Perceptual Wavelet packet transform based Wavelet Filter Banks Modeling of Human Auditory system for improving the intelligibility of voiced and unvoiced speech: A Case Study of a system development

Ranganadh Narayanam[1]

1.Research Scholar, University of Ottawa, Canada

ABSTRACT:

The objective of this project is to discuss a versatile speech enhancement method based on the human auditory model. In this project a speech enhancement scheme is being described which meets the demand for quality noise reduction algorithms which are capable of operating at a very low signal to noise ratio. We will be discussing how proposed speech enhancement system is capable of reducing noise with little speech degradation in diverse noise environments. In this model to reduce the residual noise and improve the intelligibility of speech a psychoacoustic model is incorporated into the generalized perceptual wavelet denoising method to reduce the residual noise. This 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. To calculate the bark spreading energy and temporal spreading energy the wavelet coefficients are used from which a time frequency masking threshold is deduced to adaptively adjust the subtraction parameters of the discussed method. To increase the intelligibility of speech an unvoiced speech enhancement algorithm also integrated into thesystem.

I. INTRODUCTION*:*

The performance of the automatic speech processing systems degrade drastically when confronted with a great adverse noise conditions such as background noise and micro phone distortions. For this reason there is a strong demand for quality reduction algorithms capable of operating at very low signal to noise ratio in order to combat various forms of noise distortion. The solutions can be classified into two main areas a) nonparametric; usually remove an estimate of the distortion from the noisy features, and b) statistical model based speech enhancements, statistical model based speech enhancement utilizes a parametric model (1,2) of the signal generation process. This project is based on the proposed speech enhancement system which is based on subtractive type algorithms. By subtracting the noise estimation from the noisy speech this system estimates the short time spectral magnitude of speech. The reason that this is chosen is because of the relative simplicity, in the sense that it only requires an estimate of the noise power spectrum; its high flexibility against subtraction parameters variation. Here in this project it is to emphasize the reduction of the effect of residual noise and speech distortion in the denoising process and the enhancement of the denoised speech (2,3) in high frequency to improve its intelligibility. The proposed one consists of two main functions. One is a generalized perceptual time-frequency subtraction method based on the masking properties of the human auditory system, this works in conjunction with a perceptual wavelet packet transform (PWPT) (11,10,9,1) to reduce the effect of noise contamination. The second part of it is an unvoiced speech enhancement (USE) (11,9,5,1) , which tunes a set of weights in high-frequency sub-bands to improve intelligibility of the processed speech. The main theme in this proposed method is the use of PWPT to approximate 24 critical bands of the human auditory system up to 16 khz. It enables the components of 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. This proposed method uses PWPT (1,11,2,6) to analyze to improve the perceptual quality of the final processed speech. Parametric formulation of subtractive noise reduction based on the generalized perceptual wavelet transform is the second critical step of this proposed method. In this spectral subtraction method fourier transform based gain function of the generalized spectral subtraction method and derivation of the close from expressions for the subtraction factor to optimize the trade-off in the simultaneous reduction of background noise residual noise and

speech distortion. These parameters of GPTFS (1,11, 2, 6) can then be adaptively tuned according to the noise level and the masking thresholds derived from the human auditory model in wavelet domain. To integrate GPTFS and USE in wavelet domain a new system for speech enhancement is developed.

II. A PERCEPTUAL WAVELET FILTER BANK ARCHITECTURE:

Architecture for the perceptual wavelet filter bank: To design this algorithm for enhancing speech a well built psychoacoustic model of the ear which has an unsurpassed capability to adapt to noise. In this a new human auditory model that adapts to the basic structure of traditional auditory model but replace the time invariant band pass 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 sub -bands that approximate the critical bands, efficient seven level tree structure is implemented. This is given in the Fig 1. Two channel wavelet $(1,\ 11,2,6)$ filter banks are used to split the low pass and high pass bands as opposed to only the low pass and high pass bands in the usual wavelet decomposition. Advantages: first, Smoothness property of wavelet is determined by the number of vanishing moments: more the vanishing moments the stringent bandwidth and stop band attenuation of each sub- band and can be more close approximation by using the wavelet decomposition. Second, according to the psychoacoustic study of human ears a frequency to bark transformation needs to be performed which can be accomplished in audio processing systems by dividing the frequency range into critical bands. Using the perfect reconstruction filter bank with finite length filters using different wavelets for the analysis and synthesis scaling functions. Let H(z) and G(z) be the low pass (LP) and high pass (HP) transfer functions, before the decimation by two operation in each stage of the analysis filter bank. F (z) and J(z) be the LP and HP transfer functions, after the up sampling by two operation in each stage of the synthesis filter bank. Then the analysis and synthesis filter banks are related by

$$g(n) = (-1)^n f(n) \leftrightarrow G(z) = F(-z)$$
$$j(n) = -(-1)^n h(n) \leftrightarrow J(z) = -H(-z). \quad (1)$$

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 the equation (1). According frequency selectivity related to critical band, temporal resolution of the human ear, and regularity property of wavelets debauchies wavelet basis is chosen as prototype filter and a seven stage WPT is adopted to build perceptual wavelet filter bank. $W_{j^{\prime}k}$ Represents WPT coefficient, where k is the coefficient number; j is the transform stage from which $W_{j^{\prime}k}$ 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. Fig 2 shows the difference in the critical band rate between the critical bands and the proposed perceptual wavelet packet tree structural bands. The critical band rate Z_{B} in bark is approximated by where the frequency, f, is measured in Hz.

$$Z_{\mathrm{B}} = 13\tan^{-1}(7.6\times10^{-4}f) + 3.5\tan^{-1}(1.33\times10^{-4}f)^2 \quad (2)$$

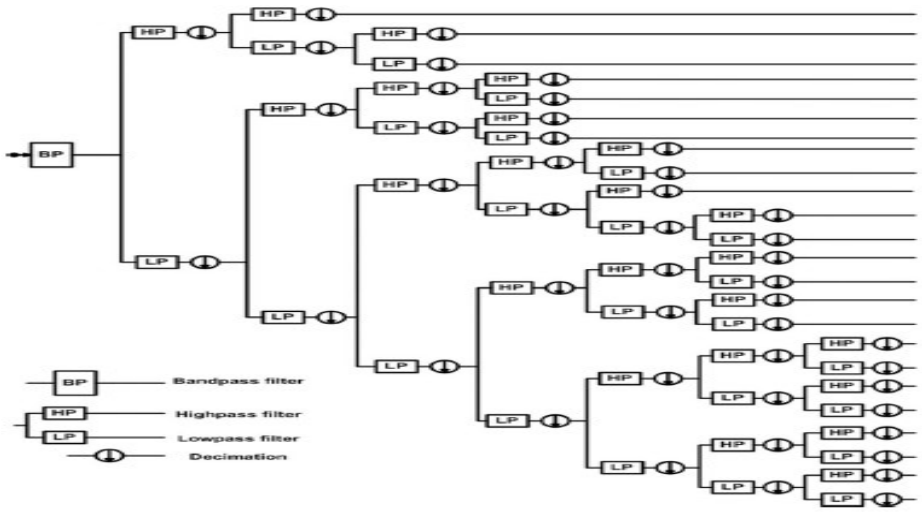

Fig 1: Perceptual wavelet packet decomposition tree (PWPT)

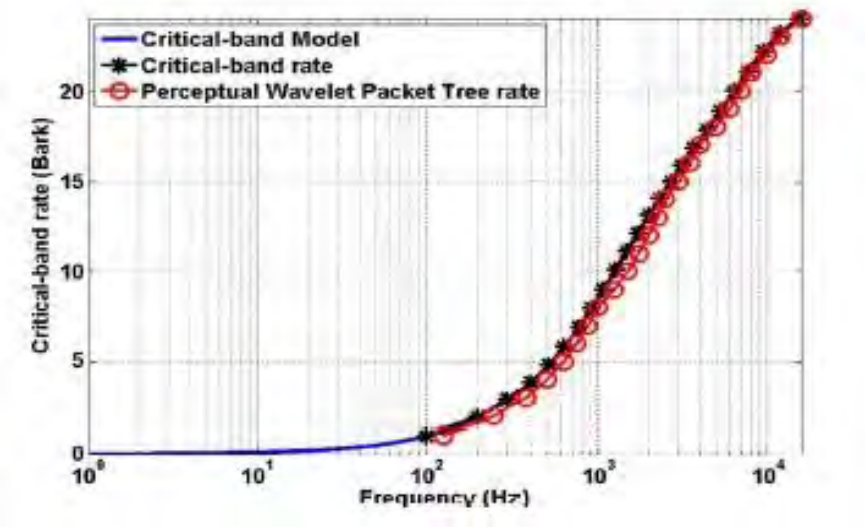

Fig 2: center frequency of critical bands and perceptual wavelet packet decomposition tree.

Table 1 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 2 Critical band rate Z and perceptual wavelet filter banks W

Z	Bark Scale			Wavelet Scale		
	[fl fu]	fc	Δf	[fl fu]	fc	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

In figure 3 it is given that the bandwidths of the critical bands and the perceptual wavelet packet tree it is form the figure 3 that the critical bands have constant width at approximately 100 hz for centre frequencies upto 500 hz, and the bandwidths increase as the centre frequency increases further. The critical bandwidth (CBW) (11,1,3,8) is hertz is calculated by

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

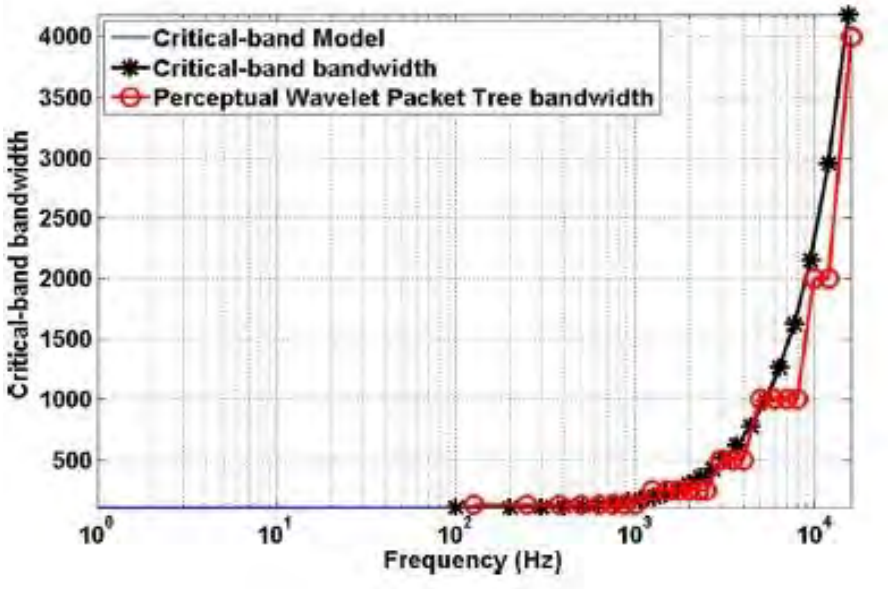

Fig. 3: Bandwidths of critical bands and perceptual wavelet packet decomposition tree.

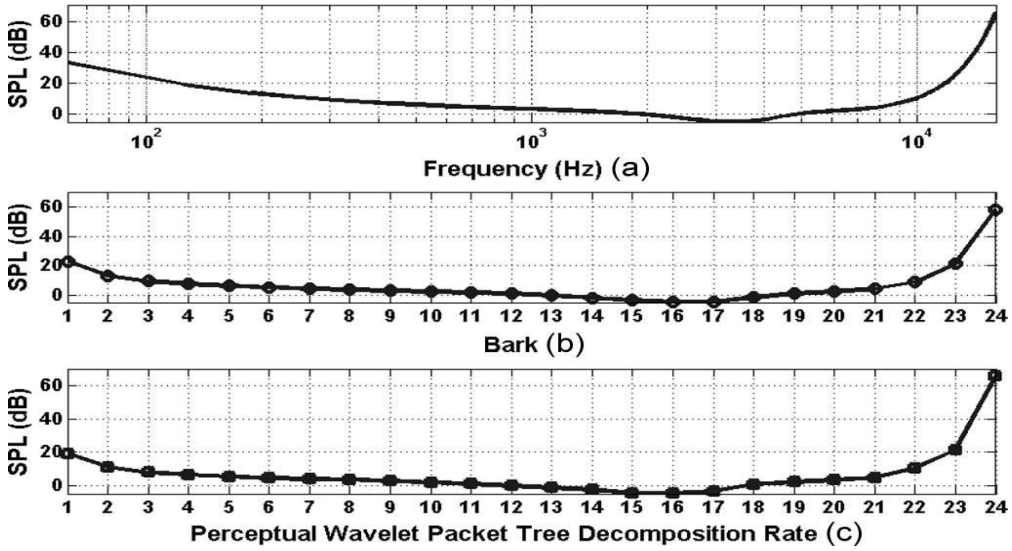

Fig 4: ATH in (a) frequency, (b) bark, and (c) perceptual wavelet packet tree scales.

Figure 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 pure tone such that it can be detected by a listener in a noiseless environment. The table II and figures given it makes clear regarding the proposed perceptual wavelet packet tree can closely mimic the experimental critical bands. The parameters of the discrete WPT (11,6,9,1) filter used to derive the plots of figures are determined based on the auditory masking properties.

III. ADAPTIVE SPEECH ENHANCEMENT SYSTEM:

The proposed adaptive speech enhancement system a new GPTFS method based on the PWPT and the human auditory perception is being discussed. Its parametric formulation is derived from the basis of the generalized Fourier spectral subtraction algorithm. The GPTFS algorithm incorporates most of the basic subtraction rules and realizes the subtraction in a broader time-frequency domain. To get better perceptual outputs more crucial information has been preserved than in the Fourier transform domain. The block diagram for this system is given Fig 5.

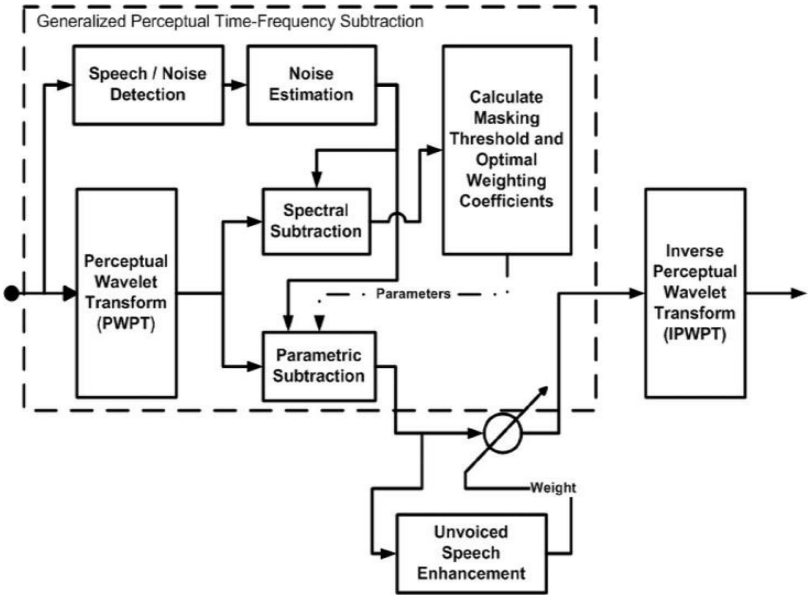

Fig 5: Architecture of the proposed speech enhancement system.

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. To calculate a time frequency masking threshold this estimate is useful. To compute an estimation of the original speech masking threshold which is masking dependent this threshold is used. This approach assumes that the high energy frames of speech will partially mask the input noise, hence reducing the need for a strong enhancement mechanism. Frames containing less

speech will undergo an overestimated subtraction. To further improve the intelligibility (2,8,9,11,1) of processed speech, an USE is applied. The processed speech is reconstructed by the inverse PWPT. During speech pauses the noise estimation is assumed to be available and is performed. The general speech pause detection algorithm is adopted for the noise spectrum estimation by tracking the power envelope dynamics. The speech pause detection algorithm has been extended for sub band processing.

A) Generalized perceptual time-frequency subtraction: noisy signal x[n] = s[n] + v[n]

x[n] and s[n] are the noisy and original speech respectively. V[n] is the additive noise. To capture the localized information of transient signal, the PWPT is applied to the noisy input speech

$$w_{j,k}(x) = w_{j,k}(s) + w_{j,k}(\nu), \qquad \forall j = 0, 1, \ldots, j_{\max} - 1 \quad k = 0, 1, \ldots, 2^j - 1 \tag{4}$$

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, noise. (j, k) in the subscript of w corresponds to its scale translation indexes. *jmax* is the maximum number of levels of wavelet decomposition. Then according to spectral subtraction method in the wavelet domain the estimated power of the enhanced speech is given by

$$|\tilde{w}_{j,k}(s)|^2 = \begin{cases} |w_{j,k}(x)|^2 - |\tilde{w}_{j,k}(\nu)|^2, & \text{if } |w_{j,k}(x)|^2 > |\tilde{w}_{j,k}(\nu)|^2 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Where $|\tilde{w}_{j,k}(s)|^2$ $|w_{j,k}(x)|^2$ $|\tilde{w}_{j,k}(\nu)|^2$, are the estimates of the power wavelet coefficients of the noise suppressed speech signal the noisy speech and the estimated noise in the wavelet domain. The time average noise spectrum is obtained as

$$|\tilde{w}_{i,j,k}(\nu)|^2 = 0.9 \cdot |w_{i-1,j,k}(\nu)|^2 + 0.1 \cdot |w_{i-1,j,k}(x)|^2 \tag{6}$$

Where i is frame index. Human speech is based on mixing voiced and unvoiced phonemes over time are the consideration taken into account for this design. For this reason s[n] is nonstationary over time interval of above 250ms, and the noise v[n] is assumed to be piecewise stationary or more stationary than s[n], which is valid for most noise encountered, then input signal x[n] is divided into 128 samples.

$$\begin{aligned} |\tilde{w}_{j,k}(s)|^2 &= |w_{j,k}(s) + w_{j,k}(\nu)|^2 - |\tilde{w}_{j,k}(\nu)|^2 \\ &= |w_{j,k}(s)|^2 + \underbrace{\left(|w_{j,k}(\nu)|^2 - |\tilde{w}_{j,k}(\nu)|^2\right)}_{\text{Noise Variations}} \\ &\quad + \underbrace{w^*_{j,k}(s) w_{j,k}(\nu) + w_{j,k}(s) w^*_{j,k}(\nu)}_{\text{Cross Products}}. \end{aligned} \tag{7}$$

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. The final gain is given by

$$\text{Gain}(j,k) = \sqrt{1 - \frac{|\tilde{w}_{j,k}(\nu)|^2}{|w_{j,k}(x)|^2}} = \sqrt{1 - \frac{1}{\text{SNR}_{\text{post}}(j,k)}} \tag{8}$$

Where $\mathrm{SNR}_{\mathrm{post}}(j,k) = |w_{j,k}(x)|^2 / |\tilde{w}_{j,k}(\nu)|^2$ is posterior SNR, which is defined as the ratio of the wavelet power spectrum of the noisy speech signal to that of the estimated noise. The input noisy speech is attenuated more heavily with decreasing *posterior* SNR and vice versa with increasing *posteriori* SNR. The residual noise is made perceptually white by using several flexible subtraction parameters. These subtraction parameters are chosen to adapt to a criterion associated with human auditory perception. It considers only the simultaneous masking property of the human auditory system. The effects of PWPT gain function's parameters have the effects in the design 1) The subtraction factor controls the amount of noise subtracted from the noisy signal. Over subtraction allows the time-frequency spectrum to be attenuated more than necessary. This factor must be selected in an appropriately. 2) The noise flooring factor [3] 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. 3) The exponent [4] determines the abruptness of the transition from pure clean speech to pure noise in noisy speech. To formally select these adaptation parameters we need to optimize the fixed gain function. To segregate the residual noise from the speech distortion a differential wavelet coefficient is defined as the difference between the wavelet coefficients of the clean speech and the enhanced speech. In this case the difference in the gain

function is that the optimization here is based on the thresholding (2,6,8,1) criterion that correrelate with both temporal and simultaneous maskings. The auditory perception is better approximated by using more complex wavelet basis and efficient filter bank structure. The optimal gain function is given by

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

$$\mathrm{Gain}_{\mathrm{opt}}(j,k) = \sqrt{\frac{T_{j,k}(Z_W)}{\sum_{k=0}^{2^{-j}-1} |w_{j,k}(\nu)|^2}}, \quad 0 \le \mathrm{Gain}(j,k) \le 1. \tag{9}$$

By equating the adaptive gain function to the optimal gain function and then considering the power subtraction the closed form expressions for the subtraction parameters alpha and beta are derived. The above equations are used to assure the subtraction parameters alpha and beta is adapted to the masking threshold of human auditory system to achieve a good trade-off between the residual noise speech distortion and background noise. In high SNR condition the parameter alpha is increased to reduce the residual noise at the expense of introducing more speech distortion. In low SNR condition the parameter beta is increased to trade (4,6,8) the residual noise reduction for an increased background noise in the enhanced speech. 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 (9,7,8,1) 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. It is given in the Fig 5.Although the GPTFS sub-system is useful for enhancing the portion of speech signal that contains most of the signal energy, for this soft thresholding is used to enhance the portion of speech signal that is in the high-energy high frequency bands. To enhance the portion of the speech that is in the high frequency range without degrading the performance of the overall system, the USE unequally weights the frequency bands to amplifyonly 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.

$$\Gamma_{j,k} = \sum_{F_j} (\tilde{w}_{j,k}(x))^2 \Big/ \sum_{n=1}^{F_{j\,\max}} \|x[n]\|^2 \tag{10}$$

$F_{j\max}$ is the total frame length. F_j is the frame length of each subband. To estimate the enhanced original speech, the assumption is that TFE of noise does not change much over time. USE spans over several frames. To amplify those high-frequency bands containing components of unvoiced speech without affecting all other high frequency bands, a threshold is defined [5]. Different wavelet coefficients of the processed speech are then emphasized (9,8,7) via their weights. $U_{j,k} = 1$. The weighted coefficients are given by

$$\hat{w}_{j,k}(s) = u_{j,k} \cdot \tilde{w}_{j,k}(s) \tag{11}$$

GPTFS either amplifies or attenuates a particular frequency band based on the estimated signal energy content in low frequency. USE is effective only when the SNR is high. In the case of low SNR, GPTFS have suppressed most energy of noise while significantly reduced the unvoiced speech at the same time. Still the succeeding USE (1,11,6,8) can still estimate the noise and tune a set of weights to somewhat enhance the unvoiced speech.

$$\hat{s}[n] = \text{IPWPT}\left(\hat{w}_{j,k}(x)\right) \tag{12}$$

Where IPWPT means inverse PWPT.

IV. *EXPERIMENTAL RESULTS:*

The discussed method 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 2) speech enhancement method incorporating a psycho acoustical model in frequency domain 3) wavelet speech enhancement method based on the teaser energy operator.

4) Perceptual time-frequency subtraction algorithm 5) single channel speech enhancement based on masking properties of human auditory system 6) parametric spectral subtraction. The 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 added to the clean speech signal with different SNRs.

A) **SNR Improvement and Itakura Saito (IS) distortion:** The amount of noise reduction is generally measured in terms the SNR improvement, given by the difference between the input and output segmental SNRs. The pre-SNR and post-SNR are defined in the equation 40. 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 $\text{SNR}_{\text{post}} - \text{SNR}_{\text{pre}}$. Which is given in 13, 14 and 15 equations.

$$\text{SNR}_{\text{post}} = \frac{1}{K}\sum_{m=0}^{K-1}\left(20\cdot\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}s(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}(s(n+Nm)-\hat{s}(n+Nm))}\right)\right) \tag{13}$$

$$\text{SNR}_{\text{pre}} = \frac{1}{K}\sum_{m=0}^{K-1}\left(20\cdot\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}s(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}v(n+Nm)}\right)\right) \tag{14}$$

$$\text{SNR}_{\text{imp}} = \frac{1}{K}\sum_{m=0}^{K-1}\left(20\cdot\log_{10}\left(\frac{\frac{1}{N}\sum_{n=0}^{N-1}\nu(n+Nm)}{\frac{1}{N}\sum_{n=0}^{N-1}(s(n+Nm)-\hat{s}(n+Nm))}\right)\right). \tag{15}$$

This equation takes into account both residual noise and speech distortion. the figures 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 level. The discussed methods has higher SNR improvement than other methods, particulary for low-input SNRs. The best noise reduction is obtained in the case of white gaussion noise, while for colored noise the improvement is less prominent (4,7,3) . According to the experiements though the SNRs are very similar at the output of the enhancement system, the listening test and speech spectrograms can produce very divergent results. So to indicate the speech quality IS distortion also have to be considered. 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

$$\mathrm{IS}(m) = \frac{\sigma_s^2(m)}{\sigma_{\hat{s}}^2(m)} \cdot \frac{\alpha_{\hat{s}}(m) R_s(m) \alpha_{\hat{s}}^{\mathrm{T}}(m)}{\alpha_s(m) R_s(m) \alpha_s^{\mathrm{T}}(m)} + \log\left(\frac{\sigma_{\hat{s}}^2(m)}{\sigma_s^2(m)}\right) - 1 \tag{16}$$

Where $\sigma_{\hat{s}}^2(m)$ and $\sigma_s^2(m)$ are the all-pole gains for the processed and clean speeches, and denotes the clean speech signal correlation matrix. Smaller value of IS implies better speech quality. The figures 6(b)-9(b) show the IS distortion of various methods in different noise environments at varying noise levels. According to all these proposed method outperforms remaining methods. The USE of the proposed system (5,6,7,8) components for the speech in the high frequency range. When the SNR is high the proposed system has excellent performance.

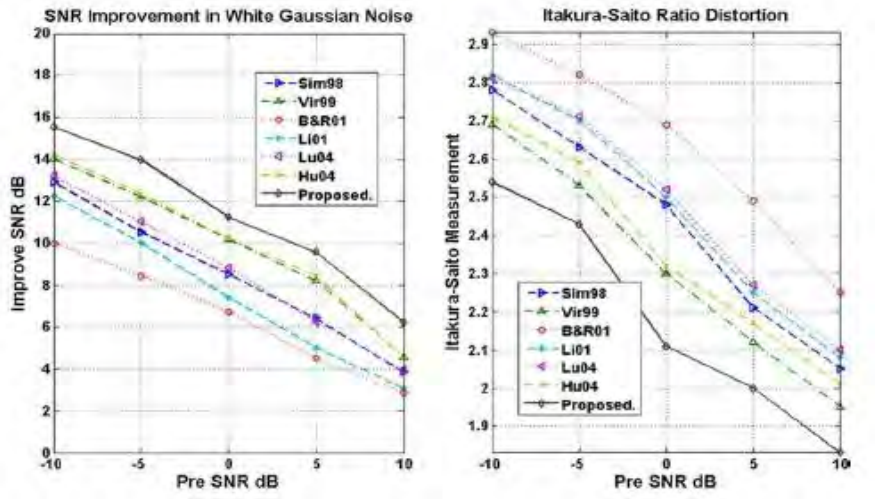

Fig. 6: Comparison of different speech enhancement methods in white Gaussian noise by (a) SNR improvement and (b) IS distortion.

A. Speech spectrograms: the above objective measures do not provide information about how speech and noise are distributed across frequency, for this reason we need to go for speech spectrograms. Fig 10 Shows the comparisons of the speech spectrograms obtained by different enhancement methods in speechlike noise. As the nonstationarity 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 critical bands of the human auditory system. By appropriately segregating (11, 10, 9,1) 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. When the noise is speech like then there are worst result.

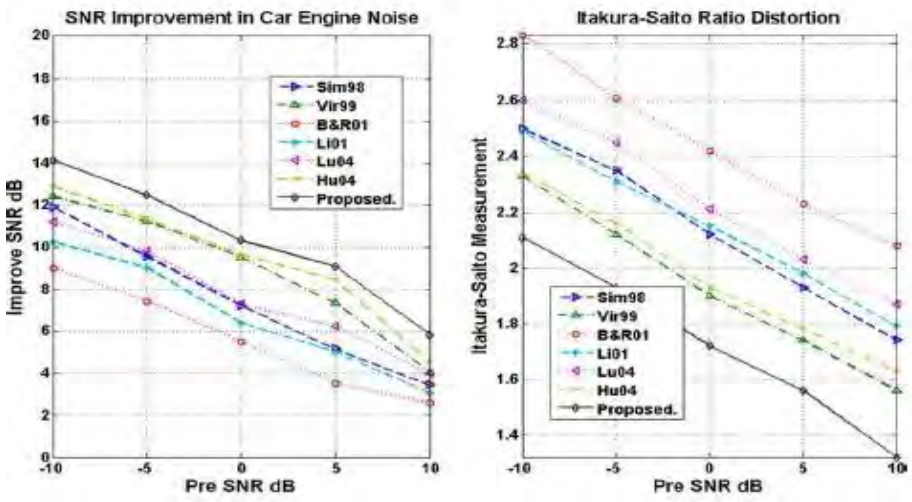

Fig. 7: Comparison of different speech enhancement methods in Volvo engine Noise by (a) SNR improvement and (b) IS distortion.

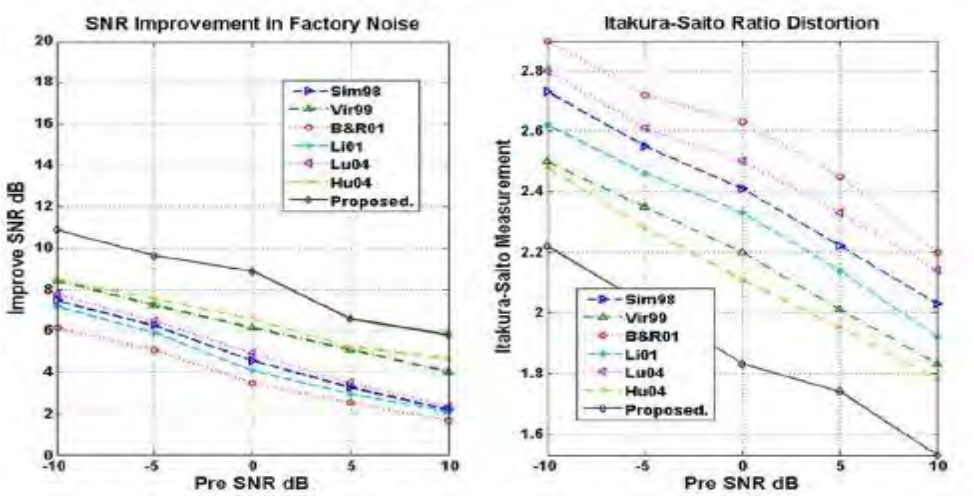

Fig. 8: Comparison of different speech enhancement methods in factory noise by (a) SNR improvement and (b) IS distortion.

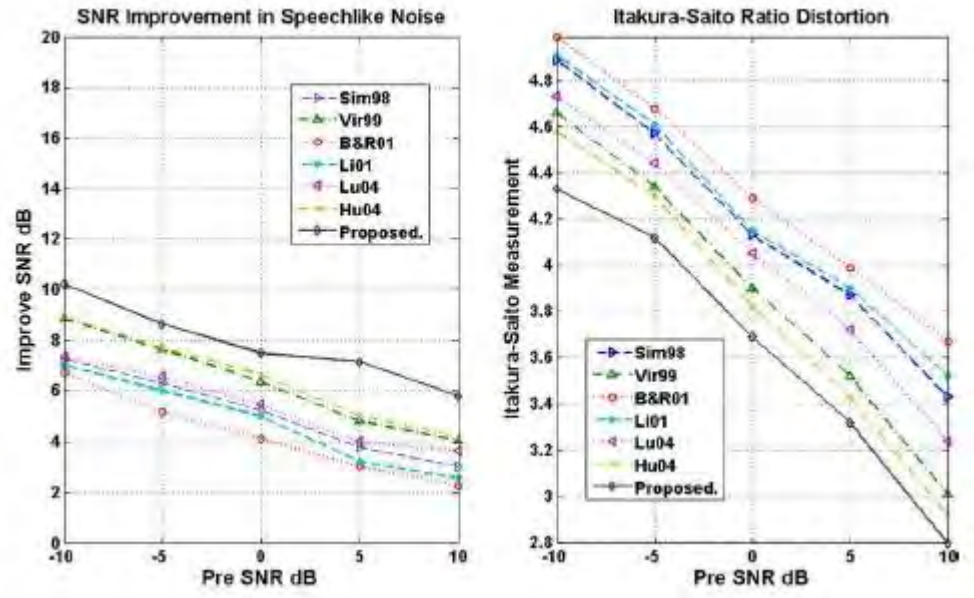

Fig. 9: Comparison of different speech enhancement methods in speechlike noise by (a) SNR improvement and (b) IS distortion.

V. CONCLUSION:

The system consists of two functional stages working cooperately to perform perceptual time- frequency subtraction by adapting the weights of the perceptual wavelet coefficientceints. The noisy speech is first decomposed into critical bands by perceptual wavelet transform. The temporal and spectral pyschoacoustic model of masking is developed to calculate the threshold to be applied to GPTFS method for noise reduction. The unvoiced speech is also enhanced by a soft thresholding scheme. Different spectral resolutions of the wavelet representation preserve the energy of the critical transcient components so that the background noises,distrotion, and residual noise can be adaptively processed by GPTFS method. Both the temporal and simultanious maskings of the tuning of subtraction parameters in the proposed GPTFS are considered. Together with the USE, the system makes an average ANR improvement of 5.5% by objective measurements, an average intelligibility improvement of 8% by subjective evaluation.

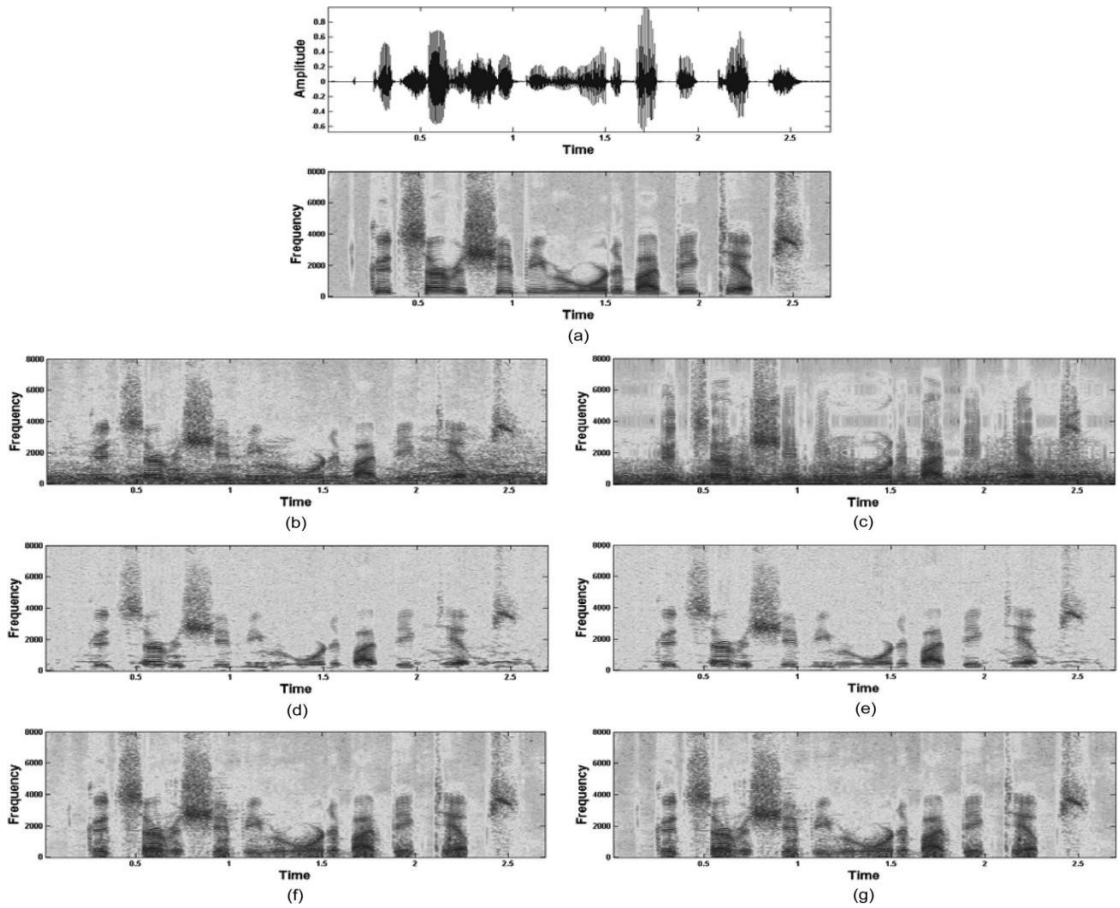

Fig. 10: Speech spectrograms. (a) Original clean speech. (b) Noisy signal (additive speech like noise at a SNR = 0 dB). (c), (d), (e) are the Speech enhanced by above specified recent methods specified (g) Speech enhanced by the proposed method.

REFERENCES

[1] Ranganadh Narayanam, "Wavelet Filter Banks Modeling of Human Auditory System for Robust Speech Enhancement", IJSER, 2012 volume 3, issure 4

[2] R. Nicole, J. Sohn, N.S. Kim, and W. Sung, "A statistical Model Based Voice Activity Detection," IEEE Signal Process. Lett., vol. 6, 1999, pp. 1-3.

[3] Javier Ramirez,Jos C segura, Carmen Benitez,Angel de la torre, Antonio Rubio, "Efficient voice activity detection algorithms using long-term speech information", J. Ram_ırez et al. / Speech Communication 42 (2004) 271– 287.

[4] I. Krekule, "zero crosssing detection of the presence of evoked responses", Electroencephalography and clinical neurophysiology, Elsevier publishing company, Amsterdam – Printed in the netherlands.

[5] GBrson Eduardo Mog, Eduardo Parente kbeiro, "Zero Crossing determination by linear interpolation of sampled sinusoidal signal.

[6] Dajani, R.H., Purcell, D., Wong, W., Kunov, H., Picton, T.W. 2005. Recording Human Evoked Potentials That Follow the Pitch Contour of a Natural Vowel. IEEE Transactions on Biomedical Engineering 52, 1614-1618.

[7] In-Chul Yoo and Dongsuk Yook, "Robust voice activity detection using the spectral spectral peaks of vowel sounds". ETRI Journal,Volume 31,Number 4,August 2009.

[8] Johnson, K.L., Nicol, G.T., Kraus, N. 2005. Brain Stem Response to Speech: A Biological Marker of Auditory Processing. Ear & Hearing 26, 424-434.

[9] Russo, N., Nicol, T., Musacchia, G., Kraus, N. 2004. Brainstem responses to speech syllables. Clinical Neurophysiology 115, 2021-2030.

[10] Galbraith GC, Arbagey PW, Branski R. Intelligible speech encoded in the human brain stem frequency-following response. NeuroReport 1995; 6: 2363-2367.

[11] M.S. John, T.W. Picton, MASTER: a Windows program for recording multipleauditory steady-state response. Computer Methods and Programs in Biomedicine 61 (2000) 125–150, Elsevier

AUTHOR BIOGRAPHY

First Author – **Mr. Ranganadh Narayanam** is an Assistant Professor in the department of Electronics & Communications Engineering in IFHE Deemed University, Hyderabad, India. Mr.Narayanam, was a research student (research scholar) in the area of "Brain Stem Speech Evoked Potentials" under the guidance of Dr. Hilmi Dajani of University of Ottawa,Canada. He was also a research student (research scholar) in The University of Texas at San Antonio under Dr. Parimal A Patel,Dr. Artyom M. Grigoryan, Dr Sos Again, Dr. CJ Qian,. in the areas of signal processing and digital systems, control systems. He worked in the area of Brain Imaging in University of California Berkeley. Mr. Narayanam has done some advanced learning in the areas of DNA computing, String theory and Unification of forces, Faster than the speed of light theory with worldwide reputed persons and world's top ranked universities. Mr. Narayanam's research interests include neurological Signal & Image processing, DSP & DIP software & Hardware design and implementations, neurotechnologies. Mr. Narayanam can be contacted at rnara100@gmail.com, rnara100@uottawa.ca, rnaraya7@uwo.ca, ranganadh.narayanam@gmail.com, rnara100@ifheindia.org