An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech

Cees H. Taal,[a] Richard C. Hendriks, and Richard Heusdens
Delft, University of Technology, 2628 CD Delft, The Netherlands

Jesper Jensen
Oticon A/S, 2765 Smørum, Denmark

(Received 17 March 2010; revised 29 June 2011; accepted 17 August 2011)

Existing objective speech-intelligibility measures are suitable for several types of degradation, however, it turns out that they are less appropriate in cases where noisy speech is processed by a time-frequency weighting. To this end, an extensive evaluation is presented of objective measure for intelligibility prediction of noisy speech processed with a technique called ideal time frequency (TF) segregation. In total 17 measures are evaluated, including four advanced speech-intelligibility measures (CSII, CSTI, NSEC, DAU), the advanced speech-quality measure (PESQ), and several frame-based measures (e.g., SSNR). Furthermore, several additional measures are proposed. The study comprised a total number of 168 different TF-weightings, including unprocessed noisy speech. Out of all measures, the proposed frame-based measure MCC gave the best results ($\rho = 0.93$). An additional experiment shows that the good performing measures in this study also show high correlation with the intelligibility of single-channel noise reduced speech.
 [DOI: 10.1121/1.3641373]

PACS number(s): 43.71.Gv, 43.72.Dv [KWG] Pages: 3013–3027

I. INTRODUCTION

Speech processing systems often introduce degradations and modifications to speech signals, e.g., quantization noise in a speech coder or residual noise and speech distortion in a noise reduction scheme. To determine the perceptual consequences of these artifacts, the algorithm at hand can be evaluated by means of a listening test or an objective machine-driven quality assessment. Although a listening test can lead to a judgment as observed by the intended group of users, such tests are often costly and time consuming. Therefore, accurate and reliable objective evaluation methods are of interest since they might replace a listening test, at least in some stages of the algorithm development process. Although it is not straightforward to describe the overall quality of a speech processing system, people tend to divide the evaluation into the attributes of speech quality (i.e., pleasantness/naturalness of speech) and speech intelligibility. The primary focus of this work is on speech intelligibility.

One of the first objective intelligibility measures was developed at AT&T Bell Labs around 1920 and eventually published by French and Steinberg (1947). Kryter (1962) made the measure better accessible by proposing a calculation scheme, which is currently known as the articulation index (AI). The basic approach of AI is to determine the signal-to-noise ratio (SNR) within several frequency bands; the SNRs are then limited, normalized and subjected to auditory masking effects and are eventually combined by computing a perceptually weighted average. This approach evolved to the speech intelligibility index (SII) and was standardized under S3.5-1997 ANSI (1997). Since AI is mainly meant for simple linear degradations, e.g., additive noise, Steeneken and Houtgast (1980) proposed the speech transmission index (STI), which is also able to predict the intelligibility of reverberated speech and non-linear distortions. For this objective measure, a noise signal with the long-term average spectrum of speech is amplitude modulated at several modulation frequencies with a cosine function and applied to the communication channel. The eventual outcome of the STI is then based on the effect on the modulation depth within several frequency bands at the output of the communication channel. While the STI is based on changes in the temporal modulation domain, the spectro-temporal modulation index (STMI) proposed by Elhilali *et al*. (2003) takes into account *joint* spectro-temporal modulations. They show that STMI is also applicable for joint spectro-temporal distortions like phase jitter distortions and phase shifts next to additive noise and reverb. The majority of recently published models are still based on the fundamentals of AI (e.g., Rhebergen and Versfeld, 2005; Kates and Arehart, 2005) and STI (for an overview see Goldsworthy and Greenberg, 2004).

In contrast to speech intelligibility, for speech-quality prediction a wide variety of objective measures are available [see, e.g., Loizou (2007) and Deller, Jr. *et al*. (1993) for an overview]. Quackenbush *et al*. (1988) evaluated a large amount of objective speech-quality measures for a wide range of degradations and proposed various new objective quality measures. Typically, these quality measures are defined for short time frames (≈25 ms), e.g., based on linear prediction coefficients and/or loudness differences in some time-frequency (TF) representation. More recently, Beerends *et al*. (2002) developed the advanced objective speech-quality measure PESQ, which can be considered as state of

[a]Author to whom correspondence should be addressed. Electronic mail: c.h.taal@tudelft.nl

the art in the field of speech-quality prediction. Several studies are available where PESQ is adjusted in order to assess the intelligibility instead of speech quality of several signal degradations such like beamforming (Beerends *et al*., 2004), low-bitrate vocoders (Beerends *et al*., 2005), and speech-enhancement systems (Kitawaki and Yamada, 2007; Yamada *et al*., 2006). Recent findings also show that other objective speech-quality measures may be used for speech-intelligibility prediction (Liu *et al*., 2008; Taal *et al*., 2009; Ma *et al*., 2009).

Although there appears to be a relation between speech quality and speech intelligibility (Preminger and Tasell, 1995), it is not that obvious that speech-quality measures can be used for speech-intelligibility assessment. For example, Liu *et al*. (2008) indicated that for SNRs below –10 dB speech may still be partly intelligible, while a lower bound for speech quality (a MOS equal to 1 indicating bad quality) is already reached. Correlation between quality and intelligibility may therefore not be present in these regions. Furthermore, there are still many types of signal degradations for which the relation between quality and intelligibility is not well understood, and perhaps not even present. For example, the quality of noisy speech may be improved by applying a single-channel noise-reduction algorithm (Hu and Loizou, 2007b), while the intelligibility is typically not improved or sometimes even decreased (Hu and Loizou, 2007a). Moreover, many objective intelligibility measures still predict incorrectly a significant intelligibility improvement after noise reduction (e.g., Ludvigsen *et al*., 1993; Dubbelboer and Houtgast, 2008; Goldsworthy and Greenberg, 2004; Taal *et al*., 2010). Only recently have new promising intelligibility measures for single-channel noise reduction been proposed by Ma *et al*. (2009), which are of great interest for the analysis of existing algorithms. However, for the development of near-future noise-reduction algorithms which aim for intelligibility improvements, these measures should be reliable for a wide variety of TF-varying gain functions applied to noisy speech and not only the ones used in conventional systems. New algorithms may involve different strategies for which it is unknown if the measures from Ma *et al*. (2009) are reliable.

In this work an evaluation is presented of objective measures for the intelligibility prediction of noisy speech processed with a technique called ideal time frequency segregation (ITFS) (Brungart *et al*., 2006). ITFS is an approach from the field of computational auditory scene analysis, simulating the remarkable properties of the auditory system to segregate a target speaker from a noisy environment. This technique is particularly of interest, since it delivers a wide variety of applied TF-weightings which can have a much stronger effect on speech intelligibility compared to single-channel noise reduction. An important reason for this difference is that ITFS assumes knowledge of the clean speech signal. Although it can therefore not be used as a practical noise-reduction algorithm (i.e., the clean speech is unknown in practice), large intelligibility improvements can be achieved with ITFS (Kjems *et al*., 2009). Moreover, the evaluation presented in this work also contains ITFS-settings which decrease the speech intelligibility of noisy speech to a larger extent than conventional noise reduction systems. The variety of signals resulting from ITFS is also demonstrated by the fact that ITFS can be applied to essentially noise-only signals, which gives fully intelligible speech (Kjems *et al*., 2009) somewhat similar to multichannel vocoded speech (Shannon *et al*., 1995). Objective measures which can correctly predict all these different aspects of ITFS are therefore expected to be robust for a wide variety of applied TF-weightings to noisy speech. Such measures may provide hints on how, and how not to process noisy speech in future algorithms which aim for intelligibility improvements. In addition, intelligibility prediction of the vocoded speech signals in ITFS is of interest in the field of cochlear implants. Namely, presenting vocoded speech to normal-hearing listeners has been a valuable method of simulating listening tests for cochlear implant users (Loizou, 1998). Hence, such reliable measures could be used, for example, in the development process of new speech-coding strategies for cochlear implants.

In total 17 objective measures are evaluated for the intelligibility prediction of ITFS-processed noisy speech. This study comprises three state-of-the-art measures for single-channel noise reduced speech as proposed by Ma *et al*. (2009), the Dau auditory model (DAU) (Christiansen *et al*., 2010), and the normalized subband envelope correlation (NSEC) (Boldt and Ellis, 2009) which both show high correlation with ITFS-processed speech, the advanced speech-quality measure (PESQ), and several conventional frame-based speech-quality measures, e.g., segmental SNR. We address some differences between quality and intelligibility prediction for ITFS-processed speech and propose a general technique which improves the performance of the frame-based quality measures when used for intelligibility assessment. From the evaluation several new promising measures for intelligibility prediction of ITFS-processed speech are revealed. To demonstrate the robustness of these measures and the generality of ITFS-processed speech, we show that they also show good prediction results for a listening test where several single-channel noise reduction algorithms are evaluated.

II. INTELLIGIBILITY DATA

The intelligibility data is obtained from a study by Kjems *et al*. (2009), where speech is degraded with various noise types at various SNRs followed by ITFS-processing as explained in Brungart *et al*. (2006). ITFS is similar to conventional noise reduction in the sense that a TF-varying gain function is applied to noisy speech. However, instead of a continuous gain function, a *binary* TF-weighting is applied to the noisy speech called the ideal binary mask (IBM) (Wang, 2005). Since details of ITFS systems differ, e.g., in thresholds used to determine the binary TF-weighting, TF-decompositions, gain values used, etc., we describe the specific system (Kjems *et al*., 2009) used to generate the speech data underlying our study.

A. Signal processing

The IBM has a value equal to one, when the instantaneous SNR within a certain TF unit exceeds a user-defined

local criterion (LC) and is zero otherwise. A mathematical description for the IBM is given as follows:

$$IBM(t,f) = \begin{cases} 1 & \text{if } T(t,f) - M(t,f) > LC \\ 0 & \text{otherwise,} \end{cases}, \quad (1)$$

where $T(t, f)$ and $M(t, f)$ denote the signal power in dBs, at time t and frequency f, for the target (clean speech) and the masker (noise only), respectively. The TF decomposition is performed at a sample rate of 20 kHz, by means of a gammatone filterbank (e.g., Patterson *et al.*, 1992) consisting of 64, 2048 tap FIR filters followed by a time segmentation of 20 ms windowed frames with an overlap of 10 ms. The gammatone filters are linearly spaced on an ERB scale between 55 and 7500 Hz The value of each TF unit is then defined as the signal energy within such a time segment. Next, the IBM is calculated, upsampled to the original sample rate, and multiplied with the noisy signal in each band. Finally, the signal is reconstructed by applying the time-reversed gammatone filters and adding the auditory bands.

B. Test material

The test signals are taken from the Dantale II corpus (Wagener *et al.*, 2003), which consists of five-word sentences all spoken by the same Danish female speaker. The sentences are of the grammatical form name-verb-numeral-adjective-noun (e.g., Ingrid owns six old jackets), where each word in the sentence is picked randomly from a list of 10 possible words. Before ITFS-processing, the speech signals are mixed with four noise types: speech shaped noise (SSN), cafeteria noise, noise from a bottling factory hall and car interior noise and mixed at three different SNRs, including the 20 and 50% speech reception threshold (SRT) and an SNR of −60 dB (The x% SRT is the SNR at which the average listener achieves x% intelligibility). The SNR of −60 dB is included for the generation of the vocoded speech signals. Kjems *et al.* (2009) performed a different listening test to determine the SRTs by finding the psychometric function for each noise type with the adaptive procedure described by Wagener *et al.* (2003), where the noisy signal energy was normalized before playback. The SRTs were then found by sampling the psychometric function where the results are shown in Table I.

Eight different values for LC are chosen, including an unprocessed condition where only the noisy speech is presented, i.e., $LC = -\infty$. LC is chosen such that the percentage of ones in the IBM varies from approximately 1.5–80%. In addition, an alternative way of calculating the IBM is included, which is only based on the clean speech. This so called target binary mask (TBM) is obtained by comparing the clean speech power with the power of a signal with the long-term spectrum of the clean speech, within a TF unit. Therefore, the noise itself is not needed in order to determine the binary mask. Note, that the TBM equals the IBM for the case that SSN is used, therefore the TBM is not included for the SSN case. In total, this results in $(4 * \text{IBM} + 3 * \text{TBM}) \times (3 * \text{SNR}) \times (8 * \text{LC}) = 168$ conditions to be tested in the listening experiment.

TABLE I. The different SNRs in dB used for each noise type (taken from Kjems *et al.*, 2009).

	SSN	bottles	cafeteria	car
20% SRT	−9.8	−18.4	−13.8	−23.0
50% SRT	−7.3	−12.2	−8.8	−20.3

C. Listening experiment

For the listening experiment, 15 normal-hearing native Danish speaking subjects participated. The correctly recognized words were recorded by an operator without providing any form of feedback. The average score for all users in each condition was consequently obtained by the average percentage of correct words.

As an example, the results for all SSN conditions are plotted in Fig. 1. Here, the percentage of correct words is plotted as a function of the mask density, i.e., the total percentage of ones in the IBM excluding noise-only regions (see Kjems *et al.*, 2009, for how the noise-only regions are defined). Note, that the rightmost point refers to a binary mask with only ones, i.e., $LC = -\infty$, which equals the condition where the noisy speech is unprocessed. It can be clearly observed that the speech can be made fully intelligible when the mask density is ≈ 20%, independent of the SNR. This is even valid for the –60 dB case, which will be a challenging condition for the objective measures, since all temporal fine structure is lost. When the mask density is lowered the intelligibility actually decreases, which can even drop below the intelligibility of the unprocessed noisy speech. This is the case for the 50% SRT signals.

III. OBJECTIVE MEASURES

An overview of the objective measures with their corresponding abbreviations and references can be found in

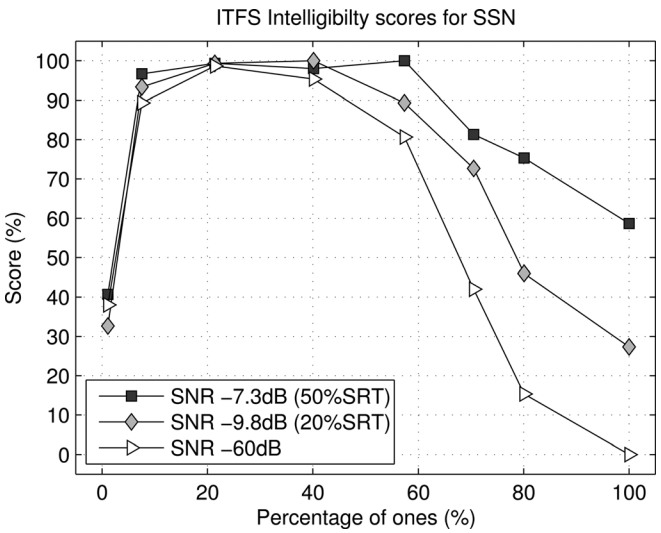

FIG. 1. Intelligibility of ITFS-processed speech, degraded with speech shaped noise (replotted from Kjems *et al.*, 2009). The percentage of correct words is plotted as a function of the mask density, i.e., the total percentage of ones in the IBM. The mask density of 100% refers to a binary mask with only ones and equals the noisy unprocessed speech.

Table II. DAU, NSEC, CSII, and CSTI are intelligibility measures, PESQ an advanced quality measure and the measures LLR, IS, CEP, SSNR, MSD, LSD, FWS1, FWS2, WSS, and PAR are speech-quality measures based on short-time ($\approx$ 20–40 ms) frames. MCC and LCC are newly proposed measures based on spectral correlation in short-time frames.

A. Preliminaries

For each of the objective measures evaluated in this study, a general descriptive notation was adopted. The outcome of an objective measure is denoted by $d(x, y)$, where x is the clean speech and y the processed speech. Let m, k, and n denote the time-frame, frequency-bin, and time-sample index, respectively. The nth sample of the mth Hann-windowed frame of x is then denoted by $x_m(n)$ and its kth DFT bin by $X_m(k)$. Similarly, $y_m(n)$ and $Y_m(k)$ represent the time frame and the DFT bin of the processed speech, respectively. Furthermore, let M, N, and K denote the total number of frames, the frame length and the total number of DFT bins, respectively. For other frequency decompositions (e.g., critical bands), the band index will be denoted by j where J equals the total number bands. For all objective measures, a sample rate of 10 kHz is used with $N = 256$ and $K = 512$, unless noted otherwise.

B. Intelligibility measures

1. Dau auditory model

The advanced auditory model developed by Dau *et al.* (1996) (DAU) has been used as an intelligibility predictor by Christiansen *et al.* (2010) and shows high correlation with ITFS-processed speech. First, the spectro-temporal internal representations of x and y are determined as described in Dau *et al.* (1996), followed by a segmentation in short-time frames within each auditory channel. Subsequently, each frame is compared by means of a correlation coefficient. Let $\Phi_{x,m}(n, j)$ and $\Phi_{y,m}(n, j)$ denote the internal representations of the complete signals x and y, respectively, for the mth frame. The measure is then simply defined as

$$d_{\mathrm{DAU}}(x,y) = \frac{1}{M}\sum_{m}\frac{\sum_{n,j}\left(\Phi_{x,m}(n,j)-\mu_{\Phi_{x,m}}\right)\left(\Phi_{y,m}(n,j)-\mu_{\Phi_{y,m}}\right)}{\sqrt{\sum_{n,j}\left(\Phi_{x,m}(n,j)-\mu_{\Phi_{x,m}}\right)^2\sum_{n,j}\left(\Phi_{y,m}(n,j)-\mu_{\Phi_{y,m}}\right)^2}}, \tag{2}$$

where $\mu\Phi_{x,m}$ and $\mu\Phi_{y,m}$ denote the average value of $\Phi_{x,m}$ and $\Phi_{y,m}$, respectively.

2. Coherence speech-intelligibility index

The coherence speech-intelligibility index (CSII) (Kates and Arehart, 2005) is based on the magnitude squared coherence function which is defined as the magnitude squared of the normalized cross-spectral density between x and y, that is,

$$|\gamma(k)|^2 = \frac{|E[X(k)Y(k)]|^2}{E\left[|X(k)|^2\right]E\left[|Y(k)|^2\right]}, \tag{3}$$

where the asterisk denotes complex conjugation and $E[\cdots]$ denotes the expectation operator. Kates and Arehart (2005) use a periodogram-based estimator for the spectral densities in Eq. (3) [e.g., $(1/M)\sum_m X_m(k)Y_m^*(k)$ estimates the cross-spectral density between $X(k)$ and $Y(k)$]. Equation (3) can be used to express the SNR within an auditory filter as follows (Kates and Arehart, 2005):

$$\mathrm{SNR}(j) = \frac{\sum_k W_j(k)|\gamma(k)|^2 E\left[|Y(k)|^2\right]}{\sum_k W_j(k)\left(1-|\gamma(k)|^2\right)E\left[|Y(k)|^2\right]}, \tag{4}$$

where W_j denotes the frequency weighting of an auditory band by means of a ro-ex filter (Kates and Arehart, 2005). The eventual CSII is then calculated by using the traditional SII (ANSI, 1997) with the SNR replaced by Eq. (4). We use the implementation as proposed by Ma *et al.* (2009), which shows high correlation with the intelligibility of single-channel noise-reduced speech [referred to as $\mathrm{CSII}_{\mathrm{mid}}$, W_4, $p = 1$ by Ma *et al.* (2009)].

TABLE II. The evaluated objective measures with their corresponding abbreviations and full names.

Objective measure name	Abbr.
Dau auditory model (Christiansen *et al.*, 2010)	DAU
Normalized subband envelope correlation (Boldt and Ellis, 2009)	NSEC
Coherence SII (Kates and Arehart, 2005)	CSII
Normalized covariance based STI (Goldsworthy and Greenberg, 2004)	CSTI
Perceptual evaluation of speech quality (Beerends *et al.*, 2002)	PESQ
Log likelihood ratio (Gray Jr and Markel, 1976)	LLR
Itakura saito distance (Itakura and Saito, 1970)	IS
Cepstral distance (Gray Jr and Markel, 1976)	CEP
Segmental SNR (Deller Jr *et al.*, 1993)	SSNR
Magnitude spectral distance	MSD
Log spectral distance	LSD
Frequency weighted SSNR (Tribolet *et al.*, 1978)	FWS1
Normalized frequency weighted SSNR (Hu and Loizou, 2008)	FWS2
Weighted spectral slope metric (Klatt, 1982)	WSS
Van de Par auditory model (van de Par *et al.*, 2005)	PAR
Magnitude spectral correlation coefficient	MCC
Log spectral correlation coefficient	LCC

3. Normalized covariance based speech transmission index

The normalized covariance based speech transmission index (CSTI) (Koch, 1992; Goldsworthy and Greenberg, 2004) shows good results for several types of nonlinear signal degradations, e.g., clipping and spectral subtraction. Let Ψ_x and Ψ_y denote the magnitude envelopes, within an octave band, of the clean and processed speech, respectively. The CSTI is then defined as the correlation coefficient between the band magnitude envelopes within an octave band of the processed and clean speech, that is,

$$r_j = \frac{\sum_m \left(\Psi_x(m,j) - \mu_{\Psi_x}\right)\left(\Psi_y(m,j) - \mu_{\Psi_y}\right)}{\sqrt{\sum_m \left(\Psi_x(m,j) - \mu_{\Psi_x}\right)^2 \sum_m \left(\Psi_y(m,j) - \mu_{\Psi_y}\right)^2}}, \quad (5)$$

This correlation coefficient is then translated to an apparent SNR (Goldsworthy and Greenberg, 2004),

$$a\mathrm{SNR}(j) = \frac{r_j^2}{1 - r_j^2}, \quad (6)$$

which is then clipped between −15 and +15 dB and normalized between 0 and 1. Let $\overline{a\mathrm{SNR}}(j)$ denote the clipped and normalized apparent SNR, the overall CSTI is then obtained by a weighted average

$$d_{\mathrm{CSTI}}(x,y) = \sum_j \overline{a\mathrm{SNR}}(j) w(j), \quad (7)$$

where we use w as proposed by Ma *et al.* (2009) to improve its performance with respect to single-channel noise reduced speech [referred to as NCM, $W_i^{(1)}$, $p = 1.5$ by Ma *et al.* (2009)].

4. Normalized subband envelope correlation

Similarly as DAU, the normalized subband envelope correlation (NSEC) (Boldt and Ellis, 2009) also shows good correlation with ITFS-processed speech (Boldt and Ellis, 2009). First, a 16 channel gammatone filterbank (80 to 8000 Hz, equally spaced on the ERB scale) is applied on the clean and processed speech, after which the normalized, compressed and highpass filtered intensity envelopes $\Lambda(m,j)$ are extracted. The eventual distance between the clean and processed speech is then defined by the normalized correlation over all time and frequency points, that is,

$$d_{\mathrm{nsec}}(x,y) = \frac{\sum_{m,j} \Lambda_x(m,j)\Lambda_y(m,j)}{\sqrt{\sum_{m,j} \left(\Lambda_x(m,j)\right)^2 \sum_{m,j} \left(\Lambda_y(m,j)\right)^2}}, \quad (8)$$

where Λ_x and Λ_y represent intensity envelopes of the clean and processed speech, respectively.

C. Speech quality measures

1. PESQ

Perceptual evaluation of speech quality (PESQ) (Beerends *et al.*, 2002) can be considered as a state of the art speech-quality predictor. Because PESQ is rather complex, we will only briefly describe its main aspects. First, the clean and processed speech are time aligned in order to compensate for any delay differences, after which both signals are processed by a psycho-acoustical model to obtain their internal representations. After global and local normalization these representations are compared resulting in so-called time-frequency dependent disturbance densities. By combining these values a PESQ-score is obtained. In this research, the wide band implementation of PESQ from (Loizou, 2007) is used.

2. Frame-based measures

The measures explained in this section are only defined for short-time frames, i.e., $d(x_m,y_m)$. For notational convenience, the frame index m is omitted for these measures and the notation $\hat{d}(x,y)$ is used instead of d (x, y). To obtain for each objective measure one total distance measure, the individual frame distances should be combined somehow. This is done by means of a simple average. However, to eliminate the influence of any outliers, first all individual frame distances are sorted, where the average is only taken over the 5–95% quantile range (Hansen and Pellom, 1998). This gives

$$d(x,y) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \hat{d}(x_m, y_m), \quad (9)$$

where $\mathcal{M}$ denotes the set of frames in the 5–95% quantile range and $|\mathcal{M}|$ its cardinality.

Several basic and well-known speech-quality measures are included like the segmental SNR (SSNR) (e.g., Loizou, 2007; Deller Jr *et al.*, 1993), where the SNR is determined within short-time frames and combined. The log-likelihood ratio (LLR) (Gray, Jr. and Markel, 1976), ceptral distance (CEP) (Gray, Jr. and Markel, 1976) and the Itakura-Saito distance (IS) (Itakura and Saito, 1970) are also common speech-quality measures, which assume that speech is an auto-regressive process for short-time segments which can be modeled with linear prediction methods. In contrast to LLR and CEP, IS is also a function of the LPC gains, which implies that a linear scaling applied on the speech will influence the outcome of the IS, which is not the case for the LLR. CEP is a function of the cepstral coefficients which can be estimated directly from the LPC coefficients (Quackenbush *et al.*, 1988). For more mathematical details for these three measures see, e.g., Quackenbush *et al.* (1988), Hansen and Pellom (1998), and Loizou (2007).

a. Critical-band based measures. Several measures evaluated in this research use a perceptually motivated frequency analysis by means of a DFT-based critical-band decomposition. This is implemented by applying an ℓ_2-norm on the critical-band filtered DFT spectrum, that is,

$$\Gamma_{x_m}(j) = \sqrt{\sum_{k=0}^{K/2} \left|H_j(k) X_m(k)\right|^2}, \quad (10)$$

where $\Gamma_{x_m}(j)$ denotes the level within the jth critical band of x_m and H represents an approximation of the magnitude spectrum of a fourth order gammatone filter (e.g., Patterson *et al.*, 1992) as described in van de Par *et al.* (2005). The signal is decomposed into 32 different filter channels equally spaced on an ERB scale ranging from 150 to 4250 Hz to include, approximately, a relevant frequency range for speech intelligibility (French and Steinberg, 1947).

One of the simplest distance measures applied on critical band spectra is the magnitude spectral distance (MSD), where an ℓ_2-norm is applied on the difference between the clean and processed magnitude spectra, that is,

$$\hat{d}_{\mathrm{MSD}}(x,y) = \sqrt{\sum_{j=0}^{J-1} \left|\Gamma_y(j) - \Gamma_x(j)\right|^2}. \tag{11}$$

The same distance measure is also applied on the log spectra [i.e., $20\log_{10}(\Gamma(j))$] denoted by log spectral distance (LSD), which is more in line with how level differences are perceived by the auditory system.

A logical extension of the SSNR is to determine an SNR within a critical band. This approach is proposed in (Tribolet *et al.*, 1978) and is known as the frequency weighted SNR (FWS) and is given by

$$\hat{d}_{\mathrm{FWS}}(x,y) = \frac{\sum_{j=0}^{J-1} w(j) 10\log_{10}\left(\frac{\Gamma_x(j)^2}{\left(\Gamma_y(j)-\Gamma_x(j)\right)^2}\right)}{\sum_{j=0}^{J-1} w(j)}, \tag{12}$$

where w denotes the AI-index weights (Kryter, 1962) as proposed by Quackenbush *et al.* (1988). An adjusted version is also included as proposed by Ma *et al.* (2009), which has better performance with single-channel noise reduced speech [referred to as fwSNRseg, $p=1$ by Ma *et al.* (2009)]. Here, before applying the critical band filters in Eq. (10), the DFT spectra of the clean and processed speech frames are first normalized to unit length in the ℓ_1-sense. Furthermore, weighting functions based on the clean speech signal are used. We denote the approach with the AI weights by FWS1 and the latter version with FWS2.

Klatt *et al.* defined a distance measure known as the weighted spectral slope metric (WSS) (Klatt, 1982), which is based on the spectral slopes in each band. First, the slope for each log-spectral critical band is calculated as follows:

$$s(j) = 20\log_{10}\Gamma(j+1) - 20\log_{10}\Gamma(j). \tag{13}$$

Then, a weighting function per band is used which is based on the level difference between the current band and the band containing the closest peak, and on the level difference between the current band and the band with the maximum peak in the spectrum, that is,

$$w(j) = \frac{c_g}{\left(c_g + \Gamma_g - 20\log_{10}\Gamma(j)\right)} \frac{c_l}{\left(c_l + \Gamma_l(j) - 20\log_{10}\Gamma(j)\right)}, \tag{14}$$

where Γ_g denotes the global maximum log-spectral magnitude of all critical bands and Γ_l the local log-spectral magnitude of the peak which is nearest to band j. The values c_g and c_l are constants which were set to 20 and 1, respectively (Klatt, 1982). The final outcome of the WSS is then defined as

$$\hat{d}_{\mathrm{WSS}}(x,y) = \sum_{j=0}^{J-1} w(j)\left(s_x(j) - s_y(j)\right)^2. \tag{15}$$

van de Par *et al.* (2005) proposed an auditory model based on spectral integration (PAR) and combines the noise-to-signal ratio within the critical bands to determine the eventual distortion outcome. The measure is defined as

$$\hat{d}_{\mathrm{PAR}}(x,y) = Nc_2 \sum_{j=0}^{J-1} \frac{\Gamma_{\varepsilon * h_{om}}(j)^2}{\Gamma_{x * h_{om}}(j)^2 + c_1}, \tag{16}$$

where $\varepsilon = y - x$, h_{om} denotes the outer-middle ear filter, and the constants c_1 and c_2 are needed for calibration. Here, the constant c_1 can be adjusted to adapt the model sensitivity and c_2 refers to the standard deviation of internal noise responsible for an absolute hearing threshold in the absence of an input signal (masker). The model is calibrated according to van de Par *et al.* (2005).

D. Additional proposed measures MCC and LCC based on spectral correlation

The correlation coefficient is a widely used outcome measure in the field of objective intelligibility assessment. In fact, all of the intelligibility measures explained in Sec. III B are based on this correlation measure. While CSTI and CSII investigate the temporal correlation within one critical band, DAU and NSEC consider the correlation in the joint spectro-temporal domain. However, no measure based only on spectral correlation has been evaluated. Note that FWS2 is perhaps the closest to such a spectral-correlation based measure and shows indeed modest correlation with speech intelligibility (e.g., Taal *et al.*, 2009; Ma *et al.*, 2009). However, FWS2 only normalizes the speech spectra energy before evaluation and does not compensate for its mean value, which is the case for the correlation coefficient. Motivated by this, a measure based on the spectral magnitude correlation coefficient (MCC) is included,

$$\hat{d}_{\mathrm{MCC}}(x,y) = \frac{\sum_{j=0}^{J-1} \left(\Gamma_x(j) - \mu_{\Gamma_x}\right)\left(\Gamma_y(j) - \mu_{\Gamma_y}\right)}{\sqrt{\sum_{j=0}^{J-1} \left(\Gamma_x(j) - \mu_{\Gamma_x}\right)^2 \sum_{j=0}^{J-1} \left(\Gamma_y(j) - \mu_{\Gamma_y}\right)^2}}, \tag{17}$$

where μ_{Γ_x} and μ_{Γ_y} denote the sample mean of the clean and processed critical band values. The same TF-decomposition is used as with the critical-band based measures. Similarly as with LSD the same procedure is also applied on the log critical-band spectra (LCC).

IV. A CRITICAL-BAND BASED NORMALIZATION PROCEDURE

For all the frame-based measures (SSNR, LLR, IS, CEP, MSD, LSD, FWS1, FWS2, WSS, PAR, MCC, LCC),

several issues can arise when using them directly for intelligibility assessment. This is caused on one hand by certain differences between speech quality and speech intelligibility prediction, but also by the nature of some of the objective measures.

The first issue is that some of these measures are sensitive to global level differences between the clean and processed speech. This is undesirable, since the intelligibility will not be affected severely when the playback level is adjusted in a listening experiment. Initial results showed indeed that the performance of several measures (e.g., SSNR, IS, FWS1, LSD, MSD) was completely dominated by these large energy difference for certain ITFS-conditions (e.g., TF-weighted noisy speech at −60 dB SNR), which led to very poor correlation with speech intelligibility. Hence, some kind of general normalization procedure is desired. Note, that the more advanced measures DAU, CSTI, NSEC, CSII, and PESQ do not have this problem, since there is already some kind of normalization procedure included.

Secondly, some of these frame-based measures are more sensitive for the frequency regions where the speech energy is dominant. This means that the low-energy high frequencies of speech (≈ 2–3 kHz) contribute less compared to lower, more powerful, frequencies (≈ 500 Hz). Although this could make sense in the field of speech-quality assessment, it turns out this is not appropriate for speech-intelligibility prediction. Several studies have shown that these high frequency components are actually of similar importance for the speech intelligibility (e.g., ANSI, 1997; Steeneken and Houtgast, 1980).

The third issue is the fact that certain high (> 5 kHz) and low frequencies (< 200 Hz) are of less importance to speech intelligibility (French and Steinberg, 1947), while they may be relevant for speech quality. Some measures are sensitive to these frequency ranges, which may bias the results after signal degradation.

To overcome these problems we use a typical procedure from the field of objective intelligibility assessment. This procedure consists of a normalization of the processed and clean critical-band envelopes by its rms-value before comparison. This approach is used for most of the STI-based (Goldsworthy and Greenberg, 2004) measures and NSEC (Boldt and Ellis, 2009). The normalization procedure is applied by prefiltering the speech signals before evaluation. In this manner, normalization can be applied to any arbitrary objective measure. Let α_j denote the normalization factor for each critical band, which equals the reciprocal of its rms value,

$$\alpha_j = \left(\frac{1}{KM} \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \left| X_m(k) H_j(k) \right|^2 \right)^{-1/2}, \tag{18}$$

where H equals the spectrum of one critical band as in Eq. (10). The normalized kth DFT bin of the mth frame, say $X'_m(k)$, is then obtained by an addition of all scaled critical bands,

$$X'_m(k) = \sum_{j=0}^{J-1} \alpha_j X_m(k) H_j(k). \tag{19}$$

The time-domain signal can now be reconstructed from the weighted short-time DFT bins by means of a simple overlap-add procedure. The processed speech y is normalized with the same procedure. The rms within each critical band is now fixed, which makes each measure insensitive for global energy differences. Furthermore, each critical band will have an equal contribution to speech intelligibility. Moreover, the total response of the sum of all critical bands will only take into account the frequency range approximately between 150 and 4500 Hz, which is roughly a relevant range for speech intelligibility.

V. EVALUATION PROCEDURE

For each ITFS condition, 30 five-word sentences are randomly chosen from the corpus, concatenated and ITFS processed. Before applying the objective measures, the silent regions are removed between the five-word sentences. To compare the results of the objective measures and the intelligibility scores, a mapping is needed in order to account for a nonlinear relation. A widely used mapping is the logistic function

$$f(d) = \frac{100}{1 + \exp(ad + b)}, \tag{20}$$

while for some measures a better fit was observed with the following function (Taal *et al.*, 2009),

$$f(d) = \frac{100}{1 + (ad + b)^c}, \tag{21}$$

where a, b, and c in Eq. (20) and Eq. (21) are free parameters, which are fitted to the intelligibility scores with a nonlinear least squares procedure, and d denotes the objective outcome. For each objective measure both mappings are evaluated, where finally the best fit is used. For evaluation we use the correlation coefficient (ρ) and a normalized version of the rms of the prediction error (σ) (RMSE),

$$\sigma = \frac{1}{100} \sqrt{\frac{1}{S} \sum_i \left(s_i - f(d_i) \right)^2}, \tag{22}$$

where s refers to an intelligibility score, S denotes the total number of processing conditions and i runs over all processing conditions. The factor 100 is included to make sure the RMSE is in the same range as the correlation coefficient. The mapping functions may not show a good fit between the intelligibility scores and the objective data for all objective measures. Therefore, the Kendall's tau (Sheskin, 2004) is also included. This outcome measure is independent of the applied (monotonic) mapping and solely tests whether there is a monotonic relation between the intelligibility scores and the objective scores.

VI. RESULTS AND DISCUSSION

For each objective measure, the RMSE, the Kendall's tau and the correlation coefficient is given in Table III,

TABLE III. RMSE (σ), Kendall's tau (τ), and correlation coefficient (ρ) for all objective measures

Name	σ	τ	ρ
PESQ	0.30	0.30	0.41
SSNR	0.27	0.38	0.58
MSD	0.16	0.70	0.88
LSD	0.32	0.19	0.30
FWS1	0.25	0.57	0.67
FWS2	0.24	0.54	0.69
WSS	0.26	0.43	0.60
PAR	0.28	0.34	0.52
MCC	0.12	0.77	0.93
LCC	0.15	0.73	0.88
LLR	0.31	0.24	0.35
IS	0.31	−0.08	0.33
CEP	0.32	0.12	0.19
DAU	0.15	0.73	0.89
CSII	0.29	0.37	0.45
CSTI	0.20	0.63	0.80
NSEC	0.15	0.74	0.89

where, except for DAU, CSII, CSTI, NSEC, and PESQ, the signals were first subjected to the proposed critical-band based normalization procedure. To give a clear overview of the differences in performance, the correlation coefficients are ranked in Fig. 2. Also the scatter plots and the fitted mapping functions are shown in Fig. 3. We can observe that the proposed measure MCC gave the best results, followed by NSEC, DAU, and LCC. The simple MSD correlated better with the intelligibility scores than various other, more advanced objective measures (e.g., CSTI). Remarkably, the more advanced measures CSII and PESQ performed relatively poor.

For the measures CSII, CSTI, and FWS2 the new band-importance functions were used as proposed by Ma *et al.* (2009). However, we also evaluated the performance with their original implementations (not shown). For CSII and CSTI we did not observe any large changes in performance, while for FWS2 the performance slightly dropped with the version proposed by Ma *et al.* (2009). In general, the conclusions made in this work hold for both implementations of each model.

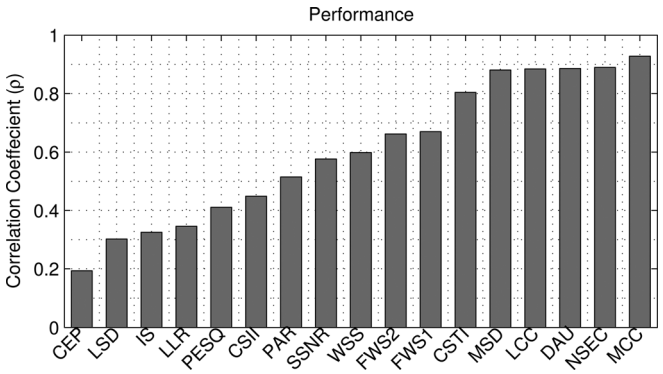

FIG. 2. Performance with respect to correlation coefficient for all objective measures (higher is better). For all measures except PSQ, CSII, CSTI, DAU, and NSEC, the speech signals are first subjected to the normalization procedure as explained in Sec. IV.

A. Detailed evaluation of intelligibility measures

Out of the four objective intelligibility measures (DAU, CSII, CSTI, NSEC), the best performance was obtained with DAU and NSEC, which both had similar values for all three outcome measures. In fact, these two measures show the best performance out of all objective measures, except for the proposed measure MCC. CSTI also performed modestly well, while CSII did not perform well.

1. DAU and NSEC

The good results of DAU and NSEC are in agreement with the results reported in (Boldt and Ellis, 2009) and (Taal *et al.*, 2009), where it was already observed that both measures appear to be good intelligibility predictors of ITFS-processed speech. Nevertheless, it was observed that both models have a similar weakness and are both more reliable for the ITFS conditions where the intelligibility score is relatively high (90–100%). To get a better insight in this behavior, an additional scatter plot of NSEC is given in Fig. 4. Here the IBM density, i.e., the percentage of ones in the binary mask, is denoted by the shading and size of the rectangular markers. A larger and brighter marker indicates a higher density IBM, where the large white squares refer to the mask density of 100%, i.e., the unprocessed noisy speech. The plot clearly illustrates that for these unprocessed conditions, the output of NSEC is much lower compared to the remaining ITFS-processed conditions. This trend is also observed when the density is lowered to 80%, which in general still have a lower objective output. As a consequence, the predicted intelligibility scores for the noisy speech conditions were underestimated. DAU has similar problems, however, from the scatter plot (not shown) it was observed that this problem was only present for the bottles noise.

2. CSTI

CSTI yielded a relatively high ranking with respect to all other objective measures. This implies that the promising results of CSTI for clipping and spectral subtraction (Goldsworthy and Greenberg, 2004), are maintained with ITFS-processed speech. Nevertheless, it is clear that the data points are less well fitted by the mapping function than, for example, DAU and NSEC. More specifically, the CSTI turns out to be less reliable for the high intelligible (90–100%) ITFS conditions than DAU and NSEC.

3. CSII

CSII performed worse than the majority of the evaluated objective measures. Figure 5 illustrates that the predicted scores for all −60 dB SNR conditions are underestimated. In fact, most prediction results for these conditions are clipped to 0, i.e., the model predicts the speech to be completely unintelligible. A similar trend occurs for the 20% SRT conditions, which generally show lower objective values than the 50% SRT conditions. This is not in line with the intelligibility scores, where specific settings of *LC* can lead to fully intelligible speech, even at low SNRs.

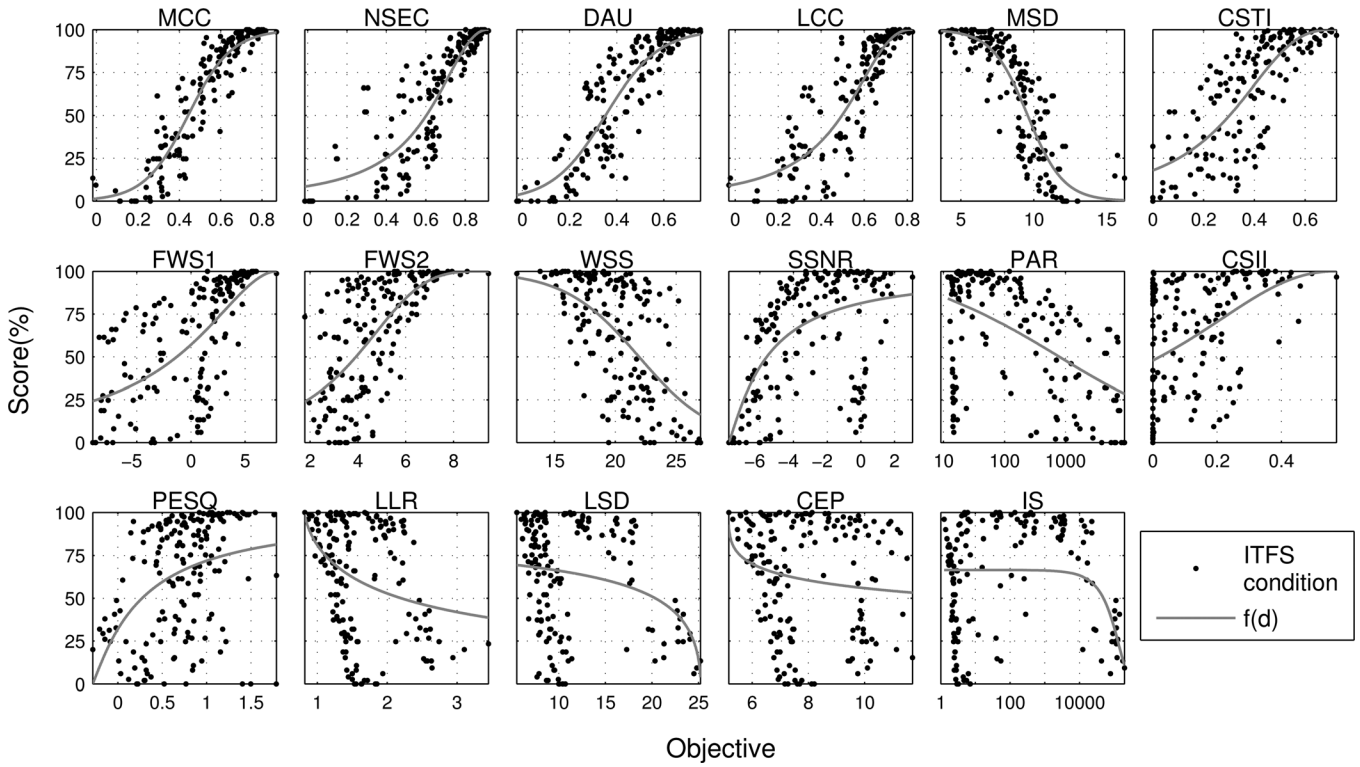

FIG. 3. Scatter plots for all objective measures together with the fitted mapping function.

A possible explanation can be given, by rewriting Eq. (3) with an independent phase and magnitude term. Let the polar representations of X and Y with magnitude a and phase θ be denoted by $a_X e^{j\theta_X}$ and $a_Y e^{j\theta_Y}$, respectively. The frequency index k is omitted for notational convenience. This gives

$$|\gamma|^2 = \frac{\left|E\left[a_X e^{j\theta_X} a_Y e^{j\theta_Y}\right]\right|^2}{E\left[\left|a_X e^{j\theta_X}\right|^2\right] E\left[\left|a_Y e^{j\theta_Y}\right|^2\right]}, \tag{23}$$

A reasonable assumption for speech is that the phase is independently distributed from its magnitude (Erkelens *et al.*, 2007). Equation (23) can then be rewritten as

$$|\gamma|^2 = \frac{E[a_X a_Y]^2}{E[a_X^2]E[a_Y^2]} \left|E\left[e^{j(\theta_X-\theta_Y)}\right]\right|^2. \tag{24}$$

The right-hand term now indicates the sensitivity for the phase difference, independently of the magnitudes.

For the situation where the clean speech magnitudes are preserved, i.e., $a_X = a_Y$, but a different uniformly distributed phase is used, the right hand term in Eq. (24) will be equal to 0. As a consequence, the CSII will report that the clean speech is not intelligible. Since the TF weighting in the ITFS procedure is real valued, the noisy phase will be preserved. Hence, the right term will be very close to zero in Eq. (24) for the case that essentially pure noise (−60 dB) is used. This is not in line with the observations described by Paliwal

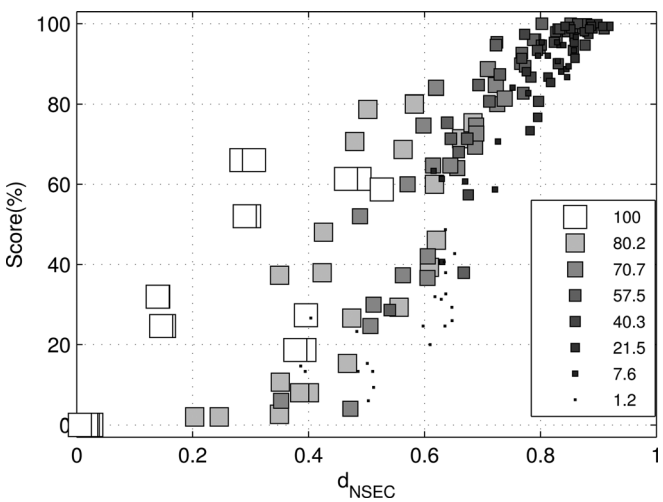

FIG. 4. Scatter plot for NSEC where the density of the IBM is highlighted by the shading and size of the markers.

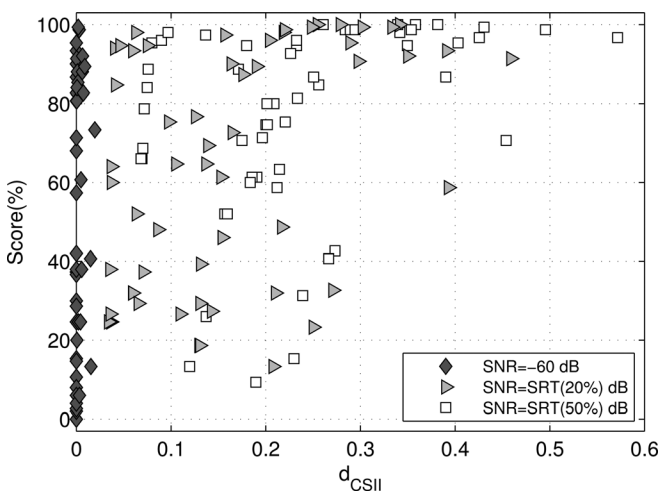

FIG. 5. Scatter plot of CSII with highlighted SNRs.

and Alsteris (2003), where it is reported that, by using a different uniformly distributed phase, the intelligibility is hardly affected.

B. Detailed evaluation of speech quality measures

1. PESQ

The low performance of PESQ was somewhat remarkable. Apparently, its high correlation with speech quality does not guarantee a good correlation with the intelligibility of the ITFS-processed speech signals. This result is different from the observations reported by Ma *et al*. (2009), where PESQ performed modestly well in terms of predicting intelligibility of single-channel enhanced noisy speech. A possible explanation for this difference is the fact that we used relatively low SNRs, compared to the higher SNRs from Ma *et al*. (2009), which were set equal to 0 and 5 dB. When lowering the SNR, PESQ will converge to a low value, predicting very poor speech quality; further lowering the SNR will have little effect on speech quality. Nevertheless, in this SNR range a lower bound for speech intelligibility is not necessarily reached yet, as was illustrated in (Liu *et al*., 2008). This explanation is also motivated by the low PESQ values, which can be observed in its scatter plot in Fig. 3. Given that PESQ is a reliable predictor of speech quality, it is therefore likely that the intelligibility of ITFS-processed speech does not correlate well with its speech quality.

2. Frame-based measures

Out of all frame-based measures the good performance of MSD was remarkable, since it is probably the simplest measure used in this research. The models FWS1 and FWS2 show modest correlation with intelligibility, which was also reported by Ma *et al*. (2009). Poorer results were obtained with WSS and SSNR. The remaining measures in ranking show poor correlation with the intelligibility of ITFS-processed signals.

MSD has approximately the same results as the complex intelligibility models DAU and NSEC. Moreover, MSD shows even better performance than the objective intelligibility measure CSTI. It is hypothesized that the proposed critical-band based normalization plays an important role for these good results (see Sec. VI C). Rather poor results were obtained with LSD. The main reason for this is that the magnitudes close to zero tend to approach minus infinity due to the log-transform. This situation occurs frequently when the IBM is sparse. This yields a large output value when evaluating the distance between processed and clean speech. The quantile-based procedure which averages all the individual frame distances [see Eq. (9)] was not sufficient to take care of these outliers.

Despite their modest correlation, the scatter plots of FWS1 and FWS2 in Fig. 3 reveal that these measures are mainly reliable for high intelligibility scores. In addition, an oversensitivity is observed for the conditions where an IBM is used with a high percentage of ones as with NSEC. This is clearly illustrated in Fig. 6, where FWS1 tends to output a lower objective score for most of the noisy unprocessed

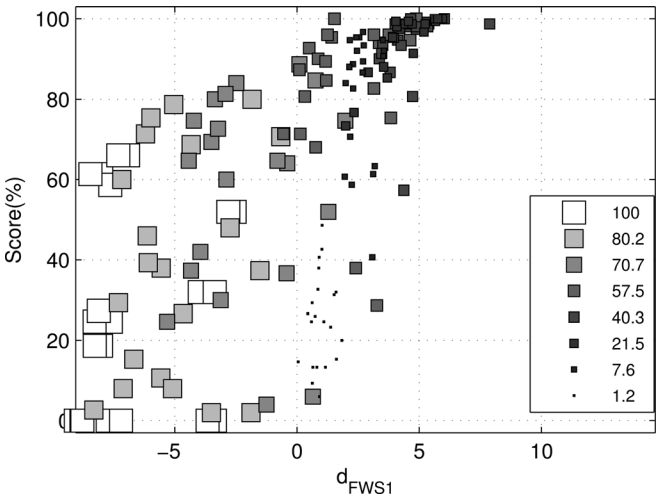

FIG. 6. Scatter plot for FWS1, where the percentage of ones of the IBM is denoted by the color and size of the markers.

speech conditions. Figure 6 also shows an additional problem, which was present for most of the SNR-based measures. Analyzing the plot reveals that for the lower mask densities (e.g., 1.2 and 7.6%), the output of FWS1 tends to converge to 0 dB. This behavior is even more present in the scatter plot of the SSNR in Fig. 3, where a cluster of points around 0 dB is observed. Indeed, it is easy to see that Eq. (12) is lower bounded by 0 dB for the case where speech information is removed, i.e., $\Gamma_y(j) < \Gamma_x(j)$. By removing speech information, the speech will eventually become unintelligible. This is not in line with the predictions of the SNR-based measures, which make them less suitable for these types of degradations. Note, that this unwanted behavior is less present with the FWS2. This is due to its normalization procedure, where the DFT spectra of the clean and processed speech frames were first normalized to unit-length in the ℓ_1-sense (Hu and Loizou, 2008). In principle, the PAR-auditory model can be interpreted as the inverse of the SNR within a critical band. However, the SNRs are not converted to a log scale, which explains the large range of scores shown in the scatter plot in Fig. 3 (Notice the log scale on the x-axis). PAR shows similar artifacts as with the SSNR, and FWS measures for the sparse IBM conditions. For PAR, these conditions tend to cluster around $d_{\text{par}} \approx 15$.

The last frame-based speech-quality measures according the ranking are LLR, CEP, and IS, which all appear to share a similar problem as with the SNR-based models. Where the SNR-based measures converged to a certain value for sparse IBMs, these measures tend to output a large value, when much speech information is removed. Similarly as with LSD, this is caused by the fact that these measures are defined in the log domain.

3. Additional proposed measures MCC and LCC based on spectral correlation

From the ranking in Fig. 2, we see that the relatively simple measure MCC has the best performance out of all objective measures. Despite its simplicity, MCC outperforms both the complex DAU model and NSEC, which makes it a

new potential measure for objective intelligibility assessment. As already mentioned, DAU and NSEC are mostly reliable for the ITFS conditions where the intelligibility score is relatively high. As shown in Fig. 7, this behavior is less present with the MCC, where the mapping shows a better fit with the data over the entire intelligibility range.

Comparable results with DAU and NSEC are obtained with LCC, which is also mainly reliable for the high intelligibility scores. Using the log spectra instead of the magnitude spectra, which is done in the MCC, the correlation with the intelligibility decreases for the evaluated ITFS conditions. Note, that DAU and NSEC also use some kind of compressive nonlinearity. In DAU this is included by means of the adaptation loops, which behave as a log transform for stationary input signals (Dau *et al.*, 1996). NSEC compresses the band intensity envelopes by raising them to the power 0.15 (Boldt and Ellis, 2009). Therefore re-investigating these band-compression stages for intelligibility assessment may be worthwhile.

C. Influence of critical-band based normalization procedure

To determine the influence of the critical-band based normalization procedure, a comparison is made with a normalization procedure based on the rms, that is $x' = x/\mathrm{rms}(x)$ and $y' = y/\mathrm{rms}(y)$. The rms-procedure is chosen since it is a straightforward and basic approach often used as an initial stage in more advanced objective measures (e.g., PESQ). Results for the three outcome measures for this experiment can be found in Table IV and in Fig. 8(a). For comparison reasons, PESQ and the four intelligibility measures are also included, denoted by the white bars (Note that these results are the same as in Fig. 2, since they were not subjected to the proposed normalization). The difference in performance is shown in Fig. 8(b), where the measures on the right indicate a stronger improvement due to the proposed critical-band based normalization procedure.

Observing the alternative ranking, none of the outcome measures of the frame-based measures have as good performance as the intelligibility measures CSTI, DAU, and NSEC. The only measure which correlates modestly with the intelligibility scores is MSD. Furthermore, as seen in Fig. 8(b), most of the frame-based measures benefit from the proposed critical-band based normalization procedure, except LSD, CEP, IS and LLR. However, also with the rms-based normalization procedure these measures turn out to be poor intelligibility predictors.

For the MCC and LCC, a clear problem was observed when the proposed critical band based normalization procedure was not included. This is caused by the already present

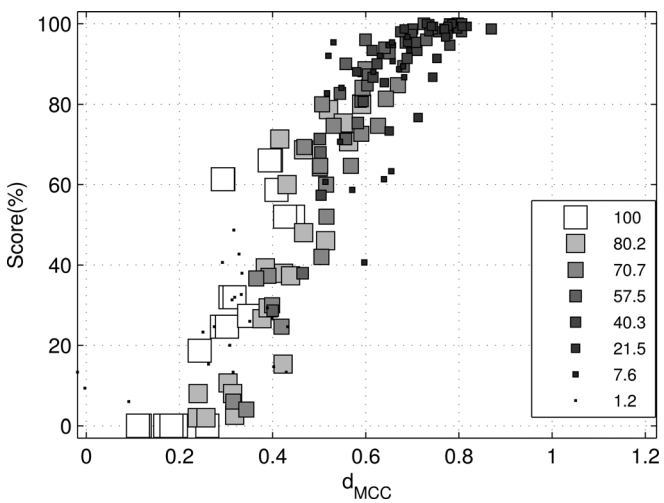

FIG. 7. Scatter plot for MCC, where the density of the IBM is denoted by the color and size of the markers.

TABLE IV. RMSE, Kendall's tau, and the correlation coefficient for all frame-based objective measures when a normalization procedure based on the rms is applied on the speech signals.

Name	σ_{rms}	τ_{rms}	ρ_{rms}
SSNR	0.30	0.28	0.44
MSD	0.24	0.54	0.70
LSD	0.30	0.24	0.41
FWS1	0.28	0.44	0.53
FWS2	0.26	0.51	0.63
WSS	0.27	0.42	0.58
PAR	0.30	0.25	0.43
MCC	0.27	0.44	0.58
LCC	0.27	0.45	0.59
LLR	0.30	0.32	0.43
IS	0.31	−0.05	0.38
CEP	0.32	0.21	0.29

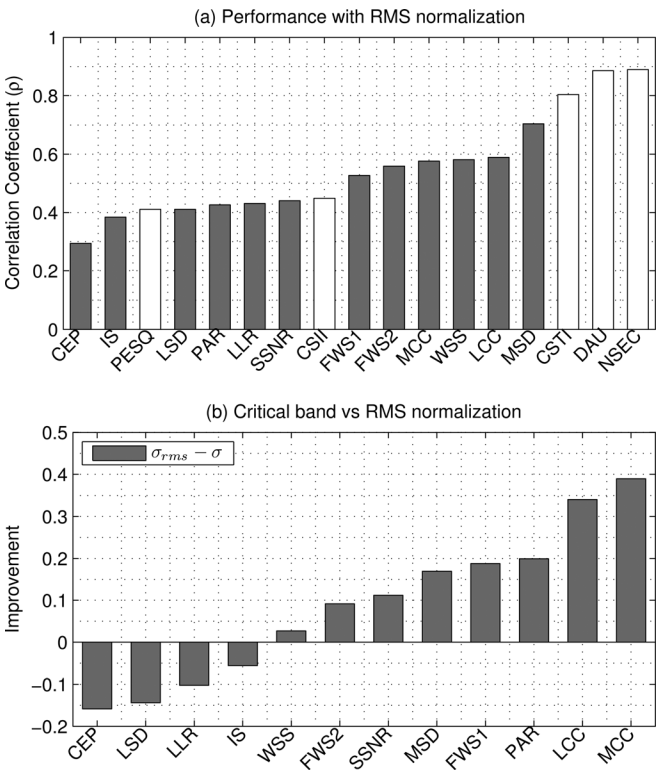

FIG. 8. (a) Ranking for all frame-based measures when normalization based on rms is used instead of the critical-band based normalization. The measures not subjected to these normalization procedures are denoted by the white colored bars. (b) The difference in performance between both normalization procedures.

correlation between the average clean and processed long-term spectra. Car noise, SSN, and cafeteria noise have a strong low-frequency content, similar to clean speech, which yields a positive correlation between their average spectra. However, the bottles noise has a strong high-frequency spectra, which shows a negative correlation with the average clean speech spectrum. This is clearly illustrated in the left plot of Fig. 9, where the noise type is denoted by the marker type. For the conditions where the speech is degraded with the bottles noise the intelligibility is underestimated, while for the remaining noise types the opposite behavior is observed. This problem is not present in the right plot, where the proposed normalization procedure is applied. After normalization the clean and processed long-term average critical-band spectra will be flat and therefore any global correlation is removed.

VII. GENERALITY OF RESULTS

From our results, several promising measures are revealed for ITFS-processed speech like MCC, NSEC, DAU, and LCC. An interesting conclusion from this evaluation is that the good performing measures are all employing a correlation coefficient in some TF-region. For example, MCC and LCC exploit spectral correlation, while CSTI looks at the correlation between the temporal envelopes within a frequency band. Moreover, DAU and NSEC are based on the correlation in the joint spectro-temporal domain. One important property of the correlation coefficient is its insensitivity to the mean value and the energy of the input signals. This probably also explains the good results obtained with the proposed normalization procedure, which eliminates the effect of the signal energy per critical band.

However, a valid question is if this correlation-based approach will also work with other TF-weighted noisy speech signals except than with ITFS, e.g., single-channel noise reduction. If we compare our findings with the results from the single-channel noise reduction evaluation of Ma *et al*. (2009), we can conclude that CSTI and FWS2 show reasonable results for both types of processing. As an initial step to indicate the robustness of the promising measures from our study (MCC, NSEC, DAU, LCC, and MSD) for TF-weighted noisy speech, an additional listening experiment is conducted where two single-channel noise reduction methods are evaluated. The prediction results from these best measures are compared with the three best performing measures from Ma *et al*. (2009), that is CSII, CSTI and FWS2, which can be considered state-of-the-art measures for intelligibility prediction for single-channel noise reduced speech. The same evaluation procedure is used as explained in Sec. V.

A. Evaluation of objective measures for single-channel noise-reduced speech

The experiment comprises unprocessed noisy speech and noisy speech processed by two different single-channel noise-reduction algorithms. That is, (1) the standard MMSE-STSA algorithm by Ephraim and Malah (1984) (EM) which was developed under the assumption that speech and noise DFT coefficients are Gaussian and (2) an improved version Erkelens *et al*. (2007) (SG), which assumes the speech and noise DFT coefficients to be super-Gaussian and Gaussian distributed, respectively. For both algorithms, the *a priori* SNR is estimated with the decision directed approach (Ephraim and Malah, 1984) with a smoothing factor of $\alpha = 0.98$. The noise PSD in EM and SG is estimated using Minimum Statistics (Martin, 2001) and the noise-tracker by Hendriks *et al*. (2010), respectively. Maximum attenuation is limited to 10 dB in both algorithms. In SG, the parameters describing the assumed super-Gaussian density of the speech DFT coefficients are $\gamma = 1$ and $\nu = 0.6$ Erkelens *et al*. (2007).

The same listening test set-up is used as in Section II. The speech signals are degraded with additive speech-shaped noise (SSN) at a sample rate of 20 kHz. Five different SNRs are considered (−8.9 dB, −7.7 dB, −6.5 dB, −5.2 dB and −3.1 dB), which were chosen such that the psychometric function of clean speech degraded by SSN [based on earlier experiments (Kjems *et al*., 2009)] was sampled approximately between 50 and 100% intelligibility.

Fifteen Danish-speaking listeners (normal hearing) were asked to judge the intelligibility of the noisy signals and the two enhanced versions. The three processing conditions (i.e., UN, EM and SG), the two noise types and the 5 SNR values make up $3 \times 2 \times 5 = 30$ conditions. For each of the 30 conditions, each listener is presented with 10 five-word sentences.

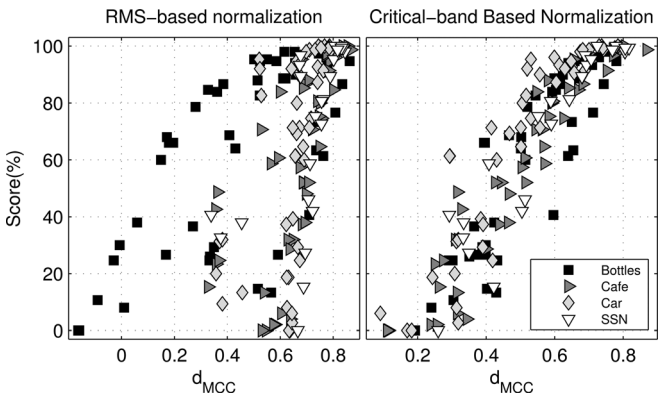

FIG. 9. Difference in performance between rms-based normalization (left plot) and critical-band based normalization (right plot) for MCC with respect to noise type.

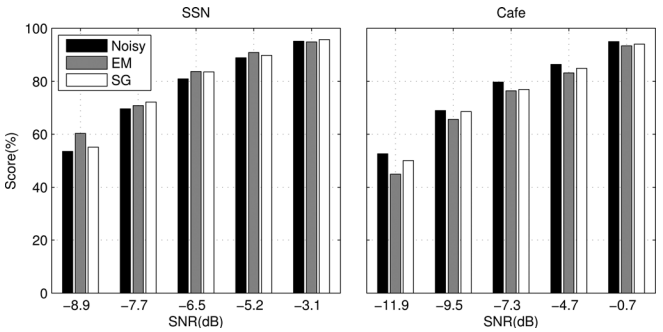

FIG. 10. Average-user intelligibility scores for unprocessed noisy (UN) speech, and two noise-reduction schemes (EM, SG) for (a) speech shaped noise and (b) cafe noise.

TABLE V. Two-way ANOVA p-values for the hypothesis that there is no effect on intelligibility due to noise reduction for both algorithms (EM, SG) and noise type (SSN, Cafe).

	EM	SG
SSN	0.2470	0.4177
Cafe	0.0702	0.4286

The results from the listening experiment are shown in Fig. 10. As can be observed, the noise-reduction algorithms have a very small effect on the speech intelligibility compared to the intelligibility of the noisy unprocessed speech. A two-way ANOVA did not showed any significant changes in intelligibility due to each noise-reduction algorithm for each noise type (see p-values in Table V). This result is in line with the conclusions from Hu and Loizou (2008) where, in general, no noise-reduction scheme could improve the intelligibility of noisy speech.

The prediction results for the objective measures are shown in Table VI. From the results we can conclude that the proposed measures MCC, MSD, and LCC also have good performance with the single-channel noise reduced signals contained in the listening test next to ITFS-processed speech. In fact, in terms of the correlation coefficient and the RMSE the proposed MCC shows similar performance as the CSII as proposed by Ma *et al*. (2009), which can be considered as a state-of-the-art intelligibility predictor of noise-reduced speech. Although not as good as the proposed measures from Ma *et al*. (2009), DAU and NSEC also show moderate correlation with this dataset. Overall it can be stated that the good performing measures for the ITFS-data set also have good performance with the single-channel noise-reduced set, but not vice-versa.

B. Other types of signal degradations

We have proposed new objective measures, which show high correlation with the intelligibility of noisy speech signals processed by a TF-varying weighting, like ITFS and single-channel noise reduction. It is not guaranteed that our results are also valid for other degradation types than TF-weighted noisy speech, e.g., reverberation. For example, Liu *et al*. (2008) showed that some measures can be very reliable for predicting the effect of speech coders on intelligibility, while the same measures may be unreliable for predicting the intelligibility of noise-reduced speech. This was also demonstrated for the CSTI by Goldsworthy and Greenberg (2004), which shows good performance for clipping and spectral subtraction but not for reverberated speech. In future research the promising measures from our research will be evaluated for other types of distortions.

TABLE VI. RMSE, Kendall's tau, and the correlation coefficient of the objective measures for intelligibility prediction of the single-channel noise-reduced speech signals.

Name	σ	τ	ρ
MCC	0.06	0.75	0.93
CSII	0.06	0.83	0.92
MSD	0.07	0.74	0.90
LCC	0.07	0.69	0.90
FWS	0.07	0.67	0.89
CSTI	0.07	0.78	0.87
DAU	0.08	0.59	0.84
NSEC	0.10	0.62	0.75

VIII. CONCLUSIONS

The focus of this study was the evaluation of various predictive models of intelligibility using ideal-time frequency segregated (ITFS) noisy speech. In total 17 objective measures were evaluated consisting of four advanced objective speech-intelligibility measures (DAU, NSEC, CSII, CSTI), an advanced speech-quality measure (PESQ), and several more conventional frame-based measures (e.g., SSNR). Several of the measures were particularly sensitive to level differences between processed and unprocessed speech. To overcome this problem a general normalization procedure based on equalizing the rms per critical band was employed. All objective measures were evaluated by means of predicting the intelligibility of 168 different conditions of noisy and ITFS-processed noisy speech signals. From these results the following conclusions can be drawn.

(1) Out of all 17 objective measures the highest correlation ($\rho = 0.93$) with speech intelligibility was obtained with the proposed frame-based measure MCC. This measure was defined as a simple correlation coefficient between the critical-band magnitude spectra of the clean and processed speech.
(2) Good results were obtained with DAU and NSEC (both with $\rho = 0.89$). Nevertheless, these measures turned out to be too sensitive for the noisy unprocessed speech compared to the TF-weighted speech. As a consequence, both measures underestimated the intelligibility for noisy speech compared to TF-weighted noisy speech.
(3) LCC and MSD frame-based measures also showed high correlations ($\rho = 0.88$).
(4) The intelligibility measure CSTI gave reasonable results ($\rho = 0.80$). Therefore, in addition to showing promising results with clipping and spectral subtraction reported by Goldsworthy and Greenberg (2004), CSTI is also a reasonable intelligibility predictor for ITFS-processed noisy speech.
(5) Poor results were obtained with the CSII, which was not a reliable intelligibility predictor for the ITFS-processed signals used in this research. This was probably due to sensitivity to the DFT phase component.
(6) The advanced objective quality-measure PESQ showed a low correlation with speech intelligibility. Since PESQ is a reliable predictor of speech quality, it is therefore likely that the intelligibility of ITFS-processed noisy speech from this study does not correlate with its speech quality.
(7) Compared with an rms-based normalization procedure, the proposed critical-band based normalization improved the correlation with intelligibility for almost all frame-based measures. In particular the measures MCC, LCC and MCD had a large performance improvement due to the proposed critical-band based normalization.

(8) The frame-based measures IS, CEP, LSD, LLR, SSNR, and PAR showed low correlation ($\rho < 0.60$) with speech intelligibility. This conclusion holds for both the proposed critical-band based normalization and the rms-based normalization procedure.
(9) The good performing measures in this study (MCC, LCC, DAU, NSEC, and FWS2) also showed high correlation with the intelligibility prediction of single-channel noise reduced speech.

ACKNOWLEDGMENTS

This research was supported by the Oticon foundation and the Dutch Technology Foundation STW.

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

Beerends, J. G., Hekstra, A. P., Rix, A. W., and Hollier, M. P. (**2002**). "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part II-psychoacoustic model," J. Audio Eng. Soc. **50**, 765–778.

Beerends, J. G., Larsen, E., Iyer, N., and van Vugt, J. M. (**2004**). "Measurement of speech intelligibility based on the PESQ approach," in *Proceedings of the Workshop Measurement of Speech and Audio Quality in Networks*.

Beerends, J. G., van Wijngaarden, S., and van Buuren, R. (**2005**). "Extension of ITU-T recommendation P. 862 PESQ towards measuring speech intelligibility with vocoders," TNO Technical Report.

Boldt, J. B., and Ellis, D. P. W. (**2009**). "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proceedings of EUSIPCO*, pp. 1849–1853.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

Christiansen, C., Pedersen, M. S., and Dau, T. (**2010**). "Prediction of speech intelligibility based on an auditory preprocessing model," Speech Commun. **52**, 678–692.

Dau, T., Püschel, D., and Kohlrausch, A. (**1996**). "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure," J. Acoust. Soc. Am. **99**, 3615–3622.

Deller, Jr., J., Proakis, J., and Hansen, J. (**1993**). *Discrete Time Processing Of Speech Signals* (Prentice Hall PTR, Upper Saddle River, NJ), pp. 580–593.

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.

Elhilali, M., Chi, T., and Shamma, S. (**2003**). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. **41**, 331–348.

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

Erkelens, J. S., Hendriks, R. C., Heusdens, R., and Jensen, J. (**2007**). "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," IEEE Trans. Audio Speech Lang. Process. **15**, 1741–1752.

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

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

Gray, Jr., A. H., and Markel, J. D. (**1976**). "Distance measures for speech processing," IEEE Trans. Acoust. Speech Signal Process. **24**, 380–391.

Hansen, J. H. L., and Pellom, B. L. (**1998**). "An effective quality evaluation protocol for speech enhancement algorithms," in *Proceedings of the Fifth International Conference on Spoken Language Processing*, Vol. 7.

Hendriks, R. C., Heusdens, R., and Jensen, J. (**2010**). "MMSE based noise PSD tracking with low complexity," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4266–4269.

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

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

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

Itakura, F., and Saito, S. (**1970**). "A statistical method for estimation of speech spectral density and formant frequencies," Electron. Commun. Jpn. **53**, 36–43.

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

Kitawaki, N., and Yamada, T. (**2007**). "Subjective and objective quality assessment for noise reduced speech," in *Proceedings of the ETSI Workshop on Speech and Noise in Wideband Communication*, pp. 1–4.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.

Klatt, D. (**1982**). "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 7.

Koch, R. (**1992**). "Auditory sound analysis for the prediction and improvement of speech intelligibility" (in German), Ph.D. thesis, Universität Göttingen.

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

Liu, W. M., Jellyman, K. A., Evans, N. W. D., and Mason, J. S. D. (**2008**). "Assessment of objective quality measures for speech intelligibility," in *Proceedings of Interspeech*, pp. 699–702.

Loizou, P. (**1998**). "Mimicking the human ear," IEEE Sign. Process. Mag. **15**, 101–130.

Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, FL), pp. 502–527.

Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**). "Evaluation of a noise reduction method—Comparison between 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.

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

Paliwal, K. K., and Alsteris, L. (**2003**). "Usefulness of phase spectrum in human speech perception," in *Proceedings of Interspeech*, pp. 2117–2120.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (**1992**). "Complex sounds and auditory images," *Auditory Physiology and Perception—Proceedings of the 9th International Symposium on Hearing*, Vol. 83, pp. 429–446.

Preminger, J., and Tasell, D. (**1995**). "Quantifying the relation between speech quality and speech intelligibility," J. Speech Lang. Hear. Res. **38**, 714.

Quackenbush, S. R., Barnwell, T. P., and Clements, M. A. (**1988**). *Objective Measures of Speech Quality* (Prentice-Hall, Englewood Cliffs, NJ), pp. 1–377.

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.

Shannon, R., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303.

Sheskin, D. J. (**2004**). *Handbook of Parametric and Nonparametric Statistical Procedures*, 3rd ed. (Chapman & Hall/CRC, Boca Raton, FL).

Steeneken, H. J. M., 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**). "On predicting the difference in intelligibility before and after single-channel noise reduction," in *International Workshop on Acoustic Echo and Noise Control* (Tel Aviv, Israel).

Taal, C. H., Hendriks, R. C., Heusdens, R., Jensen, J., and Kjems, U. (**2009**). "An evaluation of objective quality measures for speech intelligibility prediction," in *Proceedings of Interspeech*, pp. 1947–1950.

Tribolet, J. M., Noll, P., McDermott, B. J., and Crochiere, R. E. (**1978**). "A study of complexity and quality of speech waveform coders," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 586–590.

van de Par, S., Kohlrausch, A., Heusdens, R., Jensen, J., and Jensen, S. (**2005**). "A perceptual model for sinusoidal audio coding based on spectral integration," EURASIP J. Appl. Sign. Process. **2005**, 1292–1304.
Wagener, K., Josvassen, J. L., and Ardenkjaer, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.
Wang, D. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Springer, New York), pp. 181–197.
Yamada, T., Kumakura, M., and Kitawaki, N. (**2006**). "Word intelligibility estimation of noise-reduced speech," in *Proceeding of Interspeech* (ISCA), pp. 169–172.