A NEW FORWARD MASKING MODEL AND ITS APPLICATION TO SPEECH ENHANCEMENT

Teddy Surya Gunawan and Eliathamby Ambikairajah

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

ABSTRACT

This paper presents a new forward masking model, which is applied to speech enhancement. The model develops a novel expression for forward masking, where the parameters are related to the masker level, the delay and the frequency obtained by curve-fitting the psychoacoustic data. This model is then incorporated, in a novel way, into a speech enhancement scheme. Objective measures using PESQ demonstrates that our enhancement scheme, provides significant improvements over four existing speech enhancement methods, when tested with speech signals corrupted by various noises at very low signal to noise ratios. Hence, the new forward masking model provides a greater and more accurate masking threshold calculation that leads to better PESQ scores.

1. INTRODUCTION

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 [1, 2]. Furthermore, forward masking has been used for speech enhancement [3] using the speech boosting technique [4]. 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 [5] provides a reasonable approximation to the forward masking effect. However, Jesteadt's model has a deficiency in that it calculates negative amounts of masking for both long signal delays and low-level maskers. Strope et al. [6] extended the Jesteadt experiment to 120ms. By analysing both models [5, 6] we can further extend their work to model forward masking effects more accurately.

To evaluate the performance of our forward masking model, four speech enhancement algorithms were implemented: spectral subtraction [7], spectral subtraction with minimum statistics [8], speech boosting [4], and speech boosting using a simple forward masking model [3]. The PESQ (ITU-T P.862) measure was used here to benchmark the various methods.

2. NEW FORWARD MASKING MODEL

Forward masking is a time domain phenomenon in which a masker precedes the signal in time. Forward masking decays as the delay between the masker and the signal offsets, Δt, is increased, and little masking occurs beyond 200 ms. The masker-signal delay is specified between offsets.

The rate of decay in forward masking increases with the amount of masking produced for short delays. In other words, masked thresholds decrease faster with increasing masker-signal delay, as the masker level and the spectral proximity of the masker and signal increase. Forward masking by a long-duration masker lasts approximately 200 ms regardless of the initial amount of masking.

The proportionality between masker level (L_m), the delay (Δt) and frequency may be summarized by a descriptive formula that we have developed. One expression for the amount of forward masking $M(f, L_m, \Delta t)$ that fulfils these requirements is

$$M(f, L_m, \Delta t) = \frac{1}{a(f) - b(f)\log(L_m) + c(f)\log(\Delta t)} \quad (1)$$

where *a*, *b*, and *c* are parameters that are obtained by curve-fitting the psychoacoustic data in [5, 6]. To simplify the calculation, the values of *a*, *b*, and *c* are averaged across frequencies, where $a = 0.0640188$, $b = 0.0155695$, and $c = 0.00762065$. The values of *a*, *b*, and *c* were obtained from a set of 360 data points compiled from two studies [5, 6]. Equation (1) is plotted against L_m and Δt at a frequency of 500Hz in Figure 1. Similar plots can be obtained for various frequencies, thus providing a very reasonable estimation of forward masking data.

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:

$$FM(f, L_m, \Delta t, T_s) = M(f, Lm, \Delta t) + TIQ(f, T_s) \quad (2)$$

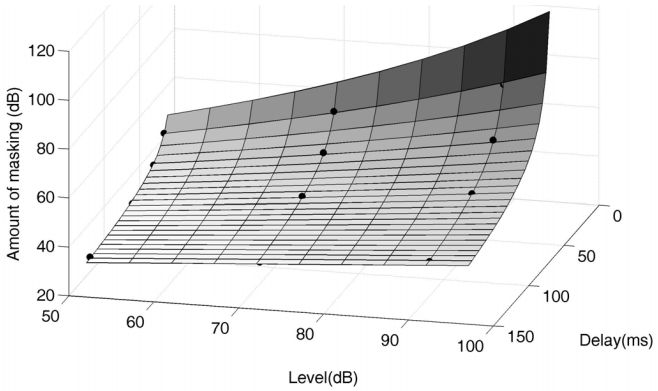

Figure 1. Amount of forward masking estimation at 500Hz

As stated in [9], the threshold in quiet is a function of frequency and signal duration. By curve-fitting a set of 120 data points compiled from [9], 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):

$$TIQ(f, T_s \geq 500) = 3.64(f)^{-0.8} + 6.5e^{-0.6(f-3.3)^2} + 0.001(f)^4 \quad (3)$$

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

$$TIQ(f, T_s) = TIQ(f, T_s \geq 500) + \left(7.53 - 6.5 \cdot 10^{-13} f^3\right) \log_{10}(500 - T_s) \quad (4)$$

3. SPEECH ENHANCEMENT

This section presents the incorporation of our model to fit the speech enhancement algorithm developed in [3]. Moreover, the forward masking threshold calculation is described.

3.1. Speech Enhancement Algorithm

Speech that has been contaminated by noise can be expressed as

$$x(n) = s(n) + v(n) \quad (5)$$

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 [3] is shown in Fig. 2. 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^{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 (Γ_m) in each subband. The gain, Γ_m, is a weighting function that amplifies the signal in band m during speech activity.

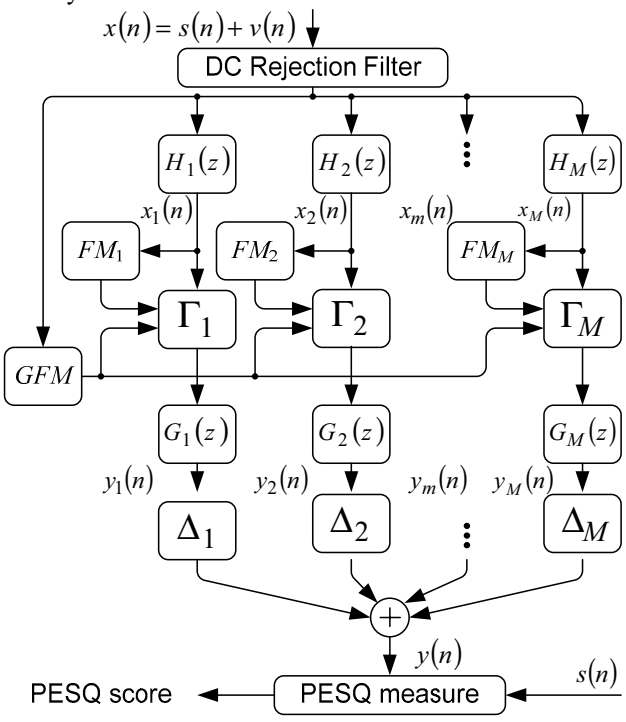

Figure 2. Speech enhancement using forward masking

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

$$y(n) = \sum_{m=1}^{M} y_m(n - \Delta_m) = \sum_{m=1}^{M} \Gamma_m x_m(n - \Delta_m) * g_m(n - \Delta_m) \quad (6)$$

Our objective is now to find a gain function, Γ_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 Equation (9). The short-term average temporal masking threshold in subband m is calculated as

$$P_m(n) = (1 - \alpha_m) P_m(n-1) + \alpha_m FM_m(n) \quad (7)$$

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 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

$$Q_m(n) = \begin{cases} (1+\beta_m)Q_m(n-1), & Q_m(n-1) \le P_m(n) \\ P_m(n), & Q_m(n-1) > P_m(n) \end{cases} \quad (8)$$

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 the *m*-th subband adapts to changes in the noise environment.

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

$$\Gamma_m(n) = \gamma_m \frac{FM_m(n)}{GFM(n)} + (1-\gamma_m)\frac{P_m(n)}{Q_m(n)} \quad (9)$$

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:

$$\Gamma_m(n) = \begin{cases} \Gamma_m(n), & \Gamma_m \le C_m \\ C_m & \Gamma_m > C_m \end{cases} \quad (10)$$

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

3.2. Forward Masking Calculation

The forward masking threshold is strongly influenced by the signals (masker) in the previous frames. The temporal information is obtained by calculating the temporal distances (T_F) between frames,

$$T_F = \frac{N_F}{F_S} \times 10^{-3} \quad \text{ms} \quad (11)$$

where N_F is the frame size and F_S is sampling frequency.

Since the longest duration of forward masking is 200 ms, then forward masking is calculated over N_{FM} successive frames as follows:

$$N_{FM} = \lfloor 200/T_F \rfloor \quad (12)$$

The forward masking threshold for each subband FM_m is then chosen as follows:

$$FM_m = \max\{FM_{m,j}\}, \quad j = 1 \ldots N_F \quad (13)$$

4. PERFORMANCE EVALUATION

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**[7], spectral subtraction with minimum statistics, **SSMS** [8], speech boosting, **SB**[4], speech boosting using forward masking model, **SBFM1**[3], and speech boosting using the proposed forward masking model, **SBFM2**.

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 [10] was utilised for the objective evaluation. Note that, the PESQ has a 93.5% correlation with subjective tests [10].

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

$$\delta = \frac{PESQ_{proc} - PESQ_{ref}}{PESQ_{ref}} \times 100\% \quad (14)$$

A total of 288 data sets from six speech files, twelve noises, and four SNRs for each method were simulated. The average quality improvement, δ, achieved by various speech enhancement methods is shown in Figure 3. Note that the δ results for various speech files and noises were averaged for -5, 0, 5, and 10 dB SNRs. From these results, the speech boosting technique using new forward masking model outperforms other methods for all SNRs.

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 δ result for each type of noise condition is shown in bold, from which it can be seen that our method using new forward masking model provides a better PESQ improvement than the four other methods tested.

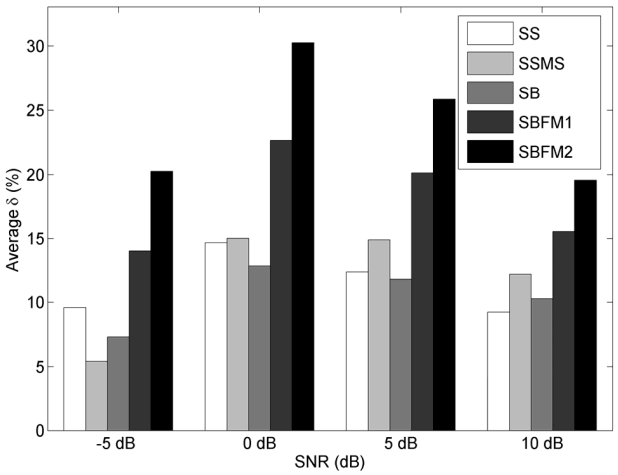

Figure 3. Average δ (%) for various algorithms

Table 1. Average PESQ improvement δ (%) for various noise types

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

Table 2. Average PESQ improvement δ (%) for different speech files

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

Table 2 shows the average of quality improvement at -5, 0, 5 and 10 dB SNRs for various speech files. The best δ 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.

5. CONCLUSIONS

In this paper, a new functional forward masking model has been proposed and incorporated into a speech enhancement algorithm. This model exploits the forward masking effect with dynamic adaptation of the auditory system. The performance of our speech enhancement algorithm employing new forward masking model was compared with four other speech enhancement methods over twelve different noise types and four SNRs. PESQ results reveal that the proposed algorithm outperforms the other algorithms by 10-17% 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.

6. REFERENCES

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

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

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

[4] N. Westerlund, "Applied Speech Enhancement for Personal Communication," in *Department of Telecommunications and Signal Processing*: Blekinge Institute of Technology, 2003.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.