magnitude of LAWP coefficients at nth node in kth frame. If $|X_k(n)| \le \beta\cdot|\hat{N}_{k-1}(n)|$ with $1.2 \le \beta \le 1.5$, noise estimation is performed by Eq. (6). Otherwise, the estimated noise information in the previous frame is used for the one in current frame.

IV. Spectral Subtraction

Berouti et al. proposed a spectral subtraction method [4] in which the amount of noise subtraction depends on the SNR of the particular frame. Unfortunately, we have one more obstacle that a great deal of real environmental noise is non-stationary. Thus, a different spectral subtraction weight is required to represent the time varying statistical information.

For that reason, we propose a new estimation method of spectral subtraction weight for each node in a frame.

4.1. Spectral Subtraction Weight

We first define a log scaled geometric mean of average spectrum magnitude of noisy speech (NSGM) and one of an estimated noise spectrum magnitude (NGM) in each kth frame,

$$NSGM(k) = \log\left(\prod_{n=0}^{NN-1}|X_k(n)|\right)^{\frac{1}{NN}} \tag{7}$$

$$NGM(k) = \log\left(\prod_{n=0}^{NN-1}|\hat{N}_k(n)|\right)^{\frac{1}{NN}} \tag{8}$$

where NN is the number of node on the adapted level of LAWP. Next, we define a geometric signal to noise ratio (GSNR) using NSGM and NGM for all frames,

$$GSNR = \frac{\sum_{k=0}^{NF-1} NSGM(k)}{\sum_{k=0}^{NF-1} NGM(k)} \tag{9}$$

where NF is the number of frame. Finally, we suggest a spectral subtraction weight for each frame,

$$\xi(k) = \left(\frac{NSGM_{\max} - NSGM(k)}{NSGM_{\max} - NSGM_{\min}}\right)\cdot\left(\frac{GSNR}{\rho}\right) \tag{10}$$

where $NSGM_{\max}$ and $NSGM_{\min}$ is the maximum and the minimum $NSGM$ for all frames. And ρ is a level controller with $2.0 \le \rho \le 3.0$.

Fig. 2 shows that the proposed spectral subtraction weight is proportional to the envelope of speech. Especially, the proposed spectral subtraction weight of 10 dB noisy speech and one of 5 dB have a very similar form. From the result, we can see that the proposed spectral subtraction weight is very robust even though the background noise is non-stationary or the SNR is low.

4.2. Modified Spectral Subtraction with the Proposed Weight

In this section, we propose a modified spectral subtraction method with the proposed LAWP, adaptive noise estimation method, and spectral subtraction weight.

At First, we define spectral subtraction gain $G_k(n)$ with

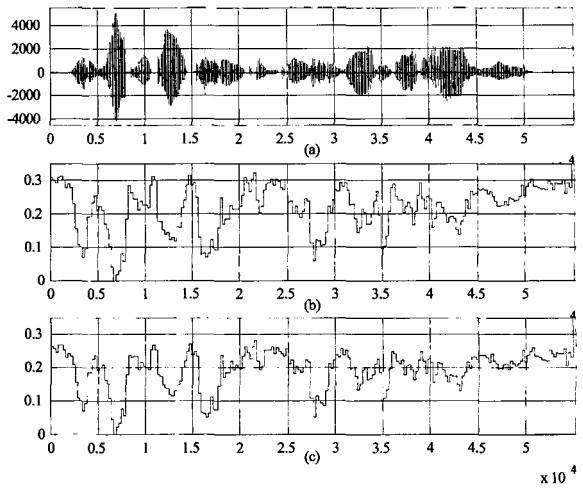


Figure 2. (a) Clean Speech ("She had your dark suit in greasy wash water all year") (b) The spectral subtraction weight of speech degraded by F-16 noise (SNR=10 dB) (c) The spectral subtraction weight of speech degraded by F-16 noise (SNR=5 dB).

the proposed methods at nth node in kth frame,

$$G_k(n) = \begin{cases} \sqrt{1 - \dfrac{(1+\xi(k))\cdot|\hat{N}_k(n)|}{|X_k(n)|}} & ,|X_k(n)| \ge (1+\xi(k))\cdot|\hat{N}_k(n)| \\ \eta\cdot\sqrt{\dfrac{|\hat{N}_k(n)|}{|X_k(n)|}} & ,otherwise \end{cases} \tag{11}$$

where $0 \le \eta \le 0.1$.

Next, we perform the modified spectral subtraction with the proposed spectral subtraction gain,

$$\hat{x}_{k,n}(i) = x_{k,n}(i)\cdot G_k(n) \tag{12}$$

which $\begin{cases} 0 \le k < No.\ of\ frame \\ 0 \le n < No.\ of\ node\ in\ ALWP \\ 0 \le i < Node\ size\ in\ ALWP \end{cases}$

$x_{k,n}(i)$denotes LAWP coefficient of noisy speech at ith coefficient in nth node and kth frame. And $\hat{x}_{k,n}(i)$ denotes the estimated LAWP coefficient.

V. Evaluation

Various noise types, from Noisex-92 database are used

Table 1. Average enhanced SNR (dB) for various noise conditions.

	Babble	F-16	Factory	Pink	Volvo
20 dB	21.06	20.57	20.87	20.95	21.37
15 dB	16.73	16.44	16.48	16.74	17.66
10 dB	12.49	12.59	12.23	12.64	14.13
5 dB	8.48	9.04	8.20	8.74	10.59
0 dB	4.70	5.73	4.43	5.29	6.99
-5 dB	0.90	2.60	0.82	2.24	2.88

n our evaluation: speech babble noise, F-16 cockpit noise, factory noise, pink noise, and Volvo car interior noise. The performance results are averaged out using 100 different utterances from the TIMIT database. Half of the utterances are taken from male speakers, and the others from female speakers.

In colored or non-stationary noise environment, the SNR cannot be used as faithful indication of speech quality. Thus we employ both objective and subjective tests for evaluation of the proposed method. In objective tests, we take enhanced SNR test in various noise conditions. Table. 1 shows well enhanced SNR performance for all noise conditions. Especially, we can see that the proposed speech enhancement method shows better performance in low SNR.

For the subjective test, we have examined the spectrogram. Particularly, the proposed method yields a good performance even in the severely corrupted spectral band. Fig. 3 shows speech spectrogram example obtained by the proposed method. In this figure, we can see that most of noise component is removed over all spectral bands. Additionally, we perform a simple listening test. Most of the listener has almost not recognized the background noise and any musical artifact from the enhanced speech up to SNR 15 dB over the various noise conditions. In the case the SNR is 10 dB, listeners recognized a little remainders of background noise. And the enhanced speeches of 5 dB, 0 dB, and -5 dB have annoyed listeners with the background noise or musical artifact.

Recently, we have reported the performance of conventional denoising methods using wavelets with various wavelet threshold estimation methods, wavelet filters, and wavelet structures (adaptive wavelet packet, discrete wavelet transform) for speech signals corrupted with colored or non-stationary noise[7]. In comparison of the conventional denoising methods with the proposed method in Table. 1, the proposed one shows considerably better results. It is because the conventional denoising methods have no consideration for time-varying characteristics of speech and non-stationary characteristic of environmental noise.

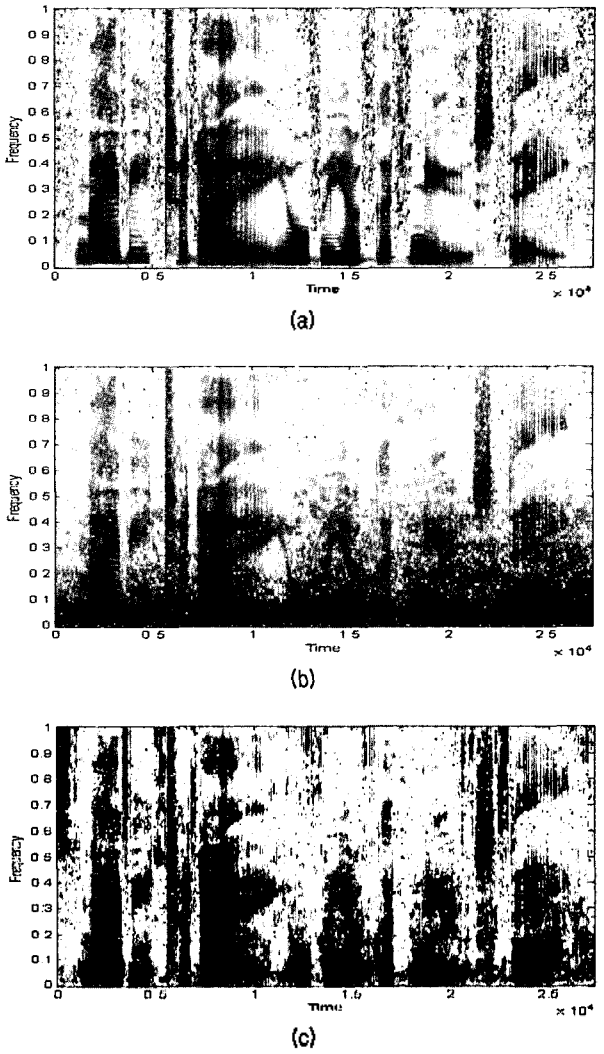


Figure 3. Spectrogram. (a) Clean speech: Maybe today'll be a good-news day. (b) Noisy speech (additive speech babble noise at a SNR=10 dB) (c) Speech enhanced by the proposed methods.

VI. Conclusions

Both objective and subjective test show good enhance-

ment performance. Especially, we note that the proposed method gives a very good quality of speech up to the SNR 10 dB. Additionally, spectrogram test shows good performance even though the SNR is low in non-stationary noise environment.

Acknowledgement

This work is supported by the faculty research fund of Hanyang University.

References

1. X. Huang, A. Acero, and H. Hon, Spoken Language Processing, Prentice Hall, 474, 2001.
2. I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. Roy. Statist. Soc. B*, 59, 319-351, 1997.
3. S.-W. Chang, Y.-H. Kwon, and S.-I. Yang, "Speech enhancement for non-stationary noise environment by adaptive wavelet packet," *ICASSP 2002*, 561-564, 2002.
4. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *ICASSP-79*, 208-211, 1979.
5. M. V. Wickerhauser, "Adapted wavelet analysis from theory to software," *A K Peters*, 1994.
6. H. G. Hirsch, and C. Ehricher, "Noise estimation techniques for robust speech recognition," *ICASSP-95*, 153-156, 1995.
7. C.-H. Oh, I.-J. Kim, S.-W. Chang, Y.-H. Kwon, and S.-I. Yang, "A study on denoising methods using wavelets for a speech with color noise," *Proceedings of ASK Regular Conference*, 21 (2(s)), 351-356, 2002.

[Profile]

• Sungwook Chang

Sungwook Chang was born in Anyang, Kyunggi, Korea in 1971. He received the B. S. and M. S. degrees in Control and Instrumentation Engineering from Hanyang University, Korea in 1997 and 1999, respectively. Currently, he is graduate student for Ph. D. degree in Electronic, Electrical, Control and Instrumentation Engineering at Hanyang University. His current research interests include speech recognition, speech enhancement, wavelets, statistical natural language processing, and bio bioinformatics. He is also a member of the Acoustical Society of Korea.

• Younghun Kwon

Younghun Kwon was born in Seoul, Korea in 1961. He received his B.S. degree in Mathematics and Physics with the greatest honors from Hanyang University, Seoul, Korea, 1984, and his M.S. and Ph. D. degrees in Physics from the University of Rochester, Rochester, New York, 1986 and 1987, respectively. Since 1995, he has been with Hanyang University and he is now an Associate Professor at Department of Physics. His current research interests include Mathematical Physics, Theoretical Physics, Artificial Intelligence, Signal Processing and Quantum computing. He is also a fellow of the International Society for Complexity, Information, and Design, and member of American Mathematical Society, Korean Mathematical Society and Korean Physical Society.

• Sung-Il Jung

Sungil Jung was born in Busan, Korea in 1972. He received the B. S. and M. S. degrees in Computer Engineering from Korea Maritime University, Korea in 2000 and 2002, respectively. Currently, he is graduate student for Ph. D. degree in Electronic, Electrical, Control and Instrumentation Engineering at Hanyang University. His current research interests include speech enhancement, speech recognition, and wavelets. He is also a member of the Acoustical Society of Korea.

• Sung-Il Yang

Sung-il Yang was born in Geosan, Chungbuk, Korea in 1956. He received his B. S. degree in Electronics Engineering with the greatest honors from Hanyang University, Seoul, Korea, 1984, and his M. S. and Ph. D. degrees in Electrical & Computer Engineering from the University of Texas, Austin, Texas, 1986 and 1989, respectively. Since 1990, he has been with Hanyang University and he is now a Professor at the School of Electrical & Computer Engineering. His current research interests include speech recognition, digital signal processing, and responsible technology. He is also a member IEEE, Korea Institute of Telematics and Electronics, and the Acoustical Society of Korea.

• Kunsang Lee

Kunsang Lee was born in Seoul, Korea in 1945. He received B. S. (1969), M. S. (1977) and Ph. D. (1985) degrees in Physics at Hanyang University at Seoul. He had taken parts in research programs of Danish Technological Institute (Denmark) and University of California at Irvine (U. S.). He was also a faculty member at National Central Polytechnic College. Since 1981, he has been with Hanyang University and is now a tenured Professor at Department of Physics. His current research interests include various topics of Applied Physics, e.g. Acoustics and Speech Recognition. He is also associated in American Physical Society, American Optical Society, Korean Physical Society, and Korean Acoustic Society.