Perceptual evaluation of noise reduction in hearing aids
Brons, I.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Perceptual effects of noise reduction by time-frequency masking of noisy speech

Inge Brons, Rolph Houben, Wouter A. Dreschler

6.1 Introduction

Understanding speech in the presence of background noise is a demanding task. Consequently, there has been extensive research in the development of noise-reduction algorithms. These algorithms should improve user satisfaction with modern communication devices (e.g., hearing aids) in noisy environments. Noise-reduction algorithms should detect and remove the background noise, without affecting the target speech. Many noise-reduction techniques are based on the time-frequency representation of the signal, thus analyzing the content of each frequency channel during each time window. For all of these individual time-frequency units, the signal-to-noise ratio (SNR) is calculated or estimated and is used to determine whether the time-frequency unit will be retained (when speech is dominant) or attenuated (when noise is dominant). The resulting time-frequency attenuation pattern is often referred to as a time-frequency mask.

Numerous studies have examined the effects of time-frequency masking on speech intelligibility in noise. Time-frequency masking has been shown to cause large improvements in intelligibility (Brungart et al. 2006; Wang et al. 2008). This improvement, however, reflects the non-realistic situation that speech and noise are given separately as input (i.e., ideal noise-tracking). In a realistic situation in which the speech and noise are not individually known, the algorithm must estimate the noise without prior knowledge of the input signal. Time-frequency masking then provides no benefit in terms of intelligibility, mainly because of errors in the noise estimation (Loizou and Kim 2011). However, intelligibility is not the only outcome that is relevant for user satisfaction. Therefore, in this study, we evaluated different types of time-frequency masking not only on speech intelligibility but also on listening effort, speech naturalness, noise annoyance, and overall preference. We determined how these outcomes were influenced by two main aspects of noise reduction: (1) the strength of attenuation and (2) the method for noise tracking. We used four different types of time-frequency masking. With the first two conditions, we compared strong (infinite) attenuation with limited attenuation. With the second, third, and fourth conditions, we compared ideal noise tracking and two types of non-ideal noise-tracking.

Our first condition is known as the ideal binary mask (IBM) (Wang 2005). This noise reduction receives speech and noise separately as input so that it does not need to estimate the noise from the mixed signal. The IBM applies a binary pattern of attenuation to the noisy signal. All of the time-frequency units that have a signal-to-noise ratio (SNR) above a specified threshold are preserved, while all of the units with a lower SNR are eliminated. Usually, the threshold is 0 dB SNR. Researchers often use the IBM
to investigate the effects of different noise-reduction parameters on intelligibility, independent of noise-estimation errors. For example, the IBM has been used to evaluate the influence of the time- and frequency resolution of noise reduction (Anzalone et al. 2006; Li and Loizou 2008a; Wang et al. 2008), the frequency range on which noise reduction is active (Anzalone et al. 2006; Li and Loizou, 2008a; Wang et al., 2009), the signal-to-noise ratio threshold below which noise reduction is applied (Li and Loizou 2008b; Kjems et al. 2009), and the type of background noise (Li and Loizou 2008b; Kjems et al. 2009).

The IBM has proven to be able to provide approximately 13 dB SNR improvement in speech intelligibility in the presence of noise (Wang et al. 2009). However, the rapid binary attenuation transitions of the IBM can introduce musical noise (Wang 2008). Musical noise has a tonal character, and it occurs because of small isolated peaks that remain in the spectrum after the signal is removed in other time-frequency units (Berouti et al. 1979). In some cases, musical noise can be more disturbing to the listener than the original distortions caused by interfering noise (Loizou 2007).

A method for reducing musical noise is to limit the attenuation so that the noise-dominated time-frequency units will be attenuated but not eliminated (Anzalone et al. 2006). Although limiting the attenuation could improve the sound quality, it could also reduce the potential intelligibility benefit. This hypothesized trade-off between the subjective perception of the sound quality and the objective benefit in terms of speech intelligibility should receive more attention in the evaluation of noise reduction algorithms (Wang 2008). Therefore, our second noise-reduction condition was an ideal mask as well, but with a tempered attenuation function (ideal tempered mask, ITM).

Tempering the attenuation function is especially useful in combination with non-ideal noise estimators, which we used in our third and fourth conditions. Because noise estimators introduce errors in the SNR estimation, applying a binary attenuation function that either retains or completely removes time-frequency units would cause not only musical noise but also additional distortions from estimation errors. Estimation errors can be classified into two types: type I errors occur when time-frequency units are wrongly classified as being speech-dominated, and type II errors occur when units are wrongly classified as being noise-dominated (Li and Loizou 2008b). We used two noise-estimation algorithms for our third and fourth time-frequency masks: for one algorithm, the type I errors dominated (Hu and Loizou 2008), and for the other algorithm, the type II errors dominated. Both of the algorithms were combined with the tempered attenuation function.
To summarize, whereas most of the studies with the IBM concentrate on its effect on intelligibility only, we evaluated this algorithm also on other perceptual aspects (listening effort, noise annoyance, speech naturalness, and overall preference). Additionally, we limited the maximum attenuation to determine whether the expected disadvantages of the IBM can be reduced and to what degree this limited attenuation reduces the advantages in terms of speech intelligibility. Finally, we replaced the ideal noise classifier by real noise estimators to determine whether this realistic noise reduction has perceptual advantages in spite of the expected lack of intelligibility improvement. These three steps resulted in the following research questions:

**Q1.** How does the IBM influence perception in terms of speech intelligibility, listening effort, noise annoyance, speech naturalness, and overall preference?

**Q2.** How do the perceptual effects differ between the IBM and an ideal mask with non-binary attenuation limited to a maximum of 10 dB (“ideal tempered mask”)?

**Q3.** How do the perceptual effects of noise reduction differ between noise reduction with and without prior knowledge of the speech and noise signals? (“non-ideal” versus “ideal” masking).

### 6.2 Methods

This study was performed in parallel with our study to investigate the perceptual effects of noise-reduction algorithms that are implemented in hearing aids. As such, the subjects, the measurement procedures and the statistical methods are identical to those described in Chapter 3.

#### 6.2.1 Subjects

Ten normal-hearing subjects (who were all university students) between 19 and 23 years of age (average = 0.8 years) participated in this study. Their hearing thresholds were 15 dB Hearing Level or better at 0.25, 0.5, 1, 2, 3, 4, 6, and 8 kHz.

#### 6.2.2 Signal processing

Table 6.1 provides an overview of the five processing conditions. Stimuli for the unprocessed condition were passed through the IBM algorithm with the attenuation set to 0 dB (i.e., no attenuation) for all of the time-frequency units, which corresponded to a linear mask with all ones and no zeros. Thus, the complete speech-in-noise signal was retained in this condition.
Table 6.1: Overview of the differences between the five processing conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Noise tracker</th>
<th>Attenuation function</th>
<th>Attenuation (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IBM</td>
<td>Ideal</td>
<td>Binary</td>
<td>0 or ∞ dB</td>
</tr>
<tr>
<td>ITM</td>
<td>Ideal</td>
<td>Gradual</td>
<td>0 to 10 dB</td>
</tr>
<tr>
<td>MartinTM</td>
<td>Martin</td>
<td>Gradual</td>
<td>0 to 10 dB</td>
</tr>
<tr>
<td>MCRA2TM</td>
<td>MCRA2</td>
<td>Gradual</td>
<td>0 to 10 dB</td>
</tr>
</tbody>
</table>

The MATLAB implementation of the IBM algorithm used in this study was provided by Loizou and was previously used in Li and Loizou (2008b) and Hu and Loizou (2008). Briefly, time-frequency units were calculated by fast Fourier transformation on 20-ms Hamming-windowed segments, with a 50% overlap between the segments. All sound files had a sample rate of 44.1 kHz, leading to an FFT size of 882 samples per frame. The SNR of each time-frequency unit was compared against a threshold of 0 dB SNR to determine whether it was retained (at a positive SNR) or eliminated (at a negative SNR). We coded speech and noise that was processed with this original IBM algorithm as condition IBM.

For the next conditions, we introduced a new attenuation function to temper the IBM. Time-frequency units with (estimated) SNR below 0 dB were attenuated by 10 dB. For higher SNRs, the attenuation decreased logarithmically as a function of SNR (see Figure 6.1), which is comparable to a Wiener filter (Loizou 2007; see also Chapter 7). This condition was coded as ITM. The choice of 10 dB attenuation was based on the maximum attenuation provided by noise-reduction algorithms in hearing aids (Hoetink et al. 2009).

![Figure 6.1: Attenuation applied by the tempered mask as a function of (estimated) SNR.](image)

We also used the tempered attenuation function for the last two conditions, but it was preceded by real noise estimation instead of ideal noise tracking. The Martin noise estimator, used in the condition coded as MartinTM, was implemented by Loizou (2007).
according to the description in Martin (2001). The Martin algorithm is a minimum tracking algorithm, which means that it makes a rough estimate of the noise level in each frequency band by tracking the minimum of the input power in that band. The second noise estimator we used is a minimum controlled recursive average algorithm (called MCRA2) and was also implemented by Loizou (2007), according to the description in Rangachari and Loizou (2006). This algorithm updates the noise estimate in each frame using a time-frequency-dependent smoothing factor that varies with the probability that speech is present.

For both noise trackers, the SNR of each time-frequency unit was estimated from the estimated noise spectrum using the decision-directed approach (Ephraim and Malah 1984). This method estimates the a priori SNR, $\hat{\varepsilon}_k$ during each time frame $(m)$ and in each frequency band $(k)$, as a weighted average of the estimated SNRs of the previous and current frame:

$$
\hat{\varepsilon}_k(m) = \alpha \frac{(G_k(m-1)Y_k(m-1))^2}{D_k^2(m-1)} + (1 - \alpha) \max\left(\frac{Y_k^2(m)}{D_k^2(m)} - 1, 0\right)
$$

The weighting factor $\alpha$ was set to 0.98. $Y_k(m-1)$ is the input spectrum of the previous frame, and $G_k(m-1)$ expresses attenuation of the previous frame according to the tempered attenuation function based on its estimated SNR. Thus, the numerator of the first fraction represents the output of the previous frame. Dividing by the estimated noise power of that frame ($D_k^2(m-1)$, the output of the noise estimator), we obtain the a posteriori SNR. The second fraction represents the estimated SNR of the current frame, with the input in the numerator and the estimated noise in the denominator. Thus, the SNR estimate is largely determined by the estimated SNR of the previous frame. This approach causes the SNR estimate to change gradually, so that the attenuation will not change radically from frame to frame, which reduces musical noise. It is also important to note that, in this approach, the SNR estimate depends on the attenuation function. The method assumes that the output of the previous frame (the numerator of the first fraction) is an estimate of the speech signal in that frame. However, even if the noise is correctly detected, our attenuation function only attenuates it by 10 dB. Thus, the output signal will still contain noise while the approach incorrectly assumes that it does not. This scenario could result in an overestimation of the SNR, leading to less attenuation than desired.
Figure 6.2 shows the four different time-frequency masks for the same sentence. For
the IBM, black pixels indicate those noise-dominated time-frequency units that were
removed (infinite attenuation), and white pixels indicate the speech-dominated units
that were retained. For the other conditions, the attenuation is shown in gray, ranging
from white (0 dB attenuation) to gray (10 dB attenuation).

Figure 6.2: Attenuation by the four noise-reduction conditions for the same input sentence. For each
processing condition, the time signals show the unprocessed signal (dark background signal) and the
processed signal (light foreground signal). The spectrogram-like plots show the attenuation pattern as
a function of time and frequency. For the IBM (a), black pixels indicate noise-dominated time-frequency
units that were removed (infinite attenuation), and white pixels are the speech-dominated units that
were retained. For the other conditions (b-d), the attenuation is color coded between white (0 dB) and
gray (10 dB).
Because noise-reduction algorithms are mainly targeted at hearing aids and other mobile devices, we limited the bandwidth of the stimuli after processing to 100-5800 Hz with elliptical filters of the seventh order.

6.2.3 Stimuli

The input signals for the noise-reduction algorithms consisted of 260 unique concatenated Dutch sentences, produced by a female speaker (Versfeld et al. 2000) in multitalker babble noise (Luts et al. 2010). We combined speech and noise at SNRs of -22, -19, -16, -13, -10, -7, -4, 0 and +4 dB, based on the A-weighted representation of the signals. Thirteen sentences (36 s) preceded the stimulus sentences to allow the noise estimators to adapt to the input signals. The noise was continuous, while the speech paused one second between successive sentences.

Stimuli for the perceptual measurements consisted of single sentences cut from the processed signals, with 0.5 s of noise before and after the sentence. The stimuli were presented diotically with Sennheiser HDA200 headphones. The noise level was 70 dB(A) for all stimuli in the unprocessed condition.

6.2.4 Intelligibility

We measured speech intelligibility as the percentage of words that the subjects repeated correctly at fixed SNRs. Each subject started with 13 training sentences containing all five processing conditions, starting at +4 dB SNR. After every three sentences, the SNR decreased one step (4 dB for the first two steps and 3 dB for the last step), terminating with an SNR of -7 dB for the last four sentences. After this training, we used one list of 13 sentences per processing condition per SNR to determine the intelligibility scores. Every new combination of algorithm and SNR started with three training sentences, followed by ten sentences that were used to calculate the percentage correct. Stimuli from all of the five processing conditions were presented at -10, -7 and -4 dB SNR. Additional measurements were performed for the IBM at -22, -19 and -16 dB SNR and for the ITM at -13 dB SNR. We balanced the order of the conditions over all of the subjects to minimize the possible effects of training on the group data. We also balanced the sentence lists over the conditions to minimize the possible effects of differences between lists.

6.2.5 Listening effort rating

The subjects rated their perceived listening effort on a nine-point rating scale that ranged from “no effort” to “extremely high effort.” This rating scale is similar to the test used in Luts et al. (2010) but differs in that our scale used five labeled buttons.
instead of seven. The five labels are based on ITU-T P. 800 methodology (ITU-T 1996). Subjects gave ratings for all five processing conditions at three SNRs (-4, 0, and +4 dB). We considered the first run of 15 ratings to be practice, and we used the subsequent three runs for analysis.

6.2.6 Paired comparison rating

We used paired-comparison rating (a two-interval, seven-alternative forced choice paradigm) to measure speech naturalness, noise annoyance, and overall preference, successively. This method was based on the ITU-T P.835 method (ITU-T 2003) in which subjects must give separate ratings for the speech signal, the background noise, and the overall quality. The ITU standard uses a rating scale to measure the quality. We chose to use paired comparisons instead because these are more sensitive to subtle differences between conditions (Böckenholt 2001).

For each pair of stimuli, the subjects answered three questions. The first time that they listened to the two fragments A and B, subjects were asked to concentrate on the speech and to rate in which of the two fragments the speech was more natural and to indicate the strength of the difference. After they made a choice, they listened to the same fragments again, now concentrating on the annoyance of the noise and selecting the least annoying fragment. The subjects could listen to both fragments again before they answered the third question. For the third question, the subjects were asked which fragment they would prefer for prolonged listening. For each of the three questions, there were seven possible answers, ranging from “A is much more natural/much less annoying/much better” to “B is much more natural/much less annoying/much better.” The seven choice categories were derived from the Comparison Category Rating method described in ITU-T P. 800 (ITU-T 1996).

All five conditions were paired with each of the other conditions, which resulted in ten different stimulus pairs. Three runs of ten comparisons were performed at both -4 and +4 dB SNR, which resulted in a total of 60 comparisons per subject (10 pairs x 3 runs x 2 SNRs). All of the subjects started with four training pairs. Subsequently, five subjects started with all of the comparisons at -4 dB SNR, and the other five subjects started at +4 dB SNR.

6.3 Results

6.3.1 Intelligibility

Figure 6.3 shows the percentage of words correctly repeated averaged over all ten subjects. For statistical analysis, we transformed the percentages of correct words to ratio-
nalized arcsine units (rau) (Studebaker 1985) and subsequently performed a repeated-measures analysis of variance (ANOVA) on the transformed data for -10, -7, and -4 dB SNR, with SNR and processing condition as fixed effects. We found significant effects of SNR (F[2,18] = 631.7, p < 0.001) and processing condition (F[4,36] = 332.8, p < 0.001), and a significant interaction between processing condition and SNR (F[8,72] = 21.0, p < 0.001). Post hoc Bonferroni-corrected pairwise comparisons showed that the scores for IBM and ITM were higher than for the other three conditions at all SNRs. At -4 dB SNR, IBM and ITM were not significantly different from each other, but at -10 and -7 dB SNR, scores for IBM were significantly higher than for ITM (Bonferroni-corrected p-values were < 0.001 for all of the differences mentioned). Intelligibility scores for the other conditions (unprocessed, MartinTM, and MCRA2TM) did not differ significantly from each other.

![Graph of intelligibility scores](image)

**Figure 6.3:** Mean percentage of words correctly repeated by the 10 subjects at the different SNRs. Error bars show the 95% confidence interval between subjects (without Bonferroni correction).

### 6.3.2 Listening effort rating

Figure 6.4 shows the mean listening-effort ratings assigned by the 10 subjects. Note that a higher value means that the listening effort was lower. To satisfy the ANOVA criteria, we transformed the listening effort ratings with an arcsine transformation. We performed a repeated measures ANOVA with SNR and processing condition as fixed effects. We found significant effects of SNR (F[2,18] = 227.4, p < 0.001) and processing condition (F[4,36] = 35.5, p < 0.001) and a significant interaction between processing condition and SNR (F[8,372] = 15.73, p < 0.001). The horizontal lines in Figure 6.4 indicate which conditions differed significantly from each other after Bonferroni correction.
6.3.3 Paired comparison rating

Figure 6.5 shows the average rating score for each processing condition. We assigned scores from -3 to 3 for each condition, according to the ITU-T recommendation P. 800 (ITU-T 1996). If the subject rated condition A slightly better than condition B, then we assigned a score of 1 to condition A and a score of -1 to condition B. Similarly, scores of -2 and +2 indicate a moderate difference, and scores of -3 and +3 indicate a major difference, and a score of 0 indicates no difference. The scale for the noise annoyance is inverted in Figure 6.5. As a result, for each outcome, a positive value means a better performance on that judgment criterion. Error bars show a 95% confidence interval between the subjects.

Because, in general, scorings do not represent a linear interval scale, we used the log-linear modeling approach for ordinal paired-comparisons described by Dittrich et al. (2004) for the statistical analysis of the paired-comparison rating data. The model is a log-linear representation of the Bradley-Terry model (Bradley and Terry 1952) and is extended for paired-comparison data with multiple response categories, including a “no difference” option. By fitting this model to the paired-comparison data, we obtained estimates of the so-called “worth” parameters, which describe the location of the five processing conditions on the subject’s preference scale. This scale can be interpreted similarly to a ratio scale, thus providing not only the ranking of preferences for the five conditions but also information regarding the strengths of the preferences.
Figure 6.5: Mean and 95% confidence intervals of the rating scores derived from the paired-comparison data for the three judgment criteria and two SNRs. We assigned values from -3 to +3 to the answers, in accordance with ITU P. 800 (ITU-T 1996). The horizontal bars indicate which processing conditions differ significantly from each other after Bonferroni correction for 10 comparisons.

We estimated the worth parameters separately for the noise annoyance, speech naturalness, and overall preference. We fitted a model for each individual run of ten comparisons, which resulted in three models per subject per SNR per judgment criterion. We tested the goodness-of-fit for all of the models by comparing the obtained model with a saturated model (a model reproducing the data perfectly). All of the p-values were >0.95, indicating a high agreement with the saturated model; thus all of the models could be accepted.
We did repeated-measures ANOVAs on the estimated worth parameters for each judgment criterion separately (noise annoyance, speech naturalness, and overall preference) with SNR and processing condition as fixed effects. The resulting F-statistics and p-values are presented in Table 6.2. We found a significant effect for processing condition for each of the three judgment criteria. The effect of SNR (+4 or -4 dB SNR) was only significant for the overall preference. The interaction between processing condition and SNR was significant for all three criteria. Because of the significant interaction between processing condition and SNR, we redid the repeated-measures ANOVA but treated each SNR separately, with processing condition as a fixed effect. The resulting values for F and p are also given in Table 6.2. In both analyses (-4 and +4 dB SNR), the effect of processing condition was significant for each judgment criterion. The horizontal lines in Figure 6.5 indicate which conditions differed significantly from each other after Bonferroni correction.

**Table 6.2:** Main analysis of variance outcomes for the paired-comparison results.

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>Noise annoyance</th>
<th>Speech naturalness</th>
<th>Overall preference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Both signal-to-noise ratios</td>
<td>Both signal-to-noise ratios</td>
<td>Both signal-to-noise ratios</td>
</tr>
<tr>
<td>Processing condition</td>
<td>4</td>
<td>6.87 &lt; 0.001</td>
<td>47.88 &lt; 0.001</td>
<td>7.90 &lt; 0.001</td>
</tr>
<tr>
<td>SNR</td>
<td>1</td>
<td>2.00 0.19</td>
<td>2.02 0.19</td>
<td>7.84 &lt; 0.05</td>
</tr>
<tr>
<td>Processing condition x SNR</td>
<td>4</td>
<td>8.35 &lt; 0.001</td>
<td>3.23 &lt; 0.05</td>
<td>5.97 &lt; 0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Effect</th>
<th>df</th>
<th>Noise annoyance</th>
<th>Speech naturalness</th>
<th>Overall preference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Both signal-to-noise ratios</td>
<td>Both signal-to-noise ratios</td>
<td>Both signal-to-noise ratios</td>
</tr>
<tr>
<td>Processing condition</td>
<td>4</td>
<td>6.81 &lt; 0.001</td>
<td>43.36 &lt; 0.001</td>
<td>6.21 &lt; 0.001</td>
</tr>
<tr>
<td>Processing condition</td>
<td>4</td>
<td>39.97 &lt; 0.001</td>
<td>6.40 &lt; 0.001</td>
<td>9.87 &lt; 0.001</td>
</tr>
</tbody>
</table>

### 6.4 Discussion

With respect to our research questions, we can summarize our findings as follows.

**Q1.** Our results show that the IBM reduces noise annoyance, causing a strong increase in intelligibility and a decrease in listening effort. However, the IBM also strongly reduces speech naturalness and is therefore not preferred over no processing at the SNRs tested (-4 and +4 dB SNR).

**Q2.** Tempering the IBM by limiting its maximum attenuation to 10 dB with a gradual instead of a binary attenuation function (ITM) removes the disadvantage of reduced speech naturalness. The tempered attenuation causes the ITM to be preferred over the IBM and over all of the other conditions. Compared to the unprocessed condition, the ITM still increases intelligibility and reduces the listening effort and noise annoyance, but to a lesser extent than the IBM.
Q3. Replacing the ideal noise tracker with real noise estimators removes the benefit of increased intelligibility, as expected. The MCRA2 noise estimator reduced the noise annoyance compared with no noise reduction and was also preferred over no noise reduction. The Martin noise estimator caused much smaller reductions in the noise annoyance and was only slightly preferred over no noise reduction at +4 dB SNR.

6.4.1 Ideal binary mask

As expected, the IBM strongly improved intelligibility. Li and Loizou (2008b) described intelligibility measurements with the same IBM implementation. In their first experiment, they investigated the effect of the threshold for retaining or removing time-frequency units on intelligibility. Because changes in this threshold have the same effect as changes in the input SNR (Brungart et al., 2006), we can interpret the results for the different thresholds as results for different input SNRs. In spite of the differences in speech material, the obtained word scores for speech in 20-talker babble agree well between Li and Loizou (2008b) and our study (we estimate from their Figure 2 that the SRT_{50} was at -19 dB SNR, compared to -21 dB SNR in our study). Additionally, Wang et al. (2009) found a 13.4 dB improvement from IBM relative to the unprocessed condition in stationary speech-shaped noise (corrected for their threshold of -6 dB), compared to 13.8 in our study. However, Brungart et al. (2006) found smaller improvements in intelligibility for speech in speech-shaped noise (we estimate from their Figure 5 (left) an SRT_{50} improvement of 5 dB). This difference can most likely be attributed partly to the differences in speech material. Brungart et al. (2006) used words embedded in a fixed carrier sentence as speech material, whereas the other studies used sentence materials.

The IBM reduced the noise annoyance more than all of the other conditions, but also reduced speech naturalness the most. Subjects described the IBM condition as “a computer-like voice” or “watery sounding”. This unnaturalness is probably the reason that the IBM was not preferred over unprocessed or other noise reduction conditions at -4 and +4 dB SNRs. However, the large error bars for the overall preference of the IBM indicate that individual subjects differ in whether they prefer IBM processing or not. Whereas the preference for the other conditions can be mainly based on the sound quality, the IBM requires a choice between low quality and high intelligibility. Which of these criteria will obtain the higher weight could differ between listening situations, and several subjects indicated that they had difficulty with the choice.

It can be expected that at +4 dB SNR intelligibility scores will be 100% for all of the conditions. However, the listening effort scores still show differences between the ideal and non-ideal conditions at this SNR. Thus, even if speech is fully intelligible, there
could be a benefit of noise reduction in terms of reduced listening effort. If the SNR decreases, the perceived effort increases for all of the conditions except for the IBM because it still removes all of the noise and provides a large intelligibility improvement. It seems that the unnaturalness of speech does not increase the perceived listening effort at SNRs high above the SRT$_{50}$.

The IBM seems optimal for improving speech intelligibility (Loizou and Kim 2011), but it can only be used in the exceptional situation in which both speech and noise are given separately as input signals. A few attempts have been made to estimate the binary mask from a single input signal in which speech and noise were mixed. For example, Kim et al. (2009) trained binary SNR classifiers to estimate the binary mask. Their approach do not require an accurate SNR estimate, but only an accurate classification of SNR $< 0$ dB and SNR $> 0$ dB. Their classifier estimated the IBM accurately enough to result in large intelligibility improvements for normal-hearing listeners. It remains to be seen whether these results can be generalized, as Brookes and Huckvale (2011) were not able to reproduce these intelligibility improvements. Estimation of the binary mask thus requires more investigation.

### 6.4.2 Ideal tempered mask

Tempering the IBM (i.e., limiting the attenuation to 10 dB and making the attenuation function gradual instead of binary) reduced the intelligibility benefit compared with IBM, but there was still an improvement in SRT$_{50}$ of 4.8 dB due to ITM compared with the unprocessed condition. Anzalone et al. (2006) also limited the attenuation of their IBM. They did not use a gradual attenuation function, and their binary attenuation was 0 or 14 dB. For speech in speech-shaped noise, their IBM improved the SRT$_{50}$ more than 7 dB for normal-hearing listeners and approximately 9 dB for hearing-impaired listeners. These values are somewhat higher than those found in our results, but multiple differences in processing strategy make it difficult to compare the studies. For example, we used a gradual instead of a binary attenuation and based the attenuation on the SNR, whereas they used the speech energy for the binary decision.

Loizou and Kim (2011) have shown in their Appendix that the IBM is optimal in that it maximizes a simplified form of the articulation index, a measure known to correlate highly with speech intelligibility. As a result, the binary infinite attenuation function is theoretically optimal for improving intelligibility. From these derivations, it follows that tempering the attenuation, as in our ITM condition, leads to sub-optimal attenuation functions for intelligibility. Our intelligibility results confirm these theoretical hypotheses. Both versions of the ideal mask (IBM and ITM) cause intelligibility scores of (almost) 100% at $-4$ dB SNR. However, the perceived listening effort is higher for
the ITM than for the IBM. In contrast to the IBM, the amount of residual noise, and thus the listening effort for the ITM, increases with a decreasing SNR. Although more effortful than the IBM, the ITM is less effortful than the unprocessed condition at all SNRs.

The paired-comparison data show that the ITM reduces noise annoyance, but in contrast to the IBM, not at the cost of speech naturalness. Taken altogether, it is not surprising that ITM is clearly preferred over all of the other conditions. The ITM combines improved speech intelligibility with less degraded sound quality.

Cao et al. (2011) investigated the effect of adding stationary noise after IBM processing and found that, for certain noise levels, the addition of noise after IBM processing can further improve the intelligibility. They hypothesized that filling the sudden silences caused by IBM processing enhances the perceived continuity of the speech, leading to an intelligibility improvement. Instead of adding noise to the processed signal, limiting the attenuation could lead to a similar effect because the limitation leads to more residual noise in the processed signal. In our study, we limited the ITM attenuation to 10 dB. The output SNR can thus be at best 10 dB higher than the input SNR. For the ITM condition, our input SNRs thus lead to output SNRs of at best -3 to +6 dB (for input SNRs of -13 to -4 dB). Cao et al. (2011) found significant intelligibility improvements for output SNRs of 8 dB and higher. Thus, for our ITM condition it is not expected that the residual noise improves speech intelligibility. To investigate possible effects of residual noise on intelligibility one would need to use the input SNRs where the IBM did not give 100% speech intelligibility (-22 to -16 dB) combined with an ITM limitation of 30 dB. This would theoretically result in output SNRs of +8 to +14 dB, the region where Cao et al. (2011) found an improvement caused by the noise. Although it could be worthwhile to determine if the ITM can be optimized, combining the required 30 dB attenuation with non-ideal noise tracking could lead to large distortions due to noise-estimation errors.

6.4.3 Ideal noise tracking compared to noise estimation

The intelligibility results from MartinTM and MCRA2TM confirm the hypothesis that the positive effects of IBM will be negated if the prior knowledge of the noise and speech signal (ideal noise tracking) cannot be used (real noise estimator). Although the attenuation function we used has proven to be able to improve the SRT$_{50}$ by roughly 5 dB (see ITM), no improvement remained in the conditions in which noise and speech had to be estimated from the mixed signal. In terms of the perceived listening effort, these realistic noise-reduction conditions also provided no benefit over the unprocessed condition. However, both MartinTM and MCRA2TM significantly reduced the
perceived noise annoyance compared to the unprocessed condition, without affecting the speech naturalness. This reduction in noise annoyance was higher for MCRA2TM than for MartinTM. MCRA2TM was, at both SNRs, preferred over the unprocessed condition, whereas MartinTM was slightly preferred only at +4 dB SNR.

The fact that MCRA2TM was preferred over unprocessed at -4 dB SNR whereas the IBM was not, is noteworthy given the fact that the intelligibility scores for MCRA2TM were approximately 19% lower than for the IBM, and the perceived listening effort was lower than for the IBM. Thus, it seems that the sound quality was more decisive for the subjects’ preference than the benefit in terms of intelligibility or listening effort. The objective intelligibility can thus be balanced against the subjectively perceived quality by modifying the attenuation function. The optimum balance depends on both the situation and the individual listeners. For non-ideal masks, it is important to take into account that modifying the attenuation function could lead to a decrease in intelligibility.

6.4.4 Comparison of different noise estimators

The paired-comparison results show a difference between the two noise-estimation conditions in the degree to which they reduce the noise annoyance. This difference is not surprising given the attenuation patterns in Figures 6.2c and 6.2d, which are very different from each other. In the MartinTM condition, the majority of noise-dominated time-frequency units were not attenuated or were only slightly attenuated. This lack of attenuation resulted from the fact that the minimum tracking method often causes an underestimation of the true noise level (Loizou 2007; Chen and Loizou 2012). Additionally, the noise-spectrum estimate was updated slowly by the Martin algorithm, resulting in the stripes in Figure 6.2c (Loizou 2007). The small number of attenuated time-frequency units was still sufficient to slightly reduce the noise annoyance and did not affect speech naturalness. At +4 dB SNR, this reduced noise annoyance caused the MartinTM to be slightly preferred over the unprocessed condition. Thus, although the changes are small, this noise reduction can improve the sound quality at higher SNRs.

In contrast, in the MCRA2TM condition, the majority of noise-dominated units were attenuated (see Figure 6.2d), resulting in a reduced perceived noise annoyance. Although MCRA2TM also attenuated many of the speech-dominated units, this attenuation was apparently not perceived as reduced speech naturalness. This condition probably sounds like an overall attenuation of the unprocessed signal. Although MCRA2TM was not able to improve intelligibility or listening effort, the subjects preferred this condition at +4 dB SNR over all of the other conditions except for the ITM, which was based on ideal noise tracking.
Li and Loizou (2008b) defined two types of noise-estimation error: type I error, which occurs when a noise-dominated time-frequency unit is retained, and type II error, which occurs when a speech-dominated unit is removed. We estimated these error percentages for the Martin and MCRA2 noise estimators with our speech and noise stimuli as input signals. For this purpose we removed the noise-only timeframes between the sentences. Similar to the IBM, we compared the estimated SNR of each time-frequency unit to the threshold of 0 dB SNR to obtain a binary mask. We compared these masks to the IBM to calculate the percentages of misclassified units (Li and Loizou, 2008b, Appendix). For an input signal of -4 dB SNR, the Martin algorithm resulted in an average of 77% type I errors (i.e., 77% of all of the noise-dominated units were classified as speech-dominated) against 16% for the MCRA2 algorithm. In contrast, the percentage of type II errors (speech-dominated units classified as noise-dominated) was 19% for the Martin algorithm and 75% for the MCRA2 algorithm.

Li and Loizou (2008b) concluded that the binary mask did not improve intelligibility if the error percentage exceeded 85% for either error type in isolation. If both error types are present, this percentage is expected to be lower. Our noise estimators both approached this percentage already for one type of error but had additional errors of the other type. This result confirms that, for our speech and noise material, no intelligibility improvement can be expected from noise reduction based on the Martin or MCRA2 noise-estimation algorithms.

Li and Loizou (2008b) also concluded that type I and type II errors have different effects on intelligibility. Whereas type I errors affected intelligibility even at low percentages, type II errors can occur in up to 60% of the speech-dominated units before they cause a substantial decrease in intelligibility. Our results also seem to indicate that type II errors, which dominated in our MCRA2TM condition, are less detrimental for the overall preference. This result suggests that, if the same attenuation function is used, a noise reduction with a noise estimator that tends to underestimate the SNR would outperform a noise estimator that tends to overestimate the SNR. This statement is further supported by Chen and Loizou (2012). They systematically introduced different degrees of SNR overestimation and underestimation in a Wiener-filter based noise-reduction algorithm. The resulting intelligibility scores confirmed that SNR overestimation for time-frequency units with negative SNR was much more harmful to speech intelligibility than SNR underestimation.

6.4.5 Limitations
Paired-comparison ratings are especially useful if differences between the stimuli are small. However, our IBM condition was clearly different from all of the other condi-
tions. The high contrasts with the IBM could have led to smaller perceived differences between the other processing conditions.

For the non-ideal conditions, both the noise-estimation algorithm and the decision-directed approach for SNR estimation determine the final attenuation and estimation errors. Our results do not allow us to distinguish between these two approaches. We can, however, calculate the influence of the decision-directed approach on the error rates. We repeated the error calculations with the weighting factor $\alpha$ set to 0 (see Equation 6.1), so that the SNR estimation was based only on the input signal and noise estimate of the current frame. Compared with an $\alpha$ of 0.98, the percentage of Type I errors increased for both estimators (from 77% to 85% for Martin and from 16% to 39% for the MCRA2). The Type II errors decreased (from 19% to 18% for Martin and from 75% to 57% for MCRA2). Thus, the decision-directed approach tends to enhance SNR overestimation. This bias is introduced by the clipping function (the max operator, see Equation (6.1)) implemented in the decision-directed approach (Chen and Loizou, 2012). Additionally, as discussed in section 6.2.2, the combination of the decision-directed approach and a limited attenuation function also leads to SNR overestimation. As a consequence, without the decision-directed approach, fewer time-frequency units will be attenuated. This strategy will retain more noise, which would probably result in more noise annoyance and less preference than with the decision-directed approach.

We used a threshold of 0 dB SNR in the attenuation function for all of the conditions, which is a common choice for the IBM (Wang 2008). Additionally, the use of the same attenuation function for the three tempered conditions allowed us to compare the effect of the different noise-tracking methods. However, one must be aware of the possible effects of the attenuation function on the results. As noted in section 6.4.1, the threshold for retaining or removing time-frequency units for the IBM is directly related to the SNR of the input signal (Kjems et al. 2009). For example, a threshold of -8 dB SNR instead of 0 dB SNR would lead to the same results for input signals at -4 dB SNR as would a 0-dB threshold for input signals at +4 dB SNR. For the non-ideal conditions, however, the SNR estimate is not independent of the attenuation function (see section 6.2.2). The attenuation function that we used was not the function for which the noise estimators were developed and tested. Thus, the optimal choice for the attenuation threshold and the maximum attenuation depends on the noise-estimation algorithm, and this choice needs further investigation.

An important application of noise reduction is in hearing aids, to make listening in noisy environments easier. For this application, our normal-hearing study population differs from the hearing-impaired target population. Subsequent investigations are required to determine the effects for listeners with sensorineural hearing losses. Previ-
ous studies have shown that the IBM also improves intelligibility for hearing-impaired listeners (Anzalone et al. 2006; Wang et al. 2009). The improvement in intelligibility might be even higher for hearing-impaired listeners than for normal-hearing listeners because the performance in noise is worse for hearing-impaired listeners, whereas performance after IBM processing is comparable between normal-hearing and hearing-impaired listeners (Wang et al. 2009). Anzalone et al. (2006) observed a difference between normal-hearing and hearing-impaired listeners in their informal comments about the speech quality. The hearing-impaired listeners appeared to be less sensitive to the reduced quality. However, as mentioned before, the binary attenuation in that study was 0 or 14 dB, so that the noise was not completely removed but instead was only attenuated by 14 dB. This change reduces the speech distortions compared to the IBM, especially for the higher input SNRs presented to the hearing-impaired listeners. Thus, it is not certain whether hearing-impaired listeners indeed had lower sensitivity to distortions or whether their stimuli contained less distortion. On the one hand, it appears reasonable that hearing-impaired listeners have a higher detection threshold for speech distortions, so that they will accept a stronger attenuation. On the other hand, avoiding speech distortions could be more important for hearing-impaired listeners because of their dependence on clean speech signals. Additionally, there are also individual differences between hearing-impaired listeners in how well they accept background noise (Mueller et al. 2006) and in the maximum attenuation strength that they prefer for noise reduction (Houben et al. 2011). More investigation is required to determine to what extent a preference for noise reduction is determined by a hearing loss and to what extent by other individual differences.

6.5 Conclusions

We conclude that, although the IBM improves speech intelligibility in noise, listeners do not prefer it over the unprocessed condition at SNRs of -4 and +4 dB because it sounds unnatural. Tempering the IBM (limiting the attenuation to a maximum of 10 dB and smoothing the attenuation function) overcomes this drawback of the IBM while maintaining an intelligibility improvement, although to a lesser extent. Other values for the maximum attenuation should be evaluated to find an optimum for the trade-off between intelligibility and sound quality. With a real noise-estimation algorithm as the basis for noise reduction, estimation errors negate the potential intelligibility benefit of noise reduction. However, such realistic noise reduction can reduce the noise annoyance that is perceived by listeners so that they prefer it over the unprocessed signal. Although noise reduction based on noise-estimation algorithms does not yet provide an objective benefit in terms of intelligibility, possible subjective benefits should receive more attention in the development and evaluation of noise-reduction algorithms.