# FORWARD MASKING THRESHOLD ESTIMATION USING NEURAL NETWORKS AND ITS APPLICATION TO PARALLEL SPEECH ENHANCEMENT

**T. S. GUNAWAN[1], O. O. KHALIFA[1], E. AMBIKAIRAJAH[2]**

[1]*Electrical and Computer Engineering Department, International Islamic University Malaysia, P.O. Box 10, Kuala Lumpur, 50728, MALAYSIA*

[2]*School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, AUSTRALIA*

*E-mail: tsgunawan@iiu.edu.my*

***ABSTRACT***: Forward masking models have been used successfully in speech enhancement and audio coding. Presently, forward masking thresholds are estimated using simplified masking models which have been used for audio coding and speech enhancement applications. In this paper, an accurate approximation of forward masking threshold estimation using neural networks is proposed. A performance comparison to the other existing masking models in speech enhancement application is presented. Objective measures using PESQ demonstrates that our proposed forward masking model, provides significant improvements (5-15 %) over four existing models, when tested with speech signals corrupted by various noises at very low signal to noise ratios. Moreover, a parallel implementation of the speech enhancement algorithm was developed using Matlab parallel computing toolbox.

***KEYWORDS***: *Human auditory system, forward masking, speech enhancement, PESQ, parallel algorithm.*

# 1. INTRODUCTION

Forward masking is a time domain phenomenon in which a masker precedes the signal in time. Forward masking psychoacoustic data depends on four dimensions, i.e. frequency, masker level, time difference between masker and maskee, and masker signal duration [1]. The current forward masking models do not fully take into account all the four dimension of forward masking data.

Functional models of the forward masking effect of the human auditory system have recently been used with success in speech and audio coding to provide more efficient signal compression [2, 3]. Furthermore, forward masking has been used for speech enhancement [4] using the speech boosting technique [5]. Instead of focusing on suppressing the noise, the speech boosting technique 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.

Jesteadt's forward masking model [6] provides a reasonable approximation to the forward masking effect. Strope *et al.* [7] extended the Jesteadt experiment to 120 ms. In

Jesteadt's and Najafzadeh's forward masking models [6, 8], only masker level and delay have been taken into account. While in [9], Gunawan and Ambikairajah have refined the model to reflect forward masking data more accurately by averaging several parameters across frequencies. Currently, the majority of these works focus on formulating mathematical models of the forward masking. Such models are often too general. Further refinement of the model requires software that can do curve-fitting of multi-dimensional data. Nevertheless, for this purpose, we utilise neural network to better approximate forward masking threshold.

To evaluate the performance of our forward masking model, five speech enhancement algorithms were implemented: spectral subtraction [10], spectral subtraction with minimum statistics [11], speech boosting [5], speech boosting using forward masking model 1 [4] and forward masking model 2 [9]. The Perceptual Evaluation of Speech Quality or PESQ (ITU-T P.862) measure was used here to benchmark the various methods.

Speech enhancement algorithm exploiting temporal masking properties of human auditory system has a very high computation requirement, especially when the noisy speech signal is long or the number of subbands is high. Recent advances in multi-core system make it a natural choice and viable option for solving high computation requirements of the speech enhancement algorithm. Therefore, the objective of this paper is two-folds: to evaluate the performance of our forward masking model in terms of enhanced speech quality and to implement and evaluate parallel speech enhancement algorithm on a multi-core system. The rest of the sections are organized as follows: Section 2 discusses the development of forward masking models using neural networks. Section 3 describes the sequential speech enhancement algorithm while Section 4 discusses the parallel implementation of speech enhancement algorithm. Experimental results and analyses are discussed in Section 5 for the sequential and parallel algorithms. Finally, Section 6 concludes this paper.

## 2. FORWARD MASKING MODELS USING NEURAL NETWORKS

Neural network has been applied for various applications within the following broad categories: function approximation (or regression analysis), classification, and data processing (filtering, clustering, blind source separation, etc). Brown *et al.* [12] applied non-recurrent neural networks for simultaneous masking modelling. In this paper, neural networks is employed to approximate the forward masking threshold for the three input parameters, including frequency, masker level, and delay.

By taking into account the threshold in quiet ( $TIQ$ ) the absolute threshold of forward masking ( $FM$ ) can be calculated using the equation we have developed below:

$$FM\left(f, L_m, \Delta t, T_s\right) = M\left(f, Lm, \Delta t\right) + TIQ\left(f, T_s\right) \quad (1)$$

As stated in [13], the threshold in quiet is a function of frequency and signal duration. By curve-fitting a set of 120 data points compiled from [13], we approximated the threshold in quiet to be as follows:

- $TIQ(f, T_s)$ for signal with long duration ($T_s \geq 500$ ms) can be approximated as ($f$ in kHz):

$$TIQ(f, T_s \geq 500) = 3.64(f)^{-0.8} + 6.5e^{-0.6(f-3.3)^2} + 0.001(f)^4 \quad (2)$$

- $TIQ(f, T_s)$ for signal with duration $Ts < 500$ ms, can be approximated as

$$TIQ(f, T_s) = TIQ(f, T_s < 500) + \left(7.53 - 6.5 \cdot 10^{-13} f^3\right) \log_{10}(500 - T_s) \quad (3)$$

The amount of forward masking $M(f, Lm, \Delta t)$ can be approximated using feed forward neural network as shown in Fig. 1.

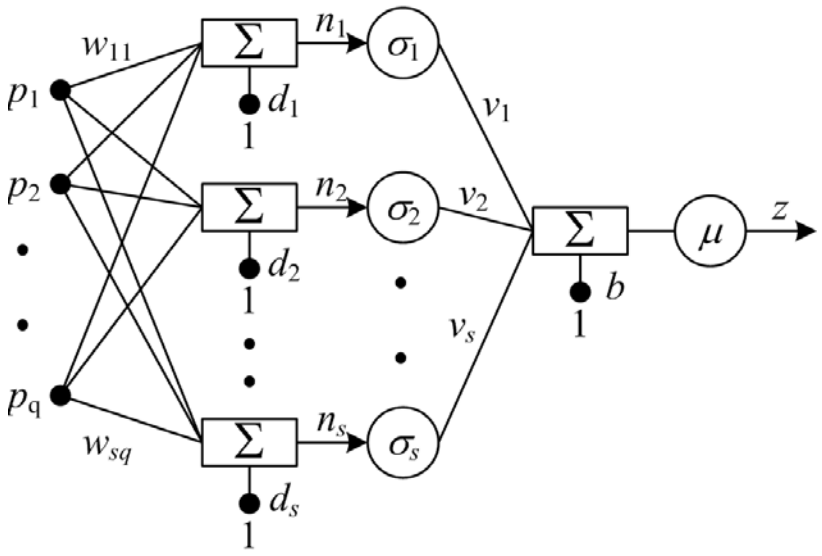


Fig. 1: Amount of forward masking modelled by neural networks.

The network configuration as shown in Fig. 1 with 1 hidden layer could approximate any function [14]. To avoid over-fitting to the training data, the Bayesian regularization as proposed in [15] was used. Figure 2 shows the amount of forward masking against $Lm$ and $\Delta t$ at frequency of 500Hz using neural network. Similar plots can be obtained for various frequencies, thus providing a more accurate estimation of forward masking data.

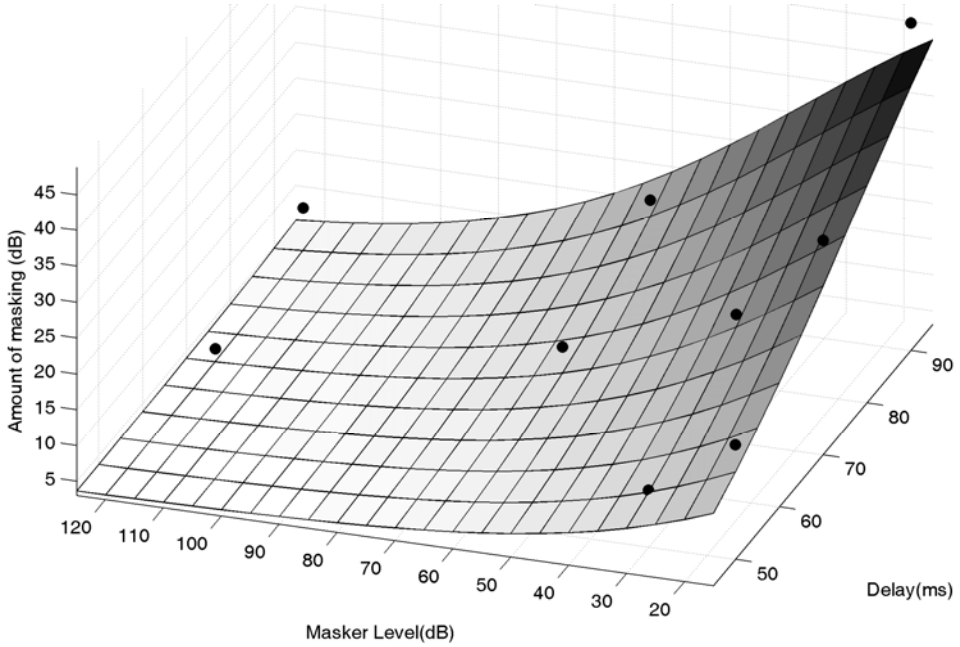


Fig. 2: Amount of forward masking estimation at 500Hz.

## 3. SPEECH ENHANCEMENT

This section presents the incorporation of our model to fit the speech enhancement algorithm developed in [4]. Speech that has been contaminated by noise can be expressed as

$$x(n) = s(n) + v(n) \tag{4}$$

where $x(n)$ is the noisy speech, $s(n)$ is the clean speech signal and $v(n)$ is the additive noise, all of which are in the discrete time domain. The objective in speech enhancement is to suppress the noise, thus resulting in an output signal $y(n)$ that has a higher signal-to-noise ratio (SNR).

The speech enhancement algorithm that incorporates forward masking [4] is shown in Fig. 3. By filtering the input signal $x(n)$ using a bank of $M$ analysis filters, the signal is divided into $M$ subbands, each denoted by $x_m(n)$, where $m$ is the subband index.

This filtering operation can be described in the time domain as $x_m(n) = x(n) * h_m(n)$ where $m = 1,\ldots,M$. and $h_m(n)$ is the impulse response of the $m^{\text{th}}$ filter. The global forward masking threshold (*GFM*) and the forward masking threshold in each subband ($FM_m$) are calculated from the noisy speech signal $x(n)$ and subband signal $x_m(m)$, respectively. The *GFM* and $FM_m$ are used to calculate the gain ($\Gamma_m$) in each subband. The gain, $\Gamma_m$, is a weighting function that amplifies the signal in band $m$ during speech activity.

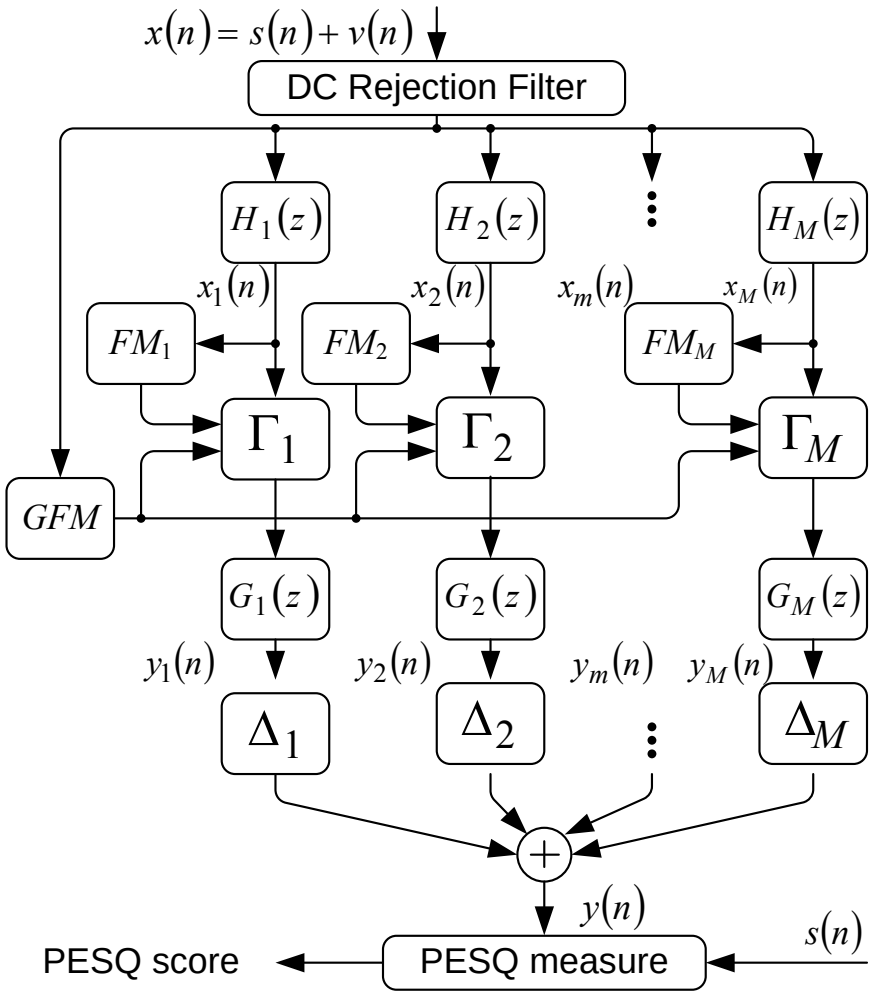


Fig. 3: Speech enhancement using forward masking.

The enhanced speech, $y(n)$, is then obtained by applying the synthesis filters, $g_m(n)$, and compensating the delay ($\Delta_m$) in each subband as follows:

$$y(n) = \sum_{m=1}^{M} y_m(n - \Delta_m) = \sum_{m=1}^{M} \Gamma_m x_m(n - \Delta_m) * g_m(n - \Delta_m) \tag{5}$$

Our objective is now to find a gain function, $\Gamma_m$, that weights the input signal subbands, $x_m(n)$, based on forward masking threshold to noise ratio (MNR). The MNR in each subband can be calculated by using the ratio of a short-term average forward masking threshold, $P_m(n)$, and an estimate of the noise floor level, $Q_m(n)$ as given in Eqn (8). The short-term average temporal masking threshold in subband *m* is calculated as

$$P_m(n) = (1 - \alpha_m) P_m(n-1) + \alpha_m FM_m(n) \tag{6}$$

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

$$Q_m(n) = \begin{cases} (1+\beta_m)Q_m(n-1), & Q_m(n-1) \le P_m(n) \\ P_m(n), & Q_m(n-1) > P_m(n) \end{cases} \tag{7}$$

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

The variables $P_m(n)$, $Q_m(n)$, $FM_m(n)$ and $GFM_m(n)$ are combined in a novel manner in order to calculate the gain function $\Gamma_m(n)$ as follows,

$$\Gamma_m(n) = \gamma_m \frac{FM_m(n)}{GFM(n)} + (1-\gamma_m)\frac{P_m(n)}{Q_m(n)} \tag{8}$$

where $0 \le \gamma_m \le 1$, i.e. $\gamma_m = 0.9, \forall m$, is a positive constant controlling the contribution of the forward masking threshold ratio and the short term MNR.

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.

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

$$\Gamma_m(n) = \begin{cases} \Gamma_m(n), & \Gamma_m \le C_m \\ C_m & \Gamma_m > C_m \end{cases} \tag{9}$$

where $C_m = 0.3529m + 2$ dB provides a suitable limiter for the gain function.

## 4. PARALLEL SPEECH ENHANCEMENT ALGORITHM

The design of an efficient parallel speech enhancement algorithm can be a challenging task. First step in the parallelization of any sequential code is to identify which part of code that takes the longest execution time. Using Matlab profiling tool, it was identified that the calculation of forward masking threshold and gain calculation for each subband (see Eqn 5 to 9) were taking the longest execution time. In this paper, we will utilize the Matlab parallel computing toolbox. The hardware used was an AMD quad core 2.5 GHz system with 2 GBytes of memory.

Master-slave paradigm is used in the parallelization. To achieve a scalable parallel implementation of speech enhancement algorithm, we used the data-parallel or single program multiple data (SPMD) programming model. A single program was written for both master and slave processes that asynchronously execute on each node. In particular,

all processes will work on different piece of data. There are two data partition schemes available in speech enhancement algorithm, time partition and frequency (subband) partition. In time partition a long noisy speech file is partitioned into smaller time and processed individually. While in frequency or subband partition, the total number of subbands is divided and distributed into a number of slaves. As the calculation of temporal masking requires the information from previous frames, it is obvious that subband partition is more appropriate for parallelization. Hence, it will be used in our implementation.

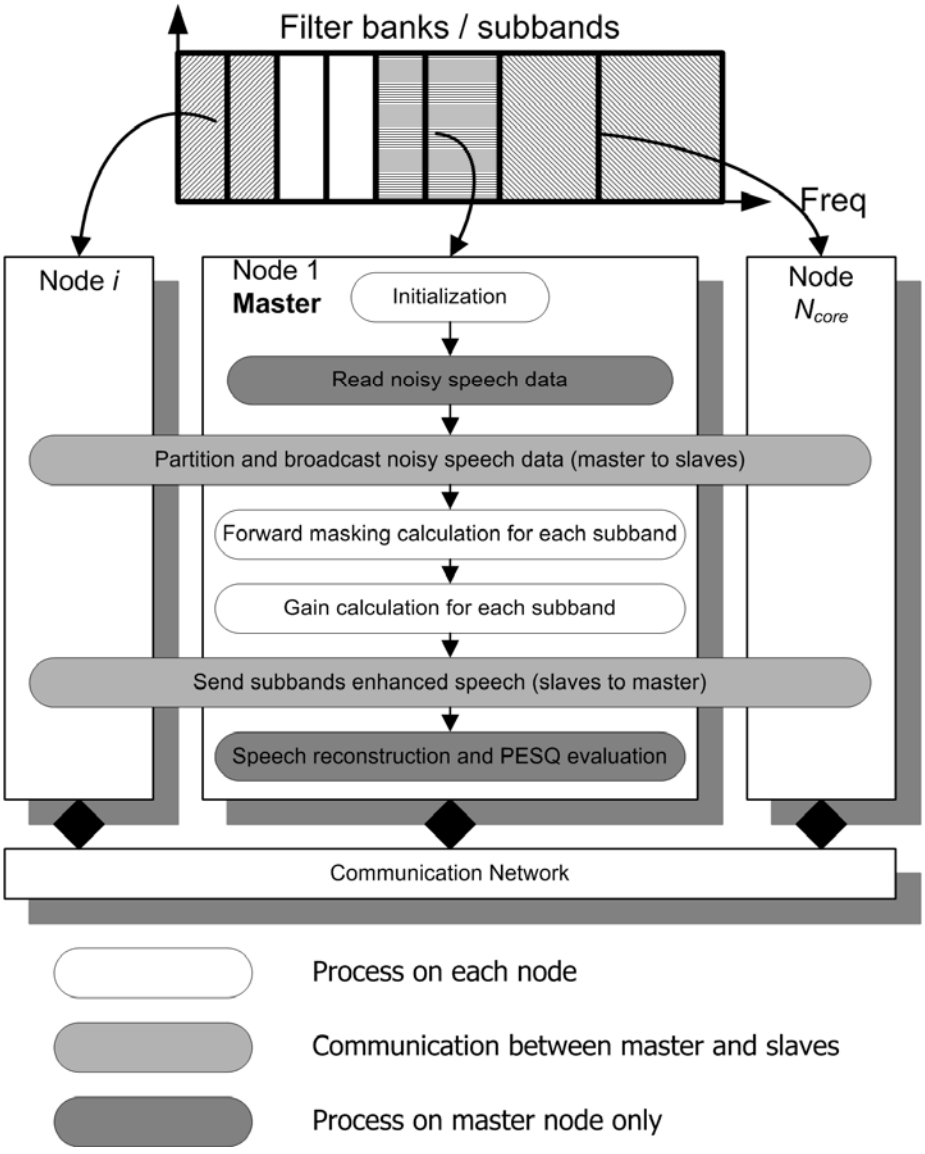


Fig. 4: Flowchart of parallel speech enhancement algorithm.

Figure 4 shows the flowchart of parallel speech enhancement algorithm that can be implemented on multi-core system and/or cluster system. Initially, the parallel program starts with initialization at every node. Of the two communication schemes available in

Matlab parallel computing toolbox, i.e. distributed array and message passing, we will use message passing scheme as it provides more flexible communication scheme. Then, a noisy speech signal is partitioned (using subband partition) and distributed to $N_{core}$ configured in master-slave fashion. After that, each slave is then filtered the noisy speech signal accordingly, calculates the forward masking threshold, determines the gain for each subband, and applies the gain for each subband signal. After obtaining denoise speech signal for each subband, each slave then sends the results to the master node. Finally, speech reconstruction and PESQ evaluation are applied at the master node.

## 5. PERFORMANCE EVALUATION

In this section, the performance of sequential code, in terms of subjective and objective quality of the enhanced speech, was evaluated. Furthermore, the performance of parallel code, in terms of speedup for various numbers of cores, was presented.

### 5.1 Subjective and Objective Quality

In order to assess the performance of the new forward masking model in enhancing speech 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 was 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**[10], spectral subtraction with minimum statistics, **SSMS**[11], speech boosting, **SB**[5], speech boosting using forward masking model 1, **SBFM1**[4], speech boosting using forward masking model 2, **SBFM2**[9], and speech boosting using the developed neural networks forward masking model, **SBFM3**.

Different types of background noises from the NOISEX-92 and AURORA database have been used - including car, white noise, pink noise, F16, factory, babble, airport, exhibition, restaurant, street, subway and train noise. The variance of noise has been adjusted to obtain -5 dB, 0 dB, 5 dB, and 10 dB SNRs.

The PESQ (Perceptual Evaluation of Speech Quality, ITU-T P.862) measure [16] was utilised for the objective evaluation. Note that, the PESQ has a 93.5% correlation with subjective tests [16]. 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, $\delta$, which measures the PESQ improvement achieved by the algorithm as follows:

$$\delta = \frac{PESQ_{proc} - PESQ_{ref}}{PESQ_{ref}} \times 100\% \tag{10}$$

A total of 288 data sets from six speech files, twelve noises, and four SNRs for each method were simulated. The average quality improvement, $\delta$, achieved by various speech enhancement methods is shown in Figure 4. Note that the $\delta$ results for various speech files and noises were averaged for -5, 0, 5, and 10 dB SNRs. From these results, the speech boosting technique incorporating neural networks forward masking model outperforms other methods for all SNRs.

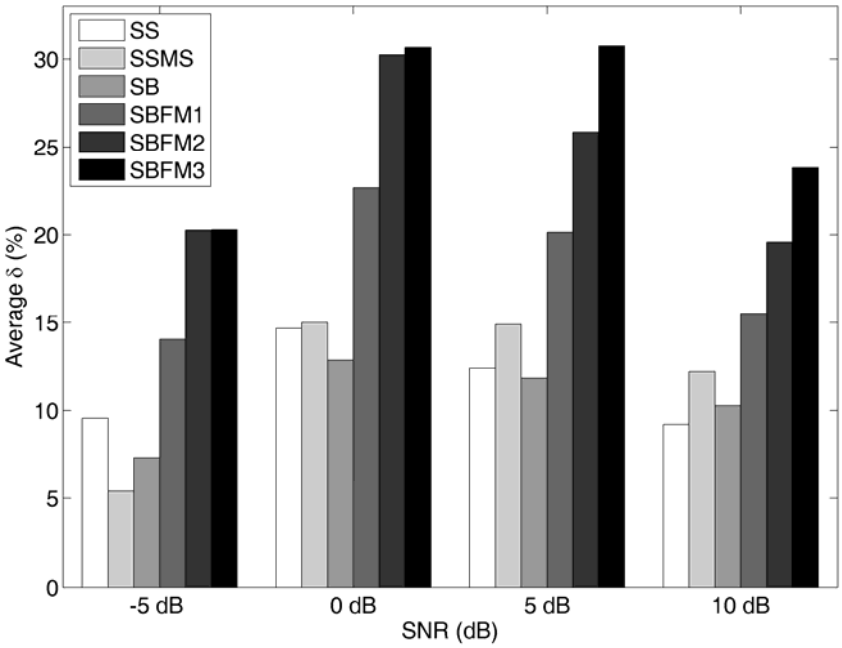


Fig. 5: Average $\delta$ (%) for various algorithms.

In order to analyse the performance of our proposed method in more detail, the average of quality improvement at -5, 0, 5, and 10 dB SNRs for various noises is shown in Table 1. The best $\delta$ result for each type of noise condition is shown in bold, from which it can be seen that our method using neural networks forward masking model provides a better PESQ improvement than the five other methods tested.

Table 2 shows the average of quality improvement at -5, 0, 5 and 10 dB SNRs for various speech files. The best $\delta$ result for each individual speech file is shown in bold. The table shows that more accurate forward masking threshold calculation leads to a better and enhanced speech quality. Furthermore, informal listening test confirm that the speech processed with the proposed algorithm sounds more pleasant to a human listener than those obtained by other algorithms.

Table 1: Average PESQ improvement $\delta$ (%) for various noise types.

| *Noise Types* | *SS* | *SSMS* | *SB* | *SBFM1* | *SBFM2* | *SBFM3* |
|---|---|---|---|---|---|---|
| *Car* | 19.88 | 18.16 | 11.80 | 21.04 | 22.09 | **23.30** |
| *White* | 17.50 | 28.33 | 15.81 | 21.58 | **34.25** | 33.67 |
| *Pink* | 22.73 | 28.90 | 16.93 | 27.41 | 37.28 | **37.32** |
| *F16* | 16.48 | 18.81 | 13.62 | 23.59 | 29.92 | **30.43** |
| *Factory* | 18.28 | 12.47 | 13.79 | 25.65 | 31.75 | **32.07** |
| *Babble* | 2.61 | 1.65 | 7.14 | 13.76 | 18.12 | **19.76** |
| *Airport* | 6.16 | 3.73 | 7.83 | 12.77 | 16.59 | **17.67** |
| *Exhibition* | 11.64 | 5.54 | 11.79 | 18.30 | 30.10 | **30.72** |
| *Restaurant* | 5.02 | 2.06 | 4.34 | 10.54 | 17.78 | **17.96** |
| *Street* | 8.59 | 9.45 | 12.82 | 18.63 | 15.86 | **17.06** |
| *Subway* | 4.29 | 7.49 | 11.57 | 20.18 | 34.42 | **34.51** |
| *Train* | 14.92 | 15.57 | 13.20 | 19.88 | 20.74 | **21.99** |

Table 2: Average PESQ improvement $\delta$ (%) for different speech files.

| *Speech Files* | *SS* | *SSMS* | *SB* | *SBFM1* | *SBFM2* | *SBFM3* |
|---|---|---|---|---|---|---|
| *English male* | 6.24 | 4.32 | 5.57 | 12.37 | 24.24 | **24.74** |
| *English female* | 8.79 | 9.08 | 9.61 | 15.65 | 26.00 | **26.23** |
| *French male* | 15.17 | 15.67 | 11.67 | 21.94 | 28.38 | **29.11** |
| *French female* | 10.83 | 11.46 | 9.36 | 14.69 | 19.31 | **19.73** |
| *German male* | 21.89 | 27.35 | 21.27 | 36.03 | 34.75 | **36.33** |
| *German female* | 11.13 | 8.20 | 12.84 | 16.00 | 21.76 | **22.08** |

### 5.2 Parallel Performance

The computing environment used in this research was AMD Phenom Quad-Core Processor 2.5 GHz system with 2 GBytes of memory. This section is intended to analyse the parallel performance of the speech enhancement algorithms in terms of parallel execution time and speed up.

Table 3: Parallel execution time and speedup for various number of processors.

| Number of Processor | Parallel Execution Time | Speedup |
|---|---|---|
| 1 | 430 seconds | 1 |
| 2 | 217 seconds | 1.98 |
| 3 | 145 seconds | 2.97 |
| 4 | 109 seconds | 3.95 |

Table 3 shows the performance of the parallel speech enhancement algorithm for 1, 2, 3, and 4 processor. For the evaluation purposes, we used female speech signal with various noises and various SNRs and take the average of parallel execution time and speedup. The parallel speech enhancement algorithm achieves almost linear speedup indicating the high efficiency on parallelization. Moreover, this could be due to the fast communication scheme between processor in which it did not affect the parallel performance. When the number of nodes is high, the communication time will affect the speedup, especially in a cluster system. Therefore, it will be interesting if we evaluated our parallel speech enhancement algorithms on a cluster system with higher number of nodes.

## 6. CONCLUSIONS

In this paper, a new forward masking model using neural networks has been proposed and incorporated into a speech enhancement algorithm. The performance of our speech enhancement algorithm employing new forward masking model was compared with five other speech enhancement methods (two other functional models of forward masking) over twelve different noise types and four SNRs. PESQ results reveal that the proposed algorithm outperforms the other algorithms by 5-15% depending on the SNR. Hence, it appears that the proposed forward masking model has good potential for speech enhancement applications across many types and intensities of environmental noise. On a quad core system, the parallel speech enhancement algorithm developed was very efficient in which almost linear speedup was achieved.

## REFERENCES

[1] J. M. Buchholz, *A Computational Model of Auditory Masking Based on Signal-Dependent Compression*, PhD Thesis, Universitat Bochum, 2002.

[2] T. S. Gunawan, E. Ambikairajah, and D. Sen, "Comparison of Temporal Masking Models for Speech and Audio Coding Applications," in International Symposium on Digital Signal Processing and Communication Systems, pp. 99-103, 2003.

[3] F. Sinaga, T. S. Gunawan, and E. Ambikairajah, "Wavelet Packet Based Audio Coding Using Temporal Masking," in Int. Conf. on Information, Communications and Signal Processing, Singapore, pp. 1380-1383, 2003.

[4] T. S. Gunawan and E. Ambikairajah, "Speech enhancement using temporal masking and fractional bark gammatone filters," in 10th International Conference on Speech Science & Technology, Sydney, pp. 420-425, 2004.

[5] N. Westerlund, *Applied Speech Enhancement for Personal Communication*, Thesis, Blekinge Institute of Technology, 2003.

[6] W. Jesteadt, S. P. Bacon, and J. R. Lehman, "Forward masking as a function of frequency, masker level, and signal delay," *Journal of Acoustic Society of America*, vol. 71, pp. 950-962, 1982.

[7] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 451-464, 1997.

[8] H. Najafzadeh, H. Lahdidli, M. Lavoie, and L. Thibault, "Use of auditory temporal masking in the MPEG psychoacoustics model 2," in the 114th Convention, Audio Engineering Society, pp., 2003.

[9] T. S. Gunawan and E. Ambikairajah, "A new forward masking model and its application to speech enhancement," in IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 149-152, 2006.

[10] S. F. Boll, "Suppresion of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113-120, 1979.

[11] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in Europe Signal Processing Conference, Edinburgh, Scotland, pp. 1182-1185, 1994.

[12] E. Brown, D. Ros, C. Bruscianelli, and B. Mila de la Roca, "Non-recurrent neural networks for auditory perceptual modelling," in IEEE International Conference on Devices, Circuits and Systems, pp. 139-143, 1995.

[13] M. Florentine, H. Fastl, and S. Buus, "Temporal integration in normal hearing, cochlear impairment, and impairment simulated by masking," *Journal of Acoustic Society of America*, vol. 84, pp. 195-203, 1988.

[14] H. Simon, *Neural Networks: A Comprehensive Foundation*, Prentice Hall PTR, 1994.

[15] F. D. Foresee and M. T. Hagan, "Gauss-Newton approximation to Bayesian regularization," in International Joint Conference on Neural Networks, pp. 1930-1935, 1997.

[16] ITU, "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," International Telecommunication Union, Geneva 2001.