Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 272–281

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Impact of SNR and gain-function over- and under-estimation on speech intelligibility

Fei Chen, Philipos C. Loizou *

Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75083-0688, USA

Received 1 March 2011; received in revised form 11 August 2011; accepted 8 September 2011
Available online 19 September 2011

Abstract

Most noise reduction algorithms rely on obtaining reliable estimates of the SNR of each frequency bin. For that reason, much work has been done in analyzing the behavior and performance of SNR estimation algorithms in the context of improving speech quality and reducing speech distortions (e.g., musical noise). Comparatively little work has been reported, however, regarding the analysis and investigation of the effect of errors in SNR estimation on speech intelligibility. It is not known, for instance, whether it is the errors in SNR overestimation, errors in SNR underestimation, or both that are harmful to speech intelligibility. Errors in SNR estimation produce concomitant errors in the computation of the gain (suppression) function, and the impact of gain estimation errors on speech intelligibility is unclear. The present study assesses the effect of SNR estimation errors on gain function estimation via sensitivity analysis. Intelligibility listening studies were conducted to validate the sensitivity analysis. Results indicated that speech intelligibility is severely compromised when SNR and gain over-estimation errors are introduced in spectral components with negative SNR. A theoretical upper bound on the gain function is derived that can be used to constrain the values of the gain function so as to ensure that SNR overestimation errors are minimized. Speech enhancement algorithms that can limit the values of the gain function to fall within this upper bound can improve speech intelligibility.

Keywords: Speech enhancement; Speech intelligibility; SNR estimation

1. Introduction

Many speech-enhancement algorithms operate in the frequency domain and are based on multiplication of the noisy speech magnitude spectrum by a gain (or suppression) function, which is designed/optimized based on certain error criteria (e.g., mean squared error). Such algorithms include the MMSE (Ephraim and Malah, 1984), logMMSE (Ephraim and Malah, 1985) and Wiener filtering (Scalart and Filho, 1996), among others (see review in (Loizou, 2007, Ch. 7)). All these algorithms rely on accurate estimates of the signal-to-noise ratio (SNR) in each frequency bin, as the gain functions are defined in terms of the spectral SNR. A well-known approach in estimating the SNR is the "decision-directed" approach proposed in (Ephraim and Malah, 1984). This SNR estimator is simply computed using the weighted average of the past SNR estimate and the present SNR estimate.

The "decision-directed" approach is computationally simple and has been found to perform quite well in noise reduction applications (Hu and Loizou, 2007). A number of studies have analyzed the "decision-directed" approach in terms of its ability to reduce musical noise (Cappe, 1994) and in terms of its smoothing behavior in low SNR conditions (Breithaupt and Martin, 2011). Others have analyzed its bias and proposed methods to compensate for it (Martin, 2005; Erkelens et al., 2007; Plapous et al., 2006). This bias is inherent in the "decision-directed"

* Corresponding author. Address: Department of Electrical Engineering, University of Texas at Dallas, 800 West Campbell Road (EC33), Richardson, TX 75080-0688, USA. Tel.: +1 972 883 4617; fax: +1 972 883 2710.
E-mail address: loizou@utdallas.edu (P.C. Loizou).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.09.002

approach, and it is introduced partly due to the clipping function (max) used for ensuring positive SNR values (Martin, 2005), and the fact that the square of the estimator of the magnitude spectrum is used rather than the estimator of the magnitude-squared spectrum (Erkelens et al., 2007). Extensions to the "decision-directed" approach have been proposed in (Cohen, 2005) using non-causal SNR estimators that made use of future noisy observations.

In summary, much work (Hu and Loizou, 2007; Cappe, 1994; Breithaupt and Martin, 2011; Martin, 2005; Erkelens et al., 2007; Plapous et al., 2006) has been done in analyzing the behavior of the "decision-directed" approach in the context of reducing musical noise as well as reducing distortions in transient conditions. The overall goal of such analysis was to improve the subjective quality of enhanced speech. Little work has been done, however, in analyzing, more generally, SNR estimation in the context of speech intelligibility. That is, the impact of errors in estimating the SNR on speech intelligibility is largely unknown. It is unclear, for instance, whether it is the SNR over-estimation errors or the SNR under-estimation errors, or both, that are harmful to speech intelligibility. Errors in estimating the SNR affect directly the estimation of the gain function, and the impact of inaccuracies in estimating the gain function on speech intelligibility is also unknown. In brief, the sensitivity analysis of errors in SNR and gain estimation is lacking from the literature, particularly as it pertains to speech intelligibility. Such a sensitivity analysis needs to be accompanied with listening studies for appropriate validation of the analysis. The focus of the present study is to accomplish just that: provide sensitivity analysis of SNR errors and examine (and confirm) the impact of such errors using intelligibility listening studies. The outcomes from the present study are important as they can provide useful insights as to how to develop better SNR estimators that can be used in statistical-model based algorithms to improve speech intelligibility.

2. Sensitivity analysis

The majority of speech-enhancement algorithms operate in the frequency domain and are based on multiplication of the noisy speech magnitude spectrum by a gain function G. In most algorithms, the gain G is a function of the *a priori* SNR (e.g., Scalart and Filho, 1996), the *a posteriori* SNR (e.g., Berouti et al., 1979) or both (e.g., Ephraim and Malah, 1984, 1985). Without loss of generality, we present next the sensitivity analysis for the Wiener gain function. Let $G_W(\xi)$ denote the Wiener gain function:

$$G_W(\xi) = \frac{\xi}{\xi + 1}, \tag{1}$$

where $\xi \triangleq E[X^2]/E[D^2]$ denotes the *a priori* SNR, and X and D are the magnitude spectra of the clean speech and noise signals respectively. Let ξ^*

$$\xi^* = \xi + \Delta\xi \tag{2}$$

denote the perturbed value of ξ. From the above two equations, it is easy to derive the change in the value of the gain function produced when perturbing the value of ξ. Such a perturbation would reflect among other things the inaccuracy in estimating ξ from the noisy observations. We define this change in the gain function as:

$$\Delta G(\xi) = G(\xi^*) - G(\xi). \tag{3}$$

For the Wiener gain function (Eq. (1)), this is given by:

$$\Delta G(\xi) = \frac{\xi^* - \xi}{(\xi^* + 1)(\xi + 1)}. \tag{4}$$

To better understand the impact of errors in the estimation of ξ on the gain function, we show in Fig. 1 the plot of the delta gain function $\Delta G(\xi)$ for different values of $\Delta\xi$ ranging from $\Delta\xi = 0.5$ to $\Delta\xi = 60$. The $\Delta G(\xi)$ function is shown for both the Wiener gain function (left panel) and the log-MMSE gain function (right panel) in Fig. 1. It is clear from Fig. 1 that small values of perturbation ($\Delta\xi$) produce relatively large changes in the gain function (at least relative to the full dynamic range of the gain function, which is 1), in the negative SNR region (i.e., for $\xi_{dB} < 0$ dB). When $\Delta\xi = 1.5$, for instance, and assuming that $\xi_{dB} < 0$ dB, the gain function is overestimated by 0.6 (note that the true value of the Wiener gain function for $\xi_{dB} < 0$ dB is close to zero), which is quite substantial given that the gain function (at least, in most cases) is bounded by 1. It is clear from Eq. (4) that when ξ is large ($\xi \gg 1$) and $\Delta\xi$ is relatively small, we have $\xi \approx \xi^*$ and therefore $\Delta G(\xi) \approx 0$. Indeed, when $\xi_{dB} > 0$ dB, $\Delta G(\xi) \approx 0$, and this is confirmed in Fig. 1. Hence, for the region where $\xi_{dB} > 0$ dB, the gain function does not seem to be influenced by errors in the estimation of ξ. This is unfortunate since most noise reduction algorithms estimate the value of ξ more accurately in the positive rather than the negative SNR region (better estimates of the speech spectrum are obtained at high SNR levels). It is noted that, although the above analysis is done in the linear domain, the changes in *a priori* SNR analyzed span across a 60-dB range.

An equivalent way of deriving the sensitivity of the gain function to perturbations of the ξ values is by differentiating the gain function with respect to ξ (Whitehead and Anderson, 2011). In doing so, we can derive plots similar to those shown in Fig. 1 for the Wiener gain function. Sensitivity is highest at lower values of ξ reaching a maximum of 1 at $\xi = 0$ and sensitivity is lowest (approaching zero) at higher values of ξ (Whitehead and Anderson, 2011). This is consistent with the shape of the curves shown in Fig. 1.

Empirical evidence supporting the fact that ξ is overestimated in the negative SNR region is provided in Fig. 2, which plots the values of ξ estimated using the "decision-directed" approach (Ephraim and Malah, 1984), and denoted as $\hat{\xi}$, against the true short-time[1] values of ξ which

[1] As we can not compute the true *a priori* SNR values ξ, short-time (instantaneous) values are used for illustration purposes. To distinguish between the two, we use the symbol $\bar{\xi}$ in Eq. (5).

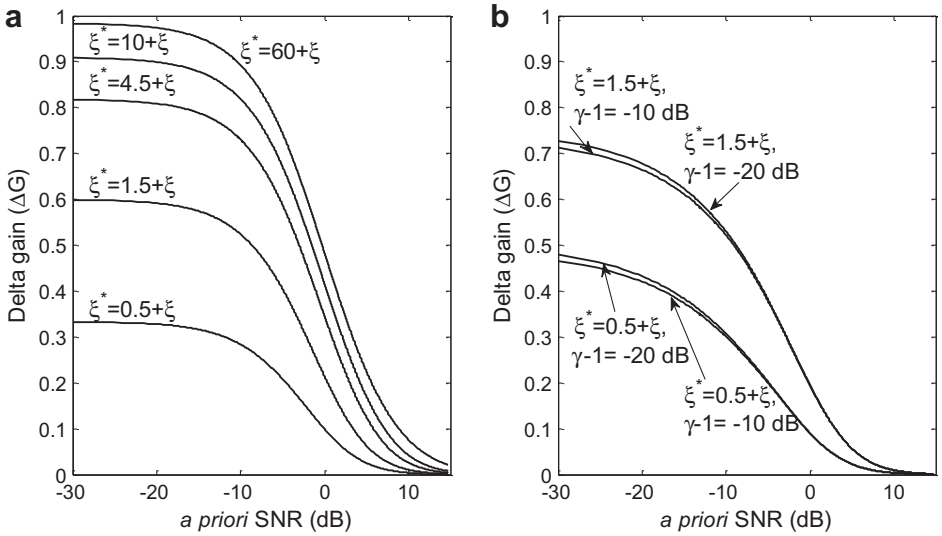

Fig. 1. Plots of the difference (ΔG) in the gain function, computed as $G(\xi^*) - G(\xi)$, for different values of the perturbed SNR (ξ^*). Panel (a) shows ΔG for the Wiener gain function and panel (b) for the log-MMSE gain function. γ in panel (b) denotes the *a posteriori* SNR.

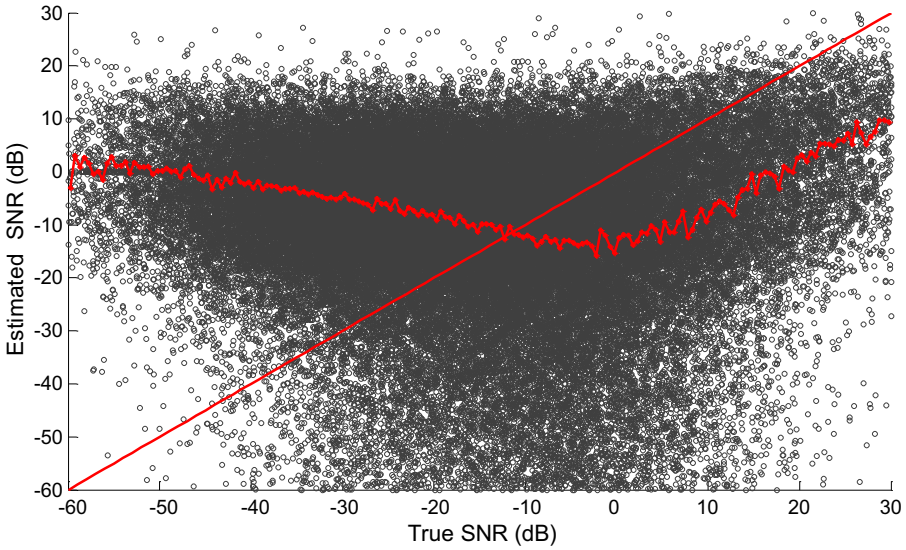

Fig. 2. Scatter plot of the true instantaneous SNR values ($\bar{\xi}$) against the corresponding estimated SNR values. The estimated SNR values were computed using the "decision-directed" approach and the Wiener gain function implemented as per (Scalart and Filho, 1996). The diagonal line indicates the perfect estimator (i.e., zero estimation error). The input global SNR was −5 dB and the background noise was babble.

are estimated according to: $\bar{\xi} = X^2/D^2$. The solid line represents (Plapous et al., 2006):

$$\hat{\xi}_{AVE} = E[\hat{\xi}|\bar{\xi}] = \int_{-\infty}^{\infty} \hat{\xi} \cdot p(\hat{\xi}|\bar{\xi}) \cdot d\hat{\xi}, \tag{5}$$

while the diagonal line represents the perfect estimator. The pattern shown in Fig. 2 was also demonstrated by others (see Plapous et al., 2006). It is clear that ξ is over-estimated for SNR < 0 dB and under-estimated for SNR > 0 dB. The ξ value is over-estimated by as much as 40–60 dB at extremely low (<-40 dB) SNR levels (see Fig. 2). The SNR over-estimation affects in turn the gain function of most statistically-based estimators (e.g., MMSE, logMMSE). Fig. 3 shows a plot of the mean of the estimated Wiener gain function against the true ξ. The bias, or shift in the Wiener gain function, relative to the true value (near 0) is clear and for this example it was substantial at low SNR levels (e.g., 0.4 at $\xi = -20$ dB SNR in Fig. 3).

To summarize, ξ is often over-estimated in the negative SNR region (see Fig. 2). As demonstrated in Fig. 1, the estimation of the gain function is particularly sensitive to perturbations of ξ in the negative SNR region. Inaccuracies

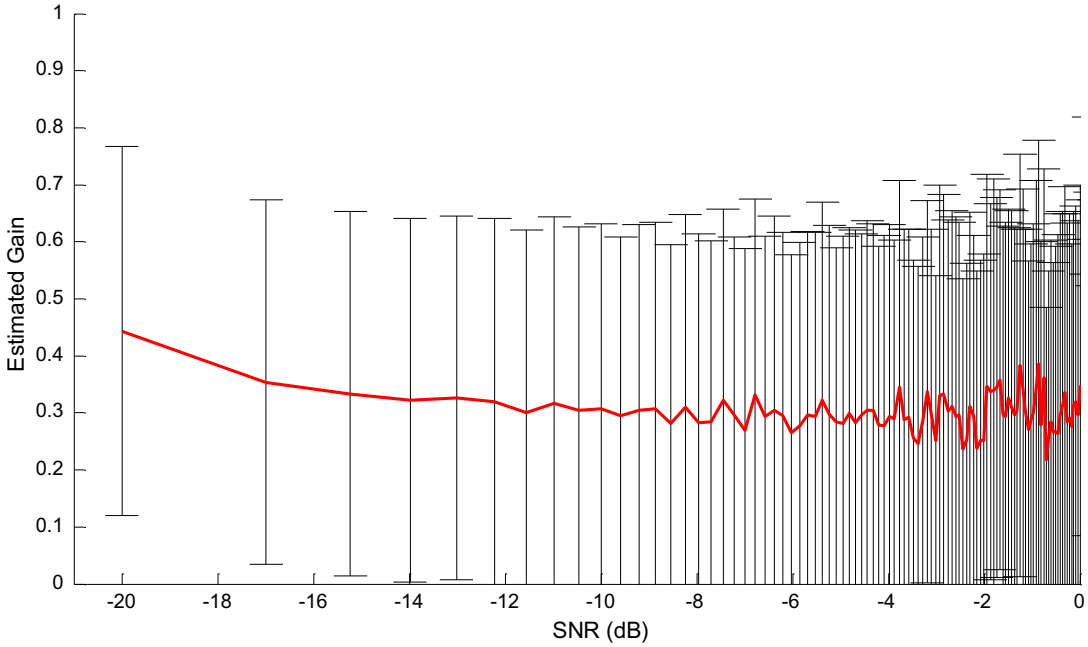

Fig. 3. Plot of the average value of the estimated gain function ($\hat{G}_W$) against the true instantaneous SNR ($\bar{\xi}$) values. The Wiener gain function was used and estimated as per (Scalart and Filho, 1996). The error bars indicate standard deviations. The input global SNR was −5 dB and the background noise was babble.

in the estimation of ξ cause an over-estimation of the gain function (see Fig. 3). But, how does that affect speech intelligibility? This is examined next.

3. Impact of SNR and gain overestimation on speech intelligibility

3.1. Conditions

To assess the impact of SNR and gain over-estimation we conducted listening studies wherein we assumed *a priori* knowledge of ξ (more precisely, we assumed *a priori* knowledge of the short-term versions of ξ). This was found necessary in order to properly control (fix) the changes in the gain function. In one set of experiments, we artificially introduced a bias in the gain function. Such a bias can be introduced by including a bias in the ξ estimation. The gain bias was introduced only in the negative SNR regions to better reflect realistic conditions (see Fig. 2). No bias in the gain function was introduced in the positive SNR region. This set of experiments simulated to some extent gain over-estimation as caused by ξ over-estimation. In a second set of experiments, we artificially introduced a bias in the gain function in the positive SNR region (no bias was introduced in the negative SNR region). More precisely, in the latter set of experiments, the gain function was purposefully attenuated.

The gain functions used in the above two experiments are shown in Fig. 4. The baseline gain function was the Wiener gain function (Eq. (1)). To create a bias in the negative SNR region, we modified the Wiener gain function as follows:

$$G_{W1} = \frac{1}{C+1}\left(\frac{\xi}{\xi+1} + C\right), \tag{6}$$

where

$$C = \frac{B}{1-B}, \tag{7}$$

and B is the amount of bias introduced in the negative SNR region. For our experiments, we considered the following values for B: 0.2, 0.4, 0.5, 0.6, and 0.7. Note that when $B = 0$ (i.e., no bias), we obtain the baseline Wiener gain function (Eq. (1)).

In the second set of experiments, we purposefully attenuated the gain function in the positive SNR region. We modified the Wiener gain function as follows:

$$G_{W2} = B \cdot \frac{\xi}{\xi+1}, \tag{8}$$

where B is the bias term. The following values of B were considered: 0.001, 0.05, 0.1, 0.2, and 0.4. Note that in the extreme case that $B = 0.001$, the gain function is effectively attenuated by 60 dB. The plots of the modified gain functions G_{W1} and G_{W2} are shown in Fig. 4 for different values of B.

3.2. Intelligibility listening tests

Eight (5 male and 3 female, mean age = 19 yrs) normal-hearing listeners participated in the listening experiments,

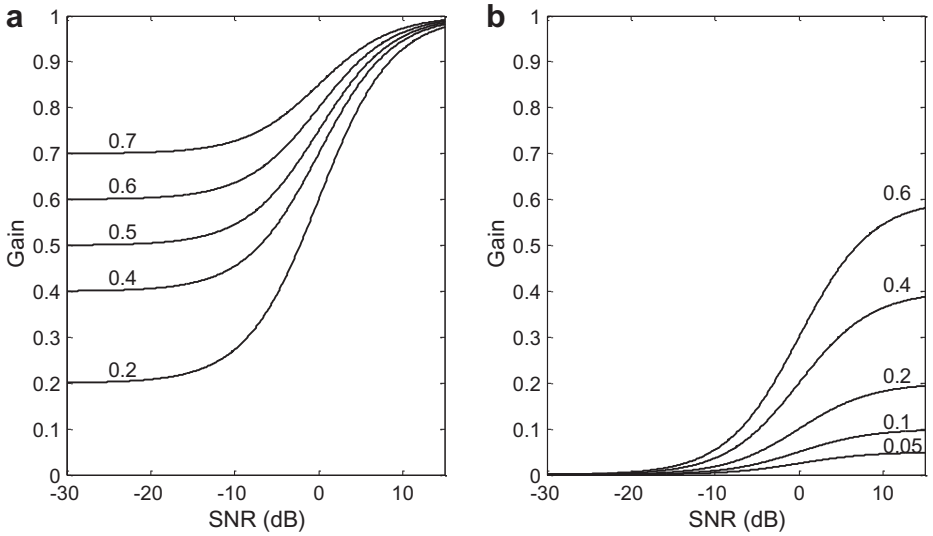

Fig. 4. Wiener gain functions modified to introduce a bias either in the negative SNR region as per Eq. (6) (panel (a)) or in the positive SNR region as per Eq. (8) (panel (b)). The numbers indicate the bias B introduced in the gain function.

and all listeners were paid for their participation. Sentences taken from the IEEE database (IEEE Subcommittee, 1969) were used for test material. The sentences in the IEEE database are phonetically balanced with relatively low word-context predictability. The sentences were recorded at a sampling rate of 25 kHz, and the recordings are available from a CD accompanying the book in (Loizou, 2007). Noisy speech was generated by adding babble noise at -10 dB and -5 dB SNR. The babble noise was produced by 20 talkers with equal number of female and male talkers. The SNR levels chosen are understandably extremely low, but they were chosen to avoid ceiling effects (e.g., performance near 100% correct), which would in term prevent us from drawing any meaningful conclusions.

Each listener participated in a total of 24 conditions (=2 SNR levels $\times$ 12 processing conditions). For each SNR level, the processing conditions included speech processed using modified Wiener filters based on: (1) 5 biased gain functions (i.e., biased by fixed bias $B = 0.2$, 0.4, 0.5, 0.6, and 0.7), and (2) 5 attenuated gain functions (i.e., attenuated by $B = 0.001$, 0.05, 0.1, 0.2, and 0.4). For comparative purposes, subjects were also presented with noise-corrupted (unprocessed) stimuli and stimuli processed by the Wiener filter implemented as per (Scalart and Filho, 1996). The noise estimation algorithm proposed in (Rangachari and Loizou, 2006) was used.

The listening experiment was performed in a sound-proof room (Acoustic Systems, Inc.) using a PC connected to a Tucker-Davis system 3. Stimuli were played to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. Before the test, each subject listened to a set of noise-corrupted sentences to be familiarized with the testing procedure. During the test, subjects were asked to write down the words they heard. Two lists of sentences (i.e., 20 sentences) were selected from the IEEE database (IEEE Subcommittee, 1969) and used for each condition, with none of the lists repeated across conditions. The intelligibility score for each condition was computed as the ratio between numbers of the correctly recognized words and the total number of words contained in 20 sentences. The order of the conditions was randomized across subjects. The testing session lasted for about 2.5 h. Subjects were given a 5-min break every 30 min during the test.

3.3. Results

The results from the intelligibility listening tests, expressed in terms of percentage of words identified correctly, are shown in Fig. 5. Panels (a) and (c) show the results from the first set of experiments, wherein the bias was introduced only in the negative SNR region. As can be seen, the gain bias had a significant effect on speech intelligibility, particularly in the extremely low SNR conditions (input SNR $= -10$ dB). Intelligibility dropped to 50% when $B = 0.4$, and to 10% when $B = 0.7$. Overall, a larger degradation in intelligibility was observed when B increased and approached the value of 1. A similar trend was also observed in the -5 dB SNR condition. High intelligibility scores were obtained in the -5 dB SNR condition compared to the scores obtained with unprocessed speech and Wiener-processed speech implemented as per (Scalart and Filho, 1996) (labeled as "Wien" in Fig. 5). This was to be expected given that in these experiments *a priori* knowledge of ξ was assumed when a gain bias was introduced. It should be noted that for the Wiener-processed speech (Scalart and Filho, 1996), the SNR values were estimated using the "decision-directed" approach.

Panels (b) and (d) show the results from the second set of experiments, wherein the bias was introduced only in the positive SNR region. Unlike the results from the first experiment, the gain bias had a minimal effect on speech

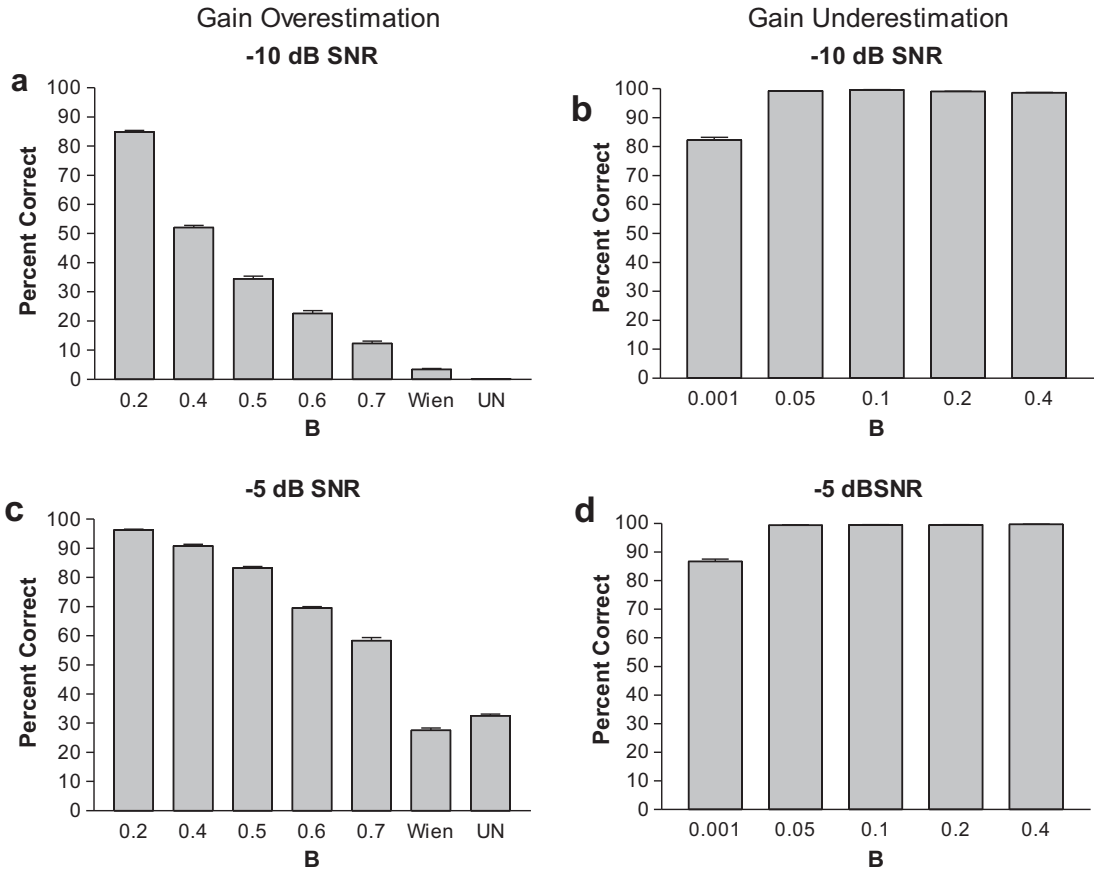

Fig. 5. Mean intelligibility scores obtained by normal-hearing listeners in the various conditions as a function of the bias B introduced in the Wiener gain function (see Fig. 4). Bars marked as 'UN' indicate intelligibility scores obtained in the un-processed noisy conditions, and bars marked as 'Wien' indicate the scores obtained with Wiener-processed speech based on the algorithm presented in (Scalart and Filho, 1996). Error bars indicate standard deviations.

intelligibility. Performance remained high (>80%) even in the extreme case where the gain function was consistently attenuated by as much as 60 dB (corresponding to $B = 0.001$). Note that in this circumstance, the gain function (plotted in linear units) was practically flat (see for instance the gain function for $B = 0.05$ in Fig. 4) across all SNR values (positive and negative). The fact that intelligibility was not affected when the gain function was underestimated in the positive SNR regions can be illustrated by examining the change in $\overline{SNR}_{ESI}$ values (see Eq. (10), Section 4) of individual frequency bins after the bias was introduced. The $\overline{SNR}_{ESI}$ metric is used here as it has been found previously (Ma et al., 2009) to correlate modestly high with intelligibility. Analysis of the $\overline{SNR}_{ESI}$ metric (Loizou and Kim, 2011) has shown that speech synthesized with spectral components of $\overline{SNR}_{ESI} > 0$ dB are more intelligible compared to speech synthesized with spectral components of $\overline{SNR}_{ESI} < 0$ dB (in fact, it can be proved in (Kim and Loizou, 2011) that spectral components with $\overline{SNR}_{ESI} < 0$ dB are always noise masked, i.e., SNR < 0 dB). Table 1 shows the average percentage of frequency bins with positive and negative $\overline{SNR}_{ESI}$ values computed before and after the bias was introduced (average percentages were based on 10 IEEE sentences). As can be seen from this table, the percentage of frequency bins with positive and negative $\overline{SNR}_{ESI}$ values remains relatively un-changed (e.g., 50.7% before bias vs. 54.0% after bias at -10 dB SNR) when the gain is underestimated in the positive SNR regions. Consequently, no change in intelligibility is expected. In contrast, the percentage of frequency bins with negative $\overline{SNR}_{ESI}$ values increases significantly (e.g., 49.3% before bias vs. 83.5% after bias at -10 dB SNR) after the bias is introduced in the negative SNR regions. More speech frequency bins are subsequently masked by noise (since $\overline{SNR}_{ESI} < 0$ dB implies SNR < 0 dB (Kim and Loizou, 2011) leading to a drop in intelligibility.

From the outcomes of the two experiments we can draw the following conclusions. In terms of preserving or improving speech intelligibility it is imperative that the gain function takes values close to 0 for SNR < 0 dB. This is necessary in order to remove masker-dominated T–F units, which are largely responsible for the loss in intelligibility. The value of the gain function in the SNR > 0 dB region had a minimal effect on intelligibility (see Fig. 5). In fact, the simplest gain function that can be considered is a binary gain function that assumes a value of 0 for SNR < 0 dB and assumes a value of 1 for SNR > 0 dB. Such binary gain functions are used in the ideal binary mask technique

Table 1
Average percentage of frequency bins with positive and negative $\overline{SNR}_{ESI}$ values (Eq. (10)) computed before and after the bias ($B = 0.7$) was introduced.

$\overline{SNR}_{ESI}$ of freq. bins	−10 dB SNR			−5 dB SNR		
	No bias ($B = 0$) (%)	$B = 0.7$ in Eq. (7) (%)	$B = 0.7$ in Eq. (8) (%)	No bias ($B = 0$) (%)	$B = 0.7$ in Eq. (7) (%)	$B = 0.7$ in Eq. (8) (%)
$\overline{SNR}_{ESI} \geqslant 0$ dB	50.7	16.5	54.0	55.2	24.0	58.7
$\overline{SNR}_{ESI} < 0$ dB	49.3	83.5	46.0	44.8	76.0	41.3

employed in computational auditory scene analysis (CASA) (Wang and Brown, 2006). The optimality of these binary gain functions has been shown in (Li and Wang, 2009; Loizou and Kim, 2011). In Loizou and Kim (2011), it was proven that these binary gain functions maximize the weighted average of the band SNRs, a metric closely related to the articulation index (AI). Consequently maximizing the articulation index ought to improve speech intelligibility, since the AI measure is highly correlated with speech reception (Kryter, 1962). Indeed, the use of such binary gain functions has been shown to yield substantial improvements in intelligibility, and this has been demonstrated in a number of studies involving normal-hearing listeners (Brungart et al., 2006; Li and Loizou, 2009). In brief, in the context of developing noise reduction algorithms, much focus needs to be placed on estimating accurately the gain function in the SNR < 0 dB region. Such algorithms are likely to improve speech intelligibility.

4. Impact of SNR overestimation on speech distortions

It was not clear from the above discussion as to whether the SNR (and gain) overestimation introduced spectral amplification distortion, spectral attenuation distortions or both. It is important to distinguish between the two since these distortions do not contribute equally to speech intelligibility loss (Loizou and Kim, 2011). More specifically, it was demonstrated in (Loizou and Kim, 2011) that the spectral amplification distortions are particularly harmful to speech intelligibility. In contrast, the spectral attenuation distortions do not impair speech intelligibility. To answer the above question, we use the signal-to-residual spectrum ratio (SNR_{ESI}) metric – also known in the literature as the frequency-weighted segmental SNR (Quackenbush et al., 1988) – as a tool. This metric has been found to correlate highly with both speech quality (Hu and Loizou, 2008) and speech intelligibility (Ma et al., 2009). The SNR_{ESI} metric can also be used to decouple the spectral amplification distortions from the spectral attenuation distortions.

The SNR_{ESI} measure can be expressed in terms of the Wiener gain function as follows (Lu and Loizou, 2010):

$$SNR_{ESI}(\xi, G) = \frac{\xi}{(1-G)^2\xi + G^2}. \tag{9}$$

Severe amplification distortions, in excess of 6 dB, are introduced when $SNR_{ESI} < 1$. More precisely, if the SNR_{ESI} is defined using short-time values of the clean and processed magnitude spectra, rather than statistical averaged spectral values (i.e., based on expected values), it is easy to show that when $\overline{SNR}_{ESI} < 1$ we have $\hat{X} > 2 \cdot X$, where $\overline{SNR}_{ESI}$ is the short-time version of Eq. (9) and is defined as (Loizou and Kim, 2011):

$$\overline{SNR}_{ESI} = \frac{X^2}{(X-\hat{X})^2} \tag{10}$$

where X denotes the clean magnitude spectrum and $\hat{X}$ the processed (via a noise reduction algorithm) magnitude spectrum obtained at given frame (the main difference between Eqs. (9) and (10) is that the first is defined using expected values while the latter is defined using short-time values of the magnitude spectra). It was demonstrated in (Loizou and Kim, 2011) via listening tests, that when $\overline{SNR}_{ESI} < 1$, speech intelligibility was severely compromised (i.e., intelligibility scores dropped to zero). This is so because T–F units that satisfy this condition ($\overline{SNR}_{ESI} < 1$) are noise masked, i.e., always have a negative SNR (see analytical proof in (Kim and Loizou, 2011)). Based on this observation, we can conclude that if the gain function falls into the $\overline{SNR}_{ESI} < 1$ region, it will severely compromise speech intelligibility. We formally define such a "forbidden" region as follows:

$$\mathbb{F} = \{G : SNR_{ESI} < 1\} \cap \{0 \leq G \leqslant 1\}. \tag{11}$$

The set shown on the right is included to ensure that the gain function is bounded. Identifying such a region is important, as the boundary of this region can serve as an upper bound for the highest value allowed for G. Using Eq. (9), and solving for G satisfying the inequality $SNR_{ESI} < 1$ we get:

$$G_F > 2\frac{\xi}{\xi + 1}. \tag{12}$$

The set of gain functions that satisfy the above equation belong to the region $\mathbb{F}$ (Eq. (11)). Fig. 6 plots the region $\mathbb{F}$ and superimposes the Wiener gain function for comparison. The shaded portion shown in Fig. 6 depicts the region $\mathbb{F}$. If the estimated gain function falls into this region, intelligibility will suffer. Unfortunately, as shown in Fig. 3, the estimated gain functions reside for the most part in this region due to SNR over-estimation. This explains the inability of current noise-reduction algorithms to improve speech intelligibility at extremely low SNR levels (see Fig. 5). Based on the above, we can define the following bounds on the estimated $\hat{G}$:

$$0 \leqslant \hat{G} \leqslant 2\frac{\xi}{\xi + 1}. \tag{13}$$

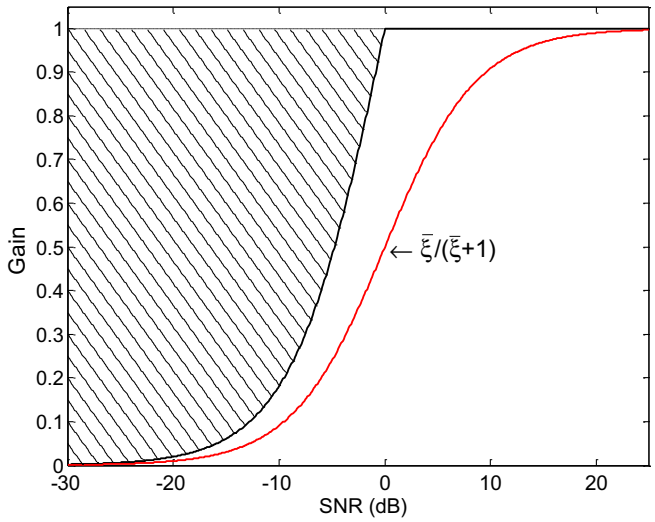

Fig. 6. Shaded portion of the graph indicates the "un-desirable" region of the gain function plotted as a function of the true instantaneous SNR ($\bar{\xi}$). When the estimated gain function falls in this region, large amplification distortions (>6 dB) are introduced in the spectrum. These distortions are largely responsible for the lack of intelligibility improvement with existing speech enhancement algorithms (Loizou and Kim, 2011). The Wiener gain function is superimposed for comparative purposes.

The upper bound is equivalent to the constraint that $\hat{X} < 2 \cdot X$. This constraint allows only attenuation distortions and limited (<6 dB) amplification distortions (Loizou and Kim, 2011). If the estimated gain function $\hat{G}$ satisfies the above inequality, then it is guaranteed that speech intelligibility will improve over that obtained by un-processed noisy speech. This was demonstrated in (Loizou and Kim, 2011; Kim and Loizou, 2011).

5. Factors contributing to SNR overestimation

So far we analyzed and discussed the detrimental effects of SNR over-estimation on speech intelligibility. But, what contributes to SNR over-estimation? This is an important question since identifying the reasons underlying SNR over-estimation can potentially lead us to the development of noise-reduction algorithms capable of improving speech intelligibility. There are at least two factors contributing to SNR over-estimation.

The first factor is attributed to the use of the "decision-directed" approach, which is often used to estimate the SNR in most statistical-model based algorithms. The "decision-directed" approach inherently yields biased estimates of the SNR. This was discussed in (Martin, 2005; Erkelens et al., 2007) and proven in (Loizou, 2007, Section 7.4.1). This bias is introduced partly due to the use of the clipping function (max) for ensuring positive SNR values, and the fact that the square of the estimator of the magnitude spectrum is used rather than the estimator of the magnitude-squared spectrum (Martin, 2005; Erkelens et al., 2007). As shown in (Erkelens et al., 2007), a $\pi/4$ bias exists even if we had access to the true signal variance. This bias, however, is not expected to be detrimental as it under-estimates the true SNR. In contrast, the bias introduced by the clipping function (max operator) may lead to over-estimation of the true SNR.

The second factor is attributed to the noise spectral variance estimation. The SNR estimate requires computation of the noise statistics, which are sometimes gathered during speech pauses or estimated/updated continuously using noise-estimation algorithms. Most noise-estimation algorithms, however, under-estimate the value of the noise spectral variance. The minimum statistics (Martin, 2001) algorithms, for instance, are designed to estimate the mean of the minimum of a set of random variables (representing past values of the noisy power spectral density). The minimum value of a set of random variables, however, is always smaller than their mean (Papoulis and Pillai, 2002). In such algorithms, the bias needs to be computed and corrected and a number of methods have been proposed to do so (Martin, 2001, 2006). Since the bias term requires knowledge of the noise variance (Martin, 2001), errors are introduced in the bias computation. Most noise tracking algorithms are unable to follow fast increases in noise level, and in those instances the noise spectral variance is under-estimated. When the noise spectral variance is under-estimated, the SNR is over-estimated since the noise term is in the denominator of the SNR calculation. Hence, SNR over-estimation is caused primarily by under-estimation of the noise spectral variance. Empirical evidence in support of this conclusion is shown in Fig. 7. This figure shows separately the scatter plots of estimated vs. true SNR values for frequency bins in which the noise spectral variance was either overestimated or underestimated (the noise-estimation algorithm in (Rangachari and Loizou, 2006) was used and the true noise variance was computed by applying a first-order recursion to the instantaneous magnitude-squared spectrum of the noise). Note that when the noise spectral variance was over-estimated, much of the SNR over-estimation errors were eliminated. In contrast, when the noise spectral variance was under-estimated, the SNR over-estimation errors were dominant. The top two panels in Fig. 7 show the histograms of SNRs of frequency bins for which the noise spectral variance was either over-estimated or under-estimated. Frequency bins corresponding to noise-overestimated bins had on the average a higher SNR, suggesting that speech intelligibility ought to be high when retaining those frequency bins. Indeed, listening experiments reported in (Kim and Loizou, 2011) confirmed that high intelligibility scores could be obtained when only retaining frequency bins that over-estimate the noise spectral variance. In contrast, when frequency bins were retained for which the noise spectral variance was under-estimated, intelligibility scores dropped to zero.

6. Conclusions

The present study analyzed the impact of errors in SNR and gain-function estimation on speech intelligibility. Listening tests indicated that SNR and gain-function

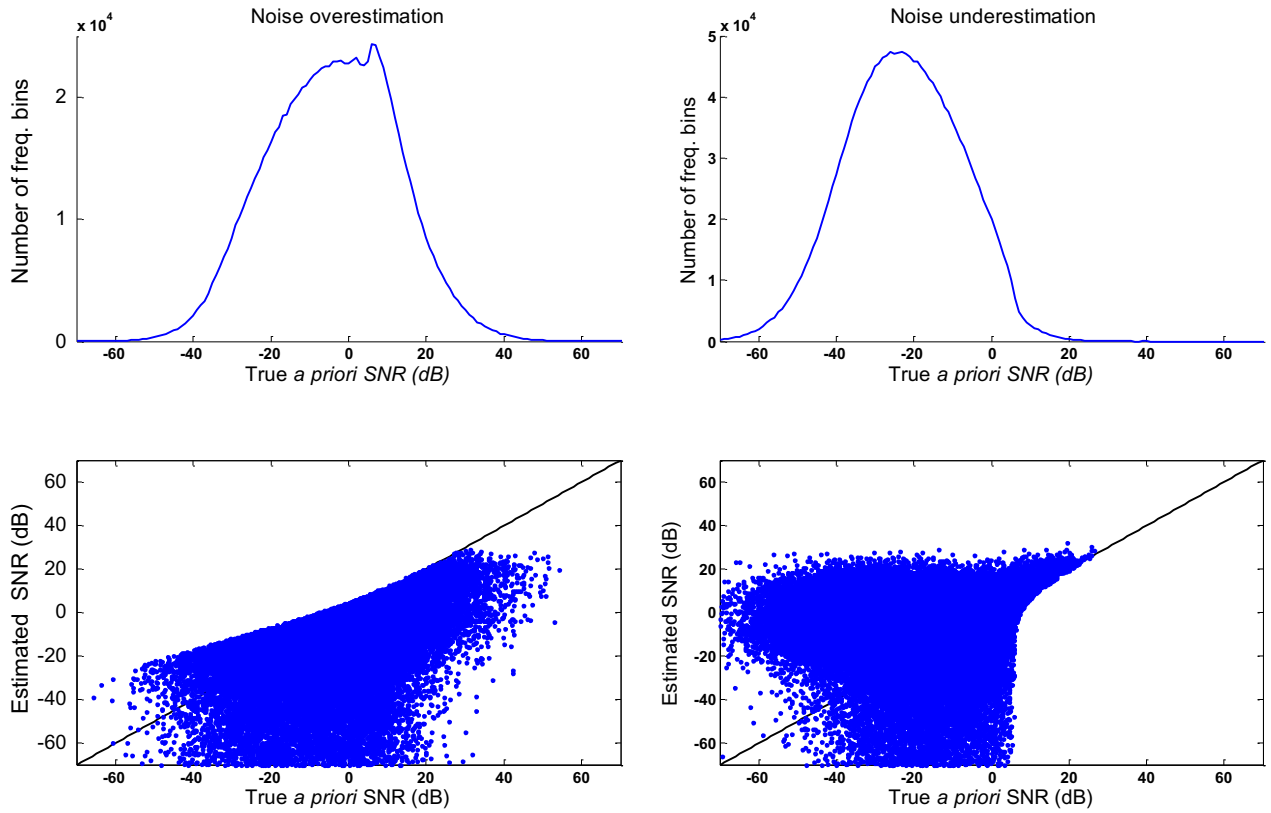

Fig. 7. Top row shows the histograms of SNR values of frequency bins wherein the noise spectral variance was either over-estimated (left) or under-estimated (right). Bottom row shows the scatter plots of the true and estimated SNR values for frequency bins wherein the noise spectral variance was over-estimated (left) or under-estimated (right).

overestimation errors in frequency bins with negative SNR are particularly harmful to speech intelligibility. The SNR overestimation errors were attributed primarily to the underestimation of the noise spectrum (Kim and Loizou, 2011), which is needed for the computation of the SNR. Most noise-estimation algorithms underestimate the value of the noise spectral variance as they are unable to follow fast increases in noise level. A theoretical upper bound (Eq. (13)) on the gain function was derived that can be used to constrain the values of the gain function so as to ensure that SNR overestimation errors are minimized. Speech enhancement algorithms that can limit the values of the gain function to fall within this upper bound can improve speech intelligibility (Loizou and Kim, 2011; Kim and Loizou, 2011). Overall, the outcomes of the present study suggest that better methods are needed to estimate the spectral SNR from noisy observations, particularly at low input SNR levels. Such methods hold promise for improving speech intelligibility (e.g., Kim et al., 2009).

Acknowledgement

This research was supported by Grant No. R01 DC010494 from the National Institute of Deafness and other Communication Disorders, NIH.

References

Berouti, M., Schwartz, M., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 208–211.

Breithaupt, C., Martin, R., 2011. Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions. IEEE Trans. Audio Speech Lang. Process. 19, 277–289.

Brungart, D., Chang, P., Simpson, B., Wang, D., 2006. Isolating the energetic component of speech-on-speech masking with ideal time–frequency segregation. J. Acoust. Soc. Amer. 120, 4007–4018.

Cappe, O., 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. 2, 346–349.

Cohen, I., 2005. Relaxed statistical model for speech enhancement and a priori SNR estimation. IEEE Trans. Speech Audio Process. 13, 870–881.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32, 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33, 443–445.

Erkelens, J., Jensen, J., Heusdens, R., 2007. A data-driven approach to optimizing spectral speech enhancement methods for various error criteria. Speech Commun. 49, 530–541.

Hu, Y., Loizou, P., 2007. Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun. 49, 588–601.

Hu, Y., Loizou, P., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 16, 229–238.

IEEE Subcommittee, 1969. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust., 17, 225–246.

Kim, G., Loizou, P., 2010. A new binary mask based on noise constraints for improved speech intelligibility. Proc. Interspeech, 1632–1635.

Kim, G., Loizou, P., 2011. Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms. J. Acoust. Soc. Amer. 130, 1581–1596.

Kim, G., Lu, Y., Hu, Y., Loizou, P., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Amer. 126, 1486–1494.

Kryter, K., 1962. Validation of the articulation index. J. Acoust. Soc. Amer. 34, 1698–1706.

Li, N., Loizou, P., 2009. Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction. J. Acoust. Soc. Amer. 123, 1673–1682.

Li, Y., Wang, D., 2009. On the optimality of ideal time–frequency masks. Speech Commun. 51, 230–239.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press LLC, Boca Raton, Florida.

Loizou, P., Kim, G., 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Acoust. Speech Signal Process. 19, 47–56.

Lu, Y., Loizou, P., 2010. Speech enhancement by combining statistical estimators of speech and noise. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 4754–4757.

Ma, J., Hu, Y., Loizou, P., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Amer. 125, 3387–3405.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9, 504–512.

Martin, R., 2005. Statistical methods for the enhancement of noisy speech. In: Benesty, J., Makino, S., Chen, J. (Eds.), Speech Enhancement. Springer, Berlin, pp. 43–64.

Martin, R., 2006. Bias compensation methods for minimum statistics noise power spectral density estimation. Signal Process. 86, 1215–1229.

Papoulis, A., Pillai, S., 2002. Probability Random Variables and Stochastic Processes, 4th ed. McGraw Hill, Inc., New York.

Plapous, C., Marro, C., Scalart, P., 2006. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 14, 2098–2108.

Quackenbush, S., Barnwell, T., Clements, M., 1988. Objective Measures of Speech Quality. Prentice-Hall, Englewood Cliffs, NJ.

Rangachari, S., Loizou, P., 2006. A noise-estimation algorithm for highly non-stationary environments. Speech Commun. 48, 220–231.

Scalart, P., Filho, J., 1996. Speech enhancement based on a priori signal to noise estimation. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 629–632.

Wang, D., Brown, G., 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley, Hoboken, NJ.

Whitehead, P., Anderson, D., 2011. Robust Bayesian analysis applied to Wiener filtering of speech. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., 5080–5083.