Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms

Philipos C. Loizou[a)]
Department of Electrical Engineering University of Texas at Dallas Richardson, Texas 75083-0688

Jianfen Ma[b)]
College of Computer Engineering and Software, Taiyuan University of Technology, Shanxi, China 030024

(Received 14 July 2010; revised 29 March 2011; accepted 7 June 2011)

The conventional articulation index (AI) measure cannot be applied in situations where non-linear operations are involved and additive noise is present. This is because the definitions of the target and masker signals become vague following non-linear processing, as both the target and masker signals are affected. The aim of the present work is to modify the basic form of the AI measure to account for non-linear processing. This was done using a new definition of the output or effective SNR obtained following non-linear processing. The proposed output SNR definition for a specific band was designed to handle cases where the non-linear processing affects predominantly the target signal rather than the masker signal. The proposed measure also takes into consideration the fact that the input SNR in a specific band cannot be improved following any form of non-linear processing. Overall, the proposed measure quantifies the proportion of input band SNR preserved or transmitted in each band after non-linear processing. High correlation ($r = 0.9$) was obtained with the proposed measure when evaluated with intelligibility scores obtained by normal-hearing listeners in 72 noisy conditions involving noise-suppressed speech corrupted in four different real-world maskers. [DOI: 10.1121/1.3605668]

PACS number(s): 43.71.Gv, 43.71.An [AA] Pages: 986–995

I. INTRODUCTION

A number of measures have been proposed in the literature to predict speech intelligibility in the presence of background noise. Among these measures, the articulation index (AI) (French and Steinberg, 1947; Fletcher and Galt, 1950; Kryter, 1962a; ANSI, 1997; Amlani *et al.*, 2002) and speech-transmission index (STI) (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985) are by far the most commonly used today for predicting speech intelligibility. The AI measure was further refined to produce the speech intelligibility index (SII) (ANSI 1997). The AI measure is based on the principle that the intelligibility of speech depends on the proportion of spectral information that is audible to the listener and is computed by dividing the spectrum into 20 bands (contributing equally to intelligibility) and estimating the weighted average of the signal-to-noise ratios (SNRs) in each band (Kryter, 1962a, 1962b; Pavlovic, 1987; ANSI, 1997; Amlani *et al.*, 2002). The SNRs in each band are weighted by band-importance functions which differ across speech materials (ANSI, 1997). The AI measure has been shown to predict successfully the effects of linear filtering and additive noise on speech intelligibility (e.g., Kryter, 1962a, 1962b).

The AI measure has, however, a number of limitations. First, it has been validated for the most part only for steady (stationary) masking noise since it is based on the long-term average spectra of the speech and masker signals. 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 to extend the AI measure to assess speech intelligibility in fluctuating maskers (Rhebergen *et al.*, 2005, 2006; Kates, 1987). Second, according to the ANSI (1997) standard, the SII measure cannot be used in conditions which include multiple sharply filtered bands of speech or sharply filtered noises. Incidentally, sharply filtered bands of speech can be produced when speech is processed via spectral-subtractive algorithms due to the non-linear thresholding of the speech envelopes (Goldsworthy and Greenberg, 2004; Loizou, 2007). Third, it cannot be applied in situations where non-linear operations are involved and additive noise is present. This is because the definitions of the target and masker signals are no longer clear following non-linear processing, as both the target and masker signals are affected. Consequently, the definition of the true output SNR, namely, the effective SNR following non-linear processing, poses great challenges. In contrast, in situations wherein speech is subjected to linear filtering operations (e.g., low-pass filter), the SNR can be determined based on the target and masker signals prior to mixing. Extensions to the AI index based on a different definition of the SNR were proposed by Kates and Arehart (2005) to predict the intelligibility of peak-clipping and center-clipping distortions, such as those introduced by hearing aids. The modified index, called the CSII index, used the base form of the SII procedure, but with the signal-to-noise ratio term

[a)]Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu
[b)]Work done while Dr. Jianfen Ma visited Prof. Loizou's lab as a research scholar.

 0001-4966/2011/130(2)/986/10/$30.00

replaced by the signal-to-distortion ratio, which was computed using the coherence function between the input and processed signals. Only a few studies have attempted to extend the AI index to handle non-linear processing (French and Steinberg, 1947; Stelmachowicz *et al.*, 1998; Kates and Arehart, 2005; Taal *et al.*, 2010). Stelmachowicz *et al.* (1998) modified the AI index to account for the differences in audibility provided by wide dynamic-range compression algorithms used in hearing aids.

The last limitation of the AI measure is also shared by the STI measure. When speech is subjected to non-linear processes such as those introduced by dynamic envelope compression (or expansion) in hearing aids, the STI measure fails to successfully predict speech intelligibility since the processing itself might introduce additional modulations which the STI measure interprets as increased SNR (Hohmann and Kollmeieir, 1995; Ludvigsen *et al.*, 1993; Van Buuren *et al.*, 1999; Goldsworthy and Greenberg, 2004). The STI measure has also failed to predict the lack of intelligibility benefit with spectral-subtractive noise reduction algorithms. Ludvigsen *et al.* (1993), for instance, have shown that in spite of the increased speech modulations and increase in STI values, the intelligibility of signals processed via the spectral-subtractive algorithm is not better than that of unprocessed signals. A number of methods (Dubbelboer and Houtgast, 2008; Goldsworthy and Greenberg, 2004) have been proposed to circumvent this limitation, but these methods cannot be easily extended to the AI measure. Both the AI and STI measures compute the SNR in each band; however, the SNR values are derived differently and the two measures are based on different principles. The STI measure (Steeneken and Houtgast, 1980) is based on the principle that the reduction in intelligibility caused by additive noise or reverberation distortions can be modeled in terms of the reduction in temporal envelope modulations. In contrast, the AI measure does not consider or account for any envelope modulation reduction/increase in its computation, but rather computes directly the SNR in each band (see review by Amlani *et al.*, 2002).

The present study takes the first step in modifying the base structure of the AI measure to handle non-linear processing, particularly when noise is present. More precisely, it considers the non-linear processing involved when the corrupted signals are processed via noise-reduction algorithms. It does not consider the non-linear distortions introduced by hearing aids (e.g., wide-dynamic compression, peak-clipping distortion) since those have been studied extensively by Kates and colleagues (Kates, 1992; Kates and Arehart, 2005; Kates, 2010). Nonetheless, the proposed model can potentially be extended to handle hearing-aid like distortions, but the focus of this article is on modeling the distortions introduced by noise-suppression algorithms. A new definition of output SNR is proposed which is used in conjunction with the traditional SNR definition to derive a new intelligibility measure. The proposed output SNR definition is designed to handle cases where the non-linear processing affects predominantly the target signal rather than the masker signal. The issues and problems surrounding the definition of SNR when non-linear processes are involved are discussed in the next section, followed by the evaluation of the proposed measure with speech intelligibility scores collected in our prior study (Hu and Loizou, 2007) assessing the evaluation of noise-reduction algorithms.

II. CHALLENGES IN DETERMINING THE EFFECTIVE SNR FOLLOWING NON-LINEAR PROCESSING

Most (if not all) noise-suppression algorithms employed for hearing aids or for other applications involve a gain reduction stage, in which the mixture envelope or spectrum is multiplied by a gain function G_k (taking typically values ranging 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) (see Fig. 9 in Chung, 2004). Figure 1 shows the signal-processing framework used in the present study. The noisy speech envelope (denoted as S + N, where S = target signal and N = masker signal) extracted at a specific band is non-linear processed to produce the output $\hat{S}$. Note that in quiet, the masker signal can be constructed using "internal noise" that is added at the appropriate level based on the absolute pure-tone hearing threshold (French and Steinberg, 1947; Pavlovic, 1987). In hearing aid applications, the processed envelope $\hat{S}$ would represent the output of a non-linear operation (e.g., dynamic range compression) which can be expressed mathematically as $\hat{S}=f(S+N)$, where $f(.)$ represents the non-linear function (e.g., compression, expansion, etc.) used. In noise-reduction applications, the $f(.)$ function would represent the noise-suppressive gain function (e.g., see Fig. 9 and Table II in Chung, 2004) that is applied to the noisy speech envelopes in order to suppress the background noise. In most cases, the gain function is highly non-linear. In the power spectral-subtractive algorithm (Boll, 1979), for instance, the gain function takes the form:

$$G_k = \max\left(0, 1 - \frac{\hat{N}_k^2}{Y_k^2}\right) \tag{1}$$

where $\hat{N}_k^2$ denotes the estimated masker power-spectrum and Y_k denotes the corrupted (S + N) spectrum (envelope) in

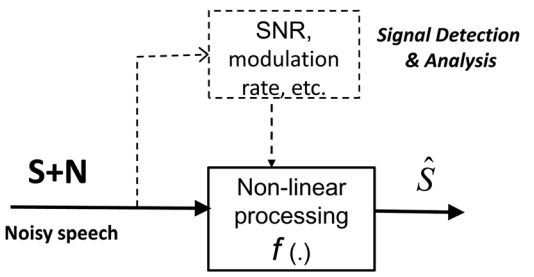

FIG. 1. Signal-processing framework used in the present study for analyzing non-linear operations in the presence of noise. The dashed block shows the additional stage used in most noise-reduction applications to compute parameters such as band SNR, modulation rate, etc. These parameters are in turn used to construct a noise-suppressive gain function. The function $f(.)$ represents generally the gain function used in noise-reduction or the non-linear function (e.g., compression function) used in hearing-aid applications.

band k (the max operator is used to ensure that the gain function is always positive). Following the computation of Eq. (1), the gain function G_k is applied to the noisy speech spectrum Y_k in band k to produce the output envelope $\hat{S}_k$, i.e., $\hat{S}_k = f(Y_k) = G_k \cdot Y_k$. In the present study, eight different non-linear functions G_k, representing eight different noise-reduction algorithms (see description in Hu and Loizou, 2007) are used to test the proposed intelligibility index.

The non-linear processing of the noisy speech envelopes (Fig. 1), whether be for noise suppression applications or dynamic-range compression as implemented in hearing aids, poses certain challenges in terms of defining the effective output band SNR based on $\hat{S}_k$. This is so because the non-linear function (e.g., gain G_k in noise-reduction applications) affects both the target and masker signals and thus we can no longer assume that the output envelope $\hat{S}_k$ always corresponds to the modified (e.g., attenuated, compressed, etc.) target signal. To see this, we can express the square of the output (non-linearly processed) envelope, i.e., $\hat{S}_k^2$, as follows:

$$\begin{aligned}\hat{S}_k^2 &= G_k^2 Y_k^2 = G_k^2 (S_k^2 + N_k^2) \\ &= G_k^2 S_k^2 + G_k^2 N_k^2 \\ &= S_T^2 + S_M^2\end{aligned} \tag{2}$$

where S_k denotes the clean signal envelope, N_k indicates the masker envelope, $S_T{}^2$ denotes the power of the modified (by non-linear processing) target component and $S_M{}^2$ (the subscript k was omitted for clarity) denotes the power of the modified masker component of $\hat{S}_k^2$. Knowing whether the target component (i.e., $S_T{}^2$) of the output envelope is dominant is important in as far as defining the effective or output band SNR.

A. Defining the output band SNR

Consider the corrupted (mixture) spectrum S + N in band k (or frequency bin k) being processed by a noise-reduction algorithm specified by the gain function G_k. Since the masker is additive, the gain function is applied to both the target spectrum S_k and the masker spectrum N_k (see Eq. (2) above). Consequently the output SNR in band k, denoted as SNRout(k), can be computed as follows:

$$\mathrm{SNRout}(k) = \frac{S_T^2}{S_M^2} = \frac{(G_k S_k)^2}{(G_k N_k)^2} = \frac{S_k^2}{N_k^2} = \mathrm{SNR}_k \tag{3}$$

where $(G_k S_k)^2$ denotes the power of the modified (by G_k) target signal in band k, $(G_k N_k)^2$ indicates the power of the modified masker signal, and $\mathrm{SNR_k}$ denotes the input band SNR as determined prior to mixing. According to the above equation, the output band SNR cannot be improved by any choice of G_k beyond the value of the input band $\mathrm{SNR_k}$. This observation partially explains the lack of intelligibility with existing noise-reduction algorithms by NH listeners (Hu and Loizou, 2007; Loizou, 2007) and hearing-impaired listeners (Bentler *et al.*, 2008), at least for algorithms that make use of gain functions to suppress the background noise. In hearing-aid applications, methods that reduce upward spread of masking can potentially be used in place of the noise-suppressive gain functions to improve speech intelligibility in noise (see review by Levitt, 1997). It is worth mentioning that while noise-reduction algorithms do not improve the SNR in a specific band, they *can* improve the *overall* SNR accumulated (and appropriately weighted) across all bands. Note that the overall SNR (computed across all bands) and the output band SNR [computed for a specific band as per Eq. (3)] are different. One strategy for improving the overall SNR (defined as the weighted sum of SNRs across all bands) is to discard bands with unfavorable (extremely low) SNRs while retaining bands with favorable SNR (see proof in Loizou and Kim, 2011). Such an approach was taken in our prior study and has been shown to improve speech intelligibility by normal-hearing listeners (Kim *et al.*, 2009) as well as by cochlear implant listeners (Hu and Loizou, 2008).

Clearly, the above definition of output band SNR is not useful as it does not involve the output or processed envelope $\hat{S}$. Alternatively, the output band SNR can be defined as follows (Ma *et al.*, 2009):

$$\overline{SNR}_k = \frac{\hat{S}_k^2}{N_k^2} \tag{4}$$

where $\overline{SNR}_k$ denotes the new definition of the output SNR in band k, and $\hat{S}_k$ denotes the processed (via, say, a noise-reduction algorithm) envelope. Similar to the AI computation, the above SNR was limited, mapped to [0,1] and weighted by band-importance functions (BIFs) in the study by Ma *et al.* (2009). The above measure, however, yielded a poor correlation ($r < 0.4$) with intelligibility scores (Ma *et al.*, 2009). We believe that it was because of the inherent ambiguity associated with non-linear processing when the G_k suppression function is applied to the noisy speech envelopes. More specifically, when the corrupted envelope is processed by a noise-reduction algorithm (via the application of the gain function G_k), it is not clear whether the resulting envelope $\hat{S}$ corresponds predominantly to say the modified (e.g., attenuated) target envelope or the modified masker envelope [see Eq. (2)]. Consequently, we cannot easily define the "true" output band SNR as we do not know beforehand whether $\hat{S}$ reflects primarily the modified masker envelope or the modified target envelope.

It is clear from the above discussion that a distinction needs to be made in Eq. (4) to reflect the scenarios in which the non-linear processing affects primarily (or predominantly) the target envelope rather than the masker envelope. If the target envelope is dominantly larger than the masker envelope (i.e., SNR $\gg$ 0 dB) then the envelope $\hat{S}$ will most likely reflect the modified target envelope (since the masker component will be extremely small), whereas if the masker envelope is dominantly larger than the target envelope (i.e., SNR $\ll$ 0 dB) then the envelope $\hat{S}$ will most likely reflect the modified masker envelope (since the target component will be extremely small). Determining, however, the appropriate SNR threshold to discriminate between these two scenarios is not straightforward given that the non-linear processing affects both the target and masker envelopes;

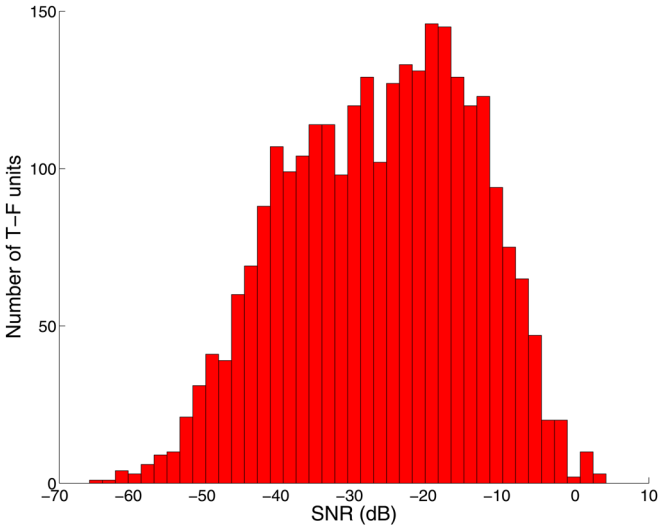

FIG. 2. (Color online) Histogram of band SNRs for corresponding bands in which $\hat{S} > S$ after noise- suppression. Band SNRs were determined for each time-frequency (T-F) unit, and accumulated over the duration of a sentence.

hence we considered an alternative strategy for making the distinction. There are two possible scenarios to consider. In the first scenario $\hat{S} < S$, suggesting attenuation of the target envelope and in the second scenario $\hat{S} > S$, suggesting amplification of the target envelope. As $\hat{S}$ gets significantly larger than S (i.e., overestimation occurs), the corresponding masker envelope also gets larger and at some point the input band SNR will become negative. This is demonstrated in Fig. 2 which shows the histogram of SNR values for all bands for which $\hat{S} > S$ (histogram was computed for one sentence processed by a spectral-subtractive algorithm at 0 dB SNR). As can be seen, the input band SNR is for the most part negative when $\hat{S} > S$. In fact, it can be proven (see Appendix) that for a certain range of gain values, the input band SNR is always negative when $\hat{S} > S$. Furthermore, it can also be proven analytically (see Appendix) that when $\hat{S} \geq 2 \cdot S$, the corresponding input band SNR is always negative. Consequently, bands for which $\hat{S} \geq 2 \cdot S$ holds should not be included since the speech information is masked. Hence, for the most part, when $\hat{S} > S$, the envelope $\hat{S}$ will likely reflect the modified masker envelope and thus should not be used in the definition of the output SNR in Eq. (4). Put differently, when $\hat{S} > S$ the masker component of $\hat{S}$ will for the most part be larger than the target component [i.e., $S_M > S_T$ in Eq. (2)], and thus $\hat{S}$ should not be used in Eq. (4).

B. Proposed intelligibility measure

In brief, as shown in Eq. (3) the output band SNR cannot exceed the input band SNR. Second, the above limitations in using Eq. (4) to compute the output band SNR can be circumvented to some extent if we identify the situations where $\hat{S}$ better reflects the effects of non-linear processing (e.g., noise reduction) on the target envelope rather than on the masker envelope. As discussed above, the processed envelope $\hat{S}$ reflects more reliably the effect of suppression on the target envelope when $\hat{S} < S$ than when $\hat{S} > S$. It seems reasonable then to restrict $\hat{S}$ in Eq. (4) to be always smaller than S, and thus consider in the computation of the proposed measure only bands in which $\hat{S} < S$. The implicit hypothesis is that those bands will contribute more to intelligibility and should thus be included. This was confirmed in listening studies (Loizou and Kim, 2011) in which normal-hearing listeners were presented with speech synthesized to contain either target attenuation distortions alone (i.e., bands with $\hat{S} < S$) or target amplification distortions alone (i.e., bands with $\hat{S} > S$). Speech synthesized to contain only target attenuation was always more intelligible, and in fact, it was found to be more intelligible than either the un-processed (noise corrupted) or processed (via the noise-reduction algorithm) speech.

After taking the above facts into account, we derive a new measure which computes the fraction or proportion of the input SNR transmitted as follows:

$$fSNR_k = \begin{cases} \dfrac{\min\left(\overline{SNR}_k, SNR_k\right)}{SNR_k} & \text{if } SNR_k \geq SNR_L \\ 0 & \text{else} \end{cases} \tag{5}$$

where $fSNR_k$ denotes the fraction (or proportion) of the input SNR transmitted (by the noise-reduction algorithm), $\overline{SNR}_k$ is given by Eq. (4), $\mathrm{SNR_k}$ is the true SNR [see Eq. (3)] and $\mathrm{SNR_L}$ denotes the smallest SNR value allowed. It is clear that $fSNR_k$, is bounded by 1, i.e., $0 \leq fSNR_k \leq 1$, and thus denotes the fraction (or proportion) of the input SNR preserved (or transmitted) by the noise-reduction algorithm in a specific band. The maximum value of 1 is attained by $fSNR_k$ when no non-linear processing (i.e., $G_k = 1$) is applied to the noisy envelopes. A value close to 1 can also be obtained when $\hat{S} \approx S$, i.e., when the noise-reduction algorithm produces an accurate estimate of the clean target spectrum. The use of minimum operation in Eq. (5) ensures that only bands for which $\hat{S} < S$ are included. Furthermore, only bands with SNR falling above a certain value (e.g., $\mathrm{SNR_L} > 0$ dB) are considered. This is necessary for two reasons. First, the condition $\hat{S} < S$ does not guarantee that the input SNR will always be positive. Second, use of $\mathrm{SNR_L} > 0$ dB always guarantees that the target component of the output envelope will always be larger than the masker component [see Eq. (2)]. Following Eq. (5), the new measure is weighted and accumulated across all bands to produce the fractional AI (fAI) index:

$$fAI = \frac{1}{\sum_{k=1}^{M} W_k} \sum_{k=1}^{M} W_k \times fSNR_k \tag{6}$$

where $\mathrm{W_k}$ denotes the weighting functions or band-importance functions applied to band k and M is the total number of bands used. Unlike the traditional AI measure, the proposed fAI measure is computed based on the weighted average of the proportion of the input SNR transmitted by the noise-suppression algorithm in each band. Note that in the traditional AI measure, the $fSNR_k$ values are replaced by the audibility functions (ranging from 0 to 1) expressing the proportion of speech information that is audible to the listener.

Given that most noise-suppression algorithms operate on short-time segments (20–30 ms frames), the above

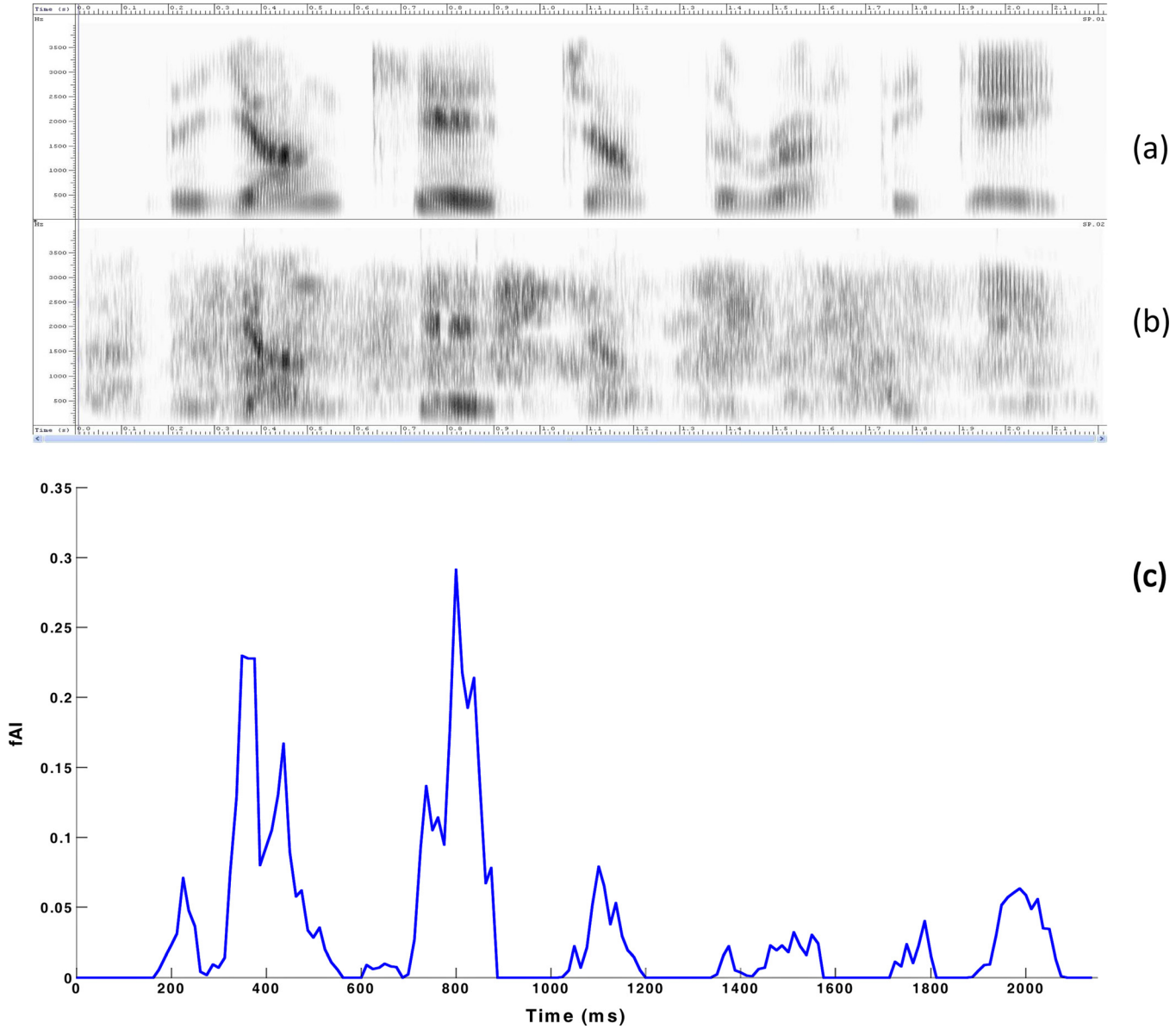

FIG. 3. (Color online) Panel (a) shows the wideband spectrogram of the IEEE sentence "The young kid jumped the rusty gate." in quiet, and panel (b) shows the sentence processed via a spectral- subtractive algorithm. The input sentence was originally corrupted by babble at 0 dB SNR. Panel (c) shows the corresponding short-term fAI values computed every 50 ms. The resulting average fAI value was 0.032.

measure [Eq. (6)] can be computed for each frame and averaged across all frames to produce a single value for each sentence. Figure 3(c) shows example short-term fAI values computed over a sentence that has been processed by the spectral-subtractive noise-suppression algorithm. The sentence was corrupted by babble at 0 dB SNR. For this example, SNR_L was set to 0 dB and no weighting was applied to the $fSNR_k$ values, i.e., $W_k = 1$. The spectrograms of the clean and processed sentences are shown in Figs. 3(a) and 3(b). From Fig. 3 we observe that the fAI value was high near the t = 400 ms and 800 ms segments, and low for the remaining words/phonemes. This was consistent with the amount of "clean" speech information (e.g., formant movements) that was evident in the spectrogram of the noise-suppressed signal [Fig. 3(b)]. We can thus, say, that the fAI measure can be regarded as a microscopic measure in as far as having the potential to predict the intelligibility of non-linearly processed speech at the phoneme or word level.

C. Implementation

The proposed *fAI* measure [Eq. (6)] was implemented as follows. The speech signals were first segmented using 50-ms duration Hamming windows with 75% overlap between adjacent frames. The critical-band spectra of the target and masker signals (prior to mixing) and the processed signals were obtained for each 50-ms frame by multiplying the FFT magnitude spectra by 25 overlapping Gaussian-shaped windows (Loizou, 2007, Ch. 11) spaced in proportion to the ear's critical bands and summing up the power within each

TABLE I. Articulation index weights (ANSI, 1997) used in the implementation of the proposed measure.

Band	Center frequencies (Hz)	Weight
1	50.0000	0.0064
2	120.000	0.0154
3	190.000	0.0240
4	260.000	0.0373
5	330.000	0.0803
6	400.000	0.0978
7	470.000	0.0982
8	540.000	0.0809
9	617.372	0.0690
10	703.378	0.0608
11	798.717	0.0529
12	904.128	0.0473
13	1020.38	0.0440
14	1148.30	0.0440
15	1288.72	0.0470
16	1442.54	0.0489
17	1610.70	0.0486
18	1794.16	0.0491
19	1993.93	0.0492
20	2211.08	0.0500
21	2446.71	0.0538
22	2701.97	0.0551
23	2978.04	0.0545
24	3276.17	0.0508
25	3597.63	0.0449

band. The center frequencies of the 25 critical-band like bands are given in Table I. Equations (4)–(6) were then used to compute the fAI values for each frame, which were in turn averaged across all frames to produce a single value for each sentence. For BIFs, we considered the ANSI weights[1] (ANSI, 1997) (see Table I) as well as the signal-dependent weighting function proposed in our prior study (Ma *et al.*, 2009):

$$W_k = (S_k)^p \tag{7}$$

where S_k denotes the target signal (prior to mixing) raised to the power p. To assess the effect of the $\mathrm{SNR_L}$ value [Eq. (5)] on performance, we varied its value from 0 dB to 14 dB.

III. INTELLIGIBILITY DATA

Data taken from the intelligibility evaluation of noise-corrupted speech processed through eight different noise-suppression algorithms and presented to normal-hearing listeners were used in the present study (Hu and Loizou, 2007). IEEE sentences (IEEE, 1969) were used as test material. The masker signals were taken from the AURORA database (Hirsch *et al.*, 2000) and included the following real-world recordings from different places: babble, car, street, and train. The maskers were added to the speech signals at SNRs of 0 and 5 dB. A total of 40 native speakers of American English were recruited for the sentence intelligibility tests.

The intelligibility study by Hu and Loizou (2007) produced a total of 72 noisy conditions including the noise-corrupted (unprocessed) conditions. The 72 conditions included non-linear distortions introduced by 8 different noise-suppression algorithms operating at two SNR levels (0 and 5 dB) in four types of real-world environments (babble, car, street, and train). Eight different gain functions were thus used to process the noisy speech (see description in Hu and Loizou, 2007) for each SNR condition. The simplified spectral-subtractive gain function depicted in Eq. (1) was only one of the 8 different gain functions tested. The intelligibility scores obtained in the 72 conditions were used in the present study to evaluate the predictive power of the proposed fAI measure.

TABLE II. Correlation coefficients, r, and prediction error (σ_e) between sentence recognition scores and the proposed fAI measure for two sets of band-importance functions (BIFs). The corresponding correlations obtained with the CSII measure are also shown for comparison.

Measure	BIF	$\mathrm{SNR_L}$ (dB)	r	σ_e
CSII	ANSI	–	0.82	0.10
fAI	ANSI	0	0.82	0.10
fAI	ANSI	6	0.86	0.09
fAI	ANSI	9	0.86	0.09
fAI	ANSI	11	0.86	0.09
fAI	ANSI	14	0.85	0.09
CSII	Eq. (7), p = 4	–	0.86	0.09
fAI	Eq. (7), p = 2	0	0.80	0.11
fAI	Eq. (7), p = 2	6	0.87	0.08
fAI	Eq. (7), p = 2	9	0.90	0.08
fAI	Eq. (7), p = 2	11	0.90	0.08
fAI	Eq. (7), p = 2	14	0.89	0.08

IV. EVALUATION OF PROPOSED INTELLIGIBILITY MEASURE

The average intelligibility scores obtained by normal-hearing listeners in 72 different noisy conditions (Hu and Loizou, 2007) were subjected to correlation analysis with the corresponding mean values obtained with the proposed fAI measure. The resulting correlation coefficients (r) and prediction errors (σ_e) are given in Table II for different values of $\mathrm{SNR_L}$ and different BIFs. For comparative purposes, we also tabulate the corresponding correlation coefficients obtained with the CSII measure for the same data and same conditions (Ma *et al.*, 2009).

As shown in Table II, the performance of the proposed fAI measure was clearly influenced by the choice of BIF and $\mathrm{SNR_L}$ value. A high correlation ($r = 0.86$) was obtained with the fAI measure when the ANSI weights were used and $\mathrm{SNR_L} = 11$ dB. The corresponding correlation obtained with the CSII measure was 0.82. Even higher correlation ($r = 0.9$) was obtained with fAI when the signal-dependent BIF [Eq. (7)] were used and $\mathrm{SNR_L} = 9$ dB or $\mathrm{SNR_L} = 11$ dB. Figure 4 shows the scatter plot of fAI values and intelligibility scores. A logistic-type function of the form (Fletcher and Galt, 1950; Amlani *et al.*, 2002):

$$I = \left(1 - 10^{-x \cdot P/Q}\right)^2$$

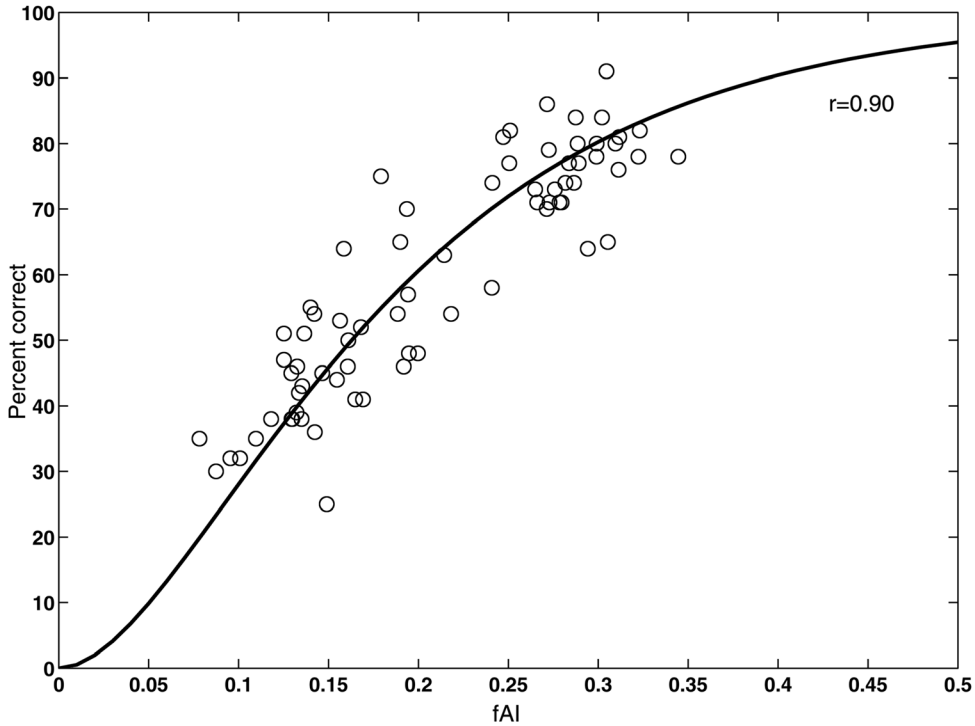

FIG. 4. Scatter plot of speech intelligibility scores and predicted fAI values for 72 noisy conditions involving noise-suppressed speech in four different masker conditions (babble, car, train and street interferences) and two SNR levels.

was used for the fitting of the fAI values, where I is the subject's intelligibility score (in proportion correct), $x = \text{fAI}$, $P = 27.5$, and $Q = 8.4$. Based on the above fitting (transfer) function, high intelligibility ($> 90\%$ correct) is predicted for fAI values greater than 0.5. Note that a similar prediction is obtained when using the AI transfer function (Fletcher and Galt, 1950, Fig. 7) for sentences. The slope of the derived transfer function is, however, shallower (for values smaller than 0.5) than the sentence AI transfer function but matches closely to the transfer function of the Modified Rhyme Test (MRT) words (Amlani *et al.*, 2002, Fig. 4). Figure 5 plots the observed scores (in percentage) against the predicted scores for all 72 conditions tested ($r = 0.9$).

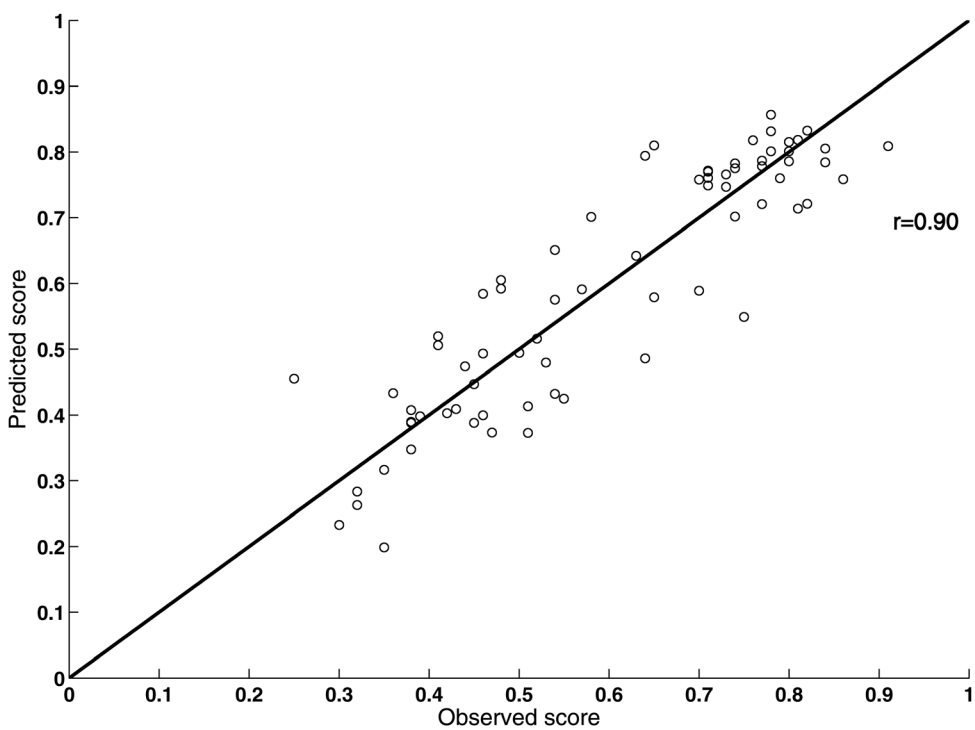

FIG. 5. Scatter plot of observed intelligibility scores (expressed in percentage) and predicted scores for the 72 noisy conditions tested.

When tested on the same dataset, the corresponding correlation with the CSII measure, implemented using the signal-dependent BIFs, was found to be 0.86. Overall, the proposed fAI measure outperformed the CSII measure in predicting the intelligibility of non-linear processed speech.

The choice of SNR_L in Eq. (5) had a clear influence in the performance of the proposed fAI measure. We believe that this was because the value of SNR_L affects how dominant are the target components (of the output envelopes) included in the computation of the fAI measure. When SNR_L is large, only dominantly large target components are allowed to enter in the computation of the fSNR values, and the "optimum" value for our test set seems to be around 9–11 dB. That is, when SNR_L is large, we have $S_T \gg S_M$ [Eq. (2)], and the output band SNR truly reflects the effect of non-linear processing on the target component. When SNR_L exceeds a certain value (11 dB in our case), however, fewer bands enter the computation of the fAI value rendering the computation of the fAI somewhat unreliable (owing to the small number of samples considered), particularly in the low input SNR conditions. As a result, the contribution of the non-linear processing in most bands is not accounted for in the computation of fAI. This explains the slight decline in performance when $SNR_L > 11$ dB. At the other end, when SNR_L is near 0 dB, the target component, although larger than the masker component, is not dominant but rather comparable in magnitude. Consequently, the output band SNR is slightly overestimated. Overall, high correlations were consistently obtained for SNR_L values in the range of 6–11 dB.

Performance improved dramatically when the signal-dependent BIFs were used [Eq. (7)]. As argued by Ma *et al.* (2009), the choice and importance of BIFs becomes more critical in situations wherein short-term processing is involved and fluctuating maskers are used, as was the case in the present study. In contrast, in the implementation of the conventional AI measure it suffices to use a single, albeit material dependent, BIF since the measure is computed based on the (single) long-term averaged spectra of the target and masker signals. As demonstrated in our prior study (Ma *et al.*, 2009), the best and simplest BIF to use is the target signal spectrum itself [Eq. (7)]. This is consistent with the notion that, for vowels, more weight should be placed on the bands containing spectral peaks as those convey information about the formants. Similarly applicable for consonants, more weight should be placed on the spectral peaks as those convey information about place of articulation (e.g., Liberman *et al.*, 1952). The power exponent p used in Eq. (7) controls the weight placed on spectral peaks and/or spectral valleys and can be optimized for different speech materials (Ma *et al.*, 2009).

The present study focused on modifying the basic form of the AI measure to account for non-linear processing. Compared to the SII index implementation (ANSI, 1997), however, which incorporates upward-spread of masking effects and level distortion factors, the fAI implementation was rather simplistic. Yet, despite these limitations, the proposed fAI measure performed quite well ($r = 0.9$). Further experiments are needed to examine whether additional improvements in performance can be obtained if upward-spread of masking effects or level distortion effects are incorporated.

From the scatter plot shown in Fig. 4 we observe that the average fAI values did not exceed the value of 0.4. It should be noted, however, that the individual short-term values of fAI sometimes exceed 0.4, but the average is biased toward lower values due to the extremely low fAI values obtained during unvoiced segments, which also happen to be the low SNR segments (see Fig. 3). This implies that on the average only 10%–40% of the input SNR is preserved or transmitted by most noise-reduction algorithms, at least for algorithms operating at the two SNR levels examined (0 and 5 dB). This average is based on values accumulated from all phonetic segments across the utterance, including vowels and consonants. The proportion of input SNR preserved for low-energy consonants was extremely low (see example in Fig. 3) compared to that of vowels. This was not surprising since the low-energy consonants are masked more easily by background noise than the high-energy vowels (Parikh and Loizou, 2005) and most noise-reduction algorithms perform poorly in segments containing consonants. It is thus possible that if noise reduction algorithms could somehow preserve or maintain a larger portion of the input SNR during the low-energy consonant segments, then improvement in speech intelligibility might be noted. Further research work is warranted to examine that.

V. CONCLUSIONS

The present study proposed a simple modification to the AI measure to account for non-linear processing (e.g., noise reduction) in the presence of additive noise. The modification was based on the following two observations. First, the input SNR in a specific band cannot be improved following any form of non-linear processing [Eq. (3)]. Second, the output (or processed) envelope $\hat{S}$ reflects more reliably the effect of suppression or non-linear processing on the target envelope when $\hat{S} < S$ than when $\hat{S} > S$. Taking the above two observations into account, a new measure (fAI) was proposed [Eq. (6)]. Only bands with input band SNR exceeding a certain threshold (denoted as SNR_L) were included in the computation [see Eq. (5)]. This ensured that the target component of the output envelope was always larger than the masker component, thereby providing a more reliable estimate of the output or effective SNR [Eq. (4)]. The proposed fAI measure was evaluated with speech intelligibility scores collected in our prior study (Hu and Loizou, 2007) involving noise-suppressed speech in four different masker conditions and two SNR levels. High correlation ($r = 0.9$) was obtained with the proposed measure when $SNR_L = 11$ dB and signal-dependent band-importance functions (Ma *et al.*, 2009) were used. In comparison, the highest correlation obtained with the CSII measure (Kates and Arehart, 2005) was $r = 0.86$ when tested in the same conditions.

ACKNOWLEDGMENTS

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

APPENDIX

In this appendix, we derive the upper bounds on the SNR when the processed signal $\hat{S}$ (following noise-reduction) is larger than the input (clean) target signal S. More precisely, we consider the scenarios when $\hat{S} > S$ or when $\hat{S} > 2{\cdot}S$. These bounds are important as they directly affect the definition of the output band SNR. Following the application of the gain function G_k on the noisy envelopes Y_k (Fig. 1), we can express the squared output (or processed) envelope $\hat{S}$ as follows:

$$\hat{S}^2 = G^2\left(S^2 + N^2\right), \tag{A1}$$

where S denotes the clean signal envelope and N indicates the masker envelope. After dividing both sides by S^2, we have:

$$\frac{\hat{S}^2}{S^2} = G^2\left(1 + \frac{1}{SNR}\right), \tag{A2}$$

where $SNR = S^2/N^2$. If $\hat{S} > S$, or equivalently $\hat{S}/S > 1$, we have the following inequality:

$$G^2\left(1 + \frac{1}{\mathrm{SNR}}\right) > 1$$
$$\mathrm{SNR} < \frac{G^2}{1 - G^2}. \tag{A3}$$

From the above, it is easy to show that for a specific range of the gain function, i.e., $0 \le G \le 0.707$, we have $\mathrm{SNR} < 1$ or equivalently $\mathrm{SNR_{dB}} < 0$ dB. When $G > 0.7$, the SNR (in dB) will be positive. Hence, for a large range of gain values, the SNR will always be negative, consistent with the histogram shown in Fig. 2.

In the scenario where $\hat{S} > 2 \cdot S$, or equivalently $\hat{S}/S > 2$, we have the following inequality:

$$\mathrm{SNR} < \frac{G^2}{2 - G^2}, \tag{A4}$$

and given that the gain function is typically bounded by 1, i.e., $0 \le G \le 1$, we have:

$$\mathrm{SNR} < 1, \tag{A5}$$

suggesting that when $\hat{S} > 2 \cdot S$, the input SNR (in dB) is always negative. This means that the masker component of $\hat{S}$ will always be larger than the target component when $\hat{S} > 2 \cdot S$.

[1]The ANSI (1997) band-importance functions (BIFs) were interpolated to account for the frequency spacing used (Table I). Due to the interpolation, the BIF values did not add up to one. This however did not affect the performance of the proposed intelligibility measure, as the measure was normalized by the sum of the BIF values [see Eq. (6)].

Amlani, A., Punch, J., and Ching, T. (**2002**). "Methods and applications of the audibility index in hearing aid selections and fitting," Trends Ampl. **6**, 81–129.

ANSI S3.5-1997 (**1997**). "Methods for calculation of the speech intelligibility index," (American National Standards Institute, NY).

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

Boll, S. F. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Proc., **27**(2), 113–120.

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.

Dubbelboer, F., and Houtgast, T. (**2008**). "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," J. Acoust. Soc. Am. **124**, 3937–3946.

Fletcher, H., and Galt, R. H. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Goldsworthy, R., and Greenberg, J. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.

Hirsch, H., and Pearce, D. (**2000**). "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA Tutorial and Research Workshop ASR2000* (ISCA, Paris, France).

Hohmann, V., and Kollmeier, B. (**1995**). "The effect of multichannel dynamic compression on speech intelligibility," J. Acoust. Soc. Am. **97**, 1191–1195.

Houtgast, T., and Steeneken, H. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.

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

Hu, Y., and Loizou, P. (**2008**). "A new sound coding strategy for suppressing noise in cochlear implants," J. Acoust. Soc. Am., **124**(1), 498–509.

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

Kates, J. (**1987**). "The short-time articulation index," J. Rehab. Res. Develop. **24**, 271–276.

Kates, J. (**1992**). "On using coherence to measure distortion in hearing aids," J. Acoust. Soc. Am. **91**, 2236–2244.

Kates, J., and Arehart, K. (**2005**). "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**, 2224–2237.

Kates, J. (**2010**). "Understanding compression: Modeling the effects of dynamic-range compression in hearing aids," Int. J. Audiol. **49**(6), 395–409.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**(3), 1486–1494.

Kryter, K. D. (**1962a**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689–1697.

Kryter, K. D. (**1962b**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698–1706.

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

Liberman, A., Delattre, P., and Cooper, F. "The role of selected stimulus variables in the perception of unvoiced stop consonants," Am. J. Psychol. **65**, 497–516 (**1952**).

Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, FL), pp. 560–570.

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

Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**). "Evaluation of a noise reduction method—Comparison of observed scores and scores predicted from STI," Scand. Audiol. Suppl. **38**, 50–55.

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

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

Pavlovic, C. V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, 413–422.

Rhebergen, K. S., and Versfeld, N. J. (**2005**). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181–2192.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.

Stelmachowicz, P., Dalzell, S., Peterson, D., Kopun, J., Lewis, D., and Hoover, B. E. (**1998**). "A comparison of threshold-based fitting strategies for nonlinear hearing aids," Ear Hear. **19**, 131–138.

Steeneken, H., and Houtgast, T. (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2010**). "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *Proc. IEEE Intern. Conf. Acoust. Speech Signal Proc.* (IEEE, Dallas, TX), pp. 4214–4217.

Van Buuren, R., Festen, J., and Houtgast, T. (**1999**). "Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality," J. Acoust. Soc. Am. **105**, 2903–2913.