Analysis of a simplified normalized covariance measure based on binary weighting functions for predicting the intelligibility of noise-suppressed speech

Fei Chen and Philipos C. Loizou[a)]
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083

(Received 23 June 2010; revised 21 September 2010; accepted 23 September 2010)

The normalized covariance measure (NCM) has been shown previously to predict reliably the intelligibility of noise-suppressed speech containing non-linear distortions. This study analyzes a simplified NCM measure that requires only a small number of bands (not necessarily contiguous) and uses simple binary (1 or 0) weighting functions. The rationale behind the use of a small number of bands is to account for the fact that the spectral information contained in contiguous or nearby bands is correlated and redundant. The modified NCM measure was evaluated with speech intelligibility scores obtained by normal-hearing listeners in 72 noisy conditions involving noise-suppressed speech corrupted by four different types of maskers (car, babble, train, and street interferences). High correlation ($r = 0.8$) was obtained with the modified NCM measure even when only one band was used. Further analysis revealed a masker-specific pattern of correlations when only one band was used, and bands with low correlation signified the corresponding envelopes that have been severely distorted by the noise-suppression algorithm and/or the masker. Correlation improved to $r = 0.84$ when only two disjoint bands (centered at 325 and 1874 Hz) were used. Even further improvements in correlation ($r = 0.85$) were obtained when three or four lower-frequency (<700 Hz) bands were selected.
 [DOI: 10.1121/1.3502473]

PACS number(s): 43.66.Ts, 43.71.Es [DKW] Pages: 3715–3723

I. INTRODUCTION

The speech transmission index (STI) (Houtgast and Steeneken, 1971; Steeneken and Houtgast, 1980) is an intelligibility metric that has been found to reliably predict the effects of reverberation as well as additive noise. The computation of the STI is based on detecting changes in signal modulation when modulated probe stimuli are transmitted through a channel of interest. The responses to probe stimuli are measured in multiple frequency bands for a range of modulation frequencies (0.63–12.7 Hz) relevant to speech. The traditional STI method has been found to perform poorly in terms of predicting the intelligibility of processed speech wherein non-linear operations (e.g., envelope compression, peak-clipping, envelope thresholding, etc.) are involved (Ludvigsen *et al.*, 1993; van Buuren *et al.*, 1999; Goldsworthy and Greenberg, 2004). A number of speech-based STI measures have been examined and analyzed by Goldsworthy and Greenberg (2004) to determine the extent to which some measures fail to predict speech intelligibility for non-linear operations. Among those, the normalized covariance measure (NCM) has been shown by Goldsworthy and Greenberg (2004) to perform better than the conventional STI method in predicting the effects of non-linear operations such as envelope thresholding or distortions introduced by spectral-subtractive algorithms. This was also confirmed by Ma *et al.* (2009) who evaluated the performance of the NCM measure with noise-suppressed speech, which generally contains various forms of non-linear distortions including the distortions introduced by spectral-subtractive algorithms. The correlation of the NCM measure with noise-suppressed speech was found to be quite high ($r = 0.89$) (Ma *et al.*, 2009).

Given the success of the NCM measure in predicting reliably the intelligibility of noise-suppressed speech containing non-linear distortions (Ma *et al.*, 2009), we consider in this study analyzing the NCM measure in terms of determining the minimum number of bands required (without compromising performance) and the shape of weighting functions to be applied to each band. We sought for a simplified NCM measure that required only a small number of bands (not necessarily contiguous) and used simple binary (1 or 0) weighting functions. The motivation behind the use of a small number of bands is that the spectral information contained in contiguous bands is correlated and redundant (Steeneken and Houtgast, 1999; Crouzet and Ainsworth, 2001). Consequently, a simple weighted summation of the individual contribution of each band (as measured by the band transmission indices) will result in an overestimation of the true information content (Steeneken and Houtgast, 1999; Musch and Buss, 2001). Steeneken and Houtgast (1999, 2002) modified the STI method by including a correction factor that accounted for the mutual dependence between adjacent octave bands. The modified STI method provided a better prediction of speech intelligibility particularly in situations with a non-contiguous frequency transfer. An iterative procedure was used by Steeneken and Houtgast (1999) to derive the optimal "redundancy-correction" factors across a number of carefully constructed conditions designed to include non-contiguous frequency transfer. A more simplified procedure is taken in the present study by examining the individual

[a)]Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu

TABLE I. The filter cut-off frequencies and AI weights (ANSI, 1997) used in the implementation of the NCM measure.

Band	Low-High cut-off frequencies (Hz)	Center frequency (Hz)	AI weight
1	300–350	325	0.0772
2	350–405	378	0.0955
3	405–466	436	0.1016
4	466–534	500	0.0908
5	534–609	571	0.0734
6	609–692	650	0.0659
7	692–784	738	0.0580
8	784–885	834	0.0500
9	885–998	942	0.0460
10	998–1123	1060	0.0440
11	1123–1261	1192	0.0445
12	1261–1413	1337	0.0482
13	1413–1583	1498	0.0488
14	1583–1770	1676	0.0488
15	1770–1977	1874	0.0493
16	1977–2207	2092	0.0491
17	2207–2461	2334	0.0520
18	2461–2743	2602	0.0549
19	2743–3055	2899	0.0555
20	3055–3400	3227	0.0514

contribution of information carried by a single or a small number of bands to speech intelligibility. Two methods are proposed for selecting a small number of bands (2–4) and the prediction power of the modified NCM measure is evaluated with the intelligibility scores collected in our prior study (Hu and Loizou, 2007). Special attention is paid to assessing the relationship between the center frequency of the selected band(s) and the effect of the masker and/or applied gain of the noise-suppression algorithm on that band. It is hypothesized that low correlations of individual bands with speech intelligibility will reflect inconsistencies in the way the noise-suppression algorithm(s) and/or masker affects (e.g., distorts) different bands (regions) of the spectrum. These inconsistencies are caused by the fact that some bands are severely distorted while other bands are effectively "cleaned" by the noise-suppression algorithm. Hence, the low correlations of individual bands (with speech intelligibility scores) might provide useful information about the regions of the spectrum and corresponding envelopes that have been heavily masked or distorted by the noise-suppression algorithm. The proposed method for band selection can thus provide diagnostic information in as far as identifying which bands are effectively suppressed (or not) by noise-reduction algorithms.

II. THE NORMALIZED COVARIANCE MEASURE (NCM)

The NCM measure is computed as follows (Holube and Kollmeier, 1996; Goldsworthy and Greenberg, 2004). The stimuli are first bandpass filtered into N bands spanning the signal bandwidth (300–3400 Hz in this study). Table I shows the filter cut-off frequencies used to decompose the signal into $N = 20$ bands. The envelope of each band is computed using the Hilbert transform and then downsampled to $2f_{\text{cut}}$ Hz, thereby limiting the envelope modulation rate to f_{cut} Hz ($f_{\text{cut}} = 12.5$ Hz in this study). An anti-aliasing low-pass filter was used prior to downsampling to eliminate aliasing artifacts. Let $x_i(t)$ and $y_i(t)$ be the downsampled envelope in the ith band of the clean signal and the processed signal, respectively. The normalized covariance in the ith frequency band is computed as

$$\rho_i = \frac{\sum_t (x_i(t) - \mu_i)(y_i(t) - v_i)}{\sqrt{\sum_t (x_i(t) - \mu_i)^2}\sqrt{\sum_t (y_i(t) - v_i)^2}}, \tag{1}$$

where μ_i and v_i are the mean values of the $x_i(t)$ and $y_i(t)$, respectively. The signal-to-noise ratio (SNR) in each band is computed as

$$\text{SNR}_i = 10\log_{10}\left(\frac{\rho_i^2}{1-\rho_i^2}\right), \tag{2}$$

and subsequently limited to the range of [−15,15] dB [as done in the computation of the SII measure (ANSI, 1997)]. The transmission index (TI) in each band is computed by linearly mapping the SNR values between 0 and 1 using the following equation:

$$\text{TI}_i = \frac{\text{SNR}_i + 15}{30}. \tag{3}$$

Finally, the transmission indices are averaged across all frequency bands to produce the NCM index:

$$\text{NCM} = \frac{\sum_{i=1}^{N} \text{TI}_i \times w_i}{\sum_{i=1}^{N} w_i}, \tag{4}$$

where $\mathbf{W} = (w_1 \ \ldots \ w_i \ \ldots \ w_N)^T$ denotes the weight vector applied to the transmission-index TI_i of N bands.

There are several methods for choosing the weight vector $\mathbf{W}$ in Eq. (4), with the most common being the articulation index (AI) weights (ANSI, 1997). Ma *et al.* (2009) proposed the use of signal-dependent weighting vectors, and more specifically, they proposed the following:

$$W_i^{(1)} = \left(\sum_t x_i^2(t)\right)^p, \tag{5}$$

$$W_i^{(2)} = \left(\sum_t (\max[x_i(t) - d_i(t), 0])^2\right)^p, \tag{6}$$

where $d_i(t)$ denotes the downsampled envelope of the scaled masker signal in the time domain (the power exponent p was varied from 0.12 to 1.5 in this study). The motivation behind the use of Eq. (5) is to place weight to each TI value in proportion to the signal energy in each band, while the motivation behind the use of Eq. (6) is to place weight to each TI value in proportion to the excess masked signal.

In the present study, we consider using a more simplified method for choosing the weights w_i for each band. More precisely, we investigate the use of a binary weight vector $\mathbf{W}_M$, where w_i in $\mathbf{W}_M$ is either set to 1 or 0, and M ($M < 20$)

is the total number of bands used with unity weight ($w_i = 1$). The weights for the remaining ($20 - M$) bands are set to zero. As mentioned in Sec. I, the rationale for choosing a subset of the N bands ($N = 20$ in our study) is that the spectral information in adjacent or nearby bands is highly correlated and therefore redundant. This could in turn diminish the performance of the NCM measure. By using binary weights in Eq. (4), we hope to answer a number of interesting questions: (1) What is the minimum number of bands needed to obtain good intelligibility prediction with the NCM measure? (2) How should these bands be chosen? (3) Does the answer to the previous two questions depend on the spectral characteristics of the masker? (4) Do the binary weights reveal a specific pattern for each masker signifying perhaps weaknesses/limitations of the noise suppression algorithms in terms of effectively suppressing background noise?

III. SPEECH INTELLIGIBILITY DATA

Data taken from the intelligibility evaluation of noise-corrupted speech processed through eight different noise-suppression algorithms by 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 and Pearce, 2000), and included the following real-world recordings from different places: Babble, car, street, and train. The maskers were artificially added to the speech signals at SNRs of 0 and 5 dB. A total of 40 normal-hearing listeners participated in the sentence intelligibility tests (Hu and Loizou, 2007). The intelligibility scores obtained from the normal-hearing listeners in a total of 72 conditions were used in the present study to evaluate the predictive power of the NCM measure implemented using binary weights.

IV. RESULTS AND DISCUSSION

The Pearson's correlation coefficient (r) was used to assess the correlation of the NCM measure with the speech intelligibility scores. The performance of the NCM measure implemented using a single band or multiple bands was examined and analyzed next.

A. Using a single band for computing the NCM measure

Figure 1 shows the correlation coefficients obtained when using a $\mathbf{W}_1$ binary vector, i.e., $\mathbf{W}_1 = (00 \ldots 1 \ldots 00)^T$, where the ith band has a weight of 1 and the remaining bands have a weight of 0. That is, the weight w_i in Eq. (4) was set to 1 for the ith band while the weights for the remaining 19 bands were set to zero. This was repeated for all 20 bands. Figure 1 reports the correlations obtained when each of the 20 bands was used in the computation of the NCM measure. The first data point in Fig. 1 indicates the correlation obtained when only band 1 was used (the remaining 19 bands were not used), the second data point in Fig. 1 indicates the correlation obtained when only band 2 was used (the remaining 19 bands were not used), and so forth. The resulting correlation coefficients ranged from a low of 0.3 (band 17) to a high of 0.8 (band 1). The baseline correlation coefficient obtained when using the ANSI weights and all 20 bands was found to be 0.82 (Ma *et al.*, 2009). Hence, the surprising finding from Fig. 1 is that high correlation can be obtained with the NCM measure even with only one band (e.g., band 1).

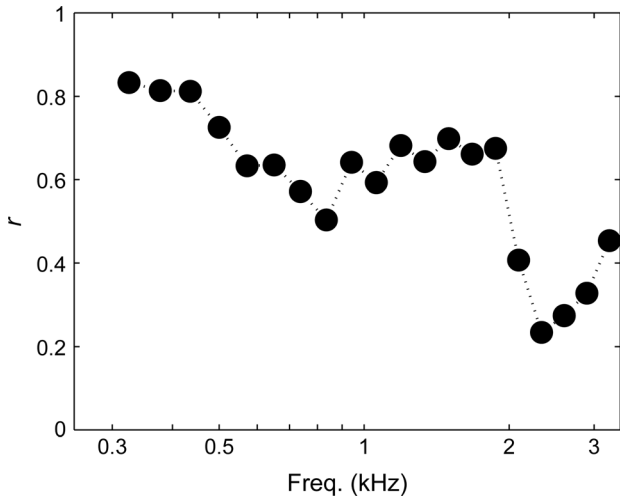

FIG. 1. The individual correlation coefficients r obtained using the modified NCM measure when only one band is used at a time.

As shown in Fig. 1, some bands exhibited low correlations with intelligibility scores while others exhibited relatively high correlation. The reasons for that were unclear at first; hence, further analysis was conducted to determine the reason. In particular, we analyzed the correlations separately for each of the four maskers. The correlations were computed based on 18 noisy conditions for each type of masker. Figure 2 shows the correlation coefficients obtained using a $\mathbf{W}_1$ binary vector for the four maskers tested, i.e., babble, car, street, and train interferences. As can be seen from Fig. 2, each masker has its own correlation pattern, which we refer to as the r-pattern. The r-pattern for babble is relatively flat, while that of train has two significant dips at bands 5 and 17. For the street interference, the lowest correlation was close to zero in band 17. The bands with low correlation differed among the car, street, and train interferences. Low correlation was obtained for the band centered near 834 Hz for the car interference, at 2334 Hz for the street interference, and near 571 and 2334 Hz for the train interference.

Figure 2 raises the question: What is the significance of the r-pattern and, perhaps more importantly, can we use these r-patterns to determine how effective the noise suppression algorithms are in reducing background noise? We believe that the frequency location of the dips in the r-pattern identifies inconsistencies (or perhaps differences) in the way the noise-suppression algorithm(s) affects (e.g., distorts) different bands (regions) of the spectrum. These inconsistencies are caused by the fact that some bands are severely distorted while other bands are effectively cleaned by the noise-suppression algorithm. In the r-pattern (Fig. 2), bands with high correlation indicate consistent performance with overall intelligibility scores, and one can view those bands as being representative of overall performance. As such, when the TI (or equivalently the effective SNR) is high in those bands, intelligibility is high, and when the TI is low in those bands, intelligibility is

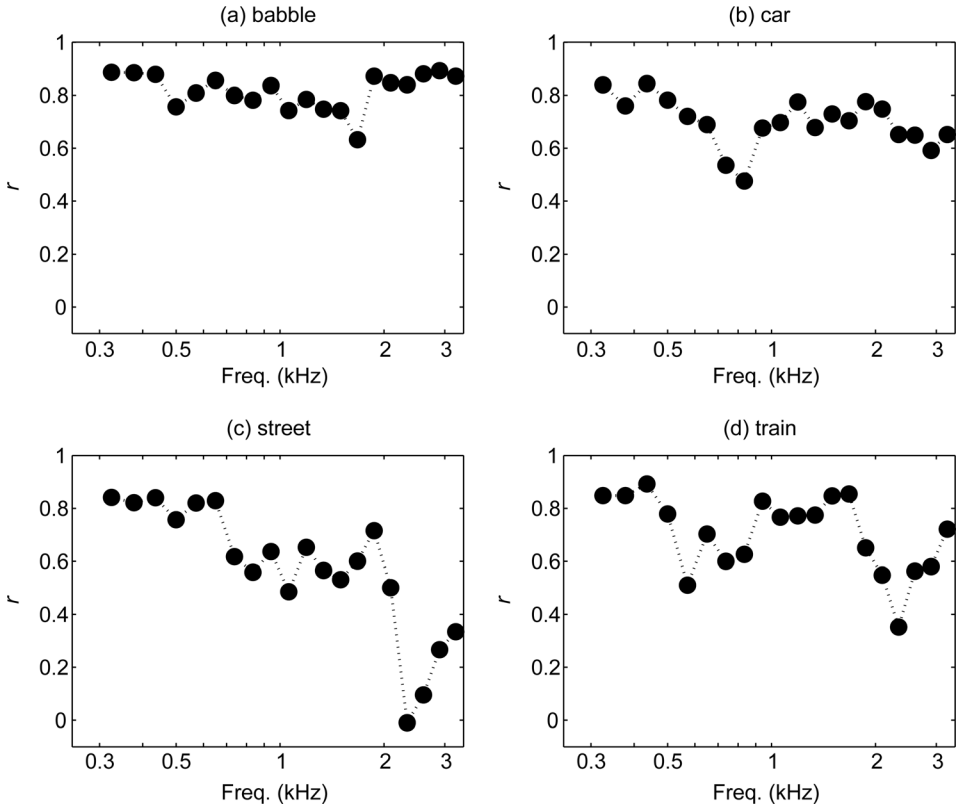

FIG. 2. The individual correlation coefficients (r-pattern) obtained for the four maskers tested using the modified NCM measure when only one band is used at a time.

low. In contrast, bands with low correlation are likely affected differently by the noise-suppression algorithm (compared to the other bands), and in a way that is inconsistent with the overall intelligibility score. Consider, for instance, a hypothetical scenario wherein a noise-suppression algorithm effectively suppresses the background noise in all bands except the last high-frequency band, which is severely distorted. In such a case, intelligibility will be mildly affected (since the majority of the bands were not distorted) and the correlations of the majority of the bands will be high. In contrast, the correlation with the single high-frequency band will be low, since the TI for that band will likely be low (due to the presence of severe distortion) and thus inconsistent with the high intelligibility scores. To illustrate this, Fig. 3 shows an example TI pattern for speech processed in the street-masker conditions. Bands 17–20 have low TI values (<0.4), suggesting that they have been distorted or not effectively enhanced, while most of the lower-frequency bands have comparatively higher TI values (>0.7). The TI values in bands 17–20 are low relative to the TI values in bands 1–16, and this difference caused the low correlations for bands 17–20 [see Fig. 2(c)]. This is so because the low TI values in bands 17–20 were not consistent with the overall intelligibility scores, in that subjects were able to recognize the sentences, despite the presence of a few distorted or noise-masked bands in the high frequencies (Hu and Loizou, 2007).

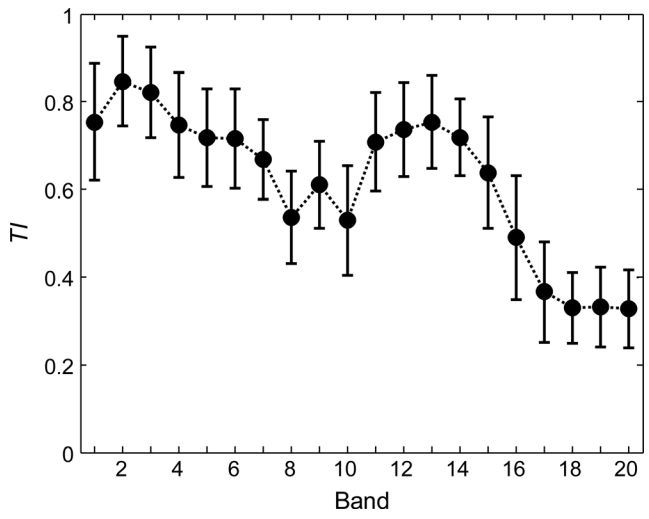

FIG. 3. Mean TI values obtained for each band using data from all street-masker conditions. Error bars indicate standard deviations.

In summary, we believe that the frequency location of the dips in the r-pattern effectively signifies the corresponding envelopes that have been severely distorted by the noise-suppression algorithm and/or the masker. In principle, a low correlation in the r-pattern could also indicate the ability of the noise reduction algorithm to effectively suppress the background noise in a particular band(s) (assuming that the remaining bands are severely distorted), however, we did not find that to be the case in our study, at least for the class of noise reduction algorithms tested. Based on the outcomes of our study, we thus believe that the low correlation in the r-pattern must be due to the poor ability of the noise reduction algorithm to suppress background noise in a specific band. To demonstrate this, we show in Fig. 4 spectrograms of four sentences, which were corrupted by four types of maskers at 0 dB SNR and processed by the spectral subtraction algorithm based on reduced-delay convolution (RDC)[1] (Gustafsson *et al.*, 2001). Figure 4 shows the spectrograms

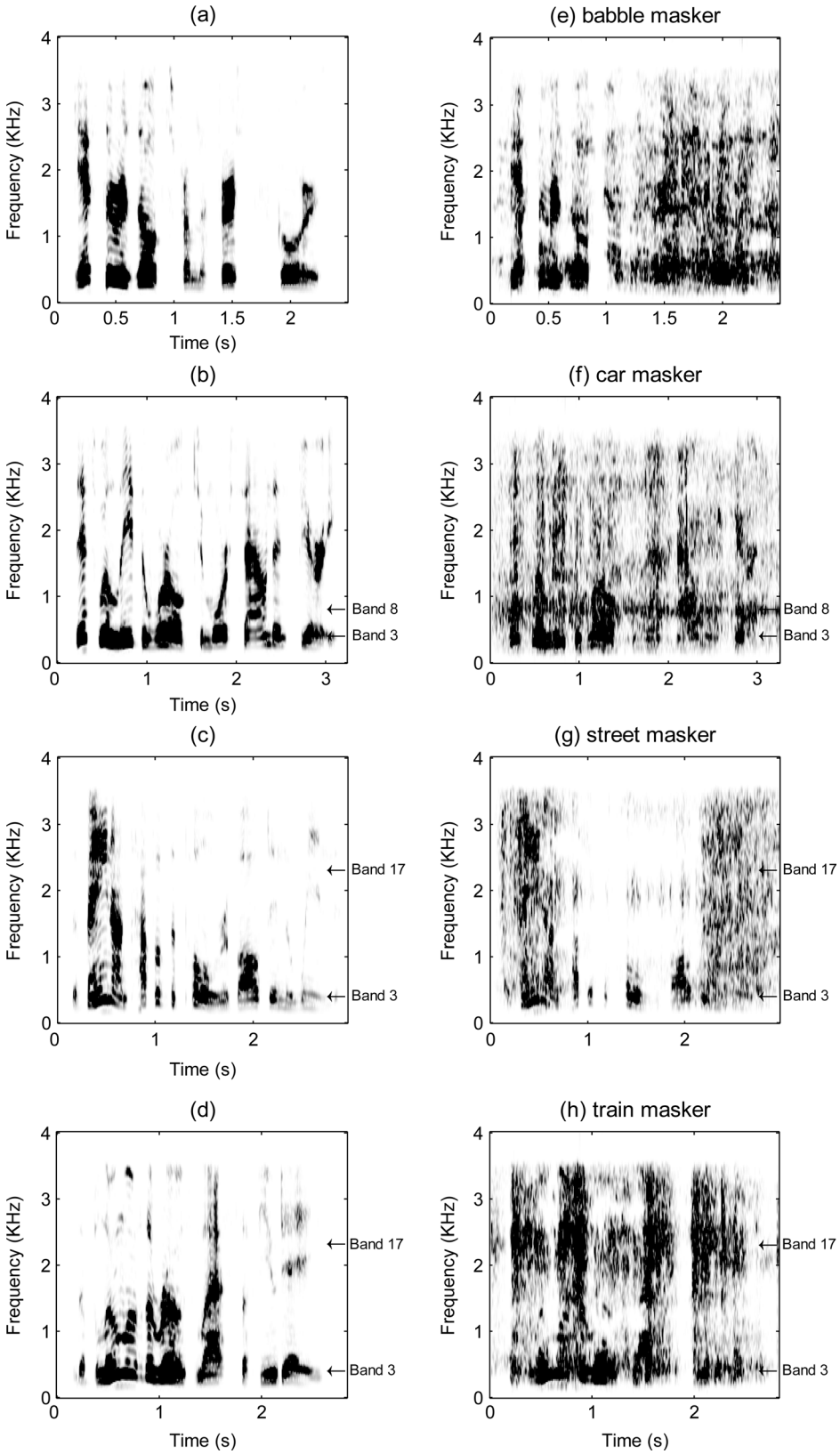

FIG. 4. The spectrograms of sentences in quiet are shown in (a), (b), (c), and (d), and the corresponding processed sentences (by the RDC noise-reduction algorithm) in four types of maskers are shown in (e), (f), (g), and (h). The sentences were originally corrupted at 0 dB SNR. Arrows point to regions (bands) of the spectrum that have been either severely distorted [e.g., band 17 in (g)] or not sufficiently enhanced by the noise-suppression algorithm [e.g., band 8 in (f)]. The center frequencies of the indicated bands are given in Table I.

of four sentences in quiet [Figs. 4(a)–4(d)] and the spectrograms of processed (by the RDC algorithm) sentences originally corrupted by four types of maskers at 0 dB SNR [Figs. 4(e)–4(h)]. As shown in Fig. 2(b), the correlation obtained for the car interference is low ($r = 0.40$) for the eighth band and high for the third band ($r = 0.83$). Accordingly, it is observed in Fig. 4(f) that the spectral region around band 8 is still heavily corrupted even after noise-suppression, while the region centered around band 3 in Fig. 4(f) is relatively unaffected and close to that of the clean stimulus in Fig. 4(b).

The differential effects of distortion introduced in different bands [e.g., bands 3 and 8 in Fig. 4(f)] by a noise-

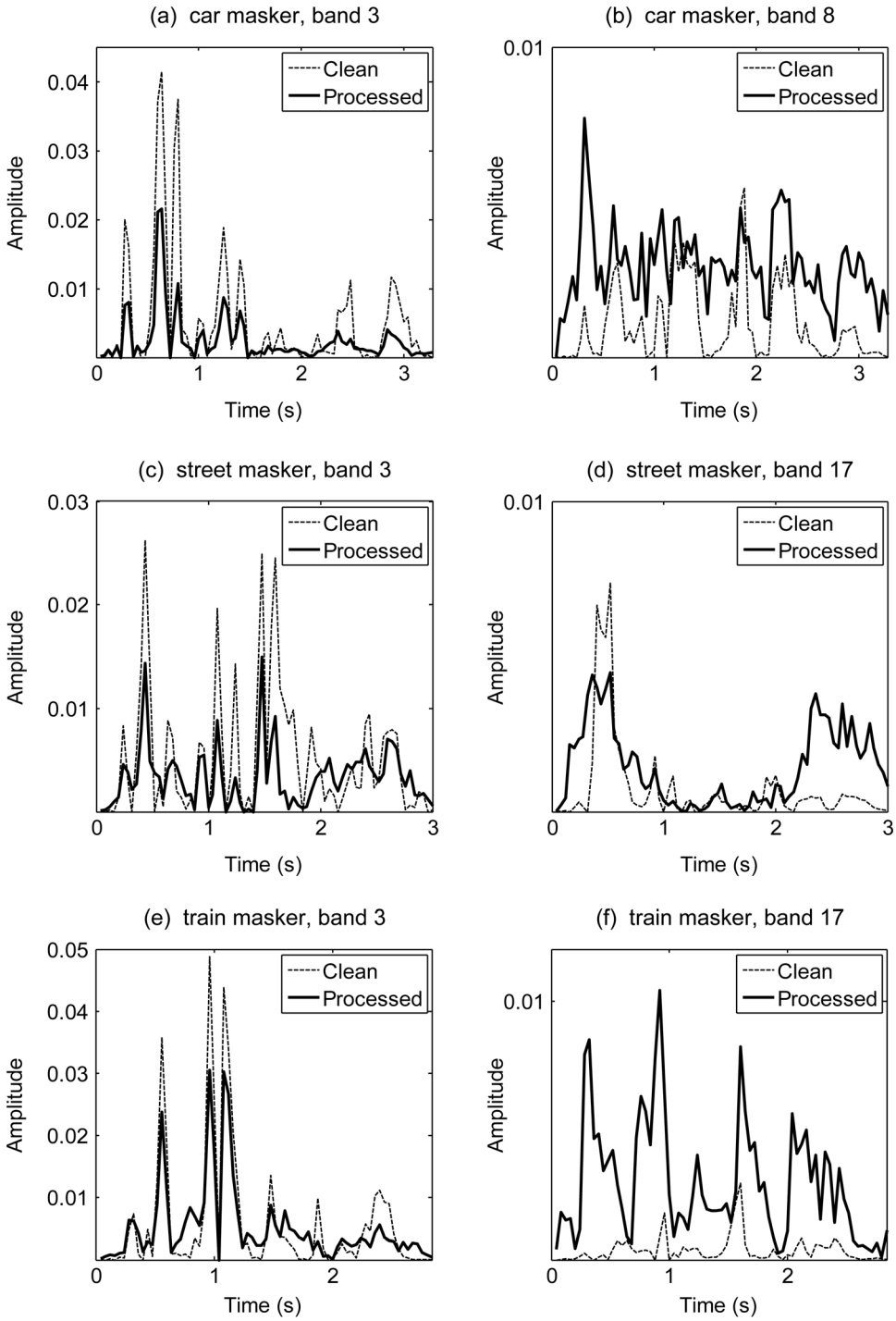

FIG. 5. Envelopes, extracted from the indicated bands, of sentences in quiet and the corresponding envelopes of noise-suppressed (by RDC algorithm) sentences originally corrupted by three types of maskers at 0 dB SNR. The resulting correlations with speech intelligibility scores of the bands shown in (a), (c), and (e) are high (refer to Fig. 2), while those in (b), (d), and (f) are low. The center frequencies of the indicated bands are given in Table I.

suppression algorithm (RDC algorithm) is also demonstrated in Fig. 5, which shows the envelopes of the clean and noise-suppressed sentences for bands with high and low correlations in the *r*-pattern. The envelopes in Figs. 5(a) and 5(b), 5(c) and 5(d), and 5(e) and 5(f) are corrupted by car, street, and train maskers, respectively, at 0 dB SNR. The output envelopes in Figs. 5(a) and 5(b) show that the background noise was suppressed more effectively for band 3 than for band 8. Similarly, the correlation coefficient is low for the 17th band and high for the 3rd band for the street and train interferences in Figs. 2(c) and 2(d). The spectrograms in Figs. 4(g) and 4(h) and the envelopes in Figs. 5(c)–5(f) both suggest that the noise-suppression algorithm performs much better for band 3 than for band 17 for the street and train interferences. Taking these observations together, we believe that the band with low *r* in the *r*-pattern in Fig. 2 is also the band in which the noise-suppression algorithm does not perform well in terms of effectively suppressing the background noise in that band or the band that is severely distorted by the noise-suppression algorithm. The spectrogram in Fig. 4(e) for the babble interference demonstrates that there is still much residual noise for *all* 20 bands after noise-suppression, which might account for the flat *r*-pattern of the babble masker in Fig. 2(a). Alternatively, we can say that all bands were affected uniformly by the noise-suppression algorithms for speech corrupted by babble, thereby yielding a flat *r*-pattern. In brief, the *r*-pattern obtained for each

TABLE II. The correlation coefficients r obtained in the various masker conditions by the NCM measure based on AI weights and weights determined by the masker-specific r-patterns (Fig. 2).

Masker	AI weights	r-pattern weights
Babble	0.91	0.91
Car	0.82	0.82
Street	0.78	0.81
Train	0.85	0.86

masker is quite informative and to some extent it is indicative of how effective or ineffective noise reduction algorithms are in suppressing noise in specific bands. This information is obtained indirectly by observing the bands with low correlation in the r-pattern. Consequently, the r-patterns can be diagnostic in terms of identifying weaknesses of noise reduction algorithms in suppressing specific types of background noise, and can thus be used to re-design and improve existing noise reduction algorithms.

Given that the r-pattern is different for each masker [see Figs. 2(a)–2(d)], we wanted to examine whether we could use it as a masker-dependent weighting function in Eq. (4) for better prediction of speech intelligibility. We thus replaced the weights w_i in Eq. (4) with the corresponding correlations given by the r-pattern (Fig. 2). Table II compares the correlations obtained with AI weights (ANSI, 1997) and weights determined from the individual r-patterns of each masker (Fig. 2). As can be seen, the prediction was improved for certain types of maskers (i.e., the street and train interferences in Table III) when using the corresponding r-patterns as weighting functions. The baseline correlation coefficient for the street noise conditions, for instance, improved from $r = 0.78$ to $r = 81$. This result suggests the possible benefit of using the r-pattern as masker-dependent weighting function to predict speech intelligibility of noise-suppressed speech.

B. Selecting multiple bands for computing the NCM measure

Figure 1 showed that high correlation can be maintained even when only one band is used in the implementation of the NCM measure. The correlation with one band was nearly as high (0.8 vs 0.82) as that obtained with 20 bands (ANSI weights). Next, we considered two different methods for selecting M out of 20 bands for implementing the NCM measure. In the first method, the r-pattern was divided into M non-overlapping sub-bands, and only the bands with the highest correlations in each sub-band were considered in the computation of the NCM measure. When $M = 3$, for instance, the following three sub-bands were used: 300–1000 Hz, 1000–2000 Hz, 2000–3400 Hz. Only bands with the highest correlations in each of the M sub-bands were incorporated in the computation of the NCM measure [Eq. (4)]. This method ensures that the selected bands are not contiguous, unless they happen to fall at the edges of two adjacent sub-bands. In the second method, the M bands with the highest correlation in the r-pattern were selected, independent of their frequency location in the spectrum. As such, the selected bands might be either contiguous or non-contiguous. The M selected bands were finally used to construct the new binary weight vector $\mathbf{W}_M$ in Eq. (4).

To assess the robustness of selecting M out of 20 bands for the implementation of the NCM measure, we used a cross-validation approach. More precisely, the dataset (i.e., 72 conditions) was divided into a training set that was used to obtain the binary weight vector $\mathbf{W}_M$ and a testing set that was used to assess the performance of the simplified NCM measure. The partitions were done as follows. The complete set of conditions was first ordered according to their intelligibility scores. The training dataset was constructed by selecting one out of every two conditions, leading to a 50%–50% partition of the training–testing datasets. Three additional training–testing dataset partitions were also implemented including 33%–67%, 25%–75%, and 20%–80% by selecting one out of every three, four, and five conditions, respectively, from the complete dataset.

Table III shows the resulting correlations with the binary weight vector $\mathbf{W}_M$ obtained using two different methods for selecting M (out of 20) bands, one based on sub-bands and one based on the M-maximum r values in the r-pattern spanning the full bandwidth (300–3400 Hz). We will refer to these two methods as sub-band and full-band M selection methods accordingly. For comparison, the correlation obtained using $M = 20$ bands and ANSI weights are also reported for the same partitions of the testing conditions. Comparing the correlations given in Fig. 1 with the $\mathbf{W}_1$ vector, we observe that increasing the number of bands improves to some extent the overall correlation. Notable improvement in correlation was noted with $M = 2$ in the sub-band method, but performance dropped for $M > 2$. We suspect that this was due to the fact that bands were forcefully selected with low correlation. Note that in the sub-band method, bands are selected from each sub-band regardless of the possibility that some

TABLE III. The correlation coefficients r obtained by the modified NCM measure based on M selected bands for the various training–testing partitions of the dataset. Correlations with the original NCM measure implemented using 20 bands and ANSI weights are also shown for comparison.

	Binary weights								
	Sub-band M selection				Full-band M selection				
Training–testing dataset partition	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	AI weights ($M = 20$)
50%–50%	0.80	0.84	0.78	0.77	0.80	0.84	0.86	0.85	0.78
33%–67%	0.78	0.84	0.83	0.81	0.78	0.72	0.80	0.83	0.86
25%–75%	0.83	0.85	0.75	0.79	0.83	0.87	0.87	0.85	0.82
20%–80%	0.82	0.85	0.74	0.78	0.82	0.86	0.86	0.88	0.84
Average	0.81	0.84	0.77	0.79	0.81	0.82	0.85	0.85	0.83

TABLE IV. The M selected bands reported in Table III in the various conditions. The center frequencies of the bands are given in Table I. Band 1, for instance, corresponds to a center frequency of 325 Hz.

	Method							
	Sub-band M selection				Full-band M selection			
Training–testing dataset partition	$M=1$	$M=2$	$M=3$	$M=4$	$M=1$	$M=2$	$M=3$	$M=4$
50%–50%	1	1/15	1/11/20	1/11/15/20	1	1/2	1/2/3	1/2/3/4
33%–67%	1	1/14	1/13/20	1/13/14/20	1	1/6	1/2/6	1/2/3/6
25%–75%	1	1/15	1/15/20	1/11/15/20	1	1/3	1/2/3	1/2/3/6
20%–80%	1	1/15	1/15/20	1/9/15/20	1	1/3	1/2/3	1/2/3/15

correlations in a specific sub-band might be small [see for instance the correlations of the higher frequency bands in Fig. 2(c)]. In contrast, correlations improved consistently in the full-band method as M increased. In most cases, $M=3$ and $M=4$ yielded the highest correlation. The resulting correlation with $M=4$ was in fact higher than that obtained with the ANSI weights ($M=20$) for most training–testing partitions. For the conditions involved in the 50%–50% partition, for instance, baseline correlation improved from $r=0.78$ to $r=0.85$ ($M=4$). On average, across all conditions, the baseline correlation improved from $r=0.83$ to $r=0.85$ ($M=4$). Overall, the full-band method ($M=3$, 4) was found to be more robust as it yielded consistently higher correlations than the baseline NCM measure implemented using $M=20$ bands and ANSI weights.

Table IV shows the corresponding bands selected in the various conditions. Interestingly, when $M=2$, a low-frequency (325 Hz) and a high-frequency (1874 Hz) band were consistently selected by the sub-band method in all conditions. These two disjoint bands alone seemed to be sufficient in terms of reliably predicting ($r=0.84$) the intelligibility of noise-suppressed speech. This outcome is consistent with that reported by Larm and Hongisto (2006), who utilized a simplified version of the STI (the rapid speech trasmission index, RASTI) to compute the envelopes from only the 500 and 2000 Hz octave bands. High correlations were obtained with RASTI. It should be pointed out, however, that the RASTI measure was evaluated using 4–5 modulation frequencies (spanning 0.7–11.2 Hz) for each octave band. Hence, a total of nine modulation-based SNR values were used to compute the RASTI index. In contrast, only two covariance-based SNR values [Eq. (2)] were used to compute the simplified NCM measure implemented using $M=2$.

For the full-band method used in the present study, low-frequency bands ($f<700$ Hz) were selected more often than high-frequency bands (Table IV). High correlation ($r=0.85$) was obtained with $M=3$, and the selected bands were all low in frequency (<500 Hz). This result is consistent with the outcomes from the study by Ma *et al.* (2009). A low-frequency version of the NCM measure was proposed that incorporated only low-frequency (100–1000 Hz) envelope information in its computation (Ma *et al.*, 2009). The correlation obtained with this measure, based only on bands 1–10, for predicting sentence recognition scores was nearly as good as that obtained with the full-bandwidth NCM measure.

Further improvement can be obtained with the full-band method if the M selected bands are weighted by the segment-dependent weighting functions given in Eqs. (5) and (6) (Ma *et al.*, 2009). The results are shown in Table V. Large improvements were particularly noted for $M=2$, 3, and 4 when the training–testing partition was 33%–67%. The average correlation with $M=4$ improved from 0.85 (based on binary weights in Table III) to 0.87 based on the signal-dependent weighting functions [Eqs. (5) and (6)].

V. CONCLUSIONS

This study presented a detailed analysis of a simplified NCM measure that was based on binary weighting functions. In order to account for the inherent redundancy in spectral

TABLE V. The correlation coefficients r obtained by the modified NCM measure based on M selected bands (full-band method) for the various training–testing partitions of the dataset. The weighting functions given in Eqs. (5) and (6) are used.

Training–testing dataset partition	$M=1$	$M=2$	$M=3$	$M=4$
50%–50%	0.80 $W_i^{(2)}, p=0.5$	0.84 $W_i^{(1)}, p=0.12$	0.86 $W_i^{(1)}, p=0.5$	0.86 $W_i^{(2)}, p=1.5$
33%–67%	0.78 $W_i^{(2)}, p=0.5$	0.77 $W_i^{(2)}, p=1.5$	0.85 $W_i^{(2)}, p=1.5$	0.86 $W_i^{(2)}, p=1.5$
25%–75%	0.83 $W_i^{(2)}, p=0.5$	0.87 $W_i^{(2)}, p=1$	0.87 $W_i^{(1)}, p=0.12$	0.87 $W_i^{(2)}, p=1.5$
20%–80%	0.82 $W_i^{(1)}, p=0.12$	0.86 $W_i^{(2)}, p=0.25$	0.86 $W_i^{(1)}, p=0.12$	0.88 $W_i^{(2)}, p=0.25$
Average	0.81	0.84	0.86	0.87

information contained in adjacent or nearby bands, two methods were proposed for selecting a small number (1–4) of disjoint (or contiguous) bands. Only the selected bands were subsequently used in the computation of the simplified NCM measure. Data taken from the intelligibility evaluation of noise-corrupted speech processed through eight different noise-suppression algorithms by normal-hearing listeners were used (Hu and Loizou, 2007) to assess the prediction power of the modified NCM measure. The following conclusions can be drawn from the present study:

(1) High correlation ($r = 0.8$) can be obtained with the modified NCM measure even when only one band (e.g., band 1) is used (Fig. 1). Further analysis revealed a masker-specific pattern of correlations when only one band was used in the implementation of the NCM measure (Fig. 2). The so-called *r*-pattern differed across the four maskers (babble, car, street, and train interferences) tested. The frequency location of the dips (minima) in the *r*-pattern identified differences (and inconsistencies) in the way the noise-suppression algorithm(s) affected different bands (regions) of the spectrum. These inconsistencies are caused by the fact that some bands are severely distorted while other bands are effectively cleaned by the noise-suppression algorithm. Overall, our data (Figs. 2 and 4) suggest that the low correlations obtained in certain bands effectively signify the corresponding envelopes that have been severely distorted by the noise-suppression algorithm and/or the masker.

(2) Further improvements in correlation were obtained when 2–4 bands (out of a total of 20 bands) were included in the computation of the modified NCM measure (Table III). Correlation improved to $r = 0.84$ when only two disjoint bands (centered at 325 and 1874 Hz) were used. Even further improvements in correlation ($r = 0.85$) were obtained when 3 or 4 lower-frequency (<700 Hz) bands were selected. This suggests that the low-frequency region of the spectrum carries critically important information about speech. The low-frequency region of the spectrum is known to carry F1 and voicing information, which in turn provides listeners with access to low-frequency acoustic landmarks of the signal (Li and Loizou, 2008). These landmarks, often blurred in noisy conditions, are critically important for understanding speech in noise as it aids listeners to better determine syllable structure and word boundaries (Stevens, 2002).

(3) The resulting correlation with $M = 4$ was higher than the baseline correlation of 0.83 obtained with the NCM measure implemented using 20 bands and the ANSI weighting functions. Further improvements in correlations (see Table V) were obtained by using signal-dependent weighting functions (Ma *et al.*, 2009) for the selected bands. The highest correlation obtained with $M = 4$ was 0.87.

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC007527 from the National Institute of Deafness and other Communication Disorders, NIH. The authors are grateful to the Associate Editor, Dr. D. Keith Wilson, and the two reviewers who provided valuable feedback that significantly improved the presentation of the manuscript.

[1]The RDC algorithm (Gustafsson *et al.*, 2001) is a spectral-subtractive algorithm that employs a gain function that is smoothed over time using adaptive exponential averaging. To circumvent the non-causal filtering due to the use of a zero-phase gain function, Gustafsson *et al.* (2001) suggested introducing a linear phase in the gain function. The RDC spectral subtraction algorithm reduced overall the processing delay to a fraction of the analysis frame duration.

ANSI (**1997**). S3.5, American National Standard Methods for Calculation of the Speech Intelligibility Index (American National Standards Institute, New York).

Crouzet, O., and Ainsworth, W. A. (**2001**). "On the various influences of envelope information on the perception of speech in adverse conditions: An analysis of between-channel envelope correlation," *Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, Aalborg, Denmark.

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.

Gustafsson, H., Nordholm, S., and Claesson, I. (**2001**). "Spectral subtraction using reduced delay convolution and adaptive averaging," IEEE Trans. Speech Audio Proc. **9**, 799–807.

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*, October 16–20, Paris, France, pp. 29–32.

Holube, I., and Kollmeier, K. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1715.

Houtgast, T., and Steeneken, H. (**1971**). "Evaluation of speech transmission channels by using artificial signals," Acustica **25**, 355–367.

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

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

Larm, P., and Hongisto, V. (**2006**). "Experimental comparison between speech transmission index, rapid speech transmission index, and speech intelligibility index," J. Acoust. Soc. Am. **119**, 1106–1117.

Li, N., and Loizou, P. (**2008**). "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," J. Acoust. Soc. Am. **124**, 498–509.

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. F., 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.

Musch, H., and Buus, S. (**2001**). "Using statistical decision theory to predict speech intelligibility. I. Model structure," J. Acoust. Soc. Am. **109**, 2896–2909.

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

Steeneken, H., and Houtgast, T. (**1999**). "Mutual dependence of the octave-band weights in predicting speech intelligibility," Speech Commun. **28**, 109–123.

Steeneken, H., and Houtgast, T. (**2002**). "Validation of the revised STI_r method," Speech Commun. **38**, 413–425.

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

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.