Published in final edited form as:
*IEEE Trans Audio Speech Lang Processing*. 2011 ; 19(1): 47–56. doi:10.1109/TASL.2010.2045180.

# Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions

**Philipos C. Loizou, IEEE[Senior Member]** and **Gibak Kim**
The authors are with the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75083-0688 USA (loizou@utdallas.edu)

## Abstract

Existing speech enhancement algorithms can improve speech quality but not speech intelligibility, and the reasons for that are unclear. In the present paper, we present a theoretical framework that can be used to analyze potential factors that can influence the intelligibility of processed speech. More specifically, this framework focuses on the fine-grain analysis of the distortions introduced by speech enhancement algorithms. It is hypothesized that if these distortions are properly controlled, then large gains in intelligibility can be achieved. To test this hypothesis, intelligibility tests are conducted with human listeners in which we present processed speech with controlled speech distortions. The aim of these tests is to assess the perceptual effect of the various distortions that can be introduced by speech enhancement algorithms on speech intelligibility. Results with three different enhancement algorithms indicated that certain distortions are more detrimental to speech intelligibility degradation than others. When these distortions were properly controlled, however, large gains in intelligibility were obtained by human listeners, even by spectral-subtractive algorithms which are known to degrade speech quality and intelligibility.

### Index Terms

speech intelligibility improvement; speech distortions; speech enhancement; ideal binary mask

## I. Introduction

Much progress has been made in the development of speech enhancement algorithms capable of improving speech quality [1,2]. In stark contrast, little progress has been made in designing algorithms that can improve speech intelligibility. The first intelligibility study done by Lim [3] in the late 1970s found no intelligibility improvement with the spectral subtraction algorithm for speech corrupted in white noise at −5 to 5 dB SNR. In the intelligibility study by Hu and Loizou [4], conducted 30 years later, none of the 8 different algorithms examined 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, but improved ease of listening and listening comfort [5] for hearing-impaired listeners. In brief, the ultimate goal of devising an algorithm that would improve speech intelligibility for normal-hearing or hearing-impaired listeners has been elusive for nearly three decades.

Little is known as to why speech enhancement algorithms, even the most sophisticated ones, do not improve speech intelligibility. Clearly, one reason is the fact that we often do not have a good estimate of the background noise spectrum, which is needed for the implementation of most algorithms. For that, accurate voice-activity detection algorithms are required. Much progress has been made in the design of voice-activity detection algorithms and noise-estimation algorithms (see review in [1, Ch. 9]), some of which (e.g., [6]) are capable of continuously tracking, at least, the mean of the noise spectrum. Noise-estimation algorithms are known to perform well in stationary background noise (e.g., car) environments. Evidence of this was provided by Hu and Loizou [4] wherein a small improvement (<10 %) in intelligibility was observed with speech processed in car environments, but not in other environments (e.g., babble). We believe that the small improvement was attributed to the stationarity of the car noise, which allowed for accurate noise estimation This suggests that accurate noise estimation can contribute to improvement in intelligibility, but that alone cannot provide substantial improvements in intelligibility, since in practice we will never be able to track accurately the spectrum of non-stationary noise. For that reason, we believe that the absence of intelligibility improvement with existing speech enhancement algorithms is not entirely due to the lack of accurate estimates of the noise spectrum.

In the present paper, we discuss other factors that are responsible for the absence of intelligibility improvement with existing algorithms. The majority of these factors center around the fact that none of the existing algorithms are designed to improve speech intelligibility, as they utilize a cost function that does not necessarily correlate with speech intelligibility. The statistical-model based algorithms (e.g. MMSE, Wiener filter), for instance, derive the magnitude spectra by minimizing the mean-squared error (MSE) between the clean and estimated (magnitude or power) spectra (e.g., [7]). The MSE metric, however, pays no attention to positive or negative differences between the clean and estimated spectra. A positive difference between the clean and estimated spectra would signify attenuation distortion, while a negative spectral difference would signify amplification distortion. The perceptual effect of these two distortions on speech intelligibility cannot be assumed to be equivalent. The subspace techniques (e.g., [8]) were designed to minimize a mathematically-derived speech distortion measure, but make no attempt to differentiate between the two aforementioned distortions. In this paper, we will show analytically that if we can somehow manage or control these two types of distortions, then we should expect to receive large gains in intelligibility. To further support our hypothesis, intelligibility listening tests are conducted with normal-hearing listeners.

## II. Imposing constraints on the estimated magnitude spectra

To gain a better understanding on the impact of the two distortions on speech intelligibility, we use an objective function that has been found to correlate highly ($r = 0.81$) with speech intelligibility [9]. This measure is the frequency-domain version of the well-known segmental SNR measure. The time-domain segmental (and overall) SNR measure has been used widely and frequently for evaluating speech quality in speech coding and enhancement applications [10,11]. Results reported in the previous studies [9,12], however, demonstrated that the time-domain SNR measure does not correlate highly with either quality or speech intelligibility. In contrast, the frequency domain version of the segmental SNR measure [13] has been shown to correlate highly with both speech quality and speech intelligibility. In the present study, we refer to this measure as the signal-to-residual spectrum measure, $SNR_{ESI}$ (defined below). The correlation of the $SNR_{ESI}$ measure with speech intelligibility was found to be 0.81 [9] and the correlation with speech quality was found to be 0.85 [12]. The two main advantages in computing the $SNR_{ESI}$ measure in the frequency domain include: (1) the use of critical-band frequency spacing for proper modeling of the frequency selectivity of normal-hearing listeners, (2) the use of perceptually motivated weighting functions which can be applied to individual bands [9]. The use of signal-dependent weighting functions in the computation of the $SNR_{ESI}$ measure was found to be particularly necessary for predicting the intelligibility of speech corrupted by (fluctuating) non-stationary noise [9]. We thus believe that it is the combination of these two attractive features in the computation of the $\mathrm{SNR}_{\mathrm{ESI}}$ measure that contributes to its high correlation with speech intelligibility.

Let $SNR_{\mathrm{ESI}}(k)$ denote the signal-to-residual spectrum ratio at frequency bin $k$:

$$SNR_{\mathrm{ESI}}(k)=\frac{X^2(k)}{\left(X(k)-\widehat{X}(k)\right)^2} \tag{1}$$

where $X(k)$ denotes the clean magnitude spectrum and $\hat{X}(k)$ denotes the magnitude spectrum *estimated* by a speech-enhancement algorithm. Dividing both numerator and denominator by $D^2(k)$, where $D(k)$, denotes the noise magnitude spectrum, we get:

$$SNR_{\mathrm{ESI}}(k)=\frac{SNR(k)}{\left(\sqrt{SNR(k)}-\sqrt{SNR_{\mathrm{ENH}}(k)}\right)^2} \tag{2}$$

where $SNR(k) \triangleq X^2(k)/D^2(k)$ is the true *a priori* SNR at bin $k$, and $SNR_{\mathrm{ENH}}(k) \triangleq \hat{X}^2(k)/D^2(k)$ is the enhanced SNR1. Figure 1 plots $SNR_{\mathrm{ESI}}(k)$ as a function of $SNR_{\mathrm{ENH}}(k)$, for fixed values of SNR. The singularity in the function stems from the fact that when $SNR(k) = SNR_{\mathrm{ENH}}(k)$, $SNR_{\mathrm{ESI}}(k) = \infty$. Figure 1 provides important insights about the contributions of the two distortions to $SNR_{\mathrm{ESI}}(k)$, and for convenience, we divide the figure into multiple regions according to the distortions introduced:

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

**Region II.** In this region, $X(k) < \hat{X}(k) \le 2 \cdot X(k)$ suggesting amplification distortion up to 6.02 dB.

**Region III.** In this region, $\hat{X}(k) > 2 \cdot X(k)$ suggesting amplification distortion of 6.02 dB or greater.

From the above, we can deduce that in the union of Regions I and II, which we denote as Region I+II, we have the following constraint:

$$\widehat{X}(k) \le 2 \cdot X(k). \tag{3}$$

The constraint in Region I stems from the fact that in this region, $SNR_{\mathrm{ENH}} \le SNR$ leading to $\hat{X}(k) \le X(k)$. The constraint in Region II stems from the fact that in this region $SNR \le SNR_{\mathrm{ENH}} \le SNR + 6.02$ dB. Finally, the condition in Region III stems from the fact that in this region $SNR_{\mathrm{ENH}} \ge SNR + 6.02$ dB. It is clear from the above definitions of these three regions that in order to maximize $SNR_{\mathrm{ESI}}$ (and consequently maximize speech intelligibility), the estimated magnitude spectra $\hat{X}(k)$ need to be contained in regions I and II (note that the trivial, but not useful, solution that maximizes $SNR_{\mathrm{ESI}}$ is $\hat{X}(k) = X(k)$). Intelligibility listening tests were conducted to test this hypothesis. If the hypothesis holds, then we expect to see large improvements in intelligibility.

It is reasonable to ask how often the above distortions occur when corrupted speech is processed by conventional speech-enhancement algorithms. To answer this question, we tabulate in Table I the frequency of occurrences of the two distortions for speech processed

[1] Note that the defined enhanced $SNR_{\mathrm{ENH}}$ is not the same as the output SNR, since the background noise is not processed separately by the enhancement algorithm.

by three different (but commonly used) algorithms at two different SNR levels. Table I provides the average percentage of frequency bins falling in each of the three regions. To compute, for instance, the percentage of bins falling in Region I we counted the number of bins satisfying the constraint in Region I, and divided that by the total number of frequency bins, as determined by the size of the discrete Fourier transform (DFT). This was done at each frame after processing corrupted speech with an enhancement algorithm, and averaging the percentages over all frames in a sentence. As can be seen, nearly half of the bins fall in Region I which is characterized by attenuation distortion, while the other half of the bins fall in Region III, which is characterized by amplification distortion in excess of 6.02 dB. A small percentage (12–18 %) of bins was found to fall in Region II which is characterized by low amplification distortion, less than 6.02 dB. The perceptual consequences of the two distortions on speech intelligibility are not clear. For one, it is not clear which of the two distortions has the most detrimental effect on speech intelligibility. Listening tests are conducted to provide answers to these questions, and these tests are described next.

## III Intelligibility listening tests

### A. Algorithms tested

The noise-corrupted sentences were processed by three different speech enhancement algorithms that included the Wiener algorithm based on *a priori* SNR estimation [14] and two spectral-subtractive algorithms based on reduced delay convolution [15]. The sentences were segmented into overlapping segments of 160 samples (20 ms) with 50% overlap. Each segment was Hann windowed and transformed using a 160-point Discrete Fourier Transform (DFT). Let $Y(k,t)$ denote the magnitude of the noisy spectrum at time frame $t$ and frequency bin $k$. Then, the estimate of the signal spectrum magnitude is obtained by multiplying $Y(k,t)$ with a gain function $G(k,t)$ as follows:

$$\widehat{X}(k,t)=G(k,t)\cdot Y(k,t). \tag{4}$$

Three different gain functions were considered in the present study. The Wiener gain function is based on the *a priori* SNR and is given by:

$$G_{\text{Wiener}}(k,t)=\sqrt{\frac{SNR_{prio}(k,t)}{1+SNR_{prio}(k,t)}} \tag{5}$$

where $SNR_{prio}$ is the *a priori* SNR estimated using the decision-directed approach as follows:

$$SNR_{prio}(k,t)=\alpha\cdot\frac{\widehat{X}^2(k,t-1)}{\widehat{P}_D^2(k,t-1)}+(1-\alpha)\cdot\max\left[\frac{Y^2(k,t)}{\widehat{P}_D^2(k,t)}-1,0\right] \tag{6}$$

where $\widehat{P}_D^2(k,t)$ is the estimate of the power spectral density of background noise and α is a smoothing constant (typically set to α = 0.98). The spectral subtractive algorithms are based on reduced delay convolution [15] and the gain functions for magnitude subtraction and power subtraction are given respectively by:

$$G_{\text{RDC\_mag}}(k,t)=\max\left[0,1-\beta/\sqrt{SNR_{post}(k,t)}\right] \tag{7}$$

$$G_{\text{RDC\_pow}}(k,t)=\sqrt{\max\left[0,1-\beta/SNR_{post}(k,t)\right]} \tag{8}$$

where $SNR_{post}(k,t)\triangleq\frac{\widehat{P}_Y^2(k,t)}{\widehat{P}_D^2(k,t)}$ denotes the *a posterior* SNR and $\widehat{P}_Y^2(k,t)$ denotes the estimate of the noisy speech power spectral density computed using the periodogram method, and β is the subtraction factor which was set to β = 0.7 as per [15]. We denote the magnitude spectral-subtraction algorithm as RDC_mag and the power spectral-subtraction algorithm as RDC_pow.

The three gain functions examined are plotted in Figure 2. As can be seen, the three algorithms differ in the shape of their gain functions. The Wiener gain function is the least aggressive, in terms of suppression, providing small attenuation even at extremely low SNR levels, while the RDC_mag algorithm is the most aggressive eliminating spectral components at extremely low SNR levels. The three gain functions span a wide range of suppression options, which is one of the reasons for selecting them. We will thus be able to test our hypotheses about the effect of the constraints on speech intelligibility with algorithms encompassing a wide range of suppression varying from aggressive to least aggressive. The RDC_mag algorithm, in particular, was chosen because it performed poorly in terms of speech intelligibility [4]. The intelligibility of speech processed by the RDC_mag algorithm was found in several noisy conditions to be significantly lower than that obtained with unprocessed (noise corrupted) speech. We will thus examine whether it is possible to obtain improvement in intelligibility with the proposed constraints, even in scenarios where the enhancement algorithm (e.g., RDC_mag) is known to perform poorly relative to unprocessed speech.

Oracle experiments were run in order to assess the full potential on speech intelligibility when the proposed constraints are implemented. We thus assumed knowledge of the magnitude spectrum of the clean speech signal. The various constraints were implemented as follows. The noisy speech signal was first segmented into 20 ms frames (with 50% overlap between frames), and then processed through one of the 3 enhancement algorithms, producing at each frame the estimated magnitude spectrum $\hat{X}(k)$. The noise estimation algorithm proposed by Rangachari and Loizou [16] was used for estimating the noise spectrum in Eq. (6)–(8). The estimated magnitude spectrum $\hat{X}(k)$ was compared against the true spectrum $X(k)$, and spectrum components satisfying the constraint were retained, while spectral components violating the constraints were zeroed-out. For the implementation of the Region I constraint, for instance, the modified magnitude spectrum, $X_M(k)$, was computed as follows:

$$X_M(k)=\begin{cases}\widehat{X}(k) & \text{if } \widehat{X}(k)<X(k)\\ 0 & \text{else}\end{cases} \tag{9}$$

An inverse discrete Fourier transform (IDFT) was finally taken of $X_M(k)$ (using the noisy speech signal's phase spectrum) to reconstruct the time-domain signal. The overlap-and-add technique was subsequently used to synthesize the signal. As shown in Eq.(9), the constraints are implemented by applying a binary mask to the *estimated* magnitude spectrum (more on this later).

Figure 3 shows example spectrograms of signals synthesized using the Region I constraints (panel d). The original signal was corrupted with babble at −5 dB SNR. The Wiener algorithm was used in this example, and speech processed with the Wiener algorithm is shown in panel c. As can be seen in panel d, the signal processed using the Region I constraints resembles the clean signal, with most of the residual noise removed and the consonant onsets /offsets made clearer.

### B. Methods and Procedure

Seven normal-hearing listeners participated in the listening experiments, and all listeners were paid for their participation. The listeners participated in a total of 32 conditions (= 2 SNR levels (−5 dB, 0 dB) × 16 (=3×5+1) processing conditions). For each SNR level, the processing conditions included speech processed using three different speech enhancement (SE) algorithms with (1) no constraints imposed, (2) Region I constraints, (3) Region II constraints, (4) Region I+II constraints, and (5) Region III constraints. For comparative purposes, subjects were also presented with noise-corrupted (unprocessed) stimuli.

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. Prior to the sentence 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 [17] and used for each condition, with none of the lists repeated across conditions. The order of the conditions was randomized across subjects. The testing session lasted for about 2 hrs. Five-minute breaks were given to the subjects every 30 minutes.

Sentences taken from the IEEE database [17] were used for test material2. The sentences in the IEEE database are phonetically balanced with relatively low word-context predictability. The sentences were originally recorded at a sampling rate of 25 kHz and downsampled to 8 kHz (the recordings are available from a CD accompanying the book in [1]). Noisy speech was generated by adding babble noise at 0 dB and −5 dB SNR. The babble noise was produced by 20 talkers with equal number of female and male talkers. To simulate the receiving frequency characteristics of telephone handsets, the speech and noise signals were filtered by the modified intermediate reference system (IRS) filters used in ITU-T P.862 [19]. Telephone speech was used as it is considered particularly challenging (in terms of intelligibility) owing to its limited bandwidth (300–3200 Hz). Consequently, we did not expect the performance to be limited by ceiling effects.

### C. Results

Figure 4 shows the results of the listening tests expressed in terms of the percentage of words identified correctly by normal-hearing listeners. The bars indicated as “UN” show the scores obtained with noise-corrupted (un-processed) stimuli, while the bars indicated as “SE” show the baseline scores obtained with the three enhancement algorithms (no constraints imposed). As shown in Figure 4, performance improved dramatically when the Region I constraints were imposed. Consistent improvement in intelligibility was obtained with the Region I constraints for all three speech-enhancement algorithms examined.

[2] A sentence recognition test was chosen over a diagnostic rhyme test [18] for assessment of intelligibility for several reasons. Sentence tests: (1) better reflect real-world communicative situations, (2) are open-set tests, and as such scores may vary from a low of 0% correct to 100% correct (in contrast, the DRT test is a closed-set test, has a chance score of 50% and needs to be corrected for chance). The sentence materials (IEEE corpus) chosen contain contextual information, however, that information is controlled by design. 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.

Performance at −5 dB SNR with the Wiener algorithm, for instance, improved from 10% correct when no constraints were imposed, to 90% correct when Region I constraints were imposed. Substantial improvements in intelligibility were also noted for the two spectral-subtractive algorithms examined. Performance with Region II constraints seemed to be dependent on the speech-enhancement algorithm used, with good performance obtained with the Wiener algorithm, and poor performance obtained with the two spectral-subtractive algorithms. Large improvements in intelligibility were obtained with Region I+II constraints for all three algorithms tested and for both SNR levels. Finally, performance degraded to near zero when Region III constraints were imposed for all three algorithms tested and for both SNR levels.

Statistical tests, based on Fisher's LSD test, were run to assess significant differences between the scores obtained in the various constraint conditions. Performance of the Wiener algorithm with Region I constraints did not differ statistically ($p>0.05$) from performance obtained with the Region I+II constraints. Similarly, performance of the RDC_pow algorithm with Region I constraints did not differ statistically ($p>0.05$) from performance obtained with the Region I+II constraints. This was found to be true for both SNR levels and for both Wiener and RDC_pow algorithms. In contrast, performance obtained with the RDC_mag algorithm with Region I constraints was significantly higher ($p<0.05$) than performance obtained with Region I+II constraints. Performance obtained with the Wiener algorithm (with no constraints) did not differ significantly ($p>0.05$) from performance obtained with unprocessed (noise corrupted) sentences for both SNR levels tested. Performance obtained at −5 dB SNR with the two spectral-subtractive algorithms did not differ significantly ($p>0.05$) from performance obtained with unprocessed (noise corrupted) sentences, but was found to be significantly ($p<0.05$) lower than performance with unprocessed sentences at 0 dB SNR. The latter outcome is consistent with the findings reported by Hu and Loizou [4].

In summary, the above analysis indicates that the Region I and Region I+II constraints are the most robust in terms of yielding consistently large benefits in intelligibility *independent* of the speech-enhancement algorithm used. Substantial improvements in intelligibility (85 percentage points at −5 dB SNR and nearly 70 percentage points at 0 dB SNR) were obtained even with the RDC_mag algorithm, which was found in our previous study [4], as well as in the present study, to degrade speech intelligibility in some noisy conditions. Of the three enhancement algorithms examined, the Wiener algorithm is recommended when imposing Region I or Region I+II constraints, as this algorithm yielded the largest gains in intelligibility for both SNR levels tested. Based on data from Table I, there does not seem to be a correlation between the numbers of frequency bins falling in the three regions with speech intelligibility gains. The RDC_pow algorithm, for instance, yielded roughly the same number of frequency bins in Region I as the Wiener filtering algorithm, yet the latter algorithm obtained larger improvements in intelligibility. We attribute the difference in performance to the shape of the suppression function.

A difference in outcomes in Region II was observed between the Wiener and spectral subtractive algorithms. Compared to the performance obtained by the subtractive algorithms in Region II, the performance of the Wiener algorithm was substantially higher. To analyze this, we examined the frequency dependence of the distortions in Region II. More precisely, we examined whether distortions in region II (as introduced by the three different algorithms) occurred more frequently within a specific frequency region. We first divided the signal bandwidth into three frequency regions: low-frequency (0–1 kHz), mid-frequency (1–2 kHz) and high frequency regions (2–4 kHz). We then computed the percentage of bins falling in each of the three frequency regions for speech processed by the three algorithms (only accounting for distortions in Region II). The results, averaged over 20 sentences, are shown in Figure 5. As can be seen from this Figure, a slightly higher percentage of bins were observed in the lower frequency region (0–1 kHz) for the Wiener algorithm compared to the spectral subtractive algorithms. The higher percentage in the lower frequency region (0–1 kHz), where the first formant frequency resides, might partially explain the better intelligibility scores. But, this difference was rather small and not enough to account for the difference in intelligibility in Region II between the Wiener algorithm and the spectral subtractive algorithms.

We continued the analysis of Region II, and considered computing the histograms of the following estimation error: $\hat{X}_{\mathrm{dB}} - X_{\mathrm{dB}}$, where the subscript indicates that the magnitudes are expressed in dB. Note that this error is always positive and is upper bounded by 6.02 dB in Region II. The resulting histograms are shown in Figure 6. As can be seen from this Figure, magnitude errors smaller than 1 dB were made more frequently by the Wiener filtering algorithm for both SNR conditions, compared to the uniformly-distributed errors (at least at −5 dB SNR) made by the spectral-subtractive algorithms. This suggests that the Wiener filtering algorithm correctly estimates the true magnitude spectra more often compared to the subtractive algorithms, at least in Region II. We believe that this could be the reason that the Wiener algorithm performed better than the subtractive algorithms in Region II.

Performance in Region III (Figure 4) was extremely low (near 0% correct) for all three algorithms tested. We believe that this was due to the excess masking of the target signal in this region. Amplification distortions in excess of 6.02 dB were introduced. In Region III, the masker overpowered the target signal, rendering it unintelligible.

## IV. Relationship between proposed residual constraints and the ideal binary mask

As shown in Eq. (9), the modified spectrum (with the proposed constraints incorporated) can be obtained by applying a binary mask to the enhanced spectrum. In computational auditory scene analysis (CASA) applications, a binary mask is often applied to the noisy speech spectrum to recover the target signal [20–23]. In this section, we show that there exists a relationship between the proposed residual constraints (and associated binary mask) and the ideal binary mask used in CASA and robust speech recognition applications (e.g., [21]). The goal of CASA techniques is to segregate the target signal from the sound mixtures, and several techniques have been proposed in the literature to achieve that [23]. These techniques can be model-based [24,25] or based on auditory scene analysis principles [26]. Some of the latter techniques use the ideal time-frequency (T-F) binary mask [20,21,27]. The ideal binary "mask" (IdBM) takes values of zero or one, and is constructed by comparing the local SNR in each T-F unit (or frequency bin) against a threshold (e.g., 0 dB). It is commonly applied to the T-F representation of a mixture signal and eliminates portions of a signal (those assigned to a "zero" value) while allowing others (those assigned to a "one" value) to pass through intact. The ideal binary mask provides the only known criterion ($SNR \geq \delta$ dB, for a preset threshold δ) for improving speech intelligibility, and this was confirmed by several intelligibility studies with normal-hearing [28,29] and hearing-impaired listeners [30,31]. IdBM techniques often introduce musical noise, caused by errors in the estimation of the time-frequency masks and manifested in isolated T-F units. A number of techniques have been proposed to suppress musical noise distortions introduced by IdBM techniques [32,33].While musical noise might be distracting to the listeners, it has not been found to be detrimental in terms of speech intelligibility. This was confirmed in two listening studies with IdBM-processed speech [28,29] and in one study with estimated time-frequency masks [34]. Despite the presence of musical noise, normal-hearing listeners were able to recognize estimated [34] and ideal binary-masked [28,29] speech with nearly 100% accuracy.

The reasons for the improvement in intelligibility with IdBM are not very clear. Li and Wang [35] argued that the IdBM maximizes the SNR as it minimizes the sum of missing target energy that is discarded and the masker energy that is retained. More specifically, it was proven that the IdBM criterion maximizes the $SNR_{\mathrm{ESI}}$ metric given in Eq. (1) [35]. The IdBM was also shown to maximize the time-domain based segmental and overall SNR measures, which are often used for assessment of speech quality. Neither of these measures, however, correlates with speech intelligibility [9]. We provide proof in the Appendix that the IdBM criterion maximizes the geometric average of the spectral SNRs, and subsequently maximizes the articulation index (AI), a metric known to correlate highly with speech intelligibility [36].

As it turns out, the ideal binary mask is not only related to the proposed residual constraints, but is also a special case of the proposed residual constraint for regions I and II. Put differently, the proposed binary mask (see example in Eq. (9)) is a generalized form of the ideal binary mask used in CASA applications. As mentioned earlier, if the estimated magnitude spectrum is restricted to fall within regions I and II, then the $SNR_{\mathrm{ESI}}$ metric will always be greater than 0 dB. Hence, imposing constraints in region I+II ensures that $SNR_{\mathrm{ESI}}$ is always positive and greater than 1 (i.e., > 0 dB). As demonstrated in Figure 4, the stimuli constrained in region I+II consistently improved speech intelligibility for all three enhancement algorithms tested. As mentioned earlier, the composite constraint required for the estimated magnitude spectra to fall in region I+II is given by:

$$\widehat{X}(k) \leq 2 \cdot X(k) \qquad (10)$$

which after squaring both sides becomes:

$$\widehat{X}^2(k) \leq 4 \cdot X^2(k). \qquad (11)$$

If we now assume that $\hat{X}(k) = Y(k)$, i.e., that the noisy signal is not processed by an enhancement algorithm, then $\hat{X}^2(k) = Y^2(k) = X^2(k) + D^2(k)$, and Eq. (11) reduces to:

$$\begin{array}{l} X^2(k) \geq \frac{1}{3}D^2(k) \\ SNR(k) \geq \frac{1}{3} \end{array} . \qquad (12)$$

In dB, the above Equation suggests that the SNR needs to be larger than a threshold of −4.77 dB. Equation (12) is nothing but the criterion used in the construction of the ideal binary mask. The only difference is that the threshold used is −4.77 dB, rather than 0 or −3 dB, which are most often used in applications of the IdBM [27]. In terms of obtaining intelligibility improvement, however, either threshold is acceptable. The previous intelligibility studies confirmed that there exists a plateau in performance when intelligibility was measured as a function of the SNR threshold [28,29,37]. In the study conducted by Li and Loizou [37], for instance, the plateau in performance (nearly 100% correct) ranged from an SNR threshold of −20 dB to 0 dB.

As shown in Figure 4, the constraint stated in Eq. (10) guarantees substantial improvement in intelligibility for all three algorithms tested. The ideal binary mask is a special case of this constraint when no enhancement algorithm is used, i.e., when no processing is applied to the noisy speech signal. Unlike the criterion used in the binary mask (Eq. (12)), the proposed constraints (Eq. (10)) do not involve the noise spectrum, at least explicitly. In contrast, the ideal binary mask criterion requires access to the true noise spectrum, which is extremely challenging to obtain at very low SNR levels (e.g., SNR<0 dB). Attempts to estimate the binary mask using existing speech enhancement algorithms met with limited success (e.g., [38,39]), and performance, in terms of detection rates was found to be relatively poor. It remains to be seen whether it is easier to estimate the proposed binary mask (e.g., Eq. (9)), given that it does not require access to the true noise spectrum.

## V. Discussion and conclusions

Current speech enhancement algorithms can improve speech quality but not speech intelligibility [4]. Quality and intelligibility are two of the many attributes (or dimensions) of speech and the two are not necessarily equivalent. Hu and Loizou [2,4] showed that algorithms that improve speech quality do not improve speech intelligibility. The subspace algorithm, for instance, was found to perform the worst in terms of overall quality [2], but performed well in terms of preserving speech intelligibility [4]. In fact, in babble noise (0 dB SNR), the subspace algorithm performed significantly better than the logMMSE algorithm [40], which was found to be among the algorithms yielding the highest overall speech quality [2].

The findings of the present study suggest two interrelated reasons for the absence of intelligibility improvement with existing speech enhancement (SE) algorithms. First, and foremost, SE algorithms do not pay attention to the two types of distortions introduced when applying the suppression function to noisy speech spectra. Both distortions are treated equally in most SE algorithms, since the MSE metric is used in the derivation of most suppression functions (e.g., [7]). As demonstrated in Figure 4, however, the perceptual effects of the two distortions on speech intelligibility are not equal. Of the two types of distortion, the amplification distortion (in excess of 6.02 dB) was found to bear the most detrimental effect on speech intelligibility (see Figure 4). Performance dropped near zero when stimuli were constrained in region III. Theoretically, we believe that this is so because this type of distortion (region III) leads to negative values of $SNR_{\mathrm{ESI}}$ (see Figure 1). In contrast, the attenuation distortion (region I) was found to yield the least effect on intelligibility. In fact, when the region I constraint was imposed, large gains in intelligibility were realized. Performance at −5 dB SNR, improved from 5% correct with stimuli enhanced with the Wiener algorithm to 90% correct when region I constraint was imposed. Theoretically, we believe that the improvement in intelligibility is due to the fact that region I always ensures that $SNR_{\mathrm{ESI}} \geq 0$ dB. Maximizing $SNR_{\mathrm{ESI}}$ ought to maximize intelligibility, given the high correlation of a weighted-version of $SNR_{\mathrm{ESI}}$ (termed fwSNRseg [9,11]) with speech intelligibility. Hence, by imposing the appropriate constraints (see Eq. (10)), we can ensure that $SNR_{\mathrm{ESI}} \geq 0$ dB, and subsequently obtain large gains in intelligibility.

Second, none of the existing SE algorithms was designed to maximize a metric that correlates highly with intelligibility. The only known metric, which is widely used in CASA, is the ideal binary mask criterion. We provided a proof in the Appendix that this metric maximizes the articulation index, an index that is known to correlate highly with speech intelligibility [36]. Hence, it is not surprising that speech synthesized based on the IdBM criterion improves intelligibility [28,29,37]. In fact, it restores speech intelligibility to the level attained in quiet (near 100% correct) even for sentences corrupted by background noise at SNR levels as low as −10 dB SNR [29]. As shown in previous Section, the IdBM criterion is a special case of the proposed constraint in region I+II, when no suppression function is applied to the noisy spectra, i.e., when $\hat{X}(k) = Y(k)$.

In summary, in order for SE algorithms to improve speech intelligibility they need to treat the two types of distortions differently. More specifically, SE algorithms need to be designed so as to minimize the amplification distortions. As the data in Figure 4 demonstrated, even spectral-subtractive algorithms can improve speech intelligibility if the amplification distortions are properly controlled. In practice, the proposed constraints can be imposed and incorporated in the derivation of the noise suppression function. That is, rather than focusing on minimizing a squared-error criterion (as done in the derivation of MMSE algorithms), we can focus instead on minimizing a squared-error criterion *subject to* the proposed constraints. The speech enhancement problem is thus converted to a constrained minimization problem. Alternatively, and perhaps, equivalently, SE algorithms need to be designed so as to maximize a metric (e.g., $SNR_{\mathrm{ESI}}$, AI) that is known to correlate highly with speech intelligibility (for a review of such metrics, see [9]). For instance, SE algorithms need to be designed to maximize $SNR_{\mathrm{ESI}}$ rather than minimize an un-constrained (mean) squared-error cost function, as done by most statistical-model based algorithms (e.g., [7]). Algorithms that maximize the $SNR_{\mathrm{ESI}}$ metric are likely to provide substantial gains in intelligibility.

## Acknowledgments

This work was supported by Grant No. R01 DC007527 from NIDCD/NIH.

## VI. Appendix

In this Appendix, we provide analytical proof that the IdBM criterion is optimal in that it maximizes the geometric average of the spectral SNRs. We also show that maximizing the geometric average of SNRs is equivalent to maximizing the articulation index (AI)[3], an objective measure used for predicting speech intelligibility [36,44].

Consider the following weighted (geometric) average of SNRs computed across $N$ frequency bins:

$$F=\sum_{j=1}^{N} I_j \cdot \mathrm{SNR}(j) \tag{A.1}$$

where $SNR(j) = 10 \log_{10}(X^2(j)/D^2(j))$ is the SNR in bin (or channel) $j$ and $I_j$ are the weights ($0 \le I_j \le 1$) applied to each frequency bin. We consider the following question: How should the weights $I_j$ be chosen such that the overall SNR (i.e., $F$) given in Eq. (A.1) is maximized? The rationale for wanting to maximize $F$, stems from the fact that $F$ is similar to the articulation index (more on this below). The optimal weights $I_j$ that maximize $F$ in Eq. 1 are given by:

$$I_j^{*}=\begin{cases} 1 & \text{if SNR}(j)>0 \\ 0 & \text{if SNR}(j) \le 0 \end{cases} \tag{A.2}$$

which is no other than the IdBM criterion. To see why the weights given in Eq. (A.2) are optimal, we can consider two extreme cases in which either $SNR(j) \le 0$ or $SNR(j) \ge 0$ in all

[3] The AI index has been shown to predict reliably speech intelligibility by normal-hearing [36] and hearing-impaired listeners [41] (the refined AI index is known as the speech intelligibility index and is documented in [42]). The AI measure, however, has a few limitations. First, the AI measure has been validated for the most part only for stationary masking noise since it is based on the long-term average spectra, computed over 125-ms intervals, of the speech and masker signals [42]. As such, it cannot be applied to situations in which speech is embedded in fluctuating maskers e.g., competing talkers. Several attempts have been made, however, to extend the AI measure to assess speech intelligibility in fluctuating maskers (e.g., see [9,43]). Second, the AI measure cannot predict synergistic effects as evident in the perception of disjoint frequency bands. This is so due to the assumption that individual frequency bands contribute independently to AI.

frequency bins. If $SNR(j) \leq 0$ in all bins, then we have the following upper bound on the value of $F$:

$$\sum_{j=1}^{N} I_j \cdot SNR(j) \leq 0. \tag{A.3}$$

Similarly, if $SNR(j) \geq 0$ in all bins, then we have the following upper bound:

$$\sum_{j=1}^{N} I_j \cdot SNR(j) \leq \sum_{j=1}^{N} SNR(j). \tag{A.4}$$

Both upper bounds (maxima) in Eq. (A.3) and (A.4) are attained with the optimal weights given in Eq. (A.2). That is, the maximum in Eq. (A.3) is attained with $I_j = 0$ (for all $j$), while the maximum in Eq. (A.4) is attained with $I_j = 1$ (for all $j$).

It is important to note that the function $F$ given in Eq. (A.1), is very similar to the articulation index defined by [42,45,46]:

$$AI=\sum_{j=1}^{M} W_j \cdot \left(\alpha \cdot \overline{SNR}(j)+\beta\right) \tag{A.5}$$

where $W_j$ are the band-importance functions ($0 \leq W_j \leq 1$), $\overline{SNR}(j)$ are the SNR values limited to the range of [−15,15] dB, $M$ is the number of critical-bands, and α, β are constants ((α = 1/ 3 0 β, = 0.5) used to ensure that the SNR is mapped within the range of [0,1]. Maximization of AI in Eq. (A.5) will yield a similar optimal solution for the weights $W_j$ as shown in Eq. (A.2), with the only difference being the SNR threshold (i.e., it will no longer be 0 dB). The AI assumes a value of 0 when the speech is completely masked and a value between 0 and 1 for SNRs ranging from −15 to 15 dB. In the original AI calculation [44] the band-importance functions $W_j$, are fixed and their values depend on the type of speech material used. In our case, the importance functions $W_j$ are not fixed, but are chosen dynamically according to Eq. (A.2) so as to maximize the geometric average of all SNRs across the spectrum. Hence, the main motivation behind maximizing $F$ in Eq. (A.1) is to maximize the articulation index (Eq. (A.5)), and consequently maximize the amount of retained information contributing to speech intelligibility. Hence, the weights $I_j$ in Eq. (A.2) used in the construction of the ideal binary mask can be viewed as the optimal band-importance function $W_j$ needed to maximize the simplified form of articulation index in Eq. (A.1). It is for this reason that we believe that the use of the IdBM criterion (Eq. (A.2)) always improves speech intelligibility [29].

## References

1. Loizou, P. Speech Enhancement: Theory and Practice. Boca Raton: Florida: CRC Press LLC; 2007.
2. Hu Y, Loizou P. Subjective comparison and evaluation of speech enhancement algorithms. Speech Communication. 2007; vol. 49:588–601. [PubMed: 18046463]
3. Lim J. Evaluation of a correlation subtraction method for enhancing speech degraded by additive noise. IEEE Trans. Acoust., Speech, Signal Proc. 1978; vol. 37(no. 6):471–472.
4. Hu Y, Loizou P. A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Am. 2007; vol. 22(no. 3):1777–1786. [PubMed: 17927437]

5. Bentler R, Wu H, Kettel J, Hurtig R. Digital noise reduction: Outcomes from laboratory and field studies. Int. J. Audiol. 2008; vol. 47(no. 8):447–460. [PubMed: 18698521]
6. Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech and Audio Processing. 2001; vol. 9(no. 5):504–512.
7. Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. 1984 Dec.; vol. ASSP-32(no. 6): 1109–1121.
8. Hu Y, Loizou P. A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. on Speech and Audio Processing. 2003; vol. 11:334–341.
9. Ma J, Hu Y, Loizou P. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 2009; vol. 125(no. 5):3387–3405. [PubMed: 19425678]
10. Grancharov, V.; Kleijn, W. Speech Quality Assessment. In: Benesty, J.; Sondhi, M.; Huang, Y., editors. Handbook of Speech Processing. Berlin: Springer Verlag; 2008. p. 83-99.
11. Quackenbush, S.; Barnwell, T.; Clements, M. Objective measures of speech quality. Englewood Cliffs, NJ: Prentice Hall; 1988.
12. Hu Y, Loizou P. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang Processing. 2008; vol. 16(no. 1):229–238.
13. Tribolet J, Noll P, McDermott B, Crochiere RE. A study of complexity and quality of speech waveform coders. Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. 1978:586–590.
14. Scalart P, Filho J. Speech enhancement based on a priori signal to noise estimation. Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. 1996:629–632.
15. Gustafsson H, Nordholm S, Claesson I. Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Trans. on Speech and Audio Processing. 2001; vol. 9(no. 8):799–807.
16. Rangachari S, Loizou P. A noise-estimation algorithm for highly non-stationary environments. Speech Communication. 2006; vol. 48:220–231.
17. IEEE Subcommittee. IEEE Recommended Practice for Speech Quality Measurements. IEEE Trans. Audio and Electroacoustics. 1969; vol. AU-17(no. 3):225–246.
18. Voiers WD. Evaluating processed speech using the Diagnostic Rhyme Test. Speech Technology. 1983 Jan/Feb.vol.:30–39.
19. ITU. Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P. 862. 2000
20. Cooke M, Green P. Recognition of speech separated from acoustic mixtures. J. Acoust. Soc. Am. 1994; (no. 96):3293.
21. Cooke M, Green P, Josifovski L, Vizinho A. Robust automatic speech recognition with missing and uncertain acoustic data. Speech Communication. 2001; vol. 34:267–285.
22. Brown G, Cooke M. Computational auditory scene analysis. Computer, Speech and Language. 1994; vol. 8:297–336.
23. Wang, D.; Brown, G. Computational auditory scene analysis: Principles, algorithms, and applications. Hoboken, NJ: Wiley; 2006.
24. Ellis, D. Model-based scene analysis. In: Wang, D.; Brown, G., editors. Computational auditory scene analysis: Principles, algorithms, and applications. Hoboken, NJ: Wiley; 2006. p. 115-146.
25. Weiss R, Ellis D. Speech separation using speaker-adapted eigenvoice speech models. Computer, Speech and Language. 2010; vol. 24(no. 1):16–29.
26. Wang, D. Feature-based speech segregation. In: Wang, D.; Brown, G., editors. Computational auditory scene analysis: Principles, algorithms, and applications. Hoboken, NJ: Wiley; 2006. p. 81-114.
27. Wang, D. On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P., editor. Speech separation by humans and machines. Norwell, MA: Kluwer Academic; 2005. p. 181-197.

28. Brungart D, Chang P, Simpson B, Wang D. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. 2006; vol. 120(no. 6):4007–4018. [PubMed: 17225427]

29. Li N, Loizou P. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. J. Acoust. Soc. Am. 2008; vol. 123(no. 3):1673–1682. [PubMed: 18345855]

30. Wang D, Kjems U, Pedersen MS, Boldt JB, Lunner T. Speech intelligibility in background noise with ideal binary time-frequency masking. J. Acoust. Soc. Am. 2009; vol. 125:2336–2347. [PubMed: 19354408]

31. Hu Y, Loizou P. A new sound coding strategy for suppressing noise in cochlear implants. J. Acoust. Soc. Am. 2008; vol. 124(no. 1):498–509. [PubMed: 18646993]

32. Araki S, Makino S, Sawada H, Mukai R. Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. 2005; vol. 3:81–84.

33. Jan T, Wang W, Wang D. A multistage approach for blind separation of convolutive speech mixtures. Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. 2009:1713–1716.

34. Kim G, Lu Y, Hu Y, Loizou P. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. 2009; vol. 126(no. 3):1486–1494. [PubMed: 19739761]

35. Li Y, Wang D. On the optimality of ideal time-frequency masks. Speech Communication. 2009; vol. 51:230–239.

36. Kryter K. Validation of the articulation index. J. Acoust. Soc. Am. 1962; vol. 34:1698–1706.

37. Li N, Loizou P. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. J. Acoust. Soc. Am. 2009; vol. 123(no. 3):1673–1682.

38. Hu Y, Loizou P. Techniques for estimating the ideal binary mask. Proc. of 11th International Workshop on Acoustic Echo and Noise Control. 2008

39. Renevey P, Drygajlo A. Detection of reliable features for speech recognition in noisy conditions using a statistical criterion. Proc. Consistent and Reliable Acoustic Cues for Sound Analysis Workshop. 2001; vol. 1:71–74.

40. Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. 1985; vol. ASSP-33(no. 2): 443–445.

41. Pavlovic CV, Studebaker GA, Sherbecoe R. An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. J. Acoust. Soc. Am. 1986; vol. 80:50–57. [PubMed: 3745665]

42. ANSI S3.5-1997. Methods for calculation of the speech intelligibility index. American National Standards Institute. 1997

43. Rhebergen K, Versfeld N, Dreschler W. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. J. Acoust. Soc. Am. 2006; vol. 120:3988–3997. [PubMed: 17225425]

44. Kryter K. Methods for calculation and use of the articulation index. J. Acoust. Soc. Am. 1962; vol. 34(no. 11):1689–1697.

45. French NR, Steinberg JC. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 1947; vol. 19:90–119.

46. Pavlovic CV. Derivation of primary parameters and procedures for use in speech intelligibility predictions. J. Acoust. Soc. Am. 1987; vol. 82:413–422. [PubMed: 3624646]

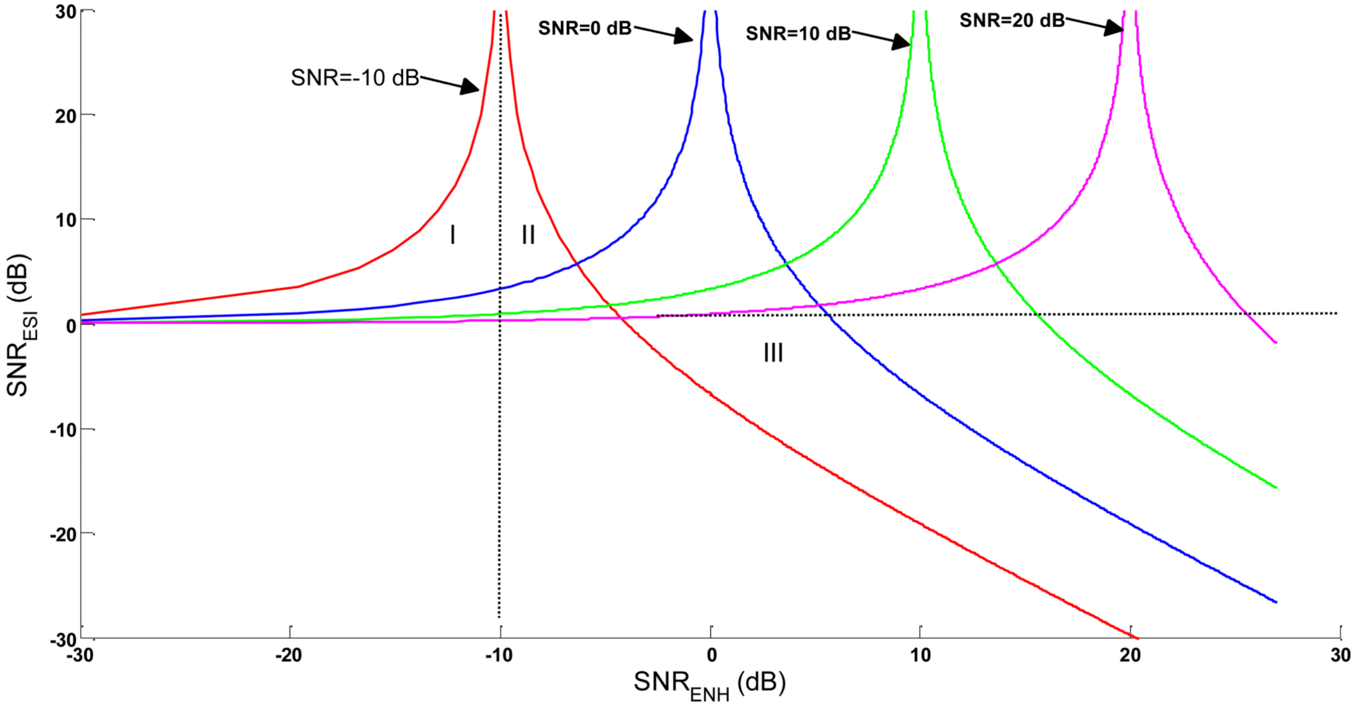

**Figure 1.**

Plot showing the relationship between SNR$_{ESI}$ and SNR$_{ENH}$ for fixed values of SNR.

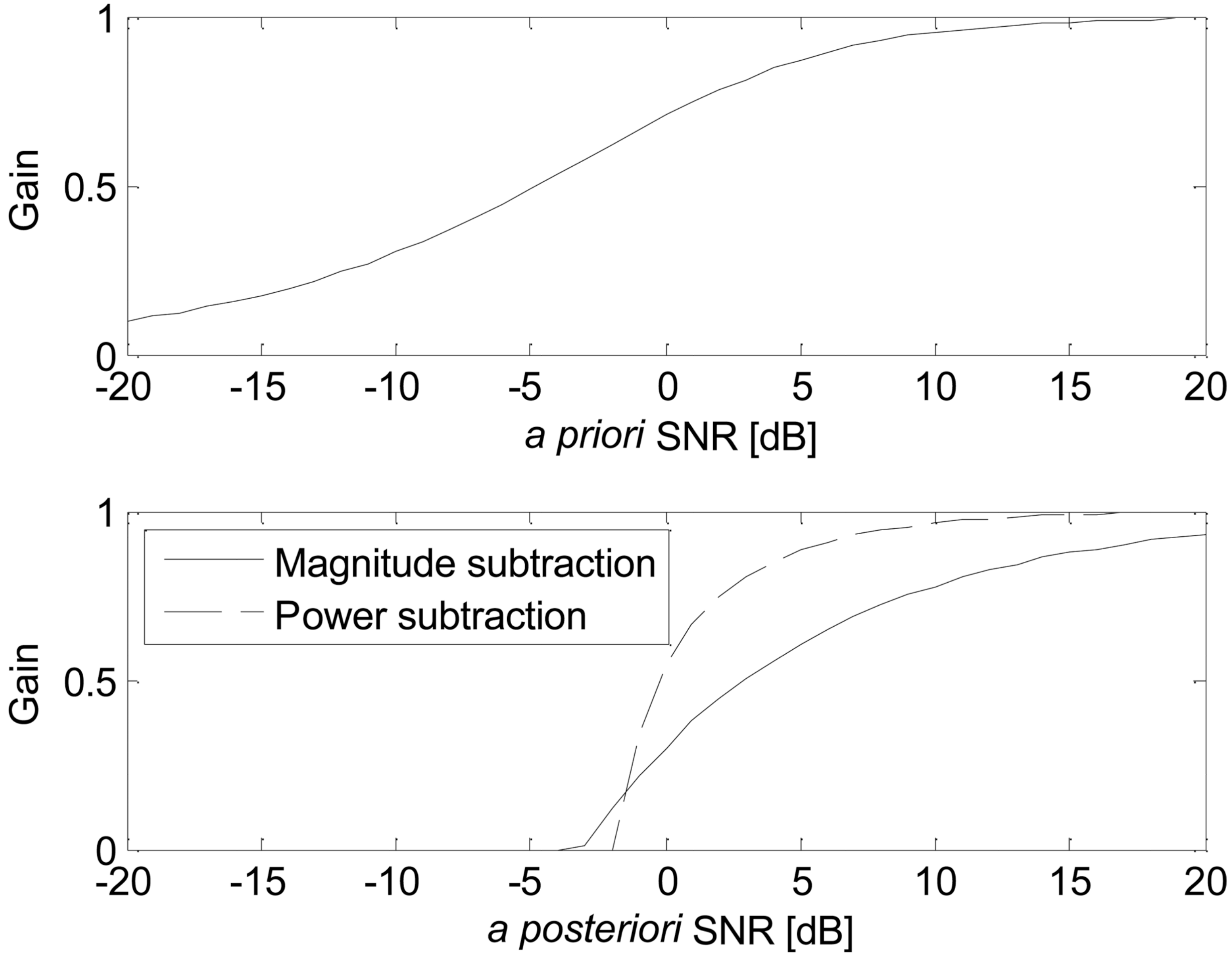

**Figure 2.**

Suppression curves of the Wiener filtering algorithm (top panel) and two spectral-subtractive algorithms (bottom panel).

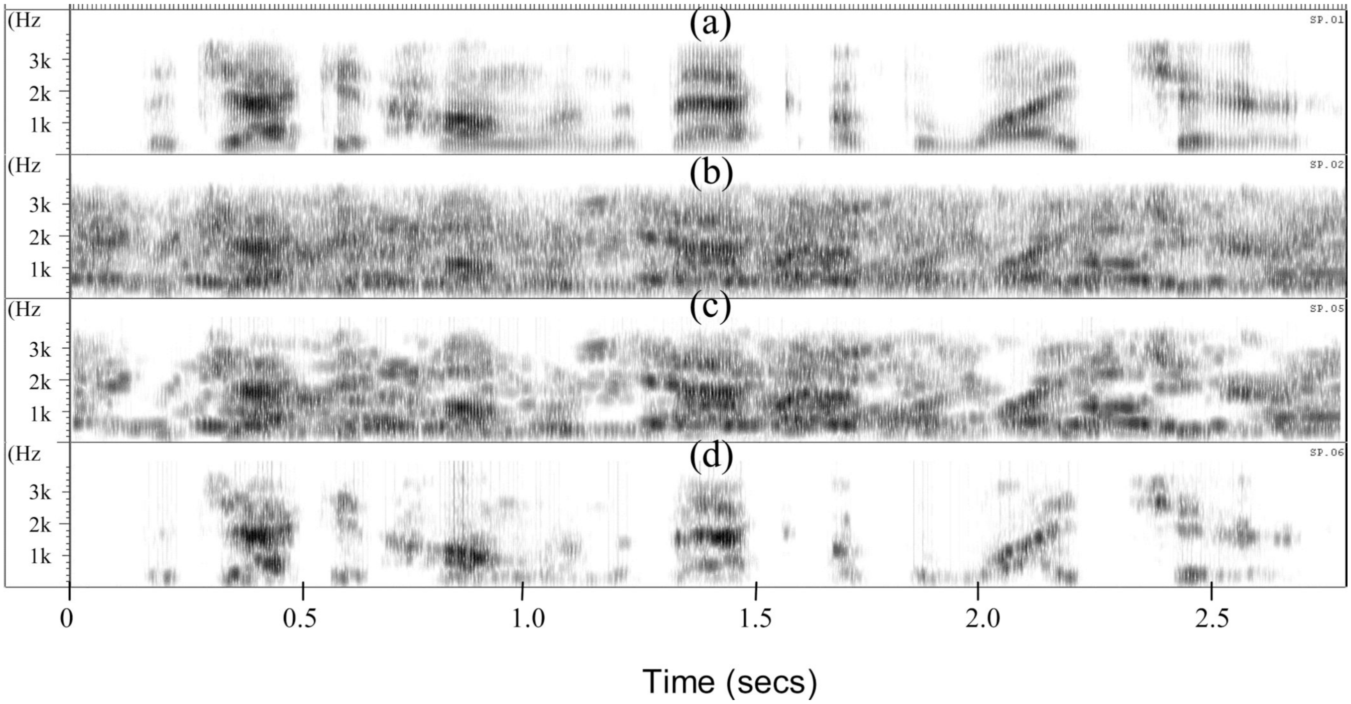

**Figure 3.**

Wide-band spectrograms of the clean signal (panel a), noisy signal in −5 dB SNR babble (panel b), signal processed by the Wiener algorithm (panel c), and signal prcessed by the Wiener algorithm after imposing the constraints in Region I (panel d).

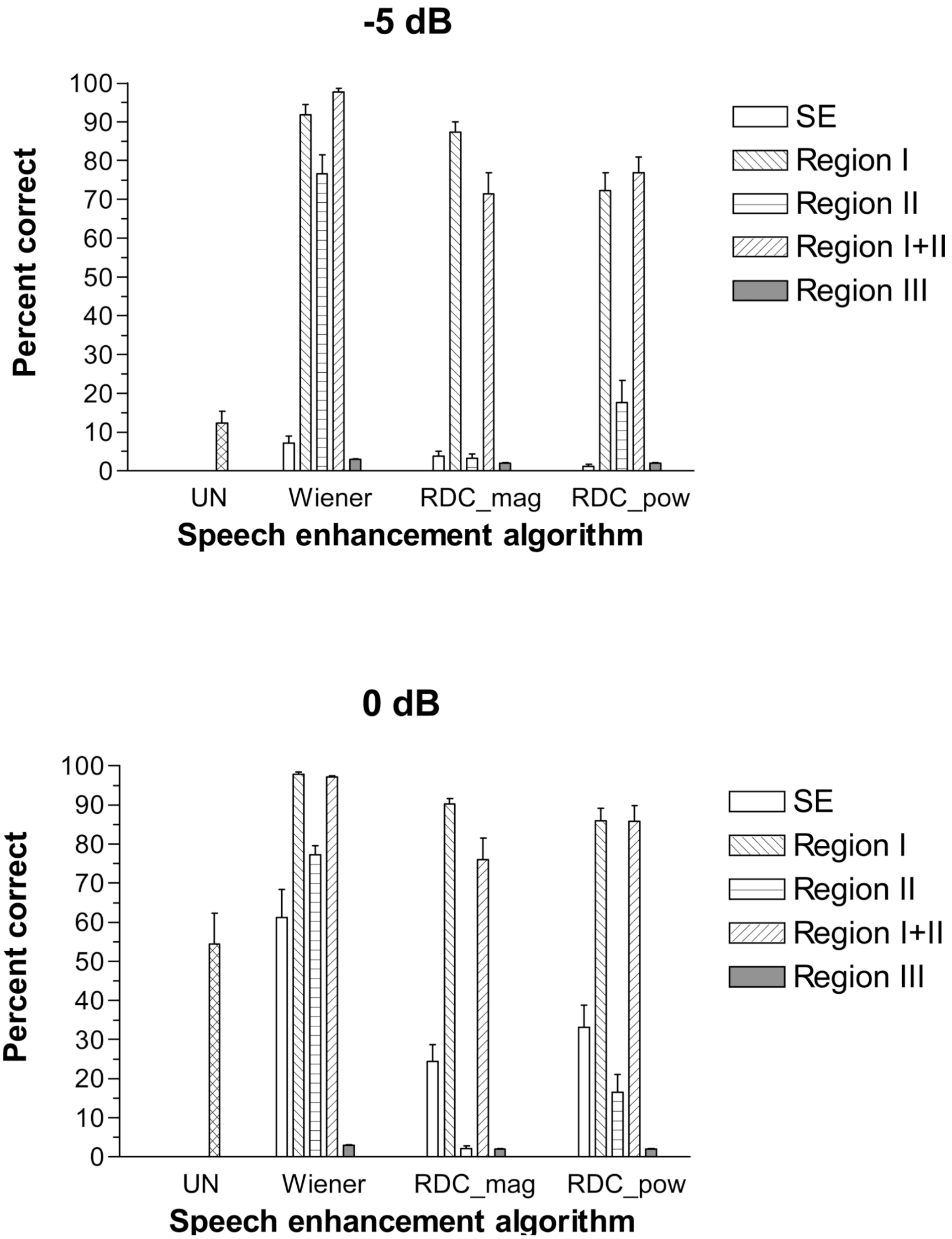

**Figure 4.**

Results, expressed in percentage of words identified correctly, from the intelligibility studies with human listeners. The bars indicated as "UN" show the scores obtained with noise-corrupted (un-processed) stimuli, while the bars indicated as "SE" show the baseline scores obtained with the three enhancement algorithms (no constraints imposed). The intelligibility scores obtained with speech processed by the three enhancement algorithms after imposing four different constraints are labeled accordingly.

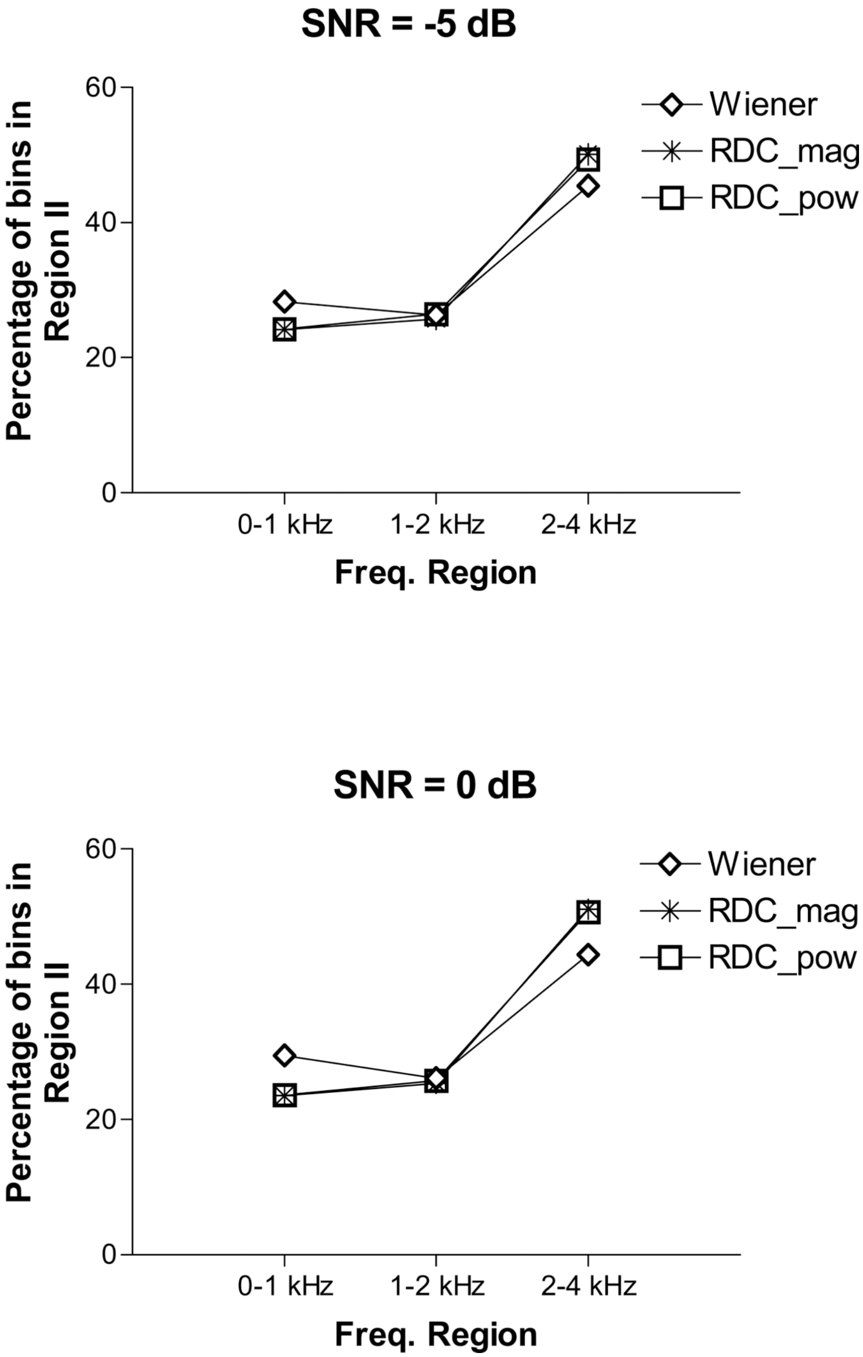

**Figure 5.**

Percentage of bins falling in three different frequency regions (Region II constraints).

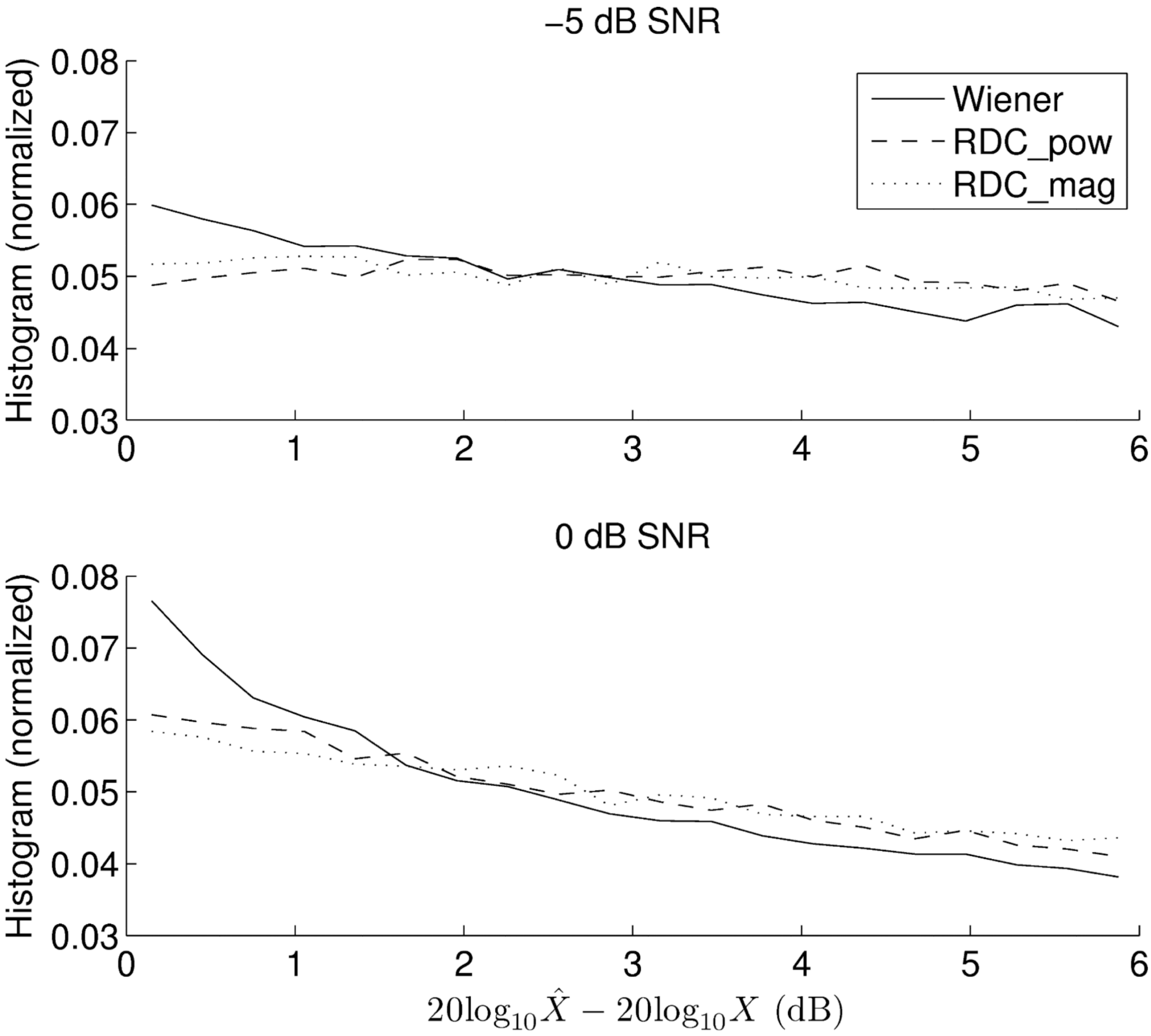

**Figure 6.**

Normalized histograms (probability mass function) of the difference between the estimated and clean speech magnitudes in Region II.

## Table 1

Percentage of frequency bins falling in the three Regions after processing noisy speech by the three enhancement algorithms.

| SNR Level | Algorithm | Region I | Region II | Region III |
|---|---|---|---|---|
| 0 dB | Wiener | 45.33 % | 14.64 % | 40.03 % |
| −5 dB |  | 37.71 % | 12.71% | 49.58 % |
| 0 dB | RDC_mag | 46.86 % | 15.45 % | 37.69 % |
| −5 dB |  | 35.88 % | 14.49 % | 49.63 % |
| 0 dB | RDC_power | 36.81 % | 18.14 % | 45.05 % |
| −5 dB |  | 28.08 % | 14.88 % | 57.04 % |