Enhancement of Speech Using Bark-Scaled Wavelet Packet Decomposition

Israel Cohen

Lamar Signal Processing Ltd.
P.O. Box 573, Yokneam Ilit 20692, Israel
icohen@lamar.co.il; http://www.AndreaElectronics.com

Abstract

In this paper, we propose a speech enhancement system, which integrates a bark-scaled wavelet packet decomposition (BS-WPD), a soft-decision gain modification and a "magnitude" decision-directed estimation technique. The BS-WPD provides an overcomplete auditory representation, having a higher frequency resolution than the critical band decomposition. Speech is estimated by Wiener filtering in the wavelet packet domain, modified by the signal presence probability. We introduce a "magnitude" decision-directed estimator for the variance of speech, which is closely related to the decision-directed estimator of Ephraim and Malah. This estimator achieves, in the established process, a better tradeoff between noise reduction and signal distortion. The proposed enhancement algorithm is tested with various noise types, and compared to a conventional log-spectral amplitude estimator. We show that noise can be further suppressed, while preserving its natural structure and the intelligibility and quality of the speech components.

1. Introduction

Speech enhancement systems must inevitably take into account aspects of human perception. Of significant importance is the particular spectral domain, in which processing of the noisy signal is carried out. Classical enhancement systems are based on uniformly spaced frequency resolutions. However, a nonuniform frequency resolution, which reflects the human auditory system, may lead to improved intelligibility and quality of the enhanced signal, when combined with an appropriate spectral gain function.

Some methods have been recently utilized critical band analysis in subtractive-type enhancement techniques [1]–[5]. Unfortunately, either the subtraction is still implemented in the short-term Fourier domain [1, 2], or the spectral resolution may be insufficient for dealing with voiced sounds [3, 6, 7]. Furthermore, subtractive-type techniques introduce a high level of musical residual noise. This entails large overestimation factor and noise-floor, which limit the overall performance of the speech enhancement system [1, 3, 4].

In this paper, we propose an enhancement method, which integrates a *bark-scaled wavelet packet decomposition* (BS-WPD), a soft-decision gain modification and a "magnitude" decision-directed estimation technique. The BS-WPD is characterized by a higher frequency resolution than the critical band decomposition, and a higher time resolution than the conventional WPD. The distance between the center frequency of one subband to the center frequency of the next subband is 1/4 Bark. The increased time resolution is obtained by including non-decimated filtering in the expansion tree. Taking into account speech presence uncertainty, the clean signal is estimated by modified Wiener filtering in the wavelet packet domain. We introduce a "magnitude" decision-directed estimator for the variance of speech, which is closely related to the decision-directed estimator of Ephraim and Malah [8]. This estimator achieves, in the present procedure, a better tradeoff between noise reduction and signal distortion. The proposed enhancement algorithm is tested with various noise types, and compared to the popular *log-spectral amplitude* (LSA) estimator [9]. We show that further noise suppression can be obtained, while preserving its natural structure and the intelligibility and quality of speech components.

2. The Bark-Scaled WPD

Let $\{\psi_n(t) : n \in \mathbb{Z}_+\}$ denote a wavelet packet family, and let $E \subset \{(\ell, n) : 0 \le \ell < L, 0 \le n < 2^\ell\}$ represent the terminal nodes of a WPD tree [10]. Then disjoint covers of $[0, 1)$ by dyadic intervals of the form $I_{\ell,n} = [2^{-\ell}n, 2^{-\ell}(n+1))$ correspond to specific wavelet packet expansions (specific sets of terminal nodes E). In particular,

$$\{\psi_{\ell,n,k} : (\ell, n) \in E,\, k \in \mathbb{Z}\}$$

where $\psi_{\ell,n,k}(t) \triangleq 2^{-\ell/2}\psi_n\left(2^{-\ell}t - k\right)$, form a basis for the signal space $\overline{span}\{\psi_0(t-k) : k \in \mathbb{Z}\}$.

A terminal node $(\ell, n) \in E$ is associated with a subband (subspace) whose center frequency and bandwidth are roughly given by [10]

$$f_{\ell,n} = 2^{-\ell}\left[GC^{-1}(n) + 0.5\right] \cdot F_s/2 \qquad (1)$$

$$\Delta_{\ell,n} = 2^{-\ell} \cdot F_s/2 \qquad (2)$$

where GC^{-1} is the inverse Gray code permutation and F_s is the sampling frequency in the signal space. To obtain critical bands with the WPD, we construct a decomposition tree such that the distance between the center frequency of one subband to the center frequency of the next subband is 1 Bark. The relation between frequency f in Hertz and critical band rate z in Bark is approximately given by [11]

$$z = 26.81/(1 + 1960/f) - 0.53\,. \qquad (3)$$

Fig. 1 shows an approximation of the bark scale by critical-band WPD (CB-WPD). The corresponding decomposition tree is depicted in Fig. 2 by fine lines. The

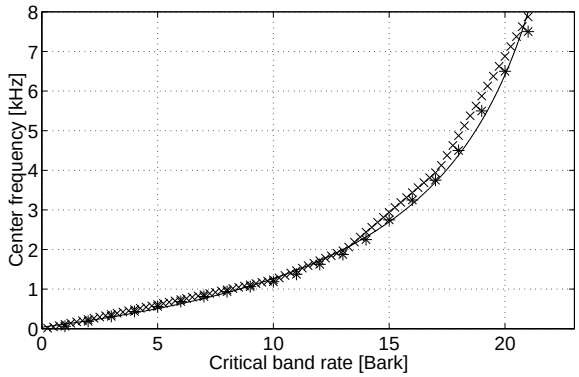

Figure 1: Approximation of the bark scale (solid) by CB-WPD ($*$) and BS-WPS ($\times$).

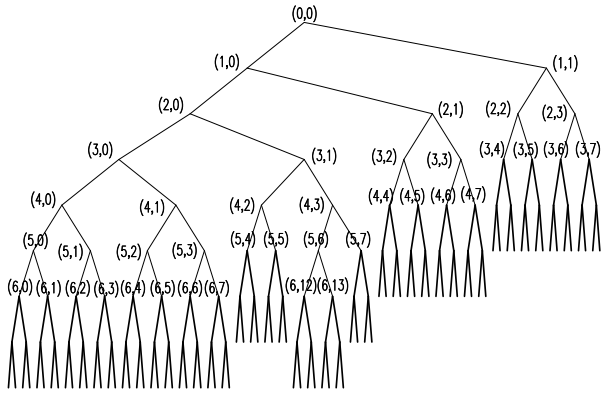

Figure 2: Wavelet packet expansion tree for the CB-WPD (the subtree depicted by fine lines) and BS-WPD. Fine and heavy lines designate, respectively, decimated and non-decimated expansions.

root of the tree $(\ell, n) = (0, 0)$ refers to the signal space. Each internal node in the tree $(\ell, n) \notin E$ is split into children-nodes $(\ell + 1, 2n)$ and $(\ell + 1, 2n + 1)$. The left and right branches, connecting a given node to its children-nodes, denote respectively low-pass and high-pass wavelet filtering followed by a 2:1 down-sampling.

The CB-WPD splits the frequency range [0 8] kHz into 21 subbands. For the application of speech enhancement, it is useful to increase the number of subbands (refine the frequency resolution) and to allow some degree of redundancy (oversampling). Accordingly, we further decompose the terminal nodes of the CB-WPD tree, each into four non-decimated subbands, by a two-level overcomplete expansion (Fig. 2). Specifically,

$$(\ell, n) \in E_{CB} \quad \Rightarrow \quad (\ell + 2, 4n + i) \in E_{BS} \qquad (4)$$

for $i = 0, \ldots, 3$, where E_{CB} and E_{BS} are the sets of terminal nodes for the CB-WPD and the BS-WPD, respectively. The additional branches, depicted in Fig. 2 by heavy lines, represent low-pass and high-pass wavelet filtering *without* down-sampling. This results in 84 subbands having a redundancy ratio equals to 4. The approximation of the bark scale by the BS-WPD is shown in Fig. 1.

For a comparison with a uniform-band WPD (UB-WPD), we also tested 8-level full tree WPD, having $2^8 = 256$ subbands. The oversampling ratio of 4 was similarly obtained by excluding the down-sampling at the two coarsest decomposition levels.

3. Spectral Enhancement

Let $x(t)$ and $d(t)$ denote speech and uncorrelated additive noise signals. The observed signal $y(t) = x(t) + d(t)$ is transformed into the wavelet packet domain, where the clean speech is estimated based on a minimum mean-square error (MMSE) criterion. Subsequently, the estimate for the clean speech is transformed back into the signal space using an inverse WPD.

The expansion coefficients of the noisy signal are given by

$$Y_{\ell,n}(k) = \left\langle y, \psi_{\ell,n,k/M} \right\rangle , \quad (\ell, n) \in E \, , \, k \in \mathbb{Z} \qquad (5)$$

where the oversampling ratio (M) is 1 for the CB-WPD and 4 for the BS-WPD and UB-WPD. Taking into account speech presence uncertainty, we have the following hypothesis testing problem:

$$\begin{aligned} H_0 \; &: \; Y_{\ell,n}(k) = D_{\ell,n}(k) \\ H_1 \; &: \; Y_{\ell,n}(k) = X_{\ell,n}(k) + D_{\ell,n}(k) \end{aligned} \qquad (6)$$

where $X_{\ell,n}(k)$ and $D_{\ell,n}(k)$ represent the expansion coefficients of the clean and noise signals, respectively, and H_0 and H_1, indicate speech absence and presence in the time index k of the subband (ℓ, n).

Assuming a Gaussian distribution for both speech and noise [8], the conditional probability density functions of the observed signal are given by

$$\begin{aligned} p(Y_{\ell,n}(k)|H_0) &= \frac{1}{\sqrt{2\pi}\sigma_{\ell,n}(k)} \exp\left\{ -\frac{Y^2_{\ell,n}(k)}{2\sigma^2_{\ell,n}(k)} \right\} \\ p(Y_{\ell,n}(k)|H_1) &= \frac{1}{\sqrt{2\,\pi(\sigma^2_{\ell,n}(k) + \lambda^2_{\ell,n}(k))}} \\ &\quad \cdot \exp\left\{ -\frac{Y^2_{\ell,n}(k)}{2(\sigma^2_{\ell,n}(k) + \lambda^2_{\ell,n}(k))} \right\} \end{aligned} \qquad (7)$$

where $\sigma^2_{\ell,n}(k) = E[D^2_{\ell,n}(k)]$ and $\lambda^2_{\ell,n}(k) = E[X^2_{\ell,n}(k)]$ denote the variances of noise and speech. Hence, the conditional signal presence probability $p_{\ell,n}(k) \triangleq P(H_1|Y_{\ell,n}(k))$, obtained using Bayes rule, is given by

$$p_{\ell,n}(k) = \left\{ 1 + \frac{\sqrt{1 + \eta_{\ell,n}(k)}}{q^{-1}_{\ell,n}(k) - 1} \exp(-\upsilon_{\ell,n}(k)/2) \right\}^{-1} \qquad (8)$$

where $q_{\ell,n}(k) \triangleq P(H_0)$ is the *a priori* probability for speech absence [12], $\upsilon \triangleq \gamma\eta/(1 + \eta)$, $\eta \triangleq \lambda^2/\sigma^2$ is the *a priori* SNR, and $\gamma \triangleq Y^2/\sigma^2$ is the *a posteriori* SNR.

For simplicity, we assume that expansion coefficients are statistically independent. Accordingly, an estimate for the clean speech, which minimizes the mean-square error, satisfies $\hat{X}_{\ell,n}(k) = E\left[X_{\ell,n}(k)|Y_{\ell,n}(k)\right]$. Based on the statistical model, this results in a modified Wiener amplitude estimator:

$$\begin{aligned} \hat{X}_{\ell,n}(k) &= E\left[X_{\ell,n}(k)|Y_{\ell,n}(k), H_1\right] p_{\ell,n}(k) \\ &= \frac{\lambda^2_{\ell,n}(k) \cdot p_{\ell,n}(k)}{\lambda^2_{\ell,n}(k) + \sigma^2_{\ell,n}(k)} Y_{\ell,n}(k) \, , \end{aligned} \qquad (9)$$

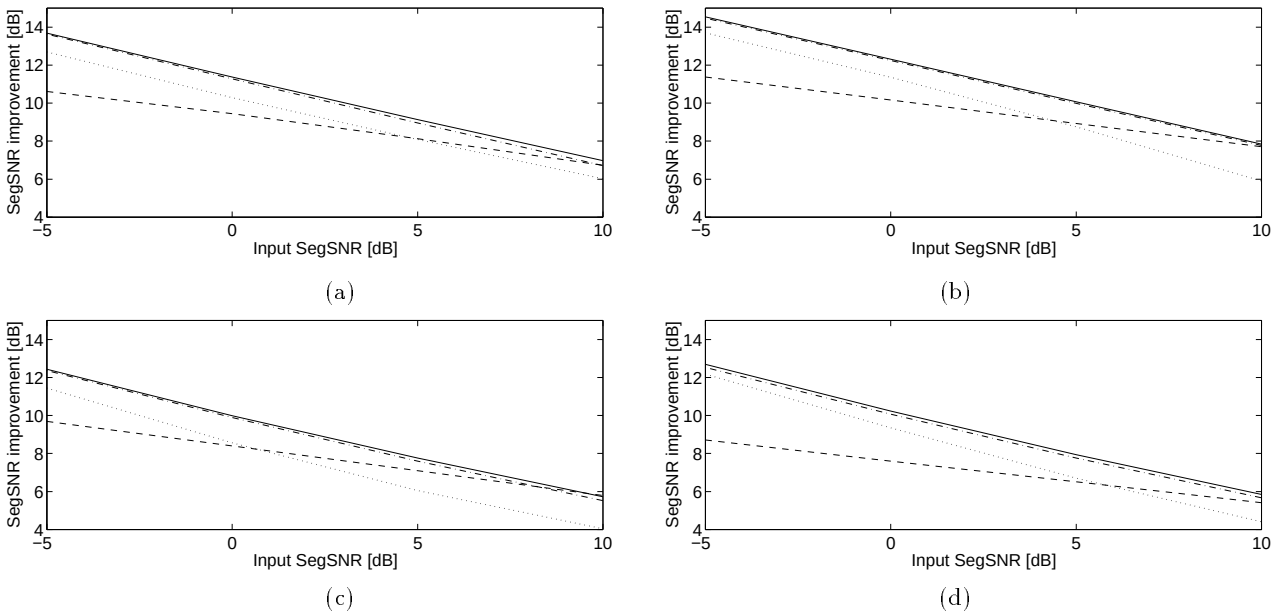

Figure 3: Average Segmental SNR improvement for various noise types and levels: (a) White Gaussian noise; (b) Car interior noise; (c) F16 cockpit noise; (d) Speech babble noise. The conventional LSA estimator is based on the STFT (dashed). The proposed estimators are based on CB-WPD (dotted), UB-WPD (dashdot), and BS-WPD (solid).

where we used $E\left[X_{\ell,n}(k)|Y_{\ell,n}(k), H_0\right] \equiv 0$. Since the variance of speech is not directly available, it needs to be estimated from the noisy speech itself. We propose the following "magnitude" decision-directed estimate:

$$\hat{\lambda}_{\ell,n}(k) = \alpha|\hat{X}_{\ell,n}(k-1)| + (1-\alpha)\max\left\{|Y_{\ell,n}(k)| - \sigma_{\ell,n}(k), 0\right\} \quad (10)$$

where α is a forgetting factor, which controls the tradeoff between noise reduction and signal distortion. This estimate is closely related to Ephraim and Malah's decision-directed estimate for the *a priori* SNR [8]:

$$\hat{\eta}_{\ell,n}(k) = \alpha\frac{\hat{X}^2_{\ell,n}(k-1)}{\sigma^2_{\ell,n}(k-1)} + (1-\alpha)\max\left\{\gamma_{\ell,n}(k) - 1, 0\right\} \quad (11)$$

However, we found the above formulation (using expression (10) rather than (11)) somehow preferable. Specifically, a lower level of musical residual noise can be achieved, while retaining the same amount of perceptual signal distortion.

Finally, the estimate for the clean speech is obtained by

$$\hat{x}(t) = \frac{1}{M}\sum_{(\ell,n)\in E,\ k\in\mathbb{Z}} \hat{X}_{\ell,n}(k)\tilde{\psi}_{\ell,n,k/M}(t) \quad (12)$$

where $\{\tilde{\psi}_n(t) : n \in \mathbb{Z}_+\}$ is the "dual" (biorthogonal) wavelet packet family of $\{\psi_n(t) : n \in \mathbb{Z}_+\}$, and $\tilde{\psi}_{\ell,n,k/M}(t) \triangleq 2^{-\ell/2}\tilde{\psi}_n(2^{-\ell}t - k/M)$.

4. Performance Evaluation

The evaluation of the proposed enhancement algorithm, and a comparison to the LSA estimator [9], consists of an objective segmental SNR measure, a subjective study of speech spectrograms and informal listening tests. Four different noise types, taken from Noisex92 database, are used in our evaluation: white Gaussian noise, car noise, F16 cockpit noise, and speech babble noise. The performance results are averaged out using six different utterances, taken from the TIMIT database. Half of the utterances are from male speakers, and half are from female speakers.

The speech signals, sampled at 16 kHz, are degraded by the various noise types with segmental SNR's in the range $[-5, 10]$ dB. The BS-WPD is implemented with the discrete Meyer wavelet filters, which provide good separation of subbands due to their regularity property. The short-time Fourier transform (STFT), for the LSA estimator, is implemented with Hanning windows of 512 samples length (32 ms) and 128 samples time step (75% overlapping). We used a forgetting factor $\alpha = 0.92$, where the proposed algorithm employs the "magnitude" decision-directed estimate (Eq. (10)), and the LSA estimator is based on the decision-directed estimate for the *a priori SNR* [8]. Additionally, we restricted the spectral gain to a minimum -25 dB, and assumed that the noise statistics is known (in practice, the noise spectrum is estimated using the *Minima Controlled Recursive Averaging* approach [13]).

Fig. 3 shows the average segmental SNR improvement obtained with the BS-WPD, CB-WPD, UB-WPD, and the LSA estimator. The estimator based on the BS-WPD yields the best results under all noise conditions. The UB-WPD provides similar results. However, subjective informal listening tests were in favor of the former, due to a higher quality of unvoiced sounds. A subjective comparison was also conducted using speech spectrograms (Fig. 4). Generally, spectral enhancement in the wavelet packet domain may result in some speech distortion, attributable to aliasing. Yet, the BS-WPD enables

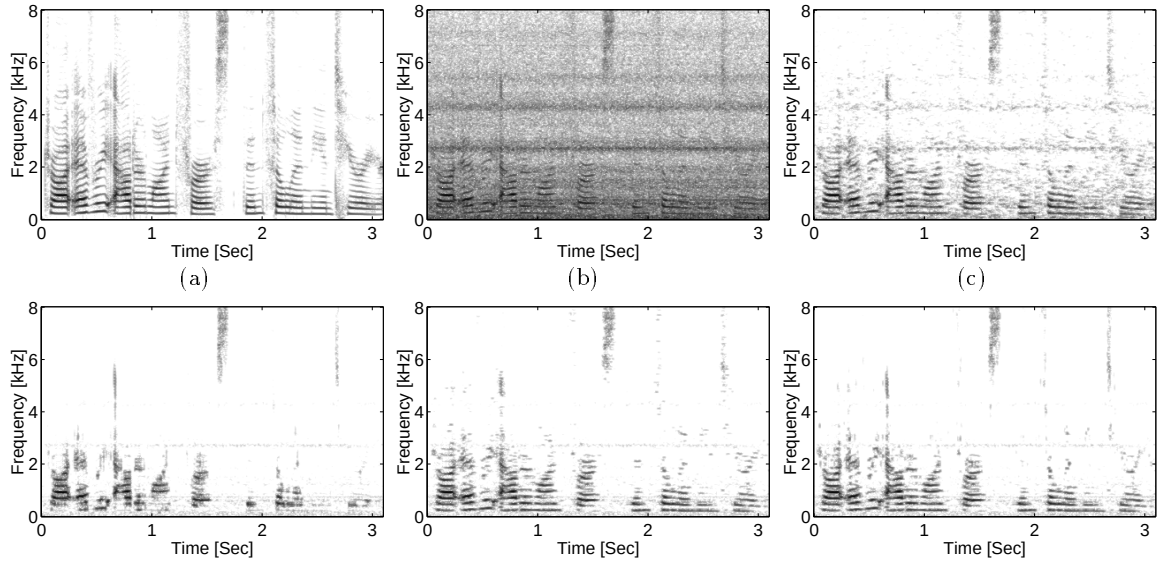

Figure 4: Speech spectrograms. (a) Original clean speech signal: "Draw every outer line first, then fill in the interior."; (b) Noisy signal (additive F16 cockpit noise at a SegSNR = 0 dB); (c) Speech enhanced with the conventional LSA estimator (SegSNR = 8.3 dB); (d) Speech enhanced with the CB-WPD (SegSNR = 6.6 dB); (e) the UB-WPD (SegSNR = 9.0 dB); and (f) the BS-WPD (SegSNR = 9.1 dB) based estimators.

a reduced level of musical residual noise, while preserving the perceptual quality of speech components.

5. Conclusion

Refining the frequency resolution of the critical band decomposition, and introducing a certain amount of redundancy, improve the speech enhancement performance. Excessive expansion of high frequency subbands, *e.g.* using the UB-WPD or STFT, not only increases the computational complexity, but also degrades the perceptual quality of unvoiced sounds. The BS-WPD provides an efficient auditory representation that is combined with the modified Wiener filtering and the "magnitude" decision-directed estimation technique. It leads to lower residual noise and higher intelligibility and quality of enhanced speech, compared to classical enhancement systems, which are based on uniform spectral decompositions. Furthermore, the redundancy in the BS-WPD can be exploited to minimize the overlapping between adjacent subbands, thereby eliminating artifacts associated with aliasing.

6. Acknowledgement

The author thanks Baruch Berdugo for valuable discussions.

7. References

[1] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 2, March 1999, pp. 126–137.

[2] L. Singh and S. Sridharan, "Speech enhancement using critical band spectral subtraction," *Proc. ICSLP*, Sydney, Australia, December 1998, pp. 2827–2830.

[3] B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms," *IEEE Trans. Signal Processing*, vol. 47, no. 6, June 1999, pp. 1622–1635.

[4] P. Dreiseeitel and H. Puder, "Speech enhancement for mobile telephony based on non-uniformly spaced frequency resolution," *Proc. EUSIPCO*, 1998, pp. 965–968.

[5] T. Gülzow, A. Engelsberg and U. Heute, "Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement," *Signal Processing*, vol. 64, no. 1, January 1998, pp. 5–19.

[6] E. Ambikairajah, G. Tattersall and A. Davis, "Wavelet transform-based speech enhancement," *Proc. ICSLP*, Sydney, Australia, December 1998, pp. 2811–2814.

[7] I. Y. Soon, S. N. Koh and C. K. Yeo, "Wavelet for speech denoising," *Proc. TENCON*, Dec. 1997, pp. 479–482.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, December 1984, pp. 1109–1121.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 2, April 1985, pp. 443–445.

[10] M. V. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, AK Peters, Ltd, Wellesley, Massachusetts, 1994.

[11] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *J. Acoust. Soc. Am.*, vol. 88, 1990, pp. 97–100.

[12] I. Cohen, "On speech enhancement under signal presence uncertainty," *Proc. 26th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-01*, Salt Lake City, Utah, 7–11 May 2001.

[13] I. Cohen and B. Berdugo, "Spectral enhancement by tracking speech presence probability in subbands," *Proc. IEEE Workshop on Hands Free Speech Communication, HSC'01*, Kyoto, Japan, 9–11 April 2001.