Perceptual speech enhancement exploiting temporal masking properties of human auditory system

Teddy Surya Gunawan a,*, Eliathamby Ambikairajah b, Julien Epps b

a Department of Electrical and Computer Engineering, International Islamic University Malaysia, Gombak, 53100 Kuala Lumpur, Malaysia
b School of Electrical Engineering and Telecommunications, The University of New South Wales, NSW 2052, Australia

Received 16 January 2007; received in revised form 1 December 2009; accepted 7 December 2009

Abstract

The use of simultaneous masking in speech enhancement has shown promise for a range of noise types. In this paper, a new speech enhancement algorithm based on a short-term temporal masking threshold to noise ratio (MNR) is presented. A novel functional model for forward masking based on three parameters is incorporated into a speech enhancement framework based on speech boosting. The performance of the speech enhancement algorithm using the proposed forward masking model was compared with seven other speech enhancement methods over 12 different noise types and four SNRs. Objective evaluation using PESQ revealed that using the proposed forward masking model, the speech enhancement algorithm outperforms the other algorithms by 6–20% depending on the SNR. Moreover, subjective evaluation using 16 listeners confirmed the objective test results.

© 2009 Elsevier B.V. All rights reserved.

Keywords: Human auditory system; Speech enhancement; Temporal masking; Subjective test; Objective test

1. Introduction

The problem of enhancing speech degraded by noise remains largely open, even though many significant techniques have been introduced over the past decades. This problem is severe when no independent information on the nature of noise degradation is available, in which case the enhancement technique must utilise only the specific properties of given speech and noise signals.

In this paper, we focus on single channel speech enhancement. This is the most difficult task, since the noise and the speech are in the same channel. Many approaches have been reported in the literature (Lim and Oppenheim, 1979; Vaseghi, 2000). The most popular method, with many variants, is spectral subtraction. Although this method reduces the noise and improves the signal-to-noise ratio (SNR), it mostly tends to introduce speech distortion and a perceptually annoying residual noise usually called musical noise. Musical noise is a special term for short sinusoids (tones) randomly distributed over time and frequency. It occurs due to imperfections in the original spectral subtraction technique and statistical inaccuracy in noise magnitude spectrum estimation.

In order to reduce musical noise, various algorithms have been developed. Some recent noise reduction techniques have exploited the known properties of the human auditory system and have resulted in good speech quality with improved intelligibility and reduced levels of musical noise (Gunawan and Ambikairajah, 2004, 2006a; Gustafsson et al., 1998; Hu and Loizou, 2003; Lin et al., 2003; Ma et al., 2006; Tsoukalas et al., 1997; Virag, 1999). Psychoacoustic exploitation of this sort has so far utilised simultaneous masking only; temporal masking properties have not been exploited.

The human auditory system acts as an analysis filter bank with a perceptually relevant frequency resolution (such as the critical band scale or ERB scale). An appropriate choice for speech denoising, therefore, is just such an

* Corresponding author.
E-mail address: tsgunawan@gmail.com (T.S. Gunawan).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.

auditory frequency scale, instead of the uniform filter bank analysis of the Short Time Fourier Transform (STFT).

The objective of this paper is to develop a novel speech enhancement algorithm exploiting temporal masking properties in very noisy conditions (SNR <10 dB). The rest of the paper is organised as follows: a review of speech enhancement techniques is described in Section 2; two forward masking models for speech enhancement application are outlined in Section 3; a novel speech enhancement method exploiting temporal masking is presented in Section 4; Section 5 describes the effect of noisy conditions to the calculation of simultaneous and temporal masking thresholds and the performance evaluation of speech enhancement techniques based on masking properties. Finally, Section 6 summarises this paper.

2. Review of single channel speech enhancement techniques

For single-channel applications, only a single microphone is available. This is a very difficult task since noise and speech are in the same channel, and noise needs to be estimated from the noisy speech. This discussion will focus on methods based on the assumption that only one input channel is available, the noise is additive, and the noise and speech signals are uncorrelated.

The majority of single-channel enhancement techniques use the spectral weighting approach (Berouti et al., 1979; Boll, 1979; Ephraim and Malah, 1984; Ephraim, 1992; Lim and Oppenheim, 1979; Vaseghi, 2000). Spectral weighting attenuates different spectral regions of the noisy speech \( x(n) \) by different factors. The result of this process is an enhanced speech signal \( \hat{s}(n) \) that contains less noise than the original one. In addition to the requirement for minimal distortion of the original speech, it is important that the residual noise, \( r(n) = s(n) - \hat{s}(n) \), i.e. the noise remaining after enhancement process, does not sound unnatural. A detailed study of the residual noise is reported in (Vary, 1985).

![Fig. 1. Proposed speech enhancement system.](image_url)

The spectral weighting (as illustrated in Fig. 1) is usually performed in a transformed domain, e.g. the frequency domain. A common transform is the Fourier transform, which provides an equidistant frequency resolution. There are also other methods such as the wavelet transform with non-equidistant spectral resolution (Ambikairajah et al., 1998; Carnero and Drygajlo, 1999), gammatone or gam-machirp filter banks (Irino, 1999), linear-phase bandsplitting (Moorer and Berger, 1986), and analysis-synthesis resonator filter banks (Gagnon and McGee, 1991). However, these methods require an appropriate voice activity detector, which further increases the computational load.

Recently, a new speech enhancement method has been reported (Westerlund, 2003). Instead of focusing on suppressing the noise, the method is focusing on enhancing the speech. The algorithm acts as a speech booster, i.e. it is only active when speech is present. As stated in (Westerlund, 2003) the algorithm has proven to be robust, flexible, versatile, and does not require a voice activity detector.

The following three broad classes of classical enhancement techniques exist and will be briefly reviewed in the following sections: (i) spectral subtraction, (ii) Wiener-type algorithms, and (iii) statistical model-based methods. A more detailed explanation can be found in the relevant literature (Ephraim, 1992; Hansen, 1999; Lim, 1983; Loizou, 2007; O'Shaughnessy, 1989; Vaseghi, 2000).

2.1. Spectral subtraction and wiener filtering

Speech that has been contaminated by noise can be expressed as

\[
x(n) = s(n) + v(n),
\]

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. Speech and noise are assumed to be uncorrelated. The objective in speech enhancement is to suppress the noise resulting in an output signal \( \hat{s}(n) \) that has a higher signal-to-noise ratio (SNR).

The enhanced speech short-time magnitude \( |\hat{S}(\omega)| \) can be obtained by subtracting the noisy speech short-time magnitude \( |X(\omega)| \) with the estimate of noise spectral magnitude \( |\hat{V}(\omega)| \) calculated during speech pauses. For the particular case of power spectral subtraction, the enhanced speech is estimated as follows:

\[
|\hat{S}(\omega)|^2 = \begin{cases} 
|X(\omega)|^2 - |\hat{V}(\omega)|^2, & \text{if } |X(\omega)|^2 > |\hat{V}(\omega)|^2, \\
0, & \text{otherwise.}
\end{cases}
\]

The phase of the noisy speech is not modified since human auditory system is insensitive to phase (Wang and Lim, 1982). The enhanced speech signal is then obtained by the following relationship:

\[
\hat{s}(n) = \text{IFFT}(|\hat{S}(\omega)| \cdot e^{j\phi(X(\omega)))}.
\]
The subtractive-type algorithms can also be examined using a filtering approach. Noisy speech is filtered with a time-varying linear filter dependent on the characteristics of the noisy speech spectrum and on the estimated noise spectrum, as follows:

$$|\hat{S}(\omega)| = \Gamma(\omega) \cdot |X(\omega)|,$$

where $\Gamma(\omega)$ is the spectral weighting function and $0 \leq \Gamma(\omega) \leq 1$. The following gain function $\Gamma(\omega)$ combines the method proposed in (Berouti et al., 1979) for noise reduction with the generalised spectral subtraction proposed in (Lim and Oppenheim, 1979):

$$\Gamma(\omega) = \begin{cases} 
(1 - \alpha \cdot \left| \frac{\hat{P}(\omega)}{X(\omega)} \right|^2)^2, & \text{if } \left| \frac{\hat{P}(\omega)}{X(\omega)} \right|^2 < \frac{1}{2 \pi \beta}, \\
\beta \cdot \left| \frac{\hat{P}(\omega)}{X(\omega)} \right|^2, & \text{otherwise.}
\end{cases}$$

This gain function allows a compromise between noise reduction, residual noise, and speech distortion by varying the following parameters:

1. **Oversubtraction factor** $\alpha (\alpha > 1)$, which attenuates the spectrum more than necessary. This leads to a reduction of residual noise peaks but also leads to an increased speech distortion.
2. **Spectral flooring** $\beta (0 \leq \beta << 1)$, which adds the background noise while reducing the residual noise.
3. **Exponents** $\gamma_1$ and $\gamma_2$, which determines the sharpness of the transition from $\Gamma(\omega) = 1$ (the spectral component is not modified) to $\Gamma(\omega) = 0$ (the spectral component is suppressed). Table 1 shows the modification of parameters $\gamma_1$ and $\gamma_2$ and the resulting algorithms.

There are many modifications found in the literature to the parameters $\alpha$, $\beta$, and $\gamma$ in order to form a compromise between noise reduction, residual noise and speech distortion. Some of the modifications include modified spectral subtraction (Berouti et al., 1979), iterative techniques to control the subtraction factor $\alpha$ (Bouquin, 1996), and non-linear spectral subtraction that modifies $\alpha$ based on the SNR in each frequency band (Lockwood et al., 1992). Further, these parameters can be adapted in time and frequency based on masking properties of the human auditory system (Gustafsson et al., 1998; Hu and Loizou, 2004; Lin et al., 2003; Ma et al., 2006; Tsoukalas et al., 1997; Virag, 1999).

Spectral subtractive algorithms normally operate in the frequency domain, however alternative methods have been developed that perform noise reduction in other domains, e.g. a signal subspace approach (Ephraim and Trees, 1995). Here the noisy speech is decomposed by a Karhunen–Loève-Transform (KLT) into a subspace connected to the signal of interest and a noise subspace. This approach differs from the spectral subtraction by the fact that a KLT is used instead of an FFT.

### 2.2. Statistical-model based speech enhancement

A review of the statistical-model based approach can be found in (Ephraim, 1992). This approach is justified when the statistics of speech and noise are not available and there is no knowledge of the best distortion measure in the perceptual sense. In (Ephraim, 1992) a composite source model has been proposed in which a finite set of statistically independent Gaussian sub-sources are chosen by a switch, controlled by a Markov chain. This is the most general statistic model of speech, known as HMM-based enhancement.

HMM-based enhancement systems allow separation between speech and noise. Moreover, introduction of a priori information about speech and modelling of noise lead to an improvement over classical methods, especially at low SNRs and for speech corrupted by non-stationary noise. However, HMM-based systems require a training phase to obtain the speech and noise models, further increasing the computational requirement. This training stage is followed by the estimation of the clean speech using Maximum A Posteriori (MAP) estimation (Hansen, 1999; Lim and Oppenheim, 1978), Minimum Mean-Squared Error (MMSE) estimation (Cohen, 2004; Ephraim and Malah, 1984), or Maximum Likelihood (ML) estimation via the Expectation Maximisation (EM) algorithm (Feder et al., 1989). MAP estimation leads to an iterative algorithm, while MMSE estimation determines the filter weights directly from the noisy signal.

The MMSE estimation is also known as Ephraim and Malah’s estimator (Ephraim and Malah, 1984). This method has the advantage of producing colourless residual noise. The gain function $\Gamma(\omega)$, see Eq. (4), of this estimator is expressed as a function of the a posteriori SNR and a priori SNR. The a posteriori SNR is defined as:

$$\text{SNR}_{\text{post}} = \frac{|X(\omega)|^2}{|\hat{V}(\omega)|^2},$$

while a priori SNR is defined as:

$$\text{SNR}_{\text{prio}} = \mathbb{E} \left[ \frac{|\hat{S}(\omega)|^2}{|\hat{V}(\omega)|^2} \right] = \lambda \cdot \frac{|\hat{S}_{\omega-1}(\omega)|^2}{|\hat{V}(\omega)|^2} + (1 - \lambda) \cdot P(\text{SNR}_{\text{post}} - 1),$$

where $|\hat{S}_{\omega-1}(\omega)|^2$ is the estimated clean speech from the previous frame, $P(y) = y$ if $y \geq 0$ and $P(y) = 0$ otherwise.

---

**Table 1** Parameters $\gamma_1$ and $\gamma_2$, and their associated algorithms.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_1 = 1$ and $\gamma_2 = 1$</td>
<td>Magnitude subtraction</td>
</tr>
<tr>
<td>$\gamma_1 = 2$ and $\gamma_2 = 0.5$</td>
<td>Power spectral subtraction</td>
</tr>
<tr>
<td>$\gamma_1 = 2$ and $\gamma_2 = 1$</td>
<td>Wiener filtering</td>
</tr>
</tbody>
</table>

---

SNR\text{post} - 1 is interpreted as the instantaneous SNR (SNR\text{post}), and \( \lambda = 0.98 \). The gain function of MMSE estimator is then can be calculated as follows:

\[
\Gamma(\omega) = \sqrt{\frac{\pi}{2}} \sqrt{\frac{1}{\text{SNR}_{\text{post}}} \frac{\text{SNR}_{\text{post}}}{1 + \text{SNR}_{\text{post}}} \cdot M\left(\frac{\text{SNR}_{\text{post}}}{1 + \text{SNR}_{\text{post}}} \right)},
\]

where \( M(\cdot) \) represents the following function:

\[
M(y) = \exp\left(-\frac{y}{2}\right) \cdot \left[(1 + y) \cdot I_0\left(\frac{y}{2}\right) + y \cdot I_1\left(\frac{y}{2}\right)\right]
\]

with \( I_0(\cdot) \) and \( I_1(\cdot) \) representing the modified Bessel functions of the zero and first order.

This algorithm has generated many of applications, and the gain function (8) and a priori SNR (7) are widely used. For example, in (Cappe, 1994) the influence of the parameter \( \lambda \) and the mechanism of elimination of musical noise with this estimator has been studied. In (Scalart and Filho, 1996), the a priori SNR was estimated adaptively for spectral subtraction and Wiener filtering. Further, a non-causal a priori estimator was proposed in (Cohen, 2004) which yields a higher quality of enhanced speech.

2.3. Enhancement based on perceptual criteria

In all the hearing sensations, masking plays an important role in the frequency domain, as well as in the time domain. The effect of auditory masking has been exploited for speech enhancement application (Hu and Loizou, 2003; Lin et al., 2003; Loizou, 2007; Ma et al., 2006; Tsoukalas et al., 1997). The basic approach is to render spectral components of annoying residual noise inaudible by forcing them to fall below the masking threshold. In this section, of the various psychoacoustically motivated speech enhancement algorithms, two algorithms were examined, i.e. spectral subtraction based algorithm (Virag, 1999) and Wiener filtering based algorithm (Lin et al., 2002). These were selected due to their historical significance (Virag, 1999) and subsequent performance improvement using a filterbank-based approach (Lin et al., 2002).

2.3.1. Virag’s method

Virag’s method (Virag, 1999) is based on the power spectral subtraction (\( \gamma_1 = 2 \) and \( \gamma_2 = 0.5 \)) as described in Eq. (5) and Table 1. In addition, the oversubtraction factor \( \alpha \) and the spectral flooring \( \beta \) are adapted in time and frequency based on simultaneous masking properties. For the current frame, the adaptation of the subtraction parameters is performed as follows:

\[
\begin{align*}
\alpha(\omega) &= F_s(\alpha_{\text{min}}, \alpha_{\text{max}}, \text{SM}(\omega)), \\
\beta(\omega) &= F_\beta(\beta_{\text{min}}, \beta_{\text{max}}, \text{SM}(\omega)),
\end{align*}
\]

where \( \alpha_{\text{min}}, \beta_{\text{min}} \) and \( \alpha_{\text{max}}, \beta_{\text{max}} \) are the minimal and maximal values of the oversubtraction and spectral flooring parameters. In (Virag, 1999), the parameters were set as:

\[ [\alpha_{\text{min}}, \alpha_{\text{max}}] = [1, 6] \text{ and } [\beta_{\text{min}}, \beta_{\text{max}}] = [0, 0.02]. \]

\( \text{SM}(\omega) \) is the simultaneous masking threshold updated each frame and is calculated as in (Johnston, 1988). \( F_s \) and \( F_\beta \) represent a maximal residual noise reduction for the minimal values of the masking threshold \( \text{SM}(\omega) \), and vice versa. In other words, \( F_s = \alpha_{\text{max}} \) if \( \text{SM}(\omega) = \text{SM}(\omega)_{\text{min}} \), and \( F_s = \alpha_{\text{min}} \) if \( \text{SM}(\omega) = \text{SM}(\omega)_{\text{max}} \). Moreover, the values of \( F_s \) between these two extreme cases are interpolated based on the value of \( \text{SM}(\omega) \). Similar considerations can be applied for \( F_\beta \).

2.3.2. Lin’s method

Lin’s method (Lin et al., 2002) is based on perceptual modification of Wiener filtering algorithm (\( \gamma_1 = 2 \) and \( \gamma_2 = 1 \)) as described in Eq. (5) and Table 1. In this method, noisy speech \( x(n) \) is decomposed into critical band signals, \( x_m(n), m = 1, \ldots, M \), using auditory filters where \( M \) is the number of critical bands. The auditory filters are chosen to correspond to the critical bands, e.g. using a gammatone filter bank instead of using a FFT. The objective here is to find a Wiener gain \( \Gamma_m \) for each critical band based on the perceptual criteria. Subsequently, each noisy subband signal is multiplied by the denoising gain \( \Gamma_m \) to obtain the denoised subband signal \( \tilde{s}_m(n) \), which is summed across all bands to yield the denoised speech, i.e.

\[
\tilde{s}(n) = \sum_{m=1}^M \Gamma_m \cdot x_m(n).
\]

In the critical band approach, the gain calculation of Wiener filtering method described in Eq. (5) can be represented as follows:

\[
\Gamma_m = \frac{\sigma_m^2}{\sigma_m^2 + \alpha \cdot \max[\sigma_m^2, \beta]},
\]

where \( \sigma_m^2 = E[\tilde{v}_m^2(n)] \), \( \sigma_m^2 = E[v_m^2(n)] \), \( \alpha \) is the oversuppression factor that allows a trade-off between signal distortion and noise reduction, and \( \beta \) is the noise floor. In (Lin et al., 2003), the auditory masking phenomenon is incorporated into Eq. (11) in order to obtain the optimal subband gain as follows:

\[
\Gamma_m = \frac{\sigma_m^2}{\sigma_m^2 + \alpha \cdot \max[(\sigma_m^2 - \eta \cdot \text{SM}_m), \beta]},
\]

where \( \eta \) is parameter that controls the amount of masking threshold, and \( \text{SM}_m \) is the simultaneous masking threshold calculated for each subband \( m, m = 1, \ldots, M \). In this method, noise is included in the perceptual criterion only if it exceeds the masking threshold. Moreover, the critical band noise \( \tilde{v}_m \) is estimated using the noise estimation method proposed in (Lin et al., 2003). Note that setting \( \alpha = 1, \beta = 0, \) and \( \eta = 0 \) yields the classical Wiener filtering approach (Linn and Oppenheim, 1979; Lin et al., 2002).

3. Temporal masking models

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 (or maskee) in time, while backward masking occurs when the masker follows the signal 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 threshold decays as the delay between the masker and the signal offsets, $\Delta t$, is increased, with little masking occurring beyond 200 ms. In this work, we used functional forward masking model based on Jesteadt’s model (Jesteadt et al., 1982) and new forward masking model based on Strope’s data (Strope and Alwan, 1997).

3.1. Forward masking Model 1

Based on the forward masking experiments carried out by Jesteadt et al. (1982), forward masking level can be well fitted to psychoacoustic data using the following equation:

$$M_f(t, m) = a(b - \log_{10} \Delta t)(Lm(t, m) - c), \quad (13)$$

where $M_f(t, m)$ is the amount of forward masking in dB in the $m$th band, $\Delta t$ is the time difference between the masker and the maskee in milliseconds, $Lm(t, m)$ 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 empirically to 0.7, 2.3, and 20, respectively. Note that, we utilised semi-formal listening tests in which it revealed that the above parameters provide good quality enhanced speech.

To evaluate the amount of forward masking, we subdivided the current frame of 32 ms into four sub-frames. The forward masking level was calculated for the $j$th sub-frame using the energy $Lm_j$, $j = 1, \ldots, 4$, accumulated over the previous frame and all sub-frames up to the current sub-frame. The forward masking threshold for sub-band $m$, $FM_m$, is then chosen as the maximum of $M_f(t, m)$ for each sub-frame calculation.

3.2. Forward masking Model 2

In this section, a new model was proposed to model forward masking effects more accurately by combining the work of (Jesteadt et al., 1982; Strope and Alwan, 1997). Jesteadt’s forward masking model, see Eq. (13), provides a reasonable approximation to the forward masking effect (Jesteadt et al., 1982). It does have a deficiency, however, in that it calculates negative amounts of masking both for long signal delays and low-level maskers. Moreover, Jesteadt’s forward masking data only cover up to 40 ms delay. At this delay, there is still significant forward masking. Therefore, Strope and Alwan (1997) extended the Jesteadt experiment to 120 ms. By analysing both psychoacoustic data (Jesteadt et al., 1982; Strope and Alwan, 1997), we can further extend their work to model forward masking effects more accurately.

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 ($Lm$), the delay ($\Delta t$) and frequency may be summarised by a descriptive formula developed in this work. One expression for the amount of forward masking $M_f(Lm, \Delta t)$ that fulfils these requirements is

$$M_f(Lm, \Delta t) = \frac{1}{a(f) + b(f) \log(Lm) + c(f) \log(\Delta t)}, \quad (14)$$

where $a$, $b$, and $c$ are parameters that are obtained by curve-fitting the psychoacoustic data in (Jesteadt et al., 1982; Strope and Alwan, 1997). To simplify the calculation, the values of $a$, $b$, and $c$ are averaged across frequencies up to 4 kHz, where $a = 0.0640188$, $b = -0.0155695$, and $c = 0.0076206$. The values of $a$, $b$, and $c$ were obtained from a set of 360 data points compiled from two studies (Jesteadt et al., 1982; Strope and Alwan, 1997). Eq. (14) is plotted against $Lm$ and $\Delta t$ at various frequencies, shown in Fig. 2. Similar plots can be obtained from various frequencies, providing a 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 developed below:

$$FM(Lm, \Delta t, T_S) = M_f(Lm, \Delta t) + TIQ(f, T_S). \quad (15)$$

As stated in (Florentine et al., 1988; Moore, 1995; Moore, 2003; Zwicker and Fastl, 1999), the threshold in quiet is a function of frequency and signal duration. By curve-fitting a set of 120 data points compiled from (Florentine et al., 1988), the threshold in quiet can be approximated using least squares minimization method as follows:

- $TIQ(f, T_S)$ for short signal duration ($T_S \gtrsim 500$ ms) can be approximated as ($f$ in kHz):

$$TIQ(f, T_S \gtrsim 500) = 3.64f^{-0.8} + 6.5e^{-0.6(f-3.3)^2} + 0.0001f^4. \quad (16)$$

- $TIQ(f, T_S)$ for long signal duration ($T_S < 500$ ms) can be approximated as:

$$TIQ(f, T_S) = TIQ(f, T_S \gtrsim 500) + (7.53 - 6.5 \times 10^{-3} f^3) \log_{10}(500 - T_S). \quad (17)$$

4. Proposed speech enhancement algorithm exploiting temporal masking

In this section, a novel speech enhancement algorithm that incorporates temporal masking is presented. The
block diagram of the proposed algorithm is shown in Fig. 3. Moreover, the analysis and synthesis filter bank used is described in more details.

4.1. Proposed speech enhancement algorithm

By filtering the input signal \( x(n) \) using a bank of \( M \) analysis filters, the noisy speech 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 \( x_m(n) \), 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.

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:

\[
\hat{s}(n) = \sum_{m=1}^{M} \delta_m(n - \Delta_m) + \sum_{m=1}^{M} \Gamma_m x_m(n - \Delta_m) * g_m(n - \Delta_m).
\]

Our aim 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 Eq. (22). The short-term average temporal masking threshold in subband \( m \) is calculated as

\[
P_m(n) = (1 - z_m) P_m(n-1) + z_m FM_m(n),
\]

where \( z_m \) is a small positive constant \( (z_m = 0.0042, \forall m) \) 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 sub-band, \( Q_m(n) \), is calculated as

\[
Q_m(n) = \begin{cases} 
(1 + \beta_m)Q_m(n - 1) & Q_m(n - 1) \leq P_m(n) \\
P_m(n) & Q_m(n - 1) > P_m(n), 
\end{cases}
\]

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

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

\[
\Gamma_m(n) = \frac{FM_m(n)}{GFM(n)} + (1 - \gamma_m) \frac{P_m(n)}{Q_m(n)},
\]

where \( 0 \leq \gamma_m \leq 1 \), (\( \gamma_m = 0.9, \forall m \)) 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. For this reason a limiter can be applied on \( \Gamma_m(n) \) as follows:

\[
\Gamma_m(n) = \begin{cases} 
\Gamma_m(n) & \Gamma_m \leq K_m \\
K_m & \Gamma_m > K_m,
\end{cases}
\]

where \( K_m = 8 \text{ dB} \approx 2.51 \) provides a suitable limiter for the gain function.

4.2. Gammatone analysis/synthesis filter bank

In this work a fractional bark gammatone filter bank was employed to filter the signal \( x(n) \) with more precision into its subband signals \( x_m(n) \). Note that a fractional Bark filter bank resolution (i.e. 0.25 and 0.5 Bark, corresponding to the filter spacing of Basic and Advanced Version of PEAQ, respectively) has been reported to provide more accurate objective measurement of perceived audio quality (ITU, 1998). Therefore it is expected that the use of fractional Bark resolution will provide more accurate masking calculations in speech enhancement.

For the analysis bank we used gammatone filters as they resemble the shape of human auditory filters (Kubin and Kleijn, 1999). The gammatone bank was implemented using FIR filters. To achieve perfect reconstruction and linear phase characteristics, the synthesis filters \( g_m(n) \) are the time reversal of the analysis filters \( h_m(n) \). The analysis filter for each subband \( m \) is obtained using the following expression:

\[
h_m(n) = a_m(nT)^{N_g-1} e^{-2nbBW_mT} \cos(2\pi f_cm nT + \varphi),
\]

where \( f_cm \) is the centre frequency for each subband \( m \), \( T \) is the sampling period, \( N_g \) is the gammatone filter order (\( N_g = 4 \)), \( BW_m \) is the critical bandwidth at a particular centre frequency, \( b = 1.65 \), and the \( a_m \) were selected for each filter so as to normalize the filter gain to \( 0 \) dB. For \( f_c = 8000 \) 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 \( m = 0, \ldots, M \).

4.3. 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 et al., 1979):

\[
z(f) = 7 \cdot a \sinh \left( \frac{f}{650} \right), \quad f(z) = 650 \cdot \sinh \left( \frac{z}{7} \right).
\]

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:

\[
M = \left\lfloor \frac{z(f_L) - z(f_U)}{dz} \right\rfloor
\]

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 ranges of \( 0-4 \) kHz. The lower, upper, and center frequency for each sub band in Bark scale can be calculated as follows:

\[
\begin{align*}
\lambda_m &= z(f_L) + (m - 1) \cdot dz, \\
\zeta_m &= \min(z(f_L) + m \cdot dz, z(f_U)), \\
\zeta_m &= \frac{1}{2}(\lambda_m + \zeta_m),
\end{align*}
\]

where \( m = 1, \ldots, M \). Subsequently, the center frequency and the bandwidth in Hz can be determined as follows:
In order to find the optimum value of $dz$ for our speech enhancement method, we evaluated the average quality improvement for two speech files (male and female English speakers) contaminated with car noise at 0, 10, and 20 dB SNRs for various $dz$ values. The processing time for various $dz$ values was also evaluated, as seen in Fig. 4. We found that setting $dz = 0.25$ provides the optimum value in terms of speech quality and processing time.

4.4. Optimum filter coefficients and delay compensation

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 compared to the high frequency filters. The length of each filter within the filterbank, $N_m$, can be optimised by evaluating the non-zero gammatone filter response in each sub-band. The optimum length of the filter $N_m$ in samples for each sub-band is given by

$$N_m = \min(N_{\text{max}}, \text{round}(f_s/f_{\text{cm}}) \cdot 25), \tag{29}$$

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

By employing the optimum length of the filter in each sub-band, $N_m$, 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 $\Delta_m = N_m - 1$, where $N_m$ is the optimum filter order calculated as in Eq. (29).

5. Performance evaluation

In this section, the performance of the proposed speech enhancement algorithm is presented. First, the calculation of simultaneous and temporal masking thresholds in noisy conditions was compared, to determine the susceptibility of both masking thresholds to corruption by noise. Moreover, the objective evaluation using PESQ and subjective evaluation conforming ITU-T P.835 (ITU, 2003) are described.

In order to assess the performance of the new forward masking model in enhancing speech signals a large number of simulations were performed. Two forward masking models were incorporated into the speech enhancement system. Six speech files were taken from the EBU SQAM data set (EBU, 1988), including male and female English speakers, male and female French speakers, and male and female German speakers. The lengths of the files were between 17 and 20 s. The sampling frequency was 8 kHz, and the frame size was 256 samples (32 ms).

Several algorithms were implemented and compared covering all three classes of speech enhancement algorithm as reviewed in Section 2. Table 2 shows the eight speech enhancement algorithms with short descriptions. Different types of background noises from the NOISEX-92 (Varga et al., 1992) and AURORA (Hirsch and Pearce, 2000) databases were used, including car, white, pink, F16, factory, babble, airport, exhibition, restaurant, street, subway and train noise. The variance of noise was adjusted to obtain $-5 \text{ dB, 0 dB, 5 dB, and 10 dB}$ SNRs according to the method described in ITU-T P.830 (ITU, 1996).

### Table 2

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS (Boll, 1979)</td>
<td>Spectral subtraction (Boll’s method)</td>
</tr>
<tr>
<td>SSSM (Virag, 1999)</td>
<td>Spectral subtraction exploiting simultaneous masking (Virag’s method)</td>
</tr>
<tr>
<td>AWF (Scalart and Filho, 1996)</td>
<td>Adaptive wiener filtering (Scalart’s method)</td>
</tr>
<tr>
<td>WFSM (Lin et al., 2003)</td>
<td>Wiener filtering exploiting simultaneous masking (Lin’s method)</td>
</tr>
<tr>
<td>SSMS (Martin, 2001)</td>
<td>Spectral subtraction with minimum statistics (Martin’s method)</td>
</tr>
<tr>
<td>SB (Westerlund, 2003)</td>
<td>Speech boosting technique (Westerlund’s method)</td>
</tr>
<tr>
<td>SBFMJ</td>
<td>Proposed speech boosting technique using forward masking model by Jesteadt et al. (1982)</td>
</tr>
<tr>
<td>SBFMG</td>
<td>Proposed speech boosting technique using novel three-parameter forward masking</td>
</tr>
</tbody>
</table>

5.1. Comparison of simultaneous and temporal masking thresholds in noisy conditions

As described in Section 2.3, the use of auditory masking in speech enhancement systems leads to a significant reduction of the unnatural structure of residual noise (Gustaf-
son et al., 1998; Lin et al., 2002; Tsoukalas et al., 1997; Virag, 1999).

In this section, the calculation of both simultaneous and temporal masking thresholds were analysed in noisy conditions using series of experiments. The objective was to determine the susceptibility of simultaneous and temporal masking thresholds in various noisy conditions.

Male and female English speakers were taken from EBU SQAM database (EBU, 1988). Two noise files from NOI-SEX-92 database (Varga et al., 1992) consisting of white noise and car noise were added to the speech files with SNRs of 0, 5, 10, and 20 dB. The simultaneous masking calculations used were MPEG-1 Psychoacoustic Model 1 (SM1) and Model 2 (SM2), and the temporal masking thresholds used were Jesteadt’s forward masking model (TMJ) and the temporal masking model described in Section 3 (TMG). The mean squared error (MSE) was then calculated between the masking threshold obtained in clean speech and the masking threshold obtained in noisy speech.

Table 3 shows the average MSE results for various SNRs. Tables 4 and 5 show the average MSE results for various noise types and speech files, respectively. In both tables the lowest MSE (less susceptible to noise) is shown in bold. We found that the MSE for the temporal masking threshold was lower than the MSE for the simultaneous masking threshold for various noise types and speech files. Moreover, the proposed forward masking model (TMG) has lower MSE (less susceptible to noise) compared with Jesteadt’s forward masking model (TMJ). Therefore, it can be concluded that the temporal masking threshold is less susceptible to noise than the simultaneous masking threshold, and the proposed forward masking model is less susceptible to noise than Jesteadt’s forward masking model for the speech data employed in this experiment.

5.2. Objective evaluation

The PESQ (Perceptual Evaluation of Speech Quality, ITU-T P.862) measure was utilised for the objective evaluation (ITU, 2001), as it provides a more accurate result compared with other objective measures (Hu and Loizou, 2008). A total of 288 data sets from six speech files, 12 noises, and four SNRs for each method were simulated.

In order to evaluate the performance of the speech enhancement algorithms, a new measure was developed to assess the improvement achieved. Suppose that we have PESQref, 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, s(n), was also measured and denoted as PESQenc. Therefore calculate \( \delta \), which measures the PESQ improvement achieved by the algorithm as follows:

\[
\delta = \frac{\text{PESQ}_{\text{enc}} - \text{PESQ}_{\text{ref}}}{\text{PESQ}_{\text{ref}}} \times 100\%.
\]

The average quality improvement, \( \delta \), achieved by various speech enhancement methods is shown in Fig. 5. Note that the results for various speech files and noises were averaged for \(-5, 0, 5, \) and \(10\) dB SNRs.

In order to analyse the performance in more detail, the average quality improvement of \(-5, 0, 5, \) and \(10\) dB SNRs for various noise types is shown in Table 6. The best result for each type of noise condition is shown in bold; it can be seen that the proposed method, using a new forward masking model, provides a better PESQ improvement than the seven other methods tested.

Table 7 shows the average quality improvement of \(-5, 0, 5, \) and \(10\) dB SNRs for various speech files. The best result for each individual speech file is shown in bold. The table shows that more accurate forward masking threshold calculation consistently leads to better enhanced speech quality. From these results, the speech boosting technique using the new forward masking model outperforms other methods for all SNRs.
5.3. Subjective evaluation

In order to complement and to verify the objective evaluation reported in the previous section, subjective evaluation using the P835Tool developed (Gunawan and Ambikairajah, 2006a,b) was performed. A subset of the total set of files was selected to reduce the length of the subjective evaluations. Furthermore, a high quality headphone set, the Sony MDRV700DJ, was used for the listening tests.

A sentence spoken by a male English speaker was corrupted using three background noise environments (car, factory, and train noises) at two levels of SNR (5 dB and 10 dB). The files were processed using eight speech enhancement algorithms. A total of 48 processed files were presented to 16 listeners for evaluation. Hence, each subject was required to rate signals 144 times. The process of rating the signal (SIG) and background (BAK) of noisy speech in a P.835 trial was designed to lead the listener to integrate the effects of both the signal and background in making their ratings of overall (OVRL) quality. More details of the scale and the evaluation method can be found in ITU-T P.835 (ITU, 2003).

Fig. 6 shows the mean scores (MOS) for SIG, BAK, and OVRL scale for the eight speech enhancement algorithms, three background noise at two levels of SNR. The mean scores for the noisy speech (unprocessed) files are also shown for reference.

Of the eight speech enhancement algorithms examined, the speech booster technique (Westerlund, 2003) and speech booster using the new forward masking model performed comparably better in OVRL scale across all SNR conditions and three types of noise. Moreover, it is found that speech booster using new forward masking model is better in OVRL scale compared to other algorithms based on perceptual criteria, such as (Virag, 1999) and (Lin et al., 2002). These overall subjective results were further compared with the objective results obtained using PESQ.

Lower signal distortion (higher SIG scores) was observed both with algorithms that used speech boosting techniques, i.e. SB, SBFM1, and SBFM2, and algorithms that used simultaneous masking properties, i.e. SSSM and WFSM. In regards to the speech boosting technique,
many listeners found that the speech signal was only slightly distorted compared with other speech enhancement algorithms.

Lower noise distortion (higher BAK scores) was obtained with algorithms that used speech boosting techniques. This is due to the fact that whilst the mechanism enhances the speech part of the signal, the noise part is retained. We conclude then that human listeners prefer natural noise rather than reduced but distorted noise found in other algorithms.

6. Conclusions

A new speech enhancement algorithm based on a short-term temporal masking threshold to noise ratio (MNR) has been presented in this paper. In the algorithm development...
phase, our proposed algorithm was compared with three other speech enhancement methods over six different noise types and three SNRs. PESQ results revealed that the proposed algorithm outperforms the other algorithms by 6–20% depending on the SNR. In the particularly demanding 0 dB SNR condition, the new technique achieves at least a 30% relative improvement in ΔPESQ.

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 the speech enhancement algorithm using the proposed forward masking model was compared with seven other speech enhancement methods over 12 different noise types and four SNRs. Objective evaluation using PESQ revealed that using the proposed forward masking model, the speech enhancement algorithm outperforms the other algorithms by 6–20% depending on the SNR. Moreover, subjective evaluation using 16 listeners is supported by the objective test results. These investigations suggest that temporal masking has a role to play in improving the accuracy of future speech enhancement systems.

Appendix A. Supplementary data


References


