Speech Enhancement using Temporal Masking and Fractional Bark Gammatone Filters

Teddy Surya Gunawan, Eliathamby Ambikairajah

School of Electrical Engineering and Telecommunications
The University of New South Wales,
NSW 2052, Australia
tsgunawan@ee.unsw.edu.au; ambi@ee.unsw.edu.au

Abstract

A speech enhancement technique based on the temporal masking properties of the human auditory system is presented. The noisy signal is divided into a number of sub-bands with fractional bark accuracy, and the sub-band signals are individually and adaptively weighted in the time domain according to a short-term temporal masking threshold to noise ratio estimate in each sub-band. Objective measures and informal listening tests demonstrate significant improvements over three well-known existing methods when tested with speech signals corrupted by various noises at signal to noise ratios of 0, 10, and 20 dB.

1. Introduction

The purpose of speech enhancement is to improve the performance of speech communication systems in noisy environments. Speech enhancement can be applied in many applications, such as in mobile communication systems, speech recognition, or hearing aids. The additive noise source may be wideband noise, in the form of a white or colored noise, or a periodic signal, such as hum noise or room reverberations.

Single channel speech enhancement is a more difficult task than multiple channel enhancement, since there is no independent source of information with which to help separate the speech and noise signals. The spectral subtraction algorithm is a well known solution to the speech enhancement (Boll 1979; Gustafsson, Nordholm, and Claesson 2001; Martin 1994; Tsoukalas, Mourjopoulos, and Kokkinakis 1997), in which noise is usually estimated during speech pauses.

Spectral subtraction is widely known to suffer from perceptible artifacts resulting from musical residual noise that is introduced into the enhanced speech by the method. In order to reduce the musical noise, various algorithms have been developed (Gustafsson et al. 2001; Tsoukalas et al. 1997; Virag 1999). In (Virag 1999) and (Tsoukalas *et al.* 1997), human auditory masking properties, i.e. simultaneous masking, were used to reduce the musical noise.

Recently, a new speech enhancement method known as speech boosting has been reported (Westerlund 2003). Instead of focusing on suppressing the noise, the method increases the relative power of the speech, thus acting as a speech booster. It is only active when speech is present, and remains idle when noise is present. As stated in (Westerlund 2003), the algorithm has proven to be robust, flexible, and versatile.

Functional models of the temporal masking effect of the human auditory system have recently been used with success in speech and audio coding to provide more efficient signal compression (Gunawan, Ambikairajah, and Sen 2003; Sinaga, Gunawan, and Ambikairajah 2003). Furthermore, a fractional bark filterbank resolution, i.e. 0.25 and 0.5 bark (Basic and Advanced Version), has been reported in (ITU 1998) to provide more accurate objective measurement of perceived audio quality (PEAQ). Therefore, it is expected that the use of fractional bark accuracy will provide more accurate temporal masking calculation in speech enhancement.

In this paper, we propose a novel speech enhancement method that employs a functional model of temporal masking, employing a fractional bark gammatone filterbank, based upon modifications to the speech boosting technique (Westerlund 2003).

To evaluate the performance of our algorithm, three other algorithms were implemented: spectral subtraction (Boll 1979), spectral subtraction with minimum statistics (Martin 1994), and speech boosting (Westerlund 2003). The PESQ (Perceptual Evaluation of Speech Quality, ITU-T P.862) measure was used here to benchmark the various methods.

2. Proposed Speech Enhancement Algorithm

Speech that has been contaminated by noise can be expressed as

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

$$x(n) = s(n) + v(n) \quad (1)$$

where $x(n)$ is the noisy speech, $s(n)$ is the clean speech signal and $v(n)$ is the additive noise source, all in the discrete time domain. As mentioned in section 1, the objective in speech enhancement is to suppress the noise resulting in an output signal $y(n)$ with a higher signal-to-noise ratio (SNR).

We propose a new speech enhancement algorithm that incorporates temporal masking, as shown in Fig. 1. By filtering the input signal $x(n)$ using a bank of *M* analysis filters, $h_m(n)$, the signal is divided into *M* sub-bands, each denoted by $x_m(n)$, where *m* is the sub-band index.

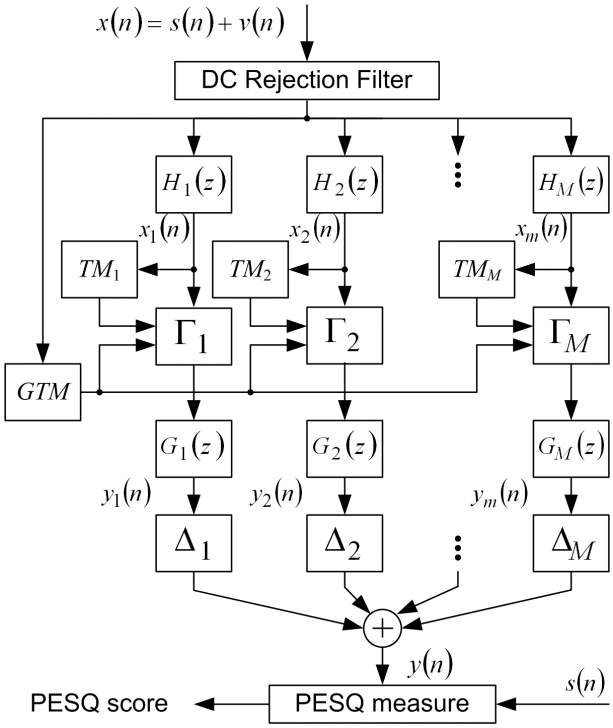

Figure 1: *Speech enhancement using temporal masking*

This filtering operation can be described in the time domain as

$$x_m(n) = x(n) * h_m(n) \quad (2)$$

where $m = 1, \ldots, M$. The global temporal masking threshold, GTM, and the temporal masking threshold in each sub-band, TM_m, are calculated from the noisy speech signal $x(n)$ and sub-band signal $x_m(m)$, respectively. The GTM and TM are used to calculate the gain (Γ_m) in each sub-band. The gain, Γ_m, is a weighting function that amplifies the signal in band *m* during speech activity.

The enhanced speech, $y(n)$, is then obtained by applying the synthesis filters, $g_m(n)$, and compensating the delay (Δ_m) in each sub-band as follows

$$y(n) = \sum_{m=1}^{M} y_m(n - \Delta_m) = \sum_{m=1}^{M} \Gamma_m x_m(n - \Delta_m) * g_m(n - \Delta_m) \quad (3)$$

Our objective is now to find a gain function, Γ_m, that weighs the input signal sub-bands, $x_m(n)$, based on temporal masking threshold to noise ratio (MNR). The MNR in each sub-band can be calculated by using the ratio of a short-term average temporal masking threshold, $P_m(n)$, and an estimate of the noise floor level, $Q_m(n)$ as given in equation (6). The short-term average temporal masking threshold in sub-band *m* is calculated as

$$P_m(n) = (1 - \alpha_m) P_m(n-1) + \alpha_m TM_m(n) \quad (4)$$

where α_m is a small positive constant (i.e. $\alpha_m = 0.0042, \forall m$) controlling the sensitivity of the algorithm to changes in temporal masking threshold, and acts as a smoothing factor. The slowly varying noise floor estimate for the *m*'th sub-band, $Q_m(n)$, is calculated as

$$Q_m(n) = \begin{cases} (1 + \beta_m) Q_m(n-1), & Q_m(n-1) \le P_m(n) \\ P_m(n), & Q_m(n-1) > P_m(n) \end{cases} \quad (5)$$

where β_m is a small positive constant (i.e. $\beta_m = 0.05, \forall m$) controlling how fast the noise floor level estimate in sub-band *m* adapts to changes in the noise environment.

The variables $P_m(n)$, $Q_m(n)$, $TM_m(n)$ and $GTM_m(n)$ are used to calculate the gain function $\Gamma_m(n)$ as follows,

$$\Gamma_m(n) = \gamma_m \frac{TM_m(n)}{GTM(n)} + (1 - \gamma_m) \frac{P_m(n)}{Q_m(n)} \quad (6)$$

where $0 \le \gamma_m \le 1$ is a positive constant controlling the contribution of the temporal masking threshold ratio and the short term MNR. Hence, the proposed algorithm still acts as a speech booster but the gain calculation $\Gamma_m(n)$ differs from (Westerlund 2003), which calculates the gain function from the short-term SNR.

In order to find the optimum γ_m, we evaluated the average quality improvement (see δ calculation in equation (18)) for a speech file (female English speaker) contaminated with car noise at 0, 10, and 20 dB SNRs at various γ_m. From the results of this experiment, shown in Figure 2, we found the optimum value to be $\gamma_m = 0.8, \forall m$.

Accepted after full paper review

Since the calculation of $\Gamma_m(n)$ involves a division, care must be taken to ensure that the quotient does not become excessively large due to a small $Q_m(n)$. In a situation with a very high MNR, $\Gamma_m(n)$ will become very large if no limit is imposed on this function.

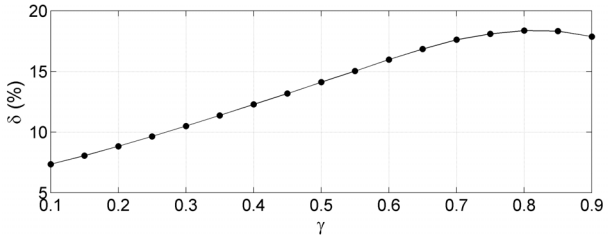

Figure 2: *Quality improvement for various* γ

Therefore, a limiter can be applied on $\Gamma_m(n)$ as follows:

$$\Gamma_m(n) = \begin{cases} \Gamma_m(n), & \Gamma_m \le C_m \\ C_m & \Gamma_m > C_m \end{cases} \quad (7)$$

where C_m is some positive constant. By using the same experiment to find the optimum γ_m, setting $C_m = 8\,dB \approx 2.51$ provides a suitable limiter for the gain function.

3. Fractional Bark Gammatone Filterbank

In this paper, a fractional bark gammatone filterbank was employed to filter the signal $x(n)$ into its sub-band signals $x_m(n)$. A DC rejection filter was applied to remove the subsonic components of the input signals. In addition, the optimum number of filter coefficients required was evaluated and the delay compensation for each sub band was calculated.

3.1. DC Rejection Filter

We designed a fourth order Butterworth high pass filter with a cut-off frequency of 20 Hz to remove the subsonic components of the input signals. The filter was implemented as a cascade of two second order IIR-filters.

$$H_{DC}(z) = \frac{1-2z^{-1}+z^{-2}}{1+az^{-1}+bz^{-2}} \times \frac{1-2z^{-1}+z^{-2}}{1+cz^{-1}+dz^{-2}} \quad (8)$$

where a = -1.9878047, b = 0.98804997, c = -1.9711486, and d = 0.97139181, for $fs = 8000$ Hz.

3.2. Gammatone Filters

For the analysis filter, we used gammatone filters as they resemble the shape of human auditory filters (Kubin and Kleijn 1999). These were implemented using FIR filters. To achieve perfect reconstruction, $g_m(n)$, are the time reverse of the analysis filters, $h_m(n)$. The analysis filter for each sub-band m is obtained using the following expression,

$$h_m(n) = a_m(nT)^{N-1} e^{-2\pi b BW_m nT} \cos(2\pi f_{cm} nT + \varphi) \quad (9)$$

where f_{cm} is the centre frequency for each sub-band m, T is the sampling period, and N is the gammatone filter order ($N = 4$). For $fs = 8000$ Hz, the total number of sub-bands, M, is dependent on the bark resolution, dz. The parameter n is the discrete time sample index, and $n = 0 \ldots Nf_m$ where Nf_m is the length of each filter within the filterbank. BW_m is the critical bandwidth at a particular center frequency, b = 1.65, and the a_m were selected for each filter so as to normalize the filter gain to 0 dB.

3.3. Spacing of the Filters

The gammatone filters were spaced linearly on the Bark scale, or critical-band rate scale. The critical band number z (in Bark) is related to the linear frequency f (in Hz), as follows (Schroeder, Atal, and Hall 1979)

$$z(f) = 7 \cdot a\sinh\left(\frac{f}{650}\right), \; f(z) = 650 \cdot \sinh\left(\frac{z}{7}\right) \quad (10)$$

The frequency borders of the filters range from $f_L = 80$ Hz to $f_U = 4000$ Hz. The widths and spacing of the filter bands correspond to a resolution of dz. The number of sub-bands M is then calculated as follows,

$$M = \left\lfloor \frac{z(f_U) - z(f_L)}{dz} \right\rfloor \quad (11)$$

A spacing of $dz = 0.5$ Bark required 34 filters, while a spacing of $dz = 0.25$ required 68 filters in order to cover the frequency range of 0 to 4 kHz. The lower, upper, and center frequency for each sub band in Bark scale can be calculated as follows,

$$\begin{aligned} z_{lm} &= z(f_L) + m \cdot dz, \\ z_{um} &= \min(z(f_L) + m \cdot dz, z(f_U)), \\ z_{cm} &= \tfrac{1}{2}(z_{lm} + z_{um}), \end{aligned} \quad (12)$$

where $m = 1, \ldots, M$. Subsequently, the center frequency and the bandwidth in Hz can be determined as follows,

$$f_{cm} = f(z_{cm}), \; BW_m = f(z_{um}) - f(z_{lm}) \quad (13)$$

In order to find the optimum value of dz for our speech enhancement method, we evaluated the average quality improvement and processing time for various dz values at 0, 10, and 20 dB SNRs, as seen in Figure 3. From Figure 3, we found that setting $dz = 0.25$ provides the optimum value in terms of speech quality and processing time. Hence, $dz = 0.25$

Accepted after full paper review

was used throughout our experiments. The frequency responses of gammatone filters for this value of $dz = 0.25$ are shown in Figure 4.

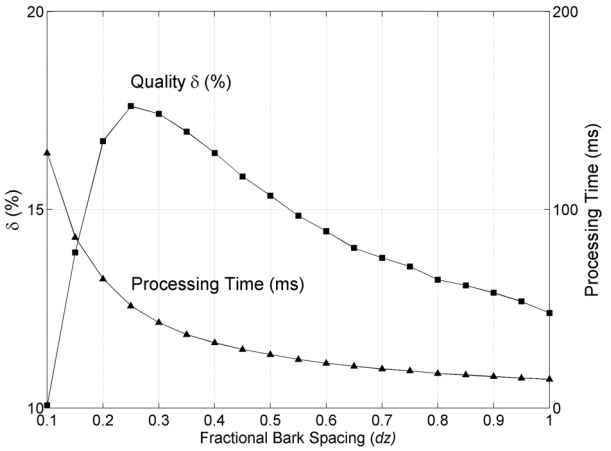

Figure 3: *Fractional bark spacing versus quality and processing time*

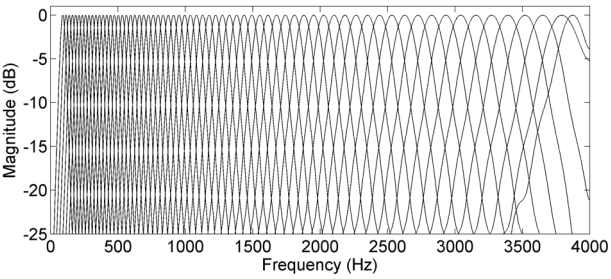

Figure 4: *¼ Bark spacing (68 filters)*

3.4. Optimum Number of Filter Coefficients (Nf_m)

The number of coefficients required to implement the analysis/synthesis filter bank depends on the impulse response of the gammatone filters. The low frequency filters need more coefficients as compared with the high frequency filters. The length of each filter within the filterbank, Nf_m, can be optimised by evaluating the non-zero gammatone filter response in each sub-band. The optimum length of the filter Nf_m in samples for each sub-band is given by

$$Nf_m = \min\left(Nf_{\max}, \text{round}\left(fs / f_{cm}\right) \cdot 25\right) \qquad (14)$$

where f_{cm} is the centre frequency of the filter in Hz and $Nf_{\max} = 1024$ is the maximum length of filter coefficients.

3.5. Delay Compensation

By employing the optimum length of the filter in each sub-band, Nf_m, the amount of filter delay accumulated by each sub-band is different. Without compensation for this delay, the reconstruction of the sub-band signal components will lead to an incoherent output signal. The total amount of delay compensation necessary for subband *m* is simply $\Delta_m = Nf_m - 1$, where Nf_m is the optimum filter order calculated as in equation 14.

4. Temporal Masking

Temporal masking is a time domain phenomenon in which two stimuli occur within a small interval of time, and plays an important role in human auditory perception. Forward temporal masking occurs when a masker precedes the signal in time, while backward masking occurs when the signal precedes the masker in time. Forward masking is the more important effect since the duration of the masking effect can be much longer, depending on the duration of the masker.

The forward masking model used in this paper is based on (Jesteadt, Bacon, and Lehman 1982), and has been used and optimised in our previous papers for speech and audio coding (Gunawan *et al*. 2003; Sinaga *et al*. 2003). Based on the forward masking experiments carried out by (Jesteadt *et al*. 1982), forward masking level *FM* can be well-fitted to psychoacoustic data using the following equation:

$$FM = a\left(b - \log_{10}\Delta t\right)\left(L - c\right) \qquad (15)$$

where *FM* is the amount of forward masking in dB, Δt is the time difference between the masker and the maskee in milliseconds, L is the masker level in dB, and *a*, *b*, and *c*, are parameters that can be derived from psychoacoustic data. To simplify the masking calculation, *a*, *b*, and *c* were set to 0.7, 2.3, and 20, respectively. Note that these parameters can be further optimised.

To evaluate the amount of forward masking, the current frame of 32 ms was subdivided into four sub-frames as shown in Figure 5. The forward masking level FM_j was calculated for the *j*th sub-frame using the energy, L_j, accumulated over the previous frame and all sub-frames up to the current sub-frame.

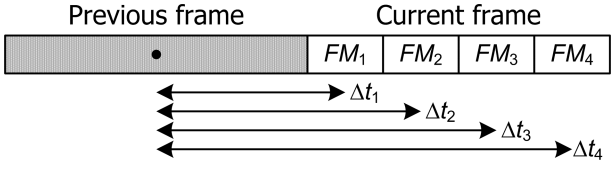

Figure 5: *Calculation of forward masking*

The temporal amount of masking *TM* is then chosen as follows

$$TM = 10^{\frac{1}{10}\max\left\{FM_1, FM_2, FM_3, FM_4\right\}} \qquad (16)$$

Note that the calculation of a temporal masking threshold every 8 ms was considered adequate since

this provides a good approximation to the decay effect that lasts around 200 ms. The temporal masking thresholds are calculated for each sub-band, $TM_1,\ldots,TM_M$, from $x_m(n)$ and GTM from $x(n)$.

5. Performance Evaluation

In order to assess the performance of the proposed algorithm to enhance noisy signals, a large number of simulations were performed. Six speech files were taken from EBU SQAM data set including English female and male speakers, French female and male speakers, and German female and male speakers. The length of the files is between 17 and 20 seconds.

The sampling frequency was 8 kHz, and the frame size was 256 samples (32 ms). Several algorithms were implemented and compared including spectral subtraction, **SS**, (Boll 1979), spectral subtraction with minimum statistics, **SSMS**, (Martin 1994), speech boosting, **SB**, (Westerlund 2003), and the proposed method speech boosting exploiting temporal masking, **SBTM**.

5.1. Addition of Noise to Test Data

Different types of background noises from the NOISEX-92 database have been used including car noise, white noise, pink noise, F16 noise, factory noise, and babble noise. The variance of noise has been adjusted to obtain SNRs in the recorded signals ranging from 0 dB to 20 dB, as follows:

$$x(n) = s(n) + \left(\sqrt{\frac{Var(s(n))}{Var(v(n))}} \times \frac{1}{10^{SNR/10}} \right) \cdot v(n) \quad (17)$$

5.2. Objective Measures

The PESQ (Perceptual Evaluation of Speech Quality) measure (ITU 2001), which was recently adopted as an ITU-T recommendation (P.862), was utilised for the objective evaluation. Other objective measures such as Itakura-Saito distortion, Articulation Index, Segmental SNR, and SNR have been correlated to subjective tests at 59%, 67%, 77%, and 24%, respectively (Quackenbush, Barnwell, and Clements 1988), while the PESQ has a 93.5% correlation with subjective tests (ITU 2001), although obviously these figures were obtained using different data sets and subjective experiments.

To evaluate the performance of the speech enhancement algorithms, we developed a new measure to assess the improvement achieved. Suppose that we have $PESQ_{ref}$ which is the PESQ score for the reference clean speech, $s(n)$, and the corrupted speech, $x(n)$. The PESQ score of the enhanced speech, $y(n)$, was also measured and denoted as $PESQ_{proc}$. Therefore, we can derive a new value, δ, which measures the PESQ improvement achieved by the algorithm as follows

$$\delta = \frac{PESQ_{proc} - PESQ_{ref}}{PESQ_{ref}} \times 100\% \quad (18)$$

A total of 108 data sets from six speech files, six noises, and three SNRs for each method were simulated. The average quality improvement, δ, achieved by various speech enhancement methods is shown in Figure 6. Note that the δ results for various speech files and noises were averaged for 0, 10, and 20 dB SNRs. From these results, the proposed temporal masking-based speech boosting method seems to outperform other methods for all SNRs.

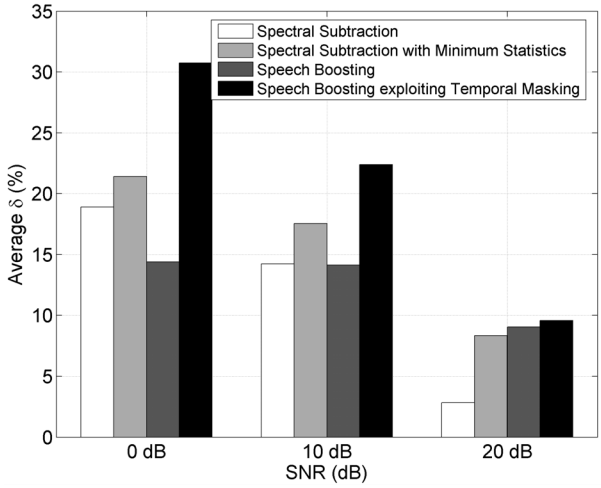

Figure 6: *Average δ (%) for various algorithms*

In order to analyze the performance of our proposed method in more detail, the average of quality improvement at 0, 10, and 20 dB SNRs for various noises is shown in Table 1.

Table 1: Average PESQ improvement δ (%) for various noise types using spectral subtraction (SS), spectral subtraction with minimum statistics (SSMS), speech boosting (SB), and speech boosting with temporal masking (SBTM).

Noise	SS	SSMS	SB	SBTM
Car noise	13.27	15.26	10.49	*17.56*
White noise	16.22	24.81	16.39	*29.76*
Pink noise	16.43	22.28	15.40	*26.60*
F16 noise	11.21	16.23	12.81	*22.15*
Factory noise	12.70	11.84	12.65	*20.20*
Babble noise	2.15	4.20	7.44	*9.12*

The best δ result for each type of noise condition is shown in italics, from which it can be seen that our proposed method provides a better PESQ improvement than the three other methods. The best improvement is

Accepted after full paper review

achieved for the white noise while the least improvement is achieved for the babble noise. The babble noise is a speech conversation in the background. Therefore, our algorithm might also misclassify and boost the babble noise as speech.

Table 2: Average PESQ improvement δ (%) for different speech files using spectral subtraction (SS), spectral subtraction with minimum statistics (SSMS), speech boosting (SB), and speech boosting with temporal masking (SBTM).

Speech	SS	SSMS	SB	SBTM
English male	8.66	12.69	9.78	*20.70*
English female	11.17	15.61	11.55	*18.58*
French male	13.82	17.31	11.71	*19.18*
French female	10.09	13.42	9.35	*14.42*
German male	18.31	25.85	19.65	*34.01*
German female	9.93	9.73	13.14	*18.51*

Table 2 shows the average of quality improvement at 0, 10, and 20 dB SNRs for various speech files. The best δ result for each individual speech files is shown in italics. While the table shows that our proposed algorithm outperforms other algorithms, it is also reveals that our algorithm improves male speech better than female speech.

6. Conclusion

We have presented a fractional bark gammatone filter for speech enhancement based on a short-term temporal masking threshold to noise ratio (MNR). The performance of our proposed algorithm was compared with three other standard speech enhancement methods over six different noise types and three SNRs. PESQ results reveal that the proposed algorithm outperforms the other algorithms by 7-15% depending on the SNR. In the particularly demanding 0 dB SNR condition, the new technique achieves at least a 40% relative improvement in delta PESQ over any of the existing methods compared. Hence, it appears that the temporal masking threshold based algorithm with fractional bark accuracy has good potential for speech enhancement applications across many types and intensities of environmental noise. Further research is required to fine tune the parameters for different speech and/or noise characteristics.

7. References

Beerends, J. G., Hekstra, A. P., Rix, A. W., & Hollier, M. P. (2002). Perceptual Evaluation of Speech Quality.

Boll, S. F. (1979). Suppresion of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing, 27*(2), pp. 113-120.

Gunawan, T. S., Ambikairajah, E., & Sen, D. (2003, December). *Comparison of Temporal Masking Models for Speech and Audio Coding Applications.* Paper presented at the International Symposium on Digital Signal Processing and Communication Systems, pp. 99-103.

Gustafsson, H., Nordholm, S. E., & Claesson, I. (2001). Spectral Subtraction Using Reduced Delay Convolution and Adaptive Averaging. *IEEE Transactions on Speech and Audio Processing, 9*(8), pp. 799-807.

ITU. (1998). *ITU-R BS.1387, Method for the Objective Measurements of Perceived Audio Quality*. Geneva: International Telecommunications Union.

ITU. (2001). *ITU-T P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. Geneva: International Telecommunication Union.

Jesteadt, W., Bacon, S. P., & Lehman, J. R. (1982). Forward masking as a function of frequency, masker level, and signal delay. *Journal of Acoustic Society of America, 71*(4), pp. 950-962.

Kubin, G., & Kleijn, W. B. (1999). *On speech coding in a perceptual domain.* Paper presented at the International Conference on Acoustic, Speech, and Signal Processing, pp. 205-208.

Martin, R. (1994). *Spectral Subtraction Based on Minimum Statistics.* Paper presented at the Europe Signal Processing Conference, Edinburgh, Scotland, pp. 1182-1185.

Quackenbush, S. R., Barnwell, T. P., & Clements, M. A. (1988). *Objective Measures of Speech Quality*. Englewood Cliffs: Prentice Hall.

Schroeder, M. R., Atal, B. S., & Hall, J. L. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of Acoustic Society of America, 66*, pp. 1647-1652.

Sinaga, F., Gunawan, T. S., & Ambikairajah, E. (2003). *Wavelet Packet Based Audio Coding Using Temporal Masking.* Paper presented at the Int. Conf. on Information, Communications and Signal Processing, Singapore.

Tsoukalas, D. E., Mourjopoulos, J. N., & Kokkinakis, G. (1997). Speech Enhancement Based on Audible Noise Suppression. *IEEE Transactions on Speech and Audio Processing, 5*(6), pp. 497-514.

Virag, N. (1999). Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System. *IEEE Transactions on Speech and Audio Processing, 7*(2), pp. 126-137.

Westerlund, N. (2003). *Applied Speech Enhancement for Personal Communication.* PhD Thesis, Blekinge Institute of Technology.

Accepted after full paper review