A generalized time–frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system

Shao, Yu; Chang, Chip Hong

2007

Shao, Y., & Chang, C. H. (2007). A generalized time–frequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 37(4), 877-889.

<https://hdl.handle.net/10356/91337>

<https://doi.org/10.1109/TSMCB.2007.895365>

IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics © copyright 2007 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder. <http://www.ieee.org/portal/site>.

A Generalized Time–Frequency Subtraction Method for Robust Speech Enhancement Based on Wavelet Filter Banks Modeling of Human Auditory System

Yu Shao and Chip-Hong Chang, *Senior Member, IEEE*

Abstract—We present a new speech enhancement scheme for a single-microphone system to meet the demand for quality noise reduction algorithms capable of operating at a very low signal-to-noise ratio. A psychoacoustic model is incorporated into the generalized perceptual wavelet denoising method to reduce the residual noise and improve the intelligibility of speech. The proposed method is a generalized time–frequency subtraction algorithm, which advantageously exploits the wavelet multirate signal representation to preserve the critical transient information. Simultaneous masking and temporal masking of the human auditory system are modeled by the perceptual wavelet packet transform via the frequency and temporal localization of speech components. The wavelet coefficients are used to calculate the Bark spreading energy and temporal spreading energy, from which a time–frequency masking threshold is deduced to adaptively adjust the subtraction parameters of the proposed method. An unvoiced speech enhancement algorithm is also integrated into the system to improve the intelligibility of speech. Through rigorous objective and subjective evaluations, it is shown that the proposed speech enhancement system is capable of reducing noise with little speech degradation in adverse noise environments and the overall performance is superior to several competitive methods.

Index Terms—Auditory masking, noise reduction, speech enhancement, wavelet.

I. INTRODUCTION

AUTOMATIC speech processing systems are increasingly employed in real-world applications, such as hands-free car kits, communicators in noisy cockpits, and voice-activated dialers for cellular phones [1], [18], [23], [24], [33], [35], but the performance of these devices in many practical situations degrades drastically when confronted with a great diversity of adverse noise conditions such as background noise and microphone distortions. Thus, there is a strong demand for quality noise reduction algorithms capable of operating at very low signal-to-noise ratio (SNR) in order to combat various forms of noise distortion. Many solutions have been developed to deal with the noise reduction problem. Generally, these solutions can be classified into two main areas: nonparametric and statistical-model-based speech enhancements. Methods from the first

category usually remove an estimate of the distortion from the noisy features, such as subtractive-type algorithms [34], and wavelet denoising [8]. The second category, statistical-model-based speech enhancement [10], utilizes a parametric model of the signal generation process. The parametric model normally describes the predictable structures and the expected patterns in the signal process [6].

In this paper, the proposed speech enhancement system is based on subtractive-type algorithms. It estimates the short-time spectral magnitude of speech by subtracting the noise estimation from the noisy speech. The main attractions of spectral subtraction are: 1) its relative simplicity, in that it only requires an estimate of the noise power spectrum, and 2) its high flexibility against subtraction parameters variation. From the basic approach of spectral magnitude subtraction introduced by Boll [4], many derivatives have been developed, such as power spectral subtraction [3], nonlinear spectral subtraction [34], generalized spectral subtraction [26], multiband spectral subtraction [13], spectral subtraction based on priori SNR [11], spectral subtraction based on the masking properties of the human auditory system [33], and single-channel speech enhancement based on a recurrent neural network [15]. However, the main problem in spectral subtraction is the presence of processing distortions caused by the random variations of noise. Musical residual noise is introduced in subtractive-type denoising methods, which will degrade the perceptibility of the processed speech. To reduce the effect of residual noise, a number of methods were considered: Boll [4] proposed magnitude averaging; Berouti *et al.* [3] proposed the oversubtraction of noise and defined a spectral floor to make residual noise inaudible; McAulay and Malpass [21] proposed soft-decision noise suppression filtering; Ephraim and Malah [9] proposed a minimum mean-square-error short-time spectral amplitude estimator; Vaseghi [34] introduced nonlinear spectral subtraction; Virag [33] made use of masking properties of the human auditory system to reduce the effect of residual noise; and Hu and Loizou [12] proposed a speech enhancement method by incorporating the psychoacoustical model in the frequency domain. However, the performances of the above methods are not satisfying in adverse environments, particularly when the SNR is very low. The reason is that in very low SNR conditions, it is difficult to suppress noise without degrading intelligibility, and without introducing residual noise and speech distortion. Therefore, processed speech can be even less perceptible than the original noisy speech. The methods that adopt the masking property of the human auditory system can reduce the effect of residual noise, but the drawback is the large computational

Manuscript received May 3, 2006; revised October 17, 2006. This paper was recommended by Associate Editor C.-T. Lin.

The authors are with the Centre for High Performance Embedded Systems, Nanyang Technological University, Singapore 63755 (e-mail: echchang@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSMCB.2007.895365

effort associated with the subband decomposition and the additional FFT analyzer required for psychoacoustic modeling. Furthermore, they do not make use of the temporal masking models.

In order to simplify the mapping of audio signals into a scale that could well preserve the time–frequency related information, the wavelet packet transform (WPT) has been incorporated recently in some proposed speech and audio coding systems [20]. Sinha and Tewfik [25] first introduced adapted wavelets to transparent audio compression; Srinivasan and Jamieson [29] proposed a high-quality audio compression using the adaptive wavelet packet decomposition and psychoacoustic modeling; Bahoura and Rouat [5] proposed modified wavelet speech enhancement based on the teager energy operator. This technique does not require an explicit estimation of the noise level or *a priori* knowledge of the SNR, which is usually needed in many popular speech enhancement methods. Li *et al.* [14] proposed the perceptual time–frequency subtraction algorithm for noise reduction in hearing aids. Veselinovic and Graupe [32] introduced a wavelet transform approach to the blind adaptive filtering of speech from unknown noises. It can reduce the noise effect well, but the complexity is high. Lu and Wang [16] applied the critical band WPT to speech enhancement. Ayat *et al.* [2] proposed a wavelet-based speech enhancement using the step-garrote thresholding algorithm. However, the qualities of the processed speech of the above methods degrade as the distortion is introduced and the high-frequency speech signals are grievously reduced by wavelet transform.

The objective of this paper is to introduce a versatile speech enhancement method based on the human auditory model. The emphases are on the reduction of the effect of residual noise and speech distortion in the denoising process and the enhancement of the denoised speech in high frequency to improve its intelligibility. The proposed speech enhancement system consists of two main functions. One is a generalized perceptual time–frequency subtraction (GPTFS) method based on the masking properties of the human auditory system. The GPTFS works in conjunction with a perceptual WPT (PWPT) to reduce the effect of noise contamination. The other is an unvoiced speech enhancement (USE), which tunes a set of weights in high-frequency subbands to improve the intelligibility of the processed speech. The following contributions have been made.

The first contribution is the use of PWPT to approximate 24 critical bands of the human auditory system up to 16 kHz. It enables the components of a complex sound to be appropriately segregated in frequency and time in order to mimic the frequency selectivity and temporal masking of the human auditory system. Previous method used WPT to approximate critical bands only up to the range of 8 kHz for telephone quality speech [25]. Critical high-frequency components in noisy speech have been analyzed by our proposed PWPT to improve the perceptual quality of the final processed speech. The second contribution is a parametric formulation of subtractive noise reduction based on the generalized perceptual wavelet transform. We refers to the Fourier transform-based gain function of the generalized spectral subtraction method [33] and derive the close form expressions for the subtraction factor and noise flooring factor, respectively, to optimize the tradeoff in the simultaneous reduction of background noise, residual noise and speech distortion. These parameters of GPTFS can

then be adaptively tuned according to the noise level and the masking thresholds derived from the human auditory model in wavelet domain. Finally, a new system structure for speech enhancement is developed to integrate GPTFS and USE in wavelet domain. The ability to process signal in both time and frequency domains of the USE in the wavelet domain makes the discrimination of the original speech in high-frequency bands from noise possible, hence further improve the intelligibility of speech.

The rest of this paper is structured as follows. In Section II, a perceptual wavelet filter bank architecture is proposed to approximate the critical band division of the human auditory system. In Section III, simultaneous and temporal maskings are unified and mapped onto the wavelet domain. These masking properties are then encapsulated in a parametric formulation of GPTFS algorithm derived in Section IV. An adaptive speech enhancement system incorporating both GPTFS algorithm and USE is presented. Rigorous evaluations and comparisons of the proposed method against other speech enhancement techniques under a wide variety of adverse conditions have been performed, using both objective and subjective measures. The results are presented and discussed in Section V. Section VI concludes this paper.

II. PROPOSED ARCHITECTURE FOR PERCEPTUAL WAVELET FILTER BANK

To design effective algorithm for enhancing speech, it needs a well build psychoacoustic model of the ear, which has an unsurpassed capability to adapt to noise. In this section, we propose a new human auditory model that adopts the base structure of traditional auditory model but replace the time-invariant bandpass filters with WPT in order to mimic the time–frequency analysis of the critical bands according to the hearing characteristics of human cochlea.

A PWPT is used to decompose the speech signal from 20 Hz to 16 kHz into 24 frequency subbands that approximate the critical bands. The decomposition is implemented with an efficient seven-level tree structure. The analysis filter bank is depicted in Fig. 1. In the proposed wavelet packet decomposition, the two-channel wavelet filter banks are used to split both the low-pass and high-pass bands [28] as opposed to only the decomposition of low-frequency bands in the usual wavelet decomposition. This binary tree decomposition is attractive due to the following reasons [30]. First, the smoothness property of wavelet is determined by the number of vanishing moments. If a wavelet with a large number of vanishing moments is used, the stringent bandwidth and stopband attenuation of each subband, as specified by the human auditory model, can be more closely approximated by the wavelet decomposition. Second, the psychoacoustic study of human ears suggests that a frequency to bark transformation needs to be performed to accurately model the frequency dependent sensitivity of human ears. Such a transformation is accomplished in audio processing systems by dividing the frequency range into critical bands.

The 24 critical bands are derived from the perfect reconstruction filter bank with finite-length filters [28] using different wavelets for the analysis and synthesis scaling functions. Let $H(z)$ and $G(z)$ be the lowpass (LP) and highpass (HP) transfer functions, respectively, before the decimation-by-two operation

Fig. 1. Perceptual wavelet packet decomposition tree.

in each stage of the analysis filter bank. Further, let $F(z)$ and $J(z)$ be the LP and HP transfer functions, respectively, after the upsampling-by-two operation in each stage of the synthesis filter bank. The analysis and synthesis filter banks are related by [28]

$$\begin{aligned} g(n) &= (-1)^n f(n) \leftrightarrow G(z) = F(-z) \\ j(n) &= -(-1)^n h(n) \leftrightarrow J(z) = -H(-z). \end{aligned} \quad (1)$$

The relationship between the LP and HP filters reduces the number of filters to be implemented for each stage of the two-channel filter bank by half. Once the LP filters, $H(z)$ and $F(z)$ are designed, the HP filters, $G(z)$ and $J(z)$ can be derived from (1).

There are 24 critical bands in the hearing range of 0–16 kHz. Owing to the property of frequency selectivity related to critical band, temporal resolution of the human ear, and regularity property of wavelets [25], Daubechies wavelet basis is chosen as prototype filter and a seven-stage WPT (the number of stages $j_{\max} = 7$, the frame length, $F_{j_{\max}} = 128$, the frame length at stage j is given by $F_j = 2^j$) is adopted to build perceptual wavelet filterbank. W_{jk} , represents WPT coefficient, where k is the coefficient number; j is the transform stage from which W_{jk} is chosen; l is the number of “temporal” coefficients in the critical band. Table I shows the mapping of the PWPT coefficients in each stage.

Table II shows the comparison of lower (f_L) and upper (f_u) frequencies, center frequency (f_c) and bandwidth (Δf) in hertz between the critical band rate and the proposed perceptual wavelet packet tree scale.

TABLE I
PERCEPTUAL WAVELET FILTER-BANKS COEFFICIENTS

Subband Z_W	l	Coefficients k_a-k_b	Transform stage j
1	1	0-0	7
2	1	1-1	7
3	1	2-2	7
4	1	3-3	7
5	1	4-4	7
6	1	5-5	7
7	1	6-6	7
8	1	7-7	7
9	2	8-9	6
10	2	10-11	6
11	2	12-13	6
12	2	14-15	6
13	2	16-17	6
14	2	18-19	6
15	4	20-23	5
16	4	24-27	5
17	4	28-31	5
18	8	32-39	4
19	8	40-47	4
20	8	48-55	4
21	8	56-63	4
22	16	64-79	3
23	16	80-95	3
24	32	96-127	2

* k_a and k_b are the coefficient indices of the first and last transform coefficients within a given critical band.

TABLE II
CRITICAL BAND RATE Z AND PERCEPTUAL WAVELET FILTER-BANKS W

Z	Bark Scale			Wavelet Scale		
	$[f_l f_u]$	f_c	Δf	$[f_l f_u]$	f_c	Δf
1	[0 100]	50	100	[0 125]	62.5	125
2	[100 200]	150	100	[125 250]	187.5	125
3	[200 300]	250	100	[250 375]	312.5	125
4	[300 400]	350	100	[375 500]	437.5	125
5	[400 510]	450	110	[500 625]	562.5	125
6	[510 630]	570	120	[625 750]	687.5	125
7	[630 770]	700	140	[750 875]	812.5	125
8	[770 920]	840	150	[875 1000]	937.5	125
9	[920 1080]	1000	160	[1000 1250]	1125	250
10	[1080 1270]	1170	190	[1250 1500]	1375	250
11	[1270 1480]	1370	210	[1500 1750]	1625	250
12	[1480 1720]	1600	240	[1750 2000]	1875	250
13	[1720 2000]	1850	280	[2000 2250]	2125	250
14	[2000 2320]	2150	320	[2250 2500]	2375	250
15	[2320 2700]	2500	380	[2500 3000]	2750	500
16	[2700 3150]	2900	450	[3000 3500]	3250	500
17	[3150 3700]	3400	550	[3500 4000]	3750	500
18	[3700 4400]	4000	700	[4000 5000]	4500	1000
19	[4400 5300]	4800	900	[5000 6000]	5500	1000
20	[5300 6400]	5800	1100	[6000 7000]	6500	1000
21	[6400 7700]	7000	1300	[7000 8000]	7500	1000
22	[7700 9500]	8500	1800	[8000 10000]	9000	2000
23	[9500 12000]	10500	2500	[10000 12000]	11000	2000
24	[12000 15500]	13500	3500	[12000 16000]	14000	4000

Fig. 2 depicts the difference in the critical band rate between the critical bands and the proposed perceptual wavelet packet tree structural bands. The referenced critical band rate Z_B in Bark is approximated by [36]

$$Z_B = 13 \tan^{-1}(7.6 \times 10^{-4}f) + 3.5 \tan^{-1}(1.33 \times 10^{-4}f)^2 \quad (2)$$

where the frequency, f , is measured in hertz.

Fig. 2. Center frequencies of critical bands and perceptual wavelet packet decomposition tree.

Fig. 3. Bandwidths of critical bands and perceptual wavelet packet decomposition tree.

The bandwidths of the critical bands and the perceptual wavelet packet tree are plotted in Fig. 3. From Fig. 3, it is evident that the critical bands have constant width at approximately 100 Hz for center frequencies up to 500 Hz, and the bandwidths increase as the center frequency increases further. The referenced critical bandwidth (CBW) in hertz is calculated by [36]

$$\text{CBW}(f) = 25 + 75(1 + 1.4 \times 10^{-6} f^2)^{0.69}. \quad (3)$$

Fig. 4 compares the absolute threshold of hearing (ATH) in hertz, critical band scale, and perceptual wavelet packet scale. The ATH characterizes the amount of energy needed in a pure tone such that it can be detected by a listener in a noiseless environment. The referenced absolute threshold is expressed in terms of sound pressure level (SPL) in decibels, and it is plotted in Fig. 4 with the following equation [36]:

$$\text{ATH}_{\text{SPL}}(f) = 3.64f^{-0.8} - 6.5e^{-0.6(f-3.3)^2} + 0.001f^4. \quad (4)$$

From the results tabulated in Table II and plotted in Figs. 2–4, it is evident that the proposed perceptual wavelet packet tree can closely mimic the experiential critical bands. The parameters of the discrete WPT filter used to derive the plots of Figs. 2–4 are determined based on the auditory masking properties, and they will be discussed in the next section.

Fig. 4. ATH in (a) frequency, (b) bark, and (c) perceptual wavelet packet tree scales.

III. UNIFICATION OF SIMULTANEOUS AND TEMPORAL MASKINGS IN WAVELET PACKET DOMAIN

The notion of a critical band is related to the phenomenon of auditory masking, whereby the perception of one sound is obscured by the presence of another in frequency or in time. There are two dominant types of masking. The simultaneous masking is the masking between two concurrent sounds, and it is best modeled in the frequency domain. The temporal masking is the masking of other sounds immediately preceding or following a sudden stimulus sound, and it is usually modeled in the time domain. After studying the limitations of a traditional auditory model using a Fourier transform, we establish a unification of the simultaneous and temporal maskings by means of PWPT.

In subtractive-type of speech enhancement, the maskee is the residual noise. To make it inaudible, it should lie below the masking threshold or threshold of just noticeable distortion [17]. Clean speech is ideal for the estimation of masking thresholds. When clean speech is not accessible, the enhanced speech processed by the spectral subtraction gives reasonably good estimate of the clean signal. Since masking is manifested by a shift of auditory threshold in signal detectability, and loudness is an important attribute of auditory perception, the noise masking threshold can be determined by the characteristics of the masker and maskee, the SPL and the frequency of the masker. The noise masking threshold for a residual noise maskee in each critical band can be calculated as follows.

A. Time and Frequency Analysis Along the Critical Band Scale

Our goals are to approximate the critical band analysis and estimate the auditory masking threshold using the PWPT. The resulting 24-band wavelet packet approximation of the critical band rate and the bandwidth approximation have been shown in Figs. 2 and 3. The input signal is decomposed into critical bands scale for time–frequency analysis. The outputs of the perceptual wavelet filterbank are squared to obtain the Bark spectrum energy and temporal energy, respectively. The

Bark energy spectrum $E_{\text{Bark}}(Z_W)$ with $1 \leq Z_W \leq 24$ is computed as

$$E_{\text{Bark}}(Z_W) = \sum_{k=k_a}^{k_b} W_{j,k}^2(Z_W) \quad (5)$$

where k_a and k_b are the coefficient indexes of the first and last transform coefficients referring to Table I. The temporal energy $E_{\text{temp}}(n)$ with $n = 0, 1, \dots, l(Z_W) - 1$ in each critical band Z_W can be derived as

$$E_{\text{temp}}(n) = W_{j,(k_a+n)}^2(Z_W). \quad (6)$$

B. Convolution With Spreading Function

In this step, the Bark spectrum energy and temporal energy are convoluted with the spreading function, respectively. The spreading of Bark energy $E'_{\text{Bark}}(Z_W)$ is calculated by the convolution of $E_{\text{Bark}}(Z_W)$ with the linear version of $\text{SF}(Z_W)$ [36]

$$E'_{\text{Bark}}(Z_W) = \frac{1}{l(Z_W)} \sum_{m=1}^{24} E_{\text{Bark}}(m) \text{SF}(Z_W - m) \quad (7)$$

where the spreading function $\text{SF}(Z_W)$ in decibels models the interaction of the masking signals between different critical bands and can be deduced as follows [27]:

$$\text{SF}(Z_W) = a + \frac{\nu + u}{2} (Z_W - Z_W(k) + c) - \frac{\nu - u}{2} \sqrt{d + (Z_W - Z_W(k) + c)^2} \quad (8)$$

where u and ν are, respectively, the upper and lower slopes in decibels per Bark; d is the peak flatness; a and c are the compensation factors needed to satisfy $\text{SF}(0) = 1$. $k \in [1, 24]$ is an integer critical band identifier. The critical band rate distance $Z_W - Z_W(k)$ varies from -24 to 24 Bark. The parameters of the spreading function, $\text{SF}(Z_W)$, which are empirically determined by the listening test, were set to $\nu = 30$ dB/Bark, $u = -25$ dB/Bark, $d = 0.3$, $c = 0.05$, and $a = 27.4$.

The temporal masking can be modeled in terms of temporal energy spreading across time. The temporal spreading energy $E'_{\text{temp}}(n)$ with $n = 0, 1, \dots, l(Z_W) - 1$ can be obtained by the convolution of the linear temporal spreading function $\text{SF}(n)$ with the temporal energy, whose calculation is similar to (7)

$$E'_{\text{temp}}(n) = \sum_{m=0}^{l(Z_W)-1} E_{\text{temp}}(m) \text{SF}\left(n - \frac{N}{l(Z_W)}m\right) \quad (9)$$

where $\text{SF}(n)$ is a parameterized spreading function similar to (8) and can be deduced as follows:

$$\text{SF}(n) = a + \frac{\nu + u}{2} (n - n(k) + c) - \frac{\nu - u}{2} \sqrt{d + (n - n(k) + c)^2}. \quad (10)$$

Corresponding to the minimum frame duration F_{min} , the parameters of the spreading function $\text{SF}(n)$ in decibels employed were set to $\nu = 20$ dB/ F_{min} (dB/ms), $u = -10$ dB/

F_{min} (dB/ms), $d = 0.3$, $c = 0.2$, and $a = 7.9$ after several listening tests.

C. Calculating the Relative Threshold Offset

Next, a relative threshold offset $O(Z_w)$ is subtracted from each critical band. It is required to distinguish between tonelike and noiselike components in speech to calculate the relative threshold offset. There are two types of noise masking threshold: the threshold of tone masking noise T_{tmn} , where $T_{\text{tmn}}(Z_W) = E'_{\text{Bark}} - 14.5 - Z_W$ dB, and the threshold of noise masking tone T_{nmt} , where $T_{\text{nmt}}(Z_W) = E'_{\text{Bark}} - 5.5$ dB. To determine whether the signal is tonelike or noiselike, the spectral flatness measure (SFM) is used. From this value, a tonality coefficient τ [17] is generated

$$\tau = \min(\text{SFM}_{\text{dB}}/\text{SFM}_{\text{dB min}}, 1) \quad (11)$$

where τ is a ratio of the SFM_{dB} , which is computed at the last stage of the decomposition to the $\text{SFM}_{\text{dB min}}$, which is an entirely tonelike reference over a frame. Using τ , the relative threshold offset $O(Z_w)$ in decibels [17] for each critical band is calculated as

$$O(Z_W) = \tau \cdot (14.5 + Z_W) + (1 - \tau) \cdot 5.5. \quad (12)$$

D. Renormalization and Comparison With the ATH

To unify the temporal and simultaneous maskings into one formulation, a temporal masking factor $\rho_{j,k}(Z_W)$ is introduced as follows:

$$\rho_{j,k}(Z_W) = E'_{\text{temp}}(k - k_a)/W_{j,k}^2(Z_W) \quad (13)$$

where $\rho_{j,k}(Z_W) \geq 1$ indicates the extent of temporal masking. If $\rho_{j,k}(Z_W) = 1$, no temporal masking is induced, otherwise, if $\rho_{j,k}(Z_W) > 1$, the temporal masking has occurred. By combining the simultaneous masking threshold $E'_{\text{Bark}}(Z_W)$ and the temporal masking factor $\rho_{j,k}(Z_W)$, the time-frequency masking threshold in the critical band Z_W can be determined as follows:

$$M_{j,k}(Z_W) = 10^{\log[E'_{\text{Bark}} \cdot \rho_{j,k}(Z_W)] - [O(Z_W)/10]}. \quad (14)$$

Finally, $M_{j,k}(Z_W)$ is compared to the ATH in each critical band, and the maximum of each value is retained as $T_{j,k}(Z_W)$

$$T_{j,k}(Z_W) = \max(M_{j,k}(Z_W), \text{ATH}(Z_W)). \quad (15)$$

The unification of the simultaneous and temporal masking properties of the human auditory system forms the basis of our GPTFS technique. It works in conjunction with the PWPT processor and an USE unit toward a speech enhancement system.

IV. PROPOSED ADAPTIVE SPEECH ENHANCEMENT SYSTEM

Variants of spectral subtraction method with varying subtraction rules have been derived from different criteria [34].

Fig. 5. Architecture of the proposed speech enhancement system.

Most of these algorithms possess parametric forms to improve the agility of their estimators. The focus is mainly on the error function used to reduce the musical residual noise in spectral domain. These methods suffer from inevitable loss of some vital information that degrades the auditory perception of enhanced speech. In this section, we develop a new GPTFS method based on the PWPT and the human auditory perception. Its parametric formulation is derived from the basis of the generalized Fourier spectral subtraction algorithm of [34]. The GPTFS algorithm incorporates most of the basic subtraction rules and realizes the subtraction in a broader time–frequency domain. More crucial information has been preserved than in Fourier transform domain, leading to better perceptual outputs than existing subtraction algorithms.

The block diagram of the proposed speech enhancement system is shown in Fig. 5. After the noisy signal $x[n]$ is decomposed by PWPT, the transform sequence is enhanced by a subtractive-type algorithm to produce the rough speech estimate. This estimate is used to calculate a time–frequency masking threshold. Using this masking threshold, a new subtraction rule, which is masking dependent, is designed to compute an estimation of the original speech. This approach assumes that the high-energy frames of speech will partially mask the input noise (high masking threshold), hence reducing the need for a strong enhancement mechanism. On the other hand, frames containing less speech (low masking threshold) will undergo an overestimated subtraction. To further improve the intelligibility of processed speech, an USE is applied. Finally, the processed speech is reconstructed by the inverse PWPT. The noise estimation is assumed to be available and is performed during speech pauses. The speech pause detection algorithm from Marzinik and Kollmeier [20] is adopted for the noise spectrum estimation by tracking the power envelope dynamics. The speech pause detection algorithm has been extended for subband processing on our proposed system.

A. Generalized Perceptual Time–Frequency Subtraction

Consider a noisy speech signal which consists of the clean speech additively degraded by noise

$$x[n] = s[n] + \nu[n] \quad (16)$$

where $x[n]$ and $s[n]$ are the noisy and original speech, respectively. $\nu[n]$ is the additive noise, which can be white or colored. Original speech and noise are assumed to be uncorrelated.

To capture the localized information of transient signal, the PWPT is applied to the noisy input speech, i.e., $\{w_{j,k}(x)\} = \text{PWPT}(x[n])$

$$w_{j,k}(x) = w_{j,k}(s) + w_{j,k}(\nu), \quad \forall j=0, 1, \dots, j_{\max}-1 \\ k=0, 1, \dots, 2^j-1 \quad (17)$$

where $w_{j,k}(x)$, $w_{j,k}(s)$, and $w_{j,k}(\nu)$ are the wavelet transform coefficients of a noisy signal, clean signal, and noise, respectively. (j, k) in the subscript of w corresponds to its scale and translation indexes. $j_{\max}$ is the maximum number of levels of wavelet decomposition.

By applying the power spectral subtraction method [34] in the wavelet domain, the estimated power of the enhanced speech is given by

$$|\tilde{w}_{j,k}(s)|^2 = \begin{cases} |w_{j,k}(x)|^2 - |\tilde{w}_{j,k}(\nu)|^2, & \text{if } |w_{j,k}(x)|^2 > |\tilde{w}_{j,k}(\nu)|^2 \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

where $|\tilde{w}_{j,k}(s)|^2$, $|w_{j,k}(x)|^2$, and $|\tilde{w}_{j,k}(\nu)|^2$ are the estimates of the power wavelet coefficients of the noise-suppressed speech signal, the noisy speech and the estimated noise, respectively, in the wavelet domain.

Since there is no prior knowledge of the parameters of $s[n]$ and $\nu[n]$, and only $x[n]$ is accessible, the noise power spectrum in (18) is to be estimated from the time-average noise magnitude during the period in which the speech signal is absent. The time-average noise spectrum can be obtained by [34]

$$|\tilde{w}_{i,j,k}(\nu)|^2 = 0.9 \cdot |w_{i-1,j,k}(\nu)|^2 + 0.1 \cdot |w_{i-1,j,k}(x)|^2 \quad (19)$$

where the subscript i is the frame index.

The design takes into account that human speech is based on mixing voiced and unvoiced phonemes over time. The voiced phonemes last for 40–150 ms and their power concentrates at below 1.5 kHz, whereas the unvoiced phonemes last for 10–50 ms and their power concentrates at around 2–8 kHz. Hence, $s[n]$ is nonstationary over time interval of above 250 ms, and the noise $\nu[n]$ is assumed to be piecewise-stationary or more stationary than $s[n]$, which is valid for most environmental noise encountered. With the above assumption, the input signal $x[n]$ is divided into 128 samples.

Substituting (17) into (18), we have

$$\begin{aligned} |\tilde{w}_{j,k}(s)|^2 &= |w_{j,k}(s) + w_{j,k}(\nu)|^2 - |\tilde{w}_{j,k}(\nu)|^2 \\ &= |w_{j,k}(s)|^2 + \underbrace{(|w_{j,k}(\nu)|^2 - |\tilde{w}_{j,k}(\nu)|^2)}_{\text{Noise Variations}} \\ &\quad + \underbrace{w_{j,k}^*(s)w_{j,k}(\nu) + w_{j,k}(s)w_{j,k}^*(\nu)}_{\text{Cross Products}}. \end{aligned} \quad (20)$$

The wavelet power spectrum of the estimated speech signal includes error terms due to the nonlinear mapping of spectral estimates, the variations of the instantaneous noise power about the estimated mean noise power, and the cross products of the original speech and noise. Since signal and noise are assumed to

be uncorrelated, and the noise wavelet coefficients are assumed to have zero mean, the expected value of the power spectrum of the estimated speech signal in the wavelet domain is given by

$$\begin{aligned} E[|\tilde{w}_{j,k}(s)|^2] &= E \left[|w_{j,k}(s)|^2 + \underbrace{(|w_{j,k}(\nu)|^2 - |\tilde{w}_{j,k}(\nu)|^2)}_{\text{Noise Variations}} \right. \\ &\quad \left. + \underbrace{w_{j,k}^*(s)w_{j,k}(\nu) + w_{j,k}(s)w_{j,k}^*(\nu)}_{\text{Cross Products}} \right] \\ &= E[|w_{j,k}(s)|^2]. \end{aligned} \quad (21)$$

Equation (21) shows that the expected value of the instantaneous power spectrum of estimated speech signal converges to that of the noise-free signal. However, it is noted that for nonstationary signals, such as speech, the objective is to recover the instantaneous or short-time signal, and only a relatively small amount of averaging can be applied. Too much averaging will obscure the temporal evaluation of the signal. The spectral subtraction in (18) can be deemed as an adaptive attenuation of noisy speech according to the *posteriori* SNR of the noisy speech signal to the estimated noise. A gain function, $\text{Gain}(j, k)$ for the zero-phase wavelet spectral subtraction filter can be defined such that its magnitude response is in the range between 0 and 1, i.e.,

$$\tilde{w}_{j,k}(s) = \text{Gain}(j, k) \cdot w_{j,k}(x), \quad 0 \leq \text{Gain}(j, k) \leq 1. \quad (22)$$

Comparing (18) and (22), we have

$$\text{Gain}(j, k) = \sqrt{1 - \frac{|\tilde{w}_{j,k}(\nu)|^2}{|w_{j,k}(x)|^2}} = \sqrt{1 - \frac{1}{\text{SNR}_{\text{post}}(j, k)}} \quad (23)$$

where $\text{SNR}_{\text{post}}(j, k) = |w_{j,k}(x)|^2 / |\tilde{w}_{j,k}(\nu)|^2$ is the *posteriori* SNR, which is defined as the ratio of the wavelet power spectrum of the noisy speech signal to that of the estimated noise. In the low SNR case, the gain is set to zero when the wavelet power spectrum of the noisy speech is less than that of the estimated noise.

The gain function in (23) acts as a *posteriori* SNR-dependent attenuator in the time-frequency domain. The input noisy speech is attenuated more heavily with decreasing *posteriori* SNR and vice versa with increasing *posteriori* SNR. This filtering function can be readily implemented since the *posteriori* SNR can be easily measured on the noisy speech. Although it can effectively filter out the background noise from the noisy speech, the processing itself introduces musical residual noise. This processing distortion is due to the random variation of noise spectrum and the nonlinear mapping of the negative or small-valued spectral estimate. To make this residual noise “perceptually white,” Virag [33] introduced several flexible subtraction parameters. These subtraction parameters are chosen to adapt to a criterion associated with the human auditory perception. The generalized spectral subtraction method in [33] is based on Fourier transform. It considers only the simultaneous masking property of human auditory system, and

the adaptation parameters are empirically determined. In what follows, we will state the generalized time-frequency subtraction method in the PWPT domain and derive the close form expressions for its optimal adaptation parameters.

From (23), the gain function of Virag’s generalized spectral subtraction algorithm [33] can be reexpressed in terms of perceptual wavelet spectral coefficients as follows:

$$\text{Gain}'(j, k) = \begin{cases} \left(1 - \alpha \cdot \left[\frac{\tilde{w}_{j,k}(\nu)}{w_{j,k}(x)}\right]^{\gamma_1}\right)^{\gamma_2}, & \text{if } \left[\frac{\tilde{w}_{j,k}(\nu)}{w_{j,k}(x)}\right]^{\gamma_1} < \frac{1}{\alpha + \beta} \\ \left(\beta \cdot \left[\frac{\tilde{w}_{j,k}(\nu)}{w_{j,k}(x)}\right]^{\gamma_1}\right)^{\gamma_2}, & \text{otherwise.} \end{cases} \quad (24)$$

The PWPT-based gain function of (24) possesses the same set of free parameters as that of [33], and their effects are summarized as follows. 1) The subtraction factor α ($\alpha \geq 1$) controls the amount of noise subtracted from the noisy signal. For full noise reduction, $\alpha = 1$ whereas for oversubtraction $\alpha > 1$. Oversubtraction allows the time-frequency spectrum to be attenuated more than necessary. This factor should be appropriately selected to provide the best tradeoff between residual noise peaks and audible distortion. 2) The noise flooring factor β ($0 \leq \beta < 1$) makes use of the addition of background noise to mask the residual noise. It determines the minimum value of the gain function. If this factor is increased, parts of the residual noise can be masked, but the level of background noise retained in the enhanced speech increases. 3) The exponent $\gamma = \gamma_1 = 1/\gamma_2$ determines the abruptness of the transition from pure clean speech to pure noise in noisy speech. For magnitude subtraction, $\gamma_1 = \gamma_2 = 1$ whereas for power subtraction, $\gamma_1 = 2$, $\gamma_2 = 0.5$.

To formally select these adaptation parameters, we will optimize the fixed gain function of (23) by minimizing the speech distortion and residue noise based on the masking threshold determined in Section III. To segregate the residual noise from the speech distortion, we define a differential wavelet coefficient as the difference between the wavelet coefficients of the clean speech and the enhanced speech. The differential wavelet coefficients can be expressed as follows:

$$\begin{aligned} \partial w_{j,k} &= \tilde{w}_{j,k}(s) - w_{j,k}(s) \\ &= \text{Gain}(j, k) \cdot w_{j,k}(x) - w_{j,k}(s) \\ &= w_{j,k}(s)(\text{Gain}(j, k) - 1) + \text{Gain}(j, k) \cdot w_{j,k}(\nu). \end{aligned} \quad (25)$$

The optimal gain function will be derived by following the same approach as in [16]. The difference is that our gain function will be optimized based on the thresholding criterion that correlate with both temporal and simultaneous maskings. The auditory perception is better approximated by using more complex wavelet basis and efficient filter bank structure. The differential wavelet coefficients can be viewed as a linear superposition of speech distortion, $\xi_{j,k}(s)$ and residual noise, $\xi_{j,k}(\nu)$ in the wavelet domain, where the speech distortion, $\xi_{j,k}(s)$ and residual noise, $\xi_{j,k}(\nu)$ are defined as

$$\xi_{j,k}(s) = w_{j,k}(s) (\text{Gain}(j, k) - 1) \quad (26)$$

$$\xi_{j,k}(\nu) = w_{j,k}(\nu) \text{Gain}(j, k). \quad (27)$$

The energies of the speech distortion and residual noise are, respectively, given by

$$E_{\xi_{j,k}(s)} = \sum_{k=0}^{2^{-j}-1} |w_{j,k}(s)|^2 \cdot |(\text{Gain}(j, k) - 1)|^2 \quad (28)$$

$$E_{\xi_{j,k}(\nu)} = \sum_{k=0}^{2^{-j}-1} |w_{j,k}(\nu)|^2 \cdot |(\text{Gain}(j, k))|^2. \quad (29)$$

Since the residual noise is not audible when its energy is smaller than the masking threshold, we will suppress only the wavelet coefficients of noisy speech when the noise level is greater than the masking threshold as in [16]. To optimize the gain function, we minimize the energy of speech distortion in wavelet domain subjected to the constraint that the energy of residual noise is kept below the masking threshold. A cost function $J(j, k)$ can be formulated according to the energies of the speech distortion and residual noise expressed in terms of wavelet coefficients.

$$J(j, k) = E_{\xi_{j,k}(s)} + \eta_{j,k}(Z_W) \cdot \{E_{\xi_{j,k}(s)} - T_{j,k}(Z_W)\} \quad (30)$$

where the Lagrangian multiplier of each subband, $\eta_{j,k}(Z_W)$ serves as a weighting factor to the residue noise.

Substituting (28) and (29) into (30), we have

$$J(j, k) = \sum_{k=0}^{2^{-j}-1} |w_{j,k}(s)|^2 \cdot |(\text{Gain}(j, k) - 1)|^2 + \eta_{j,k}(Z_W) \cdot \left\{ \sum_{k=0}^{2^{-j}-1} |w_{j,k}(\nu)|^2 \cdot |(\text{Gain}(j, k))|^2 - T_{j,k}(Z_W) \right\}. \quad (31)$$

The optimal gain function is obtained by minimizing the cost function $J(j, k)$. Taking the derivative of the cost function $J(j, k)$ with respect the Lagrangian multiplier, $\eta_{j,k}(Z_W)$ and set the result to zero, the optimal gain function is determined

$$\frac{dJ(j, k)}{d\eta_{j,k}(Z_W)} = \sum_{k=0}^{2^{-j}-1} |w_{j,k}(\nu)|^2 \cdot |(\text{Gain}(j, k))|^2 - T_{j,k}(Z_W) = 0$$

$$\text{Gain}_{\text{opt}}(j, k) = \sqrt{\frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-j}-1} |w_{j,k}(\nu)|^2}}, \quad 0 \leq \text{Gain}(j, k) \leq 1. \quad (32)$$

By equating the adaptive gain function of (24) to the optimal gain function of (32), and considering the power subtraction, i.e., $\gamma_1 = 2$ and $\gamma_2 = 0.5$, the closed-form expressions for the subtraction parameters, α and β can be derived as follows:

$$1 - \alpha \cdot \frac{1}{\text{SNR}(j, k)} = \frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-j}-1} |w_{j,k}(\nu)|^2}$$

$$\Rightarrow \alpha = \text{SNR}(j, k) \cdot \left(1 - \frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-j}-1} |w_{j,k}(\nu)|^2} \right) \quad (33)$$

$$\beta \cdot \frac{1}{\text{SNR}(j, k)} = \frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-j}-1} |w_{j,k}(\nu)|^2}$$

$$\Rightarrow \beta = \text{SNR}(j, k) \cdot \frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-j}-1} |w_{j,k}(\nu)|^2}. \quad (34)$$

Equations (33) and (34) ensure that the subtraction parameters α and β are adapted to the masking threshold of human auditory system to achieve a good tradeoff between the residual noise, speech distortion, and background noise. In high SNR condition, the parameter α is increased to reduce the residual noise at the expense of introducing more speech distortion. On the contrary, in low SNR condition, the parameter β is increased to trade the residual noise reduction for an increased background noise in the enhanced speech. To make the residual noise inaudible, the subtraction parameters are set such that the residual noise stays just below the masking threshold $T_{j,k}(Z_W)$. If the masking threshold is low, the subtraction parameters will be increased to reduce the effect of residual noise. If the masking threshold is high, the residual noise will naturally be masked and become inaudible. Therefore, the subtraction parameters can be kept at their minimal values to minimize the speech distortion.

B. Unvoiced Speech Enhancement

The intelligibility of the processed speech is further improved by an USE (see Fig. 5). Although the GPTFS subsystem is useful for enhancing the portion of speech signal that contains most of the signal energy, they sometimes degrade the high-frequency bands of low energy. To alleviate this problem, we apply a soft thresholding method to enhance the portion of speech signal that lies in the high-frequency bands [31].

Despite most energy of the speech signal is concentrated around the lower frequency bands, the speech energy in the high-frequency range may possess considerably high peak and carry crucial perception information. To enhance the portion of speech signal that lies in the high-frequency range without degrading the performance of the overall system, the USE unequally weights the frequency bands to amplify only those components with detectable peaks in the high-frequency range. The time-frequency energy (TFE), which is estimated using the wavelet coefficients, is applied in this subsystem. The TFE is computed as

$$\Gamma_{j,k} = \sum_{F_j} (\tilde{w}_{j,k}(x))^2 / \sum_{n=1}^{F_{j\max}} \|x[n]\|^2 \quad (35)$$

where $\|\cdot\|$ is the norm operator. $F_{j\max}$ is the total frame length. F_j is the frame length of each subband. The TFE mapping, Γ is a normalized energy function of wavelet coefficients at each position of the binary tree. To estimate the enhanced original speech, we assume that the TFE of noise does not change much over time. Since only noisy observation is accessible, we can only estimate the noise TFE by the minimum TFE of the observation. The high-frequency TFE peak can be deduced as

$$\tilde{\Gamma}_{j,k}(s) = \Gamma_{j,k}(x) - \tilde{\Gamma}_{j,k}(\nu) = \Gamma_{j,k}(x) - \min_{1 \leq m \leq M} \Gamma_{j,k}^m(x) \quad (36)$$

where $\Gamma_{j,k}(x)$ denote the TFE of the noisy speech which is composed of the TFEs of noise $\Gamma_{j,k}(\nu)$ and the original speech $\Gamma_{j,k}(s)$. The superscript m denotes the frame index. Unlike GPTFS, which operates on each time frame, the USE spans over several frames, where M indicates the number of frames covered in this duration.

To amplify those high-frequency bands containing components of unvoiced speech without affecting all other high-frequency bands, we define a threshold $\delta_{j,k}$. Different wavelet coefficients of the processed speech are then emphasized via their weights, $u_{j,k}$. The weight, $u_{j,k}$ is obtained by

$$u_{j,k} = \begin{cases} 1, & \hat{\Gamma}_{j,k}(s) < \delta_{j,k} \\ \sum_{m=1}^M (\Gamma_{j,k}^m(x) - \hat{\Gamma}_{j,k}(\nu)) / \hat{\Gamma}_{j,k}(\nu), & \hat{\Gamma}_{j,k}(s) > \delta_{j,k} \end{cases} \quad (37)$$

where the threshold $\delta_{j,k}$ is determined empirically through experimentation, below which the subband is noise-free and $u_{j,k}$ is set to 1. Therefore, when the noise estimate approaches zero and the speech is strong enough, $u_{j,k} = 1$. The weighted wavelet coefficients are given by

$$\hat{w}_{j,k}(s) = u_{j,k} \cdot \tilde{w}_{j,k}(s). \quad (38)$$

GPTFS either amplifies or attenuates a particular frequency band based on the estimated signal energy content in low frequency. When the SNR is high, the USE is effective. In the case of low SNR, GPTFS have suppressed most energy of noise while significantly reduced the unvoiced speech at the same time. Nevertheless, the succeeding USE can still coarsely estimate the noise and tune a set of weights to somewhat enhance the unvoiced speech. Finally, an inverse wavelet transform is applied to resynthesize the enhanced speech

$$\hat{s}[n] = \text{IPWPT}(\hat{w}_{j,k}(x)) \quad (39)$$

where IPWPT means inverse PWPT.

V. EXPERIMENTAL RESULTS

In this section, the proposed system is evaluated with speeches produced in various adverse conditions and compared against the following competitive methods: 1) speech enhancement method using perceptually constrained gain factors in critical-band-WPT (Lu04) [16]; 2) speech enhancement method incorporating a psychoacoustical model in frequency domain (Hu04) [12]; 3) wavelet speech enhancement method based on the teager energy operator (B&R01) [5]; 4) perceptual time-frequency subtraction algorithm (Li01) [14]; 5) single channel speech enhancement based on masking properties of human auditory system (Vir99) [33]; and 6) parametric spectral subtraction (Sim98) [26]. The original speeches used for the tests were taken from the DARPA TIMIT Acoustic-Phonetic Speech Database. Different background noises with different time-frequency distributions were taken from the Noisex-92 database [37]. The tested noisy environments include white Gaussian noise, pink noise, Volvo engine noise, F16 cockpit noise, factory noise, high-frequency channel noise, and speech-like noise. Noise has been added to the clean speech signal with

Fig. 6. Comparison of different speech enhancement methods in white Gaussian noise by (a) SNR improvement and (b) IS distortion.

different SNRs. The same speech pause detection algorithm from Marzinik and Kollmeier [20] is used for all algorithms being compared that require noise estimation. This speech pause detection algorithm has a low hit rate of 0.11 and 0.12, and a low false-alarm rate of 0.085 and 0.08 at SNR of -5 and -10 dB, respectively, in the speechlike noise environment.

A. SNR Improvement and Itakura-Saito (IS) Distortion

The amount of noise reduction is generally measured in terms of the SNR improvement, given by the difference between the input and output segmental SNRs. The pre-SNR and post-SNR are defined as

$$\text{SNR}_{\text{pre}} = \frac{1}{K} \sum_{m=0}^{K-1} \left(20 \cdot \log_{10} \left(\frac{\frac{1}{N} \sum_{n=0}^{N-1} s(n+Nm)}{\frac{1}{N} \sum_{n=0}^{N-1} v(n+Nm)} \right) \right) \quad (40)$$

$$\text{SNR}_{\text{post}} = \frac{1}{K} \sum_{m=0}^{K-1} \left(20 \cdot \log_{10} \left(\frac{\frac{1}{N} \sum_{n=0}^{N-1} s(n+Nm)}{\frac{1}{N} \sum_{n=0}^{N-1} (s(n+Nm) - \hat{s}(n+Nm))} \right) \right) \quad (41)$$

where K represents the number of frames in the signal and N indicates the number of samples per frame.

The segment SNR improvement is defined as $\text{SNR}_{\text{post}} - \text{SNR}_{\text{pre}}$

$$\text{SNR}_{\text{imp}} = \frac{1}{K} \sum_{m=0}^{K-1} \left(20 \cdot \log_{10} \left(\frac{\frac{1}{N} \sum_{n=0}^{N-1} v(n+Nm)}{\frac{1}{N} \sum_{n=0}^{N-1} (s(n+Nm) - \hat{s}(n+Nm))} \right) \right). \quad (42)$$

This equation takes into account both residual noise and speech distortion. Figs. 6(a)–9(a) compare the SNR improvement of various speech enhancement methods in white noise, Volvo engine noise, factory noise, and speechlike noise with different noise levels. Our proposed method produces

Fig. 7. Comparison of different speech enhancement methods in Volvo engine noise by (a) SNR improvement and (b) IS distortion.

Fig. 8. Comparison of different speech enhancement methods in factory noise by (a) SNR improvement and (b) IS distortion.

higher SNR improvement than other methods, particularly for low-input SNRs. The best noise reduction is obtained in the case of white Gaussian noise, while for colored noise the improvement is less prominent.

It has been shown by experiments that even though the SNRs are very similar at the output of the enhancement system, the listening test and speech spectrograms can produce very divergent results. Therefore, segmental SNR alone is not sufficient to indicate the speech quality and further objective measure is needed to attest the performance ascendancy. IS distortion [22] provides a higher correlation with subjective results than the SNR for speech processing systems. It is derived from the linear predictive coefficient vector $\alpha_s(m)$ of the original clean speech frame and the processed speech coefficient vector $\alpha_{\hat{s}}(m)$ as follows:

$$IS(m) = \frac{\sigma_s^2(m)}{\sigma_{\hat{s}}^2(m)} \cdot \frac{\alpha_{\hat{s}}(m)R_s(m)\alpha_s^T(m)}{\alpha_s(m)R_s(m)\alpha_s^T(m)} + \log \left(\frac{\sigma_s^2(m)}{\sigma_{\hat{s}}^2(m)} \right) - 1 \quad (43)$$

where $\sigma_s^2(m)$ and $\sigma_{\hat{s}}^2(m)$ are the all-pole gains for the processed and clean speeches, respectively, and $R_s(m)$ denotes the clean speech signal correlation matrix. The IS distortion measure gives a zero response when the distortion of two signals to

Fig. 9. Comparison of different speech enhancement methods in speechlike noise by (a) SNR improvement and (b) IS distortion.

be estimated have equal linear predictive coefficients and gain. Smaller value of IS implies better speech quality.

Figs. 6(b)–9(b) show the IS distortion of various methods in different noise environments at varying noise levels. All figures show that the proposed method outperforms other methods with significant margins. This is attributed to the holistic reduction of the background noise, residual noise and speech distortion. In low SNR, GPTFS can suppress most of the noise by appropriately tuning the subtraction parameters according to the time–frequency masking threshold. This tradeoff control helps to minimize the processing distortion in the enhanced speech. The USE of the proposed system compensates for the speech in the high-frequency range. It further improves the intelligibility and reduces the distortion. When the SNR is high, the proposed system has excellent performance. Our proposed system has successfully overcome the major drawback of traditional denoising methods, which is the tendency to deteriorate some useful components of speech, resulting in the weakening of its intelligibility.

B. Speech Spectrograms

The above objective measures do not provide information about how speech and noise are distributed across frequency. In this perspective, the speech spectrograms, which yields more accurate information about the residual noise and speech distortion than the corresponding timing waveforms are analyzed. All the speech spectrograms presented in this section used Kaiser Window of 500 samples with an overlap of 475 samples. Experiments show that the noisy phase is not perceived as long as the local SNR is greater than about 6 dB. However, at SNR = 0 dB, the effect of the noisy phase is audible and is perceived as an increased roughness.

The spectrograms of noisy and enhanced speeches obtained with the proposed method in various noise environments show that the residual noise has been greatly reduced, as long as the noise remains stationary or piecewise stationary. Fig. 10 shows the comparisons of the speech spectrograms obtained by different enhancement methods in speechlike noise. As the nonstationary of noise increases, the results of our proposed method are still better than other algorithms. This is because the proposed PWPT filter bank has closely approximated the

Fig. 10. Speech spectrograms. (a) Original clean speech. (b) Noisy signal (additive speechlike noise at a SNR = 0 dB). (c) Speech enhanced by B&R01 [5]. (d) Speech enhanced by Li01 [14]. (e) Speech enhanced by Lu04 [16]. (f) Speech enhanced by Hu04 [12]. (g) Speech enhanced by the proposed method.

critical bands of the human auditory system. By appropriately segregating the speech and noise components of the noisy speech in frequency and time, the subtraction parameters of our proposed GPTFS method adapt well to the combined temporal and simultaneous masking threshold. The supplement of USE also improves the performance of our proposed system in high-frequency noise. The worst results of the proposed method are obtained with speechlike noise. This noise is particularly difficult to handle by any method, because it has the same frequency distribution as the long-term speech.

C. Subjective Measurements

In order to validate the performances of objective evaluation, subjective listening tests are performed by subjecting the

speech to varying noise at SNR = 0 and 10 dB before they are processed by different algorithms. An informal subjective test was performed with ten listeners. Each listener gives a score between one and five to each test signal. The score represents his global perception of the residual noise, background noise and speech distortion. The scale used for these tests corresponds to the mean opinion score (MOS) scale [7]. For each tester, the following procedure has been applied: first, the clean and noisy speeches are played and repeated twice; second, each test signal, which is repeated twice for each score, is played three times in a random order. This leads to 30 scores for each test signal. The results are presented in Table III. The subjective listening tests confirm that the proposed enhancement method produces the highest quality speech perceived by the actual human listeners among the algorithms being tested.

TABLE III
COMPARISON BETWEEN INFORMAL MOS AND PESQ SCORES

Informal MOS score/ PESQ score				
Noise type, SNR _i	Noisy signal	Lu04 [13]	Hu04 [10]	Proposed
White, 0 dB	1.35/1.24	1.46/1.32	1.55/1.48	1.78/1.85
White, 10 dB	2.53/1.94	2.35/2.00	2.64/2.17	3.00/2.45
Car engine, 0 dB	1.59/1.42	1.59/1.51	1.71/1.65	1.89/1.78
Car engine, 10 dB	2.22/2.11	2.22/2.04	2.36/2.25	2.60/2.39
Factory, 0 dB	1.45/1.51	1.50/1.51	1.701/1.76	2.00/1.95
Factory, 10 dB	2.62/2.16	2.45/2.11	2.77/2.37	2.85/2.50
Speechlike, 0 dB	1.51/1.22	1.43/1.14	1.66/1.39	1.90/1.70
Speechlike, 10 dB	2.43/1.88	2.33/1.89	2.58/2.05	2.71/2.35

To prove the trend of the informal MOS scores matches well with the objective perceptual evaluation of speech quality (PESQ) measure [38], the PESQ scores are indicated in Table III. The PESQ algorithm is designed to predict the subjective opinion score of a degraded audio sample. PESQ returns a score from 4.5 to -0.5 , with higher score indicating better quality. From Table III, it is observed that the gap between the enhanced speech and noisy speech scores is similar in both MOS and PESQ scores.

VI. CONCLUSION

In this paper, a new speech enhancement system has been proposed. It can be anatomized into several powerful processing techniques that exploit the physiology of human auditory system to recover high-quality speech from noise contaminated speech. The system consists of two functional stages working cooperatively to perform perceptual time–frequency subtraction by adapting the weights of the perceptual wavelet coefficients. The noisy speech is first decomposed into critical bands by perceptual wavelet transform. The temporal and spectral psychoacoustic model of masking is developed to calculate the threshold to be applied to GPTFS method for noise reduction. Different spectral resolutions of the wavelet representation preserve the energy of the critical transient components so that the background noises, distortion, and residual noise can be adaptively processed by GPTFS method. The unvoiced speech is also enhanced by a soft-thresholding scheme. The effectiveness of the proposed system to extract a clear and intelligible speech from various adverse noisy environments in comparison with other well-known methods has been demonstrated through both objective and subjective measurements. The performance robustness under varying signal-to-noise conditions of the proposed system is attributable to the consideration of both the temporal and simultaneous maskings for the tuning of subtraction parameters in the proposed GPTFS. Together with the USE, they make an average SNR improvement of 5.5% by objective measurements, and an average intelligibility improvement of 8% by subjective evaluation.

REFERENCES

- [1] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 4, pp. 1763–1773, Aug. 2004.
- [2] S. Ayat, M. T. Manzuri, and R. Dianat, "Wavelet based speech enhancement using a new thresholding algorithm," in *Proc. Int. Symp. Intell. Multimedia, Video and Speech Process.*, Hong Kong, Oct. 20–22, 2004, pp. 238–241.
- [3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP*, Apr. 1979, vol. 4, pp. 208–211.
- [4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
- [5] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE Signal Process. Lett.*, vol. 8, no. 1, pp. 10–12, Jan. 2001.
- [6] K. Chen, "On the use of different speech representations for speaker modeling," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 301–314, Aug. 2005.
- [7] J. Deller, J. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [8] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.
- [9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
- [10] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.
- [11] M. K. Hasan, S. Salahuddin, and M. R. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Process. Lett.*, vol. 11, no. 4, pp. 450–453, Apr. 2004.
- [12] Y. Hu and P. C. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 270–273, Feb. 2004.
- [13] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE ICASSP*, May 13–17, 2002, vol. 4, p. IV-4164.
- [14] M. Li, H. G. McAllister, N. D. Black, and T. A. De Perez, "Perceptual time–frequency subtraction algorithm for noise reduction in hearing aids," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 9, pp. 979–988, Sep. 2001.
- [15] C. T. Lin, "Single-channel speech enhancement in variable noise-level environment," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 33, no. 1, pp. 137–143, Jan. 2003.
- [16] C.-T. Lu and H.-C. Wang, "Speech enhancement using perceptually-constrained gain factors in critical-band-wavelet-packet transform," *Electron. Lett.*, vol. 40, no. 6, pp. 394–396, Mar. 18, 2004.
- [17] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 314–323, Feb. 1988.
- [18] W. L. B. Jeannes, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: A survey," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 808–820, Nov. 2001.
- [19] A. Mertins, *Signal Analysis: Wavelet, Filter Banks, Time-Frequency Transforms and Applications*. Chichester, U.K.: Wiley, 1999.
- [20] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 109–118, Feb. 2002.
- [21] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.
- [22] N. Nocerino, F. Soong, L. Rabiner, and D. Klatt, "Comparative study of several distortion measures for speech recognition," in *Proc. IEEE ICASSP*, Apr. 1985, vol. 10, pp. 25–28.
- [23] A. Ortega, E. Lleida, and E. Masgrau, "Speech reinforcement system for car cabin communications," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pt. 2, pp. 917–929, Sep. 2005.
- [24] S. J. Park, C. G. Cho, C. Lee, and D. H. Youn, "Integrated echo and noise canceler for hands-free applications," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 49, no. 3, pp. 188–195, Mar. 2002.
- [25] D. Sinha and A. H. Tewfik, "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. Signal Process.*, vol. 1, no. 12, pp. 3463–3479, Dec. 1993.
- [26] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, Jul. 1998.
- [27] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, no. 6, pp. 1647–1652, Dec. 1979.
- [28] G. Strang and T. Nguyen, *Wavelets and Filter Banks*. Cambridge, MA: Wellesley-Cambridge Press, 1996.
- [29] P. Srinivasan and L. H. Jamieson, "High-quality audio compression using an adaptive wavelet packet decomposition and psychoacoustic

modeling," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 1085–1093, Apr. 1998.

- [30] Y. Shao and C. H. Chang, "A generalized perceptual time-frequency subtraction method for speech enhancement," in *Proc. IEEE ISCAS*, Kos Island, Greece, May 20–23, 2006, pp. 2537–2540.
- [31] Y. Shao and C. H. Chang, "A versatile speech enhancement system based on perceptual wavelet denoising," in *Proc IEEE ISCAS*, Kobe, Japan, May 23–26, 2005, pp. 864–867.
- [32] D. Veselinovic and D. Graupe, "A wavelet transform approach to blind adaptive filtering of speech from unknown noises," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 50, no. 3, pp. 150–154, Mar. 2003.
- [33] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.
- [34] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 2nd ed. Chichester, U.K.: Wiley, 2000.
- [35] G. D. Wu and C. T. Lin, "A recurrent neural fuzzy network for word boundary detection in variable noise-level environments," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 31, no. 1, pp. 84–97, Feb. 2000.
- [36] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Berlin, Germany: Springer-Verlag, 1990.
- [37] "Speech and noise data base," *NATO AC243-panel 3/RSG.10*, 1992. NOISEX-92.
- [38] *Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs*, p. 862, Feb. 2001. ITU-T Recommend.

Chip-Hong Chang (S'93–M'98–SM'03) received the B.Eng. degree (Hons) from the National University of Singapore, Kent Ridge, in 1989, and the M.Eng. and Ph.D. degrees from Nanyang Technological University (NTU), Singapore, in 1993 and 1998, respectively.

He served as Technical Consultant in industry prior to joining the School of Electrical and Electronic Engineering, NTU, in 1999, where he is currently an Associate Professor. He has been holding joint appointments at the university as Deputy Director of the Centre for High Performance Embedded Systems since 2000, and Program Director of the Centre for Integrated Circuits and Systems since 2003. His current research interests include low-power arithmetic circuits, constrained driven architectures for digital signal processing, and digital watermarking for IP protection. He has published three book chapters and more than 120 research papers in international refereed journals and conferences.

Yu Shao received the B.Eng. degree from the Beijing Institute of Technology, Beijing, China, in 2002. He is currently working toward the Ph.D. degree at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

His research interests include speech enhancement, speech recognition, adaptive filters, and wavelets.