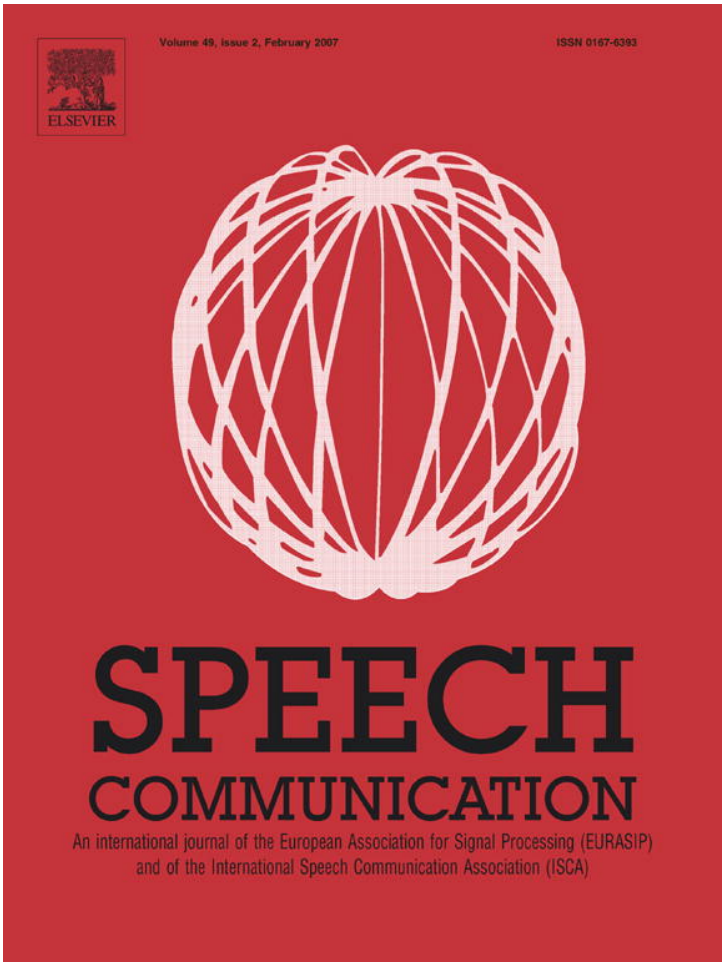

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 123–133
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Speech signal enhancement through adaptive wavelet thresholding

Michael T. Johnson [a,*], Xiaolong Yuan [b], Yao Ren [a,1]

[a] Department of Electrical and Computer Engineering, 1515 W. Wisconsin Avenue, Milwaukee, WI 53233, United States
[b] Motorola Electronics Ltd, No. 108 Jain Guo Road, Chao Yang District, Beijing 100022, PR China

Received 14 March 2006; received in revised form 29 September 2006; accepted 11 December 2006

Abstract

This paper demonstrates the application of the Bionic Wavelet Transform (BWT), an adaptive wavelet transform derived from a non-linear auditory model of the cochlea, to the task of speech signal enhancement. Results, measured objectively by Signal-to-Noise ratio (SNR) and Segmental SNR (SSNR) and subjectively by Mean Opinion Score (MOS), are given for additive white Gaussian noise as well as four different types of realistic noise environments. Enhancement is accomplished through the use of thresholding on the adapted BWT coefficients, and the results are compared to a variety of speech enhancement techniques, including Ephraim Malah filtering, iterative Wiener filtering, and spectral subtraction, as well as to wavelet denoising based on a perceptually scaled wavelet packet transform decomposition. Overall results indicate that SNR and SSNR improvements for the proposed approach are comparable to those of the Ephraim Malah filter, with BWT enhancement giving the best results of all methods for the noisiest (−10 db and −5 db input SNR) conditions. Subjective measurements using MOS surveys across a variety of 0 db SNR noise conditions indicate enhancement quality competitive with but still lower than results for Ephraim Malah filtering and iterative Wiener filtering, but higher than the perceptually scaled wavelet method.

Keywords: Adaptive wavelets; Bionic Wavelet Transform; Speech enhancement; Denoising

1. Introduction

Speech enhancement is an important problem within the field of speech and signal processing, with impact on many computer-based speech recognition, coding and communication applications. The underlying goal of speech enhancement is to improve the quality and intelligibility of the signal, as perceived by human listeners. Existing approaches to this task include traditional methods such as spectral subtraction (Boll, 1979; Deller et al., 2000), Wiener filtering (Deller et al., 2000; Haykin, 1996), and Ephraim Malah filtering (Ephraim and Malah, 1984). Wavelet-based techniques using coefficient thresholding approaches have also been applied to speech enhancement (Donoho, 1995; Guo et al., 2000), and more recently a number of attempts have been made to use perceptually motivated wavelet decompositions coupled with various thresholding and estimation methods (Bahoura and Rouat, 2001; Chen et al., 2004; Cohen, 2001; Fu and Wan, 2003; Hu and Loizou, 2004; Lu and Wang, 2003).

Recently, the Bionic Wavelet Transform (BWT) (Yao and Zhang, 2001, 2002) has been proposed as a method for wavelet decomposition of speech signals. The BWT was originally designed for applications in speech coding, with particular emphasis on the possibility of using it for encoding of cochlear implant signals. The BWT model, which will be described in more detail in the next section, is based on an auditory model of the human cochlea, capturing the non-linearities present in the basilar membrane and translating those into adaptive time-scale transformations of the underlying mother wavelet. Motivated by the communicative connection between the speech production

* Corresponding author. Tel.: +1 414 288 0631; fax: +1 414 288 5579.
E-mail addresses: mike.johnson@mu.edu (M.T. Johnson), xlyuan0514@yahoo.com (X. Yuan), yao.ren@mu.edu (Y. Ren).
[1] Tel.: +1 414 288 7451.

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.12.002

system and the auditory system, the work presented here uses the BWT in combination with existing wavelet denoising techniques to construct a new adaptive wavelet thresholding method for speech enhancement, with the idea that the improved representational capability of the BWT on speech signals could lead to better separation of signal and noise components within the coefficients and therefore better enhancement results.

In Section 2, we give a detailed overview of wavelet decompositions, wavelet thresholding techniques, and the BWT, as well as a brief discussion of the baseline enhancement methods being used for comparison. Section 3 introduces the new approach and outlines the experimental method, including the experiments, the data set, the noise models, and the evaluation metrics. Results of these experiments are presented and discussed in Section 4, followed by overall conclusions in Section 5.

2. Background

2.1. Wavelet analysis (Debnath, 2002; Jaffard et al., 2001; Walnut, 2002)

The Continuous Wavelet Transform (CWT) of a signal $x(t)$ is given by

$$X_{\mathrm{CWT}}(a,\tau) = \langle x(t), \varphi_{a,\tau}(t)\rangle = \frac{1}{\sqrt{a}}\int x(t)\varphi^*\left(\frac{t-\tau}{a}\right)\mathrm{d}t, \quad (1)$$

where τ and a represent the time shift and scale variables, respectively, and $\varphi(\cdot)$ is the mother wavelet chosen for the transform. Given that this mother wavelet satisfies a basic admissibility criteria (Daubechies, 1992), the inverse transform also exists. The idea of the wavelet originated with the Gabor Transform (Gabor, 1946), a windowed Fourier Transform designed such that the duration of the time localization window varied with frequency. A wavelet representation offers advantages over traditional Fourier analysis in that the time support of the wavelet used to perform the correlation in Eq. (1) varies as a function of scale, so that the analysis window length matches with the frequency of interest, trading off time and frequency resolution.

The time τ and scale a variables of the CWT can be discretized and in many cases still provide for complete representation of the underlying signal, provided that the mother wavelet meets certain requirements. This may be viewed as a type of multiresolution analysis, where at each scale the signal is represented at a different level of detail. In the case where the discretization of the time and scale variables is dyadic in nature, so that $a = 2^m$ and $\tau = n2^m$, an efficient implementation may be obtained through the use of a Quadrature Mirror Filter (QMF) decomposition at each level, where matching low-pass and high-pass filterbank coefficients characterize the convolution with the mother wavelet, and downsampling by 2 at each level equates to the doubling of the time interval according to scale. This multiresolution filter bank implementation is referred to as a Discrete Wavelet Transform (DWT), and exists provided that the family of wavelets generated by dyadic scaling and translation forms an orthonormal basis set. The BWT used in the current work is based on the Morlet mother wavelet, for which a direct DWT representation is not possible, so a CWT coupled with fast numerical integration techniques is used instead to generate a set of discretized wavelet coefficients.

A further generalization of the DWT is the Wavelet Packet Transform (WPT), also based on a filter bank decomposition approach. In this case the filtering process is iterated on both high and low frequency components, rather than continuing only on low frequency terms as with a standard dyadic DWT. A comparison between a DWT and a WPT is shown in Fig. 1.

The depth of the Wavelet Packet Tree shown in Fig. 1 can be varied over the available frequency range, resulting in configurable filterbank decomposition. This idea has been used to create customized Wavelet Packet Transforms where the filterbanks match a perceptual auditory scale, such as the Bark scale, for use in speech representation, coding, and enhancement (Bahoura and Rouat, 2001; Chen et al., 2004; Cohen, 2001; Fu and Wan, 2003; Hu and Loizou, 2004; Lu and Wang, 2003). The use of bark-scale WPT for enhancement has so far indicated a small but significant gain in overall enhancement quality due to this

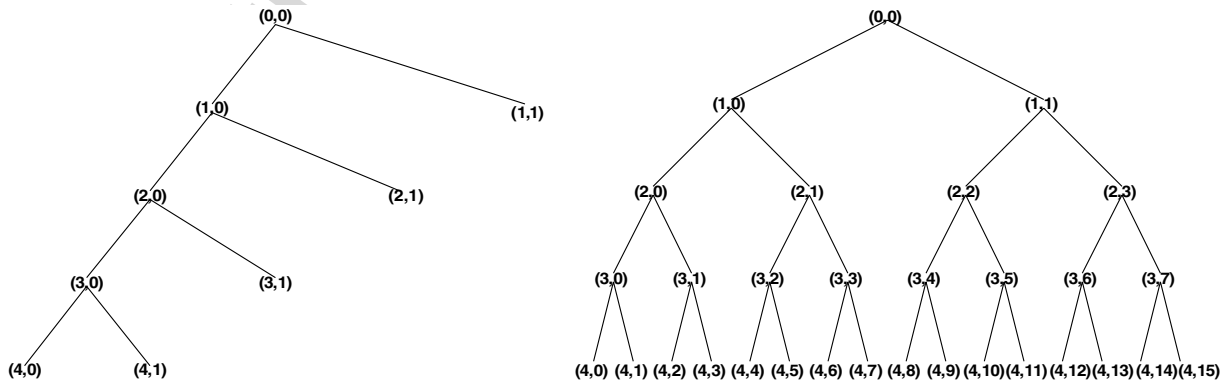

Fig. 1. A Discrete Wavelet Transform (left) and full Wavelet Packet Transform (right), represented as filterbank decompositions, with the left and right branches at each node representing a matched pair of low-pass and high-pass wavelet filters followed by downsampling.

perceptual specialization. This perceptual WPT, using auditory critical band scaling following Cohen's work (Cohen, 2001) as shown in Fig. 2, is implemented in this work as a reference method for comparison to the new technique.

The degree to which a particular set of wavelet coefficients form a useful or compact representation of a signal is a function of how well the mother wavelet matches with the underlying signal characteristics, as well as the times and scales selected. For application to signal enhancement, often referred to in the literature as wavelet denoising, the coefficient magnitudes are reduced after comparison to a threshold, as described in more detail in the following section. With a good choice of representation, this thresholding will remove noise while maintaining signal properties.

To address the fact that many types of signals have substantial non-stationarity and may not be well-represented by a single fixed set of parameters, it is possible to make the wavelet transform adaptive, such that characteristics of the transform change over time as a function of the underlying signal characteristics. There are several possible approaches to adaptive wavelet enhancement, including adaptation of the wavelet basis, adaptation of the wavelet packet configuration, direct adaptation of the time and scale variables, or adaptation of the thresholds or thresholding algorithms used. Of these, the most common approach is to use a time-varying threshold or gain function based on an *a priori* energy or SNR measure (Bahoura and Rouat, 2001; Chen et al., 2004; Cohen, 2001; Fu and Wan, 2003; Hu and Loizou, 2004; Lu and Wang, 2003).

The BWT decomposition used here is both perceptually scaled and adaptive. The initial perceptual aspect of the transform comes from the logarithmic spacing of the baseline scale variables, which are designed to match basilar-membrane spacing. Two adaptation factors then control the time-support used at each scale, based on a non-linear perceptual model of the auditory system, as described in detail in the following section.

2.2. Bionic Wavelet Transform

The BWT was introduced in (Yao, 2001, Yao and Zhang, 2001) as an adaptive wavelet transform designed specifically to model the human auditory system. The basis for this transform is the Giguere–Woodland non-linear transmission line model of the auditory system (Giguere, 1993; Giguere and Woodland, 1994), an active-feedback electro-acoustic model incorporating the auditory canal, middle ear, and cochlea. The model yields estimates of the time-varying acoustic compliance and resistance along the displaced basilar membrane, as a function of the physiological acoustic mass, cochlear frequency-position mapping, and feedback factors representing the active mechanisms of the outer hair cells. The net result can be viewed as a method for estimating the time-varying quality factor Q_{eq} of the cochlear filter banks as a function of the input sound waveform. See Giguere and Woodland (1994), Zheng et al. (1999) and Yao and Zhang (2001) for complete details on the elements of this model.

The adaptive nature of the BWT is captured by a time-varying linear factor $T(a,\tau)$ that represents the scaling of the cochlear filter bank quality factor Q_{eq} at each scale over time. Incorporating this directly into the scale factor of a Morlet wavelet, we have:

$$X_{\text{BWT}}(a,\tau) = \frac{1}{T(a,\tau)\sqrt{a}} \int x(t)\tilde{\varphi}^*\left(\frac{t-\tau}{T(a,\tau)a}\right) e^{-j\omega_0\left(\frac{t-\tau}{a}\right)}\,dt, \tag{2}$$

where

$$\tilde{\varphi}(t) = e^{-\left(\frac{t}{T_0}\right)^2} \tag{3}$$

is the amplitude envelope of the Morlet wavelet, T_0 is the initial time-support and ω_0 is the base fundamental frequency of the unscaled mother wavelet, here taken as $\omega_0 = 15{,}165.4$ Hz for the human auditory system, per Yao and Zhang's original work (Yao and Zhang, 2001). The discretization of the scale variable a is accomplished using pre-determined logarithmic spacing across the desired frequency range, so that the center frequency at each scale is given by the formula $\omega_m = \omega_0/(1.1623)^m$, $m = 0,1,2,\ldots$ For this implementation, based on Yao and Zhang's original work for cochlear implant coding (Yao and Zhang, 2002), coefficients at 22 scales,

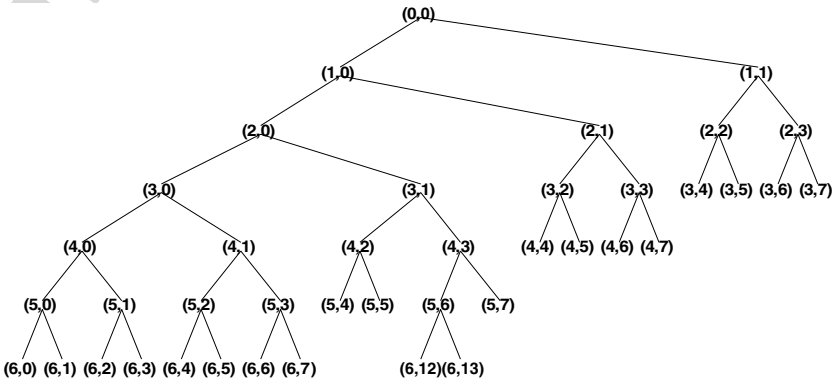

Fig. 2. Perceptually scaled Wavelet Packet Transformation, with leaf-node center frequencies following an approximately critical-band scaling.

m = 7,...,28, are calculated using numerical integration of the continuous wavelet transform. These 22 scales correspond to center frequencies logarithmically spaced from 225 Hz to 5300 Hz. (Although the scales used here match those from Yao and Zhang's original work, empirical variation of the number of scales and frequency placement showed minimal effect on the overall enhancement results.)

The BWT adaptation factor $T(a,\tau)$ for each scale and time is computed using the update equation:

$$T(a,\tau+\Delta\tau) = \frac{1}{\left(1 - G_1 \frac{C_s}{C_s + |X_{\mathrm{BWT}}(a,\tau)|}\right)\left(1 + G_2 \left|\frac{\partial}{\partial\tau} X_{\mathrm{BWT}}(a,\tau)\right|\right)}, \tag{4}$$

where G_1 is the active gain factor representing the outer hair cell active resistance function, G_2 is the active gain factor representing the time-varying compliance of the Basilar membrane, and $C_s = 0.8$ is a constant representing nonlinear saturation effects in the cochlear model (Yao and Zhang, 2001). In practice, the partial derivative of Eq. (4) is approximated using the first difference of the previous points of the BWT at that scale.

From Eq. (2), it can be seen that the adaptation factor $T(a,\tau)$ affects the duration of the amplitude envelope of the wavelet, but does not affect the frequency of the associated complex exponential. Thus, one useful way to think of the BWT is as a mechanism for adapting the time support of the underlying wavelet according to the quality factor Q_{eq} of the corresponding cochlear filter model at each scale. The key parameters T_0, G_1, and G_2 will be discussed in detail in Sections 4.1 and 4.2.

It can be shown (Yao and Zhang, 2002) that the resulting BWT coefficients $X_{\mathrm{BWT}}(a,\tau)$ can be calculated as a product of the original WT coefficients $X_{\mathrm{WT}}(a,\tau)$ and a multiplying constant $K(a,\tau)$ which is a function of the adaptation factor $T(a,\tau)$. For the Morlet wavelet, this adaptive multiplying factor can be expressed as

$$X_{\mathrm{BWT}}(a,\tau) = K(a,\tau) X_{\mathrm{WT}}(a,\tau),$$
$$K(a,\tau) = \frac{\sqrt{\pi}}{C} \frac{T_0}{\sqrt{1 + T^2(a,\tau)}}, \tag{5}$$

where C is a normalizing constant computed from the integral of the squared mother wavelet. This representation yields an efficient computational method for computing BWT coefficients directly from the original WT coefficients without needing to perform the numerical integration of Eq. (2) at each time and scale.

There are several key differences between the discretized CWT using the Morlet wavelet, used for the BWT, and a filterbank-based WPT using an orthonormal wavelet such as the Daubechies family, as used for the comparative baseline method. One is that the WPT is perfectly reconstructable, whereas the discretized CWT is an approximation whose exactness depends on the number and placement of frequency bands selected. Another difference, related to this idea, is that the Morlet mother wavelet consists of a single frequency with an exponentially decaying time support, whereas the frequency support of the orthonormal wavelet families used for DWTs and WPTs covers a broader bandwidth. The Morlet wavelet is thus more "frequency focused" along each scale, which is what permits the direct adaptation of the time support with minimal impact on the frequency support, the central mechanism of the BWT adaptation.

2.3. Wavelet denoising

If an observed signal includes measurement error or ambient noise, the result is an additive signal model given by

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}, \tag{6}$$

where $\boldsymbol{y}$ is the noisy signal, $\boldsymbol{x}$ is the original clean signal, and $\boldsymbol{n}$ is the additive noise component. It can easily be seen from Eq. (1) above that the wavelet coefficients are also additive, so that $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{N}$, where the matrix notation represents the set of coefficients across the selected times and scales. Given a well-matched wavelet representation, noise characteristics will tend to be characterized across time and scale by smaller coefficients while signal energy will be concentrated in larger coefficients. This offers the possibility of using thresholding to separate the signal from the noise. There are a wide variety of basic thresholding approaches (Antoniadis and Oppenheim, 1995; Donoho, 1995), including:

- Hard thresholding, where all coefficients below a predefined threshold value are set to zero.
- Soft thresholding, where in addition the remaining coefficients are linearly reduced in value.
- Non-linear thresholding, where a smooth function is used to map the original coefficients to a new set, avoiding abrupt value changes.

Illustrations of hard, soft, and non-linear thresholding operations are shown in Fig. 3. For any of these approaches, the threshold parameter may be either a fixed value or a level-dependent value that is a function of the wavelet decomposition level. One of the key elements for successful wavelet denoising is the selection of the thresh-

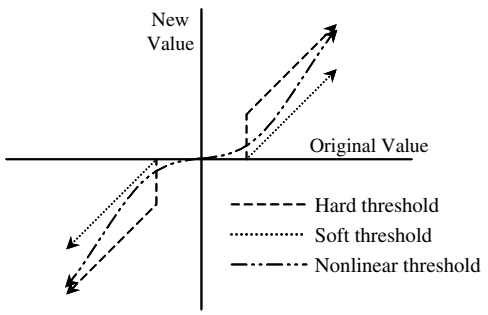

Fig. 3. Threshold mapping functions.

old. Optimization approaches for determining this value have been well-studied. Two of the most common methods are universal thresholding and Stein's unbiased risk estimator (SURE) (Donoho, 1995), typically implemented with a soft-thresholding function.

Universal thresholding uses the threshold value

$$T = \hat{\sigma}\sqrt{2\log N}, \tag{7}$$

where $\hat{\sigma}$ is an estimate of the noise variance and N compensates for signal length. The variance $\hat{\sigma}$ is typically obtained using the median absolute deviation (MAD) measure with respect to time, $\hat{\sigma} = \mathrm{MAD}(X_{\mathrm{DWT}}(n,m))/0.6745$, evaluated at a specific scale level m.

The SURE method (Donoho, 1995; Donoho and Johnstone, 1995; Johnstone and Silverman, 1997) uses the value

$$T = \underset{0\leqslant T\leqslant\hat{\sigma}\sqrt{2\log N}}{\arg\min}\left\{\hat{\sigma}^2 N + \sum_{n=1}^{N}\left[\min(x[n],T^2) - 2\hat{\sigma}^2 I(|x[n]|\leqslant T)\right]\right\}, \tag{8}$$

where $x[n]$ is the time-domain input signal and I is an indicator function. Note that the range over which the SURE threshold is considered is based on a maximum value equal to the universal threshold, so that the SURE threshold is always less than the universal threshold.

Thresholding can be done universally across all wavelet decomposition levels, referred to as level-independent thresholding, or else the threshold level can be varied at each level, level-dependent thresholding. In this case, the above formulas still apply, but with a level dependent threshold T_m calculated at each level using a scale-dependent variance estimate $\hat{\sigma}_m = \mathrm{MAD}(X_{\mathrm{DWT}}(n,m))/0.6745$.

Other thresholding approaches include minimax thresholding (Donoho, 1995), which sets a universal threshold independent of any signal information (and is therefore good primarily for completely unknown conditions), and heuristic SURE (Donoho, 1995), which uses a significance test to decide between the universal threshold and the SURE threshold.

The speech enhancement experiments presented here use the SURE method, with a level-independent soft-threshold. Empirical evaluations across the different thresholding selection methods for this task showed only minor variation in results, with slight performance benefit using the SURE approach on speech enhancement tasks compared to the universal and heuristic SURE methods.

2.4. Comparative baseline methods

To evaluate the effectiveness of using this new BWT-based method for enhancement of speech signals, we compare it to other standard approaches on this task. This includes comparing to the spectral subtraction, Wiener filtering, and Ephraim Malah filtering methods, all common approaches within this area. In addition, we also compare to a perceptually scaled WPT denoising implementation. In this section some background on these existing enhancement methods is provided.

Spectral subtraction (Boll, 1979; Deller et al., 2000) is a straightforward technique based on literal subtraction of the Fourier Transform (FT) magnitude components of the estimated noise spectrum from the signal spectrum, on a frame by frame basis. Noise components are typically estimated from regions of the signal where there is no speech present, and phase characteristics are taken directly from the noisy FT. The resulting enhancement equation is given by

$$\hat{\boldsymbol{x}} = \mathrm{IFFT}\left(\sqrt{|Y(\omega)|^2 - |N(\omega)|^2}\angle Y(\omega)\right). \tag{9}$$

Wiener filtering (Deller et al., 2000; Haykin, 1996) is accomplished by using the signal and noise spectral characteristics to estimate the optimal noise reduction filter, given by

$$H = \frac{S_x(\omega)}{S_x(\omega) - S_n(\omega)}, \tag{10}$$

where $S_x(\omega)$ and $S_n(\omega)$ are the true power spectral densities of the clean signal and noise. As with spectral subtraction, the noise spectrum is typically estimated from regions of the signal where there is no speech present. However since there is no direct way to get an estimate of the speech spectrum, this is usually accomplished via an iterative procedure where $S_x(\omega)$ is initialized using the noisy signal spectrum. In each iteration H is estimated via Eq. (10), after which the filter is applied and an improved estimate of the signal is used to determine a better clean signal spectrum estimate $S_x(\omega)$. This process is repeated until convergence, usually just a few iterations.

The Ephraim Malah filter approach (Ephraim and Malah, 1984) is based on deriving a minimum mean square estimator (MMSE) for the clean speech spectral amplitudes $\hat{X}_k = \hat{A}_k e^{j\theta_k}$ (Ephraim and Malah, 1984) or log spectral amplitudes (LSA) (Ephraim and Malah, 1985) given a complex Gaussian random variable model for the Fourier Transform coefficients of both the clean speech and the noise, assuming independence across frequency bins. The LSA estimator gives somewhat better enhancement results, using the derived estimation formula for the clean signal Fourier transform coefficient in each frequency bin given by

$$\hat{A}_k = \frac{\xi_k}{1+\xi_k} e^{\left(\frac{1}{2}\int_{v_k}^{\infty}\frac{e^{-t}}{t}dt\right)} R_k, \tag{11}$$

$$\xi_k = \frac{\lambda_x(k)}{\lambda_n(k)}, \quad v_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \quad \gamma_k = \frac{R_k^2}{\lambda_n(k)}, \tag{12}$$

where R_k is the noisy speech Fourier transform magnitude in the kth frequency bin, and $\lambda_n(k)$ and $\lambda_x(k)$ are the average noise and signal powers in each bin. Noise power $\lambda_n(k)$ is typically estimated from initial silence regions in the

waveform, while $\lambda_x(k)$ is a moving average of spectrally subtracted noisy spectra $\left(R_k^2 - \lambda_n(k)\right)$. Improved results are obtained when the *a priori* SNR ξ_k is estimated directly and includes smoothing to implicitly adjust for speech presence probability, such as via the well-known "decision-directed method". The overall result is an algorithm which adaptively tracks and adjusts estimates of both noise and signal amplitudes, and uses these estimates to adjust the degree of enhancement, which has significant impact on reducing artifacts often present in spectral subtraction and Wiener filtering. In this work, we use the standard Ephraim Malah implementation based on the LSA algorithm and decision-directed *a priori* SNR estimation (Ephraim and Malah, 1985).

For comparison to another wavelet-based approach, we have implemented a level 6 perceptually scaled WPT, following the critical-band filter bank arrangement used in (Cohen, 2001). The result is a 21-band decomposition that approximates the spacing of the auditory bark scale, very similar to the spacing used for our BWT technique. A Daubechies-5 mother wavelet is used for the decomposition, and thresholding is accomplished with the same level independent SURE approach used for the BWT.

In all four baseline methods, as well as the new approach being tested, an initial segment of the waveforms, which contains no speech, is used to estimate the noise levels. For the spectral subtraction and Wiener filtering method, this consists of power spectral estimation using the Fourier transform, while for the Ephraim Malah filter this determines initial $\lambda_n(k)$ and $\lambda_x(k)$ values. For the wavelet methods this segment is used to re-scale the signal so to provide an estimated unity noise variance, matching the assumption used in the SURE threshold selection algorithm.

3. Experimental method

The enhancement method presented here is based on applying a wavelet thresholding technique to BWT coefficients. A block diagram of the overall approach is shown in Fig. 4.

Continuous wavelet coefficients are computed using discrete convolutions at each of the 22 scales based directly on Eq. (1), with a 16 kHz sampling rate, the same rate as the speech data used for the experiments. Eqs. (4) and (5) are used to calculate the K factor representing the time-support adaptation of the continuous wavelet coefficients. The selection of the T_0, G_1, and G_2 parameters for these equations is discussed in detail in Sections 4.1 and 4.2. To ensure stability of the overall signal variance for comparison to the threshold, the K factor term is shifted to a mean value of unity (variance/range are not adjusted). Thresholding is applied using the SURE thresholding method discussed in Section 2.3. Signal reconstruction is accomplished through another discrete convolution at each scale, followed by a weighted summation across the scales.

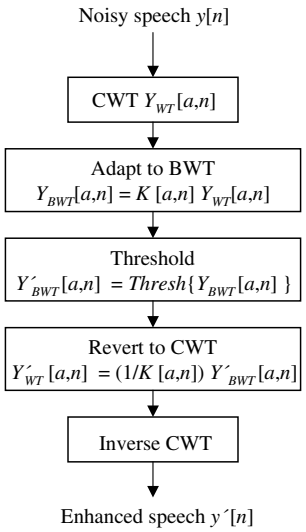

Fig. 4. Block diagram of the new BWT enhancement algorithm.

3.1. Data sets and noise characteristics

Ten utterances taken from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993) were used to evaluate the new algorithm. The sampling rate of the data is 16 kHz, and each sentence used has at least 100 ms of silence at the beginning of the utterance that can be used to estimate noise statistics for the comparative methods.

Two sets of additive noise experiments were implemented on this data. In the first, white Gaussian noise was added to the sentences at SNR levels of −10, −5, 0, +5, and +10 dB. In the second, specific noise characteristics including F-16 cockpit noise, Volvo car interior noise, ambient pink noise, and babble noise (multiple talkers), was added at a 0 dB SNR level to evaluate how well the methods work with non-white and relatively non-stationary noise sources.

Signal-to-noise ratio (SNR) and segmental signal-to-noise ratio (SSNR) (Deller et al., 2000) are used as objective measurement criteria for both sets of experiments. SSNR is computed by calculating the SNR on a frame-by-frame basis over the signal and averaging theses values, and has been shown to have a higher correlation with perceived quality than does a direct SNR metric. The formula for SSNR is

$$\mathrm{SSNR} = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\sum_{n=Nj+1}^{N(j+1)} \frac{x^2(n)}{[x(n)-\hat{x}(n)]^2}\right], \qquad (13)$$

where M is the number of frames, each of length N, and $x(n)$ and $\hat{x}(n)$ are the original and enhanced signals, respectively. SNR is computed using the inner term shown in the above equation summed across the entire signal.

For the non-stationary noise cases, a subjective perceptual measure called Mean Opinion Score (MOS) (Deller et al., 2000) was used to augment the objective SNR and SSNR measures for evaluating the perceived quality of

the enhanced waveforms. MOS is computed by having a group of listeners rate the quality of the speech on a five-point scale, then averaging the results. For these tests we used a group of 10 listeners in calculating MOS results. For all measures, results are averaged across the 10 utterances used as examples, giving a single evaluation metric for each method and noise type combination.

For spectral subtraction, Wiener filtering, and Ephraim Malah filtering, the signal is divided into 25 ms windows with 12 ms overlap between frames. Frequency analysis is done using a Hanning window, and noise estimation is accomplished using the first three frames of the signal. Coefficient thresholding for both the perceptually scaled WPT and the BWT are done using a soft level-independent thresholding function based on the SURE technique, as described in Section 2.4. Implementation was done using the Matlab Wavelet toolbox (The MathWorks Inc., 2003).

4. Results

The implementation of the BWT decomposition depends strongly on three primary parameters: T_0, G_1, and G_2. As discussed in Section 2.2, the T_0 parameter controls the base time-support of the un-adapted mother wavelet, while G_1 and G_2 control active gain factors representing outer hair cell resistance and Basilar membrane compliance, respectively. These parameters have been investigated in detail, and are discussed in Sections 4.1 and 4.2. Overall results of the enhancement experiments are presented in Section 4.3.

4.1. T_0 effects

The expression for a standard real Morlet wavelet is

$$\varphi(t) = \mathrm{e}^{-t^2/2}\cos(5t) \tag{14}$$

with frequency $\omega_0 = 5$, time support $T_0 = \sqrt{2}$, and a net time–frequency product of $T_0\omega_0 = 5\sqrt{2} \approx 7.07$. Scaling this to the base frequency of 15,165.4 Hz used in the BWT, the effective time support to match the standard Morlet would be $T_0 = (5\sqrt{2})/(2\pi(15,165.4)) \approx .00007421$. From this it can be seen that the T_0 value of 0.0005 from Yao and Zhang's original work (Yao and Zhang, 2001) is about seven times longer than that of a standard Morlet wavelet, resulting in several additional cycles in the initial mother wavelet.

Since the primary adaptation mechanism involves variation of the wavelet time support, the impact of the initial T_0 value was investigated. This was done by turning off the adaptation mechanism (i.e. $G_1 = G_2 = 0$ so that $T(a,\tau) = 1$) and investigating the SNR and SSNR resulting from thresholding the discretized CWT coefficients directly, for the AWGN noise case with input SNR = 0 db. Results are shown in Fig. 5.

It can be seen from the above plots that there is a substantial range across which the overall results are consistent, while either extremely long or extremely short time-support values substantially hurt performance. In accordance with this result, and because the time support value corresponding to a standard Morlet wavelet is in the middle part of the stable range indicated by these

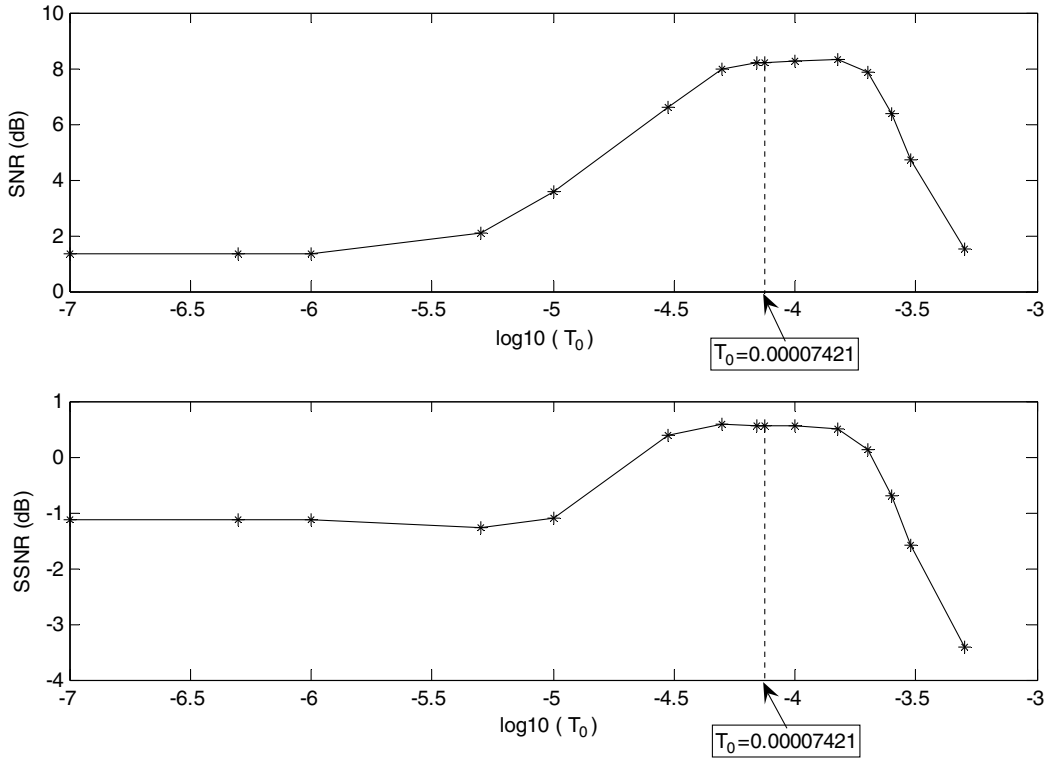

Fig. 5. SNR and SSNR versus T_0 for AWGN with initial SNR = 0 db.

results, it was decided to use this standard Morlet time-support value rather than one with extended time support, i.e. we selected $T_0 = .00007421$ for use in these experiments.

4.2. G_1 and G_2 effects

G_1 and G_2 are parameters of a non-linear adaptation model, and thus are sensitive to signal amplitude characteristics. Each of these factors independently controls the amount of adaptation for a specific component of the auditory model, with very small values resulting in almost no adaptation (neither doing harm nor providing assistance to the underlying wavelet decomposition) and very large values resulting in saturation effects that lead to excessive coefficient fluctuation and poor overall results. This was examined by varying the two factors together and plotting the corresponding SNR and SSNR results, again for the AWGN noise case with input SNR = 0 db. Results are shown in Fig. 6.

The overall result indicated above is that the SNR and SSNR results of the BWT enhancement algorithm are stable over a fairly wide range of parameter values. This would indicate that the exact choice of values is not greatly important. For the final implementation and evaluation, we selected values of $G_1 = 0.6$ and $G_2 = 75$, which are located at an approximate local maximum near the center portion of the stable region, indicated on the axes in Fig. 6.

4.3. Overall results compared to baseline methods

4.3.1. AWGN noise condition across range of SNR values

SNR and SSNR results for the white noise experiments are shown in Figs. 7 and 8. Methods compared include Ephraim Malah filtering, iterative Wiener filtering, spectral

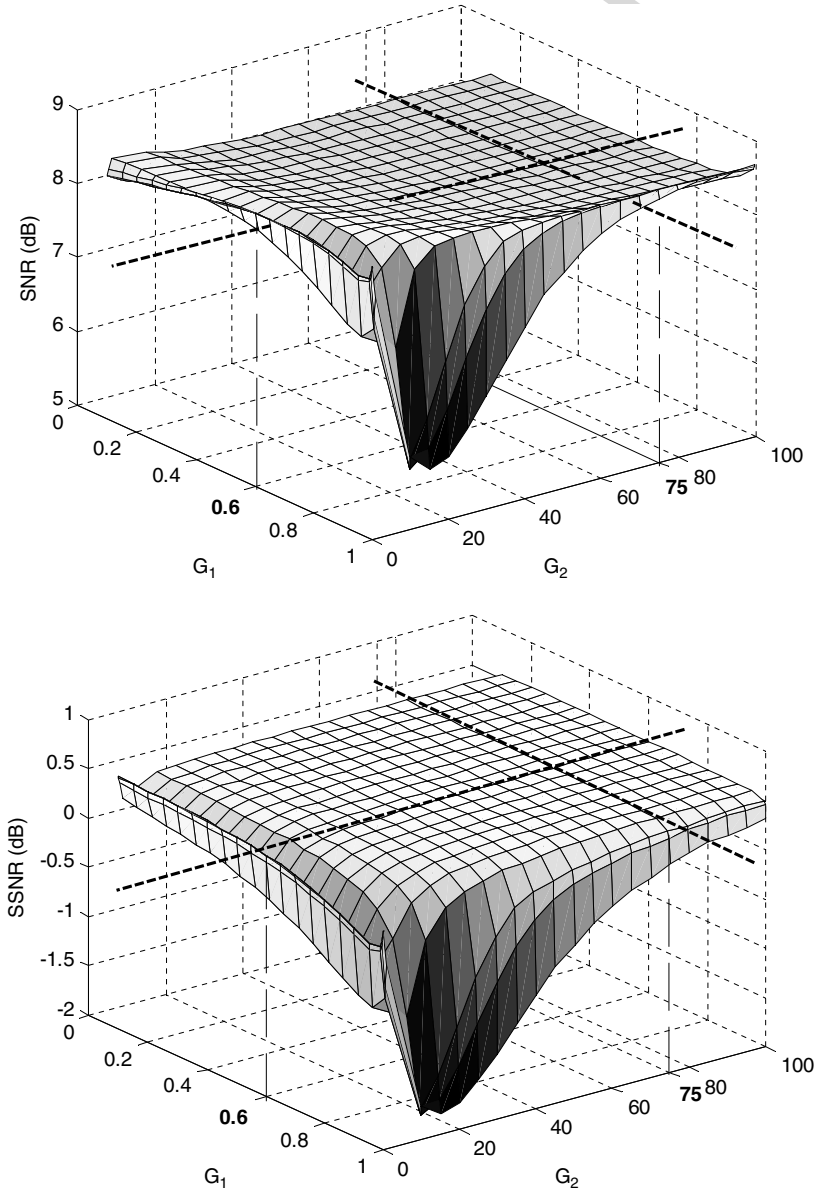

Fig. 6. SNR and SSNR versus G_1 and G_2 for AWGN with initial SNR = 0 db.

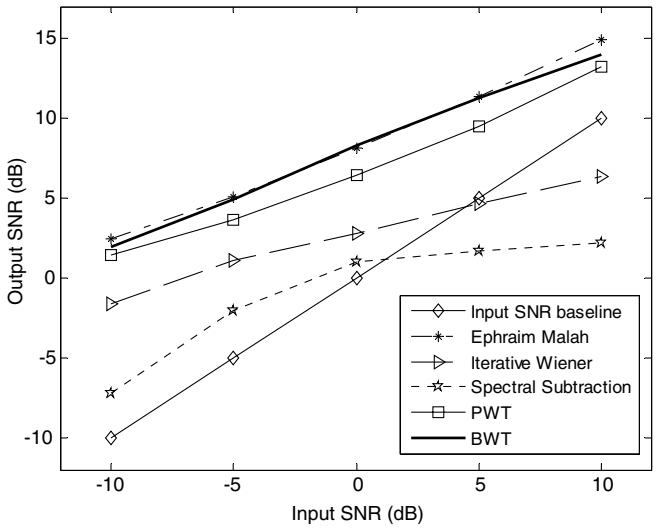

Fig. 7. SNR results for white noise case at −10, −5, 0, +5, and +10 dB SNR levels.

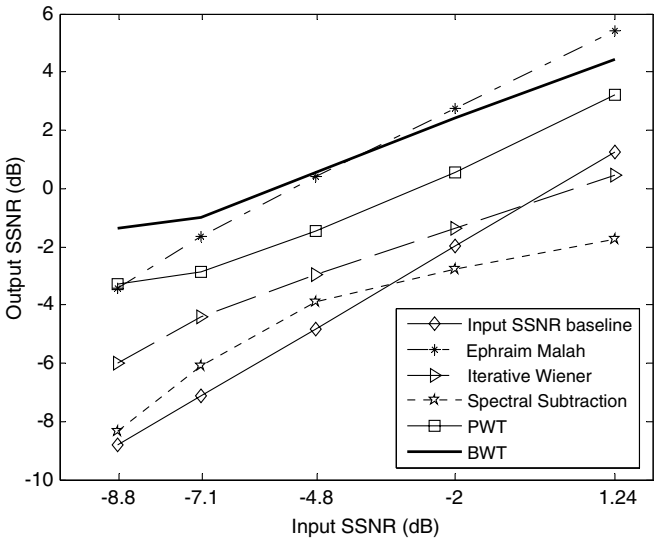

Fig. 8. SSNR results for white noise case at −10, −5, 0, +5, and +10 dB SNR levels.

subtraction, perceptually scaled WPT denoising (PWT), and the proposed BWT denoising (BWT), as well as the original noisy SNR and SSNR values for referencing degree of improvement. From these figures, the proposed BWT thresholding method and the Ephraim Malah filter clearly have the best performance for this noise condition. Each of these gave about 12 db improvement at the lower SNRs decreasing to about 8 db improvement at the higher SNRs. For SSNR, the improvement ranged from 6 to 9 db at the lower SNRs decreasing to 4–5 db at the higher SNRs. The BWT method shows the best SSNR improvement by a substantial margin at the −10 db noise case, obtaining nearly 3 db better performance than any of the other methods, and is better at the −5 db condition as well. Note that the other wavelet-based approach, the PWT method, was a close third behind the Ephraim Malah and BWT results, and in fact did slightly better than the Ephraim Malah method in terms of SSNR for the worst noise case. At the +5 db and +10 db SNR noise cases, the Ephraim Malah filter shows stronger performance than the other methods tested, including the proposed BWT method.

4.3.2. Realistic noise conditions at 0 db SNR

SNR and SSNR improvement across varying realistic noise conditions at 0 dB SNR are shown in Figs. 9 and 10. Here the results are given as net improvement, so that relative effectiveness can be seen for all four noise conditions as a function of enhancement method. Ephraim Malah filtering substantially outperforms the other methods in nearly all cases. The BWT thresholding approach outperforms the remaining three, but is competitive with Ephraim Malah only for the F-16 cockpit noise and pink

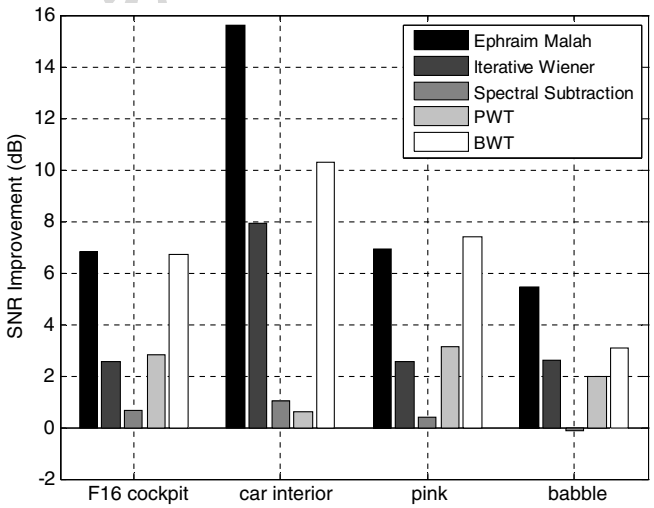

Fig. 9. SNR comparisons for varying noise conditions at 0 dB SNR.

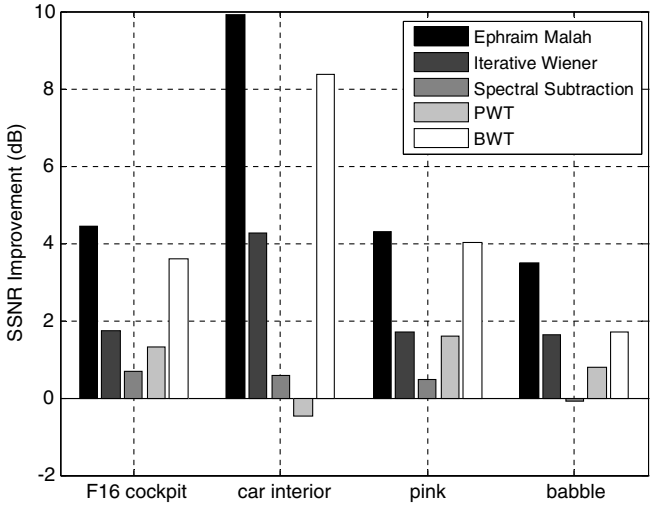

Fig. 10. SSNR comparisons for varying noise conditions at 0 dB SNR.

noise conditions. In a single instance, SSNR results for the pink noise condition, the BWT method outperforms Ephraim Malah filtering. One interesting observation is that the PWT perceptually scaled wavelet denoising approach does not compare as favorably for these realistic noise conditions as it did for the AWGN noise, with results that are closer to that of iterative Wiener filtering than to the BWT and Ephraim Malah methods, and demonstrating particularly poor performance in the car interior condition.

4.3.3. Mean opinion score results across all test conditions

Subjective results using Mean Opinion Scores (MOS) for these same conditions are shown in Fig. 11 and Table 1. Ten subjects were surveyed and asked to rate the sentences for quality, using a standard MOS scale where 5 indicates excellent quality and imperceptible distortion and 1 indicates unsatisfactory quality with annoying and objectionable distortion.

There are several interesting differences between the MOS results and the SSNR results. While the relative MOS for Ephraim Malah and BWT methods is in line with the corresponding SSNR values for the most part, both the iterative Wiener filter and the perceptually scaled wavelet denoising received much higher ratings than would have been suggested by SSNR. The net result is that the iterative Wiener filtering method was second to Ephraim Malah in overall opinion score, followed by the BWT and PWT methods. The most likely explanation for the differences between SSNR and MOS results (aside from the relatively low sample size of the MOS survey) is the presence or absence of short-term artifacts in the enhanced waveforms, which are known to have significant impact on perception but have only mild influence on SNR and SSNR values.

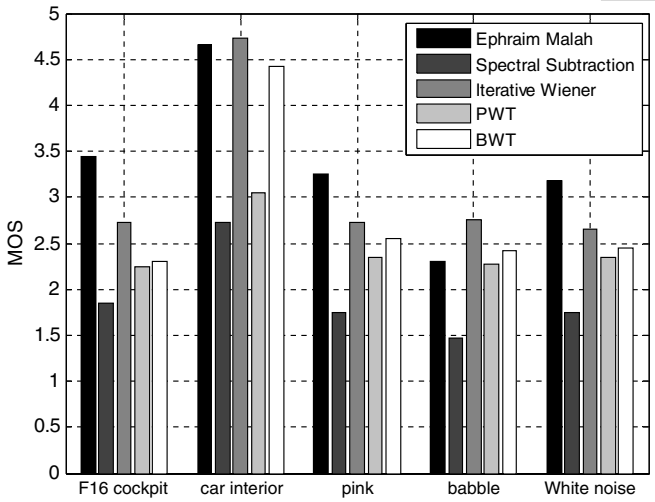

Fig. 11. MOS comparisons for varying noise conditions at 0 dB SNR.

Table 1
Average MOS by enhancement method

Method	Average MOS
Ephraim Malah	3.23
Iterative Wiener filter	3.03
Bionic wavelet transform (BWT)	2.80
Perceptual wavelet packet (PWT)	2.41
Spectral subtraction	2.07

5. Conclusions

A new method for speech signal enhancement using wavelets has been presented. This method is based on the Bionic Wavelet Transform and its incorporated Giguere–Woodland auditory model, resulting in a transform with perceptually motivated frequency scaling and adaptive time support at each scale corresponding to a non-linear model of the underlying cochlear system. Enhancement results demonstrate performance that is competitive with some of the best methods in the signal processing field, including Ephraim Malah filtering.

Future work on this approach will include the development of time-varying thresholding techniques based on tracking the *a priori* SNR, such as those used in the Ephraim Malah comparative method and in some recent perceptual wavelet denoising techniques (Cohen, 2001), as well as continued work in generalizing the BWT itself to enhancement and speech intelligibility improvement. Given the strong initial performance of the Bionic Wavelet Transform representation, continued work in this direction is expected to lead to additional improvement in overall signal enhancement.

References

Antoniadis, A., Oppenheim, G. (Eds.), 1995. Wavelets and Statistics. Springer-Verlag, New York.

Bahoura, M., Rouat, J., 2001. Wavelet speech enhancement based on the teager energy operator. IEEE Signal Process. Lett. 8 (1), 10–12.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics Speech Signal Process. 27, 113–120.

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

Cohen, I., 2001. Enhancement of speech using bark-scaled wavelet packet decomposition. Paper presented at the Eurospeech 2001, Denmark.

Daubechies, I., 1992. Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, Philadelphia.

Debnath, L., 2002. Wavelet Transforms and their Applications. Birkhauser, Boston.

Deller, J.R., Hansen, J.H.L., Proakis, J.G., 2000. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York.

Donoho, D.L., 1995. Denoising by soft thresholding. IEEE Trans. Inform. Theory 41 (3), 613–627.

Donoho, D.L., Johnstone, I.M., 1995. Adapting to unknown smoothness via wavelet shrinkage. J. Amer. Statist. Assoc. 90, 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 Process. ASSP-32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech Enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-33 (2), 443–445.

Fu, Q., Wan, E.A., 2003. Perceptual wavelet adaptive denoising of speech. Paper presented at the Eurospeech, Geneva.

Gabor, D., 1946. Theory of communications. J. Inst. Electr. Eng. Lond. 93, 429–457.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., et al., 1993. TIMIT Acoustic–Phonetic Continuous Speech Corpus: Linguistic Data Consortium.

Giguere, C. 1993. Speech processing using a wave digital filter model of the auditory periphery. Ph.D., University of Cambridge, Cambridge, UK.

Giguere, C., Woodland, P.C., 1994. A computational model of the auditory periphery for speech and hearing research. J. Acoust. Soc. Amer. 95 (1), 331–342.

Guo, D., Zhu, W., Gao, Z., Zhang, J., 2000. A study of wavelet thresholding denoising. Paper presented at the International Conference on Signal Processing, Beijing, PR China.

Haykin, S., 1996. Adaptive Filter Theory, third ed. Prentice Hall, Upper Saddle River, New Jersey.

Hu, Y., Loizou, P.C., 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Speech Audio Process. 12 (1), 59–67.

Jaffard, S., Meyer, Y., Ryan, R.D., 2001. Wavelets: Tools for Science and Technology. Society for Industrial and Applied Mathematics, Philadelphia.

Johnstone, I.M., Silverman, B.W., 1997. Wavelet threshold estimators for data with correlated noise. J. Roy. Statist. Soc., Ser. B (Gen.) 59, 319–351.

Lu, C.-T., Wang, H.-C., 2003. Enhancement of single channel speech based on masking property and wavelet transform. Speech Commun. 41 (2–3), 409–427.

The MathWorks Inc., 2003. Matlab.

Walnut, D.F., 2002. An Introduction to Wavelet Analysis. Birkhauser, Boston.

Yao, J., 2001. An active model for otoacoustic emissions and its application to time–frequency signal processing. Ph.D., The Chinese University of Hong Kong, Hong Kong.

Yao, J., Zhang, Y.T., 2001. Bionic wavelet transform: a new time–frequency method based on an auditory model. IEEE Trans. Biomed. Eng. 48 (8), 856–863.

Yao, J., Zhang, Y.T., 2002. The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations. IEEE Trans. Biomed. Eng. 49 (11), 1299–1309.

Zheng, L., Zhang, Y.-T., Yang, F.-S., Ye, D.-T., 1999. Synthesis and decomposition of transient-evoked otoacoustic emissions based on an active auditory model. IEEE Trans. Biomed. Eng. 46 (9), 1098–1106.