Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms[a)]

Gibak Kim[b)] and Philipos C. Loizou[c)]
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75080

(Received 5 January 2010; revised 30 June 2011; accepted 2 July 2011)

Most noise-reduction algorithms used in hearing aids apply a gain to the noisy envelopes to reduce noise interference. The present study assesses the impact of two types of speech distortion introduced by noise-suppressive gain functions: amplification distortion occurring when the amplitude of the target signal is over-estimated, and attenuation distortion occurring when the target amplitude is under-estimated. Sentences corrupted by steady noise and competing talker were processed through a noise-reduction algorithm and synthesized to contain either amplification distortion, attenuation distortion or both. The attenuation distortion was found to have a minimal effect on speech intelligibility. In fact, substantial improvements (>80 percentage points) in intelligibility, relative to noise-corrupted speech, were obtained when the processed sentences contained only attenuation distortion. When the amplification distortion was limited to be smaller than 6 dB, performance was nearly unaffected in the steady-noise conditions, but was severely degraded in the competing-talker conditions. Overall, the present data suggest that one reason that existing algorithms do not improve speech intelligibility is because they allow amplification distortions in excess of 6 dB. These distortions are shown in this study to be always associated with masker-dominated envelopes and should thus be eliminated.
 [DOI: 10.1121/1.3619790]

PACS number(s): 43.72.Ar, 43.72.Dv, 43.71.Es [RYL] Pages: 1581–1596

I. INTRODUCTION

Much progress has been made in the development of single-microphone noise reduction algorithms for hearing aid applications (Edward, 2004; Bentler and Chiou, 2006) and speech communication systems (Loizou, 2007). The majority of these algorithms have been found to improve listening comfort and speech quality (Baer *et al.,* 1993; Hu and Loizou, 2007b; Bentler *et al.,* 2008). In stark contrast, little progress has been made in designing single-microphone noise-reduction algorithms that can improve speech intelligibility. Past intelligibility studies conducted in the late 1970s (Lim, 1978) found no intelligibility improvement with the spectral subtraction algorithm. In the intelligibility study by Hu and Loizou (2007a), conducted nearly 30 years later, none of the eight single-microphone noise-reduction algorithms were found to improve speech intelligibility relative to un-processed (corrupted) speech. Noise-reduction algorithms implemented in wearable hearing aids revealed no significant intelligibility benefit (Levitt, 1997; Bentler *et al.,* 2008), although they have been found to improve speech quality and ease of listening in hearing-impaired listeners (e.g., Bentler *et al.,* 2008; Luts *et al.,* 2010). Some of the noise-reduction algorithms proposed for hearing aids rely on modulation spectrum filtering (Alcantara *et al.,* 2003; Bentler and Chiou, 2006), others rely on reducing the upward spread of masking (Neuman and Schwander, 1987; van Tasell and Crain, 1992) while others rely on improving the spectral contrast (e.g., Baer *et al.,* 1993). However, none of these algorithms improved consistently and substantially speech intelligibility (Tyler and Kuk, 1989; Dillon and Lovegrove, 1993; Alcantara *et al.,* 2003; Edward, 2004; Bentler *et al.,* 2008). In brief, the ultimate goal of developing (and implementing) an algorithm that would improve substantially speech intelligibility for normal-hearing and/or hearing-impaired listeners has been elusive for nearly three decades. Algorithms that have been optimized to operate in specific noisy environments have proved recently to be very promising as they have been shown to improve speech intelligibility in studies with normal-hearing listeners (Kim *et al.,* 2009; Kim and Loizou, 2010).

Our knowledge surrounding the factors contributing to the lack of intelligibility benefit with existing single-microphone noise-reduction algorithms is limited (Ephraim, 1992; Weiss and Neuman, 1993; Levitt, 1997; Kuk *et al.,* 2002; Chen *et al.,* 2006; Dubbelboer and Houtgast, 2007). In most cases we do not know how, and to what extent, a specific parameter of a noise-reduction algorithm needs to be modified so as to improve speech intelligibility. Clearly, one factor is related to the fact that we often are not able to estimate accurately the background noise spectrum, which is needed for the implementation of most single-microphone algorithms. While noise tracking or voice activity detection algorithms have been found to perform well in steady background noise (e.g., car) environments [see review in Loizou (2007, Chap. 9)], they generally do not perform well in non-stationary types of noise (e.g., multi-talker babble). The second factor is that the majority of algorithms introduce distortions,

[a)]Part of this work was presented at the International Conference on Acoustics, Speech and Signal Processing (ICASSP) in Dallas, TX, 2010.
[b)]Present address: Department of Electrical Engineering, Soongsil University, Seoul, Korea.
[c)]Author to whom correspondence should be addressed. Electronic address: loizou@utdallas.edu

which in some cases, might be more damaging than the background noise itself (Hu and Loizou, 2007a). For that reason, several algorithms have been proposed to minimize speech distortion while constraining the amount of noise distortion introduced to fall below a preset value (Ephraim and Trees, 1995; Chen *et al.*, 2006) or below the auditory masking threshold (Hu and Loizou, 2004). Aside from the distortions introduced by noise-suppression algorithms from inaccuracies in estimating the gain function, hearing aids may also introduce other non-linear distortions such as hard, soft and asymmetrical clipping distortions (Arehart *et al.*, 2007; Tan and Moore, 2008). The perceptual effect of such distortions on intelligibility are not examined in this paper. Third, non-relevant stochastic modulations arising from the non-linear noise-speech interaction can contribute to reduction in speech intelligibility, and in some cases more so than deterministic modulation reduction (Noordhoek and Drullman, 1997). In a study assessing the effects of noise on speech intelligibility, Dubbelboer and Houtgast (2007) have shown that the systematic envelope lift (equal to the mean noise intensity) implemented in spectral subtractive algorithms had the most detrimental effects on speech intelligibility. The corruption of the fine-structure and introduction of stochastic envelope fluctuations associated with the inaccurate estimates of the noise intensity and non-linear processing of the mixture envelopes further diminished speech intelligibility. It was argued that it was these stochastic effects that prevented spectral subtractive algorithms from improving speech perception in noise (Dubbelboer and Houtgast, 2007).

Most noise-reduction algorithms used in commercial hearing aids involve two sequential stages of processing (Chung, 2004; Bentler and Chiou, 2006), as shown in Fig. 1. In the first stage, the algorithm performs signal detection and analysis with the intent of identifying the presence (or absence) of speech and noise in each band. Detectors are employed to estimate the modulation rate, modulation depth, or/and SNR in each frequency band (Schum, 2003; Latzel *et al.*, 2003; Chung, 2004; Bentler and Chiou, 2006). The Siemens (Triano) hearing aid, for instance, decides whether speech is present in a particular band based on the modulation rate (Chung, 2004), while the Widex (Senso Diva) hearing aid detects speech presence based on the estimated SNR (Kuk *et al.*, 2002). In the second stage, the mixture envelope is subjected to gain reduction based on the estimated modulation rate or SNR of each band determined in the first stage. Gain reductions can range from 0 to 12 dB in some commercial hearing aids (Alcantara *et al.*, 2003), with some hearing aids equipped with several gain settings ranging from mild to severe (Chung, 2004). The amount of gain reduction is typically inversely proportional to the SNR estimated in each channel (Kuk *et al.*, 2002; Chung, 2004). In the Siemens

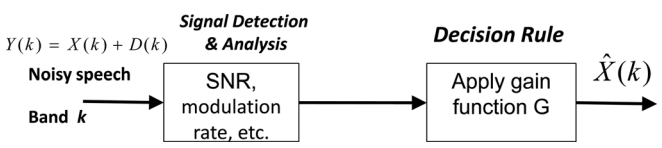

FIG. 1. Signal-processing stages involved in noise-reduction algorithms for hearing-aid applications.

(Triano) hearing aid for instance, the amount of gain reduction depends on the modulation rate/SNR and the exact amount is described by the Wiener gain function (Chung, 2004; Palmer *et al.*, 2006). The Wiener filtering algorithm (Wiener, 1949), much like many algorithms used in hearing aids (Graupe *et al.*, 1987; Kuk *et al.*, 2002; Alcantara *et al.*, 2003), applies a gain to the spectral envelopes in proportion to the estimated SNR in each frequency bin. More precisely, spectral bins with high SNR receive a high gain (close to 1), while spectral bins with low SNR, and presumably masked by noise, receive a low gain (close to 0). The Wiener gain function has also been used successfully (although under somewhat ideal conditions) for hearing impaired listeners by Levitt *et al.* (1993).

Clearly, the choice of the frequency-specific gain function is critical to the success of the noise-reduction algorithm (Kuk *et al.*, 2002; Bentler and Chiou, 2006). The frequency-specific gain function applied to the spectral mixture envelopes is far from perfect as it depends on the *estimated* SNR or *estimated* modulation rate (Kuk *et al.*, 2002; Chung, 2004). Although the intention (and hope) is to apply a small gain (near 0) only when the masker is present and a high gain (near 1) only when the target is present, that is not feasible since the target and masker signals spectrally overlap. Consequently, the target signal may in some instances be over-attenuated (to the point of being eliminated) while in other instances, it may be over-amplified. Despite the fact that the gain function is typically bounded between 0 and 1, the target signal may be over-amplified because the gain function is applied to the mixture envelopes. In brief, there are two types of envelope distortions that can be introduced by the gain functions used in most noise-reduction algorithms: amplification distortion occurring when the target signal is over-estimated (e.g., if the true value of the target envelope is say A, and the estimated envelope is $A + \Delta A$, for some positive increment ΔA), and attenuation distortion occurring when the target signal is under-estimated (e.g., the estimated envelope is $A–\Delta A$). These distortions may be introduced by any gain function independent of whether the gain is determined by the modulation rate, modulation depth, or SNR. The perceptual effect of these two distortions on speech intelligibility cannot be assumed to be equivalent, and in practice, there has to exist the right balance between these two distortions.

In the present study, we assess the impact of the two types of envelope distortions introduced by the gain function on the intelligibility of noise-suppressed speech. While these distortions will invariably affect the subjective speech quality, we focus in the present study only on the effects on intelligibility. The impact of these distortions on intelligibility was assessed in our prior study (Loizou and Kim, 2011), but using only one type of masker (babble) and for (limited bandwidth) telephone speech. Given the potential influence of signal bandwidth (e.g., Stelmachowicz *et al.*, 2007) and nature of the masker (modulated vs non-modulated) on speech intelligibility, the present article extends our prior study and assesses the effects of the two distortions using wideband speech corrupted by either steady noise or competing talker. Wideband speech is processed through a conventional noise-reduction algorithm (square-root Wiener filtering) while controlling the two types of distortions introduced. We subsequently synthesize signals

containing either only amplification distortion or only attenuation distortion. It should be noted that the processed signal from most noise-reduction algorithms used in commercially available hearing aids contain both distortions, but the individual contribution of each of the two distortions on speech intelligibility is largely unknown. It is hypothesized that only when the two types of distortions are properly controlled (limited) or eliminated, we can expect to observe a substantial benefit in intelligibility with existing noise-reduction algorithms.

II. GAIN-INDUCED DISTORTIONS AND SPEECH INTELLIGIBILITY: THEORETICAL ANALYSIS

As mentioned above, most (if not all) noise-suppression algorithms employed for hearing aids or for other applications involve a gain reduction stage (see Fig. 1), in which the mixture envelope or spectrum is multiplied by a gain function (taking values from 0 to 1) with the intent of suppressing background noise, if present. The amount of gain reduction depends, among others, on the detected modulation rate or estimated SNR, and typically no gain is applied if the estimated SNR is found to be too high (e.g., >12 dB in some hearing aids) (Chung, 2004). The shape and choice of the gain function varies across manufacturers, but independent of its shape, when the gain function is applied to the mixture envelopes (or spectra) it introduces either amplification or attenuation distortion to the envelopes. The gain-induced amplification distortion, for instance, is introduced when the envelope amplitude of the noise-suppressed signal (denoted as $|\hat{X}|$ in Fig. 1) is larger than the corresponding target envelope prior to noise corruption (indicated as $|X|$ in Fig. 1). This overamplification is caused by the presence of additive noise.

To analyze the impact of gain-induced distortions introduced by noise-reduction algorithms, on speech intelligibility, one needs to establish a relationship between distortion and intelligibility or alternatively develop an appropriate intelligibility measure. Such a measure could provide valuable insights as to whether we ought to design algorithms that would minimize the attenuation distortion, the amplification distortion or both, and to what degree. In the present study, we chose an intelligibility measure which has been found by Ma *et al.* (2009) to correlate highly ($r=0.81$) with the intelligibility of noise-suppressed speech. The intelligibility measure, denoted as the frequency-weighted segmental SNR (fwSNRseg) measure, was computed using the following equation:

$$fwSNRseg = \frac{10}{T}\sum_{t=0}^{T-1}\frac{1}{\sum_{k=1}^{K}W(k,t)} \times \sum_{k=1}^{K}W(k,t)\log_{10}\mathrm{SNR}_{\mathrm{ESI}}(k,t), \quad (1)$$

where $W(k,t)$ is the weight placed on the kth frequency band and time frame t, K is the number of frequency bands, T is the total number of time frames in the signal and $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ denotes the SigNal-to-RESIdual spectrum ratio:

$$\mathrm{SNR}_{\mathrm{ESI}}(k,t) = \frac{|X(k,t)|^2}{\left(|X(k,t)| - |\hat{X}(k,t)|\right)^2} \quad (2)$$

where $|X(k,t)|$ denotes the clean magnitude spectrum and $|\hat{X}(k,t)|$ denotes the signal magnitude spectrum *estimated* by the noise-reduction algorithm (see Fig. 1). The spectrum $|\hat{X}(k,t)|$ can be computed, for instance, by applying a gain function to the noisy speech spectrum, and it represents here the output of the noise-suppression algorithm (Fig. 1). We regard $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ as a local metric assessing the normalized "distance" between the true spectrum envelope and the processed (or estimated) spectrum. Clearly, the closer the noise-suppressed magnitude spectrum $|\hat{X}(k,t)|$ is to the true magnitude spectrum $|X(k,t)|$, the higher the value of the $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ metric, and consequently the higher value of the fwSNRseg measure [Eq. (1)]. It can be easily shown that the $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ metric can alternatively be expressed as a function of the ratio of the estimated (processed) to true magnitude spectra, i.e.,

$$\mathrm{SNR}_{\mathrm{ESI}}(k,t) = \frac{1}{\left(1-\frac{|\hat{X}(k,t)|}{|X(k,t)|}\right)^2}. \quad (3)$$

Figure 2 plots $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ as a function of the ratio of the estimated to clean magnitude spectra, i.e., $|\hat{X}(k,t)|/|X(k,t)|$. As can be seen, the values of $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ can be divided into different regions depending on whether the ratio $|\hat{X}(k,t)|/|X(k,t)|$ is smaller or larger than 1 or smaller or larger than 2. This figure provides important insights about the contributions of the two distortions on the value of the $\mathrm{SNR}_{\mathrm{ESI}}$, and for convenience, we divide the figure into three regions according to the distortions introduced.

Region I. In this region, $|\hat{X}(k,t)| \le |X(k,t)|$, suggesting only attenuation distortion.

Region II. In this region, $|X(k,t)| < |\hat{X}(k,t)| \le 2\cdot|X(k,t)|$, suggesting amplification distortion ranging from 0 to 6.02 dB.

Region III. In this region, $|\hat{X}(k,t)| > 2\cdot|X(k,t)|$, suggesting amplification distortion in excess of 6.02 dB.

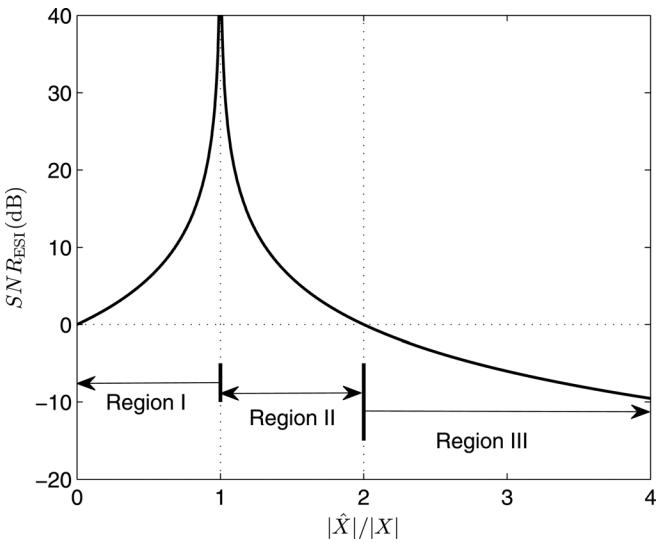

FIG. 2. Plot showing the relationship between $\mathrm{SNR}_{\mathrm{ESI}}$ and the ratio of enhanced ($|\hat{X}|$) to clean ($|X|$) spectra.

The above three regions are clearly labeled in Fig. 2. From the above, we can deduce that for the union of Regions I and II, which we denote as Region I + II, we have the following constraint:

$$\left|\hat{X}(k,t)\right| \leq 2 \cdot |X(k,t)|. \quad (4)$$

Figure 2 shows the relationship between the two envelope distortions, and their potential impact on speech intelligibility. According to this figure, in order to obtain large values for the SNR_{ESI} metric [and subsequently large values of the fwSNRseg intelligibility measure via its relationship in Eq. (1)], the envelope distortions need to be contained within Regions I and II. This is because the SNR_{ESI} metric assumes large values (and in dB, it is always positive, or 0) in Regions I and II. The assumption made here is that when the SNR_{ESI} metric attains large values across all bands, it will lead to a large overall fwSNRseg value [see Eq. (1)], and subsequently higher intelligibility. Amplification distortions in excess of 6 dB (i.e., Region III), on the other hand, can be damaging to speech intelligibility (since the SNR_{ESI} metric assumes small values in Region III, and in dB, it is negative) and consequently should be minimized. These two observations taken together imply that in order for noise-reduction algorithms to improve speech intelligibility, the amplification distortions need to be controlled in a way such that they are limited to be less than 6 dB, i.e., confined within Regions I and II. Thus, in the following experiment, we test the hypothesis that when the envelope distortions introduced by the gain function (as used by most noise-reduction algorithms) are constrained to fall within Regions I and II, substantial improvements in intelligibility are to be expected.

III. EXPERIMENT 1: EFFECT OF GAIN-INDUCED DISTORTIONS ON SPEECH INTELLIGIBILITY

In this experiment, we first process noise-corrupted sentences via a conventional noise-reduction algorithm (square-root Wiener filtering algorithm), monitor the two types of envelope distortions introduced by the gain function, and synthesize the signal accordingly by either allowing attenuation distortion alone, amplification distortion alone or both. More precisely, we constrain the distortions introduced by the gain function to fall within one of the three regions (or combinations thereof) shown in Fig. 2. The synthesized signals are presented to normal-hearing listeners for identification.

A. Methods

1. Subjects and material

Seven normal-hearing listeners were recruited for this listening experiment. They were all native speakers of American English, and were paid for their participation. Institute of Electrical and Electronics Engineers (IEEE) sentences[1] [IEEE (1969)] were used for test material, as they are phonetically balanced and have relatively low word-context predictability. The sentences were recorded at a sampling rate of 25 kHz in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The IEEE recordings are available from Loizou (2007). The sentences were corrupted by speech-shaped noise (SSN) and a single-talker (male) masker at −10, −5, and 0 dB SNRs. The speech-shaped noise was stationary having the same long-term spectrum as the sentences in the IEEE corpus. Speech produced by the same talker was used as the masker. The longest (in duration) sentence from the IEEE corpus was used for the single-talker masker. This sentence was self-duplicated and concatenated to produce a 7 sec long masker sentence. A segment of the masker was randomly cut from the masker waveforms (SSN or concatenated single-talker sentence) and mixed with the target sentences at the prescribed SNR levels. Hence, each sentence contained a different segment of the masker waveforms.

2. Signal processing

In one of the control conditions, the noise-corrupted sentences were processed by a conventional noise-suppression algorithm, namely, the Wiener algorithm (Wiener, 1949). The square-root Wiener algorithm, as implemented by Scalart and Filho (1996), was chosen as it is easy to implement, requires little computation and has been shown by Hu and Loizou (2007a, 2007b) to be equally effective, in terms of speech quality and intelligibility, as other more sophisticated noise-reduction algorithms.[2] Furthermore, the shape of the square-root Wiener gain function is similar to that used by some commercially available hearing aids (Chung, 2004), and provides a moderate amount of gain reduction [see Fig. 9 in Chung (2004)].

The corrupted sentences were first segmented into 20 ms frames, with 50% overlap between adjacent frames. Each speech frame was Hann windowed and a 500-point discrete Fourier transform (DFT) was computed. Let $Y(k,t)$ denote the noisy spectrum at time frame t and frequency band k. Then, the estimate of the signal magnitude spectrum, $\left|\hat{X}(k,t)\right|$, is obtained by multiplying $|Y(k,t)|$ with the square-root Wiener gain function $G(k,t)$ as follows:

$$\left|\hat{X}(k,t)\right| = G(k,t) \cdot |Y(k,t)|. \quad (5)$$

The square-root Wiener gain function is calculated based on the following equation:

$$G(k,t) = \sqrt{\frac{\mathrm{SNR}_{\mathrm{prio}}(k,t)}{1+\mathrm{SNR}_{\mathrm{prio}}(k,t)}}, \quad (6)$$

where SNR_{prio} is the *a priori* SNR estimated using the following recursive equation:

$$\mathrm{SNR}_{\mathrm{prio}}(k,t) = \alpha \cdot \frac{\left|\hat{X}(k,t-1)\right|^2}{\hat{\lambda}_D(k,t-1)} + (1-\alpha) \cdot \max\left[\frac{|Y(k,t)|^2}{\hat{\lambda}_D(k,t)} - 1, 0\right], \quad (7)$$

where $\hat{\lambda}_D(k,t)$ is the estimate of the background noise power spectrum and α is a smoothing constant (typically set to

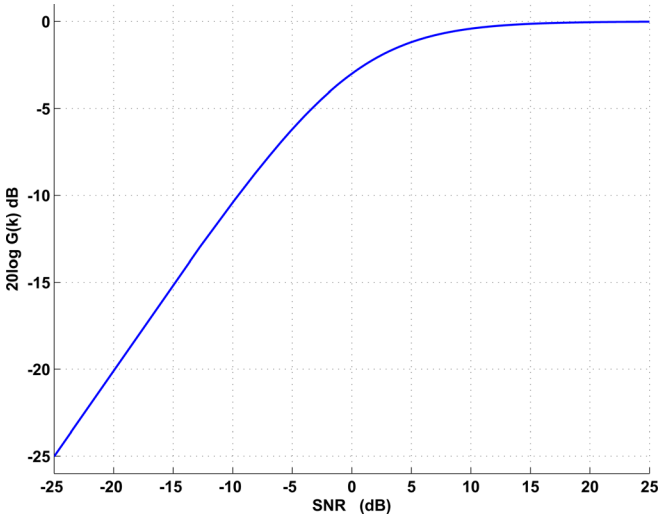

FIG. 3. (Color online) The square-root Wiener gain function used in the present study.

$\alpha = 0.98$). The noise-estimation algorithm proposed by Rangachari and Loizou (2006) was used for estimating the background noise power spectrum in Eq. (7). Following Eq. (6), an inverse DFT was applied to the processed magnitude spectrum $|\hat{X}(k,t)|$, using the phase of the noisy speech spectrum. The overlap-and-add technique was finally used to synthesize the noise-suppressed signal.

The square-root Wiener gain function is plotted in Fig. 3. Two things are worth noting about this gain function. First, the slope of the gain function is approximately 1 (at least for the region where SNR < -5 dB), in that the gain is reduced by 1 dB for every 1 dB decrease in the SNR. This corresponds to a moderate gain setting in some noise-suppression algorithms implemented in commercially available hearing aids (Chung, 2004). Second, no gain reduction is applied when the estimated SNR exceeds 15 dB, similar to the gain functions [see Fig. 9 in Chung (2004)] used in some commercially available hearing aids. In summary, the square-root Wiener gain function described in Eq. (6) is similar in many respects to those used in some hearing aids (e.g., Palmer *et al.*, 2006). It should also be noted that unlike the Wiener filter used in the study by Levitt *et al.* (1993) under ideal conditions, the square-root Wiener filter used in the present study was estimated from the mixture envelopes.

No constraints were imposed in Eq. (6) on the two types of distortions that can be incurred when applying the square-root Wiener gain function to the corrupted speech spectrum. As such, the square-root Wiener-processed sentences served as one of the two control conditions. For the remaining conditions, we assumed knowledge of the clean speech spectrum. This was necessary in order to implement the aforementioned constraints and assess the impact of the two distortions on speech intelligibility. Thus, in order to enforce the constraints, the estimated [as per Eq. (5) and Eq. (6)] magnitude spectrum $|\hat{X}(k,t)|$ was compared against the true speech spectrum $|X(k,t)|$ for each time-frequency (T-F) unit (k,t), and T-F units satisfying the constraint were retained, while T-F units violating the constrains were zeroed out. For instance, for the implementation of the Region I constraint, the modified magnitude spectrum, $|\hat{X}_M(k,t)|$, was computed as follows:

$$|\hat{X}_M(k,t)| = \begin{cases} |\hat{X}(k,t)|, & \text{if } |\hat{X}(k,t)| \le |X(k,t)| \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Following the above selection of T-F units belonging in Region I, an inverse DFT was applied to the modified spectrum $|\hat{X}_M(k,t)|$ using the phase of the noisy speech spectrum, and the overlap-and-add technique was finally used to

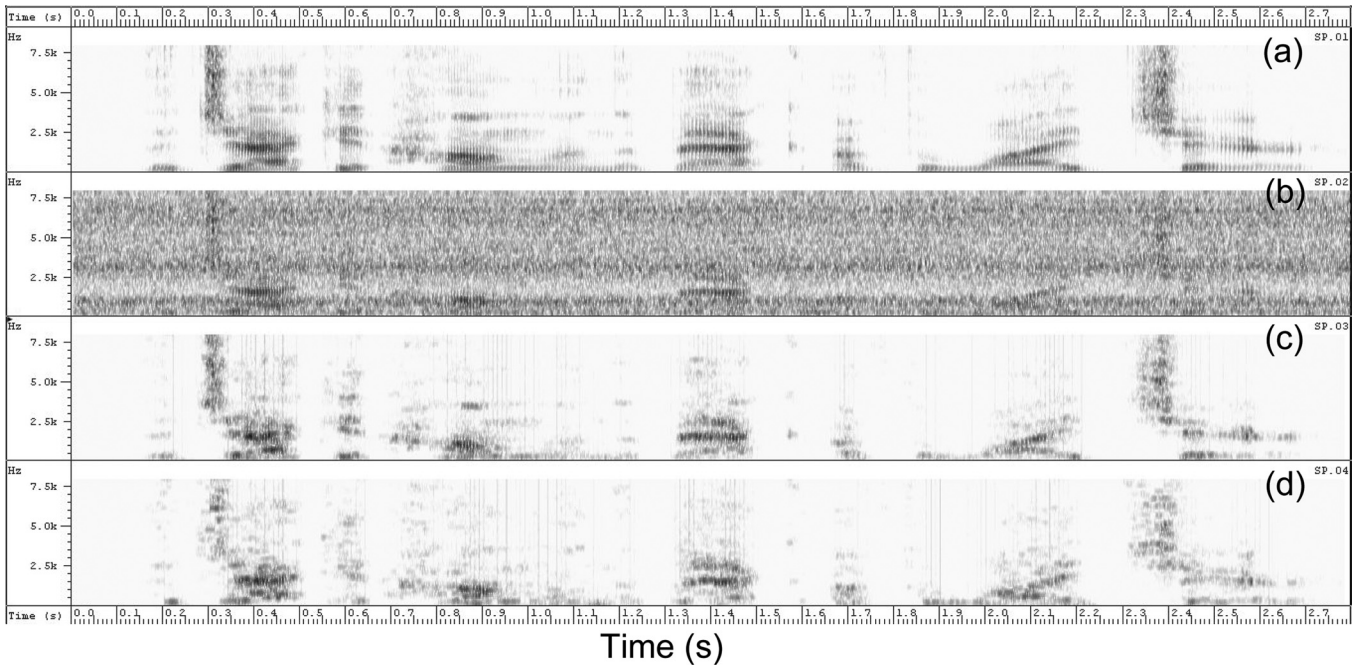

FIG. 4. Wideband spectrograms of the (a) clean signal, (b) corrupted signal (SSN masker, SNR $= -5$ dB), (c) square-root Wiener-processed signal with Region I constraints, and (d) square-root Wiener-processed signal with Region II constraints.

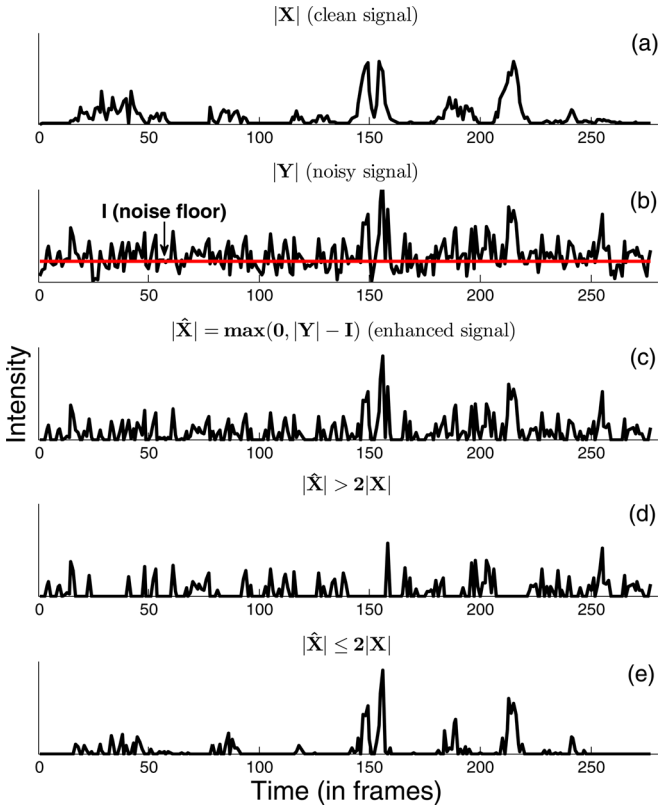

FIG. 5. (Color online) Example temporal envelopes of a band (centered at $f = 700$ Hz) processed so as to contain only amplification or attenuation distortions. (a) The clean envelope. (b) The noisy envelope corrupted at 0 dB SSN. (c) Envelope processed by a spectral subtractive algorithm. (d) The envelope containing only amplification distortions in excess of 6 dB. (e) The envelope containing only attenuation distortion and limited (< 6 dB) amplification distortion.

synthesize the noise-suppressed signal containing the prescribed envelope distortion (MATLAB implementation of the above algorithm is available from the second author). Figure 4 shows example spectrograms of a corrupted (by SSN masker at -5 dB SNR) IEEE sentence, processed and synthesized to contain only attenuation distortion (Region I) or limited amplification distortion (Region II). As can be seen, the processed signals contained adequate formant frequency information for accurate word identification. A relatively smaller number of T-F units were retained in Region II [Fig. 4(d)] compared to that in Region I [Fig. 4(c)].

Figure 5 shows example temporal envelopes for a specific band (centered at $f = 700$ Hz) containing prescribed envelope distortions. For illustrative purposes, and similar to Dubbelboer and Houtgast (2007), we show the envelopes processed via a spectral subtraction algorithm which operates by subtracting the noise floor intensity from the noisy envelope [Figs. 5(b) and 5(c)]. The resulting envelope containing only amplification distortion (in excess of 6 dB) is shown in Fig. 5(d), and the envelope containing primarily attenuation distortion and limited amplification distortion (< 6 dB) is shown in Fig. 5(e). It is clear that the envelopes constrained to lie within Region I + II [Fig. 5(e)] contain primarily speech-relevant modulations, while the envelopes constrained to fall in Region III [Fig. 5(d)] contain non-relevant stochastic modulations. These stochastic envelope fluctuations have been found in the study by Dubbelboer and Houtgast (2007) to severely impair speech intelligibility. Hence, from Fig. 5 we can conclude that the constraints imposed on the enhanced envelopes decouple to some extent the speech-relevant modulations from the stochastic envelope fluctuations.

3. Procedure

The experiments were 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 485 circumaural headphones at a comfortable listening level. The listening level was controlled by each individual but was fixed throughout all the conditions in the test for a particular subject. Prior to the sentence test, each subject listened to a set of noise-corrupted sentences to get familiarized with the testing procedure. In the single-talker masker conditions, the listeners were informed of the masker sentence, since the masker was the same talker that was used to produce the target sentences [a similar approach was taken in the study by Hawley *et al.* (2004)]. Subjects were asked to pay attention to the non-masking sentence and write down all the words they heard. Twenty sentences were used per condition, and none of the lists were repeated across conditions. The order of the conditions was randomized across subjects. The whole listening test lasted for about 3–4 h. The testing session was split into two sessions each lasting 1.5–2 h. Five-minute breaks were given to the subjects every 30 min.

The listeners participated in a total of 36 conditions ($= 3$ SNR levels $\times$ 2 types of maskers $\times$ 6 processing conditions). The six processing conditions included speech processed using the square-root Wiener algorithm with (1) no constraints imposed, (2) Region I constraints, (3) Region II constraints, (4) Region I + II constraints, and (5) Region III constraints imposed. The sixth condition included the control condition, in which the noise-corrupted sentences were left unprocessed (UN).

B. Results and discussion

The mean performance, computed in terms of percentage of words identified correctly (all words were scored), by the normal-hearing listeners are plotted in Fig. 6 for the single-talker masker [Fig. 6 (top)] and the speech-shaped noise [Fig. 6 (bottom)] conditions. The intelligibility scores obtained in the two masker conditions were separately examined and analyzed for significant effects of SNR level and type of distortion introduced. For the scores obtained in the single-talker conditions, analysis of variance (with repeated measures) indicated a significant effect of type of distortion ($F_{5,30} = 364.0$, $p < 0.0005$), significant effect of SNR level ($F_{2,12} = 90.9$, $p < 0.0005$) and significant interaction ($F_{10,60} = 18.2$, $p < 0.0005$) between the type of distortion and SNR level. For the scores obtained in the SSN conditions, analysis of variance (with repeated measures) indicated a significant effect of type of distortion ($F_{5,30} = 686.9$, $p < 0.0005$), significant effect of SNR level ($F_{2,12} = 172.2$, p

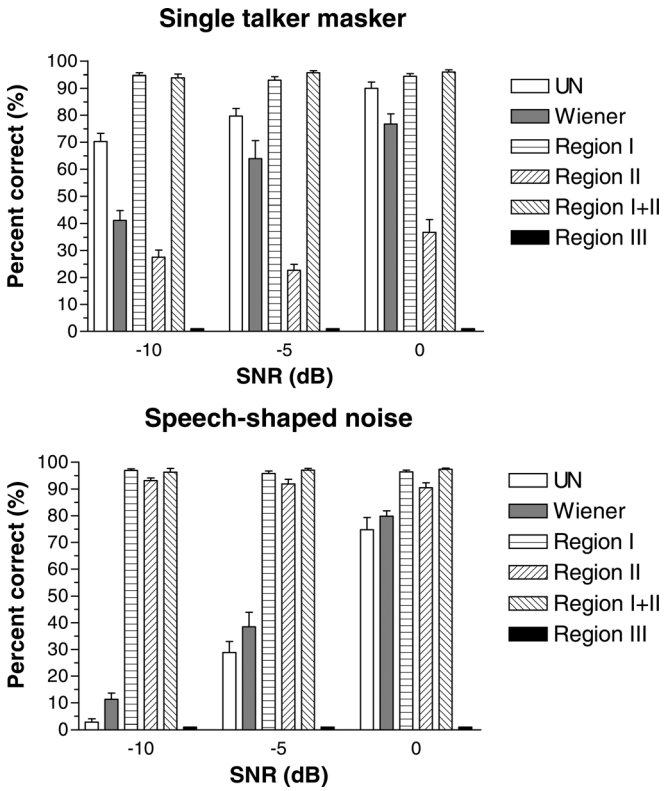

FIG. 6. Mean intelligibility scores as a function of SNR level, type of distortion and masker type. The bars labeled as "UN" show the scores obtained with noise-corrupted (unprocessed) stimuli, while the bars labeled as "Wiener" show the baseline scores obtained with the square-root Wiener algorithm (no constraints imposed). The intelligibility scores obtained with four different constraints imposed (following the square-root Wiener processing) are labeled accordingly. Error bars indicate standard errors of the mean.

<0.0005) and significant interaction ($F_{10,60} = 142.5$, $p < 0.0005$) between the type of distortion and SNR level.

As shown in Fig. 6, substantial improvements in intelligibility were obtained in nearly all conditions when the distortions were constrained to fall within Region I or Region I + II. The improvement, relative to UN and square-root Wiener-processed stimuli, was more evident in the SSN conditions. At −10 dB SNR (SSN masker), for instance, performance obtained with UN or square-root Wiener-processed sentences improved from 3% and 11% correct to nearly 100% correct when Region I constraints were imposed. Performance in Region III, in which amplification distortion in excess of 6 dB was introduced, was the lowest (near 0% correct) in all conditions and with both maskers. Performance in Region II, in which amplification distortion was limited to be lower than 6 dB, was poor (23%–37%) in the single-talker masker conditions but high ($>90\%$) in the SSN conditions.

Post hoc analysis, according to Fisher's LSD tests, was subsequently conducted to examine significant differences between conditions. For the single-talker conditions, performance with square-root Wiener-processed sentences was significantly lower ($p < 0.005$) than performance with unprocessed sentences (UN) at all three SNR levels. This was not surprising, as the computation of the square-root Wiener gain function [Eq. (6)] requires estimate of the competing talker spectrum [Eq. (7)], which is a challenging task. Performance in both Region I and Region I + II was found to be significantly higher than performance in UN conditions at the −10 and −5 dB SNR levels, but not ($p > 0.05$) at 0 dB SNR. Performance in Region II was significantly ($p < 0.005$) lower than performance in UN and square-root Wiener conditions at all SNR levels. A different pattern in results emerged in the SSN conditions. A small, but statistically significant ($p < 0.05$), improvement in intelligibility was noted at −10 and −5 dB SNR levels with the square-root Wiener-processed sentences relative to the scores obtained with unprocessed (UN) sentences. Large improvements ($p < 0.0005$) in performance, particularly at −10 and −5 dB SNR levels, were observed in the Region I, Region II, and Region I + II conditions relative to the UN and square-root Wiener conditions.

Of the two distortions introduced and examined, the attenuation distortion had the smallest effect on intelligibility. In fact, the data from the present study clearly demonstrate that substantial gains in intelligibility can be attained (see Fig. 6) when controlling and/or limiting the distortion introduced by noise suppression algorithms to be only of attenuation type. This was found to be true for both types of maskers tested. On the other hand, the impact of the amplification distortion on speech intelligibility varied across the two types of maskers tested. When the amplification distortion was limited to be smaller than 6 dB (Region II), performance was nearly unaffected in the SSN conditions, and in fact performance improved (relative to UN) and remained as high ($>90\%$) as that obtained in Region I. In contrast, performance dropped substantially (relative to UN) when the amplification distortion was limited to be smaller than 6 dB (Region II) in the single-talker conditions. When the amplification distortion was allowed to increase in excess of 6 dB, performance dropped to nearly 0% in all conditions and for both maskers. The reasons for that were not clear at first; hence, we analyzed the Region III condition further.

More precisely, we plotted the spectral SNRs for all frequency bins falling in Region III. Figure 7 shows the resulting SNR histograms computed using 20 IEEE sentences. For comparative purposes, we also plot the corresponding SNR histograms for all frequency bins falling in Region I + II. The large number of negative SNRs ($|X(k,t)| < |D(k,t)|$) in Region III suggests that the target was always masked. In fact, it can be proven analytically that Region III contains *only* masker-dominated T-F units. That is, the T-F units in Region III have *always* a negative SNR (see proof in the Appendix). This explains why performance in Region III was always near 0%. In contrast, the spectral SNR in Region I + II varied across a wide dynamic range, with nearly half of the distribution containing frequency bins with positive SNRs and the other half containing frequency bins with negative SNRs. The SNR histograms shown in Fig. 7 explain why performance in Region I + II was always higher than performance in Region III. Furthermore, we know that the SNR_{ESI} metric takes small values and is always smaller than 0 dB in Region III, while it assumes positive (≥ 0 dB) values in Region I + II. Consequently, by ensuring that the distortions remain in Region I + II we ensure that the SNR_{ESI}

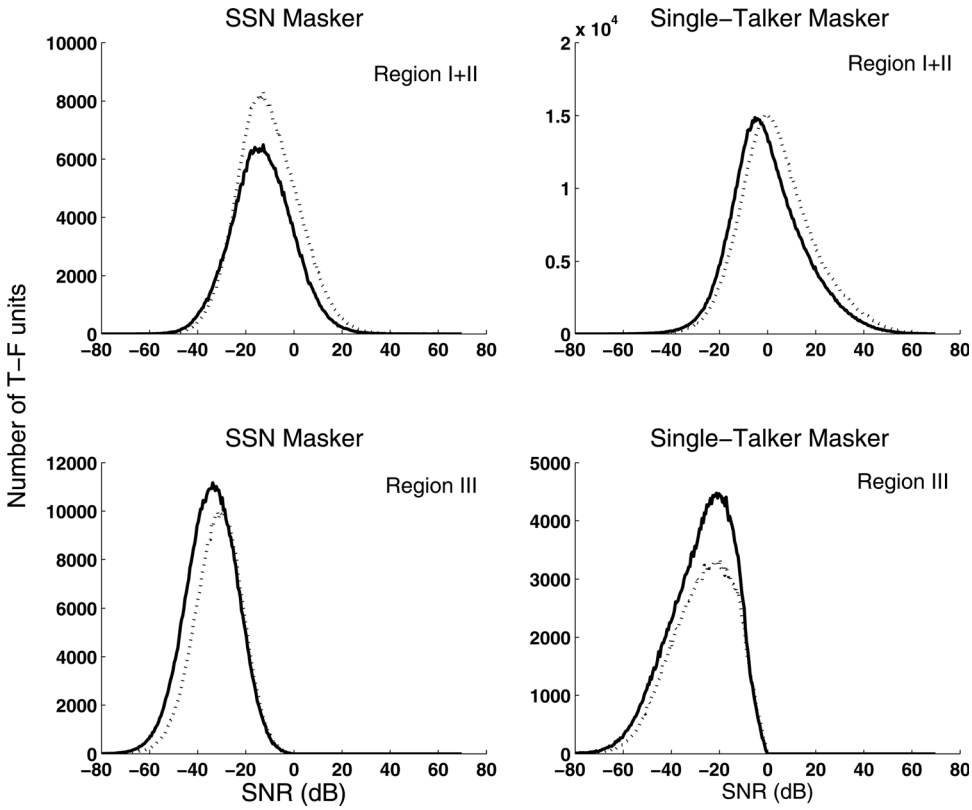

FIG. 7. Histogram of SNRs for T-F units falling in Regions (top) I + II and (bottom) III for two input SNR levels (dashed lines show input SNR = 0 dB and solid lines show input SNR = − 5 dB).

metric assumes values greater than 1, and as the present data demonstrated (Fig. 6), in doing so we can potentially maximize the intelligibility benefit.

In Region I + II, the amplification and attenuation distortions co-exist, as is often the case with distortions introduced by most (if not all) noise-reduction algorithms. However, the amplification distortion in Region I + II was limited to be lower than 6 dB (no limit was imposed on the attenuation distortion), yet large gains in intelligibility were obtained in all conditions. This suggests that one of the reasons that existing noise-reduction algorithms do not improve speech intelligibility is because they allow amplification distortions in excess

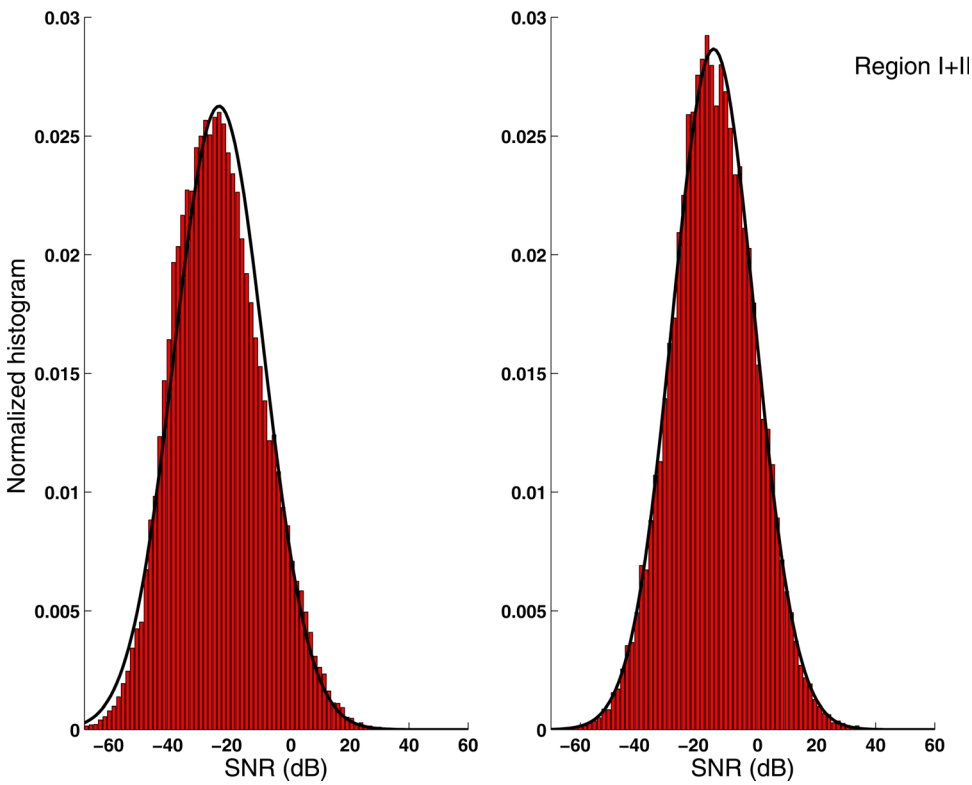

FIG. 8. (Color online) Histogram of SNRs (left) for T-F units in UN sentences and (right) for T-F units confined in Region I + II. The data were fitted with a Gaussian distribution (shown with solid lines).

of 6 dB. As shown in Fig. 7, amplification distortions in excess of 6 dB are associated with dominantly negative SNRs and subsequently with T-F units that are completely masked by background noise. Hence, by eliminating these distortions, we eliminate a large number of T-F units associated with extremely low SNRs. Consequently, we would expect that we could improve the overall SNR simply by eliminating amplification distortions in excess of 6 dB. To demonstrate this, we computed the histogram of the SNRs (computed prior to masking) of all T-F units falling in Region I + II and compared that against the corresponding SNR histogram of all T-F units of UN sentences. Figure 8 shows such a comparison for a sentence corrupted by SSN at − 5 dB SNR. The histograms were fitted to a Gaussian distribution (based on the maximum likelihood method), from which we extracted the mean of the distribution. As can be seen, the mean of the SNR distribution moved to the right (i.e., improved) from − 24 dB when all T-F units in UN sentences were included to − 14 dB when only T-F units falling in Region I + II were included. For this example, the effective SNR of Region I + II stimuli improved, on the average, by 10 dB. Hence, by simply eliminating amplification distortions in excess of 6 dB, we can improve the effective SNR of the noise-suppressed stimuli by as much as 10 dB, at least in steady background conditions.

According to Fig. 7, the signals in Region I + II contain T-F units with both positive and negative SNRs. Yet, the negative SNR T-F units did not seem to impair speech intelligibility (Fig. 6). The constraints imposed for Regions I and II provide no way of differentiating between positive and negative SNR T-F units, in terms of designing algorithms that would possibly eliminate the T-F units with negative SNRs. The constraints in Region III, however, guarantee that all T-F units falling in Region III will have negative SNR (see proof in the Appendix). Therefore, the constraints of Region III provide a mechanism which can be used by noise-reduction algorithms to eliminate low SNR T-F units and subsequently improve speech recognition. Introducing amplification distortions in excess of 6 dB is equivalent to introducing negative SNR T-F units in the processed signal, and should therefore be avoided or eliminated.

Performance in Region II was significantly higher when the masker was steady noise rather than a single-talker. There are several possible explanations for that. One possibility is that the estimation of the noise statistics needed in the square-root Wiener gain function was not done as accurately in single-talker conditions as in steady noise conditions. Estimating the noise statistics in competing-talker masking conditions is considerably more challenging than in steady-noise conditions, and this possibly influenced the number and frequency location of the T-F units falling in Region II.

Second, we considered the possibility that the number of T-F units falling in each of the three regions might explain the low performance in Region II. We thus calculated the percentage of bins falling in each Region and tabulated the percentages in Table I (these percentages represent mean values computed using 20 IEEE sentences). The percentage of T-F units falling in Region II for the single-talker (7%–8%) masker was smaller than that for the SSN masker (10%–14%). Although the difference does not seem to be large enough to fully explain the large difference in scores in Region II, the lower amount of T-F units in Region II caused a drop in intelligibility. However, for Region I and Region III which cover a much wider range of $|\hat{X}|/|X|$ (see Fig. 2) compared to Region II, no meaningful correlation or relationship was found between the percentage of T-F units falling in each region and intelligibility. A significantly larger percentage of T-F units fell in Region I in single-talker masker (64%–75%) conditions compared to SSN (21%–36%) conditions, yet the intelligibility scores obtained in both conditions were equally high. As proved above, T-F units in Region III have always negative SNR, and it is therefore not surprising that the number of Region-III units in single-talker masker conditions was significantly lower than those in SSN conditions.

Overall, attenuation distortions had a minimal effect on speech intelligibility and this was found to be clear and consistent for both maskers tested. In contrast, the effects of amplification distortions were more complex to interpret and seemed to be dependent on (a) the type of masker, (b) the amount of distortion present (for Region II it was < 6 dB and for Region III it was > 6 dB), and (c) whether these distortions co-existed with attenuation distortions (Region I + II). Despite the complexity in assessing the effects of these distortions in the various scenarios, it was clear from the present experiment that in the latter (c) scenario, when the amplification distortions were limited to be lower than 6 dB, while allowing for attenuation distortions (i.e., Region I + II), large gains in intelligibility can be obtained consistently for both maskers tested and all SNR levels.

TABLE I. Percentage of bins falling in the three regions.

	SNR	Region I	Region II	Region III
Single-talker masker	− 10 dB	64.32%	7.09%	28.59%
	− 5 dB	69.11%	8.15%	22.44%
	0 dB	74.64%	7.91%	17.45%
Speech-shaped noise	− 10 dB	20.57%	9.71%	69.72%
	− 5 dB	27.50%	11.99%	60.51%
	0 dB	36.05%	14.31%	49.64%

IV. EXPERIMENT 2: EFFECT OF AMPLIFICATION DISTORTION ALONE ON SPEECH INTELLIGIBILITY

Given the detrimental effects of amplification distortion on speech corrupted by a competing talker, we wanted to analyze it further by varying systematically the amount of distortion introduced by the gain functions. The previous experiment only examined two extreme cases in which the amplification distortion was either limited to be less than 6 dB (Region II or Region I + II) or greater than 6 dB (Region III). In the present experiment, amplification distortion is systematically varied here from a low of 2 dB to a high of 20 dB. Furthermore, unlike some of the stimuli used in the previous experiment, none of the stimuli used in the present experiment contain any attenuation distortions and this was done to assess the individual contribution of amplification distortion.

A. Methods

1. Subjects and material

Seven new normal-hearing listeners were recruited for this experiment. All subjects were native speakers of American English and were paid for their participation. The same sentence material (IEEE, 1969) was used as in Experiment 1.

2. Signal processing

To assess the impact of amplification distortion on speech intelligibility, we varied systematically the amount of distortion introduced. The corrupted signal was processed as described before (see Sec III A 2) by the square-root Wiener algorithm producing at time frame t and frequency band k the magnitude spectrum $|\hat{X}(k,t)|$. T-F units in cell (k, t) that satisfied the following constraint were retained, while the remaining were set to 0:

$$0 < 20 \cdot \log_{10} \frac{|\hat{X}(k,t)|}{|X(k,t)|} < A \ (\text{dB}), \tag{9}$$

where the positive constant A (expressed in dB) denotes the maximum amplification distortion allowed. Clearly, the smaller the value of A is, the smaller the number of T-F units retained. Note that when $0 < A \leq 6.02$ dB, the constrained region coincides with Region II, and when $A > 6.02$ dB, the constrained region includes Region II *and* part of Region III (see Fig. 2). Following the selection of T-F units according to Eq. (8), the signal was synthesized as in Experiment 1 (Sec. III A 2).

3. Procedure

Subjects participated in a total of 36 conditions (= 3 SNR levels × 2 types of maskers × 6 processing conditions). The two maskers were the same as in Experiment 1. Six processing conditions were tested corresponding to six different values of A: 2, 4, 6, 10, 15, and 20 dB. Two lists of sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions. The order of the test conditions was randomized across subjects.

B. Results and discussion

The mean performance, computed in terms of percentage of words identified correctly (all words were scored), by the normal-hearing listeners are plotted in Fig. 9 for the single-talker masker [Fig. 9 (top)] and speech-shaped noise [Fig. 9 (bottom)] conditions. The intelligibility scores obtained in the two masker conditions were separately examined and analyzed for significant effects of SNR level and amount of amplification distortion introduced. For the scores obtained in the single-talker conditions, analysis of variance (with repeated measures) indicated a significant effect of amount of amplification distortion ($F_{5,30} = 112.2$, $p < 0.0005$), significant effect of SNR level ($F_{2,12} = 30.8$, $p < 0.0005$) and significant interaction ($F_{10,60} = 19.6$, $p < 0.0005$) between amount of distortion and SNR level. For the scores obtained in the SSN conditions, analysis of variance (with repeated measures) indicated a significant effect of amount of amplification distortion ($F_{5,30} = 64.1$, $p < 0.0005$), significant effect of SNR level ($F_{2,12} = 14.8$, $p < 0.0005$) and significant interaction ($F_{10,60} = 12.2$, $p < 0.0005$) between amount of distortion and SNR level.

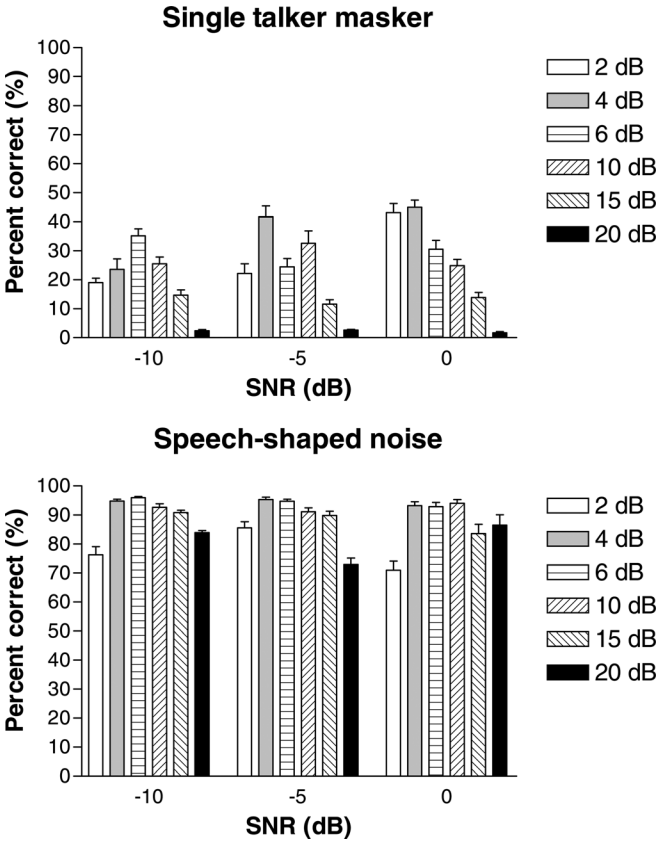

FIG. 9. Mean intelligibility scores as a function of SNR level, amount of amplification distortion and masker type. The maximum amplification distortion allowed ranged from 2 to 20 dB and is indicated accordingly. Error bars indicate standard errors of the mean.

It is clear from Fig. 9, that the amount of amplification distortion introduced affected the intelligibility of speech corrupted by the two types of maskers differently. The effect was small for speech corrupted by the SSN masker, but was quite large and significant for speech corrupted by the single-talker masker. When the constrained region coincides with Region II ($0 < A < 6.02$ dB), the lowest performance was obtained with $A = 2$ dB with the exception of one condition (0 dB single-talker). This is to be expected, since the smaller the value of A is, the smaller the number of T-F units retained and the sparser the signal is in the T-F domain. Intelligibility improved when $A = 4$ dB in nearly all conditions. *Post hoc* tests (Fisher's LSD) confirmed that the improvement, relative to $A = 2$ dB was statistically significant ($p < 0.05$). When $A \geq 6$ dB, intelligibility scores dropped significantly in the single-talker conditions, but remained high (> 80%) in the SSN conditions. It is interesting to note that in the SSN conditions, intelligibility scores remained modestly high (> 70%) at all SNR levels, even when $A = 20$ dB. It should be noted that the condition corresponding to $A = 20$ dB is not the same as the Region III condition in Experiment 1, wherein performance dropped to 0.

As shown in Fig. 2 [and Eq. (9)], the condition with $A = 20$ dB includes Region II along with part of Region III.

In summary, performance in single-talker conditions was quite susceptible to amplification distortion. Even a small amount of distortion (< 6 dB), was found to decrease performance by as much as 60 percentage points relative to the performance obtained with un-processed sentences (see Figs. 6 and 9). In contrast, no significant effect on intelligibility was observed in the SSN conditions. We attribute the differential effect of amplification distortion on two possibly interrelated reasons, as discussed previously in Sec. III B. One possibility is that the estimation of the noise statistics needed in the square-root Wiener gain function was not done as accurately in single-talker conditions as in steady noise conditions. Second, the number of T-F units falling in Region II for the single-talker conditions was smaller than the corresponding number in SSN conditions (see Table I). Subsequently, the synthesized signals in the single-talker conditions were "sparser" than the corresponding signals in SSN conditions.

At first glance, the findings from this experiment contradict those from Experiment 1. In Experiment 1, amplification distortions in excess of 6 dB (Region III) were found to be quite detrimental, while in the present experiment high intelligibility was maintained in the SSN conditions even when the amplification distortions were as large as 20 dB. The discrepancy is due to the fact that the regions examined in the two experiments are different. Experiment 1 examined Region III while Experiment 2 examined Region II plus part of Region III. Although Region II is a subset of the overall region examined, the effects of amplification distortion are complex to interpret for several reasons. First, the SNR distributions of T-F units falling in these two regions differ. Second, the number of T-F units falling in these two regions differs, and accordingly that affects the "sparsity" of the signal. Third, the accuracy in estimating the gain function in these regions also differs. We thus believe that all these factors contributed to the difference in outcomes between the two maskers.

V. EXPERIMENT 3: EFFECT OF GAIN-INDUCED DISTORTIONS ON VOWELS AND CONSONANTS

The weak consonants (e.g., stops) are masked by noise more easily and more heavily, than the high-energy vocalic segments (Parikh and Loizou, 2005; Phatak and Allen, 2007). Given that noise masks differently and to a different extent vowels and consonants, we examine in the present experiment, the impact of attenuation distortion introduced either in vowel-like segments or weak-consonant segments of the utterance. In a practical implementation of the constraints presented in Sec. II it is reasonable to expect that it would be easier to impose the constraints in voiced (e.g., vowels) rather than unvoiced (e.g., weak consonants such as stops and fricatives) segments as the former segments are easier to detect. This raises the question then as to whether we would expect to observe substantial improvements in intelligibility when the attenuation distortion is confined within the voiced segments (e.g., vowels) alone or unvoiced (e.g., stop consonants) segments alone. The present experiment is designed to answer this question.

A. Method

1. Subjects and material

Seven new normal-hearing listeners were recruited for this experiment. All subjects were native speakers of American English and were paid for their participation. The same speech material (IEEE, 1969) were used as in Experiment 1.

2. Signal processing

The IEEE sentences were phonetically transcribed into voiced or unvoiced segments using the method described in Li and Loizou (2008). Very briefly, a highly accurate F0 detector (Kawahara *et al.*, 1999) was first used to provide the initial classification of short- duration segments into voiced and unvoiced segments. The F0 detection algorithm was applied every 1 ms to the stimuli using a high-resolution fast Fourier transform (FFT) to provide for accurate temporal resolution of voiced/unvoiced boundaries. Segments with non-zero F0 values were initially classified as voiced and segments with zero F0 value were classified as unvoiced. After automatic classification, the voiced and unvoiced decisions were inspected for errors and the detected errors were manually corrected. The voiced/unvoiced segmentation of all IEEE sentences was saved in text files and is available from a CD ROM in Loizou (2007). Voiced segments included all sonorant sounds, i.e., the vowels, semivowels and nasals, while the unvoiced segments included all obstruent sounds, i.e., the stops, fricatives, and affricates.

The noise-corrupted sentences were first processed as in Experiment 1 via the square-root Wiener algorithm. The voiced/unvoiced segmentation of each sentence was retrieved from the corresponding saved text file and the square-root Wiener-processed speech spectrum was modified as per Eq. (8) to implement the Region I constraints. In one condition, the Region I constraints (allowing only attenuation distortion) were applied only to the voiced segments leaving the unvoiced segments unconstrained (but square-root Wiener processed). In another condition, the Region I constraints were applied only to the unvoiced segments leaving the voiced segments unconstrained. Following the modification of the square-root Wiener-processed spectrum as per Eq. (8), the voiced (or unvoiced) segments were synthesized using the same synthesis method described in Experiment 1.

3. Procedure

Subjects participated in a total of 24 conditions (= 3 SNR levels × 2 types of maskers × 4 processing conditions). The same two maskers were used as in Experiment 1. The four processing conditions included (1) noise-corrupted speech, (2) square-root Wiener-processed speech followed by Region I constraints applied to the whole utterance, (3) square-root Wiener-processed speech followed by Region I constraints applied only to the voiced segments (no constraints were applied to the unvoiced segments), and (4) square-root Wiener-processed speech followed by Region I

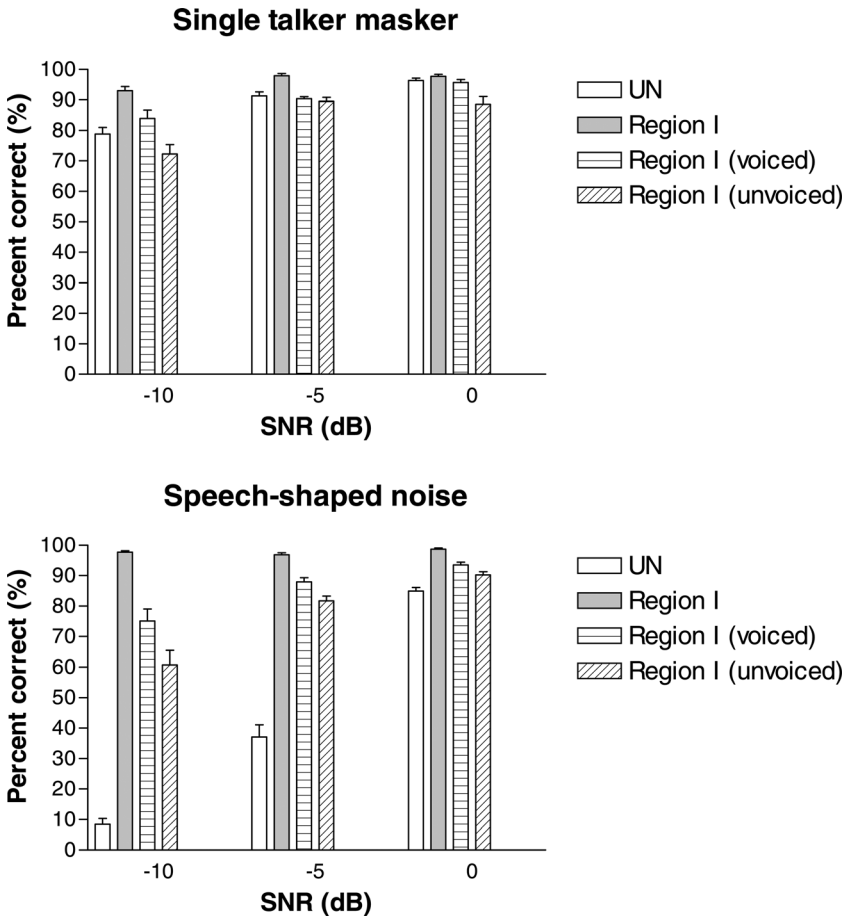

FIG. 10. Mean intelligibility scores as a function of SNR level, masker type and sound-class distortion. Attenuation distortion was limited to only voiced segments, unvoiced segments or both (Region I). Error bars indicate standard errors of the mean.

constraints applied only to the unvoiced segments (no constraints were applied to the voiced segments). Two lists of sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions. The order of the test conditions was randomized across subjects.

B. Results and discussion

The mean performance, computed in terms of percentage of words identified correctly (all words were scored), by the normal-hearing listeners are plotted in Fig. 10 for the single-talker masker [Fig. 10 (top)] and the speech-shaped noise [Fig. 10 (bottom)] conditions. The intelligibility scores obtained in the two masker conditions were separately examined and analyzed for significant effects of SNR level and sound class (voiced vs unvoiced) distortion on speech intelligibility. For the scores obtained in the single-talker conditions, analysis of variance (with repeated measures) indicated a significant effect of sound-class distortion ($F_{3,18} = 34.5$, $p < 0.0005$), significant effect of SNR level ($F_{2,12} = 56.3$, $p < 0.0005$) and significant interaction ($F_{6,36} = 101.2$, $p < 0.0005$) between sound-class distortion and SNR level. For the scores obtained in the SSN conditions, analysis of variance (with repeated measures) indicated a significant effect of sound-class distortion ($F_{3,18} = 374.0$, $p < 0.0005$), significant effect of SNR level ($F_{2,12} = 152.6$, $p < 0.0005$) and significant interaction ($F_{6,36} = 7.9$, $p < 0.0005$).

No dramatic decrease in performance (relative to the UN conditions) was noted in the single-talker conditions when the attenuation distortion was introduced in either the voiced or unvoiced segments of the utterances. Performance at -5 and 0 dB SNR, however, was limited by ceiling effects. At 0 dB SNR, performance in the unvoiced-segment conditions was significantly ($p < 0.005$) lower than performance in the control Region I condition, but nonetheless performance in the unvoiced-segment condition remained high at 88% correct.

The impact of attenuation distortion on voiced and unvoiced segments was more evident in the SSN conditions, particularly at -10 and -5 dB SNRs. Relative to the control UN condition, performance at -10 dB SNR improved from 10% (with UN) to over 70% when the attenuation distortion was introduced to the voiced segments alone. Similarly, performance improved from 10% to 60% when the attenuation distortion was introduced to the unvoiced segments alone. The same pattern was observed at -5 dB and 0 dB SNR. At 0 dB SNR, the improvement was small, but statistically significant ($p < 0.005$), according to Fisher's LSD tests conducted *post hoc*.

The most interesting outcome from the present experiment is that there was substantial benefit in intelligibility (relative to UN) in steady-noise conditions even when the attenuation distortion was limited to either voiced or unvoiced segments alone, with a larger benefit observed in the voiced-segment conditions. The large benefit observed in the unvoiced-segment conditions can be attributed to the listeners having better access to lexical segmentation cues (Stevens, 2002; Li and Loizou, 2008). These cues are critically important as they enable the listeners to better identify the

word/syllable boundaries, which are typically blurred in steady-noise backgrounds. The large benefit observed in the unvoiced-segment conditions is consistent with that observed in our prior study (Li and Loizou, 2008) wherein listeners were presented with clean unvoiced segments in otherwise corrupted sentences (no square-root Wiener processing was applied to the noise-corrupted sentences in that study).

The major contributions of the present experiment are the practical implications of applying the Region I constraints only to voiced segments. That is, signal processing algorithms can be devised that apply the proposed constraints only during voiced segments while leaving the unvoiced segments unaltered. The voiced segments are comparatively masked by noise to a lower degree than the weak consonants (e.g.,/f/). The idea is to first identify a particular speech segment as voiced or unvoiced and then apply the spectral constraints investigated in the present study only to voiced segments.

VI. DISCUSSION

Large benefits in intelligibility were obtained in Experiments 1–3 when the proposed constraints were imposed. Implementation of those constraints, however, required access to the clean envelopes (spectra), which we do not have. This raises the question: How can the proposed constraints be implemented in practice? There are two possibilities that can be explored.

First, given that the classification and retention of T-F units is binary, one can train a binary classifier to classify the mixture envelopes into two classes, those with $\mathrm{SNR}_{\mathrm{ESI}} \geq 0$ dB (belonging to Region I + II) and those with $\mathrm{SNR}_{\mathrm{ESI}} < 0$ dB (belonging to Region III). True knowledge of the target signal is only required during the training stage of the binary classifier, which can be done off-line using a large inventory of noise sources. In the testing stage, the binary classifier will have access only to the observed noisy signals and will make decisions based on features extracted from noisy observations. Based on the results from Fig. 6, T-F units with $\mathrm{SNR}_{\mathrm{ESI}} \geq 0$ dB should be retained and T-F units with $\mathrm{SNR}_{\mathrm{ESI}} < 0$ dB should be eliminated, as those are associated with SNR < 0 dB. These decisions will be made using the trained binary classifier. Such an approach was taken in Kim and Loizou (2010), wherein it was demonstrated (with normal-hearing listeners) that large gains in intelligibility can be obtained. The data from the latter study thus showed that the implementation of the proposed constraints is indeed feasible.

A second possibility is to limit or control the amount of amplification/attenuation distortions introduced by the gain functions by estimating the $\mathrm{SNR}_{\mathrm{ESI}}$ metric directly from the mixture envelopes. Equation (2) cannot be used for this purpose, as it depends on the clean envelopes, which we do not have. The $\mathrm{SNR}_{\mathrm{ESI}}$ metric at frequency band k and time frame t, however, can alternatively be expressed as:

$$\mathrm{SNR}_{\mathrm{ESI}}(k,t) = \frac{\mathrm{SNR}(k,t)}{[1-G(k,t)]^2 \cdot \mathrm{SNR}(k,t) + [G(k,t)]^2}, \qquad (10)$$

where $G(k,t)$ is the gain function applied to frequency band k and time frame t and $\mathrm{SNR}(k, t)$ is the true *a priori* SNR[3] at frequency band k and time frame t [see derivation in Lu and Loizou (2010)]. The advantage of using the above expression for $\mathrm{SNR}_{\mathrm{ESI}}(k, t)$ is that it is applicable to all gain functions, regardless of whether the gain function is determined by the modulation rate, modulation depth, and/or SNR. The SNR term in the above equation is not known in practice, but can nonetheless be estimated from the mixture envelopes. Several techniques exist for estimating the band SNR [see review in Loizou (2007, Chap. 7)] and one common approach is to use the "decision-directed" method (Ephraim and Malah, 1984). Hence, the $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ definition given above can be used to monitor and control the distortions introduced by the noise-suppressive gain functions used by most hearing aids (Chung, 2004). Clearly the success of this approach will depend on the accuracy in estimating the $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ values. The outcome of Experiment 3 (Fig. 10), however, suggests that the $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ value only needs to be computed accurately for the voiced segments (e.g., vowels) of speech, at least in steady noise conditions. These segments are known to be masked by steady additive noise to a lesser extent than the weak consonants (e.g., stops) (Parikh and Loizou, 2005).

One interesting observation about the above expression [Eq. (10)] of the $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ metric, is that when $G(k,t) = 1$ for all bands, then $\mathrm{SNR}_{\mathrm{ESI}}(k, t) = \mathrm{SNR}(k, t)$. That is, when no processing is applied to the mixture envelopes [i.e., $G(k,t) = 1$], the $\mathrm{SNR}_{\mathrm{ESI}}(k,t)$ metric returns the band SNR value. Subsequently, if $\mathrm{SNR}_{\mathrm{ESI}}(k,t) = \mathrm{SNR}(k,t)$, the fwSNRseg intelligibility measure [Eq. (1)] resembles to a simplified form of the articulation index.

As mentioned earlier, a number of noise-reduction algorithms have been designed to minimize speech distortion while limiting noise distortion (Ephraim and Trees, 1995; Hu and Loizou, 2004; Chen *et al.*, 2006). Those algorithms, however, made no distinction between attenuation vs amplification distortions, as these two distortions were lumped into one. As demonstrated in Experiment 1 (see Fig. 6), the two distortions should not be treated equally. Instead, Eq. (10) could be used to monitor and possibly eliminate excess amplification distortions, known to severely degrade intelligibility.

VII. CONCLUSIONS

Most noise-reduction algorithms used in hearing aids operate by first identifying the presence or absence of speech in the noisy signal (via modulation detection and/or SNR detection algorithms), and then applying a gain to reduce noise interference (Chung, 2004). In most cases, the gain is proportional to the SNR or modulation rate detected in each band. Since the gain applied is imperfect as it depends on the *estimated* SNR and/or modulation rate, the resulting output envelope will be either amplified or attenuated relative to the clean input envelope. The present study focused on assessing the impact of these gain-induced (amplification or attenuation) distortions on speech intelligibility in competing talker and steady noise conditions. It does not assess the impact of other unwanted nonlinear distortions introduced in hearing

aid instruments by compression or signal clipping (Arehart *et al.*, 2007). Noise-corrupted speech was processed via a conventional noise-reduction algorithm (square-root Wiener filtering) and the gain-induced distortions were confined into one of three regions: Region I containing only attenuation distortion, Region II containing only amplification distortion smaller than 6 dB, and Region III containing amplification distortion in excess of 6 dB. The following conclusions can be drawn from the present study.

(1) Of the two envelope distortions examined, the attenuation distortion (Region I) was found to have a minimal effect on speech intelligibility in both competing-talker and steady noise conditions. In fact, substantial improvements in intelligibility, relative to unprocessed (and noise-corrupted) speech, were obtained when the noise-suppressed speech contained only attenuation distortion. At -10 dB SNR (SSN masker), for instance, performance improved from 5%–10% correct with unprocessed (UN) or square-root Wiener-processed sentences to nearly 100% correct when Region I constraints (attenuation distortion) were imposed.
(2) In the absence of attenuation distortions, the impact of amplification distortions alone was complex and was found to be largely dependent on the masker and the region examined (Experiment 2).
(3) In situations wherein the attenuation and amplification distortions coexist (as is often the case with most noise-reduction algorithms), substantial gains in intelligibility can be obtained provided the amplification distortion is smaller than 6 dB (see data for Region I + II in Fig. 6). This was found to be true for both types of maskers and for all SNR conditions.
(4) Existing noise-reduction algorithms introduce both attenuation and amplification distortions, and the data from Experiment 1 suggest that one of the reasons that such algorithms do not improve speech intelligibility is because they allow amplification distortions in excess of 6 dB. It was proven empirically (see Fig. 7) and analytically (see the Appendix) that these distortions are always associated with masker-dominated (i.e., SNR <0 dB) T-F units. Therefore, the constraints of Region III provide a mechanism which can be used by noise-reduction algorithms to eliminate low SNR T-F units and subsequently improve speech recognition. Furthermore, analysis of the data in Region I + II (see Fig. 8) revealed that the effective SNR could increase by as much as 10 dB, at least in SSN conditions, when amplification distortions in excess of 6 dB are eliminated. Introducing amplification distortions in excess of 6 dB is *equivalent* to introducing negative SNR T-F units in the processed signal and should therefore be avoided or eliminated.
(5) The data from Experiment 3 indicated a substantial benefit in intelligibility (relative to UN) in steady-noise conditions even when the attenuation distortion was limited to either voiced (e.g., vowels) or unvoiced segments (e.g., stops) alone, with a larger benefit observed in the voiced-segment conditions. The large benefit observed in the unvoiced-segment conditions can be attributed to the listeners having better access to lexical segmentation cues (Stevens, 2002; Li and Loizou, 2008). The importance of Experiment 3 lies in its practical implications: signal processing algorithms can be developed that apply the proposed constraints only during the voiced segments, which are masked less by noise relative to the unvoiced segments.

Based on the above, we can conclude that in order for noise-suppression algorithms to improve speech intelligibility, it is critically important that certain distortions introduced by the gain functions are eliminated or at least properly controlled. Substantial gains in intelligibility *can* be obtained only if the gain-induced distortions are confined to be of attenuation type and amplification distortions in excess of 6 dB are eliminated, as these are associated with masker-dominated (SNR <0 dB) T-F units.

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC010494 from National Institute of Deafness and other Communication Disorders, NIH. This work was also supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0014143).

APPENDIX

In this appendix, we prove analytically that T-F units falling in Region III have always a negative SNR, that is, they are always dominated by the masker. This suggests that amplification distortion in excess of 6 dB is equivalent to a negative SNR. Recall that the Region III was defined as follows:

$$\begin{aligned} &\mathrm{SNR}_{\mathrm{ESI}}(k,t) < 1 \\ \Leftrightarrow\ &\left|\hat{X}(k,t)\right| > 2|X(k,t)|. \end{aligned} \tag{A1}$$

Given that the estimate of the target magnitude spectrum, $\left|\hat{X}(k,t)\right|$, is obtained by the square-root Wiener gain function $G(k,t)$ as per Eq. (5) and the gain function is always bounded between 0 and 1, we get the following inequalities:

$$G(k,t)|X(k,t)+D(k,t)| > 2|X(k,t)|, \tag{A2}$$

$$|X(k,t)+D(k,t)| > 2|X(k,t)|, \tag{A3}$$

$$|X(k,t)+D(k,t)|^2 > 4|X(k,t)|^2, \tag{A4}$$

$$|X(k,t)|^2+2\mathrm{Re}\{X(k,t)D^*(k,t)\}+|D(k,t)|^2 > 4|X(k,t)|^2, \tag{A5}$$

where Re $\{\cdot\}$ denotes the real part of a complex number and * denotes the complex conjugate. Since $\mathrm{Re}\{X(k,t)D^*(k,t)\} \le |X(k,t)||D(k,t)|$, we get the following inequality:

$$(3|X(k,t)|+|D(k,t)|)(|X(k,t)|-|D(k,t)|) < 0, \tag{A6}$$

which is satisfied when

$$\frac{|X(k,t)|}{|D(k,t)|} < 1. \quad \text{(A7)}$$

Based on the above equation, we conclude that T-F units in Region III always have a negative SNR. This was also verified empirically in Fig. 7.

[1]The number of words in a sentence varies from 5 to 10. The IEEE corpus contains phonetically balanced sentences and is organized into lists of 10 sentences each. All sentence lists were designed to be equally intelligible, thereby allowing us to assess speech intelligibility in different conditions without being concerned that a particular list is more intelligible than another.

[2]Other gain functions were investigated in Loizou and Kim (2011), but the choice of the gain function played a minor role when the distortions were properly controlled.

[3]The *a priori* SNR is defined mathematically as $SNR = E(|X|^2)/E(|D|^2)$, where $E(\cdot)$ denotes the expectation operator.

Alcantara, J., Moore, B., Kuhnel, V., and Launer, S. (**2003**). "Evaluation of the noise reduction system in a commercial digital hearing aid," Int. J. Audiol. **42**, 34–42.

Arehart, K., Kates, J., Anderson, M., and Harvey, L. (**2007**). "Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 1150–1164.

Baer, T., Moore, B., and Gatehouse, S. (**1993**). "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality and response times," J. Rehab. Res. Dev. **30**, 49–72.

Bentler, R., and Chiou, L. (**2006**). "Digital noise reduction: An overview," Trends Ampl. **10**, 67–82.

Bentler, R., Wu, Y., Kettel, J., and Hurtig, R. (**2008**). "Digital noise reduction: Outcomes from laboratory and field studies," Int. J. Audiol. **47**, 447–460.

Chen, J., Benesty, J., Huang, Y., and Doclo, S. (**2006**). "New insights into the noise reduction Wiener filter," IEEE Trans. Audio Speech Lang. Process. **14**, 1218–1234.

Chung, K. (**2004**). "Challenges and recent developments in hearing aids: Part, I. Speech understanding in noise, microphone technologies and noise reduction algorithms," Trends Amplif. **8**, 83–124.

Dillon, H., and Lovegrove, R. (**1993**). "Single-microphone noise reduction systems for hearing aids: A review and an evaluation," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. Studebaker and R. Hochberg (Allyn and Bacon, Needham Heights, MA), pp. 353–372.

Dubbelboer, F., and Houtgast, T. (**2007**). "A detailed study on the effects of noise on speech intelligibility," J. Acoust. Soc. Am. **122**, 2865–2871.

Edwards, B. (**2004**). "Hearing aids and hearing impairment," in *Speech Processing in the Auditory System*, edited by S. Greenberg, W. Ainsworth, A. Popper, and R. Fay (Springer Verlag, New York), pp. 339–421.

Ephraim, Y. (**1992**) "Statistical-model-based speech enhancement systems," Proc. IEEE **80**, 1526–1555.

Ephraim, Y., and 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., and Trees, H. V. (**1995**). "A signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Process. **3**, 251–266.

Graupe, D., Grosspietsch, J., and Basseas, S. (**1987**). "A single-microphone-based self adaptive filter of noise from speech and its performance evaluation," J. Rehabil. Res. Dev. **24**, 119–126.

Hawley, M., Litovsky, R., and Culling, J. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.

Hu, Y., and Loizou, P. (**2004**). "Incorporating a psychoacoustic model in frequency domain speech enhancement," IEEE Signal Process. Lett. **11**, 270–273.

Hu, Y., and Loizou, P. C. (**2007a**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.

Hu, Y., and Loizou, P. C. (**2007b**). "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun. **49**, 588–601.

IEEE. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoustics **AU-17**, 225–246.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **27**, 187–207.

Kim, G., and Loizou, P. C. (**2010**). "Improving speech intelligibility in noise using a binary mask that is based on magnitude spectrum constraints," IEEE Signal Process. Lett. **17**, 1010–1013.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.

Kuk, F., Ludvigsen, C., and Paludan-Muller, C. (**2002**). "Improving hearing aid performance in noise: Challenges and strategies," Hear. J. **55**, 34–46.

Latzel, M., Kiessling, J., and Margolf-Hackl, S. (**2003**). "Optimizing noise suppression and comfort in hearing instruments," Hear. Rev. **10**, 76–82.

Levitt, H. (**1997**). "Digital hearing aids: Past, present and future," in *Practical Hearing Aid Selection and Fitting*, edited by H. Tobin (Department of Veteran Affairs, Washington, DC), pp. xi–xxiii.

Levitt, H., Bakke, M., Kates, J., Neuman, A., Schwander, T., and Weiss, M. (**1993**). "Signal processing for hearing impairment," Scand. Audiol. **38**, 7–19.

Li, N., and Loizou, P. C. (**2008**). "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," J. Acoust. Soc. Am. **124**, 3947–3958.

Lim, J. S. (**1978**). "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," IEEE Trans. Acoust. Speech Signal Process. **26**, 471–472.

Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL), pp. 541–580.

Loizou, P. C., and Kim, G. (**2011**). "Reasons why speech enhancement algorithms do not improve speech intelligibility and suggested solutions," IEEE Trans. Audio, Speech, Lang. Process. **19**, 47–56.

Lu, Y., and Loizou, P. (**2010**). "Speech enhancement by combining statistical estimators of speech and noise," in *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing*, pp. 4754–4757.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., Dillier, N., Houben, R., Dreschler, W. A., Froehlich, M., Puder, H., Grimm, G., Hohmann, V., Leijon, A., Lombard, A., Mauler, D., and Spriet, A. (**2010**). "Multicenter evaluation of signal enhancement algorithms for hearing aids," J. Acoust. Soc. Am. **123**, 1491–1505.

Ma, J., Hu, Y., and Loizou, P. C. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**, 3387–3405.

Neuman, A., and Schwander, T. (**1987**). "The effect of filtering on the intelligibility and quality of speech in noise," J. Rehab. Res. Dev. **24**, 127–134.

Noordhoek, I., and Drullman, R. (**1997**). "Effect of reducing temporal intensity modulations on sentence intelligibility," J. Acoust. Soc. Am. **101**, 498–502.

Palmer, C., Bentler, R., and Mueller, H. (**2006**). "Amplification with digital noise reduction and the perception of annoying and aversive sounds," Trends Amplif. **10**, 95–104.

Parikh, G., and Loizou, P. (**2005**). "The influence of noise on vowel and consonant cues," J. Acoust. Soc. Am. **118**, 3874–3888.

Phatak, S., and Allen, J. (**2007**). "Consonants and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326.

Rangachari, S., and Loizou, P. C. (**2006**). "A noise-estimation algorithm for highly non-stationary environments," Speech Commun. **48**, 220–231.

Scalart, P., and Filho, J. (**1996**). "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing*, pp. 629–632.

Schum, D. (**2003**). "Noise-reduction circuitry in hearing aids: Goals and current strategies," Hear J. **56**, 32–40.

Stelmachowicz, P., Lewis, D., Choi, S., and Hoover, B. (**2007**). "Effect of stimulus bandwidth on auditory skills in normal-hearing and hearing-impaired children," Ear Hear. **28**, 483–494.

Stevens, K. (**2002**). "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891.

Tan, C.-T., and Moore, B. (**2008**). "Perception of nonlinear distortion by hearing-impaired people," Int. J. Audiol. **47**, 246–256.

Tyler, R. S., and Kuk, F. K. (**1989**). "The effects of "noise suppression" hearing aids on consonant recognition in speech-babble and low-frequency noise," Ear Hear. **10**, 243–249.
van Tasell, D., and Crain, T. (**1992**). "Noise reduction hearing aids: Release from masking and release from distortion," Ear Hear. 13, 114–121.
Weiss, M., and Neuman, A. (**1993**). "Noise reduction in hearing aids," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. Studebaker and R. Hochberg (Allyn and Bacon, Needham Heights, MA), pp. 337–352.
Wiener, N. (**1949**). *Interpolation, Extrapolation and Smoothing of Stationary Time Series* (John Wiley and Sons, New York), pp. 1–163.